O método de ponderação bayesiana de modelos para seleção de modelos

Zheng Zhangzhe

Dissertação apresentada ao
Instituto de Matemática e Estatística
da Universidade de São Paulo
para obtenção do título de
Mestre em Ciências

Programa: Estatística
Orientador: Prof. Dr. Luís Gustavo Esteves

São Paulo
Janeiro de 2023

O método de ponderação bayesiana de modelos para seleção de modelos

Zheng Zhangzhe

Esta é a versão original da dissertação
elaborada pelo candidato Zheng Zhangzhe,
tal como submetida à Comissão Julgadora.

Agradecimentos

Agradeço o meu orientador Luís Gustavo Esteves por todas as ajudas, foi sempre como amigos e me orienta com paciência. Agradecimentos infinitos aos meus pais, me deram os suportes a distância, sem eles não teria sido a concluir esse trabalho.

Resumo

Zheng Zhangzhe. **O método de ponderação bayesiana de modelos para seleção de modelos**. Dissertação (Mestrado). Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2023.

Nas pesquisas em geral, as pessoas comumente propõem um único modelo na seleção de variáveis explicativas e assumem que é o modelo final, mas isso ignora tanto a incerteza do modelo quanto em estimativas de coeficientes. Todos os modelos estatísticos tradicionais têm esse tipo de problema de "incerteza". O Bayesian Model Averaging (BMA) é um método que tem uma longa história de desenvolvimento teórico e aplicação que visa explicar diretamente a incerteza de seleção do modelo. O BMA não seleciona diretamente um único modelo final dentre os disponíveis, mas calcula uma média ponderada dos modelos possíveis baseada nas probabilidades a posteriori de tais modelos. O objetivo deste estudo é revisar o BMA e algumas de suas propriedades e aplicá-lo em alguns exemplos reais. Os resultados mostram que o BMA tem um efeito melhor do que o modelo tradicional de seleção de variáveis e melhores resultados de previsão.

Palavras-chave: "Bayesian model averaging". "incerteza do modelo". "fator de Bayes". "validação cruzada". "predição".

Abstract

Zheng Zhangzhe. **The bayesian model averaging method for model selection**. Thesis (Master's). Institute of Mathematics and Statistics, University of São Paulo, São Paulo, 2023.

In general research, people commonly propose a single model in selecting explanatory variables and assume it is the final model, but this ignores both model uncertainty and coefficient estimates. All traditional statistical models have this kind of "uncertainty" problem. Bayesian model averaging (BMA) is a method that has a long history of theoretical development and application that aims to directly deals with model selection uncertainty. BMA does not directly select a single final model from those available, but calculates a weighted average of the possible models based on the posteriori probabilities of that models. The aim of this study is to review the BMA and some of its properties and apply it to some real examples. The results show that the BMA has a better performance than the traditional variable selection methods and better forecasting results.

Keywords: "Bayesian model averaging". "model uncertainty". "Bayes factor". "cross validation". "prediction".

Lista de Figuras

Lista de Tabelas

Sumário

Apêndices

Capítulo 1

Introdução

1.1 Introdução

O "trade-off" entre viés e variância é um dos problemas relevantes na literatura de modelos estatísticos das últimas décadas. Ou seja, se um modelo for pouco especificado - i.e., se variáveis importantes não estiverem consideradas - então haverá um viés nas estimativas dos parâmetros. Por outro lado, se houver uma superespecificação - i.e., se variáveis irrelevantes forem consideradas - então suas predições podem ser prejudicadas (Kaplan, 2021) . Para solucionar este tipo de problema na seleção de modelos, existem os procedimentos baseados em regularização, que fornecem um melhor equilíbrio neste trade-off, como a regressão ridge (Hoerl e Kennard, 1970), Lasso (Tibshirani, 1996) e elastic net (Zou e Hastie, 2005). Mas esses métodos levam a um único modelo final que muitas vezes é o modelo preconcebido. Em geral, o problema de incerteza acerca do modelo é desconsiderado, o que pode levar a uma decisão errada na conclusão e a um grande risco. Alguns autores como Hoeting *et al.* (1999), Leamer (1978) e D. Draper *et al.* (1987) sugerem que é arriscado se contentar com um único modelo, em vez disso, a combinação dos modelos pode levar a um resultado mais confiável.

A estrutura bayesiana admite incerteza sobre os parâmetros e, desse modo, permite abordar o problema de seleção de modelos incorporando incerteza sobre tais quantidades. Um desses métodos, chamado Bayesian Model Averaging (BMA), será estudado nesse trabalho (Kaplan, 2021).

A proposta do BMA é usar pesos para realizar uma média ponderada de possíveis modelos. Os pesos são as probabilidades a posteriori dos modelos disponíveis. Teoricamente, ele não apenas utiliza todos os modelos possíveis, mas também usa a probabilidade a posteriori do modelo como critério para avaliar a qualidade do modelo, de modo que o problema da incerteza do modelo seja resolvido e o efeito de previsão seja mais preciso do que outros métodos de combinação de modelos.

O objetivo deste estudo é revisar o BMA, descrever suas propriedades, aplicá-lo em alguns exemplos reais e fazer a comparação com o método tradicional de seleção de variáveis e predição.

1.2 Organização do trabalho

No capítulo 2 desse trabalho é apresentado a estrutura do BMA e a vantagem de seu uso em problemas de seleção. A seguir, são apresentadas as teorias de avaliação de resultado, e também a relação do fator de Bayes com o método que diminui o custo computacional. Já no capítulo 3, é apresenta a escolha da distribuição a priori dos parâmetros e modelos no BMA e o método para determinar a probabilidade a posteriori do modelo. No capítulo 4, são apresentados os resultados usando o BMA tanto nos dados simulados quanto nos dados reais. No capítulo 5, são apresentados as conclusões e perspectivas para trabalhos futuros. Os detalhes das de cálculos são apresentados no Apêndice A.

Capítulo 2

O método BMA

A combinação de modelos na literatura estatística foi mencionada pela primeira vez por Barnard (1963). Roberts (1965) propôs uma média ponderada das distribuições a posteriori de dois modelos, semelhante ao BMA. Leamer (1978) estendeu essa ideia e propôs o conceitos básicos do BMA. Ele também destaca que o Bayesian Model Averaging explica a ideia básica da incerteza envolvida na escolha de um modelo. Kass e A. E. Raftery (1995) revisou o BMA e o custo de ignorar a incerteza do modelo.

O BMA é uma aplicação da inferência bayesiana na seleção de modelos, combinação de estimativas e problemas de previsão. Ela é uma extensão do método geral de inferência bayesiana, pois permite modelar a distribuição a priori do modelo, e também utilizar o teorema de Bayes para obter a probabilidade a posteriori do modelo e a distribuição a posteriori dos parâmetros. O método é baseado no cálculo da ponderação do modelo utilizando as probabilidades a posteriori dos modelos disponíveis e a distribuição priori de cada modelo.

Neste capítulo, apresentamos as incertezas do modelo e o motivo do uso do BMA. Em seguida são apresentadas a formalização do BMA e as teorias de avaliação de resultado, e também a relação do fator de Bayes com o método que diminui o custo computacional.

2.1 Incerteza de modelo estatístico e parâmetros

Os modelos lineares são comumente usados em pesquisas. A discussão sobre a incerteza dos modelos estatísticos nesta parte é baseada principalmente em modelos lineares. Especificamente, um modelo linear pode ser expresso da forma:

$$y = f(x) + \epsilon$$

,

em que y é o vetor de variável resposta, x é a matriz de variáveis explicativas independentes, $f(x)$ é uma função que mede a relação entre x e y, e ϵ representa um vetor de erros aleatórios. Nesta expressão, estamos focados mais em $f(x)$. Por exemplo, no modelo

linear geral, $f(x)$ assume a forma de uma combinação linear mais simples da variáveis explicativas, ou seja, $f(x)$ corresponde ao produto de x e o seu vetor de coeficientes β. Em outros modelos lineares generalizados, $f(x)$ pode ser algum tipo de transformação (por exemplo, transformação logística). A incerteza do parâmetro do modelo acima depende de ϵ. Determinamos a distribuição condicional da variável y via assumindo uma distribuição para ϵ, a partir da qual podemos estabelecer o intervalo de variação do coeficiente estimado, ou seja, o intervalo de confiança. A incerteza do modelo vem de $f(\cdot)$. Por exemplo, se utilizamos os termos quadráticos, cúbicos e outros termos de variáveis no modelo linear, a relação entre y e x pode ser apresentada em várias formas de modelo como linear, parabólico e dentre outros.

No modelo acima, a incerteza do parâmetro é geralmente expressa pelo erro padrão, pois não podemos obter os coeficientes exatos de parâmetros, só estimativas. No entanto, a forma de como utilizar os erros padrão depende do objetivo do estudo. Para a maioria dos estudos, os erros padrão servem principalmente para construção de intervalos de confiança e testes de hipóteses. Por exemplo, ao observar se o valor zero está dentro do intervalo de confiança, o pesquisador pode observar a significância estatística do coeficiente de regressão ao nível α definido. É certo que se o pesquisador prestar atenção apenas ao efeito do tratamento de uma determinada variável (ou seja, se é significativamente diferente de zero), essa operação é apropriada. No entanto, se o objetivo do pesquisador não é testar a hipótese, mas usar um modelo estatístico para estimar o valor de coeficiente da variável, o erro padrão pode dizer ao pesquisador não se ela é estatisticamente significativa, mas os fatores previstos com base neste modelo, a faixa de variação que o valor da variável pode refletir. Por exemplo, suponha que estimamos um modelo de regressão linear simples $E(y) = \beta x$, onde a estimativa pontual do coeficiente de regressão β é 0,6 e o intervalo de confiança de 95% é 0,4-0,7. Neste caso, se usarmos x para prever y, em geral, podemos pensar que o valor esperado de y está entre $0{,}4x$ e $0{,}7x$. Em outras palavras, quando usamos modelos estatísticos para fazer previsões, o valor previsto de y varia à incerteza do próprio parâmetro β.

Diferente da incerteza do parâmetro, a incerteza do modelo enfatiza mais nas formas do próprio modelo. A diversidade de formas de modelo é comum em diferentes pesquisas, pois muitos pesquisadores tendem a ajustar vários modelos estatísticos no processo de análise de dados. Por exemplo, ao analisar os retornos econômicos da educação, um pesquisador pode tentar inserir diferentes variáveis explicativas, resultando em diferentes modelos estatísticos (por exemplo, um modelo incorpora a variável residência, enquanto o outro não). Essa prática é muito comum nos estudos quando os resultados da análise estatística são reportados. O leitor pode ver apenas um dos muitos modelos alternativos, ou seja, um único modelo ótimo é selecionado pelo pesquisador de propósito segundo algum critério específico. Assim, outros modelos alternativos são ignorados. Essa incerteza do modelo pode levar ao que o economista Leamer (1983) chama de "o problema do horizonte". Leamer acredita que os pesquisadores de ciências sociais devem garantir um horizonte extremamente amplo para reconhecer e demonstrar a complexidade e incerteza no processo de ajuste de modelos estatísticos. Caso contrário, a pesquisa consiste inevitavelmente num modelo mais desejado baseado em dados. Como resultado, as suas conclusões de pesquisa podem perder a credibilidade. Por exemplo, Young (2009) reanalisou o estudo de religião e relações econômicas de Barro e McCleary (2003) com as variáveis 'Church attendance'

e 'belief in hell' e descobriu que seu processo de ajuste de modelo teve apenas uma pequena mudança, mas sua conclusão não se sustentou mais, isto é, a variável 'Church attendance' não tem impacto negativo e a variável 'belief in hell' não tem impacto positivo para o crescimento econômico. Além disso, os resultados são inconsistentes ao longo do tempo, e a relação entre religiosidade e crescimento econômico não é válida no ocidente. Outro exemplo semelhante na área da economia: Magnus e Morgan (1999) convidaram diferentes pesquisadores a usar modelos estatísticos para estimar a demanda do cliente por um produto ao mesmo tempo. Os pequisadores chegaram a conclusões diferentes devido as diferenças nos modelos. Todos esses estudos mostram que de fato existe uma incerteza do modelo na pesquisa em ciências.

Como esse problema é comum, como mostrar claramente os múltiplos modelos alternativos aos pesquisadores e como escolher os modelos alternativos tornou-se uma das tarefas importantes das pesquisas em ciências. O método mais utilizado é o BMA (Kaplan, 2021). Esse método nasceu na área estatística (Leamer, 1978), e foi adotado pela economia (Wright, 2008; Horvath, 2011; Arin e Braunfels, 2018), ciência política (Montgomery e Nyhan, 2010) e medicina (Viallefont *et al.*, 2001; Genell *et al.*, 2010), dentre outras áreas.

Nos próximas seções e capítulos, vamos apresentar o método BMA e as vantagens do BMA.

2.2 Teoria

Suponha que M=$\{M_1...M_Q\}$ é o conjunto de Q modelos alternativos, D é o conjunto de dados e θ_j é o vetor de parâmetros associado o modelo M_j. Sejam L(D|θ_j, M_j) a função de verossimilhança para o parâmetro θ_j gerada pelo conjunto de dados D, e P($\theta_j|M_j$) a priori do parâmetro θ_j do modelo M_j, j=1,...,Q.

Com o teorema de Bayes, obtemos a distribuição de probabilidade a posteriori do parâmetro θ_j no modelo M_j, j=1,...,Q:

$$P(\theta_j|D, M_j) = \frac{L(D|\theta_j, M_j)P(\theta_j|M_j)}{\int L(D|\theta_j, M_j)P(\theta_j|M_j)d\theta_j}$$

A função de verossimilhança marginal P(D|M_j) sob o modelo M_j é calculado por:

$$P(D|M_j) = \int P(D, \theta_j|M_j)d\theta_j = \int L(D|\theta_j, M_j)P(\theta_j|M_j)d\theta_j$$

Agora, consideramos que cada modelo tem a probabilidade a priori P(M_j), j = 1,2,...,Q, com P(M_j) ≥ 0 e $\sum_{j=1}^{Q} P(M_j) = 1$, usando o teorema de Bayes, segue que:

$$P(M_j|D) = \frac{P(D|M_j)P(M_j)}{P(D)}$$

Assim, temos:

$$P(M_j|D) = \frac{P(D|M_j)P(M_j)}{\sum_{u=1}^{Q} P(D|M_u)P(M_u)}$$

Se y é a variável resposta de interesse de n indivíduos que queremos tratar, isto é, $y = (y_1, y_2, ..., y_n)$, temo que a esperança e a variância de y que podem ser escrita como (Hoeting *et al.*, 1999):

$$E(y|D) = \sum_{j=1}^{Q} E(y|D, M_j)P(M_j|D) = \sum_{j=1}^{Q} \hat{y}_j P(M_j|D),$$

onde $\hat{y}_j = E(y|D, M_j)$, e

$$\begin{aligned} Var(y|D) &= E(y^2|D) - E(y|D)^2 \\ &= \sum_{j=1}^{Q} E(y^2|D, M_j)P(M_j|D) - E(y|D)^2 \\ &= \sum_{j=1}^{Q} (Var(y|D, M_j) + E(y|D, M_j)^2)P(M_j|D) - E(y|D)^2 \end{aligned}$$

Desta forma, podemos estimar o n+1 ésima variável resposta y_{n+1} da seguinte forma:

$$\hat{y}_{n+1} = \sum_{j=1}^{Q} E(y_{n+1}|D, M_j)P(M_j|D)$$

em que $E(y_{n+1}|D, M_j)$ é a estimativa do y_{n+1} sob o M_j, e $P(M_j|D)$ é a probabilidade a posteriori do modelo M_j dado o conjunto de dados D. Note que $P(M_j|D)$ pode ser vista como o peso da estimativa de y_{n+1} a posteriori sob o modelo M_j na expressão de $\hat{y}$.

No caso de regressão linear, seja o conjunto $A_i \subseteq M$ em que $A_i = \{M_j : j = 1, ..., Q; \beta_i \neq 0\}$, e o β_i é o i-ésimo coeficiente.

A probabilidade a posteriori de parâmetro β_i é a soma das probabilidade a posteriori dos modelos que incluem a variável associada a β_i, ou seja:

$$P(\beta_i \neq 0|D) = \sum_{A_i} P(M_j|D)$$

em que A_i é o conjunto de modelos nos quais $\beta_i \neq 0$.

Por outro lado, a distribuição a posteriori do parâmetro β_i dado que $\beta_i \neq 0$ é dada por

$$P(\beta_i|D, \beta_i \neq 0) = \frac{\sum_{A_i} P(\beta_i|D, M_j)P(M_j|D)}{P(\beta_i \neq 0|D)}$$

Podemos reescrever a formula acima em

$$P(\beta_i|D, \beta_i \neq 0) = \sum_{A_i} P(\beta_i|D, M_j)P^{'}(M_j|D),$$

em que $P^{'}(M_j|D)$ é a probabilidade a posteriori do j ésimo modelo condicional ao evento $\beta_i \neq 0$. Note que

$$P^{'}(M_j|D) = \frac{P(M_j|D)}{P(\beta_i \neq 0|D)}$$

Então, a esperança e a variância de β_i:

$$E(\beta_i|D, \beta_i \neq 0) = \sum_{A_i} E(\beta_{i(j)})P^{'}(M_j|D) \approx \sum_{A_i} \hat{\beta}_{i(j)}P^{'}(M_j|D)$$

$$\begin{aligned} Var(\beta_i|D, \beta_i \neq 0) &= \sum_{A_i}[Var(\beta_{i(j)}) + E(\beta_{i(j)})^2]P^{'}(M_j|D) - E(\beta_i|D, \beta_i \neq 0)^2 \\ &\approx \sum_{A_i}[\sigma^2_{\beta_{i(j)}} + \hat{\beta}^{\,2}_{i(j)}]P^{'}(M_j|D) - E(\beta_i|D, \beta_i \neq 0)^2 \end{aligned}$$

em que $\hat{\beta}_{i(j)}$ e $\sigma^2_{\beta_{i(j)}}$ são os estimadores de máxima verossimilhança e a variância de β_i em M_j (A. E. Raftery, 1995).

A seguir, exemplificamos o uso das expressões acima considerando que a probabilidade a priori do modelo uniforme, isto é, $P(M_j) = \frac{1}{Q}$ para j = 1,2,...,Q.

Exemplo: Suponha que temos um conjunto de dados D de n indivíduos com 2 variáveis explicativa $\{x_1, x_2\}$ e uma variável resposta y, onde $x_i = (x_{1,i}, x_{2,i}, ..., x_{n,i})$ para i=1,2 e $y = (y_1, y_2, ...y_n)$. Assim, temos $2^2 = 4$ modelos candidatos, que são:

$$\begin{aligned} M_1 &: y = \beta_0 + \epsilon \\ M_2 &: y = \beta_0 + \beta_1 x_1 + \epsilon \\ M_3 &: y = \beta_0 + \beta_2 x_2 + \epsilon \\ M_4 &: y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon \end{aligned}$$

Suponhamos que

$$\begin{aligned} P(D|M_1) &= 0,1 \\ P(D|M_2) &= 0,2 \\ P(D|M_3) &= 0,3 \\ P(D|M_4) &= 0,4 \end{aligned}$$

e que

$$P(M_1) = P(M_2) = P(M_3) = P(M_4) = \frac{1}{4}$$

Assim, temos que

$$P(M_1|D) = 0,1$$
$$P(M_2|D) = 0,2$$
$$P(M_3|D) = 0,3$$
$$P(M_4|D) = 0,4$$

Assim, para uma nova observação y_{n+1} com covariáveis $\{x_1 = x_{n+1,1}, x_2 = x_{n+1,2}\}$, temos a predição de y_{n+1} da nova observação:

$$\hat{y}_{n+1} = \sum_{j=1}^{4} \hat{y}_{M_j} P(M_j|D)$$

$$= \hat{\beta}_{0M_1} * 0,1 + (\hat{\beta}_{0M_2} + \hat{\beta}_{1M_2} x_{n+1,1}) * 0,2 + (\hat{\beta}_{0M_3} + \hat{\beta}_{2M_3} x_{n+1,2}) * 0,3 + (\hat{\beta}_{0M_4} + \hat{\beta}_{1M_4} x_{n+1,1} + \hat{\beta}_{2M_4} x_{n+1,2}) * 0,4$$

em que $\hat{\beta}_{iM_j}$ representa os estimadores de máxima verossimilhança de β_i em j-ésimo modelo para i = 0,1,2 e j = 1,...,4.

2.3 Otimalidades

Na prática, podem existir infinitas soluções para tratar um problema de modelagem. Por isso, precisamos de uma forma de avaliar a otimalidade do BMA.

Comparando com um modelo único, uma medida de avaliação do desempenho de modelos é o log predictive score do scoring rule (A. Raftery e Zheng, 2003). Consideremos que uma função S: $\triangle Y \bullet Y \longrightarrow R$ ($\bullet$ representa um produto cartesiano), em que Y é uma variável aleatória e $\triangle Y$ é o conjunto de distribuições de probabilidade para Y, para uma distribuição $p \in \triangle Y$, temos o log predictive score dado por $S(p, y) = logp(y)$.

Segundo as propriedades da divergência de informação Kullback-Leibler, temos o critério de scoring logarítmica (Good, 1952) (a prova da formula será apresentado em Apêndice A):

$$E[log(\sum_{Q} P(y|M_j, D)P(M_j|D))] \geq E[log(P(y|M_i, D))],$$

para i = 1,...,Q e $\sum_{Q} P(y|M_j, D)P(M_j|D)$ é a distribuição preditiva sob o BMA.

Assim, temos a conclusão que o log predictive score do BMA é melhor que de qualquer modelo único M_i (A. E. Raftery *et al.*, 1997; A. Raftery e Zheng, 2003).

Outro critério segundo o que o BMA é ótimo é usando a métrica baseada nas diferenças entre predição e valor real da variável resposta. Como o objetivo geral de modelagem no problema de predição é minimizar os erros, uma função comum para a avaliação é a função de perda quadrática negativa. Suponha que u é uma função que associa às quantidades a e b o valor:

$$u(a, b) = -(a - b)^2$$

Em inferência Bayesiana, se a variável resposta y tem uma distribuição de probabilidade a posteriori $P(\cdot|D)$ e se w é uma possível predição de y, então a melhor solução e a predição dela w, a melhor solução w^* é o w que maximiza a função:

$$w^* = \underset{w \in W}{argsup} \int_{\Omega} u(y, w) p(y|D) dw$$

Com a função de perda quadrática negativa, temos $w^* = E(y|D)$ que é a esperança de y dado um conjunto de dados D. Por outro lado,

$$E(y|D) = \sum_{j=1}^{Q} E(y|M_j, D) p(M_j|D) = \sum_{j=1}^{Q} \hat{y}_j p(M_j|D)$$

onde $\hat{y}_j$ é a esperança de y sob modelo M_j, ou seja, a melhor solução w^* pode ser escrita como $w^* = \sum_{j=1}^{Q} \hat{y}_j p(M_j|D)$, que é a previsão de y sob o BMA. Isto é, o BMA oferece a mesma resposta que uma decisão Bayesiana sob perda quadrática (Bernardo e A. F. M. Smith, 1994; Clyde e Iversen, 2013).

2.4 Fator de Bayes

O fator de Bayes é uma medida bayesiana para a seleção de modelo. Uma vantagem do fator de Bayes é que os modelos não precisam ser aninhados. O Fator de Bayes oferece uma medida de relevância entre os modelos diferentes (Kass e A. E. Raftery, 1995). Suponha que temos 2 modelos canditados, M_i e M_j, a medida para apoiar o melhor dos modelos entre si usando o fator de Bayes pode ser definido como a razão das probabilidades à posteriori:

$$B_{i,j} = \frac{p(M_i|D)}{p(M_j|D)} = \frac{p(D|M_i)}{p(D|M_j)} \frac{p(M_i)}{p(M_j)}$$

No caso em que as probabilidades à priori são uniformes, temos que o fator pode ser resumido por:

$$B_{i,j} = \frac{p(D|M_i)}{p(D|M_j)}$$

Com essa medida, é possível fazer a comparação entre os dois modelos i e j. Por exemplo, quando $B_{i,j} = 10$, uma possível interpretação é a de que o modelo M_i é dez vezes mais relevante do que M_j (Wasserman, 2000).

Considere novamente os Q possíveis modelos M_j, j=1, ...Q, fixando o primeiro modelo M_1, é possível também obter a probabilidade a posteriori do modelo via seguinte formula:

$$p(M_j|D) = \frac{B_{j,1}p(M_j)}{\sum_{l=1}^{Q} B_{l,1}p(M_l)}$$

Continuando o exemplo de 2.2, temos:

$$B_{1,1} = \frac{p(D|M_1)}{p(D|M_1)} = 1$$
$$B_{2,1} = \frac{p(D|M_2)}{p(D|M_1)} = 2$$
$$B_{3,1} = \frac{p(D|M_3)}{p(D|M_1)} = 3$$
$$B_{4,1} = \frac{p(D|M_4)}{p(D|M_1)} = 4$$

e então:

$$P(M_1|D) = \frac{1}{10} = 0,1$$
$$P(M_2|D) = \frac{2}{10} = 0,2$$
$$P(M_3|D) = \frac{3}{10} = 0,3$$
$$P(M_4|D) = \frac{4}{10} = 0,4$$

Ou seja, nesse exemplo, é possível obter a probabilidade a posteriori de cada modelo usando o fator de Bayes.

2.5 Consideração do custo computacional

Pensando o custo computacional, o BMA pode ter dois problemas na implantação (Hoeting *et al.*, 1999):

1. A complexidade de calcular a função integral de $P(D|M_j)$;

2. Número de modelos na prática;

Para resolver o problema 1, pode-se usar o método de Laplace, que tem sido efetivamente usado para calcular fator de Bayes (Kass e A. E. Raftery, 1995). Em alguns casos, tal procedimento é equivalente uma aproximadação via BIC (Tierney e Kadane, 1986; A. E. Raftery, 1995).

Com relação o problema 2, o número de modelos candidatos pode ser bem grande na prática. Por exemplo, se existem p variáveis, então temos 2^p modelos. Uma solução interessante que pode resolver o problema será critério de Occam's window.

A ideia de Occams windows foi sugerido pelo Madigan e A. E. Raftery (1994). O objetivo deles é reduzir a quantidade de modelos candidatos no BMA. Por exemplo, em um estudo de A. E. Raftery (1995), o número de modelos candidatos foi reduzido de 32768 para 14 modelos de acordo com a Occams windows. A ideia dele pode ser escrita da seguinte forma:

1. Os modelos com uma probabilidade a posterior muito menor (por exemplo, C vezes menor, onde C é determinado pelo pesquisador) do que o modelo mais provável, eles são eliminados. Seja

$$A^{'} = \{M_j : \frac{max_l\{p(M_l|D)\}}{p(M_j|D)} \leq C\},$$

o conjunto de modelos não eliminados segundo o critério apresentado em 1.

2. Se o modelo reduzido (com menos variáveis explicativas) tiver uma probabilidade a posterior maior, o modelo complexo (com mais variáveis explicativas mas contém as variáveis explicativas do modelo reduzido) que include as variáveis do modelo reduzido é eliminado. Seja

$$B = \{M_j : \exists M_l \in A^{'}, M_l \subset M_j, \frac{p(M_l|D)}{p(M_j|D)} > 1\},$$

o conjunto de modelos não eliminados segundo o critério apresentado em 2.

Finalmente, obtemos o conjunto $A = A^{'} \backslash B$ de possíveis modelos. Ou seja, os modelos considerados para BMA são aqueles que estão em $A^{'}$ mas não em B. A formula do BMA pode ser simplificado como:

$$p(\tilde{y}|D,A) = \sum_{M_j \in A} p(\tilde{y}|M_j,D)p(M_j|D,A)$$

onde $\tilde{y}$ é a variável de interesse, e a $p(M_j|D,A)$ é a probabilidade a posteriori do j ésimo modelo no conjunto A.

2.6 Avaliação via do método cross validation

Na modelagem de aprendizado de máquina, é prática comum dividir os dados em conjuntos de treinamento e teste. O conjunto de teste é independente do treinamento e não participa do treinamento, sendo usado para a avaliação do modelo final. No Capitulo 4, vamos usar esse método para calcular a medida de desempenho em avaliação de predição.

No processo de treinamento, o problema de overfitting geralmente ocorre, ou seja, o modelo pode corresponder bem com os dados de treinamento, mas não pode prever bem os dados fora do conjunto de treinamento. Se os dados de teste forem usados para ajustar os parâmetros do modelo neste momento, é equivalente a conhecer algumas informações dos dados de teste durante o treinamento, o que afetará a precisão do resultado final da avaliação. A prática usual é dividir uma parte dos dados de treinamento como dados de validação para avaliar o efeito de treinamento do modelo.

Os dados de validação são obtidos dos dados de treinamento, mas não participam do treinamento, de modo que a validação do modelo com os dados fora do conjunto de treinamento pode ser de forma mais relativamente objetiva.

A cross validation também conhecida como validation loop, foi mencionada em Stone (1974) e Hastie *et al.* (2009), o método divide os dados originais em k grupos (k-Fold), usa cada dado de um subconjunto como um conjunto de validação e os dados dos k-1 grupos restantes de subconjunto como um conjunto de treinamento, para que os modelos k sejam obtidos. Esses modelos k são avaliados no conjunto de validação respectivamente, e o erro final MSE (Mean Squared Error) é a média do erro de cross validation.

No k-Fold, divida todo o conjunto de treinamento S em k subconjuntos disjuntos. Assumindo que o número de observação de treinamento em S é m, então cada subconjunto tem $\frac{m}{k}$ observações de treinamento e o subconjunto correspondente é $\{S_1, S_2, ..., S_k\}$. Cada vez treinamos com o tipo de modelo M_i com k-1 subconjuntos $\{S_1, S_2, ..., S_{j-1}, S_{j+1}, ..., S_k\}$ e deixamos o subconjunto S_j em teste, assim, obtemos uma função $h_{i,j}$ e para cada subconjunto S_j, temos um erro $\hat{\epsilon}_{S_j}(h_{i,j})$, ou seja, o erro final para o tipo de modelo M_i (Stone, 1974; Hastie *et al.*, 2009) seria:

$$\hat{\epsilon}_{final} = \frac{1}{k}\sum_{j=1}^{k} \hat{\epsilon}_{S_j}(h_{i,j})$$

Baseado nisso, podemos escolher o melhor tipo de modelo M_i^*, e treinamos com o conjunto de dados inteiro S. No final, obtemos a melhor função estimada h_i^*.

Capítulo 3

Distribuição a priori para parâmetros e modelos

Neste capitulo, serão apresentadas as opções de distribuição a priori para os parâmetros do modelo e para o próprio modelo de regressão.

Existe duas partes da distribuição a priori para o método BMA. A especificação da priori consiste em duas etapas: uma é a seleção de distribuição de priori de parâmetros em cada modelo, e outra a é a seleção de distribuição de priori de cada modelo. Uma dificuldade de implantação de BMA é que precisamos especificar todas as prioris de cada modelo.

3.1 Distribuição a priori de modelo

Sobre a parte de priori de modelos, hoje os mais comuns seriam a priori uniforme, a priori binomial e a priori Beta-binomial, detalhadas a seguir.

1. Priori Uniforme

A escolha de distribuição a priori de modelo bayesiano sempre foi um desafio. A escolha mais simples é a priori uniforme. Caso existam p variáveis explicativas, a priori uniforme especifica probabilidade 2^{-p} para cada um dos 2^p modelos, ou seja, cada modelo tem a mesma probabilidade de priori.

2. Priori Bernoulli (Hoeting *et al.*, 1999)

A suposição de priori uniforme para todos os modelos é uma escolha "neutra" quando há poucas informações de antemão. Há outra opção que generaliza a priori uniforme: a priori Bernoulli. Essa opção foi usado na seleção de variáveis em regressão linear em George e McCulloch (1993) e foi sugerido para BMA em modelos Cox.

$$P(M_j) = \prod_i^p \pi_i^{\gamma_{ji}}(1-\pi_i)^{1-\gamma_{ji}}, \pi_i \in [0,1], \forall i = 1, ..., p$$

em que γ_{ji} é 1 se a variável j fica dentro o modelo M_j, e caso contrário, γ_{ji} é 0. Um caso especial de distribuição a priori Bernoulli é quando $\pi_i = 0,5$ para i = 1,..,p, que produz a

priori uniforme como em 1. Segundo Hoeting *et al.* (1999), é comum assumir $\pi_i < 0,5$ para todos j, assim gera uma penalidade na probabilidade a priori para os modelos que têm muitas variáveis explicativas.

No exemplo do capítulo 2.2 para p = 2, nesse caso, temos Q = 2^2 = 4 modelos, e temos as seguintes probabilidades a priori:

$$P(M_1) = \prod_i^2 \pi_i^{\gamma_{1i}}(1-\pi_i)^{1-\gamma_{1i}} = (1-\pi_1)(1-\pi_2)$$
$$P(M_2) = \prod_i^2 \pi_i^{\gamma_{2i}}(1-\pi_i)^{1-\gamma_{2i}} = \pi_1(1-\pi_2)$$
$$P(M_3) = \prod_i^2 \pi_i^{\gamma_{3i}}(1-\pi_i)^{1-\gamma_{3i}} = (1-\pi_1)\pi_2$$
$$P(M_4) = \prod_i^2 \pi_i^{\gamma_{4i}}(1-\pi_i)^{1-\gamma_{4i}} = \pi_1\pi_2$$

Quando $\pi_i = 0,5$, temos

$$P(M_1) = P(M_2) = P(M_3) = P(M_4) = 2^{-2} = \frac{1}{4}$$

3. Priori Binomial (Zeugner e Feldkircher, 2015)

Observar que $\pi_i = \pi^*, \forall i = 1, ..., p$, temos que para o modelo M_j que tem q variáveis explicativas, a probabilidade priori seria

$$P(M_j) = \pi^{*q}(1-\pi^*)^{p-q},$$

Essa é a priori binomial para o modelo. Com a suposição de distribuição binomial, a esperança do número de variáveis explicativas q no modelo é $\pi^* p$, e se a $\pi^* = \frac{1}{2}$, a priori binomial simplifica para a priori unforme. Na prática, sugere-se $\pi^* < \frac{1}{2}$, assim os modelos com poucas variáveis explicativas podem ter uma probabilidade a priori de modelo maior que possam impactar mais no resultado final de BMA.

4. Priori Beta-binomial (Ley e M. F. Steel, 2009)

Suponha que no caso 3, π^* é uma variável aleatória que segue a distribuição Beta com parâmetros a e b, ou seja $\pi^* \sim Beta(a, b)$ com a,b > 0 . Nesse caso, temos a esperança e variança de número de variáveis explicativas $\bar{q}$:

$$E(\bar{q}) = \frac{a}{a+b}p$$
$$Var(\bar{q}) = \frac{ab(a+b+p)}{(a+b)^2(a+b+1)}p$$

3.2 Distribuição a priori de parâmetros no modelo de regressão

A g-priori de Zellner é um tipo de priori objetiva que foi desenvolvida por Zellner (1986). Ela é uma Normal-Gamma conjugada natural para coeficientes de regressão β sob o modelo de regressão linear normal.

Suponha que y é uma variável resposta, x é a matriz de covariáveis, $x^{'}$ é a matriz transposta de x, β_j é o vetor de coeficiente de variáveis explicativas no modelo M_j e ϵ é o erro aleatório, temos:

$$y = x^{'}\beta_j + \epsilon$$

onde os erros ϵ são independentes e identicamente distribuídas $N(0, \sigma^2)$

Para cada modelo M_j, a g-priori pode ser escrita como a seguinte forma (Feldkircher *et al.*, 2009):

$$\beta_j|\sigma^2, M_j, g \sim N(0, g\sigma^2(x^{'}x)^{-1}))$$

onde g é um parâmetro de escala, conhecido e positivo.

A g-priori tem uma consistência com o aumento do tamanho da amostra para descobrir assintoticamente o modelo verdadeiro e o g como um termo que penaliza mais modelos com mais variáveis do tamanho do modelo em fator de Bayes e BIC, a seguir, será apresentado. Aplicando no BMA, temos a posteriori

$$P(M_j|D, g) = \frac{P(D|M_j, g)P(M_j)}{P(D)} = \frac{P(D|M_j, g)P(M_j)}{\sum_{i=1}^{Q} P(D|M_i, g)P(M_i)},$$

e o fator de Bayes entre o modelo M_j e o modelo M_l é

$$B_{j,l} = (1+g)^{\frac{q_l - q_j}{2}}\left(\frac{1 - \frac{g}{1+g}R_j^2}{1 - \frac{g}{1+g}R_l^2}\right)^{-\frac{N-1}{2}}$$

em que q_l e q_j correspondendo o número de variáveis explicativas dos modelos M_l e M_j, $R^2_{j(l)}$ é o coeficiente de determinação do modelo $M_{j(l)}$, e $(1+g)^{\frac{q_l - q_j}{2}}$ é a função de penalidade. Note que quanto maior g, $\left(\frac{1-\frac{g}{1+g}R_j^2}{1-\frac{g}{1+g}R_l^2}\right)^{-\frac{N-1}{2}}$ fica cada vez mais afastado de 1 e as probabilidades a posteriori são mais concentradas em uns "super modelos". Os detalhes podem ser consultados em Feldkircher *et al.* (2009). A seguir são apresentadas opções de g com valor fixado:

1. Priori de informações da Unidade (UIP)

Kass e Wasserman (1995) (Wasserman, 2000) propunham a priori de informações da Unidade (UIP) para os parâmetros do modelo. Nessa priori, suponha que o g=N, onde N é o número de observações. Liang *et al.* (2008) mostra que a combinação de priori UIP e priori uniforme de modelo tem a melhor performance nos dados simulados .

2. Priori de critério de Inflação de Risco (RIC)

O critério de Inflação de Risco é um critério para a avaliação de procedimentos de

seleção de variáveis em regressão múltipla. Foster e George (1994) mostraram que a seleção do modelo com a maior probabilidade a posteriori do modelo equivale a selecionar o modelo com maior RIC quando $g = p^2$, onde p é o número de variáveis explicativas.

3. Critério de Inflação de Risco de Benchmark (BRIC)

A BRIC é uma combinação de UIP e RIC, onde $g = max(N, p^2)$. Fernández *et al.* (2001) investiga varias escolhas possíveis de g e concluí que tomar $g = max(N, p^2)$ leva a resultados razoáveis, isso é usar $g = p^2$ quando $N \leq p^2$ e $g = N$ quando $N > p^2$

Para g não fixo, temos:

1. Bayes empírica local (local EB)

O método local EB estima um g para cada modelo pela máxima verossimilhança com base nos dados observados. Assim, temos a estimativa de $\hat{g}_j = argmax(0, F_j - 1)$ com a estatística F do modelo M_j, em que $F_j = \frac{R_j^2(N-p_j-1)}{(1-R_j^2)p_j}$, R_j^2 é o coeficiente de determinação de regressão e p_j é o número de variável explicativas do modelo M_j (Liang *et al.*, 2008).

2. Bayes empírica Global (Global EB)

O método Global EB assume um g comum para todos modelos disponíveis. Ele estima uma média de g baseado na função verossimilhança nos dados observados, ou seja

$$\hat{g} = argmax \sum_j P(M_j) \frac{(1+g)^{\frac{(N-p_j-1)}{2}}}{[1+g(1-R_j^2)]^{\frac{(N-1)}{2}}}$$

3. Prioris de Hyper-g (Hyper-g priors)

As prioris de Hyper-g foram uma proposta para data-dependent shrinkage. Assume que

$$f(g) = \frac{a-2}{2}(1+g)^{-\frac{a}{2}}, g > 0.$$

Assim,

$$\frac{g}{1+g} \sim Beta(1, \frac{a}{2} - 1), a > 2$$

em que $\frac{g}{1+g}$ é chamado fator de contração(shrinkage). Quando a = 4, a distribuição a priori de fator de contração será a distribuição uniforme (Liang *et al.*, 2008; Feldkircher *et al.*, 2009).

3.3 Obtenção da estimativa da posteriori do modelo

A estimativa da probabilidade a posteriori do modelo pode ser calculado de certas formas. A seguir, vamos apresentar a estimativa via Critério de Informação Bayesiano,

e também a formula fechada da probabilidade a posteriori no caso com umas suposições.

O Critério de Informação Bayesiano (BIC) (Schwarz, 1978) considera uma função de log de penalidade que depende do tamanho da amostra e do número de parâmetros estimados no modelo. Isto é,

$$BIC(M) = -2ln(\hat{L}) + kln(n),$$

onde a $\hat{L}$ é o máximo da função de verossimilhança do modelo M e n é o tamanho da amostra e k é o número parâmetros estimados no modelo. Isto é,

A função a posteriori do modelo M pode ser aproximada (Claeskens e Hjort, 2008) como (o desenvolvimento da formula será apresentado em Apêndice A):

$$p(M|D) \approx exp(-\frac{1}{2}BIC)p(M).$$

Ou seja, para cada modelo M_j, temos:

$$p(M_j|D) \approx \frac{exp(-\frac{1}{2}BIC_j)p(M_j)}{\sum_{l=1}^{K} exp(-\frac{1}{2}BIC_l)p(M_l)}.$$

No caso de distribuição uniforme para o modelo:

$$p(M_j|D) \approx \frac{exp(-\frac{1}{2}BIC_j)}{\sum_{l=1}^{K} exp(-\frac{1}{2}BIC_l)}$$

Além da aproximação do BIC, também é possível escrever a distribuição a posteriori em uma fórmula fechada. Vamos mostrar a fórmula da probabilidade a posteriori de um modelo com suposições adicionais sobre os parâmetros com variância desconhecida.

No caso da variância desconhecida e considerando o parâmetro (β, σ^2) no modelo M_j (o desenvolvimento da formula será apresentado em Apêndice A), suponhamos que

$$\beta|\sigma^2 \sim N_p(m_0, \sigma^2 V_0)$$

e

$$\sigma^2 \sim IG(a_0, b_0),$$

ou seja, suponhamos que a distribuição a priori de (β, σ^2) é Normal-Inversa-Gamma com parâmetros (m_0, V_0, a_0, b_0)

$$(\beta, \sigma^2) \sim N_pIG(m_0, V_0, a_0, b_0),$$

Murphy (2007) e Hoyos (2020) mostram que

$$\beta, \sigma^2 | y, x \sim N_p IG(m^*, V^*, a_1, b_1),$$

em que

$$\begin{aligned} V^* &= (V_0^{-1} + X^T X)^{-1} \\ m^* &= V^*(V_0^{-1} m_0 + X^T y) \\ a_1 &= a_0 + \frac{n}{2} \\ b_1 &= b_0 + \frac{m_0^T V_0^{-1} m_0 + y^T y - m^{*T} V^{*-1} m^*}{2} \end{aligned}$$

Assim,

$$\begin{aligned} P(D|M_j) &= \int L(D|\beta, \sigma^2, M_j) P(\beta, \sigma^2 | M_j) d\beta d\sigma^2 \\ &= \frac{b_0^{a_0} \Gamma(a_1) |V^*|^{\frac{1}{2}}}{(2\pi)^{\frac{n}{2}} (b_1)^{a_1} \Gamma(a_0) |V_0|^{\frac{1}{2}}} \end{aligned}$$

Capítulo 4

Exemplos

Neste capítulo, serão apresentados os resultados de comparação de seleção de modelos e desempenho de previsão entre o BMA e os modelos tradicionais de regressão tanto nos dados simulados como para dados reais. Note que os dados reais são abertos na internet.

4.1 Exemplo com dados simulados

Nesta seção, vamos apresentar os resultados de comparação entre o BMA e o modelo regressão linear em dados simulados.

Para testar e comparar o BMA com a regressão linear clássica, estabelecemos 10 variáveis independentes de diferentes distribuições, denominadas X_1 a X_{10}. Dentre elas, X_1 a X_3 vem da distribuição normal, X_4 e X_8 vem da distribuição Poisson, X_5 vem da distribuição exponencial, X_6 vem da distribuição Beta, X_7 vem da distribuição binomial, X_9 vem da distribuição Gamma e X_{10} vem da distribuição uniforme, conforme especificado abaixo:

$$X_1 \sim N(0, 1)$$

$$X_2 \sim N(4, 2)$$

$$X_3 \sim N(0.5, 1)$$

$$X_4 \sim Pois(3)$$

$$X_5 \sim Exp(10)$$

$$X_6 \sim Beta(3, 10)$$

$$X_7 \sim B(10, 0.5)$$

$$X_8 \sim Pois(40)$$

$$X_9 \sim Gamma(10, 20)$$

$$X_{10} \sim U(0, 1)$$

Construímos a variável resposta y baseado nas variáveis X_1, X_2, X_3 acima. Isto é:

$$y = X\beta + \epsilon$$

em que $X = (1, X_1, X_2, X_3)$, $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)^T$ e $\epsilon \sim N(0, 1)$. Além disso, estabelecemos também $\beta_0 = 2, \beta_1 = -1.5, \beta_2 = 0.45, \beta_3 = -3$, ou seja, para o i-ésimo y, temos:

$$y_i = 2 - 1.5X_{1i} + 0.45X_{2i} - 3X_{3i} + \epsilon_i$$

Baseado nas suposições acima, geramos uma sequência de amostra com dados simulados para o estudo de comparação do modelo BMA com o modelo linear.

Usamos a regra padrão na probabilidade a posteriori para decidir a significância de uma variável explicativa da Tabela 4.1 (Kass e A. E. Raftery, 1995; A. E. Raftery, 1995):

Valor da posteriori	Nível de evidência
< 50%	não há
50% - 75%	fraco
75% - 95%	positivo
95% - 99%	forte
> 99%	muito forte

Tabela 4.1: *Regra padrão de nível de evidencia da probabilidade a posteriori*

Para o cálculo das probabilidade a posteriori, foi considerado uma priori uniforme para os modelos e a g-prior com g = 'UIP' (capítulo 3) para os parâmetros dos modelos. Observamos que quando n = 100 (Tabela 4.2 e Figura 4.1), a variável X_4 é considerada como variável significativa ao nível 0.05, e outras variáveis X_7 e X_9 são consideradas como variáveis significativas ao nível 0.1. No método BMA, $P(\beta_{X_4} \neq 0) = 0.7$, ou seja, não tem uma evidência tão forte para dizer que a variável X_4 é significativa e para as outras duas variáveis, X_7 e X_9, as probabilidade a posteriori de tais parâmetros serem diferentes de zero são pequenas. Ou seja, o BMA diz que X_7 e X_9 não são variáveis significativas, o que está de acordo com a simulação acima. Para outros casos, isso também acontece (Tabela 4.2 e Figura 4.1): para n=500, os p-valores de X_4, X_5 e X_7 são próximos de 0.1; no caso n=1000, o p-valor da X_9 é próximo de 0.1, e para n=2000, os p-valors de X_5 e X_7 são próximo de 0.1. As correspondentes probabilidades a posteriori do BMA se mantem como não significativas. Note que na Figura 4.1 e na Figura 4.2, as retas de corte são de p-valor = 0.1 no exio X e P(coeficiente $\neq$ 0) = 0.75 no eixo Y.

Olhamos também outros casos com n = 5000, 10000, 20000 e 50000 (Figura 4.2). O BMA mantem o mesmo resultado de não misturar as variáveis 'aleatórias' e as variáveis que participam da simulação do y. Em outras palavras, com esses dados simulados, o BMA consegue selecionar as variáveis de maneira mais satisfatória, mais próximo do modelo de fato considerado (com 3 variáveis explicativas).

Variável	p-valor	P($\beta \neq 0$)
Intercepto	0.08	1.00
X1	<0.01	1.00
X2	<0.01	1.00
X3	<0.01	1.00
X4	0.54	0.16
X5	1.00	0.12
X6	0.92	0.12
X7	0.91	0.12
X8	0.67	0.17
X9	0.33	0.26
X10	0.89	0.13

(a) *n = 50*

Variável	p-valor	P($\beta \neq 0$)
Intercepto	0.02	1.00
X1	<0.01	1.00
X2	<0.01	1.00
X3	<0.01	1.00
X4	0.02	0.70
X5	0.25	0.05
X6	0.66	0.05
X7	0.07	0.29
X8	0.33	0.10
X9	0.09	0.23
X10	0.51	0.05

(b) *n = 100*

Variável	p-valor	P($\beta \neq 0$)
Intercepto	<0.01	1.00
X1	<0.01	1.00
X2	<0.01	1.00
X3	<0.01	1.00
X4	0.34	0.08
X5	0.87	0.06
X6	0.66	0.07
X7	0.61	0.06
X8	0.50	0.07
X9	0.32	0.09
X10	0.91	0.06

(c) *n = 200*

Variável	p-valor	P($\beta \neq 0$)
Intercepto	<0.01	1.00
X1	<0.01	1.00
X2	<0.01	1.00
X3	<0.01	1.00
X4	0.12	0.14
X5	0.11	0.12
X6	0.50	0.05
X7	0.13	0.18
X8	0.16	0.12
X9	0.21	0.11
X10	0.22	0.10

(d) *n = 500*

Variável	p-valor	P($\beta \neq 0$)
Intercepto	<0.01	1.00
X1	<0.01	1.00
X2	<0.01	1.00
X3	<0.01	1.00
X4	0.53	0.04
X5	0.97	0.03
X6	0.34	0.05
X7	0.40	0.04
X8	0.44	0.04
X9	0.11	0.09
X10	0.71	0.03

(e) *n = 1000*

Variável	p-valor	P($\beta \neq 0$)
Intercepto	<0.01	1.00
X1	<0.01	1.00
X2	<0.01	1.00
X3	<0.01	1.00
X4	0.64	0.02
X5	0.10	0.07
X6	0.77	0.02
X7	0.12	0.06
X8	0.25	0.04
X9	0.62	0.02
X10	0.26	0.04

(f) *n = 2000*

Tabela 4.2: *P-valores e probabilidades a posteriori para o exemplo 4.1*

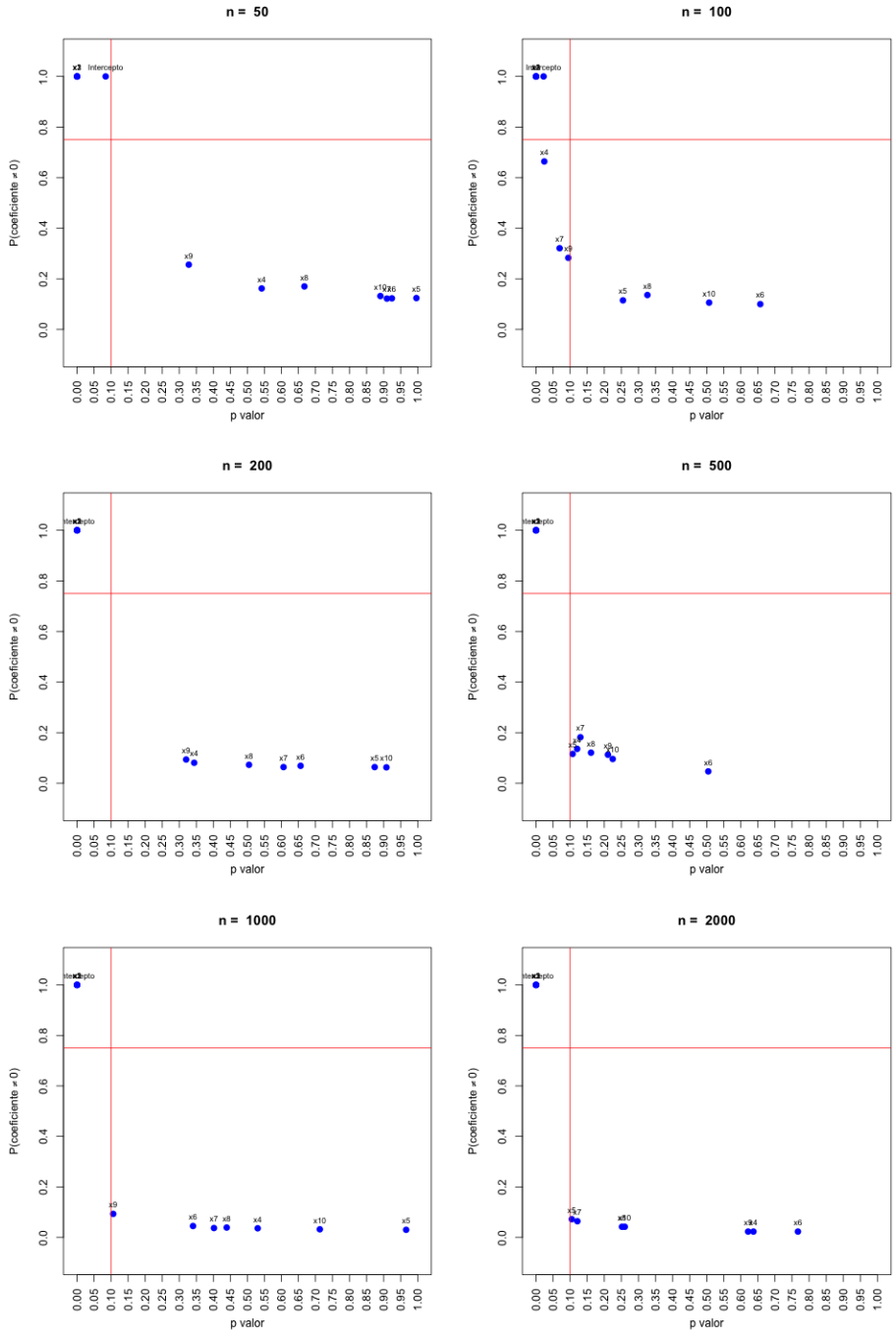

Figura 4.1: *Gráficos da probabilidade a posteriori e p-valor para o exemplo de simulação com n de 50 a 2000.*

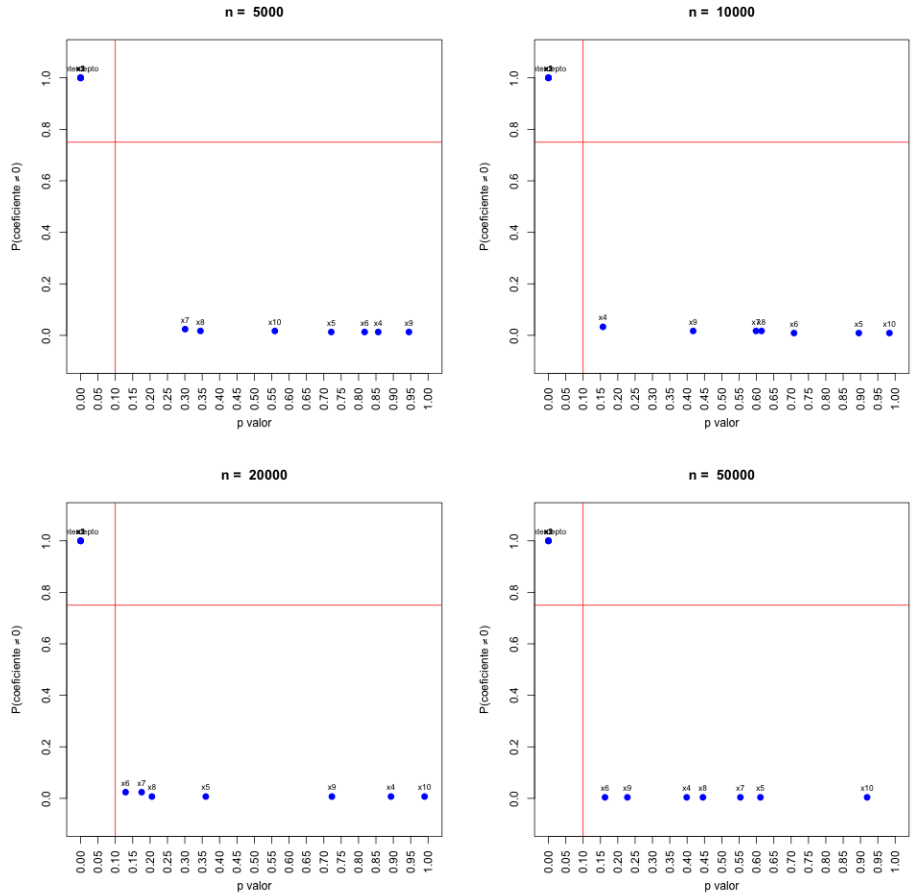

Figura 4.2: *Gráficos da probabilidade a posteriori e p-valor para o exemplo de simulação com n de 5000 a 50000.*

Além da comparação de p-valor com probabilidade a posteriori, também comparamos a predição da variável y dos 2 métodos. Usamos a cross validation com k = 5 (k-fold=5) para comparar o erro quadrado médio (EQM). Tirando o caso n = 500, o BMA sempre tem EQM menor que o modelo de regressão linear como na Tabela 4.4 e Figura 4.3. Refazendo tal comparação com k = 3, observa-se que o EQM do BMA é menor que o EQM do modelo linear. No entanto, o ganho do método BMA diminue à medido que cresce o tamanho da amostra.

Observação: testamos diferentes k de 3 a 7 nesse exemplo. No tanto para manter o padrão de k igual nos outros exemplos com dados reais, vamos mostrar somente os resultados com k = 5. Os demais são apresentados no apêndice B.

n	EQM-BMA	EQM-linear
50	0.8729	1.0632
100	0.8266	0.8377
200	0.9080	0.9570
500	1.0327	1.0281
1000	0.9817	0.9867
2000	0.9838	0.9874
5000	1.0272	1.0292
10000	1.0182	1.0192
20000	1.0022	1.0025
50000	1.0140	1.0143

Tabela 4.4: *Comparação de EQM via cross validation com k = 5*

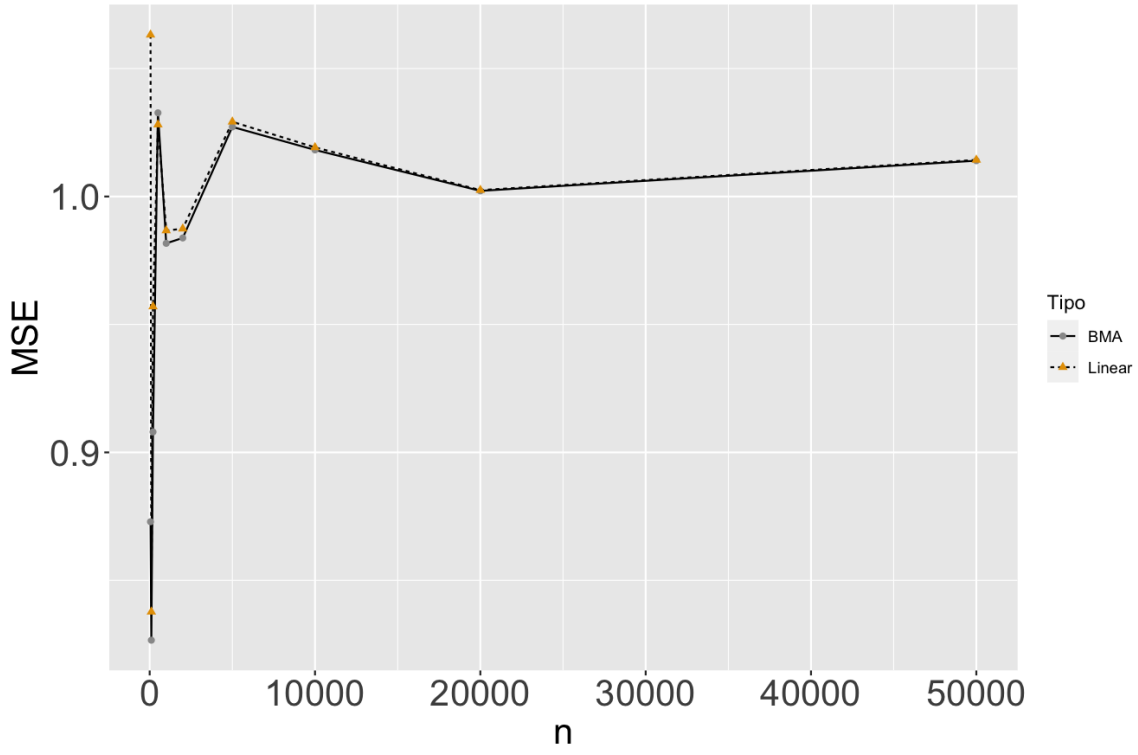

Figura 4.3: *Gráficos de EQM via cross validation com k = 5*

4.2 Análise com dados de concreto

Nesta seção, vamos apresentar os resultados de comparação entre BMA e modelo regressão linear em dados de concreto. Este conjunto de dados foi extraído do Kaggle em https://www.kaggle.com/datasets/prokaggler/concrete-data.

4.2.1 Descrição dos dados

O concreto é o material mais importante na engenharia civil. Esse estudo visou estudar a resistência do concreto em função de ingredientes que tais como cimento, escória de alto forno, cinzas volantes, água, superplastificante, agregado graúdo e agregado miúdo. O conjunto de dados de concreto contem 1030 observações e 9 variáveis, sendo que 8 são variáveis explicativas e 1 é a variável resposta, em que:

Variáveis explicativas
Cement: Cimento medido em kg em uma mistura de 1 m^3
Blast: Explosão medida em kg em uma mistura de 1 m^3
Fly ash: Cinza volante medida em kg em uma mistura de 1 m^3
Water: Água medida em kg em uma mistura de 1 m^3
Superplasticizer: Superplastificante medido em kg em uma mistura de 1 m^3
Coarse Aggregate: Agregado Grosso medido em kg em uma mistura de 1m^3
Fine Aggregate: Agregado Fino medido em kg em uma mistura de 1 m^3
Age: dia (1~365)

Variável resposta
Strength: Resistência à compressão do concreto medida em MPa

4.2.2 Comparação entre modelos

A Figura 4.4 e a Tabela 4.5 apresentam a participação das variáveis em cada modelo e os 5 primeiros modelos com maiores probabilidades a posteriori do modelo (PMP). Para o cálculo das probabilidade a posteriori, foi considerado uma priori uniforme para os modelos e a g-prior com g = 'UIP' para os parâmetros dos modelos. Na Figura 4.4, a largura do eixo X de cada modelo é o peso da probabilidade a posteriori. A cor vermelha significa que o coeficiente da variável é positivo e o azul que o coeficiente da variável é negativo. Quando o retângulo é branco significa que a variável não participa no modelo. Na Tabela 4.5, ✓ representa a participação da variável explicativa em modelo (As fórmulas dos 5 modelos são apresentadas no Apêndice B.2). Usamos o Occam's window (seção 2.5) com C = 1000 resultou numa redução de 2^8 modelos para 8 modelos. Observamos que as variáveis 'Cement','Blast', 'Fly ash', 'Water' e 'Age' participam em todos modelos mais importantes, e a 'Superplasticizer' aparece em alguns desses modelos. A 'Coarse Aggregate' e a 'Fine Aggregate' só aparecem nos modelos 4 e 5 (Tabela 4.5) que tem a probabilidade a posteriori menor.

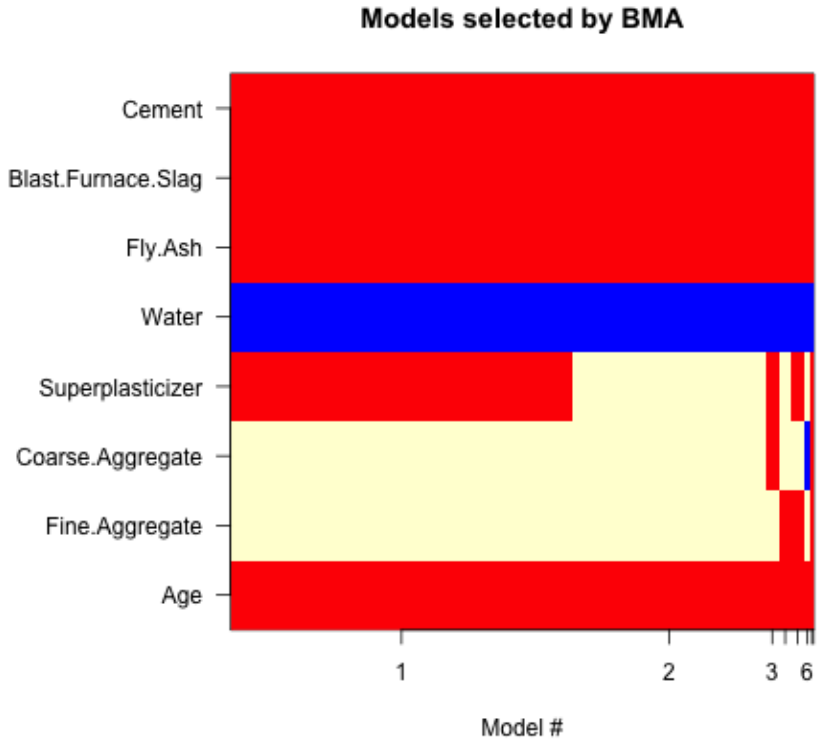

Figura 4.4: *Representação gráfica dos modelos segundo o BMA*

Modelo	Cement	Blast.Furnace.Slag	Fly.Ash	Water	Superplasticizer	Coarse.Aggregate	Fine.Aggregate	Age	PMP
model 1	✓	✓	✓	✓	✓			✓	58.6%
model 2	✓	✓	✓	✓				✓	33.1%
model 3	✓	✓	✓	✓	✓	✓		✓	2.3%
model 4	✓	✓	✓	✓			✓	✓	2.2%
model 5	✓	✓	✓	✓	✓		✓	✓	2.1%

Tabela 4.5: *Os 5 modelos com os maiores valores de PMP*

A Tabela 4.6 apresenta a comparação entre o p-valor do teste para $\beta = 0$ e a probabilidade a posteriori de $\beta \neq 0$. Considerando o nível significância $\alpha = 0.1$, o modelo linear obtido é sem intercepto e todas variáveis são significativas. Por outro lado, o modelo BMA diz que a variável 'Coarse Aggregate' e 'Fine Aggregate' quase não têm impacto para a resistência de concreto e a variável 'Superplasticizer' não é muito importante (P($\beta \neq 0$) = 0.63) segundo Tabela 4.1.

Variável	p-valor	P($\beta \neq 0$)
Intercepto	0.3804	1.00
Cement	<0.01	1.00
Blast.Furnace.Slag	<0.01	1.00
Fly.Ash	<0.01	1.00
Water	<0.01	1.00
Superplasticizer	<0.01	0.63
Coarse.Aggregate	0.0544	0.04
Fine.Aggregate	0.0595	0.05
Age	<0.01	1.00

Tabela 4.6: *P-valores e probabilidades a posteriori para o exemplo 4.2*

A Figura 4.5 apresenta os gráficos de dispersão entre a variável reposta e as duas variáveis 'Coarse Aggregate' e 'Fine Aggregate', com correlações de Pearson iguais a -0.1649 e -0.1672, respectivamente. Observamos que, de fato, não parece que cada umas dessas variáveis tem uma relação forte com a variável resposta.

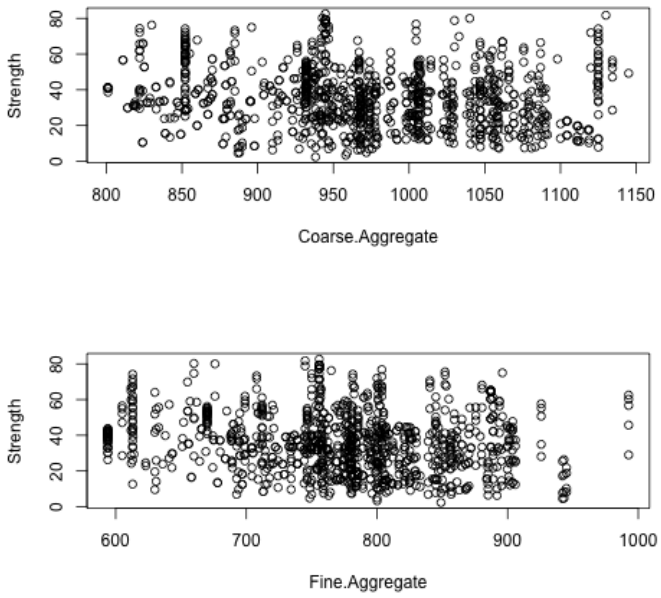

Figura 4.5: *Gráfico de dispersão das variáveis Coarse.Aggregate e Fine.Aggregate com a variável resposta com p-valor baixo e P($\beta \neq 0$) baixo*

A Figura 4.6 apresenta o gráfico de dispersão entre a variável resposta e a variável explicativa 'Superplasticizer'. Pelo gráfico, observamos relação entre 'Strength' e a 'Superplasticizer', mas a relação não é muito forte por causa de ter os nulos.

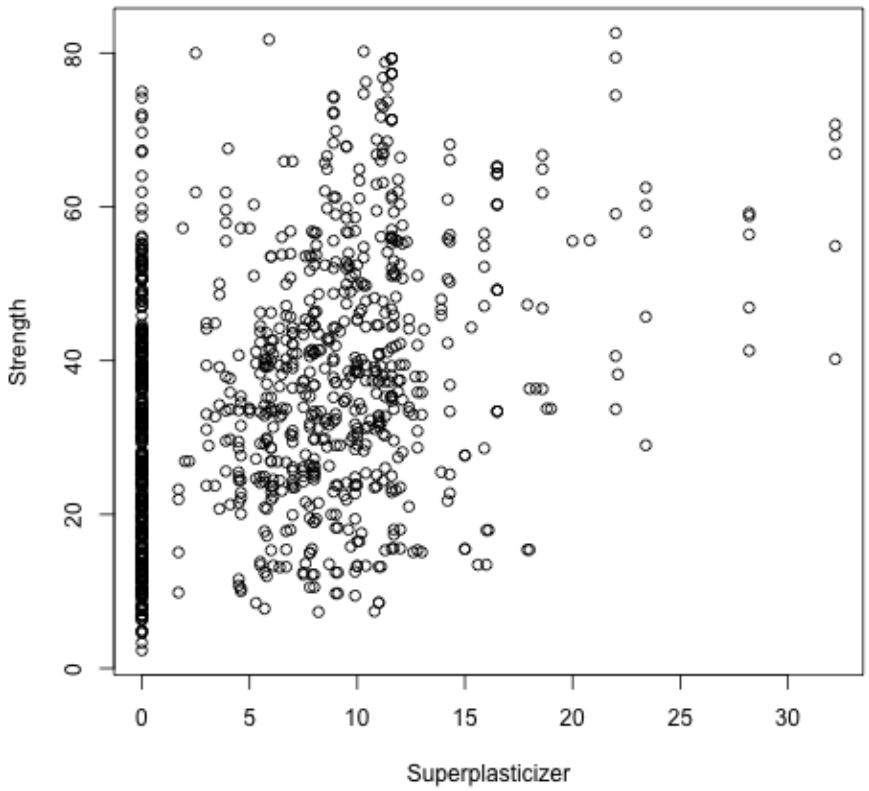

Figura 4.6: *Gráfico das variáveis as variável Superplasticizer com a variável resposta com p-valor baixo e P(β ≠ 0) baixo*

A Tabela 4.7 apresenta as correlações de Pearson entre as variáveis explicativas e resposta.

	Superplasticizer	Coarse.Aggregate	Fine.Aggregate	Strength
Superplasticizer	1.0000	-0.2660	0.2227	0.3661
Coarse.Aggregate	-0.2660	1.0000	-0.1785	-0.1649
Fine.Aggregate	0.2227	-0.1785	1.0000	-0.1672
Strength	0.3661	-0.1649	-0.1672	1.0000

Tabela 4.7: *Tabela de correlação de Pearson*

A Tabela 4.8 apresenta o resultado do EQM na predição usando cross validation com k = 5. Notamos que não existe diferença entre esses três modelos na predição (vale observar que

com o método stepwise regression - backward elimination na regressão linear, os p-valores das duas variáveis 'Coarse.Aggregate' e 'Fine.Aggregate' são menor que 0.01).

Modelos	EQM
BMA	109.57
Linear	109.15
Linear - backward elimination	109.05

Tabela 4.8: *Comparação de EQM com k = 5*

4.3 Análise com dados de nota de alunos

Nesta seção, vamos apresentar os resultados de comparação entre BMA e modelo regressão linear em dados de nota de alunos. Este conjunto de dados foi extraído do Kaggle em https://www.kaggle.com/datasets/dipam7/student-grade-prediction.

4.3.1 Descrição de dados

Estes dados abordam o desempenho dos alunos no ensino secundário de duas escolas portuguesas. Os atributos de dados incluem notas dos alunos, características demográficas, sociais e relacionadas à escola e foram coletados por meio de relatórios e questionários escolares. O conjunto de dados de nota dos alunos contem 395 observações e 28 variáveis, sendo que 27 são variáveis explicativas e 1 é a variável resposta, em que:

Variáveis explicativas
school: escola do aluno (binária: 'GP' - Gabriel Pereira ou 'MS' - Mousinho da Silveira)
age: idade do aluno (numérica: de 15 a 22)
address: tipo de endereço residencial do aluno (binária: 'U' - urbano ou 'R' - rural)
famsize: tamanho da família (binária: 'LE3' - menor ou igual a 3 ou 'GT3' - maior que 3)
Pstatus: status de coabitação dos pais (binária: 'T' - morando juntos ou 'A' - separados)
Medu: escolaridade da mãe (numérica: 0 - nenhum, 1 - ensino fundamental (4^{ε} ano), 2 - 5^{ε} ao 9^{ε} ano, 3 - ensino médio ou 4 - ensino superior)
Fedu: escolaridade do pai (numérica: 0 - nenhum, 1 - ensino fundamental (4^{ε} ano), 2 - 5^{ε} ao 9^{ε} ano, 3 - ensino médio ou 4 - ensino superior)
Mjob: trabalho da mãe (nominal: 'professor', 'saúde' relacionado, 'serviços' civis (por exemplo, administrativo ou policial), 'at_home' ou 'outros')
Fjob: trabalho do pai (nominal: 'professor', 'saúde' relacionado, 'serviços' civis (por exemplo, administrativo ou policial), 'at_home' ou 'other')
reason: motivo para escolher esta escola (nominal: perto de 'casa', escola 'reputação', preferência 'curso' ou 'outro') guardian: tutor do aluno (nominal: 'mãe', 'pai' ou 'outro')
traveltime: tempo de viagem de casa para escola (numérica: 1 - <15 min., 2 - 15 a 30 min., 3 - 30 min. a 1 hora, ou 4 - >1 hora)
studytime: tempo de estudo semanal (numérica: 1 - <2 horas, 2 - 2 a 5 horas, 3 - 5 horas)
failures: número de falhas de classe passadas (numérica: n se 1<=n<3, senão 4)
famsup: apoio educacional familiar (binária: sim ou não)
paid: aulas extras pagas dentro da disciplina do curso (Matemática ou Português) (binário: sim ou não)
activities: atividades extracurriculares (binária: sim ou não)
nursery: frequentou creche (binária: sim ou não)
higher: quer fazer o ensino superior (binária: sim ou não)
internet: Acesso à Internet em casa (binária: sim ou não)
famrel: qualidade das relações familiares (numérica: de 1 - muito ruim a 5 - excelente)
freetime: tempo livre depois da escola (numérica: de 1 - muito baixo a 5 - muito alto)
goout: sair com amigos (numérica: de 1 - muito baixo a 5 - muito alto)
Dalc: consumo de álcool no dia de trabalho (numérica: de 1 - muito baixo a 5 - muito alto)
Walc: consumo de álcool no fim de semana (numérica: de 1 - muito baixo a 5 - muito alto)
health: estado de saúde atual (numérica: de 1 - muito ruim a 5 - muito bom)

absences: número de faltas escolares (numérica: de 0 a 93)

Variáveis resposta
y: nota final (numérica: de 0 a 20)

4.3.2 Tratamento de dados

O conjunto de dados tem bastante variáveis explicativas categóricas. Para facilitar a comparação, tratamos essas variáveis como variáveis "dummies"(1 ou 0 para cada categoria) e removemos a primeira categoria como a referência da variável (N. R. Draper e H. Smith, 1998).

4.3.3 Comparação entre modelos

A Figura 4.7 e a Tabela 4.9 apresentam a participação das variáveis em cada modelo e os 5 primeiros modelos com maiores probabilidades a posteriori do modelo (PMP). Para o cálculo das probabilidade a posteriori, foi considerado uma priori uniforme para os modelos e a g-prior com g = 'UIP' para os parâmetros dos modelos. Na Figura 4.7, observamos que a única variável explicativa que aparece em todos modelos é a variável 'failures', e algumas variáveis explicativas não aparecem por causa de corte de número de modelos segundo Occam's window (resultou numa redução de 2^{36} modelos para 827 modelos com C = 1000). Na Tabela 4.7, são apresentadas apenas as variáveis explicativas que aparecem em ao menos um dos 5 primeiros modelos com maiores PMP e ✓ representa a participação da variável explicativa em modelo (As fórmulas dos 5 modelos são apresentadas no Apêndice B.2).

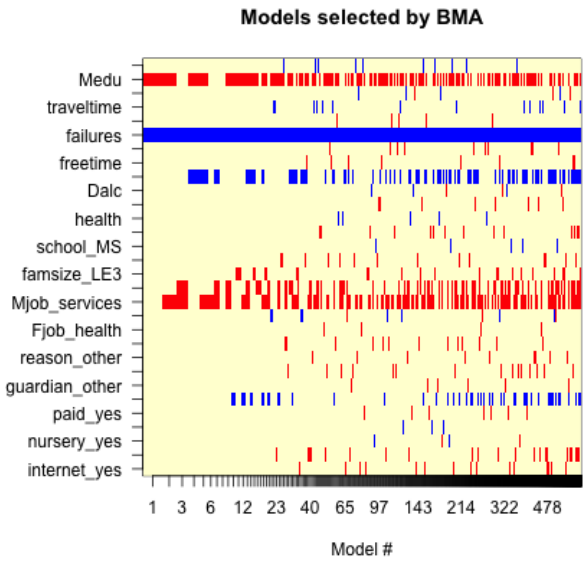

Figura 4.7: *Representação gráfica dos modelos segundo o BMA*

Modelo	Medu	failures	goout	Mjob_health	Mjob_services	PMP
model 1	✓	✓				4.4%
model 2	✓	✓			✓	3.2%
model 3		✓		✓	✓	2.8%
model 4	✓	✓	✓			2.6%
model 5	✓	✓	✓		✓	1.8%

Tabela 4.9: *Os 5 modelos com os maiores valores de PMP*

A Tabela 4.10 e a Tabela 4.11 apresentam a comparação entre o p-valor do teste para $\beta = 0$ e a probabilidade a posteriori de $\beta \neq 0$. Considerando o nível significância $\alpha = 0.1$, as varáveis 'freetime','goout' e 'famsup_yes' são significativas, mas as probabilidades a posteriori para essas variáveis são bem próximas de 0 pelo BMA (comparando com a Tabela 4.1). Ou seja, o BMA aponta que essas variáveis não impactam a variável resposta.

Variável	p-valor	P($\beta \neq 0$)
Intercepto	0.0032	1
age	0.1148	0.05
Medu	0.1540	0.66
Fedu	0.5935	0.02
traveltime	0.4768	0.05
studytime	0.2145	0.02
failures	0.0000	1
famrel	0.2317	0.02
freetime	0.0985	0.04
goout	0.0053	0.36
Dalc	0.5649	0.01
Walc	0.1304	0.02
health	0.3309	0.02
absences	0.2182	0.03
school_MS	0.4725	0.01
address_U	0.4656	0.05
famsize_LE3	0.1567	0.12
Pstatus_T	0.7509	0
Mjob_health	0.2170	0.33
Mjob_other	0.8169	0
Mjob_services	0.2031	0.55
Mjob_teacher	0.6703	0.04
Fjob_health	0.7671	0.02
Fjob_other	0.8027	0
Fjob_services	0.8207	0
Fjob_teacher	0.2823	0.04
reason_home	0.7671	0
reason_other	0.4916	0.02
reason_reputation	0.2764	0.04
guardian_mother	0.9117	0
guardian_other	0.6272	0.02
famsup_yes	0.0265	0.18
paid_yes	0.6394	0.02
activities_yes	0.5007	0.01
nursery_yes	0.7127	0.01
higher_yes	0.3022	0.06
internet_yes	0.5014	0.04

Tabela 4.10: *P-valores e probabilidades a posteriori para o exemplo 4.3*

Variável	p-valor	P($\beta \neq 0$)
freetime	0.0985	0.04
goout	0.0053	0.36
famsup_yes	0.0265	0.18

Tabela 4.11: *Variáveis significativas segundo o modelo linear mas não significativas segundo o BMA*

A Figura 4.8 e a Tabela 4.12 apresentam, respectivamente, o gráfico de boxplot da variável resposta segundo os níveis da variável explicativa 'famsup_yes' e as estatísticas descritivas da variável resposta por grupo dessa variável explicativa, onde Q_i represento i-ésimo quartil i=1,3. Pelo gráfico, observamos que parece não existir associação entre a nota de alunos e o apoio educacional familiar (famsup_yes).

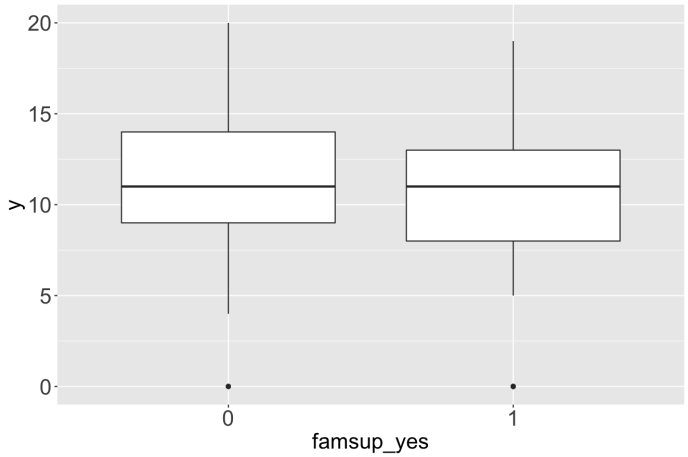

Figura 4.8: *Gráfico de boxplot da variável nota de alunos segundo famsup_yes*

Grupo	Mínimo	Q_1	Mediana	Média	Q_3	Máximo
0	0	9	11	10.64	14	20
1	0	8	11	10.27	13	19

Tabela 4.12: *Tabela de medidas descritiva da variável nota de alunos segundo famsup_yes*

A Figura 4.9 e a Tabela 4.13 apresentam, respectivamente, o gráfico de boxplot da variável reposta segundo os níveis da variável explicativa 'freetime' e as estatísticas descritivas da variável resposta por grupo dessa variável explicativa, onde Q_i representa i-ésimo quartil i = 1,3. Pelo gráfico, observamos que parece não existir associação entre a nota de alunos e níveis de tempo livre depois da escola (freetime).

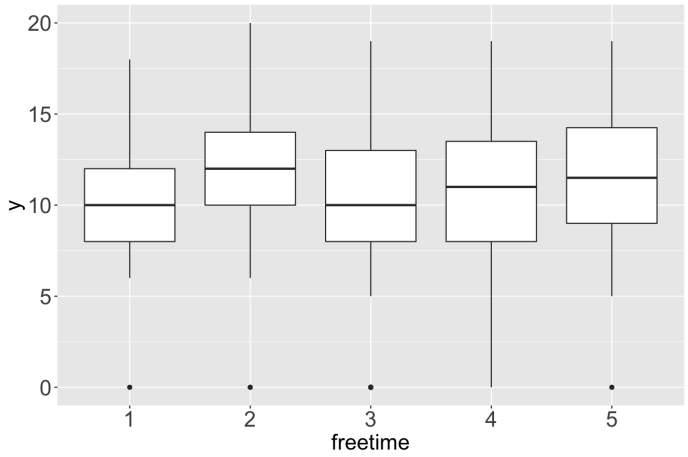

Figura 4.9: *Gráfico de boxplot da variável nota de alunos segundo freetime*

Grupo	Mínimo	Q_1	Mediana	Média	Q_3	Máximo
1	0	8	10	9.842	12	18
2	0	10	12	11.56	14	20
3	0	8	10	9.783	13	19
4	0	8	11	10.43	13.50	19
5	0	9	11.50	11.30	14.25	19

Tabela 4.13: *Tabela de medidas descritiva da variável nota de alunos segundo freetime*

A Figura 4.10 e a Tabela 4.14 apresentam apresenta, respectivamente, o gráfico de boxplot da variável resposta segundo os níveis da variável explicativa 'goout' e as estatísticas descritivas da variável resposta por grupo dessa variável explicativa, onde Q_i representa i-ésimo quartil i = 1,3 . Pelo gráfico, observamos que não parece que existir associação entre a nota de alunos e níveis de sair com amigos (goout).

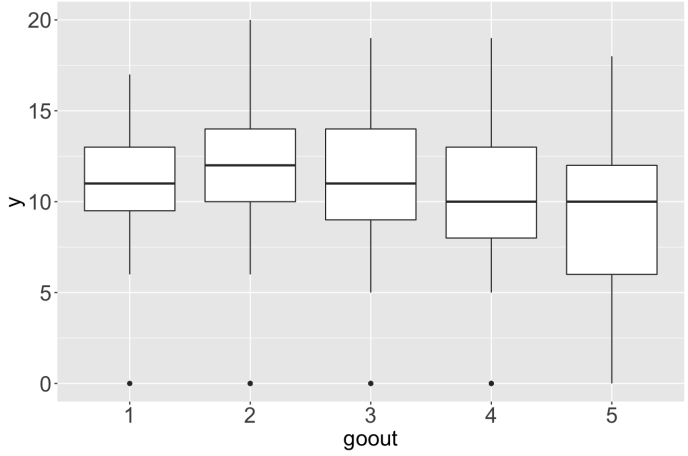

Figura 4.10: *Gráfico de boxplot da variável nota de alunos segundo goout*

Grupo	Mínimo	Q_1	Mediana	Média	Q_3	Máximo
1	0	9.50	11	9.87	13	17
2	0	10	12	11.19	14	20
3	0	9	11	10.96	14	19
4	0	8	10	9.651	13	19
5	0	6	10	9.038	12	18

Tabela 4.14: *Tabela de medidas descritiva da variável nota de alunos segundo goout*

Por fim, a Tabela 4.15 apresenta o resultado de EQM usando cross validation com k = 5. Observamos que o BMA tem um ganho na predição comparando com o modelo linear, mas não há diferença quando comparado com a regressão linear usando stepwise (vale nota que com o método stepwise regression - backward elimination na regressão linear, apenas as duas variáveis 'failures' e 'goout' são significativas no modelo com p-valores menor que ao nível 0.1).

Modelos	EQM
BMA	18.52
Linear	19.85
Linear backward elimination	18.34

Tabela 4.15: *Comparação de EQM com k = 5*

4.4 Análise com dados de ataque cardíaco

Nesta seção, vamos apresentar os resultados de comparação entre BMA e modelo regressão logística em dados de doenças cardíacas. Este conjunto de dados foi extraído do Kaggle em https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset.

4.4.1 Descrição de dados

Nesse estudo, a análise é realizada usando dados publicamente disponíveis para doenças cardíacas. Deseja-se avaliar quais fatores influenciam na ocorrência de doença cardíaca. Este conjunto de dados contém 303 observações. Para cada individuo são registrados 12 variáveis sendo que 11 são variáveis explicativas e 1 é a variável resposta, em que:

Variáveis explicativas
age: a idade da pessoa em anos
sex: o sexo da pessoa (1 = masculino, 0 = feminino)
cp: tipo de dor no peito
— Valor 0: assintomático
— Valor 1: angina atípica
— Valor 2: dor não anginosa
— Valor 3: angina típica
trestbps: A pressão arterial em repouso da pessoa (mm Hg na admissão ao hospital)
chol: a medição de colesterol da pessoa em mg/dl
fbs: glicemia em jejum da pessoa (> 120 mg/dl, 1 = verdadeiro; 0 = falso)
restecg: resultados eletrocardiográficos em repouso
— Valor 0: mostrando hipertrofia ventricular esquerda provável ou definitiva pelos critérios de Estes
— Valor 1: normal
— Valor 2: com anormalidade da onda ST-T (inversões da onda T e/ou elevação ou depressão de ST > 0,05 mV)
oldpeak: depressão de ST induzida pelo exercício em relação ao repouso (’ST’ refere-se a posições no gráfico de ECG.)
slope: inclinação do segmento ST no pico do exercício — 0: inclinação descendente; 1: plano; 2: subida
0: downsloping; 1: flat; 2: upsloping
ca: O número de vasos principais (0–3) coloridos por fluorosopia
thal: Um distúrbio do sangue chamado talassemia
— Valor 0: NULL (retirado do conjunto de dados anteriormente
— Valor 1: defeito fixo (sem fluxo sanguíneo em alguma parte do coração)
— Valor 2: fluxo sanguíneo normal
— Valor 3: defeito reversível (observa-se um fluxo sanguíneo, mas não é normal)

Variáveis resposta
y: Doença cardíaca (1 = não, 0 = sim)

4.4.2 Tratamento de dados

O conjunto de dados tem bastantes variáveis explicativas categóricas. Para facilitar a comparação, tratamos essas variáveis como variáveis "dummies"(1 ou 0 para cada categoria) e removemos a primeira categoria como a referência (N. R. Draper e H. Smith, 1998).

4.4.3 Comparação entre modelos

A Figura 4.11 e a Tabela 4.16 apresentam a participação das variáveis em cada modelo e os 5 primeiros modelos com maiores probabilidades a posteriori do modelo (PMP). Para o cálculo das probabilidade a posteriori, foi considerado uma priori uniforme para os modelos e a g-prior com g = 'UIP' para os parâmetros dos modelos. Na Figura 4.11, observamos que algumas variáveis explicativas aparecem em todos modelos, ou seja, são variáveis significativas no BMA. Neste exemplo, o Occam's window resultou numa redução de 2^{20} modelos para 760 modelos com C = 1000. Na Tabela 4.16, são apresentadas apenas as variáveis explicativas que aparecem em ao menos um dos 5 primeiros modelos com maiores PMP e ✓ representa a participação da variável explicativa em modelo (As fórmulas dos 5 modelos são apresentadas no Apêndice B.2).

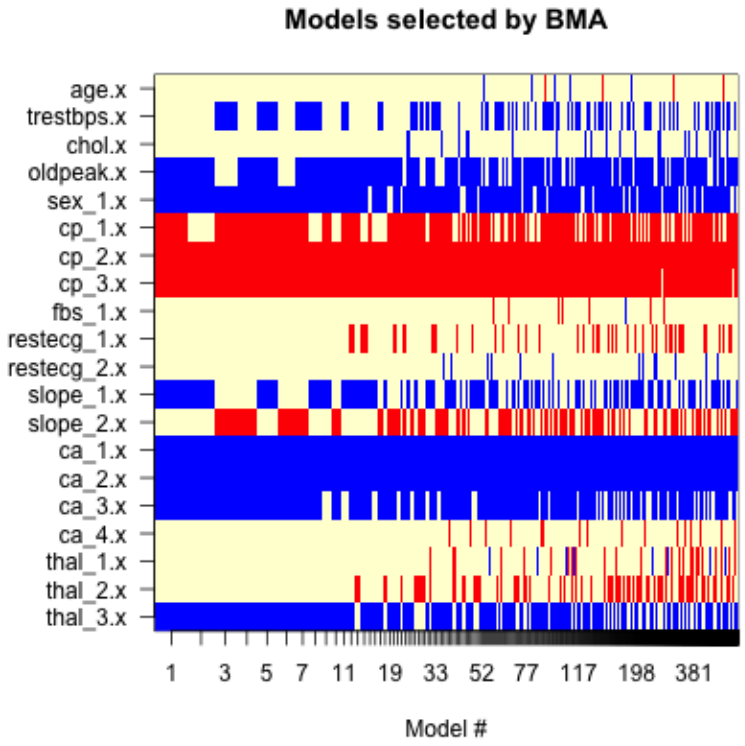

Figura 4.11: *Representação gráfica dos modelos segundo o BMA*

Modelo	trestbps	oldpeak	sex_1	cp_1	cp_2	cp_3	slope_1	slope_2	ca_1	ca_2	ca_3	ca_4	thal_1	thal_2	thal_3	PMP
model 1		✓	✓	✓	✓	✓	✓		✓	✓	✓				✓	5.8%
model 2		✓	✓		✓	✓	✓		✓	✓	✓				✓	4.4%
model 3	✓		✓	✓	✓	✓		✓	✓	✓	✓				✓	3.9%
model 4		✓	✓	✓	✓	✓		✓	✓	✓	✓				✓	3.5%
model 5	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓				✓	3.4%

Tabela 4.16: *Os 5 modelos com os maiores valores de PMP*

A Tabela 4.17 e a Tabela 4.18 apresentam a comparação entre o p-valor do teste para $\beta = 0$ e a probabilidade a posteriori de $\beta \neq 0$. Considerando o nível significância $\alpha = 0.1$, a varável explicativa 'trestbps' é significativa sob o modelo linear mas a probabilidade a posteriori é baixa sob o BMA, ou seja, o BMA aponta que essa variável não impacta muito a variável resposta. Outra variável explicativa, 'thal_3', não é significativa em p-valor, mas a probabilidade a posteriori é alta, ou seja, o BMA aponta que essa variável é importante para a variável resposta mas o p-valor não (comparando com a Tabela 4.1).

Variável	p-valor	P($\beta \neq 0$)
Intercepto	0.4747	1.00
age	0.6201	0.04
trestbps	0.0487	0.36
chol	0.4308	0.07
oldpeak	0.0418	0.78
sex_1	0.0026	0.89
cp_1	0.0204	0.72
cp_2	0.0000	1.00
cp_3	0.0001	0.99
fbs_1	0.4924	0.04
restecg_1	0.2385	0.14
restecg_2	0.7028	0.05
slope_1	0.2866	0.55
slope_2	0.3744	0.48
ca_1	0.0000	1.00
ca_2	0.0001	1.00
ca_3	0.0118	0.85
ca_4	0.5396	0.05
thal_1	0.3447	0.09
thal_2	0.3148	0.23
thal_3	0.6643	0.81

Tabela 4.17: *P-valores e probabilidade a posteriori para o exemplo 4.3*

Variável	p-valor	P($\beta \neq 0$)
trestbps	0.0487	0.36
thal_3	0.6643	0.81

Tabela 4.18: *As variáveis que tem diferentes resultados nos dois métodos*

A Figura 4.12 e a Tabela 4.19 apresentam o gráfico de boxplot e as medidas descritivas para a variável explicativa 'trestbps' segundo a variável y. Pelo gráfico, observamos que não existir uma relação entre a doença cardíaca e a pressão arterial em repouso da pessoa. E na Tabela 4.19, os valores de média e mediana da variável explicativa 'trestbps' não mudam segundo os níveis da variável reposta: observamos que para os pacientes que tem

doença cardíaca, as medidas não tem muita diferença das medidas dos pacientes que não tem doença.

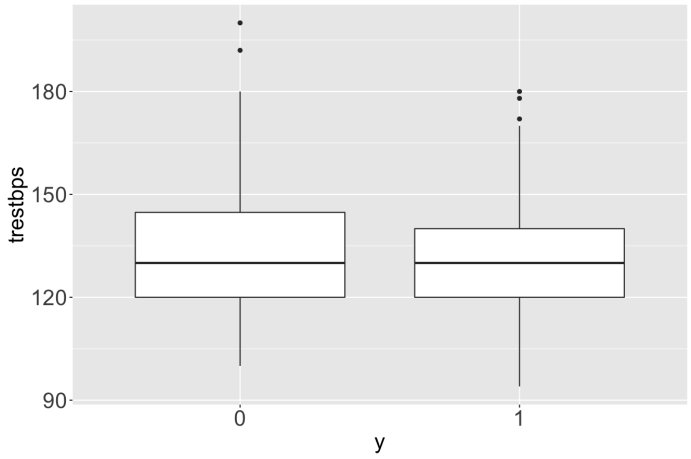

Figura 4.12: *Gráfico de boxplot da variável trestbps segundo variável y*

Doença cardíaca	Mínimo	Q_1	Mediana	Média	Q_3	Máximo
0(sim)	0	9	11	10.64	14	20
1(não)	0	8	11	10.27	13	19

Tabela 4.19: *Tabela de medidas descritiva da variável trestbps segundo variável y*

A Tabela 4.20 é a tabela de contingência entre a variável thal_3 e a variável reposta. Aplicamos o teste qui-quadrado para detectar a existência ou não de associação entre as variáveis. O p-valor é menor que 0.01, isto é, há dependência de associação entre de doença cardíaca e defeito reversível de talassemia.

	Doença cardíaca = 0	Doença cardíaca = 1	Total
thal_3 = 0	49(26.3%)	137(73.7%)	186(100%)
thal_3 = 1	89(76.1%)	28(23.9%)	117(100%)

Tabela 4.20: *Tabela de contingência 2x2 para as variáveis y e thal_3*

Por fim, a Tabela 4.21 apresenta o resultado de área sob a Curva de ROC (AUC) (Hanley e Mcneil, 1983) usando cross validation com k = 5. Observamos que o BMA tem um ganho na predição tanto em relação ao modelo de regressão logística usual quanto em relação ao modelo considerando o método stepwise backward (vale observar que com o método stepwise regression - backward elimination na regressão logística, não muda o resultado das variáveis significativas).

Modelos	AUC
BMA	0.9121
Logística	0.8830
Logística backward elimination	0.8926

Tabela 4.21: *Comparação de AUC com k = 5*

Capítulo 5

Conclusões e perspectivas futuras

A pesquisa cientifica contemporânea de diversas áreas baseia-se cada vez mais em métodos quantitativos como, por exemplo, em modelos estatísticos. Ao ajustar modelos estatísticos a dados específicos, os pesquisadores podem atualizar a incerteza de modelo e a incerteza dos coeficientes do modelo. Essas incertezas não receberam atenção suficiente nas pesquisas.

Neste trabalho, foram apresentadas a motivação da aplicação de BMA em estudos, uma revisão do método BMA, os resultados da aplicação dele em alguns conjunto de dados, bem como a comparação entre o BMA e o modelo de regressão clássica. Com os resultados, observamos que vários fatores podem impactar o estudo sobre quando usar modelos clássicos de regressão, como tamanho da amostra, a escolha de α comparando com p-valor, etc. Além disso, o BMA mostra uma estabilidade de seleção de variáveis explicativas melhor independente desses fatores (tanto nos dados simulados quando em dados reais) e resultados melhores também na predição.

M. F. J. Steel (2017) anotou que:
a) O BMA é ótimo (sob M-closed, isto é, quando o modelo verdadeiro é um dos modelos consideráveis) em previsão.
b) O BMA é fácil de implementar em situações onde o espaço do modelo é grande com alguns algoritmos como Occam's windows.
c) O BMA leva a interpretações substantivamente valiosas de probabilidades de modelos a posteriori e probabilidades de inclusão a posteriori.

Finalmente, o estudo da incerteza dos modelos estatísticos também ajudam a replicação da pesquisa, pois o código do estudo fornecido pelo pesquisador já incluindo a preferência e outras "horizonte"(como seleção de variáveis, tratamentos dos dados, etc.) (Leamer, 1983) dele. Essencialmente, qualquer modelo estatístico é definido pelo pesquisadores. Portanto, qualquer modelo estatístico é inevitavelmente subjetivo. Porém, com o crescimento do tamanho de conjunto de dados e número de variáveis, as pesquisas devem continuar a considerar essa questão da incerteza do modelo.

5.1 Perspectiva

Davidson e Fan (2006) aponta que BMA é pouco usado comparando com boosting e bagging por causa de a) ser computacionalmente ineficiente e b) o boosting e o bagging são provados que funcionam para vários problemas. Mas o BMA é cada vez mais usado na última década. Para estudos futuros de BMA, por um lado, é de interesse aprimorar a eficiência do método. Embora o desenvolvimento dos computadores tenha permitido o processamento de modelos que envolvem bilhões de dados, devido à enorme quantidade de modelos, o BMA ainda não é um método muito eficiente em big data. Considerando custo computacional, uma solução é a aproximação via pseudo-BMA (Yao *et al.*, 2018) para o cálculo de probabilidades a posteriori. Por outro lado, podemos estudar comparativamente o BMA com os métodos de ensemble learning como boosting, bagging e stacking na seleção de modelos. Davidson e Fan (2006) mostra que mesmo boosting e bagging são mais usados e fáceis de implementar, mas ainda o BMA tem a melhor performance quando há a incerteza de modelo significativa, Clarke (2003) aponta que o BMA é sempre melhor que stacking não bayesiano quando os dados via simulados. Além disso, Jiyuan *et al.* (2017) mostra que BMA tem o melhor performance de predição do que LASSO com algumas distribuições a priori de parâmetros e Porwal e A. E. Raftery (2022) mostra que BMA tem a performance computacional parecido com LASSO. Comparando com demais métodos de regularização, BMA é o melhor em g= $\sqrt{n}$ (seção 3.2).

Pelos estudos nesse trabalho, o BMA tem o resultado mais robusto na estimativa de parâmetros do modelo. Comparando com outros métodos, também tem uma teoria por atrás mostra que é melhor que qualquer único modelo.

Por fim, para trabalhos futuros, sugerimos as seguintes direções:
a) Estudar a relação entre o C (de Occam's window) e as variáveis significativas no BMA, ou determinar um número ótimo de variáveis significativas via C.
b) Nesse trabalho, só aplicamos a priori uniforme com a g-priori de 'UIP', poderia testar outros g-prioris.
c) Somente testamos BMA com as regressões em relação lineares, é possível estudamos também com diferentes formatos do modelo que não seja linear, por exemplo, $y = \beta^2 X$, $y = \beta^3 X$,... etc ou via próprio variáveis explicativas como $y = \beta X^2$, $y = \beta X^3$,... etc.
d) O BMA pode levar a uma ponderação com um número grande de modelos. É possível desenvolver o BMA com outros algoritmos mais rápidos na implantação do método.

Apêndice A

A.1 Otimalidade do BMA

Vamos verificar que a distribuição preditiva para a variável resposta Y, que é uma ponderação das preditivas dos modelos em M, possui log predictive score esperado maior que o log predictive score esperado de qualquer modelo único. Assim, para qualquer distribuição g para Y, temos dado a informação da amostra D, que

$$\begin{aligned} E[S(p(\cdot|D))] - E[S(g,y)|D] &= \int log(p(y|D))p(y|D)dy - \int log(g(y))p(y|D)dy \\ &= \int [log(p(y|D)) - log(g(y))]p(y|D)dy \\ &= \int p(y|D)log\frac{p(y|D)}{g(y)}dy \\ &= KLD(p(\cdot|D), g), \end{aligned}$$

onde KLD(f,g) é a distância de Kullback-Leibler entre f e g. Como sabemos que KLD(f,g)≥ 0 (desigualdade de Gibbs em MacKay (2003)), resulta que

$$E[S(p(\cdot|D), y)|D] \geq E[S(g,y)|D] \quad (*)$$

Lembrando que

$$p(y|D) = \sum_{j=1}^{Q} p(y|M_j, D)p(M_j|D)$$

e definindo

$$g_i(y) = p(y|M_j, D) = \sum_{j=1}^{Q} \alpha_j^{(i)} p(y|M_j, D),$$

onde $\alpha_j^{(i)}$ = 1 se j = i e $\alpha_j^{(i)}$ = 0, se j≠i, conclui-se de (*) que $E[S(p(\cdot|D), y)|D] \geq E[S(g_i, y)|D]$,

isto é, para qualquer i = 1,...,Q, o log predictive score esperado da distribuição preditiva é maior ou igual ao log predictive socre esperado da distribuição preditiva considerando unicamente o modelo M_i (Zhou, 2011; Clyde e Iversen, 2013; A. Raftery e Zheng, 2003).

A.2 BIC aproximação

Vamos provar o resultado da seção 3.3 fazendo do uso do método do Laplace para integral

$$P(D) = \int P(D|\theta)P(\theta)d\theta,$$

onde $\theta = (\theta_1, \theta_2, ..., \theta_d)$. Seja

$$g(\theta) = logP(D|\theta)P(\theta)$$

Pelo desenvolvimento em séries de Taylor, temos, em torno do ponto de máximo θ^*, que

$$g(\theta) = g(\theta^*) + (\theta - \theta^*)^T g^{'}(\theta^*) + \frac{1}{2}(\theta - \theta^*)^T g^{''}(\theta^*)(\theta - \theta^*) + o(||\theta - \theta^*||^2)$$

em que $g^{'}(\theta) = (\frac{\partial g(\theta)}{\partial \theta_1}, ..., \frac{\partial g(\theta)}{\partial \theta_d})^T$ é o vetor de derivadas parciais de primeiro ordem de $g(\theta)$, e $g^{''}(\theta)$ é a matriz Hessiana de derivadas parciais de segunda ordem em (i,j), com $\frac{\partial^2 g(\theta)}{\partial \theta_i \partial \theta_j}$. Temos $g^{'}(\theta) = 0$ quando $g(\theta)$ atingir o valor máximo. Assim:

$$g(\theta) \approx g(\theta^*) + \frac{1}{2}(\theta - \theta^*)^T g^{''}(\theta^*)(\theta - \theta^*)$$

Temos

$$P(D) \approx \int exp[g(\theta^*) + \frac{1}{2}(\theta - \theta^*)^T g^{''}(\theta^*)(\theta - \theta^*)]d\theta$$

Ou seja,

$$P(D) = \int exp[g(\theta)]d\theta \approx exp[g(\theta^*)] \int exp[\frac{1}{2}(\theta - \theta^*)^T g^{''}(\theta^*)(\theta - \theta^*)]d\theta$$

Mas $\int exp[\frac{1}{2}(\theta - \theta^*)^T g^{''}(\theta^*)(\theta - \theta^*)]d\theta$ é a integral de densidade da distribuição normal de vetor médias θ^* e matriz de covariância $g^{''}(\theta^*)$. Assim, temos:

$$P(D) \approx \int exp[g(\theta)]d\theta \approx exp[g(\theta^*)](2\pi)^{\frac{d}{2}}|A|^{-\frac{1}{2}} = P(D|\theta^*)P(\theta^*)(2\pi)^{\frac{d}{2}}|A|^{-\frac{1}{2}}$$

em que d é o número de parâmetros no modelo e $A = -g^{''}(\theta^*)$ (Tierney e Kadane, 1986).

Aplicando o log na última expressão acima,

$$logP(D) = logP(D|\theta^*) + logP(\theta^*) + \frac{d}{2}log(2\pi) - \frac{1}{2}log|A| + O(n^{-1}),$$

em que $O(n^{-1})$ é o erro de aproximação e $nO(n^{-1})$ converge para um valor constante quando $n \longrightarrow \infty$.

Para uma amostra suficiente grande, $\theta^* \approx \hat{\theta}$ que é um o estimador de máxima verossimilhança, e $A \approx ni$, em que i é a esperança de matriz de informação de Fisher para uma observação (A. E. Raftery, 1995). Ela é uma matriz (dxd) com elementos de forma $-E[\frac{\partial^2 P(y_1|\theta)}{\partial\theta_i\partial\theta_j}|\theta = \hat{\theta}]$, que é o valor esperado de y_1 com θ fixado e ainda, $|A| \approx n^d i$ (A. E. Raftery, 1995). Assim

$$logP(D) = logP(D|\hat{\theta}) + logP(\hat{\theta}) + \frac{d}{2}log(2\pi) - \frac{d}{2}logn - \frac{1}{2}log|i| + O(n^{-\frac{1}{2}})$$

Suponha que, a priori, θ tem a distribuição normal multivariada com a média $\hat{\theta}$ e matriz de variância i^{-1}. Como esses termos $logP(\hat{\theta})$, $\frac{d}{2}log(2\pi)$ e $\frac{1}{2}log|i|$ não dependem do n, em média, contém a mesma quantidade de informação que uma única observação. Então $logP(\hat{\theta}) = \frac{d}{2}log(2\pi) + \frac{1}{2}log|i|$ (A. E. Raftery, 1995). Assim, temos

$$logP(D) = logP(D|\hat{\theta}) - \frac{d}{2}logn + O(n^{-\frac{1}{2}}),$$

onde o erro $O(n^{-\frac{1}{2}})$ converge para 0 quando $n \longrightarrow \infty$.

Sendo que $BIC = -2ln(\hat{L}) + kln(n)$, para o modelo M_k, temos

$$\begin{aligned} P(M_k|D) &= \frac{P(D|M_k)P(M_k)}{\sum_{j=1}^{K} P(D|M_j)P(M_j)} \\ &\approx \frac{exp(-\frac{1}{2}BIC_k)p(M_k)}{\sum_{j=1}^{K} exp(-\frac{1}{2}BIC_j)p(M_j)} \end{aligned}$$

A.3 Formula fechada de obter a posteriori dos parâmetros no modelo linear

Considerar um modelo de regressão linear múltipla

$$y = X\beta + \epsilon$$

em que $y = (y_1, y_2, ..., y_n)^T$ um vetor $nx1$ de observações, e $X = (x_1, x_2, ..., x_n)^T$ uma matriz nxp de covariáveis, ou seja, a matriz de design, com $x_i = (1, x_{i1}, x_{i2}, ..., x_{i(p-1)})^T$, (β, σ^2) o vetor de parâmetros, com $\beta = (\beta_0, \beta_1, ..., \beta_{p-1})$, e $\epsilon = (\epsilon_1, \epsilon_2, ..., \epsilon_n)$ um vetor $nx1$ de erros aleatórios.

A função de verossimilhança nesse caso é:

$$f(y|\beta, \sigma^2, x) = (2\pi\sigma^2)^{-\frac{n}{2}} exp\{-\frac{1}{2\sigma^2}(y - x\beta)^T(y - x\beta)\}$$

No caso, a variância de ϵ é σ^2 desconhecida (Murphy, 2007; Hoyos, 2020). Suponhamos que:

$$\sigma^2 \sim IG(a_0, b_0),$$

isto é,

$$g(\sigma^2) = \frac{(b_0)^{a_0}}{\Gamma(a_0)}(\sigma^2)^{-(a_0+1)} \exp\{-\frac{b_0}{\sigma^2}\}\mathbb{I}_{R^+}(\sigma^2),$$

e

$$\beta|\sigma^2 \sim N_p(m_0, \sigma^2 V_0),$$

isto é,

$$g(\beta|\sigma^2) = (2\pi)^{-\frac{p}{2}}|V_0|^{-\frac{1}{2}}(\sigma^2)^{-\frac{p}{2}} \exp\{-\frac{1}{2\sigma^2}[(\beta - m_0)^t V_0^{-1}(\beta - m_0)]\}.$$

Ou seja,

$$(\beta, \sigma^2) \sim N_{p+1}IG(m_0, V_0, a_0, b_0)$$

cuja densidade é

$$g(\beta, \sigma^2) = \frac{b_0^{a_0}}{(2\pi)^{\frac{p}{2}}|V_0|^{\frac{1}{2}}\Gamma(a_0)}(\sigma^2)^{-(a_0+\frac{p}{2}+1)} exp\{-\frac{1}{2\sigma^2}[(\beta - m_0)^T V_0^{-1}(\beta - m_0) + 2b_0]\}\mathbb{I}_{R^+}(\sigma^2)$$

Usando a função de verossimilhança e a distribuição a priori, temos que

$$
\begin{aligned}
f(\beta, \sigma^2|y, x) &= \frac{f(y|\beta, \sigma^2, x)g(\beta, \sigma^2)f(x)}{f(y|x)f(x)} \\
&\propto f(y|\beta, \sigma^2, x)g(\beta, \sigma^2) \\
&\propto (\sigma^2)^{-(a_1+\frac{p}{2}+1)} \exp\{-\frac{1}{2\sigma^2}[(\beta - m^*)^t V^{*-1}(\beta - m^*) + 2b_1]\}\mathbb{I}_{R^+}(\sigma^2)
\end{aligned}
$$

em que

$$
\begin{aligned}
V^* &= (V_0^{-1} + X^T X)^{-1} \\
m^* &= V^*(V_0^{-1} m_0 + X^T y) \\
a_1 &= a_0 + \frac{n}{2} \\
b_1 &= b_0 + \frac{m_0^t V_0^{-1} m_0 + y^T y - m^{*T} V^{*-1} m^*}{2}
\end{aligned}
$$

Supondo que $X^T X$ seja uma matriz não singular, temos ainda que

$$m^* = V^*(V_0^{-1} m_0 + X^T X\hat{\beta}),$$

onde $\hat{\beta}$ é o vetor de estimativa da β. Ou seja, a distribuição posteriori de (β, σ^2) é Normal Gama Inversa (Hoyos, 2020), isto é,

$$\beta, \sigma^2|y, x \sim N_p IG(m^*, V^*, a1, b1)$$

Apêndice B

B.1 Erro quadrado médio de dados simulados com diferentes valores de k (k-fold)

n	EQM-BMA	EQM-linear	n	EQM-BMA	EQM-linear
50	1.1073	1.6025	50	0.9892	1.1549
100	0.7760	0.8030	100	0.8445	0.8748
200	0.8668	0.9098	200	0.8837	0.9630
500	1.0214	1.0475	500	1.0253	1.0226
1000	0.9864	0.9993	1000	0.9872	1.0009
2000	0.9875	0.9905	2000	0.9855	0.9894
5000	1.0284	1.0301	5000	1.0275	1.0297
10000	1.0184	1.0193	10000	1.0180	1.0185
20000	1.0019	1.0022	20000	1.0020	1.0022
50000	1.0139	1.0141	50000	1.0140	1.0142

(a) *k = 3* **(b)** *k = 4*

n	EQM-BMA	EQM-linear	n	EQM-BMA	EQM-linear
50	1.0661	1.2205	50	0.9586	1.2769
100	0.8082	0.8639	100	0.8113	0.8447
200	0.8916	0.9422	200	0.9038	0.9504
500	1.0266	1.0263	500	1.0222	1.0170
1000	0.9801	0.9960	1000	0.9812	0.9957
2000	0.9830	0.9897	2000	0.9847	0.9865
5000	1.0284	1.0305	5000	1.0266	1.0283
10000	1.0179	1.0189	10000	1.0177	1.0187
20000	1.0018	1.0020	20000	1.0016	1.0019
50000	1.0140	1.0142	50000	1.0139	1.0141

(c) *k = 6* **(d)** *k = 7*

Tabela B.1: *Comparação de EQM via cross validation*

B.2 Detalhes dos modelos mais prováveis do capítulo 4

Os 5 modelos mais prováveis no exemplo da seção 4.2(Tabela 4.5):

$$Model1 : y = \beta_0 + \beta_1 X_{Cement} + \beta_2 X_{Blast.Furnace.Slag} + \beta_3 X_{Fly.Ash} + \beta_4 X_{Water} + \beta_5 X_{Superplasticizer} + \beta_8 X_{Age}$$

$$Model2 : y = \beta_0 + \beta_1 X_{Cement} + \beta_2 X_{Blast.Furnace.Slag} + \beta_3 X_{Fly.Ash} + \beta_4 X_{Water} + \beta_8 X_{Age}$$

$$Model3 : y = \beta_0 + \beta_1 X_{Cement} + \beta_2 X_{Blast.Furnace.Slag} + \beta_3 X_{Fly.Ash} + \beta_4 X_{Water} + \beta_5 X_{Superplasticizer} + \beta_6 X_{Cparse.Aggregate} + \beta_8 X_{Age}$$

$$Model4 : y = \beta_0 + \beta_1 X_{Cement} + \beta_2 X_{Blast.Furnace.Slag} + \beta_3 X_{Fly.Ash} + \beta_4 X_{Water} + \beta_7 X_{Fine.Aggregate} + \beta_8 X_{Age}$$

$$Model5 : y = \beta_0 + \beta_1 X_{Cement} + \beta_2 X_{Blast.Furnace.Slag} + \beta_3 X_{Fly.Ash} + \beta_4 X_{Water} + \beta_5 X_{Superplasticizer} + \beta_7 X_{Fine.Aggregate} + \beta_8 X_{Age}$$

Os 5 modelos mais prováveis no exemplo da seção 4.3(Tabela 4.9):

$$Model1 : y = \beta_0 + \beta_1 X_{Medu} + \beta_2 X_{failures}$$

$$Model2 : y = \beta_0 + \beta_1 X_{Medu} + \beta_2 X_{failures} + \beta_5 X_{Mjob\_services}$$

$$Model3 : y = \beta_0 + \beta_2 X_{failures} + \beta_4 X_{Mjob\_health} + \beta_5 X_{Mjob\_services}$$

$$Model4 : y = \beta_0 + \beta_1 X_{Medu} + \beta_2 X_{failures} + \beta_3 X_{goout}$$

$$Model5 : y = \beta_0 + \beta_1 X_{Medu} + \beta_2 X_{failures} + \beta_3 X_{goout} + \beta_5 X_{Mjob\_services}$$

Os 5 modelos mais prováveis no exemplo da seção 4.4(Tabela 4.16):

$$Model1 : y = \beta_0 + \beta_2 X_{oldpeak} + \beta_3 X_{sex\_1} + \beta_4 X_{cp\_1} + \beta_5 X_{cp\_2} + \beta_6 X_{cp\_3} + \beta_7 X_{slope\_1} + \beta_9 X_{ca\_1} + \beta_{10} X_{ca\_2} + \beta_{11} X_{ca\_3} + \beta_{15} X_{thal\_3}$$

$$Model2 : y = \beta_0 + \beta_2 X_{oldpeak} + \beta_3 X_{sex\_1} + \beta_5 X_{cp\_2} + \beta_6 X_{cp\_3} + \beta_7 X_{slope\_1} + \beta_9 X_{ca\_1} + \beta_{10} X_{ca\_2} + \beta_{11} X_{ca\_3} + \beta_{15} X_{thal\_3}$$

$$Model3 : y = \beta_0 + \beta_1 X_{trestbps} + \beta_3 X_{sex\_1} + \beta_4 X_{cp\_1} + \beta_5 X_{cp\_2} + \beta_6 X_{cp\_3} + \beta_8 X_{slope\_2} + \beta_9 X_{ca\_1} + \beta_{10} X_{ca\_2} + \beta_{11} X_{ca\_3} + \beta_{15} X_{thal\_3}$$

$$Model4 : y = \beta_0 + \beta_2 X_{oldpeak} + \beta_3 X_{sex\_1} + \beta_4 X_{cp\_1} + \beta_5 X_{cp\_2} + \beta_6 X_{cp\_3} + \beta_8 X_{slope\_2} + \beta_9 X_{ca\_1} + \beta_{10} X_{ca\_2} + \beta_{11} X_{ca\_3} + \beta_{15} X_{thal\_3}$$

$$Model5 : y = \beta_0 + \beta_1 X_{trestbps} + \beta_2 X_{oldpeak} + \beta_3 X_{sex\_1} + \beta_4 X_{cp\_1} + \beta_5 X_{cp\_2} + \beta_6 X_{cp\_3} + \beta_7 X_{slope\_1} + \beta_9 X_{ca\_1} + \beta_{10} X_{ca\_2} + \beta_{11} X_{ca\_3} + \beta_{15} X_{thal\_3}$$

Referências

[Arin e Braunfels 2018] K. Peren Arin e Elias Braunfels. "The resource curse revisited: a bayesian model averaging approach". Em: *Energy Economics* 70 (2018), pgs. 170–178. issn: 0140-9883. doi: https://doi.org/10.1016/j.eneco.2017.12.033 (citado na pg. 5).

[Barnard 1963] G. A. Barnard. "New methods of quality control". Em: *Journal of the Royal Statistical Society. Series A (General)* 126.2 (1963), pgs. 255–258. issn: 00359238. url: http://www.jstor.org/stable/2982365 (acesso em 25/10/2022) (citado na pg. 3).

[Barro e McCleary 2003] Robert J Barro e Rachel McCleary. *Religion and Economic Growth*. Rel. técn. 9682. National Bureau of Economic Research, 2003 (citado na pg. 4).

[Bernardo e A. F. M. Smith 1994] José M. Bernardo e Adrian F. M. Smith. *Bayesian Theory*. JOHN WILEY & SONS, LTD, 1994 (citado na pg. 9).

[Claeskens e Hjort 2008] Gerda Claeskens e Nils Lid Hjort. *Model Selection and Model Averaging*. Cambridge University Press, 2008. doi: 10.1017/CBO9780511790485 (citado na pg. 17).

[Clarke 2003] Bertrand Clarke. "Comparing bayes model averaging and stacking when model approximation error cannot be ignored". Em: *The Journal of Machine Learning Research* 4 (nov. de 2003), pgs. 683–712. doi: 10.1162/153244304773936090 (citado na pg. 46).

[Clyde e Iversen 2013] Merlise Clyde e Ed Iversen. "Bayesian model averaging in the m-open framework". Em: jan. de 2013, pgs. 484–498. isbn: 0191647004. doi: 10.1093/acprof:oso/9780199695607.003.0024 (citado nas pgs. 9, 48).

[Davidson e Fan 2006] Ian Davidson e Wei Fan. "When efficient model averaging out-performs boosting and bagging". Em: vol. 4213. Set. de 2006, pgs. 478–486. isbn: 978-3-540-45374-1. doi: 10.1007/11871637_46 (citado na pg. 46).

[D. Draper *et al.* 1987] David Draper, James S. Hodges, Edward E. Leamer, Carl N. Morris e Donald B. Rubin. *A Research Agenda for Assessment and Propagation of Model Uncertainty*. RAND Corporation, 1987 (citado na pg. 1).

[N. R. Draper e H. Smith 1998] Norman R. Draper e Harry Smith. *Applied Regression Analysis*. Abr. de 1998, pgs. 299–326. ISBN: 978-0-471-17082-2 (citado nas pgs. 32, 39).

[Feldkircher *et al.* 2009] Martin Feldkircher, Martin e Zeugner Stefan. "Benchmark priors revisited: on adaptive shrinkage and the supermodel effect in bayesian model averaging". Em: *IMF Working Papers 09/202, International Monetary Fund* 09 (jan. de 2009). DOI: 10.5089/9781451873498.001 (citado nas pgs. 15, 16).

[Fernández *et al.* 2001] Carmen Fernández, Eduardo Ley e Mark F.J. Steel. "Benchmark priors for bayesian model averaging". Em: *Journal of Econometrics* 100.2 (2001), pgs. 381–427. ISSN: 0304-4076. DOI: https://doi.org/10.1016/S0304-4076(00)00076-2 (citado na pg. 16).

[Foster e George 1994] Dean P. Foster e Edward I. George. "The risk inflation criterion for multiple regression". Em: *The Annals of Statistics* 22.4 (1994), pgs. 1947–1975. DOI: 10.1214/aos/1176325766 (citado na pg. 16).

[Genell *et al.* 2010] Anna Genell, Szilard Nemes, Gunnar Steineck e Paul W. Dickman. "Model selection in medical research: a simulation study comparing bayesian model averaging and stepwise regression". Em: *BMC Medical Research Methodology* 10 (dez. de 2010) (citado na pg. 5).

[George e McCulloch 1993] Edward I. George e Robert E. McCulloch. "Variable selection via gibbs sampling". Em: *Journal of the American Statistical Association* 88.423 (1993), pgs. 881–889. DOI: 10.1080/01621459.1993.10476353 (citado na pg. 13).

[Good 1952] I. J. Good. "Rational decisions". Em: *Journal of the Royal Statistical Society. Series B (Methodological)* 14.1 (1952), pgs. 107–114. ISSN: 00359246 (citado na pg. 8).

[Hanley e Mcneil 1983] J.A. Hanley e Barbara Mcneil. "A method of comparing the areas under receiver operating characteristic curves derived from the same cases". Em: *Radiology* 148 (out. de 1983), pgs. 839–43. DOI: 10.1148/radiology.148.3.6878708 (citado na pg. 42).

[Hastie *et al.* 2009] Trevor Hastie, Robert Tibshirani e Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Jan. de 2009, pgs. 241–249. ISBN: 9780387848570. DOI: 10.1007/978-0-387-84858-7 (citado na pg. 12).

[Hoerl e Kennard 1970] Arthur E. Hoerl e Robert W. Kennard. "Ridge regression: applications to nonorthogonal problems". Em: *Technometrics* 12.1 (1970), pgs. 69–82. ISSN: 00401706 (citado na pg. 1).

[Hoeting *et al.* 1999] Jennifer A. Hoeting, David Madigan, Adrian E. Raftery e Chris T. Volinsky. "Bayesian model averaging: a tutorial (with comments by m. clyde, david draper and e. i. george, and a rejoinder by the authors)". Em: *Statistical Science* 14.4 (1999), pgs. 382–417. doi: 10.1214/ss/1009212519 (citado nas pgs. 1, 6, 11, 13, 14).

[Horvath 2011] Roman Horvath. "Research & development and growth: a bayesian model averaging analysis". Em: *Economic Modelling* 28.6 (2011), pgs. 2669–2673. issn: 0264-9993 (citado na pg. 5).

[Hoyos 2020] Alejandra Estefanía Patiño Hoyos. "Adaptive significance levels in linear regression models". Tese de dout. São Paulo, SP, Brasil: Institute of Mathematics e Statistics, University of São Paulo, 2020 (citado nas pgs. 18, 51, 52).

[Jiyuan *et al.* 2017] Wang Jiyuan, Peng Geng e Wang. Shouyang. "Model selection on tourism forecasting: a comparison between bayesian model averaging and lasso". Em: *African Journal of Business Management* 11 (2017), pgs. 158–167 (citado na pg. 46).

[Kaplan 2021] David Kaplan. "On the quantification of model uncertainty: a bayesian perspective". Em: 86 (2021), pgs. 215–238. doi: 10.1007/s11336-021-09754-5 (citado nas pgs. 1, 5).

[Kass e A. E. Raftery 1995] Robert E. Kass e Adrian E. Raftery. "Bayes factors". Em: *Journal of the American Statistical Association* 90.430 (1995), pgs. 773–795. issn: 01621459 (citado nas pgs. 3, 9, 11, 20).

[Kass e Wasserman 1995] Robert E. Kass e Larry Wasserman. "A reference bayesian test for nested hypotheses and its relationship to the schwarz criterion". Em: *Journal of the American Statistical Association* 90.431 (1995), pgs. 928–934. issn: 01621459 (citado na pg. 15).

[Leamer 1978] Edward E. Leamer. *Specification Searches: Ad Hoc Inference With Nonexperimental Data*. 1ª ed. New York: Wiley, 1978 (citado nas pgs. 1, 3, 5).

[Leamer 1983] Edward E. Leamer. "Let's take the con out of econometrics". Em: *The American Economic Review* 73.1 (1983), pgs. 31–43. issn: 00028282 (citado nas pgs. 4, 45).

[Ley e M. F. Steel 2009] Eduardo Ley e Mark F.J. Steel. "On the effect of prior assumptions in bayesian model averaging with applications to growth regression". Em: *Journal of Applied Econometrics* 24.4 (2009), pgs. 651–674. doi: https://doi.org/10.1002/jae.1057 (citado na pg. 14).

[Liang *et al.* 2008] Feng Liang, Rui Paulo, German Molina, Merlise A Clyde e Jim O Berger. "Mixtures of g priors for bayesian variable selection". Em: *Journal of the American Statistical Association* 103.481 (2008), pgs. 410–423. doi: 10.1198/016214507000001337 (citado nas pgs. 15, 16).

[MacKay 2003] David J. C. MacKay. *Information theory, inference, and learning algorithms*. 2003, pgs. 34, 44. ISBN: 0521642981 9780521642989 (citado na pg. 47).

[Madigan e A. E. Raftery 1994] David Madigan e Adrian E. Raftery. "Model selection and accounting for model uncertainty in graphical models using occam's window". Em: *Journal of the American Statistical Association* 89.428 (1994), pgs. 1535–1546. ISSN: 01621459 (citado na pg. 11).

[Magnus e Morgan 1999] Jan R. Magnus e Mary S. Morgan. *Methodology and Tacit Knowledge: Two Experiments in Econometrics*. 1ª ed. New York: Wiley, 1999 (citado na pg. 5).

[Montgomery e Nyhan 2010] Jacob M. Montgomery e Brendan Nyhan. "Bayesian model averaging: theoretical developments and practical applications". Em: *Political Analysis* 18.2 (2010), pgs. 245–270. ISSN: 10471987, 14764989. (Acesso em 23/07/2022) (citado na pg. 5).

[Murphy 2007] Kevin Murphy. "Conjugate bayesian analysis of the gaussian distribution". Em: (nov. de 2007) (citado nas pgs. 18, 51).

[Porwal e A. E. Raftery 2022] Anupreet Porwal e Adrian E. Raftery. "Comparing methods for statistical inference with model uncertainty". Em: *Proceedings of the National Academy of Sciences* 119.16 (2022), e2120737119. DOI: 10.1073/pnas.2120737119. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.2120737119. URL: https://www.pnas.org/doi/abs/10.1073/pnas.2120737119 (citado na pg. 46).

[A. Raftery e Zheng 2003] Adrian Raftery e Yingye Zheng. "Discussion: performance of bayesian model averaging". Em: *Journal of the American Statistical Association* 98 (fev. de 2003), pgs. 931–938. DOI: 10.1198/016214503000000891 (citado nas pgs. 8, 48).

[A. E. Raftery 1995] Adrian E. Raftery. "Bayesian model selection in social research". Em: *Sociological Methodology* 25 (1995), pgs. 111–163. ISSN: 00811750, 14679531. (Acesso em 23/07/2022) (citado nas pgs. 7, 11, 20, 50).

[A. E. Raftery *et al.* 1997] Adrian E. Raftery, David Madigan e Jennifer A. Hoeting. "Bayesian model averaging for linear regression models". Em: *Journal of the American Statistical Association* 92.437 (1997), pgs. 179–191. DOI: 10.1080/01621459.1997.10473615 (citado na pg. 8).

[Roberts 1965] Harry V. Roberts. "Probabilistic prediction". Em: *Journal of the American Statistical Association* 60.309 (1965), pgs. 50–62. DOI: 10.1080/01621459.1965.10480774 (citado na pg. 3).

[Schwarz 1978] Gideon Schwarz. "Estimating the dimension of a model". Em: *The Annals of Statistics* 6.2 (1978), pgs. 461–464. ISSN: 00905364 (citado na pg. 17).

[M. F. J. Steel 2017] Mark F. J. Steel. "Model averaging and its use in economics". Em: (2017). DOI: 10.48550/ARXIV.1709.08221 (citado na pg. 45).

[Stone 1974] M. Stone. "Cross-validatory choice and assessment of statistical predictions". Em: *Journal of the Royal Statistical Society. Series B (Methodological)* 36.2 (1974), pgs. 111–147. ISSN: 00359246 (citado na pg. 12).

[Tibshirani 1996] Robert Tibshirani. "Regression shrinkage and selection via the lasso". Em: *Journal of the Royal Statistical Society. Series B (Methodological)* 58.1 (1996), pgs. 267–288. ISSN: 00359246 (citado na pg. 1).

[Tierney e Kadane 1986] Luke Tierney e Joseph B. Kadane. "Accurate approximations for posterior moments and marginal densities". Em: *Journal of the American Statistical Association* 81.393 (1986), pgs. 82–86. DOI: 10.1080/01621459.1986.10478240 (citado nas pgs. 11, 49).

[Viallefont *et al.* 2001] Valerie Viallefont, Adrian Raftery e Sylvia Richardson. "Variable selection and bayesian model averaging in case-control studies". Em: *Statistics in medicine* 20 (nov. de 2001), pgs. 3215–30. DOI: 10.1002/sim.976 (citado na pg. 5).

[Wasserman 2000] Larry Wasserman. "Bayesian model selection and model averaging". Em: *Journal of Mathematical Psychology* 44.1 (2000), pgs. 92–107. ISSN: 0022-2496. DOI: https://doi.org/10.1006/jmps.1999.1278 (citado nas pgs. 10, 15).

[Wright 2008] Jonathan H. Wright. "Bayesian model averaging and exchange rate forecasts". Em: *Journal of Econometrics* 146.2 (2008), pgs. 329–341. ISSN: 0304-4076. DOI: https://doi.org/10.1016/j.jeconom.2008.08.012 (citado na pg. 5).

[Yao *et al.* 2018] Yuling Yao, Aki Vehtari, Daniel Simpson e Andrew Gelman. "Using stacking to average bayesian predictive distributions (with discussion)". Em: *Bayesian Analysis* 13.3 (2018), pgs. 917–1007. DOI: 10.1214/17-BA1091 (citado na pg. 46).

[Young 2009] Cristobal Young. "Model uncertainty in sociological research: an application to religion and economic growth". Em: *American Sociological Review* 74.3 (2009), pgs. 380–397. DOI: 10.1177/000312240907400303 (citado na pg. 4).

[Zellner 1986] Arnold Zellner. "On assessing prior distributions and bayesian regression analysis with g-prior distributions". Em: *Bayesian Inference & Decision Techniques* 6 (1986), pgs. 233–243 (citado na pg. 15).

[Zeugner e Feldkircher 2015] Stefan Zeugner e Martin Feldkircher. "Bayesian model averaging employing fixed and flexible priors: the bms package for r". Em: *Journal of Statistical Software* 68.4 (2015), pgs. 1–37. DOI: 10.18637/jss.v068.i04 (citado na pg. 14).

[Zhou 2011] Shouhao Zhou. “Bayesian Model Selection in terms of Kullback-Leibler discrepancy”. Tese de dout. New York, NY, USA: Graduate School of Arts e Sciences, Columbia University, 2011 (citado na pg. 48).

[Zou e Hastie 2005] Hui Zou e Trevor Hastie. “Regularization and variable selection via the elastic net”. Em: *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 67.2 (2005), pgs. 301–320. issn: 13697412, 14679868 (citado na pg. 1).