Universidade de Brasília
IE - Instituto de Exatas
Departamento de Estatística

Uso da capacidade preditiva como critério bayesiano de adequação de modelos

Gabriel Hideki Vatanabe Brunello

Brasília
2017

Gabriel Hideki Vatanabe Brunello

Uso da capacidade preditiva como critério bayesiano de adequação de modelos

Documento apresentado ao Departamento de Estatística da Universidade de Brasília, como requisito para a obtenção do título de Mestre em Estatística.

Orientador(a): Dr: Eduardo Yoshio Nakano

Brasília
2017

Brunello, G. H. V.
Uso da capacidade preditiva como critério bayesiano de adequação de modelos
57 páginas
Dissertação - Instituto de Exatas da Universidade de Brasília. Departamento de Estatística.

1. Inferência Bayesiana
2. Qualidade de Modelo
3. Leave-one-out

I. Universidade de Brasília. Instituto de Exatas. Departamento de Estatística.

Comissão Julgadora:

Prof. Dr: Gustavo Leonel Gilardoni Avalle

Prof. Dr: Carlos Alberto de Bragança Pereira

Prof. Dr: Eduardo Yoshio Nakano
Orientador

Agradecimentos

Primeiramente aos meus pais, pelo apoio que me deram durante todos esses anos, sempre me incentivando a estudar e me esforçar.

A todos os professores e professoras que contribuíram direta e indiretamente na minha formação, tanto no período da graduação como agora no mestrado. Em especial, agradeço ao meu orientador, professor Eduardo Nakano, pela paciência e por toda a sabedoria transmitida.

Aos professores membros da banca examinadora, Gustavo Gilardoni e Carlos Pereira, por aceitarem o convite para participar da defesa.

Aos meus colegas do mestrado e do trabalho, pelos conselhos e o motivações.

A Roberto, um irmão que conheci na estatística, pela amizade e companheirismo em todos esses anos de universidade e fora dela.

Ao meu grande amor, Karine, pela atenção, carinho e estar sempre ao meu lado, me apoiando e incentivando a superar minhas dificuldades.

Por fim, agradeço a todos que de forma direta ou indireta contribuíram para o meu crescimento profissional e acadêmico.

Resumo

Certificar-se de que o modelo probabilístico proposto representa satisfatoriamente o problema é um dos principais passos na modelagem estatística, pois a escolha de um modelo que não esteja bem ajustado pode provocar prejuízos irreparáveis com a tomada de uma decisão errada. Frequentemente, o objetivo da modelagem estatística é a predição de novas observações, fazendo com que avaliar a acurácia do modelo seja imprescindível. Porém, métodos que analisam a capacidade preditiva de um modelo não são muito utilizados por sua complexidade. Este trabalho apresentou uma adaptação para a metodologia de verificação da capacidade preditiva de um modelo proposta por Gelfand (1996), que apesar de simples e intuitiva, não permitia a validação de modelos de uma maneira objetiva. A adaptação possibilitou a definição de um critério de rejeição de modelos, por meio de estudos de simulação, proporcionando um meio de discriminação imparcial para a metodologia. O desenvolvimento da proposta foi realizado sob uma perspectiva bayesiana de inferência, expondo os conceitos utilizados em sua elaboração e apresentando os procedimentos necessários para sua aplicação. A metodologia proposta foi aplicada a base de dados reais para exemplificar sua utilização, possibilitando verificar a praticidade de sua aplicação em situações reais.

Palavras-chave: Inferência Bayesiana, Qualidade de Modelo, Leave-one-out

Abstract

To ensure that a proposed probability model is a good representation of the problem is one of the main steps of statistical modelling, since choosing a model that does not have a good fit may lead to wrong decisions. Often, the aim of the statistical modelling is the prediction of new observations, making it necessary to ensure the model accuracy. This work provides an adjustment to Gelfand (1996) methodology to validate the model predictive capacity, which, although simple, does not allow an objective form of validation. The adjustment allowed the definition of a model rejection criterion, providing an impartial method to ensure model accuracy. The development of the adjustment was done on a bayesian inference approach, presenting the employed concepts and the necessary procedures to the application. The methodology was tested on a real database, exhibiting the practicality of the method on real applications.

Keywords: Bayesian Inference, Model Quality, Leave-one-out

Lista de Figuras

Lista de Tabelas

Sumário

Lista de Abreviações

Abreviações

CPO Ordenada Preditiva Condicional

DPA Distribuição da Proporção de Acertos

HPD Highest Posterior Density

IC Intervalo de Credibilidade

LOO Leave-One-Out

LPML Log Pseudo Marginal Likelihood

MCMC Markov Chain Monte Carlo

MLG Modelo Linear Generalizado

Capítulo 1

Introdução

Para a tomada de decisões, o entendimento do problema é fundamental, mas eventualmente envolve a compreensão de grande quantidade de dados, que devido ao seu volume, apresenta complexa estrutura de relacionamentos. Nessas situações, perceber importantes relações entre dados pode não ser trivial, fazendo com que seja necessária a aplicação de metodologias de análise.

A modelagem estatística é uma ferramenta que pode ser utilizada para facilitar a compreensão, resumindo dados em informações que geram conhecimento sobre o problema de interesse. Ela é o processo em que se tenta representar, de forma simplificada, algum fenômeno ou evento através de modelos probabilísticos. Esses modelos são generalizações da realidade baseadas em conjuntos de dados com características do problema, sendo utilizados para predizer o comportamento de novos acontecimentos do evento.

Por ser uma aproximação da realidade, a modelagem estatística é suscetível a erros, de forma que torna-se necessário verificar se o modelo utilizado representa satisfatoriamente o problema de interesse. Um modelo mal especificado prejudica a qualidade das informações obtidas, fazendo com que sejam imprecisas, o que ocasiona conclusões equivocadas.

Existem diversas maneiras de se verificar a qualidade de um modelo, mas a grande maioria apresenta critérios de classificação subjetivos ou complexa elaboração, dificultando a sua utilização em aplicações cotidianas.

Dessa forma, este trabalho apresentará uma proposta de metodologia bayesiana para verificar a qualidade de um modelo estatístico baseado em sua capacidade preditiva, isto

é, a eficiência do modelo em predizer valores para novos acontecimentos do problema de interesse.

A vantagem da proposta é ser de simples entendimento e, por avaliar a capacidade preditiva do modelo, não se baseia apenas na adequação aos dados já observados, o que facilita a aplicação e incentiva a utilização em problemas de tomada de decisões.

Apesar de ser uma metodologia de simples entendimento, até a presente data esta metodologia não possui um critério objetivo de classificação da qualidade do modelo. Por esse motivo, o presente trabalho estudou seu comportamento por meio de aplicações simuladas em modelos lineares generalizados com distribuição Exponencial.

Com os resultados do estudo, buscou-se elaborar um critério objetivo de classificação que permitisse um uso prático e imparcial da metodologia.

Os objetivos deste trabalho foram apresentar uma proposta de metodologia bayesiana de adequação de modelos baseada em sua capacidade preditiva, estudar o comportamento dessa metodologia em modelos lineares generalizados com distribuição Exponencial e elaborar um critério objetivo para a metodologia que classifique a qualidade de um modelo.

O critério de classificação proposto foi obtido por meio de dados simulados e o mesmo foi ilustrado em um conjunto de dados reais obtido na literatura. Todas as simulações e análises foram realizadas por meio do software livre R.

Capítulo 2

Qualidade de Modelos Estatísticos

Certificar-se de que o modelo probabilístico proposto é adequado foi considerado por Box (1980) como um dos principais passos na modelagem estatística, pois a escolha de um modelo que não esteja bem ajustado pode provocar prejuízos irreparáveis com a tomada de uma decisão errada.

Qualidade é um conceito pessoal, pois algo que é adequado para alguns pode não ser satisfatório para outros, fazendo com seja necessário a elaboração de um critério objetivo de avaliação de modelos.

2.1 Conceitos de Qualidade

Cada problema possui necessidades específicas que devem ser incorporadas durante a modelagem, buscando representar da melhor maneira possível o comportamento real do evento de interesse. Consequentemente, diversos critérios foram elaboradas para avaliar a qualidade de um modelo, cada um aferindo uma característica que considera mais relevante.

Inúmeras técnicas para a verificação da qualidade de um modelo podem ser encontradas na literatura, mas cada uma avalia um critério específico, fazendo com que seja necessário conhecer qual o mais apropriado para o problema, pois atender a um critério não garante o atendimento dos outros.

Assim, para facilitar o entendimento, os critérios utilizados nas principais técnicas de avaliação de qualidade estão descritos abaixo:

- ***Adequabilidade do Modelo (Model Adequacy)***: verifica se o modelo atende os pressupostos probabilísticos assumidos em sua elaboração;
- ***Aderência do Modelo (Goodness of Fit ou Lack of Fit)***: avalia a discrepância entre os valores observados e os valores esperados pelo modelo;
- ***Seleção de Modelos (Model Selection ou Model Specification)***: realiza de comparações entre múltiplos modelos, identificando qual deles apresenta o melhor desempenho;
- ***Acurácia do Modelo (Model Accuracy)***: analisa a capacidade preditiva do modelo, isto é, a capacidade de predizer corretamente observações não utilizadas.

2.2 Verificação de Modelos

A avaliação da qualidade de um modelo é uma área com vasta literatura no paradigma frequentista, dispondo de diversas técnicas que permitem avaliar a qualidade de maneira objetiva. O livro de D'Agostino (1986) por exemplo, apresenta um levantamento dos mais importantes testes de ajustamento clássicos.

De maneira oposta, no cenário bayesiano a literatura é recente e os métodos disponíveis ainda são restritivos ou complexos, fazendo com que essa parte essencial da análise estatística seja ocasionalmente negligenciada.

Analisar a qualidade do modelo em um contexto bayesiano, ao contrário do caso clássico, não depende da adequação da função de verossimilhança utilizada, mas sim da adequação da distribuição *a posteriori* utilizada, pois qualquer inferência de interesse para o problema será calculada a partir dela. A distribuição *a posteriori* é a ponderação entre a informação prévia sobre o assunto e a obtida pelos dados, contendo toda a informação conhecida sobre o problema.

É sugerido em Paulino et al. (2003) que a qualidade de um modelo bayesiano seja avaliada por sua distribuição preditiva, pois caso os dados não se adequem a distribuição preditiva, espera-se que o modelo não seja adequado.

Vehtari et al. (2012) reune diversas técnicas para a avaliação da qualidade de um modelo em um contexto bayesiano, detalhando a elaboração e uso de cada uma. No entanto, as

técnicas mais utilizadas são: os Processos Dirichlet (Polidoro, 2014), o *posterior preditive check* (Gelman et al., 2014) e a *Log Pseudo Marginal Likelihood* (Chen et al., 2008). Na qual cada técnica emprega uma abordagem e um critério diferente de validação da qualidade.

Os Processos Dirichlet (Ferguson, 1973) são utilizados na estimação de um modelo bayesiano não-paramétrico, que é comparado com o modelo proposto por meio do Fator de Bayes (Kass R. E., 1995), verificando se a diferença entre eles é significativa. Esta é uma técnica de aderência do modelo que utiliza a distância entre o valor estimado pelo modelo proposto e o estimado pelo modelo não paramétrico como critério de qualidade, tendo como desvantagem um complexo processo para sua elaboração. O Processo Dirichlet, suas variações e outros modelos são estudados em Polidoro (2014).

A metodologia mais utilizada atualmente para a verificação de qualidade de modelos bayesianos é o *posterior predictive check* (Gelman et al., 2014), que verifica se alguma estatística T do modelo é compatível com a observada empiricamente. O *posterior predictive check* é uma técnica de aderência do modelo, fazendo uma dupla utilização dos dados, pois são aproveitados tanto na estimação do modelo quanto na comparação com a estatística do teste. A subjetividade na elaboração da estatística T é uma das maiores críticas enfrentadas por essa abordagem, pois deve ser definida de acordo com cada problema. Mais informações sobre a técnica podem ser encontradas em Gelman et al. (2014), Sinharay e Stern (2003) e Kruschke (2013).

O método *Log Pseudo Marginal Likelihood* (LPML) consiste em realizar a avaliação por meio da Ordenada Preditiva Condicional, CPO (Chen et al., 2008), que é a densidade preditiva da observação no modelo estimado sem ela. Esta metodologia realiza uma seleção de modelos por meio de sua capacidade preditiva, calculando uma estatística que indica qual o melhor modelo a ser utilizado. No entanto, a estatística obtida não permite avaliar se o modelo utilizado é um bom modelo, apenas se ele é melhor que os outros ao qual foi comparado.

Frequentemente, o objetivo da modelagem estatística é a predição de novas observações, fazendo com que avaliar a acurácia do modelo seja imprescindível. Porém, verificar a capacidade preditiva de um modelo não é um critério usualmente utilizado, pois necessita de uma grande quantidade de dados, separando uma parcela para estimação e outra para

predição, ou uso de técnicas de reamostragem, que estimam vários modelos retirando em cada um deles algumas observações para que sejam preditas. No entanto, o uso de alternativas como a programação em paralelo, o desenvolvimento de técnicas de aproximação ou até mesmo a implementação de programações mais eficientes, tornam cada vez mais viável a sua aplicação.

Gelman et al. (1996) sugere um procedimento em que se calcula um intervalo preditivo de probabilidade $1 - \alpha$ para observações não utilizadas na modelagem, verificando a quantidade de observações que se encontram dentro desses intervalos, pois ela deve ser próxima da credibilidade $100(1 - \alpha)\%$ definida. Este procedimento, apesar de intuitivo, não é muito utilizado pois o critério de rejeição é subjetivo, dado que a discriminação da qualidade pode variar conforme o ponto de vista do usuário.

Este trabalho busca adaptar a metodologia sugerida por Gelman et al. (1996), pois é uma maneira intuitiva de se verificar a acurácia de um modelo. Para esse fim, alterações serão feitas em alguns dos passos do método, possibilitando a definição de uma regra objetiva para a validação de modelos.

Capítulo 3

Metodologia

Na elaboração da proposta de um critério bayesiano para a adequação de modelos utilizando a capacidade preditiva, a compreensão de técnicas que serão descritas neste capítulo é essencial, pois a proposta se baseia em conceitos de inferência bayesiana e técnicas de validação cruzada, sendo estudada com aplicações em modelos lineares generalizados com distribuição exponencial.

3.1 Distribuição Exponencial

A distribuição Exponencial é uma distribuição contínua de probabilidade que possui suporte nos números reais positivos e assimetria positiva. Ela descreve o tempo entre eventos independentes que ocorrem a uma taxa de tempo constante, sendo amplamente utilizada em aplicações estatísticas nas áreas de saúde e confiabilidade.

A densidade de probabilidade da distribuição é dada por

$$f(x|\lambda) = \lambda e^{-\lambda x}, x > 0, \lambda > 0, \tag{3.1}$$

e sua forma é apresentada na Figura 3.1.

A média e a variância da distribuição são, respectivamente,

$$E\left[X|\lambda\right] = \frac{1}{\lambda}, \tag{3.2}$$

e

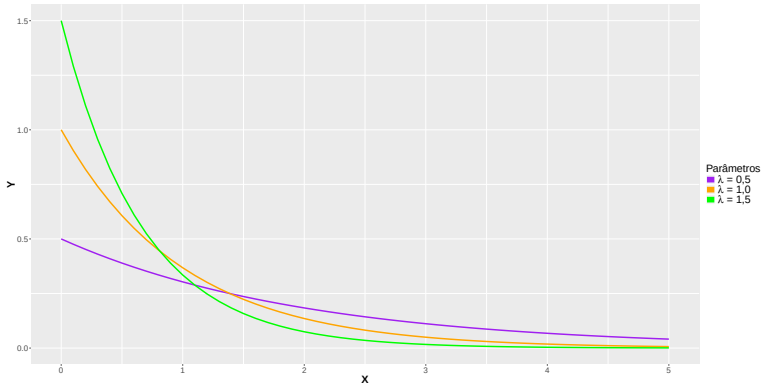


Figura 3.1: Densidade da distribuição Exponencial

$$Var\left[X|\lambda\right] = \frac{1}{\lambda^2}. \tag{3.3}$$

A distribuição Exponencial é um caso particular da distribuição Gama, dada por

$$f(x) = \frac{\lambda^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}, x > 0, \lambda > 0, \alpha > 0, \tag{3.4}$$

quando o parâmetro de forma α é igual a 1.

3.2 Inferência Bayesiana

Em um estudo, devido a experiência ou conhecimento, é natural que pesquisadores possuam diferentes graus de incerteza em relação a um parâmetro de interesse. No paradigma bayesiano, a incerteza a respeito de um parâmetro θ pode ser representada por meio de um modelo probabilístico que é atualizado à medida em que se obtém mais informações sobre o problema.

Sendo θ um parâmetro de interesse desconhecido e a informação conhecida sobre ele esteja resumida probabilisticamente em uma distribuição de probabilidades $h(\theta)$, uma quantidade aleatória observada $X = x$, relacionada a θ por meio da distribuição amostral $f(x|\theta)$, pode ser utilizada para diminuir incertezas em relação ao parâmetro θ.

Todavia, é importante notar que a distribuição *a priori* $h(\theta)$ deve ser representada por uma forma funcional e seus parâmetros, chamados de *hiperparâmetros*, devem ser identificados de acordo com o conhecimento prévio do problema.

O processo de atualização da incerteza utilizando a informação de $f(x|\theta)$ é realizado por meio do Teorema de Bayes,

$$h(\theta|x) = \frac{f(x|\theta)h(\theta)}{\int h(\theta)f(x|\theta)d\theta}, \tag{3.5}$$

que utiliza o conhecimento *a priori* $h(\theta)$ do problema e a informação adquirida observando X para obter estimativas da distribuição *a posteriori* $h(\theta|x)$.

A distribuição *a posteriori* $h(\theta|x)$ contém toda a informação probabilística a respeito do parâmetro e é usualmente representada como

$$h(\theta|x) \propto f(x|\theta)h(\theta), \tag{3.6}$$

pois o denominador da expressão 3.5 não depende de θ e é apenas uma constante normalizadora.

Por exemplo, considerando uma amostra X_1,X_2,...,X_n de uma distribuição Exponencial com parâmetro θ (3.1), com distribuição *a priori* $\theta \sim Gama(\alpha,\beta)$, tem-se que a distribuição *a posteriori* é dada por

$$h(\theta|x) \propto \theta^n e^{-\theta\sum x_i}\theta^{\alpha-1}e^{-\beta\theta} = \theta^{\alpha+n-1}e^{-\theta(\beta+\sum x_i)}, \tag{3.7}$$

que é uma distribuição $Gama(\alpha+n,\beta+\sum x_i)$.

A partir da distribuição *a posteriori* é possível obter estimativas para o parâmetro, de maneira pontual ou intervalar, que usualmente é o principal objetivo em uma pesquisa.

3.2.1 Estimação Pontual

Em uma estimação pontual, uma função perda $L(a,\theta|x)$ deve ser definida, associando pesos aos resultados das possíveis decisões a serem tomadas. Sendo a uma possível estimativa de θ, e com base em $L(a,\theta|x)$, se calcula o valor de a que minimiza o valor esperado dessa função.

A função de perda quadrática é a mais utilizada pois é obtida pela média *a posteriori*,

$$\hat{\theta} = E\left[\theta|x\right] = \int_{\Theta} \theta h(\theta|x)d\theta, \tag{3.8}$$

que consiste em uma medida mais intuitiva, usualmente prática de se determinar e gera estimativas próximas aos estimadores de máxima verossimilhança.

Para a distribuição *a posteriori* $Gama(\alpha + n, \beta + \sum x_i)$ (3.7), tem-se que a estimativa pontual da média *a posteriori* é dada por

$$\hat{\theta} = \int_{\Theta} \theta h(\theta|x) d\theta = \int_{\Theta} \frac{(\beta + \sum x_i)^{\alpha+n}}{\Gamma(\alpha + n)} \theta^{\alpha+n-1} e^{-(\beta+\sum x_i)\theta} d\theta = \frac{\alpha + n}{\beta + \sum x_i}. \tag{3.9}$$

3.2.2 Estimação Intervalar

Na estimação intervalar especifica-se um intervalo de credibilidade (IC) que fornece a probabilidade *a posteriori* do parâmetro pertencer a ele, isto é, C é um intervalo com $100\gamma\%$ de credibilidade para θ se $P(\theta \in C) \geq \gamma$.

Devido a sua definição, infinitos intervalos com a mesma credibilidade podem ser criados, fazendo com que seja necessário definir um critério para a elaboração do intervalo que será calculado.

Os critérios que serão utilizados neste trabalho são :

- **Simétrico:** Um intervalo C de $100\gamma\%$ de credibilidade para θ é Simétrico se $P[\theta|x \leq c_1] = P[\theta|x \geq c_2] = \frac{1-\gamma}{2}$, onde c_1 e c_2 são os limites de C e $c_1 < c2$.
- **Highest Posterior Density (HPD):** Um intervalo C de $100\gamma\%$ de credibilidade para θ é HPD se $C = \{\theta \in \Theta : h(\theta|Y) \geq k(\gamma)\}$, no qual $k(\gamma)$ é a maior constante tal que $P(\theta \in C) \geq \gamma$.

3.2.3 Distribuição Preditiva

Após a obtenção da distribuição *a posteriori* para θ, realizar predições a respeito de novas observações é, usualmente, uma das finalidades da análise. Para predizer um valor não observado y relacionado a θ, utiliza-se a distribuição preditiva, pois aplica toda a informação já obtida sobre o parâmetro na criação de estimativas pontuais ou intervalares para a nova observação.

A distribuição preditiva é dada por

$$p(y|x) = \int f(y,\theta|x)d\theta = \int f(y|\theta)h(\theta|x)d\theta = E_{\theta|x}\left[f(y|\theta)\right], \tag{3.10}$$

e é de extrema importância para a inferência bayesiana, pois é apenas por meio dela que se torna possível extrapolar os dados e obter estimativas de valores ainda não observados.

Como exemplo, considere $X = [X_1,X_2,...,X_n]$ uma amostra que, dado λ, segue uma distribuição exponencial com parâmetro λ, isto é,

$$f_X(x_i|\theta) = \lambda e^{\lambda x_i}, i = 1,2,...,n. \tag{3.11}$$

Assumindo, a priori, que $\lambda \sim Gama(a,b), a,b > 0$, é fácil de ver que $\lambda|X \sim Gama(a + n, b + \sum X_i)$.

Assim, a distribuição preditiva de uma nova observação Y é dada por

$$F_Y[y|X] = P\left[Y \leq y|X\right] = E_{\lambda|X}\left[P[Y \leq y|X]\right] = E_{\lambda|X}\left[1 - e^{\lambda y}\right] = 1 - E_{\lambda|X}\left[e^{\lambda y}\right] = 1 - \int_0^\infty e^{-\lambda y}\frac{(b+\sum x_i)^{a+n}}{\Gamma(a+n)}\lambda^{a+n-1}e^{-\lambda(b+\sum x_i)}d\lambda = 1 - \left(\frac{b+\sum x_i}{b+\sum x_i+y}\right)^{a+n}, y > 0, \tag{3.12}$$

e a densidade preditiva de uma nova observação é dada por

$$f_Y(y|x) = \frac{\partial}{\partial y}F_Y(y|x) = (a+n)\frac{(\beta+\sum x_i)^{a+n}}{(\beta+\sum x_i+y)^{a+n+1}}, y > 0. \tag{3.13}$$

Dessa forma, o quantil q da preditiva $Y|X$ é

$$y_q = \left(b + \sum x_i\right)\left[(1-q)^{\frac{-1}{a+n}} - 1\right]. \tag{3.14}$$

Logo, o intervalo simétrico com $100\gamma\%$ de credibilidade do valor preditivo y, dado a amostra $X = (X_1,...,X_n)$ é

$$IC_{100\gamma\%} : \left[\left(b + \sum x_i\right)\left[\left(1 - \frac{\gamma}{2}\right)^{\frac{-1}{a+n}} - 1\right]; \left(b + \sum x_i\right)\left[\left(\frac{\gamma}{2}\right)^{\frac{-1}{a+n}} - 1\right]\right]. \tag{3.15}$$

Em situações nas quais a distribuição preditiva é de difícil obtenção, pode-se estimá-la numericamente utilizando métodos de Monte Carlo (Hammersley e Handscomb, 1964):

1. Para $j = 1,...,J$, amostre $\theta^{[j]}$ da distribuição *a posteriori* $h(\theta|x)$;

2. Com os valores $\theta^{[j]}$, amostre $y^{[j]}$ da distribuição $\pi(y|\theta^{[j]})$;

3. Então $y^1,y^2,...,y^J$ serão amostras iid da distribuição preditiva *a posteriori*.

3.3 Regressão Linear

A análise de regressão linear é uma metodologia estatística que possui como objetivo verificar a existência de uma relação funcional entre uma variável dependente Y e uma ou mais variáveis independentes $X_1, X_2, ..., X_k$.

O modelo de regressão linear é dado por

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_k X_{ik} + e_i, \tag{3.16}$$

que pode ser escrito na forma matricial como

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}, \tag{3.17}$$

$$\begin{bmatrix} y_1 \\ y_2 \\ .. \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & .. & x_{1k} \\ 1 & x_{21} & x_{22} & .. & x_{2k} \\ .. & .. & .. & .. & .. \\ 1 & x_{n1} & x_{n2} & .. & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ .. \\ \beta_k \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ .. \\ e_n \end{bmatrix},$$

onde $\boldsymbol{X}$ é a matriz de variáveis dependentes; $\boldsymbol{\beta}$ é o vetor de parâmetros, que representam o efeito das variáveis $\boldsymbol{X}$ na variável dependente Y, e e representa o erro do modelo entre o valor estimado $\hat{Y}$ e o valor observado Y.

Assumindo que $e_i \overset{iid}{\sim} N(0,\sigma^2)$, a função de verossimilhança para uma amostra de tamanho n apresenta a seguinte forma:

$$L(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\beta}, \sigma^2) \propto (\sigma^2)^{-\frac{n}{2}} e^{\left(-\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})\right)}, \tag{3.18}$$

em que $\boldsymbol{\beta}$ e $\sigma \geq 0$ são os parâmetros a serem estimados. Aqui, $\boldsymbol{y} = (y_1,...,y_n)$ é o vetor de valores observados.

Como no modelo bayesiano os parâmetros $\boldsymbol{\beta}$ e $\sigma > 0$ são tratados como variáveis aleatórias, as seguintes distribuições *a priori* foram consideradas: $\boldsymbol{\sigma}^2 \sim GI(c,d)$, sendo GI a distribuição Gama Inversa dada por

$$f(\sigma^2;c,d) = \frac{d^c}{\Gamma(c)}(\sigma^2)^{-c-1}e^{-\frac{d}{\sigma^2}}, \sigma^2 > 0, \tag{3.19}$$

e $\boldsymbol{\beta}|\boldsymbol{\sigma^2} \sim \boldsymbol{N}_k(\boldsymbol{a}, \sigma^2\boldsymbol{B^{-1}})$, onde $\boldsymbol{N}_k$ é a distribuição normal k-variada

$$f_x(x_1,...,x_k|\boldsymbol{a},\boldsymbol{B}) = \frac{1}{\sqrt{(2\pi)^k|\boldsymbol{B}|}}e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{a})^T\boldsymbol{B}^{-1}(\boldsymbol{x}-\boldsymbol{a})}, \tag{3.20}$$

com vetor de médias $\boldsymbol{a}$ e matriz de covariâncias $\boldsymbol{B}$, sendo $\boldsymbol{a}, \boldsymbol{B}, c, d$ são hiperparâmetros conhecidos.

A distribuição *a posteriori* é dada por:

$$\begin{aligned}
\pi(\boldsymbol{\beta},\sigma^2|\boldsymbol{y},\boldsymbol{X}) &\propto L(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\beta},\sigma^2) \times GI(c,d) \times N_k(\boldsymbol{a},\sigma^2\boldsymbol{B^{-1}}) \\
&\propto (\sigma^2)^{-\frac{n}{2}}e^{\left(-\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X\beta})^T(\boldsymbol{y}-\boldsymbol{X\beta})\right)}(\sigma^2)^{-\frac{k}{2}}e^{\left(-\frac{1}{2\sigma^2}\left((\boldsymbol{\beta}-\boldsymbol{a})^T\boldsymbol{B}(\boldsymbol{\beta}-\boldsymbol{a})\right)\right)}(\sigma^2)^{-(c+1)}e^{-\frac{d}{\sigma^2}} \\
&\propto (\sigma^2)^{\left(-\frac{k}{2}\right)}e^{\left(-\frac{1}{2\sigma^2}(\boldsymbol{\beta}-\boldsymbol{\mu})^T(\boldsymbol{X}^T\boldsymbol{X}+\boldsymbol{B})(\boldsymbol{\beta}-\boldsymbol{\mu})\right)}(\sigma^2)^{\left(-\frac{n+2c}{2}-1\right)}e^{\left(-\frac{2d+\boldsymbol{y}^T\boldsymbol{y}-\boldsymbol{\mu}^T(\boldsymbol{X}^T\boldsymbol{X}+\boldsymbol{B})\boldsymbol{\mu}+\boldsymbol{a}^T\boldsymbol{B}\boldsymbol{a}}{2\sigma^2}\right)} \\
&\propto N_k(\boldsymbol{\mu},\sigma^2\boldsymbol{\Lambda^{-1}}) \times GI(\vartheta,\xi),
\end{aligned} \tag{3.21}$$

com,

$\boldsymbol{\mu} = (\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{B})^{-1}(\boldsymbol{Ba} + \boldsymbol{X}^T\boldsymbol{y})$,

$\boldsymbol{\Lambda} = (\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{B})$,

$\vartheta = c + \frac{n}{2}$,

$\xi = d + \frac{1}{2}(\boldsymbol{y}^T\boldsymbol{y} + \boldsymbol{a}^T\boldsymbol{Ba} - \boldsymbol{\mu}^T\boldsymbol{\Lambda\mu})$.

A distribuição *a posteriori* (3.21) pode ser escrita como $\rho(\boldsymbol{\beta}|\sigma^2,\boldsymbol{y},\boldsymbol{X})\rho(\sigma^2|\boldsymbol{y},\boldsymbol{X})$, permitindo que seja estimada numericamente, utilizando o método do amostrador de Gibbs (Gelfand e Smith, 1990).

Em um contexto bayesiano, as predições são feitas através da distribuição preditiva (3.10), diferente do caso clássico em que se substitui as estimativas de máxima verossimi-

lhança na distribuição condicional.

Para o caso de regressão linear, a predição de um ponto $\boldsymbol{\omega} = \{1, \omega_1, \omega_2, ..., \omega_k\}$ é dada por:

$$E_{\boldsymbol{\beta},\sigma^2|\boldsymbol{X}}\left[\boldsymbol{\omega\beta}\right]. \tag{3.22}$$

Esta distribuição preditiva pode ser obtida numericamente utilizando o algoritmo apresentado na Seção 3.2.3, que simula amostras independentes da distribuição, permitindo a obtenção de estimadores pontuais ou intervalares.

3.4 Modelos Lineares Generalizados

Existem situações em que se deseja ajustar um modelo de regressão linear para determinada variável, mas esta não é bem representada por uma distribuição Gaussiana. Para estes casos a alternativa é a utilização de Modelos Lineares Generalizados (MLG), que permitem o uso de diferentes distribuições de probabilidade na variável resposta, propiciando um melhor ajuste dos modelos (Nelder e Baker, 1972).

Os MLG associam por meio de uma função de ligação η, um preditor linear $\boldsymbol{X\beta}$ a uma variável resposta Y, o que permite a utilização de outras distribuições de probabilidade, como por exemplo, a distribuição Exponencial, que só aceita valores positivos como resposta.

Utilizando uma função de ligação $\eta = log(\frac{1}{\lambda}) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k$ é possível ajustar um Modelo Linear Generalizado para a distribuição Exponencial com $\lambda = e^{-\eta}$, que apresenta função de verossimilhança dada por

$$L(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\beta}) \propto \prod_{i=1}^{n} e^{-\eta_i} e^{-y_i e^{-\eta_i}} = \prod_{i=1}^{n} e^{-\eta_i - y_i e^{-\eta_i}} = \prod_{i=1}^{n} e^{-(\beta_0 + x_{i1}\beta_1 + ... + x_{ik}\beta_k) - y_i e^{-(\beta_0 + x_{i1}\beta_1 + ... + x_{ik}\beta_k)}}. \tag{3.23}$$

Para este trabalho, utilizou-se uma distribuição *a priori* $\boldsymbol{N}_k(\boldsymbol{\mu}, I\boldsymbol{\sigma})$. A distribuição *a posteriori* é então dada por

$$\pi(\boldsymbol{\beta}|\boldsymbol{X},\boldsymbol{y}) \propto L(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\beta}) \times N_k(\boldsymbol{\mu}, I\boldsymbol{\sigma})$$

$$\propto \prod_{i=1}^{n} e^{-\eta_i - y_i e^{-\eta_i}} \prod_{j=1}^{p} e^{\left(-\frac{(\beta_j-\mu_j)^2}{2\sigma_j^2}\right)}$$

$$= \prod_{i=1}^{n} e^{(\beta_0+x_{i1}\beta_1+...+x_{ik}\beta_k) - y_i e^{(\beta_0+x_{i1}\beta_1+...+x_{ik}\beta_k)}} e^{\left(-\frac{\{(\beta_0+x_{i1}\beta_1+...+x_{ik}\beta_k)-\mu_i\}^2}{2\sigma^2}\right)}. \quad (3.24)$$

Amostras dessa distribuição *a posteriori* podem ser obtidas utilizando Cadeias de Markov via Monte Carlo (MCMC) por meio do algoritmo de Metropolis-Hastings (Hastings, 1970), que permite a simulação de amostras de uma distribuição de probabilidade desde que se conheça seu núcleo.

Para o caso de um modelo linear generalizado com distribuição Exponencial, a predição de um ponto $\boldsymbol{\omega} = \{1, \omega_1, \omega_2, ..., \omega_k\}$ em um cenário bayesiano é dada por

$$E_{\boldsymbol{\beta}|\boldsymbol{X}}\left[\frac{1}{\lambda}\right] = E_{\boldsymbol{\beta}|\boldsymbol{X}}\left[e^{-\boldsymbol{\omega}\boldsymbol{\beta}}\right] = E_{\boldsymbol{\beta}|\boldsymbol{X}}\left[e^{-(\beta_0+\omega_1\beta_1+..+\omega_k\beta_k)}\right], \quad (3.25)$$

e pode ser obtida numericamente utilizando o algoritmo apresentado na Seção 3.2.3, permitindo a obtenção de estimadores pontuais ou intervalares.

3.5 Validação Cruzada Preditiva

Avaliar a qualidade preditiva de um modelo envolve predizer informações não utilizadas em sua estimação, mas nem sempre a quantidade de dados disponíveis permite a separação em duas bases. Sendo assim, a validação cruzada pode ser empregada para realizar a avaliação nessas situações.

A Validação Cruzada, ou *Cross-Validation*, consiste em estimar o modelo m vezes e em cada uma delas retirar algumas das observações, obtendo uma estatística para cada observação não utilizada. Para o caso em que as observações são retiradas uma de cada vez e o modelo é estimado n vezes, a técnica é chamada de *leave-one-out*.

3.5.1 *Leave-One-Out*

A retirada de apenas um dado observado de cada vez permite que a capacidade pre-

ditiva do modelo seja avaliada sem muitos prejuízos na qualidade da estimação, por este motivo, o *leave-one-out* (LOO) é uma técnica amplamente utilizada na validação da qualidade preditiva. A Figura 3.2 apresenta um exemplo gráfico do algoritmo LOO.

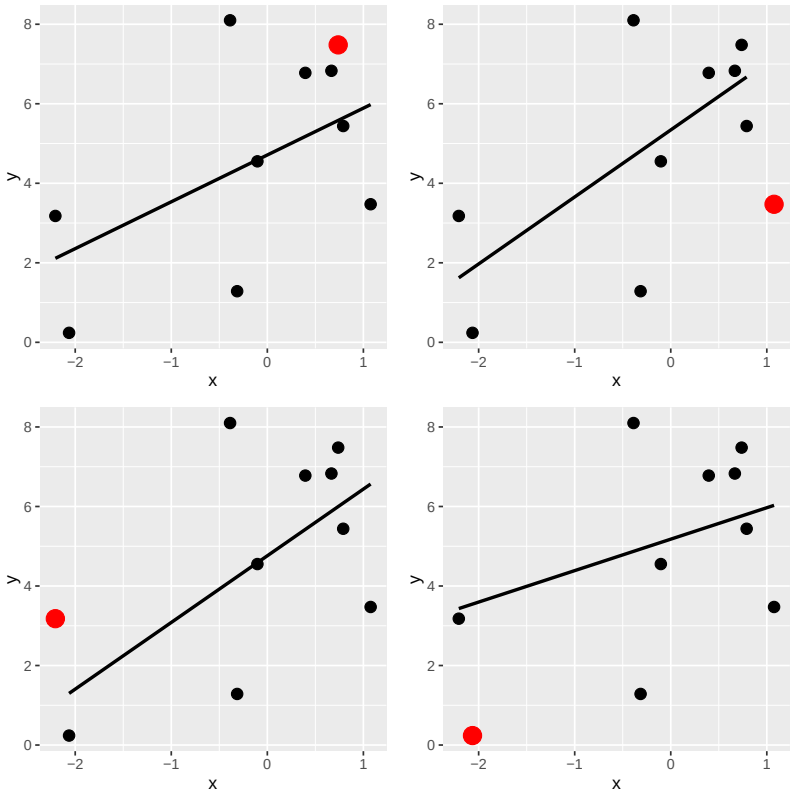


Figura 3.2: Exemplo de leave-one-out

Métodos de validação cruzada em um contexto bayesiano foram propostos por diversos autores, como Geisser e Eddy (1979), Gelfand e Smith (1990), dentre outros. Uma revisão desses e de outros métodos preditivos e de adequação de modelos em um contexto bayesiano é apresentada por Vehtari et al. (2012).

Ao obter as predições $\widehat{\boldsymbol{y}}$, a estatística calculada para o valor depende da metodologia utilizada, podendo ser a densidade do ponto ou até mesmo a distância entre o valor real e o valor predito.

Capítulo 4

Proposta de Análise de Capacidade Preditiva

Este trabalho propõe uma adaptação da abordagem sugerida por Gelman et al. (1996) para a avaliação da qualidade de um modelo por meio da capacidade preditiva de sua distribuição *a posteriori*.

A utilização da distribuição *a posteriori* garante que o modelo final é adequado, pois métodos que avaliam apenas a função verossimilhança não garantem que a distribuição *a priori* utilizada seja propícia, podendo prejudicar os resultados do modelo final.

O método proposto consiste em aplicar a técnica *leave-one-out* para avaliar a capacidade do modelo de classificar corretamente novas observações. A classificação é feita calculando, para cada observação y_i, o estimador intervalar predito C_i por meio do modelo M_i, que é o modelo estimado no i-ésimo passo da validação cruzada e não utiliza a observação $y_i, i = 1, 2,..., n$.

Caso o valor real esteja dentro do intervalo predito, ele é classificado como acerto ($\kappa = 1$) ou, caso contrário, como erro ($\kappa = 0$), isto é,

$$\kappa_i = \begin{cases} 1, & y_i \in C_i \\ 0, & c.c. \end{cases}, i = 1, 2,..., n \tag{4.1}$$

Sendo assim, a estatística calculada pela metodologia é a proporção de intervalos que contém o valor real, ou seja, a proporção de acertos obtida, que é definida por,

$$\text{Prop. de Acertos} = \sum_{i=1}^{n} \frac{\kappa_i}{n}. \tag{4.2}$$

A técnica *leave-one-out* evita a dupla utilização dos dados, como ocorre no *Posterior Preditive Check*. Além disso, uso de estimadores intervalares facilita a especificação de uma proporção de acertos esperada para o modelo, pois para um intervalo com $100\gamma\%$ de credibilidade, o percentual de acertos deve ser aproximadamente $\gamma\%$. Portanto, um valor distante de $\gamma\%$ é um indício de que o modelo não possui uma boa capacidade preditiva e não é adequado para representar o problema.

Como a quantidade de intervalos de credibilidade que predizem corretamente as observações deve ser próxima da credibilidade utilizada, optou-se pelo nível $\gamma = 0{,}5$, pois é o ponto ideal para verificar a simetria da estatística.

4.1 Algoritmo

Dada uma amostra de observações y_1,...,y_n em que cada observação y_i possui uma covariável x_i $(i = 1, 2, ..., n)$ associada a ela, os passos utilizados para a aplicação da metodologia são descritos a seguir:

1. Estimar um modelo M_i retirando a observação i da amostra;
2. Obter a distribuição preditiva do modelo M_i;
3. Com a distribuição preditiva do modelo M_i, estimar o intervalo de 50% de credibilidade para uma nova observação;
4. Verificar se o valor y_i, retirado da amostra, está dentro do intervalo predito;
5. Repetir os passos 1 a 4 para todas as observações da amostra;
6. Contar quantos intervalos de credibilidade previram corretamente.

Para a obtenção de uma aproximação numérica da distribuição preditiva mencionada no segundo passo, o seguinte procedimento pode ser utilizado:

Sendo M_i o modelo estimado sem a observação y_i com vetor de parâmetros $\beta_i = (\beta_{i0}, \beta_{i1},..., \beta_{ik})$,

1. Para $j = 1,...,J$, amostre $\boldsymbol{\beta}_i^{[j]} = (\beta_{i0}^{[j]},\beta_{i1}^{[j]},...,\beta_{ik}^{[j]})$ da distribuição *a posteriori* do modelo M_i;

2. Para cada valor de $\boldsymbol{\beta}_i^{[j]}$, gere um $y_i^{[j]}$ de uma Exponencial($\lambda_i^{[j]}$) com $\lambda_i^{[j]} = e^{X_i\beta_i^{[j]}}, j = 1, 2,..., J$;

3. Então $y_i^{[1]},y_i^{[2]},...,y_i^{[J]}$ serão amostras iid da distribuição preditiva *a posteriori* para y_i.

O procedimento simula amostras da distribuição preditiva do modelo M_i, permitindo a obtenção de qualquer estatística para uma nova obervação y_i. O intervalo simétrico de $100\gamma\%$ de credibilidade para y_i, por exemplo, pode ser obtido calculando o percentis $100\frac{\gamma}{2}\%$ e $100(1-\frac{\gamma}{2})\%$ das amostras $y_i^{[1]},y_i^{[2]},...,y_i^{[J]}$.

4.2 Aplicação Simulada

A Figura 4.1 apresenta dados gerados de uma amostra de tamanho $n = 100$ de uma distribuição Exponencial com uma única covariável X e intervalos, HPD e Simétrico, de 50% de credibilidade calculados com a metodologia proposta.

O modelo de regressão com distribuição exponencial foi escolhido por possuir apenas um parâmetro, λ, facilitando o entendimento da proposta e a testando em uma situação de distribuição assimétrica.

Em uma distribuição assimétrica, como é o caso da Exponencial, os intervalos HPD e Simétrico apresentarão regiões distintas mesmo possuindo a mesma credibilidade, como pode ser visto na Figura 4.1. O fato do intervalo HPD não conter a média se deve a credibilidade utilizada e a assimetria da distribuição exponencial, pois o intervalo HPD é dependente da moda e não da média da distribuição.

Neste exemplo, o modelo de regressão exponencial apresentou um bom ajuste preditivo, uma vez que as proporções de acerto foram de 47% e 53% para os intervalos de 50% de credibilidade Simétrico e HPD, respectivamente. É importante ressaltar que uma proporção de acertos muito superior à credibilidade utilizada não é algo benéfico, pois indica

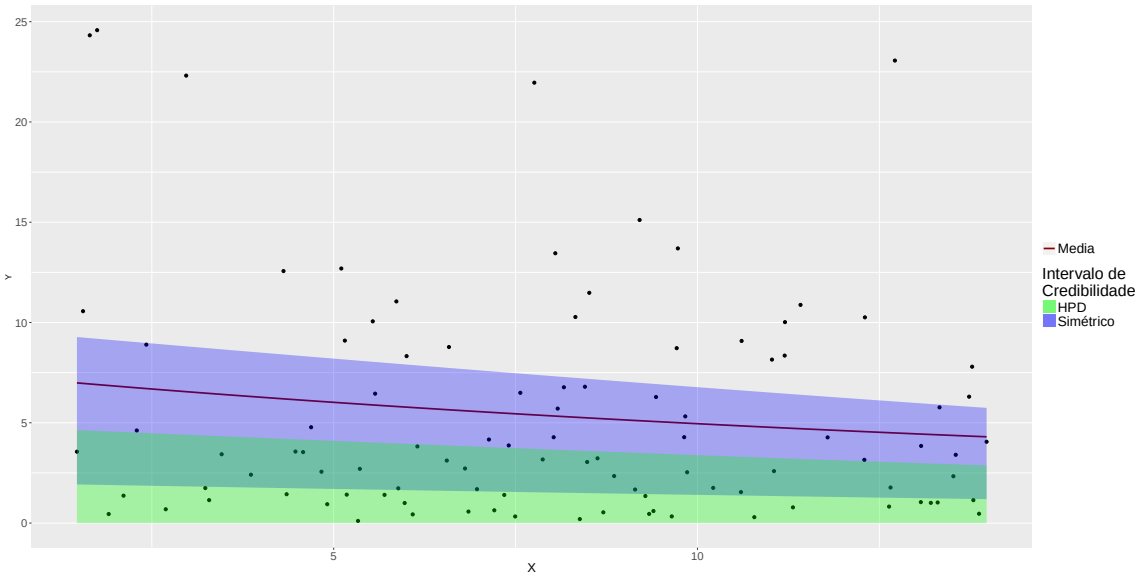


Figura 4.1: Intervalos de Credibilidade em uma Amostra Exponencial

que o modelo não possui uma boa acurácia da mesma maneira que uma proporção de acertos muito inferior.

O método proposto apresenta uma lógica semelhante à *Log Pseudo Marginal Likelihood* (Chen et al., 2008), mas com objetivos diferentes, pois a LPML realiza comparações entre modelos e a proposta verifica se um modelo é capaz de predizer bem o problema.

Para o estudo de simulações, a aproximação sugerida por Chen et al. (2012) para o cálculo do *leave-one-out* não foi utilizada, pois ela não faz uso da distribuição *a priori*, podendo não ser uma boa aproximação para pequenos tamanhos amostrais caso a distribuição *a priori* esteja mal especificada.

Uma inconveniência do método é seu alto custo computacional, pois necessita da estimação de um modelo para cada observação da amostra, tornando-o ineficiente em grandes bases de dados.

Outra dificuldade é que a proximidade entre a proporção de acertos e a credibilidade γ é um fator subjetivo, pois 44% de acertos pode ser razoável para algumas pessoas, mas para outras não. Por este motivo, é essencial solucionar o problema da subjetividade para o uso eficaz da metodologia. Portanto, para definir o quão próximo o percentual de acertos deve estar do nível de credibilidade adotado, simulações foram realizadas para estudar o comportamento da distribuição da proporção de acertos (DPA) da metodologia.

Capítulo 5

Estudos de Simulação

O método proposto para avaliar a capacidade preditiva de um modelo calcula a proporção de valores preditos corretamente e a utiliza como estatística de qualidade, verificando se o valor observado é factível para a credibilidade escolhida e rejeitando o modelo quando a proporção observada é improvável.

A aplicação da metodologia em situações reais necessita da definição de um ponto crítico que será utilizado como critério de rejeição de modelos, porém, a dificuldade na definição desse ponto consiste no mesmo poder ser influenciado por vários fatores, como por exemplo, o tipo de intervalo (HPD ou Simétrico), o tamanho da amostra utilizada na estimação, dentre outros.

Para a definição do critério, os fatores considerados mais relevantes foram:

1. O efeito do tipo de intervalo (HPD ou Simétrico);
2. Tipo de covariáveis utilizadas (numéricas ou categóricas);
3. Número de parâmetros do modelo;
4. Tamanho Amostral.

Estes fatores foram escolhidos por possivelmente influenciarem o comportamento da distribuição da proporção de acertos (DPA), fazendo com que estudar seus efeitos seja essencial na definição do ponto crítico. A fim de verificar quais desses fatores influenciam

efetivamente a *DPA*, foram realizados estudos de simulação utilizando modelos de regressão exponencial.

Todas as simulações foram realizadas com o software R, utilizando o algoritmo de Metropolis−Hastings (Hastings, 1970) para o cálculo da distribuição *a posteriori* do modelo em cada passo do *leave-one-out*. O núcleo da distribuição *a posteriori* foi o apresentado em (3.25) com distribuições *a priori* difusas $N(0{,}100)$ para os parâmetros β's. Para a aplicação do MCMC, utilizou-se a função *MCMCmetrop1R* do pacote *MCMCpack* com 20.000 passos na cadeia, *burn-in* de 1.000 e saltos de tamanho 3.

5.1 Tipo de Intervalo

Para identificar o efeito do tipo de intervalo, foram geradas 1.000 amostras de tamanho 100 de uma distribuição exponencial com parâmetro $\lambda = e^{-0{,}7+1{,}1X}$, sendo X gerado por uma distribuição $Uniforme(a = 1{,}b = 5)$. Em cada amostra obtida, o método proposto foi aplicado para os intervalos HPD e Simétricos com 50% de credibilidade.

A *DPA* para cada caso é apresentada na Figura 5.1, mostrando que apesar de ambas apresentarem em média 50% de acertos, a variância da distribuição empírica dos acertos dos intervalos Simétricos foi maior que a dos intervalos HPD. Isso ocorreu devido à natureza assimétrica da distribuição Exponencial, como observado na Figura 4.1, pois o intervalo HPD se concentra em torno da moda da distribuição, diferente do intervalo Simétrico.

Os dois tipos de intervalo apresentaram um comportamento simétrico, que era o efeito desejado ao se definir a credibilidade como 50%. Caso a credibilidade utilizada fosse 95%, a *DPA* provavelmente teria uma forma com forte assimetria negativa.

Para verificar se o método consegue identificar modelos mal ajustados, i.e., com baixa capacidade preditiva, foram geradas amostras de uma distribuição $Gama(\alpha = 1{,}5$ e $\beta = e^{-0{,}7+1{,}1X})$. A distribuição Gama foi utilizada por se tornar uma distribuição Exponencial quando seu parâmetro de forma α é 1.

A Figura 5.2 apresenta a *DPA* das amostras geradas a partir das distribuições Gama e Exponencial utilizando intervalos Simétricos e HPD, permitindo identificar diferenças no comportamento entre distribuições para os tipos de intervalo utilizado.

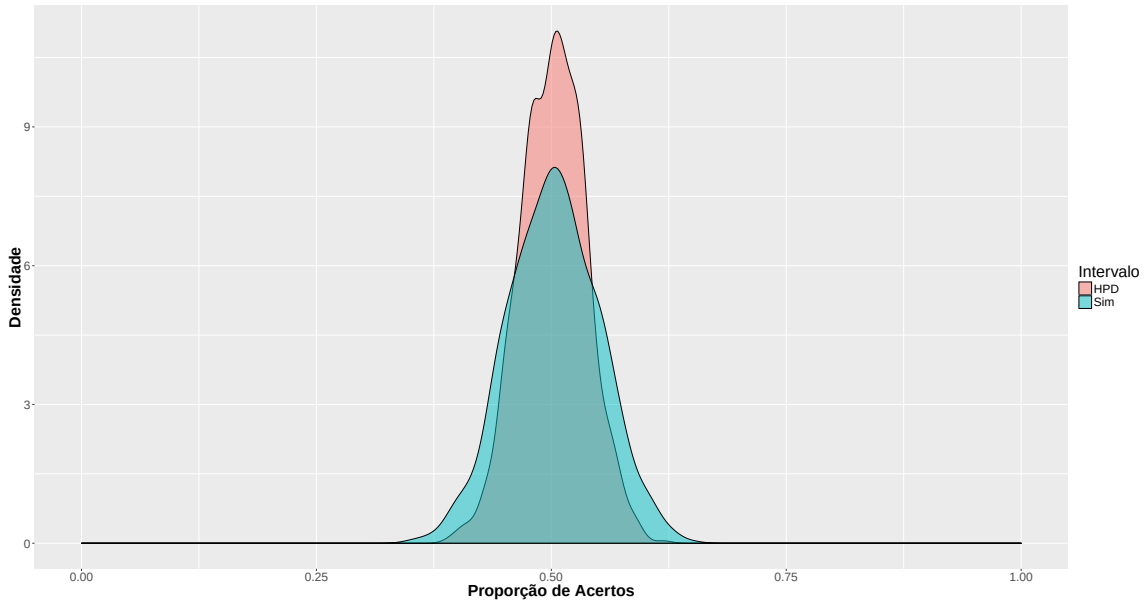


Figura 5.1: Distribuição da proporção de acertos com $\gamma = 0{,}5$

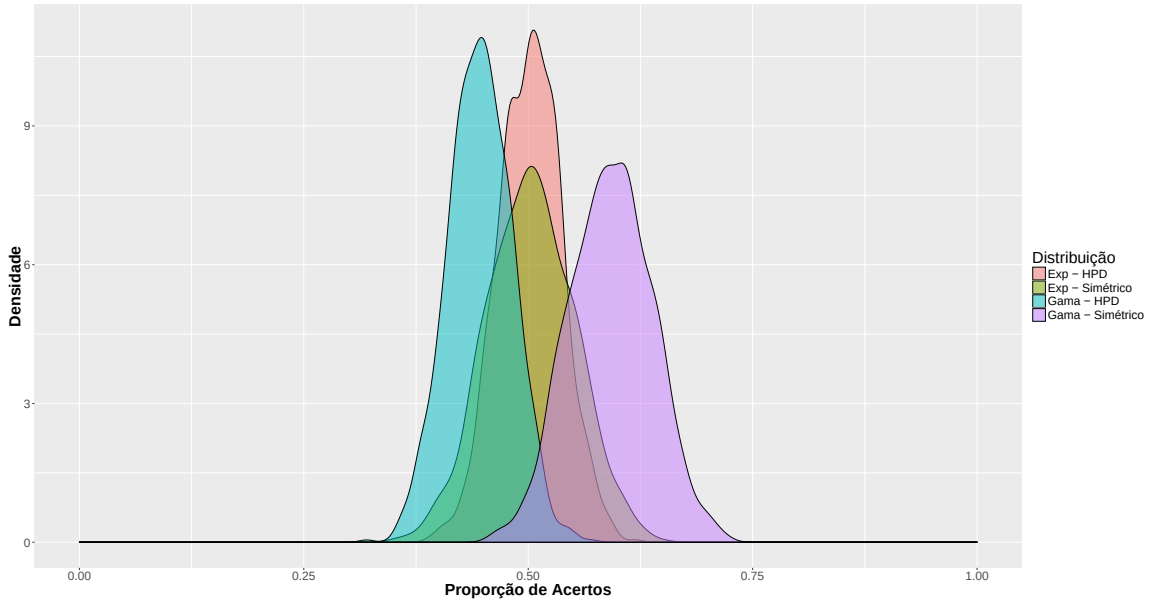


Figura 5.2: Comparação da *DPA* das distribuições Gama e Exponencial

Inesperadamente, a diferença de comportamento da *DPA* dos intervalos Simétrico e HPD das distribuições Gama, apresentando valores demasiadamente distintos. O intervalo HPD para a distribuição Gama se concentrou em torno de 44,6% de acertos, enquanto o intervalo Simétrico se concentrou em 59,2%. A variância mais elevada no intervalo Simétrico se manteve também para a distribuição Gama.

Os dois tipos de intervalo apresentarem o mesmo poder de discriminação, visto que a

interseção entre as distribuições exponenciais e Gama, com os mesmo tipos de intervalo, possuem áreas semelhantes, indicando que qualquer um dos tipos de intervalo pode ser utilizado sem prejuízo à qualidade da metodologia.

Devido à facilidade de obtenção e interpretação, este trabalho utilizará apenas o intervalo Simétrico para a definição do critério de rejeição de modelos, por esse motivo as simulações para estudo dos demais fatores serão realizadas apenas com intervalos Simétricos.

5.2 Tipo de Covariável

O tipo de covariável utilizada, númerica ou categórica, pode influenciar a *DPA* proposta pela metodologia, da mesma maneira que o tipo de intervalo. Dessa forma, para identificar o efeito do tipo de covariável, foram geradas 1.000 amostras de tamanho 100 de uma distribuição exponencial com parâmetro $\lambda = e^{-0,7+1,1X}$, sendo X gerado por uma distribuição $Bernoulli(p = 0,3)$.

As amostras geradas foram comparadas com os resultados obtidos pela simulação anterior, que utilizou a mesma distribuição com uma covariável numérica.

A comparação das distribuições é apresentada na Figura 5.3, mostrando que as duas apresentaram *DPA* muito próximas, sendo que a diferença na convergência da média na covariável categórica ocorreu, provavelmente, por insuficiência de simulações.

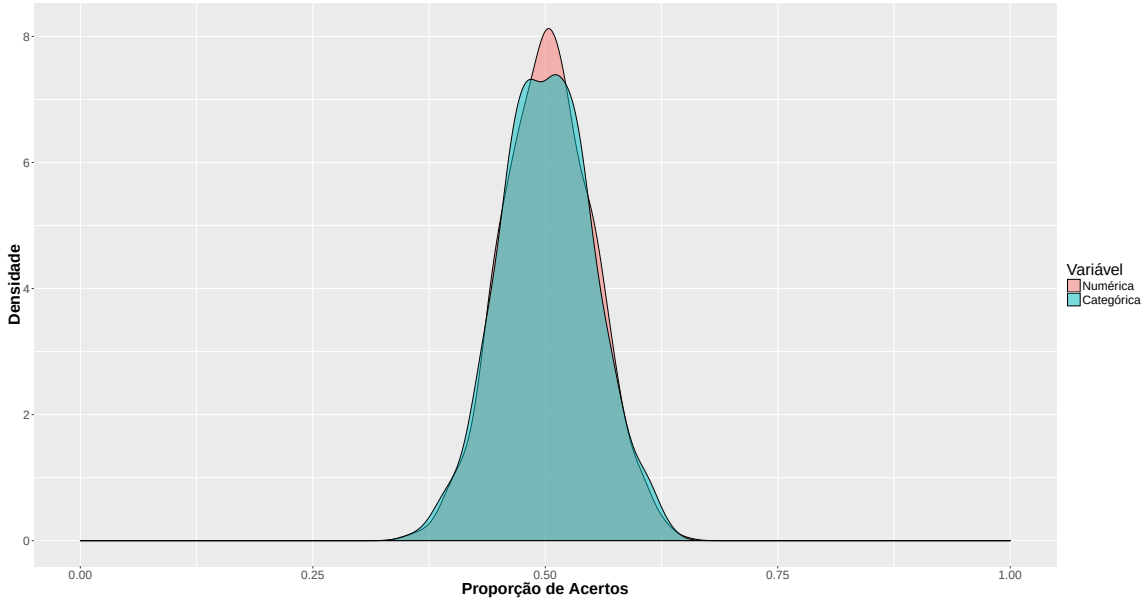


Figura 5.3: Comparação da *DPA* utilizando diferentes tipos de covariável

Com o resultado observado, é possível acreditar que o tipo de covariável utilizada é um fator que não influencia a *DPA* do modelo que está sendo avaliado.

5.3 Número de Parâmetros

O número de parâmetros utilizados no modelo é um fator que pode influenciar a *DPA*, pois modelos saturados normalmente possuem baixa capacidade preditiva. Para testar esse fator foram simuladas amostras de distribuições exponencias com 1 a 4 covariáveis, tendo cada uma 1.000 amostras de tamanho 100. Os preditores lineares utilizados nas simulações foram:

$$
\begin{aligned}
&C_1 : -0{,}3 + 0{,}7X_1; \\
&C_2 : -0{,}3 + 0{,}7X_1 - 1{,}2X_2; \\
&C_3 : -0{,}3 + 0{,}7X_1 - 1{,}2X_2 + 1{,}1X_3; \\
&C_4 : -0{,}3 + 0{,}7X_1 - 1{,}2X_2 + 1{,}1X_3 - 0{,}7X_4,
\end{aligned}
$$

sendo $X_1 \sim Uniforme(a = 1{,}b = 5)$, $X_2 \sim Bernoulli(p = 0{,}2)$, $X_3 \sim Normal(\mu = 0{,}\sigma = 1)$ e $X_4 \sim Bernoulli(p = 0{,}7)$.

Os resultados das simulações são apresentados na Figura 5.4 indicaram que o número de covariáveis não parece influenciar a *DPA* do modelo, pois todas as distribuições apresentaram comportamentos semelhantes. Entretanto, é importante ressaltar que foram utilizados apenas 4 parâmetros em amostras de tamanho 100, de forma que o problemas de saturação não puderam ser avaliados.

5.4 Tamanho Amostral

Para verificar o impacto do tamanho amostral na metodologia foram geradas 1.000 amostras de tamanhos $n = 40, 60, 80$ e 100 de uma distribuição exponencial com parâmetro $\lambda = e^{-0{,}7+1{,}1X}$, sendo X gerado por uma distribuição $Uniforme(a = 1{,}b = 5)$.

A Figura 5.5 apresenta o resultado das simulações, mostrando que a *DPA* é influenciada pelo tamanho amostral, e aparenta diminuir a variabilidade à medida em que o tamanho amostral cresce.

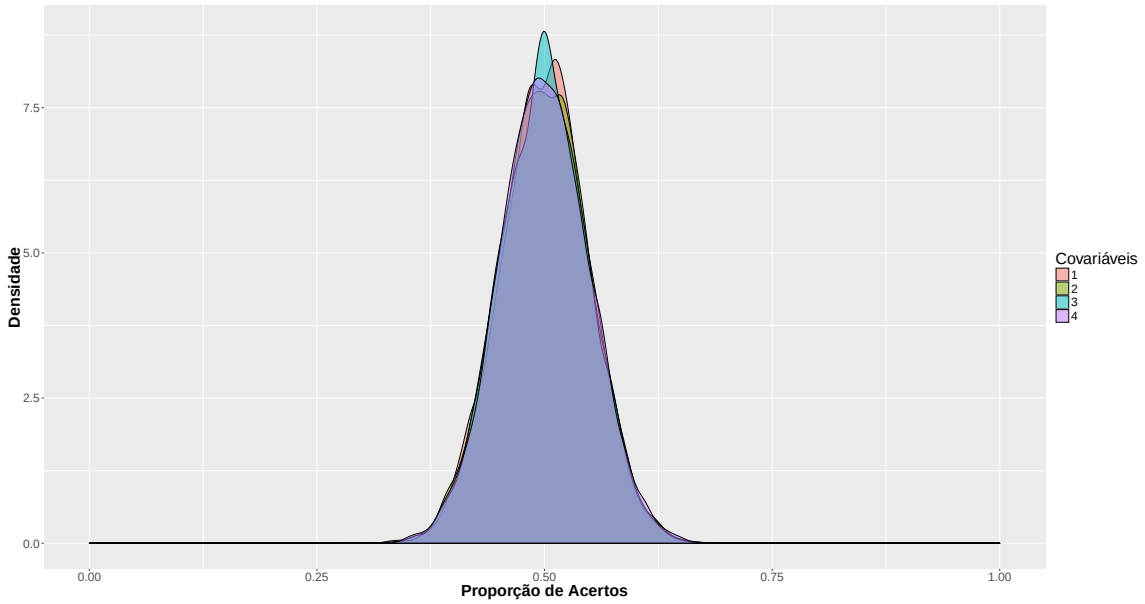


Figura 5.4: Comparação da *DPA* utilizando diferentes quantidades de covariáveis

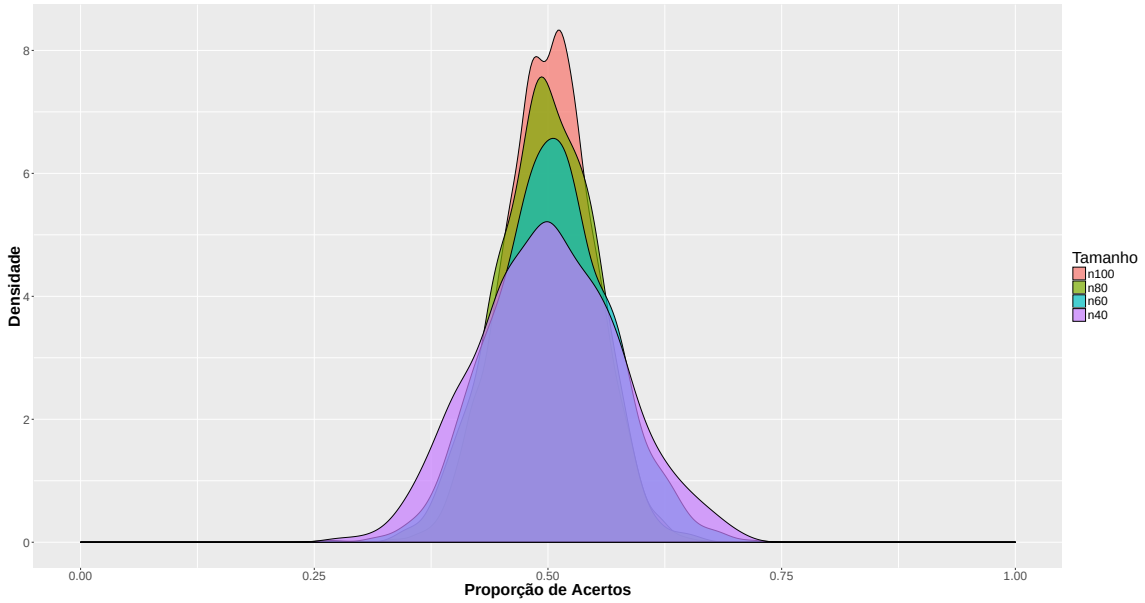


Figura 5.5: Comparação da *DPA* utilizando diferentes tamanhos amostrais

A diminuição da variabilidade era um resultado esperado, visto que quanto maior o tamanho amostral, maior será a convergência para o valor esperado da *DPA*, que é 0,5 devido à credibilidade utilizada, fazendo com que se diminua a variabilidade à medida que se cresce o tamanho da amostra.

O resultado da simulação indica que é necessária a utilização do tamanho amostral na elaboração do critério de rejeição da metodologia, pois influencia diretamente a variabili-

dade da *DPA* do modelo.

Apesar de variabilidades diferentes, as distribuições continuaram a apresentar um comportamento simétrico em todos os tamanhos amostrais, porém o decrescimento da variabilidade não aparenta ser linear.

5.5 Simulações Cruzadas

Os resultados obtidos nas simulações anteriores indicam que, dos fatores de influência considerados, apenas o tamanho amostral afeta a *DPA* do modelo. Para confirmar esses resultados é necessário olhar também o comportamento da *DPA* nas diferentes interações dos fatores de influência, pois é possível que algum fator apresente efeito quando combinado com algum dos demais, como por exemplo, o tipo de covariável apresentar efeito na *DPA* quando o tamanho da amostra é pequeno.

Para verificar o efeito dos possíveis cruzamentos entre os fatores, foram simuladas amostras de tamanho n, considerando 4 combinações de parâmetros com 1 a 5 preditores cada, totalizando 20 possíveis cruzamentos. Gerou-se 1.000 amostras para cada um desses cruzamentos, totalizando 20.000 amostras, e para cada uma dessas amostras geradas a metodologia foi aplicada.

Os valores de n simulados foram de 10 a 40, 50, 60, 70,..., 140 e 150 por questões de custos computacionais, pois cada amostra realiza n MCMC's, elevando demasiadamente o tempo de múltiplas simulações.

O organograma retratado na Figura 5.6 evidencia a estrutura utilizada na simulação e são apresentados também as combinações de parâmetros utilizados.

As combinações foram escolhidas com o intuito de diversificar ao máximo os valores dos parâmetros e covariáveis utilizadas. Segue abaixo as 4 combinações consideradas:

$$C_1 : \boldsymbol{X}^T\boldsymbol{\beta} = -0{,}7 + 1{,}1X_1 - 0{,}6X_2 + 0{,}2X_3 + 1X_4 - 1{,}5X_5$$
$$C_2 : \boldsymbol{X}^T\boldsymbol{\beta} = 1 - 1{,}3X_1 + 0{,}4X_2 - 0{,}2X_3 + 0{,}9X_4 - 0{,}3X_5$$
$$C_3 : \boldsymbol{X}^T\boldsymbol{\beta} = -0{,}3 + 0{,}7X_1 - 1{,}2X_2 + 1{,}1X_3 - 0{,}7X_4 + 1X_5$$
$$C_4 : \boldsymbol{X}^T\boldsymbol{\beta} = 1{,}7 - 0{,}8X_1 + 0{,}1X_2 + 0{,}6X_3 - 0{,}8X_4 - 1{,}1X_5$$

Nas combinações, $X_1 \sim Uniforme(a = 0{,}b = 5)$, $X_2 \sim Bernoulli(p = 0{,}2)$, $X_3 \sim$

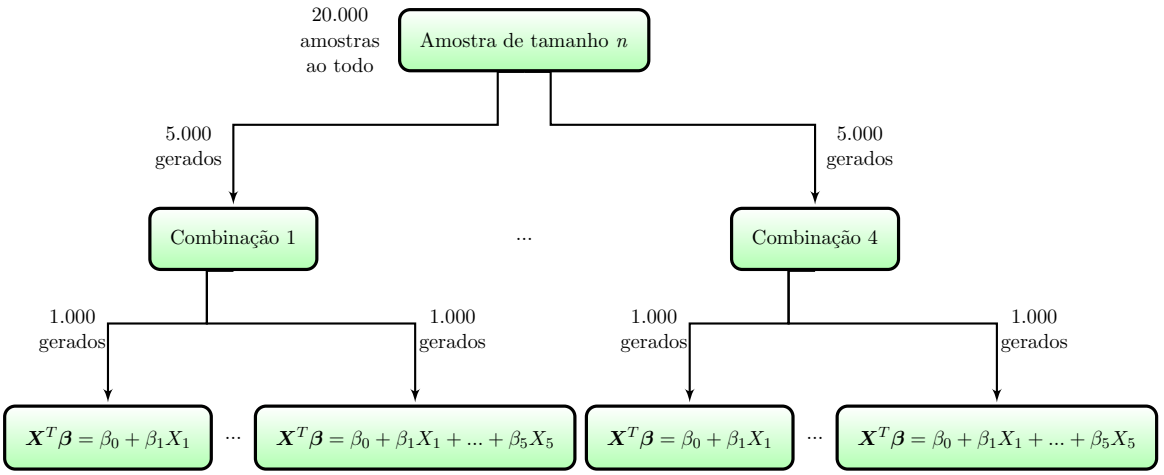


Figura 5.6: Estrutura da simulação

$Normal(\mu = 0{,}\sigma^2 = 1)$, $X_4 \sim Bernoulli(p = 0{,}7)$ e $X_5 \sim Uniforme(a = 0{,}b = 5)$.

Os resultados das simulações são demonstrados nas Figuras 5.7 e 5.8, que apresentam a média e o desvio padrão da *DPA* de cada combinação de fatores simuladas.

Pelos gráficos é possivel verificar que em tamanhos amostrais pequenos, grandes disparidades foram observadas entre os modelos com diferentes números de covariáveis, pois as simulações com maiores números de covariáveis apresentaram médias e desvios mais elevados que os demais. Este resultado é natural devido a saturação do modelo para pequenas amostras, onde se tenta estimar muitos parâmetros com poucas observações, fazendo com que o modelo ajustado fique com baixa capacidade preditiva.

À medida em que cresce o tamanho amostral, as diferenças por número de covariáveis deixam de acontecer e todos convergem para o mesmo valor, tanto na média quanto no desvio padrão. Pelos gráficos, quando o modelo apresenta número de covariáveis menor do que aproximadamente 20% do tamanho amostral, a *DPA* não é afetada por esse fator.

As diferentes combinações de parâmetros utilizados não aparentam afetar a *DPA* do modelo visto que ela apresentou valores bem distribuidos entre as simulações realizadas. Isso reforça o resultado obtido (Seção 5.2) de que os tipos de covariáveis não influenciam a *DPA* do modelo.

Outro resultado esperado é a convergência do desvio da proporção de acertos para zero e a média converge para 0,5, uma vez que quanto maior o tamanho amostral, maior é a concentração da proporção de acertos ao redor da credibilidade utilizada.

Para verificar se todas as *DPAs* simuladas apresentaram comportamento simétrico,

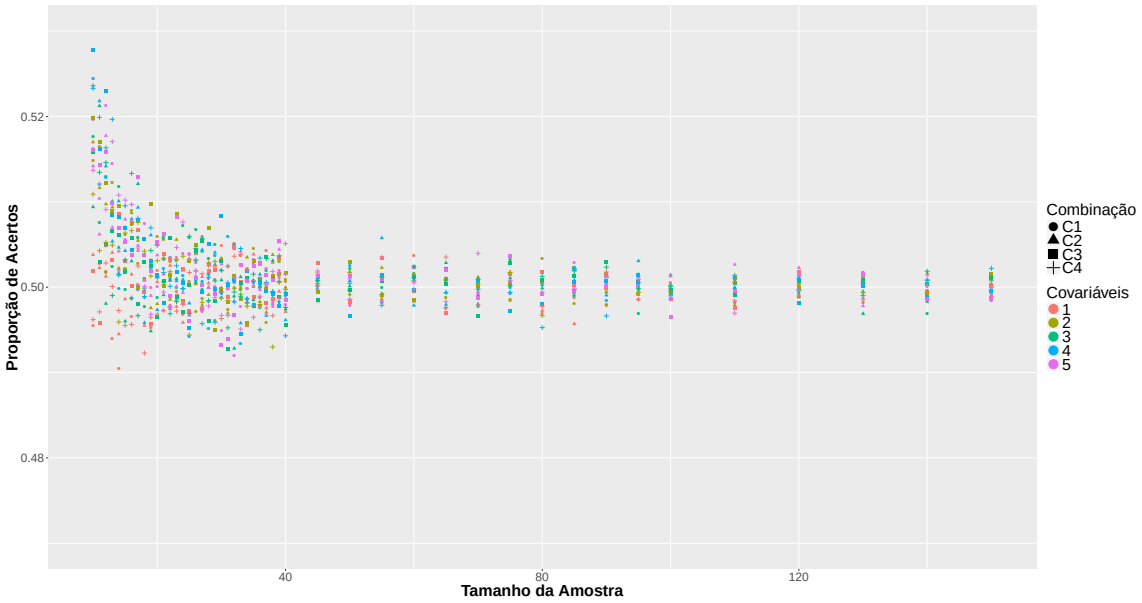


Figura 5.7: Média da *DPA* por Combinação e Número de Covariáveis

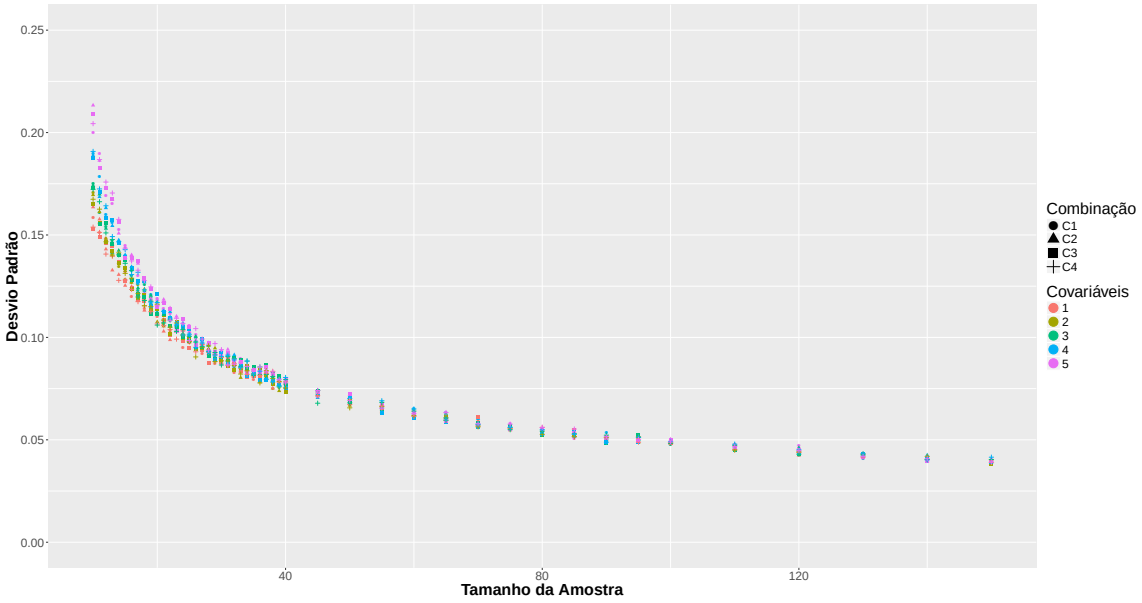


Figura 5.8: Desvio Padrão da *DPA* por Combinação e Número de Covariáveis

calculou-se o coeficiente de assimetria pelo número de covariáveis, tipo de combinação e tamanho amostral. O coeficiente de assimetria é uma maneira prática de se verificar a simetria de uma distribuição, pois a distribuição é simétrica quando o coeficiente é igual a 0, por esse motivo este coeficiente foi utilizado para a verificação.

A Figura 5.9 apresenta o resultado dos coeficientes de assimetria calculado para as simulações realizadas. Percebe-se pelo gráfico que todos os valores se concentram próximos

de 0, indicando evidências de simetria em todas as *DPAs* simuladas.

As oscilações em torno de zero são consequência do número de simulações realizadas para cada caso, mesmo assim, valores entre -0,25 e 0,25 são muito próximos de um comportamento simétrico e podem ser aproximados sem prejuízo.

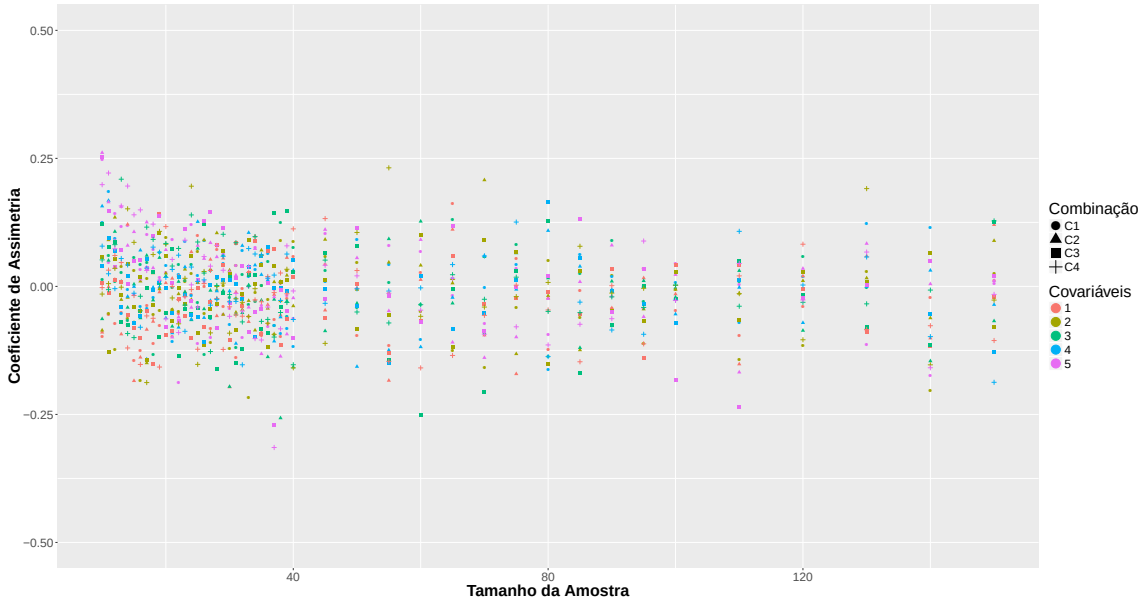


Figura 5.9: Assimetria da *DPA* por Número de Covariáveis

Com os resultados das simulações cruzadas foi possível validar os resultados obtidos anteriormente (Seções 5.2 a 5.4), de que apenas o fator tamanho amostral afeta a *DPA* do modelo. Por esse motivo apenas o tamanho amostral n será utilizado para a definição do critério de rejeição do modelo.

A Figura 5.10 apresenta a média e o desvio padrão por tamanho amostral das *DPAs* simuladas. Cada tamanho amostral apresentado no gráfico contém 20.000 simulações, o que melhora a aproximação das estimativas.

Pelo gráfico percebe-se que a média das *DPAs* está centrada em 0,5, consiste na credibilidade usada, e que o desvio padrão converge para zero à medida em que o tamanho amostral cresce, semelhante ao comportamento apresentado nas Figuras 5.7 e 5.8.

Foi calculada também a assimetria para as *DPAs* agregadas apenas pelo tamanho amostral e os resultados estão na Figura 5.11, que mostra coeficientes de assimetria muito próximos de 0, permitindo considerar as distribuições como simétricas.

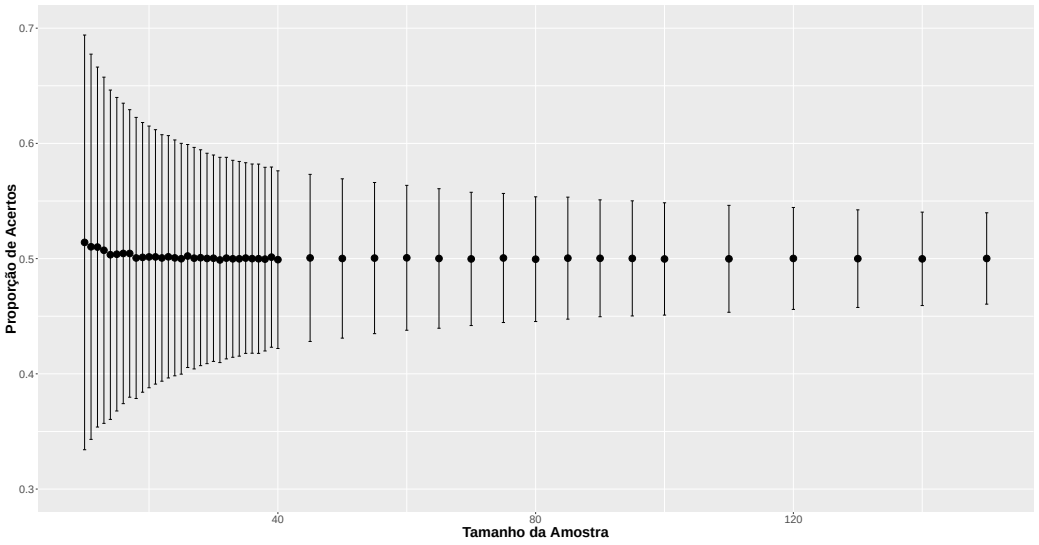


Figura 5.10: Média e Desvio Padrão da *DPA* por Tamanho de Amostra

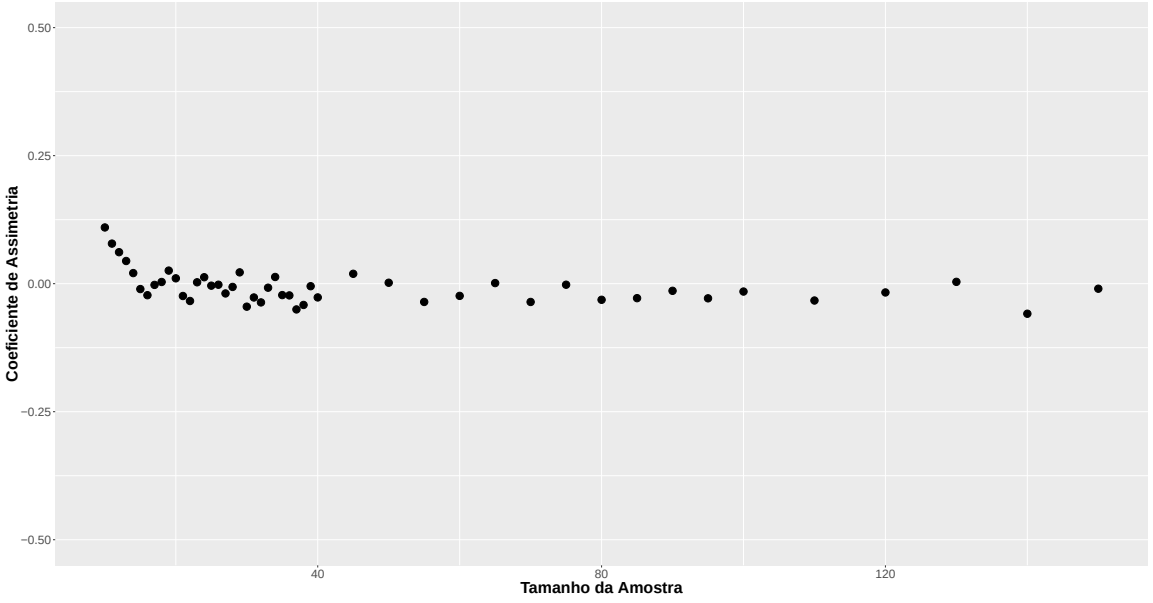


Figura 5.11: Assimetria da *DPA* por Tamanho de Amostra

Os resultados explicitados nessa seção permitiram identificar que apenas o fator tamanho amostral precisa ser utilizado na elaboração do critério de rejeição da metodologia, o que facilita sua elaboração e generalização.

Uma vantagem da *DPA* dos modelos apresentarem comportamento simétrico (e centrado em 0,5) é permitir a elaboração um critério de rejeição mais intuitivo e prático de se calcular, pois utilizará a mesma regra para a definição dos pontos críticos superior e inferior.

Capítulo 6

Critério de Rejeição

A elaboração do critério de rejeição de modelos é fundamental para a metodologia proposta, pois permite identificar, de maneira objetiva, quando a proporção de acertos calculada é diferente da esperada para a distribuição proposta.

Utilizando as simulações e resultados obtidos na Seção 5.5, tem-se que apenas o tamanho amostral se mostrou como um fator de influência na *DPA* dos modelos, fazendo com que seja desnecessário utilizar os fatores tipo de covariável e número de covariáveis. Como este trabalho aborda apenas o uso de intervalos simétricos, o critério de rejeição será elaborado apenas para esse tipo de intervalo.

Como critério de rejeição, optou-se por utilizar como ponto crítico os percentis empíricos obtidos pelas *DPAs* simuladas. Esses percentis são afetados diretamente pelo tamanho amostral, fazendo com que tamanhos diferentes de amostra possuam pontos críticos diferentes. Dessa maneira, os valores de proporção de acerto entre os percentis e a média serão considerados como aceitáveis para o modelo proposto, rejeitando o modelo apenas quando a proporção de acertos é mais extrema que os percentis.

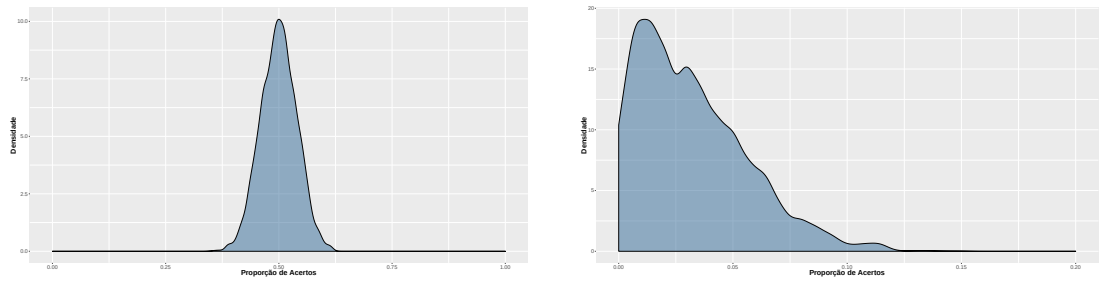
Pelo fato das distribuições simuladas terem apresentado comportamento simétrico, é possível definir uma margem de erro para apenas um lado, pois sendo α o nível de credibilidade utilizado, a distância de 0,5 para o percentil $100\frac{\alpha}{2}\%$ é equivalente à distância para o percentil $100(1-\frac{\alpha}{2})\%$, totalizando um intervalo que contém $100(1-\alpha)\%$ de credibilidade. Dessa maneira, apenas é necessária a obtenção de um erro ξ que representará a distância máxima aceitável entre proporção de acertos obtida e o valor 0,5.

De forma a evitar situações em que, devido a imprecisões na simulação, a distribuição apresentasse erros diferentes para os percentis, a seguinte operação foi aplicada nas *DPAs*. Sendo x a proporção de acertos,tem-se a seguinte transformação:

$$\begin{cases} 0{,}5 - x, & \text{se } x < 0{,}5 \\ x - 0{,}5, & \text{se } x \geq 0{,}5. \end{cases} \tag{6.1}$$

Essa operação faz com que todos os valores se concentrem em apenas um lado e desloca a origem da distribuição no ponto 0, permitindo a obtenção de apenas um valor para o erro ξ, ao invés de dois.

A Figura 6.1 mostra uma *DPA* antes e depois da operação, sendo necessária a obtenção dos percentis $100\frac{\alpha}{2}\%$ e $100\left(1-\frac{\alpha}{2}\right)\%$ para a primeira e apenas o percentil $100(1-\alpha)\%$ para a segunda.



(a) *DPA* sem alteração (b) *DPA* com alteração

Figura 6.1: *DPA* empírica antes e depois de operação

A Figura 6.2 apresenta o valor crítico de ξ da metodologia calculado para $\alpha = 0{,}05$. Apesar de ser possível identificar alguns pontos críticos através do gráfico, algumas inconsistências podem ser observadas, como por exemplo, o decrescimento do ponto crítico quando o tamanho amostral aumenta não é contínuo. Um problema é a impossibilidade de se obter o ponto para tamanhos amostrais que não foram simulados, pois apenas graficamente não é possível extrapolar os valores.

Para evitar os problemas apresentados na Figura 6.2, ajustou-se um modelo de mínimos quadrados para os erros (ξ) considerando a raiz quadrada do tamanho amostral como variável explicativa. A utilização do modelo ajustado permite a extrapolação para valores não observados e deixa o ponto de corte com um comportamento contínuo, evitando que ξ não apresente comportamento monótono decrescente em relação ao tamanho amostral n.

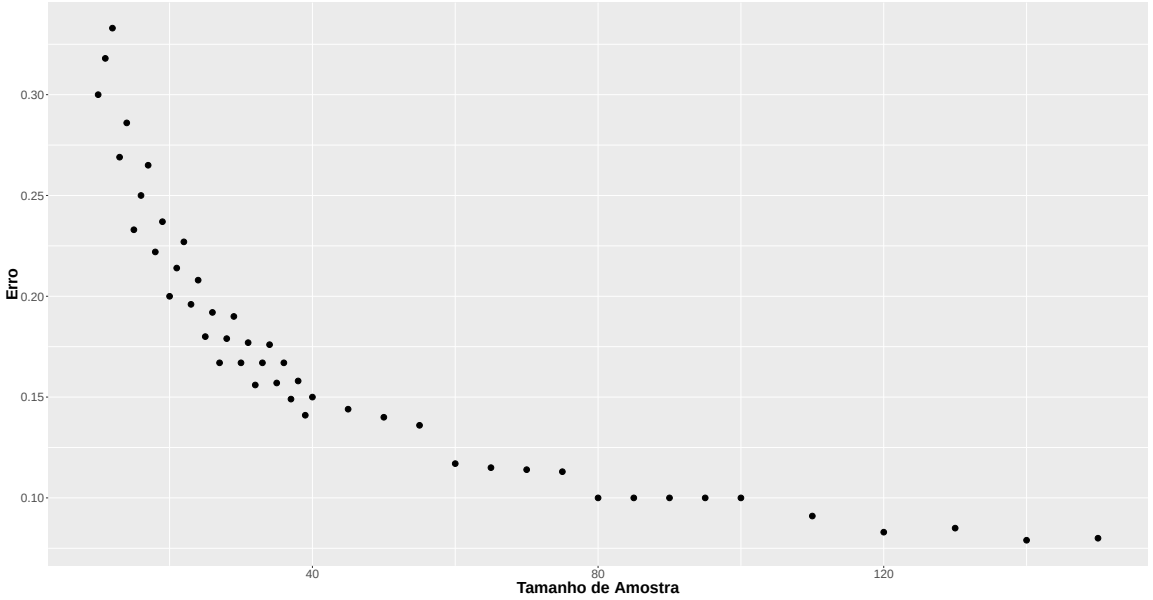


Figura 6.2: Erro ξ para $\alpha = 0{,}05$

O modelo de mínimos quadrados utilizado foi $\xi = \frac{\beta_1}{\sqrt{n}}$, tendo sido baseado no mesmo utilizado pelo teste de Kolmogorov-Smirnov para adequabilidade de ajustamento (Massey Jr, 1951). Como os valores de ξ de amostras menores que 20 possuem um decaimento muito maior que o restante, foi estimado um modelo $\xi = \beta_0 + \frac{\beta_1}{\sqrt{n}}$ apenas para as observações menores que 20, para que as estimativas não fossem prejudicadas.

Os parâmetros estimados para os modelos com $\alpha = 0{,}2; 0{,}1; 0{,}05$ e 0,01 estão apresentados na Tabela 6.1. O modelo para $n < 20$ será utilizado apenas para a suavização dos valores, enquanto o modelo para $n \geq 20$ será utilizado para a suavização e extrapolação dos valores.

Tabela 6.1: Parâmetros estimados para suavização dos erros

Modelo $n < 20$

α	β_0^*	β_1^*
0,01	-0,183	2,045
0,05	-0,043	1,171
0,10	-0,109	1,310
0,20	-0,011	0,718

Modelo $n \geq 20$

α	β_1
0,01	1,262
0,05	0,955
0,10	0,807
0,20	0,631

As Figuras 6.3, 6.4, 6.5 e 6.6 apresentam os ajustes dos modelos nos erros para as diferentes credibilidades. Os gráficos evidenciam ajustes satisfatórios, indicando que o modelo ajustado pelo método de mínimos quadrados conseguiu se adequar bem à curva.

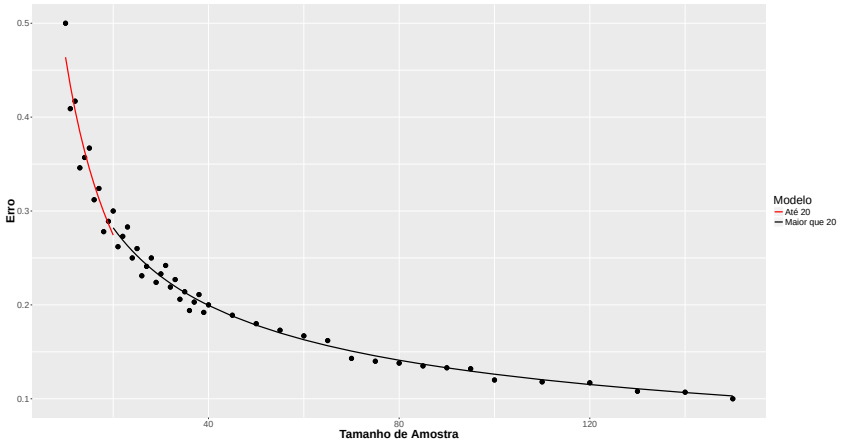


Figura 6.3: Curvas de regressão do erro ξ para $\alpha = 0{,}01$

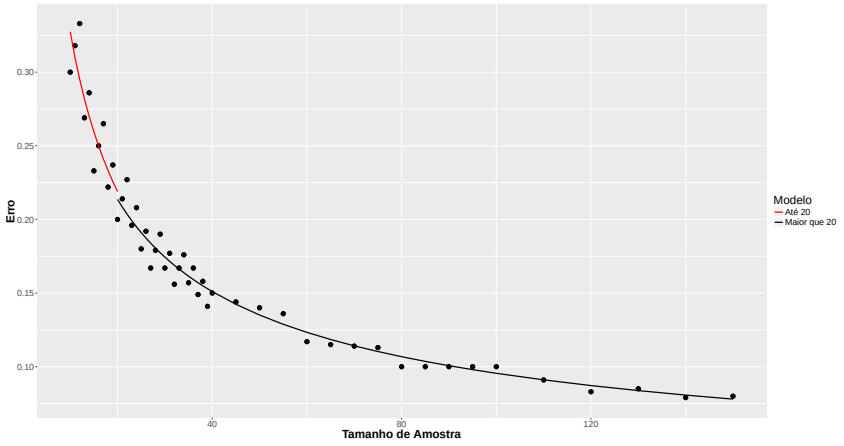


Figura 6.4: Curvas de regressão do erro ξ para $\alpha = 0{,}05$

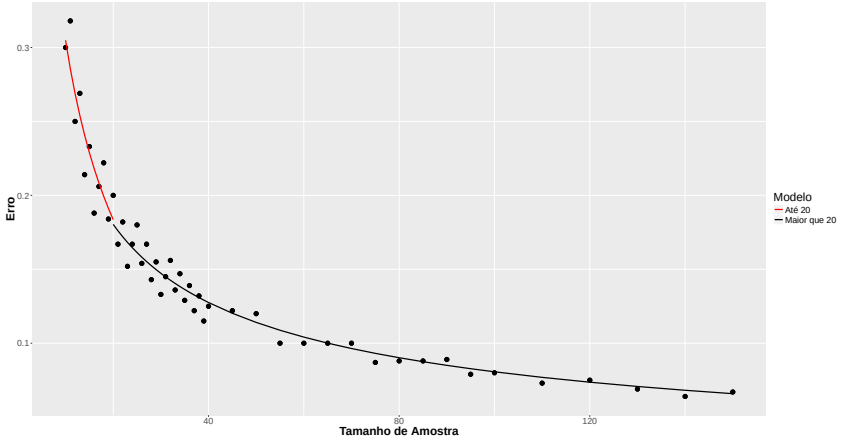


Figura 6.5: Curvas de regressão do erro ξ para $\alpha = 0{,}10$

A interseção entre os dois modelos apresentou valores próximos em todas as situações, tornando desnecessário o uso de alguma correção para suavizar a curva de regressão.

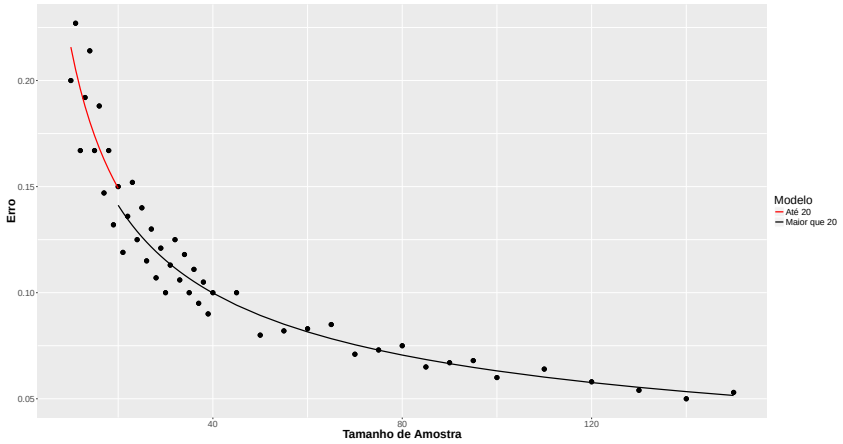


Figura 6.6: Curvas de regressão do erro ξ para $\alpha = 0{,}20$

Os pontos críticos para a metodologia são exibidos nas Tabelas 6.2 e 6.3, onde a primeira apresenta o valor de ξ obtido nas simulações e a segunda o o valor ξ suavizado pelo modelo de mínimos quadrados. As tabelas indicam qual o erro máximo que a proporção de acertos do modelo pode assumir, como por exemplo, para uma amostra de tamanho 75 com 95% de credibilidade, a proporção de acertos deve estar entre 0,39 e 0,61 ($0{,}5 \pm 0{,}11$), caso contrário o modelo testado não possui uma boa capacidade preditiva para o problema.

É importante lembrar que a distribuição da proporção de acertos é discreta, pois uma amostra tamanho 4 por exemplo, só pode assumir os valores $\frac{1}{4}, \frac{2}{4}, \frac{3}{4}$ e $\frac{4}{4}$. Apesar da distribuição ser discreta, a suavização é uma aproximação contínua dos valores, fazendo com que ξ eventualmente assuma valores que não existem na distribuição.

Com o auxílio da Tabela 6.3 é possível utilizar a metodologia proposta em uma base de dados com qualquer tamanho amostral para os níveis de credibilidade $\alpha = 0{,}2; 0{,}1; 0{,}05$ e 0,01, podendo utilizar as aproximações sugeridas para extrapolar os tamanhos de amostra que não estão disponibilizadas nas tabelas.

Valores Críticos da Metodologia

Tabela 6.2: Valores de ξ Simulados

	α			
n	**0,01**	**0,05**	**0,10**	**0,20**
10	0,500	0,300	0,300	0,200
11	0,409	0,318	0,318	0,227
12	0,417	0,333	0,250	0,167
13	0,346	0,269	0,269	0,192
14	0,357	0,286	0,214	0,214
15	0,367	0,233	0,233	0,167
16	0,312	0,250	0,188	0,188
17	0,324	0,265	0,206	0,147
18	0,278	0,222	0,222	0,167
19	0,289	0,237	0,184	0,132
20	0,300	0,200	0,200	0,150
21	0,262	0,214	0,167	0,119
22	0,273	0,227	0,182	0,136
23	0,283	0,196	0,152	0,152
24	0,250	0,208	0,167	0,125
25	0,260	0,180	0,180	0,140
26	0,231	0,192	0,154	0,115
27	0,241	0,167	0,167	0,130
28	0,250	0,179	0,143	0,107
29	0,224	0,190	0,155	0,121
30	0,233	0,167	0,133	0,100
31	0,242	0,177	0,145	0,113
32	0,219	0,156	0,156	0,125
33	0,227	0,167	0,136	0,106
34	0,206	0,176	0,147	0,118
35	0,214	0,157	0,129	0,100
36	0,194	0,167	0,139	0,111
37	0,203	0,149	0,122	0,095
38	0,211	0,158	0,132	0,105
39	0,192	0,141	0,115	0,090
40	0,200	0,150	0,125	0,100
45	0,189	0,144	0,122	0,100
50	0,180	0,140	0,120	0,080
55	0,173	0,136	0,100	0,082
60	0,167	0,117	0,100	0,083
65	0,162	0,115	0,100	0,085
70	0,143	0,114	0,100	0,071
75	0,140	0,113	0,087	0,073
80	0,138	0,100	0,088	0,075
85	0,135	0,100	0,088	0,065
90	0,133	0,100	0,089	0,067
95	0,132	0,100	0,079	0,068
100	0,120	0,100	0,080	0,060
110	0,118	0,091	0,073	0,064
120	0,117	0,083	0,075	0,058
130	0,108	0,085	0,069	0,054
140	0,107	0,079	0,064	0,050
150	0,100	0,080	0,067	0,053

Tabela 6.3: Valores de ξ Suavizados

	α			
n	**0,01**	**0,05**	**0,10**	**0,20**
10	0,464	0,327	0,305	0,216
11	0,434	0,310	0,286	0,205
12	0,407	0,295	0,269	0,196
13	0,384	0,282	0,254	0,188
14	0,364	0,270	0,241	0,181
15	0,345	0,259	0,229	0,174
16	0,328	0,250	0,218	0,168
17	0,313	0,241	0,208	0,163
18	0,299	0,233	0,199	0,158
19	0,286	0,226	0,191	0,153
20	0,274	0,219	0,184	0,149
21	0,275	0,208	0,176	0,138
22	0,269	0,204	0,172	0,135
23	0,263	0,199	0,168	0,132
24	0,258	0,195	0,165	0,129
25	0,252	0,191	0,161	0,126
26	0,247	0,187	0,158	0,124
27	0,243	0,184	0,155	0,122
28	0,238	0,181	0,152	0,119
29	0,234	0,177	0,150	0,117
30	0,230	0,174	0,147	0,115
31	0,227	0,172	0,145	0,113
32	0,223	0,169	0,143	0,112
33	0,220	0,166	0,140	0,110
34	0,216	0,164	0,138	0,108
35	0,213	0,161	0,136	0,107
36	0,210	0,159	0,134	0,105
37	0,207	0,157	0,133	0,104
38	0,205	0,155	0,131	0,102
39	0,202	0,153	0,129	0,101
40	0,199	0,151	0,128	0,100
45	0,188	0,142	0,120	0,094
50	0,178	0,135	0,114	0,089
55	0,170	0,129	0,109	0,085
60	0,163	0,123	0,104	0,082
65	0,156	0,118	0,100	0,078
70	0,151	0,114	0,096	0,075
75	0,146	0,110	0,093	0,073
80	0,141	0,107	0,090	0,071
85	0,137	0,104	0,087	0,068
90	0,133	0,101	0,085	0,067
95	0,129	0,098	0,083	0,065
100	0,126	0,096	0,081	0,063
110	0,120	0,091	0,077	0,060
120	0,115	0,087	0,074	0,058
130	0,111	0,084	0,071	0,055
140	0,107	0,081	0,068	0,053
150	0,103	0,078	0,066	0,052
$n \geq 40$	$\frac{1,262}{\sqrt{n}}$	$\frac{0,955}{\sqrt{n}}$	$\frac{0,807}{\sqrt{n}}$	$\frac{0,631}{\sqrt{n}}$

6.1 Regra de Decisão

Com base nos resultado obtidos neste capítulo, é possível elaborar um teste de hipóteses para a metodologia, formalizando uma maneira objetiva de se identificar se existem evidências de que o modelo utilizado não possui boa capacidade preditiva para o problema.

Considerando as hipóteses

$$\begin{cases} H, & \text{O modelo M apresenta uma boa capacidade preditiva} \\ H_a, & \text{O modelo M não apresenta uma boa capacidade preditiva} \end{cases} \tag{6.2}$$

Segundo a metodologia deste trabalho, a hipótese H é rejeitada, com um nível de credibilidade α se $\xi_{\text{obs}} > \xi_{\text{crítico}}$, onde $\xi_{\text{crítico}}$ depende da credibilidade α e do tamanho da amostra n (Tabela 6.3).

Dessa forma, é possível a utilização da metodologia de forma prática em qualquer problema real, pois agora possui um critério único para comparação, evitando divergências nos resultados advindas de critérios subjetivos.

Capítulo 7

Estudo de Caso

Neste capítulo a metodologia foi aplicada em duas bases de dados reais, avaliando a capacidade de um Modelo Linear Generalizado Exponencial predizer novas ocorrências dos problemas apresentados.

7.1 Base *Baby Boom*

7.1.1 Descrição dos Dados

A base *Baby Boom*, apresentada por Dunn (1999), contém dados sobre o tempo, em minutos, entre nascimentos de bebês durante um período de 24 horas no hospital de Brisbane, Austrália, totalizando 43 tempos observados. Os dados são apresentados na Tabela 7.1.

Tabela 7.1: Dados *Baby Boom*

Tempo entre nascimentos				
59	27	47	45	27
14	37	9	25	14
37	2	61	18	13
62	55	1	29	19
68	86	26	15	54
2	14	13	38	70
15	4	28	2	28
9	40	77	2	
157	36	26	19	

Para verificar se os dados realmente seguem uma distribuição Exponencial, eles foram

avaliados de maneira descritiva através de um Q-Q Plot, que mostra a distância entre o quantil observado e o quantil teórico de uma distribuição Exponencial.

A Figura 7.1 contém o Q-Q Plot dos dados e indica que possuem valores condizentes com o de uma distribuição Exponencial.

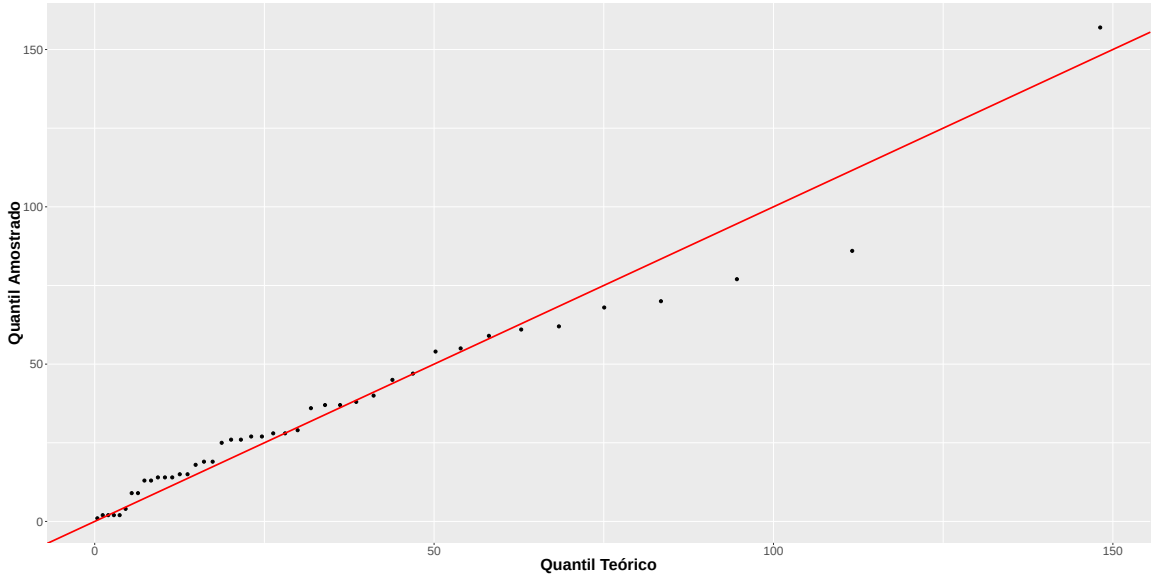


Figura 7.1: Q-Q plot da distribuição Exponencial

Como os dados não possuem covariáveis, foram realizados os testes estatísticos de Kolmogorov-Smirnov (Massey Jr, 1951) e de Lilliefors (Lilliefors, 1969) para uma verificação inferencial da adequação de uma distribuição Exponencial. Os resultados dos testes são apresentados na Tabela 7.2.

Tabela 7.2: Teste de ajustamento de uma distribuição exponencial para a base *Baby Boom*

	Kolmogorov-Smirnov	Lilliefors
p-valor	0,1375	0,1498

Pelos resultados apresentados verificamos que não existem evidências de que os dados não sejam advindos de uma distribuição Exponencial.

Esta base de dados foi escolhida por ser um exemplo pequeno e simples de dados que seguem uma distribuição exponencial.

7.1.2 Aplicação da Metodologia

A metodologia proposta foi aplicada aos dados da base *Baby Boom* utilizando uma priori difusa $N(0{,}100)$ para β_0, simulando 100.000 amostras com saltos de tamanho 5 e burn-in de 10.000 no MCMC.

O modelo utilizado para a aplicação é o exponencial com parâmetro $\lambda = e^{-\beta_0}$. Por se tratar de um modelo sem covariáveis, o mesmo depende apenas de β_0 para predizer o tempo entre nascimentos, T.

A Tabela 7.3 e a Figura 7.2 apresentam os resultados do modelo estimado para cada observação retirada no LOO da aplicação.

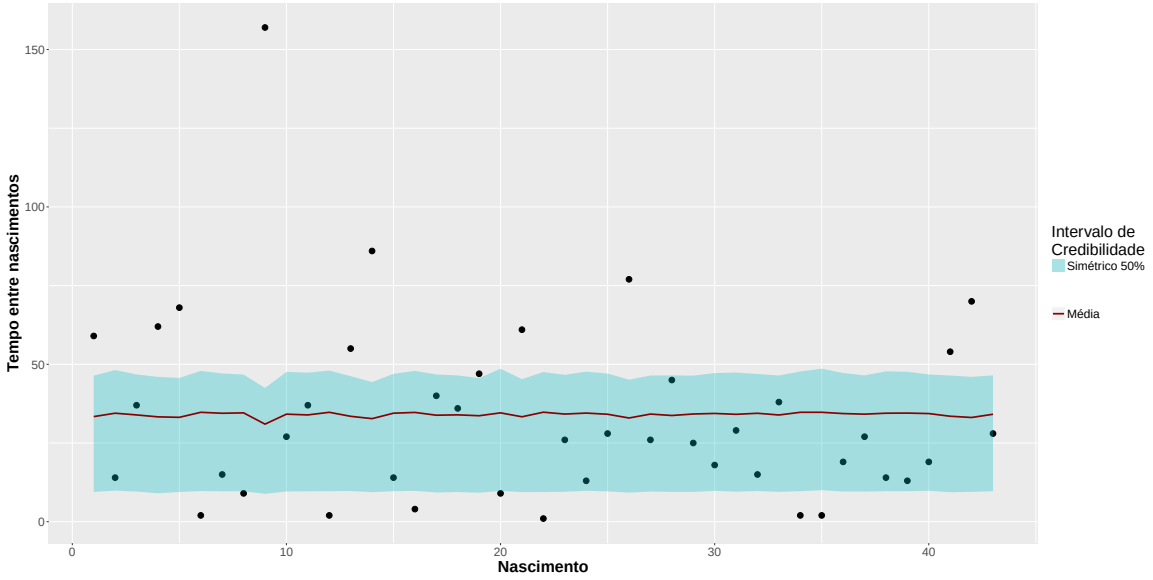


Figura 7.2: Modelo Exponencial com distribuição *a priori* difusa

Os resultados obtidos para cada tempo foram próximos dos obtidos para o modelo completo (com todas observações), com exceção do 9º nascimento que apresentou grande alteração no λ estimado, mas foi mantido na análise. Por serem calculados numericamente, é natural que tempos iguais possuam pequenas variações nas estimativas apresentadas.

A coluna *Acerto* na Tabela 7.3 indica se o *Tempo* da observação retirada estava dentro do intervalo predito pelo LOO, fazendo com que a estatística do teste para a metodologia seja o total de "S" na coluna acertos dividido pelo tamanho amostral.

Tabela 7.3: Resultados do *LOO* da metodologia na base *Baby Boom*

Nascimento Retirado	Tempo	β_0	λ^{-1}	Intervalo Predito	Acerto
1	59	3,499	33,084	9,633 - 45,808	N
2	14	3,532	34,200	9,837 - 47,100	S
3	37	3,516	33,638	9,436 - 47,238	S
4	62	3,497	33,013	9,158 - 45,731	N
5	68	3,492	32,868	9,211 - 46,314	N
6	2	3,540	34,474	9,738 - 48,057	N
7	15	3,530	34,132	9,942 - 47,967	S
8	9	3,535	34,291	9,613 - 47,975	N
9	157	3,424	30,693	8,730 - 42,504	N
10	27	3,522	33,838	9,704 - 47,179	S
11	37	3,516	33,638	9,657 - 46,570	S
12	2	3,540	34,474	9,734 - 47,654	N
13	55	3,503	33,203	9,464 - 46,208	N
14	86	3,479	32,418	9,736 - 44,344	N
15	14	3,532	34,200	9,898 - 47,949	S
16	4	3,539	34,432	9,873 - 48,011	N
17	40	3,512	33,522	9,723 - 46,246	S
18	36	3,516	33,645	9,615 - 46,713	S
19	47	3,509	33,407	9,623 - 46,542	N
20	9	3,535	34,291	9,382 - 47,472	N
21	61	3,498	33,046	9,518 - 46,082	N
22	1	3,541	34,509	9,740 - 47,586	N
23	26	3,523	33,897	9,709 - 46,932	S
24	13	3,531	34,167	9,698 - 47,401	S
25	28	3,522	33,851	9,537 - 47,034	S
26	77	3,486	32,649	9,192 - 45,046	N
27	26	3,523	33,897	9,560 - 47,354	S
28	45	3,510	33,457	9,870 - 47,158	S
29	25	3,524	33,908	9,525 - 46,662	S
30	18	3,530	34,114	9,575 - 47,213	S
31	29	3,520	33,786	9,439 - 46,933	S
32	15	3,530	34,132	9,880 - 46,924	S
33	38	3,514	33,587	9,499 - 46,909	S
34	2	3,540	34,474	9,788 - 47,803	N
35	2	3,540	34,474	9,739 - 47,288	N
36	19	3,528	34,060	9,740 - 47,453	S
37	27	3,522	33,838	9,635 - 46,727	S
38	14	3,532	34,200	9,740 - 47,523	S
39	13	3,531	34,167	9,758 - 48,161	S
40	19	3,528	34,060	9,934 - 47,964	S
41	54	3,502	33,178	9,527 - 46,438	N
42	70	3,492	32,861	9,476 - 45,602	N
43	28	3,522	33,851	9,318 - 46,979	S
Modelo Completo		3,514	33,980	9,463 - 46,764	

A Tabela 7.4 apresenta a proporção de acertos e o erro ξ obtidos para os dados, que são as estatísticas necessárias para utilização do critério de rejeição da metodologia.

Tabela 7.4: Proporção de Acertos e Erro ξ da Metodologia

Proporção de acertos	$\xi_{\text{observado}}$
0,558	0,058

Utilizando a Tabela 6.3 é possivel obter o ξ crítico para uma amostra de tamanho $n = 43$, utilizando a aproximação sugerida para $n > 40$. Os valores críticos são apresentados na Tabela 7.5.

Tabela 7.5: Valores Críticos para $n = 43$

α	0,01	0,05	0,10	0,20
$\xi_{\text{crítico}}$	0,192	0,146	0,123	0,096

Com as informações obtidas, tem-se que $\xi_{\text{observado}} < \xi_{\text{crítico}}$ para $\alpha = 0{,}05$. Assim, não existem evidências a 95% de credibilidade de que o modelo Exponencial não é um modelo com boa capacidade preditiva para o problema.

A aplicação da metodologia para um exemplo sem covariáveis apresenta resultados semelhantes ao teste de Lilliefors para distribuição exponencial (Lilliefors, 1969), não indicando evidências de que os dados não sejam advindos de uma distribuição exponencial.

É importante lembrar que a definição da distribuição *a priori* impacta diretamente os resultados da distribuição *a posteriori*, sendo que uma má especificação da mesma será identificada pela metodologia. Para exemplificar, definiu-se $N(6{,}1)$ como distribuição *a priori* de β_0 e os resultados do LOO são apresentados na Figura 7.3.

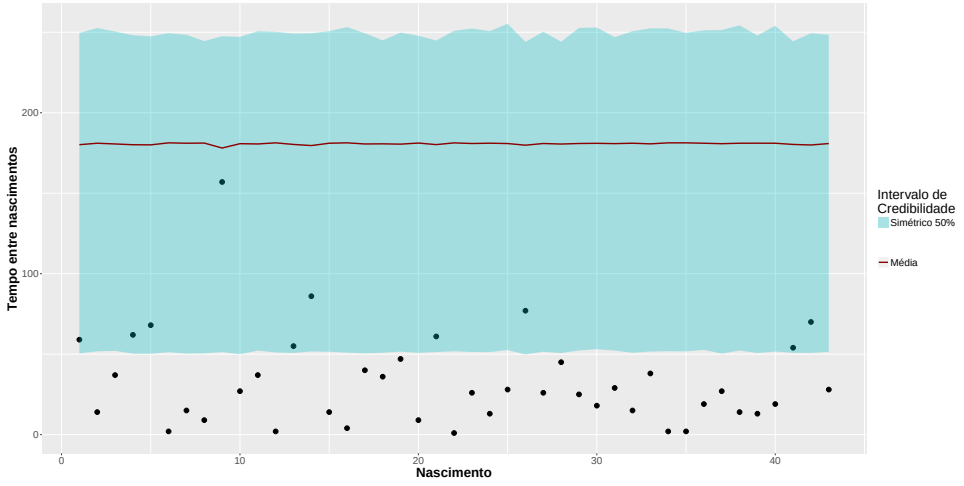


Figura 7.3: Modelo Exponencial com distribuição *a priori* $N(6{,}1)$

Pela Figura 7.3 é possível identificar que a má especificação da distribuição *a priori* resultou em intervalos preditivos distantes dos obtidos na Figura 7.2.

A propoção de acertos no exemplo foi 0,233 com $\xi_{\text{obs}} = 0{,}267$, indicando que com 95% de credibilidade, este modelo não possui boa capacidade preditiva para o problema. Se a aproximação proposta por Chen et al. (2012) tivesse sido utilizada para a obtenção do resultado do LOO, esse tipo de problema não seria identificado, pois ela não faz uso da distribuição *a priori* na estimação, o que levaria a conclusões erradas sobre o modelo.

7.2 Base *Leucemia*

7.2.1 Descrição dos Dados

A base *Leucemia*, apresentada por Hand et al. (1993), contém dados sobre o tempo de morte (em semanas) e a quantidade de glóbulos branco (medida em unidades de 10.000) para dois grupos de pacientes com leucemia, totalizando 33 observações. Os dados são apresentados na Tabela 7.6.

Tabela 7.6: Dados Leucemia

AG Positivo		AG Negativo	
Glóbulos Branco	Tempo	Glóbulos Branco	Tempo
0,23	65	0,44	56
0,075	156	0,3	65
0,43	100	0,4	17
0,26	134	0,15	7
0,6	16	0,9	16
1,05	108	0,53	22
1	121	1	3
1,7	4	1,9	4
0,54	39	2,7	2
0,7	143	2,8	3
0,94	56	3,1	8
3,2	26	2,6	4
3,5	22	2,1	3
10	1	7,9	30
10	1	10	4
5,2	5	10	43
10	65		

O modelo utilizado para aplicação é $T \sim$Exp(λ), com

$$\lambda = e^{-(\beta_0+\beta_1 GB+\beta_2 AG)}$$

sendo T o tempo até a morte do paciente por Leucemia, GB o número de Glóbulos Branco e AG o tipo de Leucemia ($AG\ Positivo = 1$ e $AG\ Negativo = 0$).

A metodologia proposta foi aplicada aos dados da base *Leucemia* utilizando uma priori difusa $N(0{,}100)$ para β_0, β_1 e β_2, simulando 100.000 amostras com saltos de tamanho 5 e burn-in de 10.000 no MCMC. Os resultados do LOO da verificação de capacidade preditiva pela metodologia são apresentados na Tabela 7.7 e na Figura 7.4.

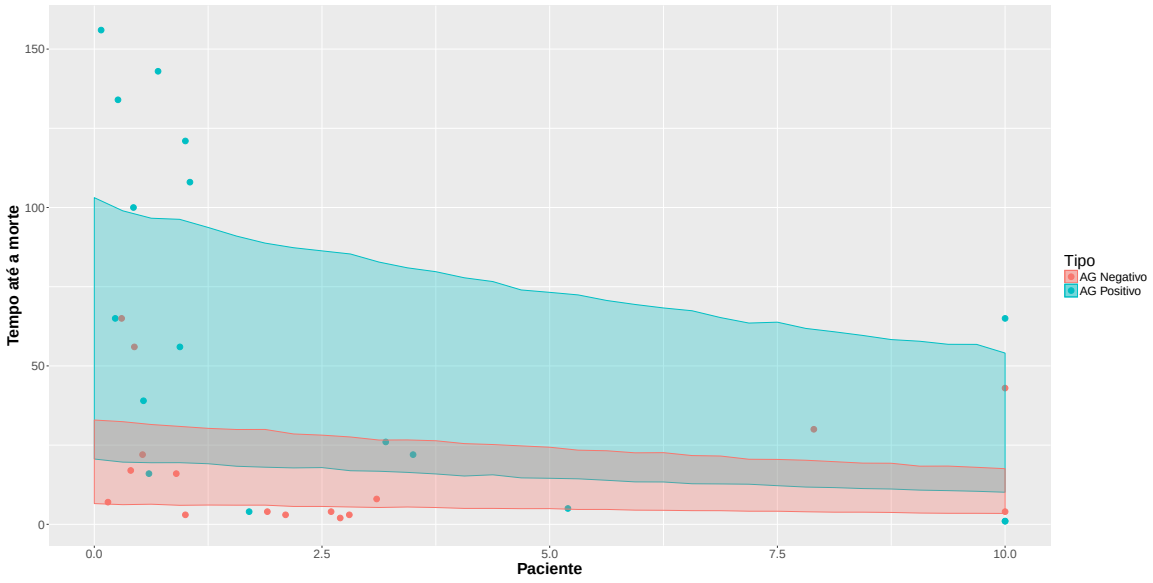


Figura 7.4: Intervalos de 50% de credibilidade obtidos no *LOO*

Tabela 7.7: Resultados do *LOO* da metodologia na base *Leucemia*

Paciente	Y	Tipo	Glóbulos Branco	β_0	β_1	β_2	λ^{-1}	Intervalo Predito	Acerto
1	65	1	0,230	3,165	1,123	-0,066	71,769	19,631 - 101,345	S
2	156	1	0,075	3,140	1,050	-0,060	65,706	17,792 - 91,030	N
3	100	1	0,430	3,160	1,089	-0,064	68,137	19,235 - 96,377	N
4	134	1	0,260	3,148	1,068	-0,061	66,686	18,736 - 93,256	N
5	16	1	0,600	3,168	1,167	-0,068	73,290	19,919 - 102,164	N
6	108	1	1,050	3,160	1,080	-0,065	64,831	18,273 - 90,130	N
7	121	1	1,000	3,155	1,070	-0,063	64,209	18,221 - 90,613	N
8	4	1	1,700	3,163	1,181	-0,065	68,887	19,329 - 96,511	N
9	39	1	0,540	3,166	1,143	-0,066	71,792	19,865 - 99,624	S
10	143	1	0,700	3,145	1,053	-0,062	63,744	17,813 - 88,817	N
11	56	1	0,940	3,163	1,126	-0,065	68,577	19,310 - 96,200	S
12	26	1	3,200	3,156	1,156	-0,063	60,949	16,756 - 84,807	S
13	22	1	3,500	3,151	1,169	-0,062	60,464	16,938 - 84,505	S
14	1	1	10,000	3,088	1,223	-0,045	47,558	12,668 - 70,333	N
15	1	1	10,000	3,088	1,223	-0,045	47,558	12,668 - 68,868	N
16	5	1	5,200	3,134	1,191	-0,058	55,793	15,509 - 78,343	N
17	65	1	10,000	3,302	0,941	-0,094	27,104	7,049 - 39,590	N
18	56	0	0,440	2,975	1,284	-0,049	19,171	5,375 - 27,543	N
19	65	0	0,300	2,929	1,315	-0,044	18,455	5,068 - 26,474	N
20	17	0	0,400	3,189	1,098	-0,068	23,615	6,462 - 33,303	S
21	7	0	0,150	3,236	1,059	-0,072	25,151	6,998 - 35,849	S
22	16	0	0,900	3,185	1,100	-0,066	22,776	6,299 - 32,327	S
23	22	0	0,530	3,164	1,124	-0,065	22,855	6,435 - 32,245	S
24	3	0	1,000	3,243	1,050	-0,071	23,865	6,599 - 33,669	N
25	4	0	1,900	3,237	1,061	-0,070	22,268	6,378 - 31,422	N
26	2	0	2,700	3,233	1,054	-0,068	21,100	5,991 - 30,074	N
27	3	0	2,800	3,230	1,057	-0,068	20,885	5,843 - 29,303	N
28	8	0	3,100	3,209	1,074	-0,067	20,124	5,649 - 27,977	S
29	4	0	2,600	3,227	1,055	-0,068	21,103	5,976 - 29,842	N
30	3	0	2,100	3,234	1,056	-0,070	21,924	6,062 - 31,291	N
31	30	0	7,900	3,118	1,184	-0,077	12,338	3,402 - 17,499	N
32	4	0	10,000	3,169	1,096	-0,055	13,697	3,722 - 19,551	S
33	43	0	10,000	3,118	1,291	-0,124	6,507	1,739 - 9,336	N
	Modelo Completo			3,157	1,115	-0,0637			

Pela Figura 7.4 é possível identificar que a capacidade preditiva para indivíduos com tipo de leucemia *AG Positivo* não foi satisfatória, pois grande parte dos pontos está fora do intervalo de 50% de credibilidade, indicando um possível mal ajuste do modelo.

A Tabela 7.8 apresenta a proporção de acertos e o erro ξ obtidos para os dados, que são as estatísticas necessárias para utilização do critério de rejeição da metodologia.

Tabela 7.8: Proporção de Acertos e Erro ξ da Metodologia

Proporção de acertos	$\xi_{\text{observado}}$
0,333	0,167

Utilizando a Tabela 6.3 é possivel obter o ξ crítico para uma amostra de tamanho $n = 33$. Os valores críticos são apresentados na Tabela 7.9.

Tabela 7.9: Valores Críticos para $n = 33$

α	0,01	0,05	0,10	0,20
$\xi_{\text{crítico}}$	0,220	0,166	0,140	0,110

Com as informações obtidas, tem-se que $\xi_{\text{observado}} > \xi_{\text{crítico}}$ para $\alpha = 0{,}05$. Assim, com 95% de credibilidade rejeita-se a hipótese de que o modelo Exponencial utilizado é um modelo com boa capacidade preditiva para o problema.

O $\xi_{\text{observado}}$ poderia ser um modelo utilizado caso a credibilidade empregada fosse $\alpha = 0{,}01$, pois o $\xi_{\text{crítico}}$ seria 0,220, não rejeitando a hipótese. Entretanto, esse tipo de decisão deve ser definida previamente pelo pesquisador que possui mais conhecimento sobre o problema.

É importante lembrar que rejeitar o modelo não significa que a distribuição de probabilidades utilizada não ajusta bem o problema, pois é possível que com a utilização de outras variáveis explicativas a capacidade preditiva do modelo aumente.

Atualmente, uma maneira de se avaliar este modelo seria pela utilização dos resíduos de Cox-Snell (Cox e Snell, 1968). Segundo Lawless (2011), os resíduos de Cox-Snell vêm de uma população homogênea e devem seguir uma distribuição exponencial com média 1.

A Figura 7.5 contém o Q-Q Plot que compara os resíduos de Cox-Snell calculados para o modelo utilizado na base de Leucemia com os quantis teóricos de uma distribuição Exponencial(1).

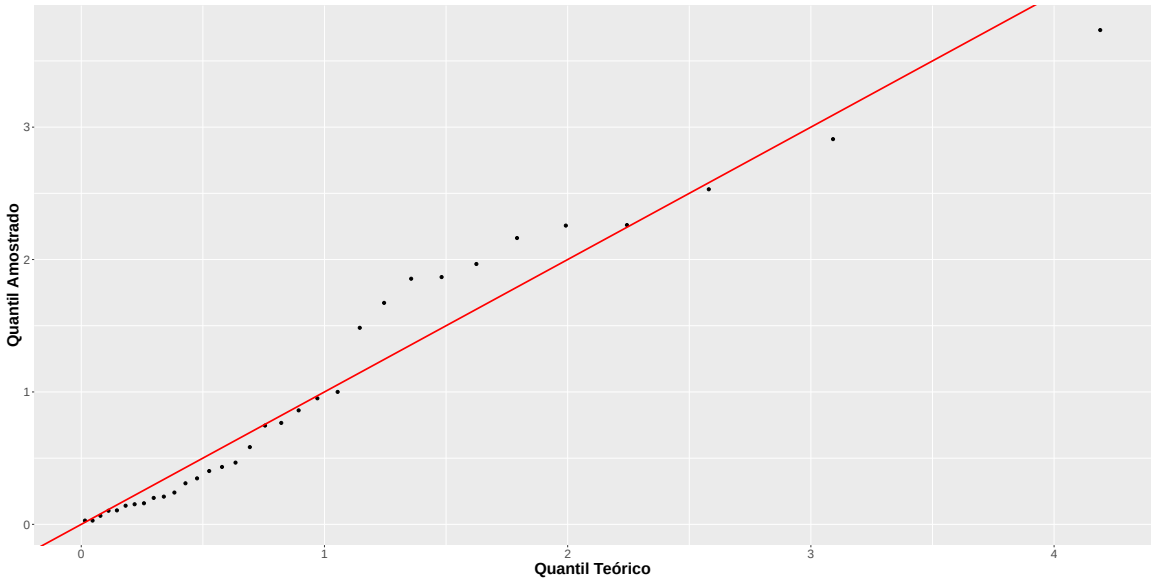


Figura 7.5: Q-Q plot do Resíduo de Cox-Snell com da distribuição Exponencial(1)

Pelo Q-Q Plot, o resíduo de Cox-Snell apresentou leve diferença do esperado para uma distribuição Exponencial(1). A Tabela 7.10 apresenta o resultado do teste de Kolmogorov-Smirnov e de Lilliefors para o resíduo de Cox-Snell.

Tabela 7.10: Teste de ajustamento de uma distribuição exponencial para a base *Leucemia*

	Kolmogorov-Smirnov	Lilliefors
p-valor	0,120	0,484

Baseado nos testes, não se rejeita a hipótese de que o resíduo de Cox-Snell segue uma distribuição Exponencial(1), consequentemente, não se rejeita a hipótese de que o modelo utilizado seja um bom modelo.

Apesar da metodologia proposta e o a análise de resíduos apresentarem resultados divergentes sobre o modelo utilizado, é importante lembrar que analisam características diferentes, sendo que a metodologia proposta analisa a acurácia do modelo e o resíduo de Cox-Snell avalia a aderência do modelo.

Os resultados reforçam que não necessariamente um modelo com boa aderência aos dados possui boa capacidade preditiva, sendo necessário avaliar qual a finalidade da modelagem, para que se possa empregar uma avaliação adequada para o modelo proposto.

Capítulo 8

Conclusão

Este trabalho apresentou uma adaptação para a metodologia proposta por Gelfand (1996), que apesar de simples e intuitiva, não permitia a validação de modelos de uma maneira objetiva. A adaptação possibilitou a definição de um critério de rejeição de modelos, proporcionando um meio de discriminação imparcial para a metodologia, visto que anteriormente a discriminação poderia variar dependendo do ponto de vista do usuário.

O deselvolvimento da proposta foi realizado sob uma perspectiva bayesiana de inferência, expondo os conceitos utilizados em sua elaboração e apresentando os procedimentos necessários para sua aplicação.

Os estudos realizados permitiram a elaboração de um critério de rejeição simples para a metodologia, pois verificou-se que a definição de regiões críticas só depende do tamanho amostral. Apesar do critério ter sido especificado apenas para modelos lineares generalizados com distribuição exponencial, os resultados incentivam o estudo do comportamento da metodologia em modelos que utilizem outras distribuições.

As aplicações em dados reais permitiram verificar que a metodologia é de simples aplicação e possui bons resultados, identificando quando o modelo não apresenta uma boa capacidade preditiva e apresentando resultados semelhantes a testes já muito utilizados.

Acredita-se que a facilidade de implementação e sua baixa complexidade tornem a metodologia proposta uma alternativa atraente para a avaliação de modelos, incentivando mais pesquisas nessa área.

8.1 Propostas Futuras

Nesta seção, apresentamos algumas propostas para a continuação deste trabalho

- Formalização estatística, por meio de testes de hipóteses, para relações assumidas no Capítulo 5 sobre o comportamento da *DPA*.
- Definir as regiões críticas para outras distribuições de probabilidade, buscando comportamentos semelhantes em distribuições da família exponencial.
- Implementação de métodos de aproximação do *leave-one-out* que tenham bons resultados para pequenos tamanhos amostrais.
- Adequação da metodologia para a utilização de dados censurados.

Apêndice A

Programação

```
require(MCMCpack)

## Posteriori para MCMC
expreg <- function(betas, Y, X, prioris){
   eta <- X %*% betas
   lambda <- exp(-eta)
   sum( (log(lambda) - lambda * Y) - sum( ( betas - prioris[1] )^2 / (2 *
       prioris[2]^2) )  )
}

##Função Leave-One-Out
louSim <- function(ii){
      i = as.numeric(ii)
      base <- baseoriginal[-i,]
      Xdata <- as.matrix(base[,-ncol(base)])
      yvector <- base$Y
      capture.output(
         posterior <- MCMCmetrop1R(expreg , theta.init=rep(1,ncol(Xdata)),
                 X=Xdata, Y=yvector, prioris = c(0,100), thin=3, mcmc=20000,
                 burnin=1000, verbose=0, logfun=TRUE),
         file='NUL'
      )
      posteriori <- data.frame(posterior)
```

```
      lambda <- exp(-(as.matrix(posteriori) %*%
          as.numeric(baseoriginal[i,-ncol(baseoriginal)])))
   #Intervalo preditivo
      ypred <- data.frame(sapply(lambda, function(k){rexp(1,rate=k)}))
      pred.sim <- quantile(ypred[[1]],c(0.25,0.75))
      valor <- baseoriginal[i,'Y']
      ComparaSim <- ifelse(valor > pred.sim[[1]] & valor < pred.sim[[2]],T,F)
      Compara <- mean(ComparaSim)
   #Barra de progresso
      setWinProgressBar(pb, i, title=paste( round(i/total*100, 0),"%"))
   #Resultado
      c(valor, as.numeric(baseoriginal[i,-ncol(baseoriginal)]),
          colMeans(posteriori), pred.sim, Compara, confianca )
}

## Gerando base de dados
B0 <- -2 ; B1 <- 0.02
baseoriginal <- data.frame(X0 = 1,X1 = runif(30,1,14))
confianca <- 0.5
baseoriginal$Y <- sapply(( B0 + B1 * baseoriginal$X1 ),
    function(u){rexp(1,exp(u))})

##Aplicando Leave-One-Out
total <- nrow(baseoriginal)
pb <- winProgressBar(title = "Progresso", min = 0, max = total, width = 300)
resultado = data.frame(t(sapply(1:nrow(baseoriginal),tryCatch(louSim, error =
    function(e){cat("ERROR :",conditionMessage(e), "\n")}))))
names(resultado) =
    c('Y','X0','X1','B0','B1','LimInf','LimSup','ComparaSim','Confianca')
close(pb)

## Calculando proporção
resultado = transform(resultado, Acertos = mean(resultado$ComparaSim))
resultado
```

Referências Bibliográficas

Box, G. E. (1980). Sampling and bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society. Series A (General)*, pages 383–430.

Chen, M.-H., Huang, L., Ibrahim, J. G., e Kim, S. (2008). Bayesian variable selection and computation for generalized linear models with conjugate priors. *Bayesian analysis (Online)*, 3(3):585.

Chen, M.-H., Shao, Q.-M., e Ibrahim, J. G. (2012). *Monte Carlo methods in Bayesian computation*. Springer Science & Business Media.

Cox, D. R. e Snell, E. J. (1968). A general definition of residuals. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 248–275.

D'Agostino, R. B. (1986). *Goodness-of-fit-techniques*, volume 68. CRC press.

Dunn, P. K. (1999). A simple data set for demonstrating common distributions. *Journal of Statistics Education*, 7(3).

Ferguson, T. S. (1973). A bayesian analysis of some nonparametric problems. *The annals of statistics*, pages 209–230.

Geisser, S. e Eddy, W. F. (1979). A predictive approach to model selection. *Journal of the American Statistical Association*, 74(365):153–160.

Gelfand, A. E. (1996). Model determination using sampling-based methods. *Markov chain Monte Carlo in practice*, pages 145–161.

Gelfand, A. E. e Smith, A. F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, 85(410):398–409.

Gelman, A., Carlin, J. B., Stern, H. S., e Rubin, D. B. (2014). *Bayesian data analysis*, volume 2. Taylor & Francis.

Gelman, A., Meng, X.-L., e Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica sinica*, pages 733–760.

Hammersley, J. e Handscomb, D. (1964). Monte carlo methods. *Methuen*.

Hand, D. J., Daly, F., McConway, K., Lunn, D., e Ostrowski, E. (1993). *A handbook of small data sets*, volume 1. cRc Press.

Hastings, W. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika*.

Kass R. E., R. A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90:773–795.

Kruschke, J. K. (2013). Posterior predictive checks can and should be bayesian: Comment on gelman and shalizi,?philosophy and the practice of bayesian statistics? *British Journal of Mathematical and Statistical Psychology*, 66(1):45–56.

Lawless, J. F. (2011). *Statistical models and methods for lifetime data*, volume 362. John Wiley & Sons.

Lilliefors, H. W. (1969). On the kolmogorov-smirnov test for the exponential distribution with mean unknown. *Journal of the American Statistical Association*, 64(325):387–389.

Massey Jr, F. J. (1951). The kolmogorov-smirnov test for goodness of fit. *Journal of the American statistical Association*, 46(253):68–78.

Nelder, J. A. e Baker, R. J. (1972). Generalized linear models. *Encyclopedia of Statistical Sciences*.

Paulino, C. D. M., Turkman, M. A. A., e Murteira, B. (2003). *Estatística bayesiana*.

Polidoro, M. J. F. P. (2014). Metodologia bayesiana e adequação de modelos.

Sinharay, S. e Stern, H. S. (2003). Posterior predictive model checking in hierarchical models. *Journal of Statistical Planning and Inference*, 111(1):209–221.

Vehtari, A., Ojanen, J., et al. (2012). A survey of bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys*, 6:142–228.