

# ABORDAGEM BAYESIANA PARA POLINÔMIOS FRACIONÁRIOS

Dennison Célio de Oliveira Carvalho

Tese apresentado à Universidade Estadual Paulista “Júlio de Mesquita Filho” para obtenção do título de Doutor em Biometria.

BOTUCATU  
São Paulo - Brasil  
Fevereiro – 2019

# ABORDAGEM BAYESIANA PARA POLINÔMIOS FRACIONÁRIOS

**Dennison Célio de Oliveira Carvalho**

Orientadora: Profa. Dra. **Miriam Harumi Tsunemi**

Coorientadora: Profa. Dra. **Luzia Aparecida Trinca**

Tese apresentado à Universidade Estadual Paulista “Júlio de Mesquita Filho” para obtenção do título de Doutor em Biometria.

BOTUCATU  
São Paulo - Brasil  
Fevereiro – 2019

# Ficha Catalográfica

FICHA CATALOGRÁFICA ELABORADA PELA SEÇÃO TÉC. AQUIS. TRATAMENTO DA INFORM.  
DIVISÃO TÉCNICA DE BIBLIOTECA E DOCUMENTAÇÃO - CÂMPUS DE BOTUCATU - UNESP  
BIBLIOTECÁRIA RESPONSÁVEL: LUCIANA PIZZANI-CRB 8/6772

Carvalho, Dennison.

Abordagem Bayesiana para polinômios fracionários /  
Dennison Carvalho. - Botucatu, 2019

Tese (doutorado) - Universidade Estadual Paulista  
"Júlio de Mesquita Filho", Instituto de Biociências de  
Botucatu

Orientador: Miriam Harumi Tsunemi

Coorientador: Luzia Aparecida Trinca

Capes: 10202064

1. Polinômios. 2. Estatística. 3. Monte Carlo, Método  
de. 4. Análise de regressão logística. 5. Teoria Bayesiana  
de decisão estatística.

Palavras-chave: Métodos de Monte Carlo via Cadeias de  
Markov; Monte Carlo Hamiltoniano; Regressão não linear.

## Agradecimentos

À Deus por ter me concedido a vida e não me deixar desistir diante das inúmeras tribulações que passei durante o período do doutorado;

Aos meus pais Amélia Carvalho e Raimundo Carvalho por terem me educado com tanta sabedoria e simplicidade. Sem vocês jamais chegaria até aqui;

À minha família, em especial à minha irmã Diane Ellen e à minha amada sobrinha Alice Carvalho, fontes de inspiração nos dias mais difíceis;

À minha orientadora Profa. Dra. Miriam Tsunemi pelos ensinamentos e por ter sido tão humana e compreensível quando mais precisei;

À minha co-orientadora Profa. Dra. Luzia Trinca por aceitar o desafio de me orientar às pressas diante das dificuldades que tive no decorrer de todo o processo e pelos ensinamentos;

A todos meus amigos e amigas que contribuíram de alguma forma no dia a dia para realização deste trabalho;

A todos os professores e professoras que contribuíram de alguma forma na minha formação;

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) e a Fundação Amazônia de Amparo a Estudos e Pesquisas (Fapespa) pelo apoio financeiro dado através do programa de formação doutoral docente (Prodoutoral).



# Sumário

	Página
<b>LISTA DE FIGURAS</b>	<b>vii</b>
<b>LISTA DE TABELAS</b>	<b>ix</b>
<b>LISTA DE QUADROS</b>	<b>xi</b>
<b>RESUMO</b>	<b>xiv</b>
<b>SUMMARY</b>	<b>xvi</b>
<b>1 INTRODUÇÃO</b>	<b>1</b>
<b>2 REVISÃO DE LITERATURA</b>	<b>5</b>
2.1 Regressão polinomial . . . . .	5
2.2 Regressão linear bayesiana . . . . .	6
2.3 Inferência bayesiana - aspectos computacionais . . . . .	9
2.3.1 <i>MCMC</i> . . . . .	9
2.3.2 Monte Carlo Hamiltoniano ( <i>HMC</i> ) . . . . .	11
2.3.3 Algoritmo <i>NUTS</i> . . . . .	13
2.3.4 Avaliação da convergência . . . . .	15
2.3.5 Implementação computacional - <code>rjags</code> e <code>rstan</code> . . . . .	17
2.4 Polinômios fracionários clássicos . . . . .	18
2.4.1 Ajuste do modelo . . . . .	21

	v
2.4.2 Escolha do modelo <i>FP</i> . . . . .	21
2.4.3 Polinômios fracionários multivariado . . . . .	22
2.4.4 Escala e centralização . . . . .	24
2.5 Polinômios fracionários bayesianos . . . . .	25
2.5.1 Definição do modelo . . . . .	26
2.5.2 Distribuição a priori . . . . .	27
2.5.3 Distribuição a posteriori . . . . .	28
2.5.4 Inferência . . . . .	29
<b>3 METODOLOGIA</b>	<b>31</b>
3.1 Um breve estudo para investigar o desempenho da proposta bayesiana de Bové & Held (2011) . . . . .	31
3.2 Uma proposta de modelos <i>FP</i> Bayesianos . . . . .	32
<b>4 RESULTADOS</b>	<b>43</b>
4.1 Exploração dos <i>FPs</i> sob a abordagem bayesiana . . . . .	43
4.1.1 Aplicações para o conjunto de dados IMC (Royston & Sauerbrei, 2008)	44
4.1.2 Aplicações para o conjunto de dados <i>PSA</i> (Royston & Sauerbrei, 2008)	46
4.1.3 Aplicações para o conjunto de dados ozônio (Bové & Held, 2011) . . .	48
4.1.4 Avaliação de percentagem de acertos das abordagens <i>FP</i> via simulações	50
4.2 Avaliação da modelagem <i>FP</i> via regressão não linear bayesiana . . . . .	53
4.2.1 Estudo de casos simulados de modelos <i>FP1</i> . . . . .	54
4.2.2 Estudos de casos simulados de modelos <i>FP2</i> modificado segundo a abor- dagem proposta . . . . .	59
4.2.3 Aplicação para o conjunto de dados IMC . . . . .	62
4.2.4 Aplicação para o conjunto de dados aranha . . . . .	64
<b>5 CONCLUSÕES</b>	<b>67</b>
<b>ANEXOS</b>	<b>70</b>

**REFERÊNCIAS BIBLIOGRÁFICAS**

## Lista de Figuras

	Página
1    Ajuste de quatro modelos para os dados da pesquisa sobre gordura corporal em função do IMC (índice de massa corporal) disponível em Royston & Sauerbrei (2008). . . . .	20
2    Valores da transformação logística de $\theta$ segundo equação (38) para $-3 \leq \theta \leq 3$ . . . . .	36
3    Relação entre valores preditos e a resposta observada para os modelos aplicados aos dados IMC, sob diferentes metodologias. . . . .	46
4    Relação entre os valores preditos e a resposta observada para os modelos aplicados aos dados <i>PSA</i> , sob diferentes metodologias . . . . .	48
5    Relação entre os valores preditos e a resposta observada para os modelos aplicados aos dados ozônio, sob diferentes metodologias . . . . .	50
6    Dados simulados sob modelos <i>FP1</i> com diferentes potências. . . . .	55
7    Dados simulados sob modelos <i>FP1</i> com diferentes potências. . . . .	57
8    Comparação dos dados observados $y$ com os dados replicados da distribuição preditiva $y_{rep}$ . . . . .	59
9    Dados simulados sob modelos <i>FP2</i> com potência diferentes, acordo com a técnica <i>HMC</i> . . . . .	60
10   Comparação dos dados observados $y$ com os dados replicados da distribuição preditiva $y_{rep}$ . . . . .	62
11   Comparação dos dados observados $y$ com os dados replicados da distribuição preditiva $y_{rep}$ , dados IMC. . . . .	63

12	Diagrama de dispersão do comprimento (mm) e peso (mg) de aranhas. .	64
13	Comparação dos dados observados $y$ com os dados replicados da distribuição preditiva $y_{rep}$ , para o conjunto de dados aranhas. . . . .	66
14	Traço dos dados simulados sob modelos $FP1$ com diferentes potências. .	80
15	Traço dos dados simulados sob modelos $FP1$ com diferentes potências. .	81
16	Traço dos dados simulados sob modelos $FP2$ com diferentes potências. .	82
17	Traço dos dados IMC e aranha. . . . .	83

## Lista de Tabelas

	Página
1 Estimativas dos parâmetros e medidas de qualidade dos modelos ajustados - dados IMC (Royston & Sauerbrei, 2008) . . . . .	45
2 Estimativas dos parâmetros e medidas de qualidade dos modelos ajustados, sob diferentes abordagens - dados <i>PSA</i> (Royston & Sauerbrei, 2008)	47
3 Estimativas dos parâmetros e medidas de qualidade dos modelos ajustados - dados ozônio (Bové & Held, 2011) . . . . .	49
4 Porcentagem de acertos no grau do <i>FP</i> segundo a abordagem de ajuste de modelo quando a função subjacente é um <i>FP1</i> (resultados de 1.000 simulações) . . . . .	52
5 Porcentagem de acertos no grau do <i>FP</i> segundo a abordagem de ajuste de modelo quando a função subjacente é um <i>FP2</i> (resultados de 5.000 simulações) . . . . .	53
6 Propriedades dos ajustes, segundo a abordagem proposta, para modelos <i>FP1</i> , dados simulados com potências diferentes, de acordo com as técnicas <i>HMC</i> e <i>MCMC</i> . . . . .	56
7 Propriedades dos ajustes, segundo a abordagem proposta, para modelos <i>FP1</i> , dados simulados com potências diferentes, de acordo com a técnica <i>HMC</i> . . . . .	58
8 Propriedades dos ajustes, segundo a abordagem proposta, para modelos <i>FP2</i> , dados simulados com potência diferentes, de acordo com a técnica <i>HMC</i> . . . . .	61

9	Estimativas dos parâmetros sob a abordagem proposta, para o conjunto de dados IMC, utilizando a técnica <i>HMC</i> . . . . .	63
10	Estimativas dos parâmetros sob a abordagem proposta, para o conjunto de dados aranha, utilizando a técnica <i>HMC</i> . . . . .	65

## Lista de Quadros

	Página
1    Descrição do código para o modelo <i>FP1</i> proposto, no <b>rstan</b> . . . . .	38
2    Descrição do código para o modelo <i>FP2</i> proposto, no <b>rstan</b> . . . . .	40



## Siglas e abreviações

<i>JAGS</i>	Just another Gibbs Sampler (Apenas outro amostrador de Gibbs)
<i>OLS</i>	Ordinary least squares (Mínimos Quadrados Ordinários)
BE	backward elimination (Eliminação passo atrás)
bfp	Bayesian Fractional Polynomials (Polinômios Fracionários Bayesianos)
BMA	Bayesian model average (Ponderação Bayesiana de modelos)
bmi	Body-mass index (Índice de massa corporal)
BUGS	<i>Bayesian inference using Gibbs sampling</i> (Inferência Bayesiana usando o amostrador de Gibbs)
FP	Fractional Polynomials (Polinômios Fracionários)
FSP	function select procedure (Procedimento para seleção da função FP)
HMC	Halmitonian Monte Carlo (Monte Carlo Hamiltoniano)
MAP	maximum a posteriori (Máxima Densidade a Posteriori)
MCMC	Markov chain Monte Carlo (Métodos de Monte Carlo via Cadeias de Markov)
MFP	mulivariable FP procedure (Polinômios Fracionários Multivariados)
MLE	maximum likelihood estimation (estimador de máxima verossimilhança)
mlg	Modelos Lineares Generalizados

MSPE mean squared prediction error (Erro quadrático médio de predição)

NUTS *No-U-Sampler*

pbfm Percentage body fat content (Percentual de gordura no corpo)

PSA Prostate-Specific Antigens (Antígenos específicos da próstata)

RMSPE means of the root mean squared prediction error (Raíz quadrada do erro quadrático médio de predição)

Stan Sampling through adaptive neighborhoods (Amostragem através de vizinhanças adaptativas)

# ABORDAGEM BAYESIANA PARA POLINÔMIOS FRACIONÁRIOS

Autor: DENNISON CÉLIO DE OLIVEIRA CARVALHO

Orientadora: Profa. Dra. MIRIAM HARUMI TSUNEMI

Coorientadora: Profa. Dra. LUZIA APARECIDA TRINCA

## RESUMO

Em inúmeras situações práticas a relação entre uma variável resposta e uma ou mais covariáveis é curvada. Dentre as diversas formas de representar esta curvatura, Royston e Altman (1994) propuseram uma extensa família de funções denominada de Polinômios Fracionários (*Fractional Polynomials* - *FP*). Bové e Held (2011) implementaram o paradigma bayesiano para *FP* sob a suposição de normalidade dos erros. Sua metodologia é fundamentada em uma distribuição a priori *hiper - g* (Liang et al., 2008), que, além de muitas propriedades assintóticas interessantes, garante uma predição bayesiana de modelos consistente. Nesta tese, compara-se as abordagens clássica e Bayesiana para *PF* a partir de dados reais disponíveis na literatura, bem como por simulações. Além disso, propõem-se uma abordagem Bayesiana para modelos *FPs* em que a potência, diferentemente dos métodos usuais, pode assumir qualquer valor num determinado intervalo real e é estimada via métodos de simulação *HMC* (Monte Carlo Hamiltoniano) e *MCMC* (Métodos de Monte Carlo

via Cadeias de Markov). Neste modelo, para o caso de um  $FP$  de segunda ordem, ao contrário dos modelos atualmente disponíveis, apenas uma potência é estimada. Avalia-se este modelo a partir de dados simulados e em dados reais, sendo um deles com transformação de Box-Cox.

**Palavras-chave:** *Regressão não linear, Monte Carlo Hamiltoniano, Métodos de Monte Carlo via Cadeias de Markov.*

# BAYESIAN APPROACH TO FRACTIONAL POLYNOMIALS

Author: DENNISON CÉLIO DE OLIVEIRA CARVALHO

Adviser: Profa. Dra. MIRIAM HARUMI TSUNEMI

Co-Adviser: Profa. Dra. LUZIA APARECIDA TRINCA

## SUMMARY

In many practical situations the relationship between the response variable and one or more covariates is curved. Among the various ways of representing this curvature, Royston and Altman (1994) proposed an extended family of functions called Fractional Polynomials (FP). Bové and Held (2011) implemented the Bayesian paradigm for FP on the assumption of error normality. Their methodology is based on a *hyper-g* prior distribution, which, in addition to many interesting asymptotic properties, guarantees a consistent Bayesian model average (BMA). In addition, a Bayesian approach is proposed for FPs models in which power, unlike the usual methods, can obtain any numerical real interval value and is estimated via HMC (Monte Carlo Hamiltonian) and MCMC (Markov chain Monte Carlo). In this model, in the case of a second-order FP, unlike the currently available models, only one power is estimated. This model is evaluated from simulated data and real data, one of them with Box-Cox transformation.

**Keywords:** *nonlinear regression, Monte Carlo Hamiltonian, Markov chain Monte Carlo.*

# 1 INTRODUÇÃO

O termo regressão foi criado por Francis Galton. Em um artigo famoso, Galton verificou que, embora existisse uma tendência de que pais altos tivessem filhos altos e pais baixos tivessem filhos baixos, a estatura média das crianças nascidas de pais com uma dada altura tendia a mover-se ou “regredir” à altura média da população como um todo<sup>1</sup>. Em outras palavras, a altura dos filhos de pais mais altos ou mais baixos que o padrão tende a mover-se no sentido da altura média da população. A lei da regressão universal de Galton foi confirmada por seu amigo Karl Pearson, que coletou mais de mil registros de altura de membros de grupos familiares<sup>2</sup>. Ele constatou que a altura média dos filhos de um grupo de pais altos era menor do que a de seus pais e que a altura média de um grupo de filhos de pais baixos era maior do que a de seus pais; portanto, filhos de pais altos ou baixos “regrediam” igualmente à altura média de todos os homens. Nas palavras de Galton, isso era uma “regressão à mediocridade” (Gujarati & Poter, 2011).

Uma interpretação moderna, segundo Gujarati & Poter (2011), é que a análise de regressão diz respeito ao estudo da dependência de uma variável, a variável dependente, em relação a uma ou mais variáveis, as variáveis explanatórias, visando estimar e/ou prever o valor médio (da população) da primeira em termos dos valores conhecidos ou fixados (em amostragens repetidas) das segundas.

As aplicações dos modelos de regressões são numerosas e ocorrem em

---

<sup>1</sup>Galton, Francis. “Family likeness in stature”. Proceedings of Royal Society, Londres, 1886. v. 40, p. 42-72.

<sup>2</sup>Pearson, K.; Lee, A. “On the laws of inheritance”. Biometrika, nov. 1903. v. 2, p. 357-462.

quase todos os campos, incluindo engenharia, ciências físicas e químicas, economia, gestão, ciências biológicas e ciências sociais. Na verdade, a análise de regressão pode ser a técnica estatística mais utilizada (Montgomery et al., 2012).

Segundo Gordon (2015) modelos de regressão com apenas uma variável preditora (explanatória) é tipicamente chamada de regressão simples ou bivariada (bivariada porque duas variáveis estão envolvidas - a variável de saída, frequentemente denotada por  $Y$ , e a variável de entrada, frequentemente denotada por  $X$ , também conhecida como variável preditora ou regressora). No centro dos modelos de regressão bivariada esta a geometria de uma linha reta. Uma linha reta é definida por um intercepto, que denota o ponto em que a linha passa pelo eixo  $Y$ , e um coeficiente angular, que mede o quanto  $Y$  muda quando  $X$  aumenta em uma unidade.

Apenas dois pontos distintos são necessários para ajustar uma linha reta. Se tivermos três pontos distintos, então poderíamos inserir uma parábola (um modelo de segunda ordem). Se escolhermos encaixar apenas a linha reta, então notamos que em geral, a linha reta não passa pelos três pontos. Nós normalmente assumimos que esse fenômeno é devido a erro. Por outro lado, o verdadeiro mecanismo subjacente poderia ser realmente quadrático. Polinômios de ordem 2 ou superior são amplamente utilizados em situações em que a resposta é curvilínea, que é uma importante classe da regressão chamada regressão polinomial (Montgomery et al., 2012).

Existem diversos tipos de modelos de regressão como os descritos por Royston & Sauerbrei (2008): regressão com erros normais, regressão logística, regressão de Cox, modelos lineares generalizados (*mlg*) e regressão não linear. Temos inúmeras alternativas para estudar a relação entre uma variável de saída e uma ou mais preditoras, quando esta é curvada. Os polinômios quadráticos e cúbicos são usados para este fim, no entanto, estes polinômios de baixa ordem são limitados. Uma família mais geral de modelos denominada Polinômios Fracionários, proposta



por Royston & Altman (1994), em que um, dois ou mais termos na forma  $x^p$  são ajustados, em que  $p$  é escolhido dentro de um pequeno conjunto de valores inteiros e não inteiros, é uma alternativa para ajustar dados com curvatura. Royston & Sauerbrei (2008) desenvolveram a metodologia de *FP* para modelos com erros normais e estenderam, com pequenas modificações, para modelos lineares generalizados, modelos de sobrevivência, etc. A metodologia foi implementada no pacote `mfp` (original por Gareth Ambler e modificado por Axel Benner, 2015), no *software* R (R Core Team, 2018).

Bové & Held (2011) propuseram uma abordagem Bayesiana para *FP* com erros normais, implementada no pacote `bfp` (Bové & Held, 2011), também no *software* R. Nesta abordagem mantiveram o conjunto discreto para as potências e fazem o uso de uma distribuição priori *hiper-g* (Liang et al., 2008) para os coeficientes das regressoras que garante uma distribuição fechada para a distribuição a posteriori.

O uso de um conjunto discreto para as potências tem a grande vantagem de evitar as dificuldades de estimação de parâmetros não lineares e talvez seja a única alternativa de lidar com inúmeras variáveis regressoras. Entretanto, para uma ou algumas regressoras, a ampliação do conjunto discreto para um intervalo real pode representar uma grande flexibilização na família de curvas disponíveis para analisar dados.

Nesse contexto, esta tese tem dois objetivos principais. O primeiro deles é fazer um estudo de comparação entre as propostas disponíveis, utilizando-se para isso, dados publicamente disponíveis. Esse estudo é complementado com a simulação de casos de modelos conhecidos para se verificar a porcentagem de acertos na estimação do grau do *FP* em ambas abordagens, clássica e bayesiana.

O segundo objetivo é propor e avaliar uma abordagem de estimação Bayesiana para modelos *FPs* nos quais os parâmetros potências podem assumir valores reais dentro de um intervalo. Para os casos de curvas de grau 2, propõe-se também o uso de uma modificação na definição do *FP* que requer a estimação de apenas

uma potência por variável regressora. A abordagem proposta é avaliada por alguns estudos de casos simulados e aplicação a conjuntos de dados reais.

## 2 REVISÃO DE LITERATURA

Neste capítulo é apresentada uma breve revisão de literatura sobre modelos de regressão polinomial, regressão linear bayesiana e aspectos computacionais sobre inferência bayesiana, em que são descritos detalhes sobre os métodos de simulação utilizados nesta tese. Posteriormente, apresentam-se as abordagens clássica e bayesiana para *FPS*.

### 2.1 Regressão polinomial

Rawlings et al. (1998) comentam que uma simples extensão do modelo linear envolvendo uma variável independente é o modelo polinomial de segunda ordem (quadrático), dado por

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon, \quad (1)$$

com a suposição usual de  $\epsilon \sim \text{Normal}(0, \sigma^2)$ .

Note que este modelo é um caso especial do modelo de regressão múltipla, em que  $X_1 = X$  e  $X_2 = X^2$ . Um modelo de regressão polinomial de alta ordem pode ser escrito como

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \cdots + \beta_p X^p + \epsilon. \quad (2)$$

O modelo em (1) e em (2) pode ser estendido para mais de uma regressora, embora devido ao possível mal condicionamento da matriz do modelo, causado pelos termos de potências altas, e dificuldade de interpretação, apenas polinômios de ordem

baixa são viáveis na prática. Sob as condições usuais, homocedasticidade dos erros e não correlação, os métodos de estimação de mínimos quadrados ordinários (*Ordinary least squares* - *OLS*) são apropriados para este tipo de modelo. Detalhes do método, com expressões e propriedades dos estimadores, além de técnicas de diagnóstico do ajuste, sob a perspectiva de inferência frequentista, podem ser encontradas em Draper & Smith (2014), Montgomery et al. (2012), Gujarati & Poter (2011), entre outros.

## 2.2 Regressão linear bayesiana

Na abordagem bayesiana, o grau de incerteza inicial sobre os parâmetros desconhecidos do vetor  $\boldsymbol{\theta} \in \Theta$  de um modelo deve ser representado por um modelo de probabilidade para  $\boldsymbol{\theta}$ , isto é,  $\boldsymbol{\theta}$  é uma variável aleatória. Este modelo de probabilidade, que representa a incerteza inicial ou a priori sobre  $\boldsymbol{\theta}$  é denotado por  $f(\boldsymbol{\theta})$  e chamado de distribuição a priori de  $\boldsymbol{\theta}$ . Deste modo,  $f(\boldsymbol{\theta})$  indica o grau de credibilidade dado pelo pesquisador ao vetor de parâmetro de interesse, em que  $f(\boldsymbol{\theta})$  pode ser uma função de probabilidade, no caso em que  $\boldsymbol{\theta}$  é discreto, ou uma função densidade de probabilidade no, caso contínuo.

Uma das maiores dificuldades iniciais da inferência bayesiana está na escolha da distribuição a priori. Estas distribuições são de fundamental importância, visto que, diferentes escolhas podem gerar resultados divergentes para uma mesma verossimilhança (mais detalhes em Paulino et al., 2003, Gelman et al., 2014, Ghosh et al., 2006, dentre outros).

No entanto, existem diversos métodos para construção, escolha e comparação de prioris (Paulino et al., 2003). Entre as recomendações básicas para a especificação de distribuições a priori, tem-se características como: respeitar o espaço paramétrico; conduzir a uma posteriori própria; refletir o conhecimento do especialista, dentre outros.

Ao realizar inferências sobre o vetor de parâmetros não observável  $\boldsymbol{\theta}$ , os métodos clássicos se baseiam em probabilidades associadas com diferentes amostras  $\mathbf{X}$ , que poderiam ocorrer para algum valor fixo, mas desconhecido, do parâmetro  $\boldsymbol{\theta}$ . É o que ocorre quando se fazem inferências com base nas distribuições amostrais de certas estatísticas (Paulino et al., 2003). Já na metodologia bayesiana, todo o processo de inferência se baseia na distribuição a posteriori de  $\boldsymbol{\theta}$ , dada por

$$f(\boldsymbol{\theta}|\mathbf{x}) = \frac{f(\mathbf{x}|\boldsymbol{\theta})f(\boldsymbol{\theta})}{\int_{\Theta} f(\mathbf{x}|\boldsymbol{\theta})f(\boldsymbol{\theta})d\boldsymbol{\theta}}, \quad \boldsymbol{\theta} \in \Theta, \quad (3)$$

na qual  $f(\mathbf{x}|\boldsymbol{\theta})$  representa a distribuição conjunta da amostra. Como o denominador da equação (3), conhecido como distribuição preditiva, não depende de  $\boldsymbol{\theta}$ , este funciona como uma constante normalizadora de  $f(\boldsymbol{\theta}|\mathbf{x})$ . Sendo assim, a equação (3) pode ser escrita como o produto da verossimilhança pela priori, a menos de uma constante,

$$f(\boldsymbol{\theta}|\mathbf{x}) \propto f(\mathbf{x}|\boldsymbol{\theta})f(\boldsymbol{\theta}), \quad \boldsymbol{\theta} \in \Theta. \quad (4)$$

Segundo Gelman et al. (2014), grande parte da análise estatística aplicada é baseada em técnicas de regressão linear simples. Tais técnicas podem ser pensadas sobre o paradigma bayesiano, em que, pode-se fazer inferências a posteriori para os parâmetros de um modelo de regressão, dada uma distribuição a priori não-informativa.

Segundo Paulino et al. (2003), a notação matricial para o modelo de regressão linear é dado por:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad (5)$$

em que  $\mathbf{Y}$  é um vetor coluna  $n \times 1$  de observações  $Y_i$ ,  $\mathbf{X}$  é uma matriz  $n \times p$  cuja primeira coluna é constituída por elementos 1, seguida pelos vetores das  $p-1$  colunas das observações sobre as variáveis independentes,  $\boldsymbol{\beta}$  é o vetor de parâmetros  $p \times 1$  a ser estimado e  $\boldsymbol{\epsilon}$  é o vetor de erros  $n \times 1$ .

Gelman et al. (2014) descrevem o modelo de regressão linear simples em notação vetorial:

$$\mathbf{y}|\boldsymbol{\beta}, \sigma, X \sim N(X\boldsymbol{\beta}, \sigma^2\mathbf{I}), \quad (6)$$

em que  $\mathbf{I}$  é a matriz identidade  $n \times n$ .

É comum usar a seguinte distribuição conjunta a priori não-informativa

$$p(\boldsymbol{\beta}, \sigma^2|\mathbf{X}) \propto \sigma^{-2}. \quad (7)$$

Desta forma, a distribuição condicional a posteriori do vetor de parâmetros  $\boldsymbol{\beta}$ , dado  $\sigma$ , é

$$\boldsymbol{\beta}|\sigma, \mathbf{y} \sim N(\hat{\boldsymbol{\beta}}, V_{\boldsymbol{\beta}}\sigma^2), \quad (8)$$

em que

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \quad (9)$$

$$V_{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1}. \quad (10)$$

A distribuição a posteriori de  $\sigma^2$  pode ser escrita como:

$$p(\sigma^2|\mathbf{y}) = \frac{p(\boldsymbol{\beta}, \sigma^2|\mathbf{y})}{p(\boldsymbol{\beta}|\sigma^2, \mathbf{y})}, \quad (11)$$

que pode ser visto como uma distribuição qui-quadrado inversa escalonada

$$\sigma^2|\mathbf{y} \sim \text{Inv-}\chi^2(n - p, s^2), \quad (12)$$

em que

$$s^2 = \frac{1}{n - p}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}). \quad (13)$$

Segundo Gelman et al. (2014), a distribuição a posteriori marginal de  $\boldsymbol{\beta}|y$ , integrando em  $\sigma$ , é *t-Student* multivariada com  $n - p$  graus de liberdade, mas, na prática, raramente usamos este fato ao fazer inferências por simulação, uma vez que para caracterizar a distribuição posterior conjunta podemos fazer inferências simulando de  $\sigma$  e então  $\boldsymbol{\beta}|\sigma$ . Com estas configurações, os estimadores  $\hat{\boldsymbol{\beta}}$  e  $s$  coincidem com os estimadores da inferência clássica.

## 2.3 Inferência bayesiana - aspectos computacionais

### 2.3.1 *MCMC*

Na inferência bayesiana, é comum não encontrar uma solução analítica para a distribuição a posteriori. Nestas situações, métodos de simulação da distribuição a posteriori são frequentemente usados. A ideia básica dos métodos *MCMC* consiste em obter amostras da distribuição a posteriori de forma iterativa, sem que se conheça sua densidade.

Os métodos *MCMC* foram desenvolvidos por físicos em meados de 1950. Estes métodos permitem obter distribuições a posteriori para parâmetros e variáveis latentes (variáveis não observáveis) de modelos complexos. Em paralelo, a melhora dos computadores pessoais nos anos 90 e *softwares* amigáveis com diferentes programas baseados na linguagem de programação *BUGS*<sup>3</sup>, surgiram. Esses desenvolvimentos impulsionaram o uso da abordagem bayesiana em análise de dados, particularmente em genética e ecologia (Korner-Nievergelt et al., 2015).

Existem diversos algoritmos *MCMC*, dentre os quais destaca-se os mais usuais: algoritmo Metropolis-Hasting e o amostrador de Gibbs, que são apresentados a seguir.

O algoritmo de Metropolis-Hastings é um termo geral para uma família de métodos de simulação *MCMC* que são úteis para amostragem a partir de distribuições posteriores (mais detalhes em Gelman et al., 2014).

De acordo com Robert & George (2010) o algoritmo de Metropolis-Hastings associado à densidade objetiva (alvo)  $f$  e à densidade condicional  $q$  produz uma cadeia de Markov  $(X^{(t)})$  através do seguinte núcleo de transição:

Dado  $x^{(t)}$ ,

1. Gere  $Y_t \sim q(y|x^{(t)})$ .

---

<sup>3</sup>bayesian inference using Gibbs sampling (Inferência bayesiana usando o amostrador de Gibbs)

2. Tome

$$X^{(t+1)} = \begin{cases} Y_t, & \text{com probabilidade } \rho(x^{(t)}, Y_t), \\ x^{(t)}, & \text{com probabilidade } 1 - \rho(x^{(t)}, Y_t), \end{cases}$$

em que

$$\rho(x, y) = \min \left\{ \frac{f(y) q(x|y)}{f(x) q(y|x)}, 1 \right\}.$$

De acordo com Paulino et al. (2003), o amostrador de Gibbs é um caso especial de um algoritmo Metropolis-Hastings. Ele é útil principalmente em situações em que a priori é multidimensional.

Seja  $\boldsymbol{\theta}^0 = (\theta_1^{(0)}, \dots, \theta_k^{(0)})$  um valor arbitrário inicial para o vetor  $\boldsymbol{\theta}$ .

Procede-se iterativamente da seguinte forma:

1. Obtém-se  $\theta_1^{(1)}$  de  $h(\theta_1|x, \theta_2^{(0)}, \dots, \theta_k^{(0)})$ ,

Obtém-se  $\theta_2^{(1)}$  de  $h(\theta_2|x, \theta_1^{(1)}, \theta_3^{(0)}, \dots, \theta_k^{(0)})$ ,

Obtém-se  $\theta_3^{(1)}$  de  $h(\theta_3|x, \theta_1^{(1)}, \theta_2^{(1)}, \theta_4^{(0)}, \dots, \theta_k^{(0)})$ ,

...

Obtém-se  $\theta_k^{(1)}$  de  $h(\theta_k|x, \theta_1^{(1)}, \dots, \theta_{k-1}^{(1)})$ .

Completa-se assim uma iteração do esquema e uma transição de  $\boldsymbol{\theta}^{(0)}$  para  $\boldsymbol{\theta}^{(1)} = (\theta_1^{(1)}, \dots, \theta_k^{(1)})$ .

2. O esquema anterior é repetido com  $\boldsymbol{\theta}^{(1)}$  anteriormente obtido, como vetor inicial, para obter um novo vetor de  $\boldsymbol{\theta}^{(2)}$  e assim haver uma transição de  $\boldsymbol{\theta}^{(1)}$  para  $\boldsymbol{\theta}^{(2)} = (\theta_1^{(2)}, \dots, \theta_k^{(2)})$ .



3. Itera-se  $t$  vezes este ciclo de geração de observações aleatórias de cada uma das distribuições condicionais completas, produzindo assim  $\boldsymbol{\theta}^{(0)}, \dots, \boldsymbol{\theta}^{(t)}$ .

Quando  $t \rightarrow \infty$ ,  $(\theta_1^{(t)}, \dots, \theta_k^{(t)})$  tende em distribuição para um vetor aleatório cuja função densidade de probabilidade conjunta é  $f(\boldsymbol{\theta}|x)$ . Em particular,  $\theta_i^{(t)}$  tende em distribuição para uma quantidade aleatória cuja densidade é  $f(\theta_i|x)$  (densidade marginal superior a posteriori de  $\theta_i$ ) e

$$\frac{1}{t} \sum_{i=1}^t g(\boldsymbol{\theta}^{(i)}) \rightarrow E[g(\boldsymbol{\theta})|x], \quad (14)$$

quase certamente para qualquer função  $g(\cdot)$ , em que  $E[g(\boldsymbol{\theta})|x]$  representa o valor esperado de  $g(\boldsymbol{\theta})$  em relação à distribuição a posteriori de  $f(x|\boldsymbol{\theta})$ .

### 2.3.2 Monte Carlo Hamiltoniano (*HMC*)

O *HMC* é computacionalmente muito mais dispendioso que o Metropolis ou o amostrador de Gibbs. Mas suas propostas são tipicamente muito mais eficientes. Como resultado, não são necessárias tantas amostras para descrever a distribuição posteriori. E à medida que os modelos se tornam mais complexos - milhares ou dezenas de milhares de parâmetros - o *HMC* pode realmente ofuscar outros algoritmos (McElreath, 2015).

Uma ineficiência inerente ao amostrador de Gibbs e ao algoritmo Metropolis é seu comportamento de passeio aleatório. As simulações podem levar um longo tempo indo e vindo e se movendo ao longo da distribuição alvo. Reparametrização e regras de salto eficientes podem melhorar a situação, mas para modelos complicados esse comportamento de passeio aleatório local permanece, especialmente para distribuições alvo de alta dimensão (Gelman et al., 2014).

Segundo Gelman et al. (2014), o *HMC* utiliza uma ideia da física para suprimir o comportamento de passeio aleatório local no algoritmo Metropolis, permitindo assim que ele se mova muito mais rapidamente por meio da distribuição alvo.

Para cada componente  $\theta_j$  no espaço paramétrico, o *HMC* adiciona uma variável “impulso”  $\phi_j$ . Ambos  $\theta$  e  $\phi$  são então atualizados juntos em um novo algoritmo Metropolis, no qual a distribuição de salto para  $\theta_j$  é determinada em grande parte por  $\phi$ . Cada iteração do *HMC* prossegue através de várias etapas, durante as quais a posição e o momento evoluem com base em regras que imitam o comportamento da posição. Os passos podem se mover rapidamente onde possível, através do espaço de  $\theta$ , e até mesmo transformar cantos no espaço paramétrico para preservar o total de “energia” da trajetória. O *HMC* também é chamado de Monte Carlo híbrido porque combina *MCMC* e métodos de simulação determinística.

No *HMC* a densidade a posteriori  $p(\theta|y)$  é aumentada por uma distribuição independente  $p(\phi)$  no momento, definindo então a distribuição conjunta  $p(\theta, \phi|y) = p(\phi)p(\theta|y)$ . Simula-se a partir da distribuição conjunta, mas o interesse é apenas na distribuição de  $\theta$ ; o vetor  $\phi$  é então uma variável auxiliar, introduzido apenas para permitir que o algoritmo se mova mais rapidamente através do espaço paramétrico.

A seguir, apresenta-se um resumo do algoritmo *HMC* de acordo com Gelman et al. (2014):

1. Atualização de  $\phi$  com um sorteio aleatório de sua distribuição posterior - que, conforme especificado, é igual à sua distribuição a priori,  $\phi \sim N(0, M)$ , em que  $M$  é a matriz massa, chamada desta forma por analogia ao modelo Hamiltoniano dinâmico da física.
2. Atualização simultânea de  $(\theta, \phi)$  envolvendo  $L$  “leapfrog”<sup>4</sup> saltos, cada salto dimensionado por um fator  $\epsilon$  (intervalo de tempo ou tamanho do salto). Em cada passo *leapfrog*,  $\theta$  e  $\phi$  são alterados. As  $L$  etapas *leapfrog* são as seguintes:

---

<sup>4</sup>Algoritmo de integração numérica que é especificamente adaptado para fornecer resultados estáveis para sistemas de equações Hamiltonianas (Stan Development Team, 2017).

- a) Uso do gradiente<sup>5</sup> do log da densidade a posteriori de  $\theta$  para fazer um passo intermediário de  $\phi$ :

$$\phi \leftarrow \phi + \frac{1}{2}\epsilon \frac{\partial \log p(\theta|y)}{\partial \theta}.$$

- b) Use o vetor “momento”  $\phi$  para atualizar o vetor “posição”  $\theta$ .

$$\theta \leftarrow \theta + \epsilon M^{-1} \phi.$$

$M$  é a matriz massa, a covariância da distribuição momento  $p(\phi)$ .

- c) Novamente, use o gradiente de  $\theta$  para fazer a atualização intermediária de  $\phi$ :

$$\phi \leftarrow \phi + \frac{1}{2}\epsilon \frac{\partial \log p(\theta|y)}{\partial \theta}.$$

3. Sendo  $\theta^{(t-1)}, \phi^{(t-1)}$  os valores do parâmetro e do vetor momento no início do processo *leapfrog* e  $\theta^*, \phi^*$  o valor depois de  $L$  passos. Na etapa aceita-rejeita, calcula-se:

$$r = \frac{p(\theta^*|y)p(\phi^*)}{p(\theta^{t-1}|y)p(\phi^{t-1})}.$$

4. Conjunto

$$\theta^{(t)} = \begin{cases} \theta^*, & \text{com probabilidade } \min(r, 1) \\ \theta^{(t-1)}, & \text{caso contrário.} \end{cases}$$

### 2.3.3 Algoritmo *NUTS*

A escolha do número de saltos *leapfrog* e o tamanho de cada salto pode ser complexa. Hoffman & Gelman (2014) propuseram o algoritmo *NUTS* (*No-U-Sampler*), uma extensão do *HMC* que escolhe o tamanho do salto automaticamente

---

<sup>5</sup>se  $\theta$  é d-dimensional o gradiente é  $\frac{\partial \log p(\theta|y)}{\partial \theta} = \left( \frac{\partial \log p(\theta|y)}{\partial \theta_1}, \dots, \frac{\partial \log p(\theta|y)}{\partial \theta_d} \right)$

evitando o comportamento de passeio aleatório, problema comum no algoritmo Metropolis e amostrador de Gibbs. Os autores afirmam que, empiricamente, o *NUTS* é tão eficiente quanto o método *HMC* padrão.

De acordo com Hoffman & Gelman (2014), no *HMC* deve-se encontrar um valor para o tamanho do passo  $\epsilon$  que não seja muito pequeno (o que desperdiçaria computação tomando etapas minúsculas desnecessariamente) nem muito grande (o que prejudicaria o cálculo, causando altas taxas de rejeição). Uma abordagem padrão é ajustar  $\epsilon$  de modo que a probabilidade média de aceitação de Metropolis do *HMC* seja igual a algum valor  $\delta$ . Para *HMC*, define-se um critério  $h^{HMC}(\epsilon)$  tal que

$$H_t^{HMC} \equiv \min \left\{ 1, \frac{p(\tilde{\theta}^t, \tilde{r}^t)}{p(\theta^{t-1}, r^{t,0})} \right\}; \quad h^{HMC}(\epsilon) \equiv \mathbb{E}_t[H_t^{HMC}|\epsilon],$$

no qual  $\tilde{\theta}^t$  e  $\tilde{r}^t$  são a posição e o momento propostos na  $t$ -ésima iteração da cadeia de Markov,  $\theta^{t-1}$  e  $r^{t,0}$  são a posição inicial e o impulso (reamostrado) para a iteração da cadeia de Markov,  $H_t^{HMC}$  é a probabilidade de aceitação desta  $t$ -ésima proposta do *HMC* e  $h^{HMC}$  é a probabilidade de aceitação média esperada da cadeia em equilíbrio para um  $\epsilon$  fixo.

Segundo Hoffman & Gelman (2014), como não há uma única etapa de aceitação/rejeição no *NUTS*, devemos definir uma estatística alternativa para a probabilidade de aceitação de Metropolis. Para cada iteração, definimos a estatística  $H_t^{NUTS}$  e sua expectativa quando a cadeia atingiu o equilíbrio como

$$H_t^{NUTS} \equiv \frac{1}{|\mathcal{B}_t^{final}|} \sum_{\theta, r \in \mathcal{B}_t^{final}} \min \left\{ 1, \frac{p(\theta, r)}{p(\theta^{t-1}, r^{t,0})} \right\}; \quad h^{NUTS} \equiv \mathbb{E}_t[H_t^{NUTS}],$$

no qual  $\mathcal{B}_t^{final}$  é o conjunto de todos os estados explorados durante a duplicação final da iteração  $t$  da cadeia de Markov e  $\theta^{t-1}$  e  $r^{t,0}$  são a posição inicial e o impulso (reamostrado) para a iteração da cadeia de Markov.  $H_t^{NUTS}$  pode ser entendido como a probabilidade de aceitação média que o *HMC* daria aos estados de momento de posição explorados durante a iteração final de duplicação.

### 2.3.4 Avaliação da convergência

Um procedimento comum nos métodos de simulação em inferência bayesiana é avaliar a convergência das cadeias geradas. Duas importantes estatísticas - comuns tanto nos métodos *MCMC* quanto nos métodos *HMC* - muito conhecidas na literatura são o  $\hat{R}$  e o  $n_{eff}$ , denominados fator de redução de escala potencial e número efetivo de amostras, respectivamente.

Diversos autores propõem calcular  $\hat{R}$  de maneiras ligeiramente diferentes e implementações nos *softwares* variam, mas todos os  $\hat{R}$  têm em comum o fato de que valores muito maiores que 1 indicam não-convergência, segundo Korner-Nievergelt et al. (2015).

Segundo Gelman et al. (2014), para cada escalar estimado  $\psi$ , rotula-se as simulações como  $\psi_{ij}$  ( $i = 1, \dots, n$ ;  $j = 1, \dots, m$ , sendo  $m$  o número de cadeias e  $n$  o comprimento de cada cadeia) e calcula-se  $B$  e  $W$ , as variações entre e dentro da sequência:

$$B = \frac{n}{m-1} \sum_{j=1}^m (\bar{\psi}_{\cdot j} - \bar{\psi}_{\cdot\cdot})^2,$$

$$\text{com } \bar{\psi}_{\cdot j} = \frac{1}{n} \sum_{i=1}^n \psi_{ij} \text{ e } \bar{\psi}_{\cdot\cdot} = \frac{1}{m} \sum_{j=1}^m \bar{\psi}_{\cdot j}; \quad \text{e}$$

$$W = \frac{1}{m} \sum_{j=1}^m s_j^2,$$

$$\text{com } s_j^2 = \frac{1}{n-1} \sum_{i=1}^n (\psi_{ij} - \bar{\psi}_{\cdot j})^2.$$

Pode-se estimar  $var(\psi|y)$ , a variância a posteriori marginal da estimativa, pela média ponderada de  $W$  e  $B$ ,

$$\widehat{var}^+(\psi|y) = \left( \frac{n-1}{n} \right) W + \frac{1}{n} B. \quad (15)$$

Gelman et al. (2014) comentam que esta quantidade superestima a variância a posteriori marginal assumindo que a distribuição inicial é apropriadamente superdispersa. Entretanto, é não viesada sob estacionariedade (isto é, se a distribuição inicial for igual à distribuição alvo), ou no limite  $n \rightarrow \infty$ . Isso é análogo à estimativa clássica de variância com amostragem por conglomerados. Enquanto isso, para qualquer  $n$  finito, a variância “dentro”  $W$  deve ser uma subestimativa de  $\text{var}(\psi|y)$  porque as sequências individuais não tiveram tempo para variar ao longo de toda a distribuição alvo e, como resultado, terão menos variabilidade; no limite como  $n \rightarrow \infty$ , a expectativa de  $W$  se aproxima de  $\text{var}(\psi|y)$ .

De acordo com Gelman et al. (2014), monitora-se a convergência da simulação iterativa estimando o fator pelo qual a escala da distribuição atual para  $\psi$  pode ser reduzida se as simulações continuarem no limite  $n \rightarrow \infty$ . Esta redução de escala potencial é estimada por

$$\hat{R} = \sqrt{\frac{\widehat{\text{var}}^+(\psi|y)}{W}}, \quad (16)$$

que declina para 1 quando  $n \rightarrow \infty$ . Se a redução de escala potencial é alta, então temos razão para acreditar que prosseguir com simulações adicionais pode melhorar nossa inferência sobre a distribuição alvo da estimativa escalar associada.

Gelman et al. (2014) comentam que é usual calcular o tamanho efetivo da amostra usando a seguinte fórmula assintótica para a variância da média de uma sequência correlacionada:

$$\lim_{n \rightarrow \infty} mn \text{var}(\bar{\psi}..) = \left(1 + 2 \sum_{t=1}^{\infty} \rho_t\right) \text{var}(\psi|y), \quad (17)$$

em que  $\rho_t$  é a autocorrelação da sequência  $\psi$  no *lag*  $t$ . Se a simulação  $n$  fosse independente da cadeia  $m$ , então  $\text{var}(\bar{\psi}..)$  seria simplesmente  $\frac{1}{mn} \text{var}(\psi|y)$  e o tamanho da amostra seria  $mn$ . Na presença de correlação, define-se o tamanho efetivo da

amostra como

$$n_{eff} = \frac{mn}{1 + 2 \sum_{t=1}^{\infty} \rho_t}. \quad (18)$$

Segundo Gelman et al. (2014), quando  $\hat{R}$  é próximo de 1 e  $n_{eff}$  é maior que 10 por cadeia para todas as estimativas escalares de interesse, basta coletar  $mn$  simulações e tratá-las como uma amostra da distribuição alvo.

### 2.3.5 Implementação computacional - rjags e rstan

Segundo Gelman et al. (2014), *BUGS* é um programa desenvolvido por estatísticos do *Medical Research Council* em Cambridge, Inglaterra. De acordo com Plummer (2012), a linguagem *BUGS* atualmente é implementada em três tipos de *software*: *OpenBUGS*, *WinBUGS* e *JAGS*<sup>6</sup>. O *OpenBUGS* e o *WinBUGS* rodam em sistemas operacionais Windows, enquanto o *JAGS* foi construído especificamente para rodar em múltiplas plataformas, incluindo Mac OS e Unix. Embora todos os três programas usem essencialmente a mesma sintaxe, o *OpenBUGS* e o *WinBUGS* são executados em uma interface gráfica elaborada, enquanto o *JAGS* é executado apenas a partir da linha de comando de um shell Unix ou R.

O **rjags** (Plummer, 2018) e o **rstan** (Stan Development Team, 2018) são pacotes do R para ajuste de modelos bayesianos. Os algoritmos *MCMC* são a base do pacote **rjags**, sendo o amostrador de Gibbs o algoritmo implementado. Todos os programas citados anteriormente seguem esta mesma lógica: para simular da distribuição a posteriori, usam métodos *MCMC*.

O *Stan*<sup>7</sup> também é um *software* para ajuste de modelos bayesianos que usa diversas interfaces, como por exemplo, o **rstan**. Sua linguagem é semelhante a *BUGS*, além de similaridades com programação em C. O *software* traduz o modelo

---

<sup>6</sup>*Just another Gibbs Sampler* (Apenas outro amostrador de Gibbs).

<sup>7</sup>Sampling through adaptive neighborhoods (Amostragem através de vizinhanças adaptativas).

escrito em sua linguagem para **C++** antes de compilar e mostrar os resultados para o usuário. Diferentemente dos *softwares* e pacotes citados anteriormente, no **rstan**, para simular amostras da distribuição a posteriori, usa-se o algoritmo *NUTS*.

## 2.4 Polinômios fracionários clássicos

Os Polinômios Fracionários foram introduzidos por Royston & Altman (1994) como uma extensão dos modelos de regressão polinomial. O uso destes modelos é interessante quando a suposição de linearidade entre a variável resposta e as covariáveis não é válida. De acordo com Royston & Altman (1994) a relação entre uma variável resposta e uma ou mais covariáveis é frequentemente curvada e, para representar esta curvatura é comum o uso de polinômios, tipicamente quadráticos. Contudo, polinômios de baixa ordem oferecem formas limitadas e, de alta ordem podem não ajustar bem nos valores extremos das covariáveis. Neste contexto, o uso de *FPs*, em que as potências são restritas a um pequeno conjunto de valores inteiros e não inteiros, são uma alternativa para ajuste de modelos aos dados que apresentam algum tipo curvatura.

Segundo Royston & Sauerbrei (2008), uma transformação *FP1* (*FP* de grau 1) de um argumento positivo  $x > 0$  com potência  $p$  é definida como  $x^p$ , em que  $p$  pertence ao conjunto de potências  $S = \{-2; -1; -0,5; 0; 0,5; 1; 2; 3\}$  proposto por Royston & Altman (1994), com  $x^0 = \log x$ . Um *FP1* é definido por:

$$\varphi_1^*(x; p) = \beta_0 + \beta_1 x^p = \beta_0 + \varphi_1(x; p), \quad (19)$$

em que  $\beta_0$  e  $\beta_1$  são os parâmetros a serem estimados, e  $\varphi_1(x; p) = \beta_1 x^p$ .

Para um *FP2*, ou *FP* de segunda ordem, com vetor de parâmetros  $\beta = (\beta_1, \beta_2)^T$  e de potências  $\mathbf{p} = (p_1, p_2)$  ou para  $p_1 = p_2$  (potência repetida), temos:

$$x^{\mathbf{p}} = x^{(p_1, p_2)} = \begin{cases} (x^{p_1}, x^{p_2}), & \text{para } p_1 \neq p_2 \\ (x^{p_1}, x^{p_1} \log x), & \text{para } p_1 = p_2 \end{cases} \quad (20)$$



e neste caso o modelo *FP2* é escrito da seguinte maneira:

$$\varphi_2^*(x; \mathbf{p}) = \beta_0 + \beta_1 x^{p_1} + \beta_2 x^{p_2} = \beta_0 + x^{\mathbf{p}} \boldsymbol{\beta} = \beta_0 + \varphi_2(x; \mathbf{p}), \quad (21)$$

em que  $\varphi_2(x; \mathbf{p}) = x^{\mathbf{p}} \boldsymbol{\beta}$ .

Royston & Sauerbrei (2008) descrevem a definição geral de um Polinômio Fracionário de grau  $m$  (*FPm*) com potências  $\mathbf{p} = (p_1 \leq \dots \leq p_m)$  pela seguinte relação de recorrência: seja  $h_0(x) = 1$  e  $p_0 = 0$ , para uma covariável temos então

$$\varphi_m^*(x; \mathbf{p}) = \beta_0 + \varphi_m(x; \mathbf{p}) = \sum_{j=0}^m \beta_j h_j(x), \quad (22)$$

tal que

$$h_j(x) = \begin{cases} x^{p_j}, & \text{se } p_j \neq p_{j-1} \\ h_{j-1}(x) \log x, & \text{se } p_j = p_{j-1} \end{cases}, \quad (23)$$

para  $j = 1, 2, \dots, m$ . Por exemplo, para  $m = 2$  e  $\mathbf{p} = (-1, 2)$ , temos que  $h_1(x) = x^{-1}$ ,  $h_2(x) = x^2$ . Para  $\mathbf{p} = (2, 2)$ ,  $h_1(x) = x^2$ ,  $h_2(x) = x^2 \log x$ .

Na modelagem múltipla, o uso de *FP3* ou funções de ordem superior raramente é sensato, uma vez que a instabilidade é muito maior. Alguns dos *softwares* que possuem a técnica *FP*, estão restritos às análises com funções *FP1* ou *FP2*. Vale a pena considerar as famílias *FP1* e *FP2* detalhadamente, porque, modelos com grau superior a 2 são raramente necessários na análise múltipla. Polinômios fracionários com  $m \leq 2$  oferecem muitas melhorias potenciais no ajuste em comparação com polinômios convencionais do mesmo grau (Royston & Sauerbrei, 2008).

Para ilustrar graficamente, foi feito o ajuste de quatro modelos de uma pesquisa sobre gordura corporal disponível em Royston & Sauerbrei (2008) - dados IMC. A partir da Figura 1, pode-se notar diferenças razoáveis entre estes. Entretanto, o melhor ajuste foi o *FP1* com potência  $p = -1$ , para o qual o preditor de  $E(Y)$  é  $\hat{\beta}_0 + \hat{\beta}_1 x^{-1}$ . Este modelo, comparado aos demais, apresenta maior coeficiente de

determinação ( $R^2 \approx 0,86$ ) bem como, menor deviance  $D$ , que é definido como

$$D = -2 \times \log L(\boldsymbol{\theta}; \mathbf{x}),$$

em que  $L(\boldsymbol{\theta}; \mathbf{x})$  é a função verossimilhança da amostra dada por:  $L(\boldsymbol{\theta}; \mathbf{x}) = f(x_1; \boldsymbol{\theta}) \times f(x_2; \boldsymbol{\theta}) \times \dots \times f(x_n; \boldsymbol{\theta}) = \prod_{i=1}^n f(x_i; \boldsymbol{\theta})$ , e  $\boldsymbol{\theta}$  o vetor de todos os parâmetros.

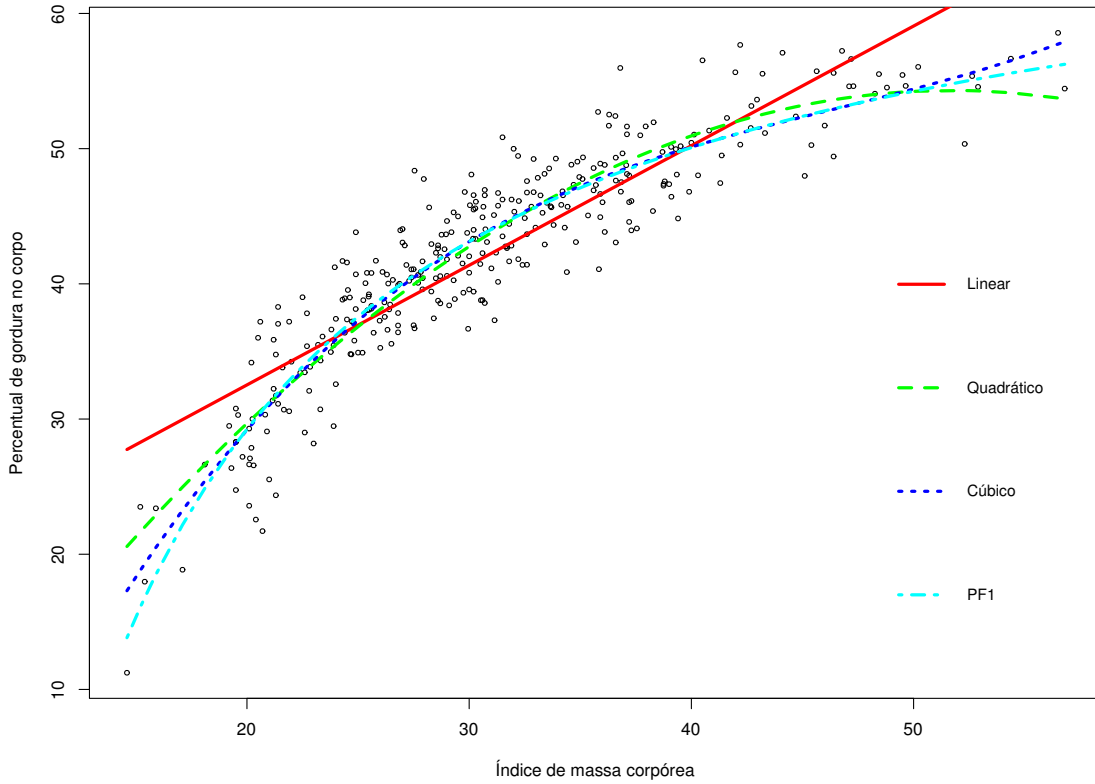


Figura 1: Ajuste de quatro modelos para os dados da pesquisa sobre gordura corporal em função do IMC (índice de massa corporal) disponível em Royston & Sauerbrei (2008).

### 2.4.1 Ajuste do modelo

Os modelos *FPs* são ajustados utilizando-se o método de máxima verossimilhança. Para qualquer potência  $\mathbf{p}$ , encontrar o estimador de máxima verossimilhança equivale a encontrar o  $\boldsymbol{\beta}$  que maximiza a função de verossimilhança de modelos com preditor linear  $\beta_0 + x^{\mathbf{p}}\boldsymbol{\beta}$ . Para uma dada classe (*FP1*, *FP2*, etc.) isto é feito para cada possível  $\mathbf{p}$  com potência em  $S$ , o melhor modelo ajustado é aquele com maior verossimilhança. Para um *FP1*, oito modelos são ajustados. Para um *FP2* trinta e seis modelos são examinados.

De acordo com Royston & Altman (1994), para um determinado  $m$ , o melhor vetor de potências  $\hat{\mathbf{p}} = (\hat{p}_1, \hat{p}_2, \dots, \hat{p}_m)$  está associado à maior verossimilhança, ou equivalentemente, a menor deviance  $D$ . Então  $\hat{\mathbf{p}}$  pode ser considerado como o estimador de máxima verossimilhança de  $\mathbf{p}$  restrito ao conjunto  $S$ . Suponha que cada elemento  $m$  de  $\mathbf{p}$  possa variar continuamente, ao invés de ser restrito ao conjunto  $S$ . Então o *FP*  $\beta_0 + \varphi_m(x; \mathbf{p})$  é um modelo não linear com  $2m + 1$  parâmetros  $p_1, p_2, \dots, p_m; \beta_0, \beta_1, \dots, \beta_m$ . Seja  $D(m, \mathbf{p})$  a deviance de um *FPm* com potência  $\mathbf{p}$  e seja  $D(0)$  a deviance do modelo nulo só com o intercepto  $\beta_0$ . Seja  $\hat{\mathbf{p}}$  o estimador de máxima verossimilhança de  $\mathbf{p}$ . Sob a hipótese nula que  $\boldsymbol{\beta} = \mathbf{0}$ , a distribuição de  $D(0) - D(m, \hat{\mathbf{p}})$  é aproximadamente  $\chi^2$  com  $2m$  graus de liberdade.

Seja  $\mathbf{p}_{FP}$  o *MLE* (*maximum likelihood estimation* - estimador de máxima verossimilhança) de  $\mathbf{p}$  com potências restritas em  $S$ . Então  $D(m, \hat{\mathbf{p}}) \leq D(m, \mathbf{p}_{FP})$ . A diferença de deviance (ou modelo  $\chi^2$  ou estatística da razão de verossimilhança)  $\Delta D_{FP} = D(0) - D(m, \mathbf{p}_{FP})$  é também aproximadamente distribuída com  $\chi^2$  com  $2m$  graus de liberdade (Royston & Sauerbrei, 2008).

### 2.4.2 Escolha do modelo *FP*

Royston & Sauerbrei (2008) estabeleceram o procedimento para seleção da função *FP* (*FSP* - *function select procedure*) para escolha do melhor modelo *FP*

como segue:

- 1) Teste o melhor *FP2* para  $x$  com nível de significância  $\alpha$  contra o modelo nulo usando quatro graus de liberdade (dois coeficientes de regressão e duas potências). Se o teste não for significativo, pare, concluindo que o efeito de  $x$  é não significativo ao nível  $\alpha$ . Caso contrário, continue.
- 2) Teste o melhor *FP2* para  $x$  ao nível de significância  $\alpha$  contra o modelo linear usando três graus de liberdade. Se o teste não for significativo, pare, concluindo que o modelo final é uma reta. Caso contrário, continue.
- 3) Teste o melhor *FP2* para  $x$  ao nível de significância  $\alpha$  contra o melhor *FP1* usando dois graus de liberdade. Se o teste não for significativo, o modelo final é o *FP1*. Caso contrário, o modelo final é o *FP2*. Fim do procedimento.

No passo 1 testa-se a associação global de  $x$  com a saída. No passo 2 examina-se a evidência de não-linearidade. No passo 3 escolhe-se entre um modelo não-linear mais simples ou um mais complexo.

Antes de iniciar o *FSP* o usuário deve escolher o nível de significância e o grau do modelo *FP* mais complexo permitido. É comum a escolha de  $\alpha = 0,05$  e *FP2* ( $m=2$ ) (Royston & Sauerbrei, 2008).

### 2.4.3 Polinômios fracionários multivariado

Para Sauerbrei & Royston (2017), as principais questões da abordagem *FP* multivariada são: seleção de variáveis com eliminação passo atrás (*backward elimination* - *BE*) e seleção de uma função *FP* para modelar a não-linearidade.

Royston & Sauerbrei (2008) definiram um procedimento de *FP* multivariado (*multivariable FP procedure* - *MFP*) combinando *BE* com *FSP*. Dois níveis de significância são necessários:  $\alpha_1^*$  para seleção de variáveis com *BE*, e  $\alpha_2^*$  para comparar o ajuste de funções dentro do *FSP*. A notação  $MFP(\alpha_1^*, \alpha_2^*)$  denota a seleção de

um modelo usando  $\alpha_1^*$  para *BE* e  $\alpha_2^*$  para *FSP*, com  $BE(\alpha_1^*)$  e  $FSP(\alpha_2^*)$  representando aspectos individuais. O caso comum  $\alpha_1^* = \alpha_2^* = \alpha^*$  é abreviado como  $MFP(\alpha^*)$ .

Nos passos do  $MFP(\alpha_1^*, \alpha_2^*)$ , que serão descritos a seguir, todos os testes de significância são da estatística da razão de verossimilhança ( $\chi^2$ ). No *software* R (R Core Team, 2018), o pacote **mfp** foi implementado com este algoritmo.

1. Valores  $P$  nominais  $\alpha_1^*$  e  $\alpha_2^*$  são escolhidos. Valores típicos são  $\alpha_1^* = \alpha_2^* = 0,05$ . Os valores podem diferir entre as variáveis. Tomando  $\alpha_1^* = 1$  para uma determinada variável, “força-a” a entrar no modelo (sem seleção de variável). Tomando  $\alpha_2^* = 1$  para uma variável contínua, força a função *FP* permitida mais complexa a ser ajustada para ela (sem seleção de função).
2. O máximo de graus de liberdade para a função *FP* são escolhidos; por exemplo, quatro, dois e um significa funções *FP2*, *FP1* e linear, respectivamente. O padrão sugerido é quatro graus de liberdade.
3. O modelo linear completo é ajustado. A “ordem de visita” dos preditores é determinada de acordo com o valor  $P$  para omitir cada preditor do modelo. O preditor mais significativo é visitado primeiro e o menos por último. Suponha que as variáveis  $x_1, x_2, \dots, x_k$  foram organizadas nesta ordem, que é retida em todos os ciclos do procedimento.
4. Seja  $c = 0$ , para inicializar o contador de ciclos.
5. Seja  $j = 1$ , para inicializar o contador de variáveis dentro de cada ciclo.
6. Se  $x_j$  é contínuo, vá para o passo 7. Caso contrário,  $x_j$  é categórico ou binário. A significância global da(s) sua(s) variável(s) *dummy* (variáveis indicadoras) é testada no nível  $\alpha_1^*$ . Todas as outras variáveis atualmente no modelo estão incluídas como termos de ajuste. Se  $x_j$  é significativo, é retido, caso contrário, descartado. Vá para o passo 8.
7. O passo 1 do *FSP* (Seção 2.1.2) é aplicado a  $x_j$  ao nível  $\alpha^* = \alpha_1^*$ . Se  $x_j$  não

é significativo, ele é excluído. Caso contrário, os passos 2 e 3 do *FSP* são aplicados ao nível  $\alpha^* = \alpha_1^*$ , para escolher uma função *FP* ou linear. Todas as outras variáveis atualmente no modelo são incluídas como termos de ajuste. No caso de um *FP* com potências  $\mathbf{p}$  sendo escolhido,  $x_j$  é representado por variáveis transformadas  $x_j^{\mathbf{p}}$  em etapas subsequentes em que outras variáveis são consideradas.

8. A inclusão ou eliminação de  $x_j$  se aplica até que  $x_j$  seja reconsiderado no próximo ciclo.
9. Seja  $j = j + 1$ . Se  $j \leq k$ , retorne ao passo 6 para processar o próximo preditor. Caso contrário, continue na etapa 10.
10. Seja  $c = c + 1$ . O  $c$ -ésimo ciclo é completo. Se  $c > c_{\text{máx}}$ , pare (um valor prático de  $c_{\text{máx}} = 5$ ). Caso contrário, verifique se as variáveis incluídas e as transformações *FP* foram alteradas do ciclo  $c - 1$  para o ciclo  $c$ . Nesse caso, volte ao passo 5 para iniciar um novo ciclo. Caso contrário, pare e relate as estimativas atuais do modelo. Fim do procedimento

Normalmente, o *MFP* requer dois, três ou ocasionalmente quatro ciclos para convergência. Falta de convergência envolve oscilação entre dois ou mais modelos, e é extremamente raro (Royston & Sauerbrei, 2008).

#### 2.4.4 Escala e centralização

Um aspecto importante no ajuste de um modelo *FP* é que os valores transformados de  $x^{\mathbf{p}}$  devem ser positivos. Segundo Royston & Sauerbrei (2008), escalonar e centralizar estes valores reduz as chances de imprecisões ou dificuldades no ajuste do modelo. A centralização assegura que o intercepto  $\beta_0$  retenha uma interpretação relevante. A seguinte transformação escalar está implementado no

pacote `mfp` do R:

$$\begin{aligned}\text{lrange} &= \log_{10}[\max(x) - \min(x)] \\ \text{scale} &= 10^{\text{signal}(\text{lrange})\text{int}(|\text{lrange}|)} \\ x^* &= \frac{x}{\text{scale}},\end{aligned}$$

em que `lrange` é o log da amplitude dos dados, `signal(lrange)` é o sinal do `lrange` e `int(|lrange|)` é o inteiro mais próximo abaixo de  $x$ . A quantidade `scale` é uma potência inteira de 10. A transformação *FP* de  $x^*$  é então centrada na média dos valores observados de  $x^*$ .

## 2.5 Polinômios fracionários bayesianos

O paradigma bayesiano para *FPs* foi implementado por Bové & Held (2011) e disponibilizado no pacote `bfp` (<http://r-forge.r-project.org/projects/bfp>) para modelos com erros normalmente distribuídos. A metodologia usa um modelo linear bayesiano com uma priori *hiper-g*, proposta por Liang et al. (2008), e combina seleção de variáveis com modelagem paramétrica de efeitos aditivos. Os autores apresentaram um algoritmo de Monte Carlo via cadeias de Markov (*Markov chain Monte Carlo - MCMC*) para a exploração do modelo. Sob a suposição de que cada covariável  $x_i$  ( $i = 1, 2, \dots, k$ ) tem efeito linear na média da resposta  $y$ , temos formalmente que

$$\eta(\mathbf{x}) := \mathbb{E}(y|\mathbf{x}) = \beta_0 + \sum_{i=1}^k \beta_i x_i, \quad (24)$$

em que  $\mathbf{x} = (x_1, x_2, \dots, x_k)^T$ . Tal formulação pode conduzir para inferências incorretas se a verdadeira relação é distante de ser linear para certo  $x_i$ . Uma generalização imediata que retém o efeito aditivo é substituir  $\beta_i x_i$  por  $f_i(x_i)$  na equação (24), isto é,

$$\eta(\mathbf{x}) = \beta_0 + \sum_{i=1}^k f_i(x_i). \quad (25)$$

Um  $FP$  de grau  $m$  com potências  $p_1 \leq \dots \leq p_m$  e com respectivos coeficientes  $\beta_1, \dots, \beta_m$  é dado por  $f^m(x_i; \boldsymbol{\beta}, \mathbf{p}) = \sum_{j=1}^m \beta_j h_j(x_i)$ . Ao usar  $f^m(x_i; \boldsymbol{\beta}, \mathbf{p})$  no lugar da função  $f_i$ , o modelo em (25) resulta na mesma forma da equação (24):

$$\eta(\mathbf{x}) = \beta_0 + \sum_{i=1}^k f^{m_i}(x_i; \boldsymbol{\beta}_i, \mathbf{p}_i) = \beta_0 + \sum_{i=1}^k \sum_{j=1}^{m_i} \beta_{ij} h_{ij}(x_i), \quad (26)$$

em que  $h_{ij}(x_i)$  representa a relação de recorrência dada pela equação (23) da  $j$ -ésima transformação  $FP$ , da  $i$ -ésima covariável  $x_i$ ;  $\beta_{ij}$  é o parâmetro da  $i$ -ésima covariável  $x_i$ , do  $j$ -ésimo  $FP$ ;  $\mathbf{p}_i$  é o vetor de potências da  $i$ -ésima covariável e  $\beta_0$  é o intercepto.

### 2.5.1 Definição do modelo

Segundo Bové & Held (2011), considere o modelo linear com intercepto

$$\mathbf{y} = \beta_0 + \mathbf{B}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad (27)$$

em que a matriz  $\mathbf{B} = (B_i(\mathbf{x}_j))_{ji}$  ( $n \times p$ ) com índices linha  $j = 1, \dots, n$  e coluna  $i = 1, \dots, p$  é uma função do vetor  $\mathbf{x}_j$  dado pelos valores das variáveis explicativas referentes à  $j$ -ésima observação; e  $\mathbf{y}$ ,  $\boldsymbol{\varepsilon}$  e  $\boldsymbol{\beta}$  são vetores com tamanho  $n$ ,  $n$  e  $p$ , respectivamente.

As funções básicas  $B_i(\mathbf{x}_j)$  são escolhidas como as transformações  $h_{ij}$  em (26), referentes ao vetor de parâmetros  $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_k)^T$ , em que  $\boldsymbol{\beta}_i = (\beta_{i1}, \dots, \beta_{im_i})$ , com  $i = 1, \dots, p$ , e a abordagem de  $FP$  foi incorporada na estrutura do modelo linear. As transformações  $h_{ij}$  são determinadas pelo vetor de potências  $\mathbf{p}_1, \dots, \mathbf{p}_k$  a partir da definição (23), de modo que cada  $FP$  múltiplo possa ser representado pelo vetor  $\boldsymbol{\theta} = (\mathbf{p}_1, \dots, \mathbf{p}_k)$  de tuplas ordenadas, com  $\mathbf{p}_i = (p_{i1} \leq p_{i2} \leq \dots \leq p_{im_i})$ .

Sob a suposição de que  $\boldsymbol{\varepsilon} \sim \mathbf{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n)$ , em que  $\mathbf{I}_n$  denota a matriz identidade de dimensão  $n$ ,  $\mathbf{y}$  também segue uma distribuição normal multivariada com a mesma matriz de covariância de  $\boldsymbol{\varepsilon}$  e vetor de médias  $\boldsymbol{\mu} = \mathbf{1}_n \beta_0 + \mathbf{B}\boldsymbol{\beta}$ , que



determina a função de verossimilhança  $f(\mathcal{D}|\boldsymbol{\beta}, \beta_0, \sigma^2)$ , em que  $\mathcal{D} = \{y_j, \mathbf{x}_j\}_{j=1}^n$  denota os dados observados.

O espaço paramétrico  $\Theta$  do modelo contém todos os  $\boldsymbol{\theta}$  que cumprem a restrição do conjunto dado em  $S$ . Note que  $\boldsymbol{\theta}$  é de dimensão variável  $p_{\boldsymbol{\theta}} := \sum_{i=1}^k m_i$ . Para o modelo nulo  $p_{\boldsymbol{\theta}} = 0$ , porque a  $i$ -ésima tupla  $\mathbf{p}_i$  é vazia se a covariável  $x_i$  não é inclusa no modelo ( $m_i = 0$ ). As quantidades que dependem do modelo, a partir daqui, serão subscritas com  $\boldsymbol{\theta}$ . As colunas de covariáveis designadas  $\mathbf{B}_{\boldsymbol{\theta}}$  são centralizadas tal que  $\mathbf{1}_n^T \mathbf{B}_{\boldsymbol{\theta}} = \mathbf{0}_{p_{\boldsymbol{\theta}}}^T$ , para garantir que o intercepto  $\beta_0$  seja um parâmetro comum com interpretação idêntica em todos os modelos.

### 2.5.2 Distribuição a priori

Bové & Held (2011) usaram a priori *hyper-g* de Liang et al. (2008), a qual é construída usando uma priori de Jeffreys para a variância  $\sigma^2$ . Condicional em  $\sigma^2$ ,  $g > 0$ , atribui-se uma priori imprópria vaga ao intercepto  $\beta_0$  e uma distribuição normal multivariada com vetor médio nulo e matriz de covariância  $\sigma^2 g (\mathbf{B}_{\boldsymbol{\theta}}^T \mathbf{B}_{\boldsymbol{\theta}})^{-1}$  aos demais coeficientes em  $\boldsymbol{\beta}_{\boldsymbol{\theta}}$

$$f(\sigma^2) \propto (\sigma^2)^{-1};$$

$$f(\beta_0, \boldsymbol{\beta}_{\boldsymbol{\theta}} | \sigma^2, g) \propto (\sigma^2 g)^{-\frac{p_{\boldsymbol{\theta}}}{2}} |\mathbf{B}_{\boldsymbol{\theta}}^T \mathbf{B}_{\boldsymbol{\theta}}|^{\frac{1}{2}} \times \exp \left\{ -\frac{1}{2\sigma^2 g} \|\mathbf{B}_{\boldsymbol{\theta}} \boldsymbol{\beta}_{\boldsymbol{\theta}}\|^2 \right\}.$$

O fator de covariância  $g > 0$  é assumido independente de  $\sigma^2$  com densidade a priori  $f(g) = \frac{a-2}{2}(1+g)^{-\frac{a}{2}}$ , em que  $a \in (3, 4]$  garante que a média a posteriori  $\mathbb{E}(g|\boldsymbol{\theta}, \mathcal{D})$  é finita para qualquer modelo dado  $\boldsymbol{\theta}$ . Além disso, a distribuição a priori implícita em  $t = g/(g+1)$  desloca o vetor  $\boldsymbol{\mu}$  em direção ao intercepto  $\beta_0$ .

Para  $n \geq p_{\boldsymbol{\theta}} + 3$ , o fator de Bayes de um modelo com  $R_{\boldsymbol{\theta}}^2 \rightarrow 1$  versus o modelo nulo pode crescer em paralelo sem restrição, em que  $R_{\boldsymbol{\theta}}^2$  é o coeficiente de determinação estimado por *OLS* com componentes  $\hat{\beta}_o^{OLS} = \bar{y}$  e  $\hat{\boldsymbol{\beta}}_{\boldsymbol{\theta}}^{OLS} =$

$(\mathbf{B}_\theta^T \mathbf{B}_\theta)^{-1} \mathbf{B}_\theta^T \mathbf{y}$ . Além disso, Bové & Held (2011) afirmam que, se o verdadeiro modelo não é o nulo, a máxima densidade a posteriori (*MAP*; ver anexo 2) é um modelo consistente para o verdadeiro modelo quando  $n \rightarrow \infty$ . Esta priori também produz uma ponderação bayesiana de modelos (*BMA*; ver anexo 2) consistente para predição de novas respostas.

Prioris independentes para as transformações *FP* podem ser especificadas assumindo  $f(\boldsymbol{\theta}) = f(\mathbf{p}_1) \times \dots \times f(\mathbf{p}_k)$ . Então o modelo a priori pode ser formulado como

$$f(\mathbf{p}_i) = (p_{i1}, \dots, p_{im_i} | m_i) f(m_i) = d(m_i)^{-1} (m_{max} + 1)^{-1}, \quad (28)$$

em que  $d(m_i)$  é o número de diferentes *FP*s, de grau  $m_i$  e  $m_{max}$  é o grau máximo do *FP*.

### 2.5.3 Distribuição a posteriori

A densidade a posteriori dos parâmetros  $\beta_0$ ,  $\boldsymbol{\beta}_\theta$  e  $\sigma^2$  para um modelo específico  $\boldsymbol{\theta}$  e fator de covariância  $g$  é

$$f(\beta_0, \boldsymbol{\beta}_\theta, \sigma^2 | \mathcal{D}, g) \propto (\sigma^2)^{-\left(\frac{n+p_\theta}{2} + 1\right)} (g)^{-\frac{p_\theta}{2}} \times \exp \left\{ -\frac{1}{2\sigma^2} \left[ \|\mathbf{y} - \boldsymbol{\mu}\|^2 + \frac{1}{g} \|\mathbf{B}_\theta \boldsymbol{\beta}_\theta\|^2 \right] \right\}.$$

Pode-se mostrar que o núcleo da distribuição acima pertence à distribuição normal gama inversa (ver Denison et al., 2002), de modo que

$$\beta_0, \boldsymbol{\beta}_\theta, \sigma^2 | \mathcal{D}, g \sim N_{p_\theta+1} GI \left( \mathbf{m}_\theta, \mathbf{V}_\theta, \frac{n-1}{2}, c_\theta \right), \text{ no qual}$$

$$\mathbf{V}_\theta = \begin{pmatrix} n^{-1} & \mathbf{0}_{p_\theta}^T \\ \mathbf{0}_{p_\theta} & \frac{g}{g+1} (\mathbf{B}_\theta^T \mathbf{B}_\theta)^{-1} \end{pmatrix}, \quad (29)$$

$$\mathbf{Z}_\theta = (\mathbf{1}_n, \mathbf{B}_\theta), \quad (30)$$

$$\mathbf{m}_\theta = \mathbf{V}_\theta \mathbf{Z}_\theta^T \mathbf{y} = \begin{pmatrix} \bar{y} \\ \frac{g}{g+1} \hat{\boldsymbol{\beta}}_\theta^{OLS} \end{pmatrix} \text{ e} \quad (31)$$

$$c_{\boldsymbol{\theta}} = \mathbf{y}^T [\mathbf{I}_n - \mathbf{Z}_{\boldsymbol{\theta}} \mathbf{V}_{\boldsymbol{\theta}} \mathbf{Z}_{\boldsymbol{\theta}}^T] \mathbf{y} / 2. \quad (32)$$

A densidade marginal a posteriori do fator de encolhimento  $t = g/(g+1)$  é  $f(t|\mathcal{D}) \propto (1-t)^{(p_{\boldsymbol{\theta}}+a-2)/2-1} (1-R_{\boldsymbol{\theta}}^2 t)^{-(n-1)/2}$ . Fazendo a transformação de variáveis  $u = (1-R_{\boldsymbol{\theta}}^2)/(1-R_{\boldsymbol{\theta}}^2 t)$  temos que  $\tilde{F}_{\boldsymbol{\theta}}(q) \propto \int_0^q f(t|\mathcal{D}) dt \propto \mathcal{B}_{\boldsymbol{\theta}} \left( \frac{1-R_{\boldsymbol{\theta}}^2}{1-R_{\boldsymbol{\theta}}^2 q} \right) - \mathcal{B}_{\boldsymbol{\theta}}(1-R_{\boldsymbol{\theta}}^2)$ , em que  $\mathcal{B}_{\boldsymbol{\theta}}$  é a função de distribuição acumulada de uma variável aleatória Beta com parâmetros de forma  $(n-p_{\boldsymbol{\theta}}-a+1)/2$  e  $(p_{\boldsymbol{\theta}}+a-2)/2$ . A normalização da função de distribuição do fator de encolhimento é  $\tilde{F}_{\boldsymbol{\theta}}(1) = 1 - \mathcal{B}_{\boldsymbol{\theta}}(1-R_{\boldsymbol{\theta}}^2)$ , produzindo a posteriori  $F_{\boldsymbol{\theta}(q)} = \tilde{F}_{\boldsymbol{\theta}}(q)/\tilde{F}_{\boldsymbol{\theta}}(1)$  e a inversa da função de distribuição acumulada pode ser encontrada por

$$F_{\boldsymbol{\theta}}^{-1}(p) = \left( 1 - \frac{1-R_{\boldsymbol{\theta}}^2}{\mathcal{B}_{\boldsymbol{\theta}}^{-1}(p + (1-p)\mathcal{B}_{\boldsymbol{\theta}}(1-R_{\boldsymbol{\theta}}^2))} \right) / R_{\boldsymbol{\theta}}^2. \quad (33)$$

#### 2.5.4 Inferência

Inferências no espaço paramétrico  $\boldsymbol{\Theta}$  de todos os possíveis modelos  $\boldsymbol{\theta}$  se baseiam na distribuição a posteriori

$$f(\boldsymbol{\theta}|\mathcal{D}) = \frac{f(\mathcal{D}|\boldsymbol{\theta})f(\boldsymbol{\theta})}{f(\mathcal{D})}. \quad (34)$$

Bové & Held (2011) comentam que a *hyper-g* é conveniente porque conduz a uma forma fechada da função de verossimilhança marginal  $f(\boldsymbol{\theta}|\mathcal{D})$  de um modelo  $\boldsymbol{\theta}$ . De acordo com Liang et al. (2008) temos que

$$f(\mathcal{D}|\boldsymbol{\theta}) = \frac{f(\mathcal{D}|g, \boldsymbol{\theta})f(g)}{f(g|\mathcal{D}, \boldsymbol{\theta})} \propto \frac{{}_2F_1\left(\frac{n-1}{2}; 1; \frac{p_{\boldsymbol{\theta}}+a}{2}; R_{\boldsymbol{\theta}}^2\right)}{p_{\boldsymbol{\theta}}+a-2}, \quad (35)$$

no qual  ${}_2F_1(a; b; c; z) = \frac{\Gamma(c)}{\Gamma(b)\Gamma(c-b)} \int_0^1 \frac{t^{b-1}(1-t)^{c-b-1}}{(1-tz)^a} dt$  é a função hipergeométrica gaussiana.

O método *MCMC* adaptado para o problema é apresentado na seção 3.1 de Bové & Held (2011). Uma vez obtida a distribuição a posteriori, a seleção do modelo é realizada em três etapas hierárquicas:

1. Extraia um modelo da distribuição a posteriori estimada  $\hat{f}(\boldsymbol{\theta}|\mathcal{D})$ ;
2. Tome amostras do fator de encolhimento  $t = g/(1 + g)$  a partir de  $f(t|\mathcal{D}, \boldsymbol{\theta})$  usando (33);
3. Obtenha amostras do intercepto  $\beta_0$  e do vetor de coeficientes  $\boldsymbol{\beta}_\theta$  a partir da distribuição t de Student

$$\beta_0|\mathcal{D}, \boldsymbol{\theta}, g \sim t\left(\bar{y}, \frac{2c_\theta}{n(n-1)}, n-1\right),$$

e

$$\boldsymbol{\beta}_\theta|\mathcal{D}, \boldsymbol{\theta}, g \sim t_{p_\theta}\left(\frac{g}{g+1}\hat{\boldsymbol{\beta}}_\theta^{OLS}, \frac{2c_\theta g}{(n-1)(g+1)}(\mathbf{B}_\theta^T \mathbf{B}_\theta)^{-1}, n-1\right).$$

Amostras de combinações lineares de  $\boldsymbol{\beta}_\theta$ , especialmente pontos  $f_i(x_i)$  das curvas do *FP*, são facilmente obtidas no último passo.

### 3 METODOLOGIA

O desenvolvimento metodológico nesta pesquisa envolveu duas fases. A primeira constituiu nas aplicações das propostas de Royston & Sauerbrei (2008) e co-autores, no contexto de inferência clássica, e de Bové & Held (2011), no contexto de inferência bayesiana. Os detalhes dos passos seguidos nesta fase estão descritos na Seção 3.1.

A segunda fase abordou a especificação de um modelo Bayesiano, com transformações *FP*, relaxando a restrição do conjunto discreto para os valores das potências utilizado tanto por Royston & Sauerbrei (2008) quanto por Bové & Held (2011). Nesta versão, o modelo se encaixa na classe de regressão não linear, no qual o suporte para cada potência é um intervalo real. Uma modificação na forma do *FP2* é utilizada visando uma simplificação em termos do número de parâmetros. Os detalhes são apresentados na Seção 3.2.

#### 3.1 Um breve estudo para investigar o desempenho da proposta bayesiana de Bové & Held (2011)

Conforme apresentado na seção 2.3, os *FPs* foram estudados apenas para o conjunto  $S$  proposto por Royston & Altman (1994). Na abordagem bayesiana de Bové & Held (2011), foi usada a priori *hiper-g* por apresentar diversas propriedades interessantes, como por exemplo, garantir uma posteriori fechada para a verossimilhança marginal do modelo. Para cada potência  $p$ , foi assumida uma distribuição

uniforme discreta, para que assim, cada potência do conjunto  $S$  tenha a mesma chance de ser escolhida.

Usando conjuntos de dados utilizados pelos autores, um univariado e dois multivariados, inicialmente investigou-se os ajustes para as duas metodologias. O conjunto univariado (*Research Body Fat* de Luke et al. (1997) - dados IMC) e um dos multivariados (*Prostrate Cancer* de Stamey et al. (1989) - dados *PSA* (*Prostate-Specific Antigens*)) foram utilizados por Royston & Sauerbrei (2008), disponíveis em: <http://biom131.imbi.uni-freiburg.de/biom/Royston-Sauerbrei-book/#datasets>. O segundo multivariado (dados sobre concentração atmosférica de ozônio - dados ozônio) foi utilizado por Bové & Held (2011), disponível em: <https://github.com/cran/bfp/tree/master/data>.

A intenção inicial foi adquirir experiência nos ajustes dos modelos e destacar possíveis similaridades e diferenças entre os ajustes, do ponto de vista prático. Os três conjuntos de dados foram então analisados ajustando-se os modelos via os pacotes `mfp` e `bfp`.

Os resultados desses estudos conduziram à busca sobre desempenho dos métodos na estimação do grau do polinômio. Similarmente a um estudo de simulação executado por Royston & Sauerbrei (2008) para avaliar essa propriedade de seu método, realizou-se um breve estudo comparando-se também o método de Bové & Held (2011).

## 3.2 Uma proposta de modelos *FP* Bayesianos

Royston & Sauerbrei (2008) afirmam que as estimativas são mais fáceis de serem obtidas para um *FP*, com um número limitado de potências. Por exemplo, para um *FP1*, apenas oito modelos são necessários para encontrar a melhor potência, enquanto que para um *FP2*, somente 36. Além disso, os autores comentam que as potências selecionadas em  $S$  podem ser vistas como uma aproximação da melhor

potência possível no intervalo  $[-2, 3]$ .

Todavia, Royston & Sauerbrei (2008) declaram que às vezes é necessário suplementar o conjunto  $S$  para incluir, por exemplo, a potência  $1/3$ , se  $x$  tiver dimensão de volume, por exemplo, e que, em geral, o conjunto  $S$  pode não conter valores suficientes para detectar uma função não-linear, em todas as aplicações. Shkedy et al. (2006), por exemplo, argumentam que o conjunto de potências  $S$  não oferece flexibilidade suficiente para seu modelo.

Na tentativa de contribuir para a ampliação da flexibilidade de curvaturas disponíveis, neste trabalho, propõem-se um modelo não linear em que as potências dos polinômios não se restringem ao conjunto  $S$ , podendo ser quaisquer valores dentro do intervalo  $[-3, 3]$ , por exemplo, e são estimadas por medidas resumo da distribuição a posteriori por meio dos métodos de simulação apresentados no Capítulo 2. A proposta tem particular potencial no caso de poucas regressoras. Para muitas regressoras, provavelmente a saída é mesmo a discretização, devido ao problemas de convergência associados aos modelos não lineares.

Mesmo utilizando o espaço discreto, o processo de estimação das potências nos polinômios fracionários pode falhar por diversos motivos, dentre eles, tamanho da amostra pequeno, presença de valores discrepantes nos dados, presença de multicolinearidade, presença de valores influentes. Para cada variável regressora, mesmo limitando-se ao  $FP2$ , tem-se dois parâmetros extras para estimar, indicando que tamanhos de amostras modestos e poucos níveis distintos nas variáveis regressoras podem ser problemáticos. Considerando a variável regressora  $x_j$ , um modelo parcimonioso, que acomoda curvas flexíveis é o modelo de transformação para  $x_j$  dado por

$$\varphi(x_j) = \beta_j x_j^{\theta_j} + \beta_{jj} x_j^{2\theta_j}, \quad j = 1, 2, \dots, k, \quad (36)$$

com  $x_j^{\theta_j} = \log x_j$  se  $\theta_j = 0$ , de forma que para a regressora  $x_j$  apenas uma potência é estimada. Na descrição que segue, por simplicidade de notação, considera-se apenas uma regressora e o índice  $j$  será removido.

Seja  $\theta$  uma variável aleatória que representa a potência a ser estimada do modelo de regressão simples

$$y = \beta_0 + \varphi(x), \quad (37)$$

com  $\varphi(x)$  dado em (36). Comparado ao *FP2* usual (ver expressão (21)) esse modelo é mais parcimonioso por envolver estimação de uma única potência e também oferece um grande variedade de curvas úteis na prática.

A partir deste modelo pode-se construir uma regressão bayesiana, utilizando-se priores para  $\theta$ . O conjunto  $S$ , suporte para a potência  $\theta$ , pode ser substituído por um intervalo assumindo-se, por exemplo,  $\theta$  com uma distribuição a priori uniforme em  $[-3, 3]$ , ou alguma outra apropriada.

Para estimar os coeficientes deste modelo, bem como a potência  $\theta$ , é necessário encontrar a distribuição a posteriori. Todavia, não há solução analítica para tais distribuições, sendo necessário partir para soluções por meio dos métodos de simulação como os abordados no Capítulo 2.

Em estudos preliminares do modelo em (37), observou-se que atribuindo-se prioris não informativas para os coeficientes de regressão e, uma distribuição  $U_c(-3, 3)$  ou uma distribuição  $N(0, 3)$  para a potência  $\theta$ , o ajuste não converge.

Gelman & Hill (2006) declaram que no processo de simulação, se os valores iniciais dos parâmetros forem próximos de  $10^{-4}$  ou  $10^6$ , os algoritmos comuns, Metropolis e Gibbs podem falhar. Desta forma, buscou-se não dar valores iniciais das cadeias semelhantes a estes. Entretanto, o problema de não convergência permaneceu, mesmo aumentando consideravelmente o número de iterações.

Uma alternativa para resolver este problema é a reparametrização do modelo. Segundo Bates & Watts (1988), quando o espaço de valores é restrito, digamos

$$a \leq \theta \leq b,$$



pode-se utilizar a seguinte transformação logística:

$$\theta = a + \frac{b - a}{1 + e^{-\Phi}}, \quad (38)$$

tal que o modelo na forma reparametrizada depende de  $\Phi$  que pode variar continuamente nos reais.

Desta forma o modelo em (37) passa a ser escrito como:

$$y = \beta_0 + \beta_1 x \left( a + \frac{b - a}{1 + e^{-\Phi}} \right) + \beta_2 x^2 \left( a + \frac{b - a}{1 + e^{-\Phi}} \right). \quad (39)$$

A Figura 2 apresenta os valores da transformação logística de  $\theta$  para diferentes valores de  $\Phi$ . Nela pode-se verificar que, quando  $\Phi \rightarrow \infty$  ou  $\Phi \rightarrow -\infty$  (equação (38)) fica garantido que  $\theta \in (a, b)$ . No caso de seguir as recomendações de Royston & Altman (1994) para os valores da potência os limites são  $a = -3$  e  $b = 3$ , porém agora podendo assumir qualquer valor nesse intervalo.

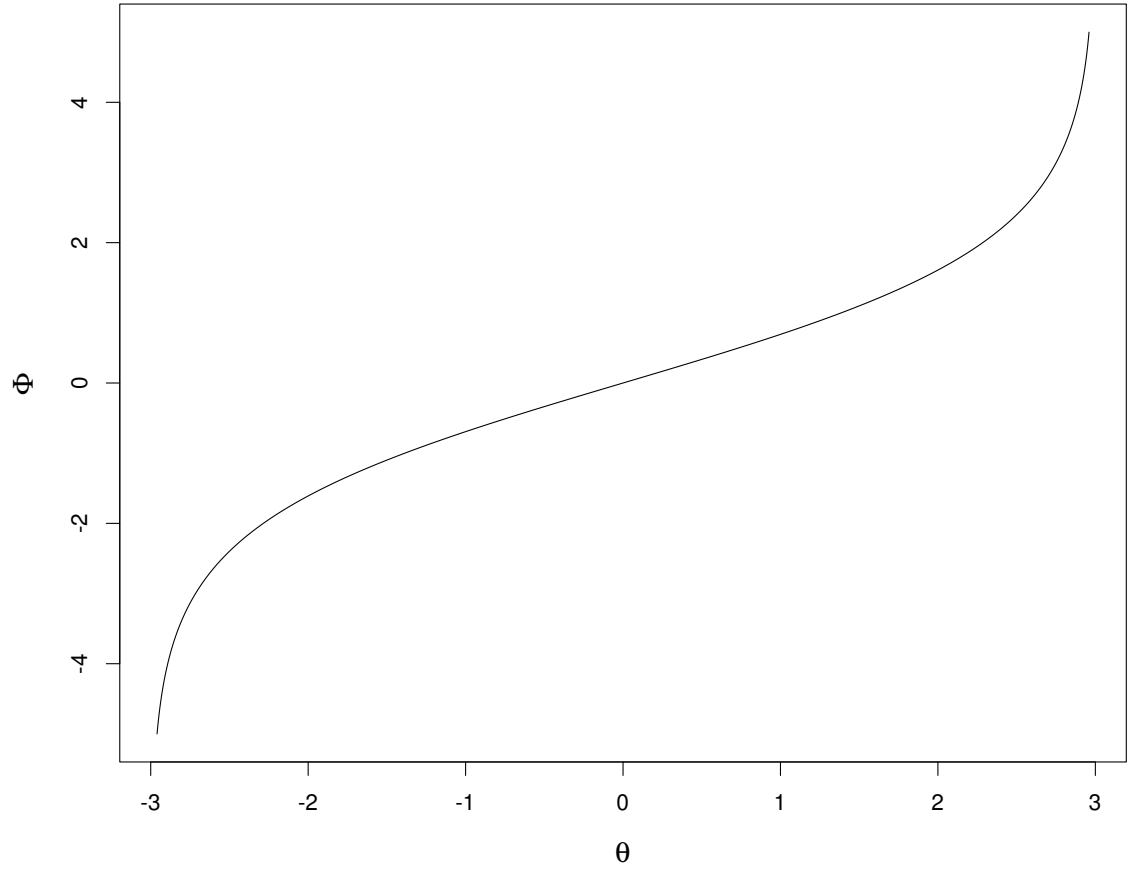


Figura 2: Valores da transformação logística de  $\theta$  segundo equação (38) para  $-3 \leq \theta \leq 3$ .

Alguns estudos com dados simulados indicaram convergência do *MCMC* sob o modelo reparametrizado. Entretanto, a convergência, dependendo dos valores iniciais para as cadeias, não foi sempre garantida. Como foi exposto no Capítulo 2, os métodos *MCMC* podem apresentar comportamento de passeio aleatório. Aumentar o número de iterações e reparametrização de parâmetros, que seriam as soluções para estes problemas, não foi o suficiente. Desta forma, buscou-se uma alternativa aos métodos *MCMC* e, encontrou-se uma possibilidade nos métodos *HMC*.

Utilizando-se os métodos *HMC*, mais especificamente o algoritmo *NUTS*, para estimação dos parâmetros do modelo proposto reparametrizado, a convergência foi alcançada em diversos cenários. Todas as estimativas foram obtidas a partir dos

pacotes `rjags` e `rstan`.

No `rstan`, o qual tem implementado o *HMC*, o modelo é escrito em blocos, basicamente três: *data*, *parameters* e *model*. No bloco *data* declara-se os dados que serão utilizados para análise. No bloco *parameters* são declarados os parâmetros que são estimados via *HMC* ou *NUTS*. No caso deste trabalho, utilizou-se o *NUTS*, visto que, conforme seção 2.3.3, este algoritmo evita o comportamento de passeio aleatório. Finalmente, no bloco *model* o modelo bayesiano é descrito com as especificações das respectivas priors e verossimilhança.

Além desta formulação básica, como o modelo proposto recebeu uma reparametrização, acrescenta-se o bloco *transformed parameters* em que são declaradas as devidas transformações do modelo. No Quadro 1 temos o código para o modelo proposto de ordem 1 (idem à expressão em (39) porém sem o segundo termo) e no Quadro 2 o código para o modelo de ordem 2.

**Quadro 1:** Descrição do código para o modelo *FP1* proposto, no *rstan*

```

#write("
data {
  int <lower=1> N;
  vector[N] x;
  vector[N] y;
}
parameters {
  real beta0;
  real beta1;
  real <lower=0> sigma;
  real phi;
}
transformed parameters {
  real <lower=-3,upper=3> theta;
  theta = -3 + ((6)/(1+exp(-phi)));
}
model {
  beta0 ~ normal(0, 1000);
  beta1 ~ normal(0, 1000);
  phi ~ normal(0, 100);
  sigma ~ cauchy (0 , 2.5);
  for (n in 1:N)
  y[n]~normal(beta0 + pow(x[n],theta) * beta1, sigma);
}
generated quantities {
  real y_rep[N];

```

```
for (n in 1:N) {  
  y_rep[n] = normal_rng(beta0 + pow(x[n],theta) * beta1, sigma);  
}  
}",  
#"mod_0.stan")
```

Observe que, além dos blocos citados anteriormente, temos também o bloco *generated quantities*, em que declara-se os comandos para obter a distribuição preditiva a posteriori usada para avaliar o ajuste da distribuição a posteriori do modelo bayesiano, denotada por  $y_{rep}$  (Gelman et al., 2014).

**Quadro 2:** Descrição do código para o modelo *FP2* proposto, no *rstan*

```

#write("
data {
  int <lower=1> N;
  vector[N] x;
  vector[N] y;
}
parameters {
  real beta0;
  real beta1;
  real beta11;
  real <lower=0> sigma;
  real phi;
}
transformed parameters {
  real <lower=-3,upper=3> theta;
  theta = -3 + ((6)/(1+exp(-phi)));
}
model {
  beta0 ~ normal(0, 200);
  beta1 ~ normal(0, 10);
  beta11 ~ normal(0, 10);
  phi ~ normal(0, 100);
  sigma ~ cauchy(0, 2.5);
  for (n in 1:N)

```

```

y[n]~normal(beta0 + pow(x[n],theta) * beta1 +
pow(x[n],2*theta) * beta11, sigma);
}
generated quantities {
real y_rep[N];
for (n in 1:N) {
y_rep[n] = normal_rng(beta0 + pow(x[n],theta) * beta1 +
pow(x[n],2*theta) * beta11, sigma);
}
}";
#"mod_0_1.stan")
mod_0_1 = "mod_0_1.stan"

```

Para avaliar a metodologia proposta no ajuste destas versões dos modelos, inicialmente, simulou-se nove conjuntos de dados, sendo 7 de curvas de primeiro grau e 2 de segundo grau. Especificou-se uma combinação de valores dos parâmetros dos modelos de forma a reproduzir padrões curvilíneos de apelo prático, ou seja, que não raro, aparecem em dados reais. Para os ajustes do modelo de primeira ordem, utilizou-se prioris normais vagas para os coeficientes de regressão e para  $\Phi$ . Quanto à distribuição a priori para  $\sigma$ , para ambos os modelos, seguindo as recomendações de Gelman et al. (2008), utilizou-se a distribuição half-Cauchy. Já para o modelo de ordem dois, foi necessário utilizar prioris não tão vagas para os coeficientes de regressão, para garantir a convergência do modelo.

Por fim, dois conjuntos de dados foram reanalisados seguindo essa proposta de modelagem. Um deles é o conjunto dados IMC já descrito. O outro é oriundo de um estudo para modelar o peso de aranhas em função do comprimento do corpo (Stropa & Trinca, 2005 - dados aranha). Neste conjunto, devido a indicação

de heterogeneidade de variâncias, o modelo bayesiano proposto inclui também a possibilidade de transformação de Box-Cox (Box & Cox, 1964).



## 4 RESULTADOS

Este capítulo está dividido em duas partes. Na primeira parte foi feita a exploração das técnicas clássica e bayesiana para *FPs*, comparando-as a partir de medidas de acurácia utilizadas em Bové & Held (2011), a partir de dados disponíveis em Royston & Sauerbrei (2008) e Bové & Held (2011). Posteriormente, fez-se um estudo preliminar de simulação para avaliar a percentagem de acerto no grau do *FP* para o caso univariado. Royston & Sauerbrei (2008) fizeram avaliações dos ajustes via técnicas clássicas que são, portanto, complementadas para os modelos bayesianos.

A segunda parte deste capítulo trata da aplicação e estudo do modelo proposto na seção 3.2, através de simulação de casos e modelagem de conjuntos de dados reais.

### 4.1 Exploração dos *FPs* sob a abordagem bayesiana

Utilizaram-se as abordagens clássica e bayesiana para ajustar modelos aos dois conjuntos de dados disponíveis em Royston & Sauerbrei (2008) e um disponível em Bové & Held (2011), comparando-se os resultados das abordagens clássica e bayesiana. Na metodologia bayesiana, Bové & Held (2011) separam em torno de 10% dos dados para usar como dados de teste para fazer previsões e, a partir da raiz do erro quadrático médio de predição (*RMSPE*; ver anexo 2), fazem a comparação dos modelos. Na abordagem clássica, as estimativas dos coeficientes do *FP*, bem como das potências, são feitas por *MLE*, enquanto que, na abordagem bayesiana, estas estimativas são a moda da distribuição a posteriori e, a potência escolhida é

aquela que apresenta o maior *MAP*. Os resultados são apresentados nas subseções 4.1.1, 4.1.2 e 4.1.3. No anexo 1 encontra-se a descrição das variáveis dos conjuntos de dados multivariados apresentados neste capítulo, e no anexo 2 encontram-se as fórmulas para os cálculos do *MAP*, *BMA* e *RMSPE*, que são usados na seleção de covariáveis e na seleção de modelos.

Na subseção 4.1.4 são apresentados os resultados das simulações para avaliar o desempenho dos métodos no acerto do grau do *FP*. Todos os ajustes dos modelos mostrados a seguir foram feitos a partir dos pacotes `mfp` e `bfp` disponíveis no *software* R (R Core Team, 2018).

#### 4.1.1 Aplicações para o conjunto de dados IMC (Royston & Sauerbrei, 2008)

O primeiro conjunto de dados, com 327 observações, refere-se ao estudo da relação entre percentual de gordura no corpo (*pbfm*) e índice de massa corporal (*bmi*) em amostras de negros de três países, Nigéria, Jamaica e USA (Royston & Sauerbrei, 2008), ilustrando o caso univariado (única regressora). Esse conjunto de dados já foi analisado pela técnica *FP* clássica pelos autores, mas para realizar uma comparação com a proposta de Bové & Held (2011), os dados são reanalisados usando estratégias de separar os dados em amostras de treino e teste, com o objetivo de explorar e entender melhor a técnica bayesiana.

A Tabela 1 apresenta as estimativas dos coeficientes do modelo *FP1* ajustado, assim como o coeficiente de determinação e o *RMSPE*. Nota-se que as estimativas foram muito próximas nos três modelos, sendo que o intercepto dos modelos bayesianos *MAP* e *BMA* foi mais próximo da média da covariável *bmi* ( $\approx 30,94$ ), já que, no modelo bayesiano, todas as covariáveis são centralizadas na média, após transformação, conforme destacado sobre as colunas  $\mathbf{B}_\theta$  (equação (27)) da subseção 2.5.1. O melhor ajuste em ambas abordagens foi o *FP1* com potência  $p = -1$ , sendo que o *MAP* apresentou o menor *RMSPE*. Então, o modelo ajustado é da forma

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x^{-1},$$

em que  $\hat{\beta}_0$  e  $\hat{\beta}_1$  são as estimativas da moda da distribuição a posteriori, ou de *MLE*, conforme o método. Deve-se destacar que tanto o pacote **mfp** quanto **bfp** usam o fator de escala em  $x$ , se apropriado numericamente (subseção 2.4.4). No caso deste exemplo,  $x = bmi/10$ .

Tabela 1: Estimativas dos parâmetros e medidas de qualidade dos modelos ajustados - dados IMC (Royston & Sauerbrei, 2008)

Parâmetros	<i>MFP</i>	<i>MAP</i>	<i>BMA</i>
$\beta_0$	71,0627	42,3152	42,3152
$\beta_1$	-83,7001	-83,5651	-83,5651
$R^2$	0,8650	0,8650	0,8650
<i>RMSPE</i>	3,2764	<b>3,2753</b>	3,2829

A Figura 3 apresenta os valores preditos para os dados IMC para a amostra teste. Pode-se verificar que os ajustes estão muito próximos, o que justifica os erros de predição tão semelhantes entre os três modelos.

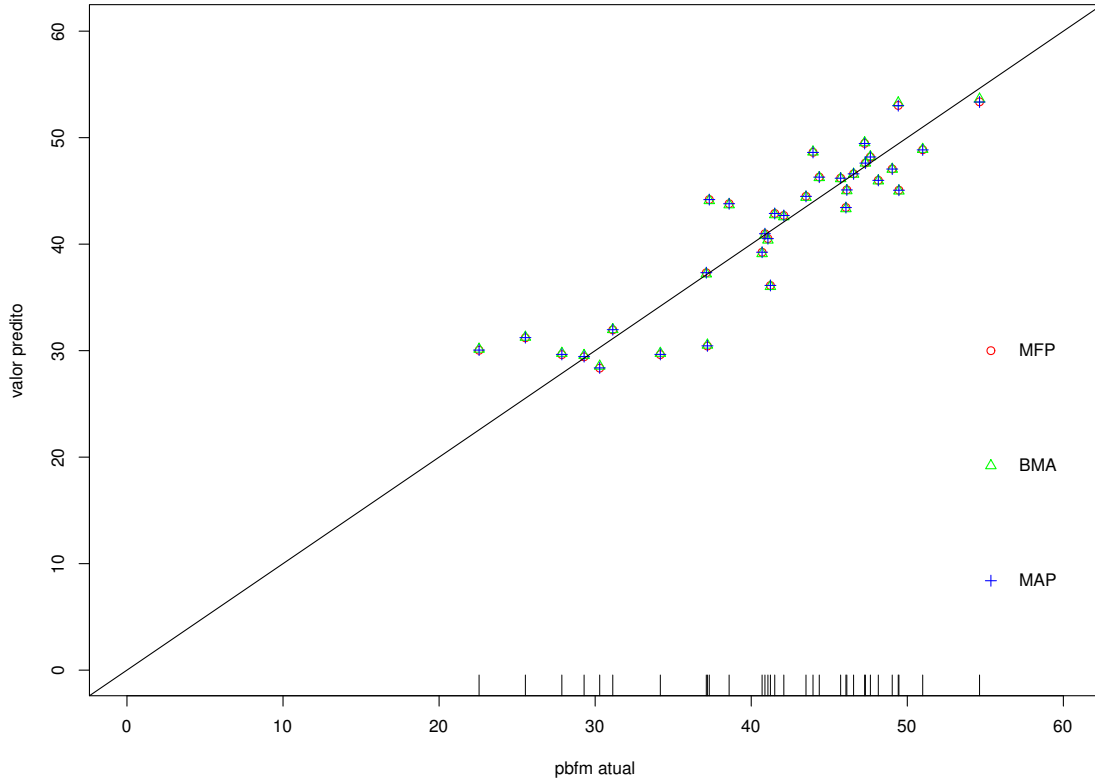


Figura 3: Relação entre valores preditos e a resposta observada para os modelos aplicados aos dados IMC, sob diferentes metodologias.

#### 4.1.2 Aplicações para o conjunto de dados *PSA* (Royston & Sauerbrei, 2008)

Para o caso múltiplo, utilizaram-se os dados de câncer de próstata, com 97 observações, disponível em Royston & Sauerbrei (2008), em que a variável resposta é o log da concentração de *PSA*. Neste conjunto de dados, somente a abordagem clássica havia sido utilizada.

A Tabela 2 apresenta as potências selecionadas e as estimativas dos coeficientes do modelo *FP2* ajustado, assim como o coeficiente de determinação e o *RMSPE*. Neste problema um pouco mais complexo envolvendo 7 regressoras, nota-

se mais diferenças entre os resultados. Enquanto a abordagem clássica indica a importância de 4 regressoras, a bayesiana seleciona 6. O grau da transformação coincide para 3 das regressoras, embora as estimativas das potências se diferenciam.

Apesar de o modelo bayesiano ter apresentado coeficiente de determinação um pouco maior ( $R^2 \approx 0,73$  contra  $R^2 \approx 0,70$  do clássico), o ajuste *MFP* apresentou menor erro de predição, talvez pelo fato de, neste caso, resultar em um modelo mais parcimonioso. Portanto o modelo clássico selecionado foi

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_2 x_2 + \hat{\beta}_{41} x_4^{-0,5} + \hat{\beta}_{42} x_4^3 + \hat{\beta}_{51} x_5 + \hat{\beta}_{71} \log x_7 + \hat{\beta}_{72} x_7^2,$$

em que, por exemplo, os coeficientes  $\hat{\beta}_{41}$  e  $\hat{\beta}_{42}$  são as estimativas de máxima verossimilhança do efeito da covariável  $x_4$  para as potências  $\hat{p}_1 = -0,5$  e  $\hat{p}_2 = 3$ , respectivamente.

Tabela 2: Estimativas dos parâmetros e medidas de qualidade dos modelos ajustados, sob diferentes abordagens - dados *PSA* (Royston & Sauerbrei, 2008)

Preditor	<i>MFP</i>				<i>MAP</i>			
	$p_1$	$p_2$	Coef. 1	Coef. 2	$p_1$	$p_2$	Coef. 1	Coef. 2
$x_1$	-	-	-	-	-0,5	-	-0,3387	-
$x_2^*$	-	-	0,6495	-	-	-	1,6519	-
$x_3$	-	-	-	-	0	-	2,4707	-
$x_4$	-0,5	3	-0,3833	0,0772	-0,5	2	0,0534	-0,0243
$x_5$	1	-	1,6228	-	0,5	-	0,6693	-
$x_6$	-	-	-	-	-	-	-	-
$x_7$	0	2	0,0451	-0,2137	2	3	0,0013	0,1342
Intercepto			2,6594				2,4050	
$R^2$			0,7010				0,7268	
<i>RMSPE</i>			<b>0,6381</b>				1,7324 (1,5579**)	

\* $x_2$  é uma variável binária. \*\* *RMSPE* do *BMA*.

A Figura 4 apresenta os valores preditos para os dados de câncer de cada modelo e mostra que as predições do *MAP* e *BMA* estão, na maioria dos casos, mais distantes dos valores observados do que as predições do *MFP*, explicando as diferenças de *RMSPE* entre as três abordagens.

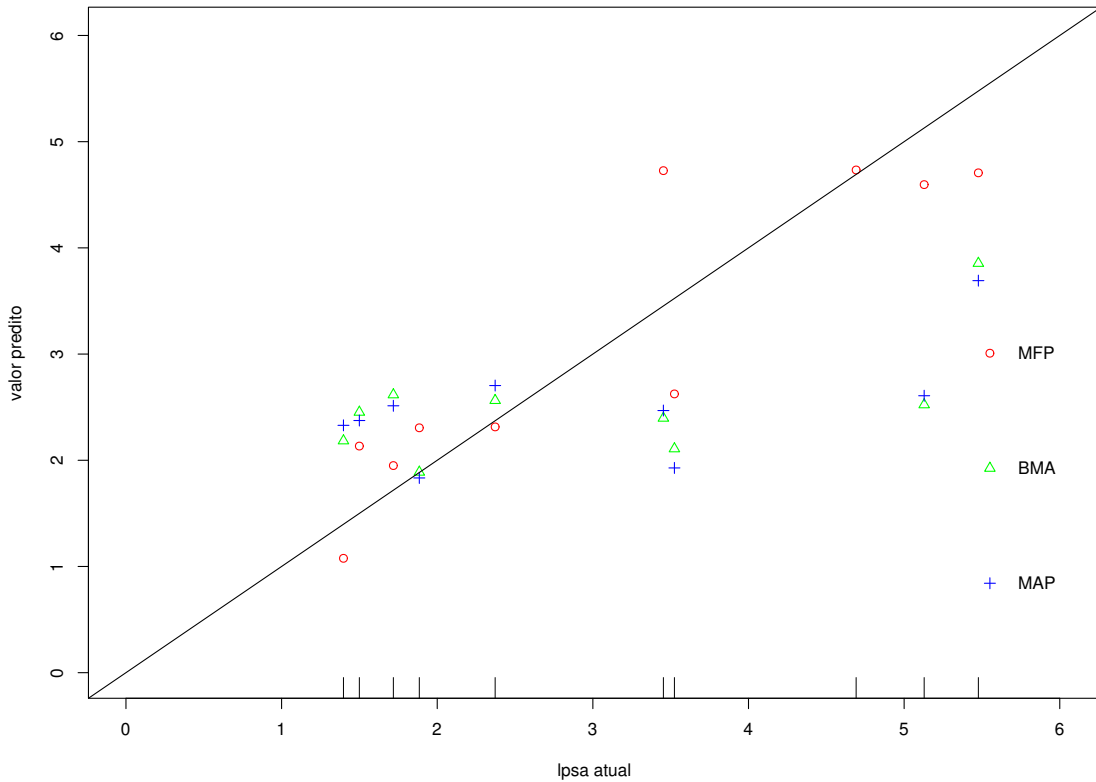


Figura 4: Relação entre os valores preditos e a resposta observada para os modelos aplicados aos dados *PSA*, sob diferentes metodologias

#### 4.1.3 Aplicações para o conjunto de dados ozônio (Bové & Held, 2011)

Os *FPs* Bayesianos foram aplicados por Bové & Held (2011) no conjunto de dados de ozônio com 366 observações, previamente analisado por Breiman & Friedman (1985). A variável resposta é concentração atmosférica de ozônio e existem 9 possíveis variáveis regressoras contínuas. Nesta seção, aplica-se a técnica *FP* clássica

e compara aos resultados prévios de Bové & Held (2011). Algumas variáveis sofreram transformações prévias para corrigir os problemas de escala.

A Tabela 3 apresenta as potências selecionadas e as estimativas dos coeficientes do modelo *FP2* ajustado, assim como o coeficiente de determinação e o *RMSPE*. Neste caso, em ambas abordagens as mesmas covariáveis foram selecionadas, havendo pequenas diferenças nas potências estimadas, bem como nos coeficientes de regressão. Entretanto o modelo *MAP* ajustado apresentou menor erro de predição, cuja equação é

$$\begin{aligned}\hat{y} = & \hat{\beta}_0 + \hat{\beta}_{01}x_0 + \hat{\beta}_{02}x_0 \log x_0 + \hat{\beta}_{61}x_6^{0,5} + \hat{\beta}_{71}x_7^3 + \hat{\beta}_{81} \log x_8 + \hat{\beta}_{82}x_8^{0,5} + \hat{\beta}_{91}x_9^3 \\ & + \hat{\beta}_{10,1}x_{10}^{-0,5} + \hat{\beta}_{10,2}x_{10}^{-0,5} \log x_{10}.\end{aligned}$$

Tabela 3: Estimativas dos parâmetros e medidas de qualidade dos modelos ajustados - dados ozônio (Bové & Held, 2011)

Preditor	<i>MFP</i>				<i>MAP</i>			
	$p_1$	$p_2$	Coef. 1	Coef. 2	$p_1$	$p_2$	Coef. 1	Coef. 2
$x_0$	1	1	9,4878	-7,5323	1	1	9,3022	-7,3588
$x_6$	1	-	5,0327	-	0,5	-	6,8282	-
$x_7$	3	-	22,0825	-	3	-	22,0072	-
$x_8$	0	0,5	4,5140	-8,5111	0	0,5	4,6545	-8,6977
$x_9$	3	-	-2,1614	-	3	-	-2,1024	-
$x_{10}$	-0,5	0	-3,5168	-3,8493	-0,5	-0,5	8,0089	2,0601
Intercepto	11,7308				11,7733			
$R^2$	0,7768				0,7770			
<i>RMSPE</i>	3,7777				<b>3,7327</b> / (3,7396)*			

\* *RMSPE* do *BMA*.

A Figura 5 apresenta os valores preditos para os dados ozônio de cada modelo. Nela pode-se verificar que os ajustes foram bem próximos para os três

modelos.

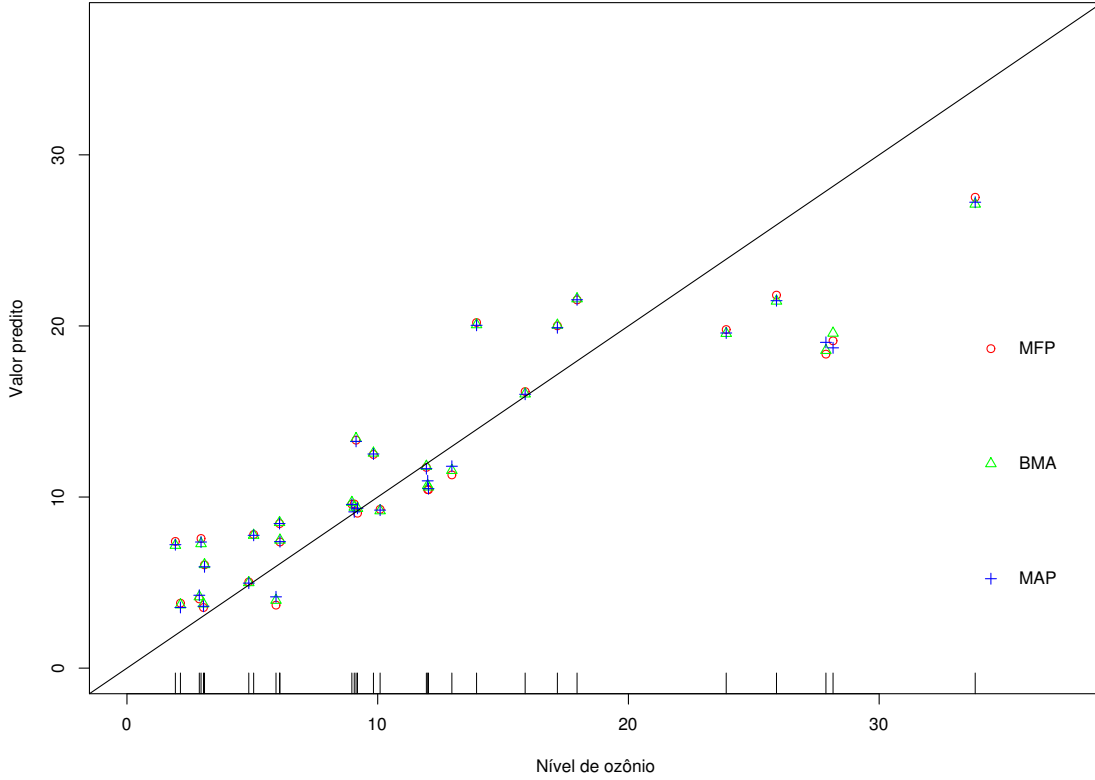


Figura 5: Relação entre os valores preditos e a resposta observada para os modelos aplicados aos dados ozônio, sob diferentes metodologias

#### 4.1.4 Avaliação de percentagem de acertos das abordagens *FP* via simulações

Nesta subseção fez-se um pequeno estudo de simulação para avaliar o comportamento da metodologia bayesiana na identificação do grau do *FP*, para o caso univariado. As funções subjacentes simuladas foram do tipo *FP1* e *FP2* para uma única covariável  $x$ , utilizados por Royston e Saurbrei (2008). Para *FP1* gerou-se  $\log X \sim N(0; 1)$  e  $Y|x \sim N(\beta \log x; 1)$  e para *FP2*,  $X \sim N(4; 1)$  e  $Y|x \sim N(\beta(x - 4)^2; 1)$ . O valor de  $\beta$  foi obtido pela relação  $R^2 = \text{var}(X\beta) / [\sigma^2 + \text{var}(X\beta)] =$



$\beta^2 \text{var}(X) / [\sigma^2 + \beta^2 \text{var}(X)]$ , na qual  $R^2$  é a variabilidade explicada, fixada em 0,1; 0,2 e 0,3. Como  $\sigma^2 = 1$ , temos  $R^2 = \beta^2 / (1 + \beta^2)$  e então  $\beta = [R^2 / (1 + R^2)]^{0,5}$ . Os tamanhos das amostras variam com  $n = 25, 50$  e  $100$ . Em cada caso 1.000 simulações foram efetuadas, sendo que os dados de treino foram alterados 5 vezes, totalizando assim 5.000 simulações. Aplicou-se, então, as duas abordagens de ajuste do *FP* com grau máximo permitido de 2.

A Tabela 4 apresenta os resultados de ajustes pelo método clássico e Bayesiano das mil simulações de um *FP1*. Nela pode-se verificar que, mesmo quando a correlação entre as variáveis é fraca, ou seja,  $\beta$  pequeno e tamanho de amostra pequeno, a proporção de acerto da verdadeira função subjacente é alta, acima de 84%, sendo que a abordagem clássica apresenta melhores resultados em todos os cenários estudados. Uma possível explicação para esses resultados é a tendência da abordagem bayesiana de Bové & Held (2011) resultar em modelos mais complicados, já sinalizada nas aplicações.

Tabela 4: Porcentagem de acertos no grau do  $FP$  segundo a abordagem de ajuste de modelo quando a função subjacente é um  $FP1$  (resultados de 1.000 simulações)

			Abordagem clássica			Abordagem bayesiana		
			Número de potências estimadas			Número de potências estimadas		
$R^2$	$\beta$	$n$	0	1	2	0	1	2
0,1	0,33	25	4,4	94,5	1,1	0,1	84,2	15,7
		50	0,1	98,4	1,5	0,0	84,3	14,7
		100	0,0	97,4	2,3	0,0	88,6	11,4
0,2	0,50	25	0,0	99,5	0,5	0,0	87,6	12,4
		50	0,0	97,9	2,1	0,0	88,7	11,3
		100	0,0	97,6	2,4	0,0	91,7	8,3
0,3	0,65	25	0,0	99,1	0,9	0,0	90,5	9,5
		50	0,0	98,3	1,7	0,0	89,9	10,1
		100	0,0	97,4	2,6	0,0	93,2	6,8

A Tabela 5 apresenta os resultados dos desempenhos dos ajustes pelos métodos estudados em mil simulações de um  $FP2$ . Neste caso a abordagem bayesiana apresenta melhores resultados, mesmo quando temos correlação e coeficiente baixos, bem como o tamanho da amostra pequeno, a proporção de acerto é acima de 80%, exceto quando  $R^2 = 0,1$ ,  $\beta = 0,33$  e  $n = 25$ , em que a proporção fica em torno de 56%.

Tabela 5: Porcentagem de acertos no grau do  $FP$  segundo a abordagem de ajuste de modelo quando a função subjacente é um  $FP2$  (resultados de 5.000 simulações)

			Abordagem Clássica			Abordagem bayesiana		
			Número de potências			Número de potências		
			Estimadas			Estimadas		
$R^2$	$\beta$	$n$	0	1	2	0	1	2
0,1	0,33	25	83,3	10,4	6,3	23,8	19,9	56,3
		50	50,6	16,4	33,0	8,5	11,2	80,3
		100	14,1	9,4	76,5	1,1	2,2	96,7
0,2	0,5	25	62,4	21,2	16,4	11,2	15,8	73,0
		50	16,9	15,3	67,8	1,5	3,3	95,2
		100	1,9	1,9	96,7	0,1	0,2	99,7
0,3	0,65	25	41,2	24,8	34,0	5,3	10,9	83,9
		50	5,3	8,1	86,6	0,4	1,3	98,3
		100	0,0	0,6	99,4	0,0	0,0	100,0

## 4.2 Avaliação da modelagem $FP$ via regressão não linear bayesiana

Para avaliar a abordagem proposta, na seção 3.2, inicialmente fez-se alguns estudos de casos via simulação de  $FP1$  e de  $FP2$ , agora definidos como na equação (39), para uma covariável. Para ajustar os modelos propostos tentou-se utilizar as técnicas  $HMC$  e  $MCMC$ . Para o caso de primeira ordem obteve-se convergência com ambas as técnicas. Já para os modelos de segunda ordem apresentam-se somente os resultados obtidos via  $HMC$ , visto que não houve convergência do ajuste com o  $MCMC$ . Os resultados são apresentados na seção 4.2.1 para  $FP1$  e 4.2.2 para  $FP2$ .

Aplicações da abordagem proposta foram feitas nos dados IMC e aranha.

O conjunto de dados aranha refere-se aos coletados no Jardim Botânico de Botucatu-SP, IBB, Unesp, conforme referenciado em Stropa & Trinca (2005). Ambas aplicações remetem a uma curva de primeiro grau, sendo que apenas a técnica *HMC* apresentou convergência dos ajustes. Os resultados são apresentados em 4.2.3 e 4.2.4.

#### 4.2.1 Estudo de casos simulados de modelos *FP1*

Para avaliar a abordagem proposta no modelo em (36), de primeira ordem, simulou-se diversas curvaturas com as potências do conjunto  $S$ . Para todas as curvas geradas escolheu-se os valores de coeficientes de regressão de tal maneira que as curvas tivessem padrões semelhantes àqueles que frequentemente aparecem em dados reais conforme os diversos apresentados nos trabalhos de Royston & Sauerbrei (2008) e co-autores.

Em particular os valores usados foram  $\beta_0 = 1,5$  e  $\beta_1 = 1,2$ , com  $x \sim N(2; 0, 5^2)$  e  $n = 100$ . Desta forma, pode-se representar os modelos simulados da seguinte maneira

$$y_i = 1,5 + 1,2x^\theta + \epsilon_i, \quad (40)$$

com  $\epsilon_i \sim N(0; 0,09^2)$ .

A Figura 6 apresenta os dados simulados para um *FP1* com potências  $\theta = -2$ ,  $\theta = -1$  e  $\theta = -0,5$ . Apresenta-se na Tabela 6 as estimativas do modelo segundo a abordagem proposta para os casos simulados. Foram usadas 10.000 iterações no **rjags**, que utiliza os métodos *MCMC*, enquanto que no **rstan**, que utiliza os métodos *HMC*, foram necessárias apenas 1.000 iterações. Em ambas metodologias, utilizando-se prioris não-informativas para todos os parâmetros, as estimativas foram bem semelhantes e próximas dos valores verdadeiros. Além disso, o  $\hat{R}$  foi próximo de 1, caracterizando a convergência do ajuste tanto no *HMC* quanto no *MCMC*, isto pode ser visto também a partir dos traços que estão no anexo 3.

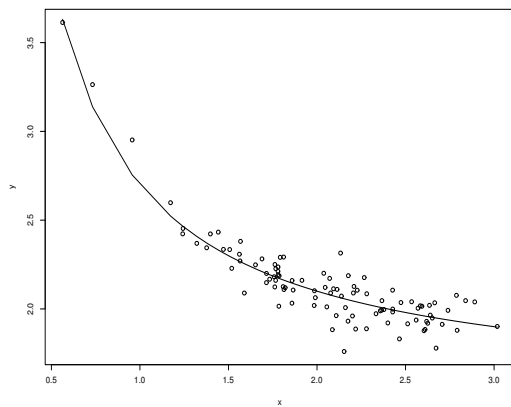
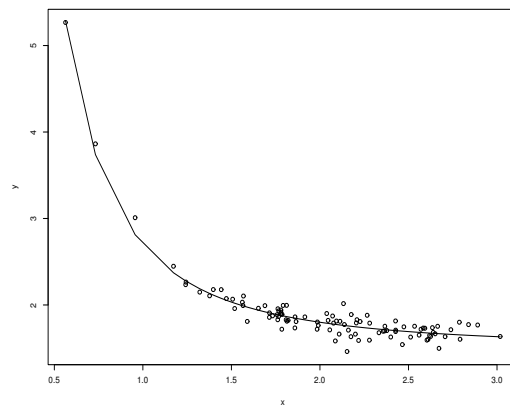
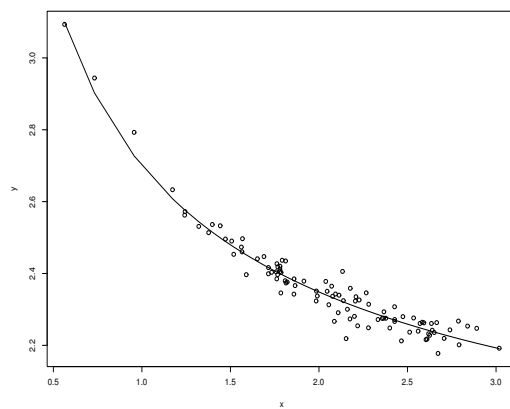
(a)  $\theta = -2$ .(b)  $\theta = -1$ .(c)  $\theta = -0,5$ .

Figura 6: Dados simulados sob modelos *FP1* com diferentes potências.

Tabela 6: Propriedades dos ajustes, segundo a abordagem proposta, para modelos *FP1*, dados simulados com potências diferentes, de acordo com as técnicas *HMC* e *MCMC*

Parâmetros	<i>HMC</i>			<i>MCMC</i>		
	Estimativas*	$\hat{R}$	$n_{eff}$	Estimativas*	$\hat{R}$	$n_{eff}$
$\beta_0 = 1,5$	1,471	1,003	409	1,473	1,001	4700
$\beta_1 = 1,2$	1,266	1,004	414	1,264	1,001	3600
$\phi = -1,609$	-1,541	1,006	425	-1,543	1,001	6800
$\theta = -2$	-1,939	1,006	420	-1,941	1,001	6800
$\beta_0 = 1,5$	1,458	1,001	424	1,467	1,007	800
$\beta_1 = 1,2$	1,269	1,002	436	1,259	1,008	480
$\phi = -0,693$	-0,689	1,000	415	-0,696	1,012	360
$\theta = -1$	-0,992	1,000	416	-1,002	1,011	380
$\beta_0 = 1,5$	1,528	1,014	268	1,525	1,024	130
$\beta_1 = 1,2$	1,180	1,014	275	1,183	1,018	130
$\phi = -0,336$	-0,360	1,015	302	-0,359	1,017	130
$\theta = -0,5$	-0,534	1,015	300	-0,532	1,017	130

\*Média da distribuição a posteriori.

A Figura 7 apresenta os dados simulados para modelos *FP1* com potências  $\theta = 0,5$ ,  $\theta = 1$  e  $\theta = 2$ . Apresenta-se na Tabela 7 as estimativas dos parâmetros do modelo proposto para estes dados. Para estas potências, só houve convergência nos métodos *HMC*. Observa-se que estimativas foram próximas do valor verdadeiro.

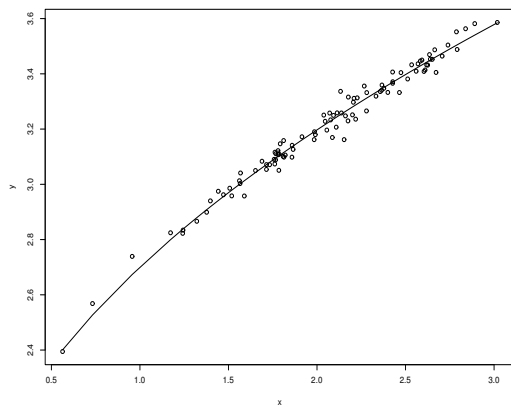
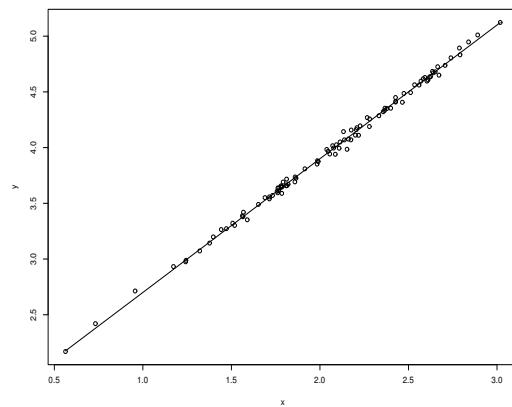
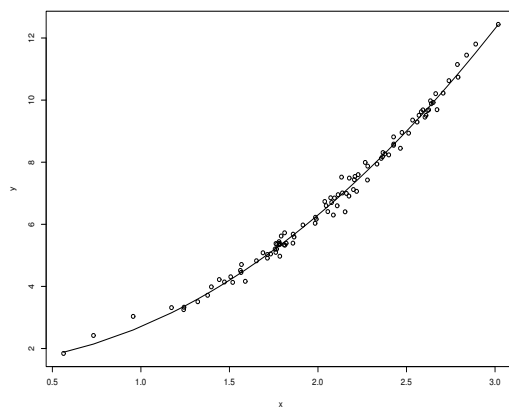
(a)  $\theta = 0,5$ .(b)  $\theta = 1$ .(c)  $\theta = 2$ .Figura 7: Dados simulados sob modelos *FP1* com diferentes potências.

Tabela 7: Propriedades dos ajustes, segundo a abordagem proposta, para modelos *FP1*, dados simulados com potências diferentes, de acordo com a técnica *HMC*

Parâmetros	<i>HMC</i>		
	Estimativas*	$\hat{R}$	$n_{eff}$
$\beta_0 = 1,5$	1,704	1,005	333
$\beta_1 = 1,2$	1,007	1,005	331
$\Phi = 0,336$	0,387	1,005	353
$\theta = 0,5$	0,574	1,005	352
$\beta_0 = 1,5$	1,590	1,004	269
$\beta_1 = 1,2$	1,124	1,004	262
$\Phi = 0,693$	0,722	1,003	266
$\theta = 1$	1,039	1,003	266
$\beta_0 = 1,5$	1,789	1,004	385
$\beta_1 = 1,2$	1,033	1,004	376
$\Phi = 1,609$	1,769	1,004	368
$\theta = 2$	2,124	1,004	372

\*Média da distribuição a posteriori.

A Figura 8 apresenta a comparação entre a densidade dos valores observados  $y$  com os dados replicados da distribuição preditiva  $y_{rep}$  para os casos *FP1*. Foram replicadas 1500 densidades preditivas para cada modelo, sendo que, na Figura 8, apenas 200 são apresentadas para melhor visualização para dois casos, modelos com  $\theta = -2$  e  $\theta = -0,5$ , segundo o método *HMC*. Nos dois casos, pode-se notar que o modelo está bem ajustado aos dados.



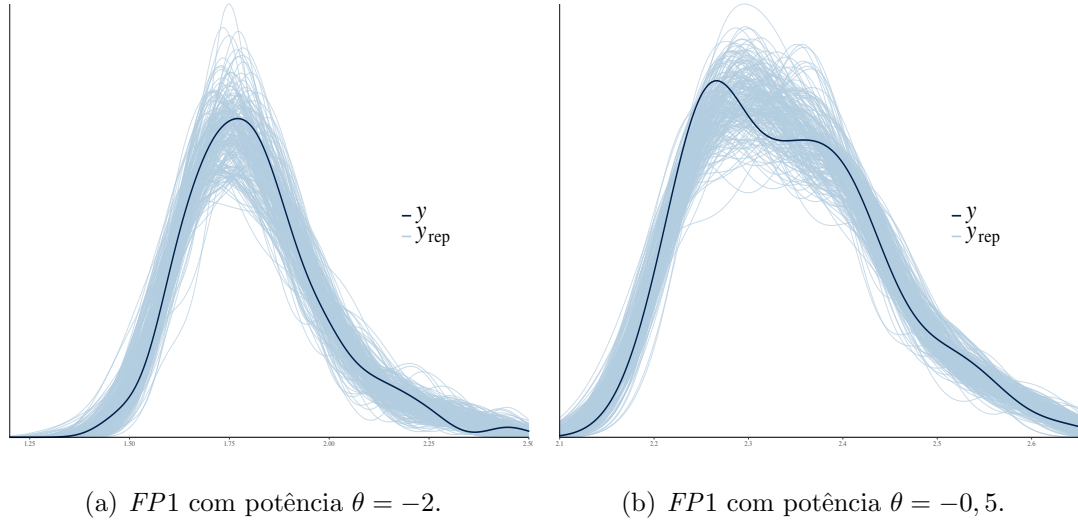


Figura 8: Comparação dos dados observados  $y$  com os dados replicados da distribuição preditiva  $y_{rep}$ .

#### 4.2.2 Estudos de casos simulados de modelos *FP2* modificado segundo a abordagem proposta

Para avaliar o modelo em (36), de segunda ordem, simulou-se dados de dois modelos quadráticos com as transformações com potência  $\theta = -0,5$  e  $\theta = 0,5$ . Para  $\theta = -0,5$ , utilizou-se  $\beta_0 = 150$ ,  $\beta_1 = 1$  e  $\beta_{11} = -0,3$ . Para  $\theta = 0,5$ , utilizou-se  $\beta_0 = 150$ ,  $\beta_1 = 1,2$  e  $\beta_{11} = -0,3$ , escolhidos de tal maneira que as curvas caracterizassem padrões de interesse prático. Em ambas simulações,  $x$  é uma sequência de 100 números de 0,1 até 10. Desta forma, pode-se representar os modelos simulados da seguinte maneira

$$y_i = 150 + x^\theta - 0,3x^{2\theta} + \epsilon_i, \quad \text{para } \theta = -0,5 \text{ e} \quad (41)$$

$$y_i = 150 + 1,2x^\theta - 0,3x^{2\theta} + \epsilon_i, \quad \text{para } \theta = 0,5, \quad (42)$$

com  $\epsilon_i \sim N(0; 0,05^2)$ .

A Figura 9 apresenta os dados simulados para  $FP2$  com potência  $\theta = -0,5$  e  $\theta = 0,5$  e as curvas esperadas. Apresenta-se na Tabela 8 o desempenho da abordagem proposta para estes dados. Foram usadas 1.000 iterações no **rstan**, que utiliza os métodos *HMC*. Verifica-se, a partir desta tabela, que os coeficientes estimados foram bem próximos dos valores verdadeiros. Além disso, o  $\hat{R}$  foi próximo de 1, caracterizando a convergência dos ajustes, o que pode ser confirmado também pela análise dos traços que estão no anexo 3.

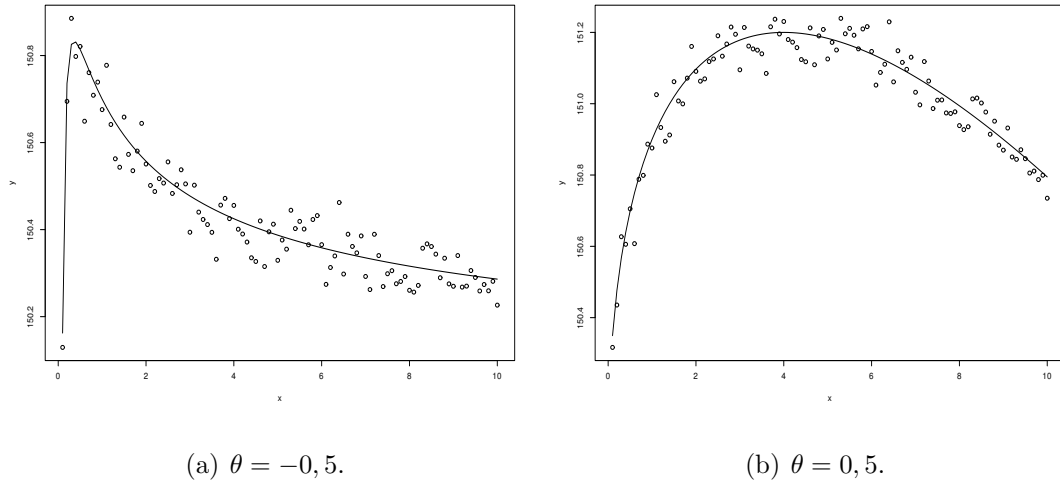


Figura 9: Dados simulados sob modelos  $FP2$  com potência diferentes, acordo com a técnica *HMC*.

Tabela 8: Propriedades dos ajustes, segundo a abordagem proposta, para modelos *FP2*, dados simulados com potência diferentes, de acordo com a técnica *HMC*

Parâmetros	<i>HMC</i>		
	Estimativas	$\hat{R}$	$n_{eff}$
$\alpha_0 = 150$	150,001	1,019	206
$\alpha_1 = 1$	0,978	1,019	214
$\alpha_{11} = -0,3$	-0,295	1,020	205
$\phi = 0,336$	-0,339	1,015	214
$\theta = -0,5$	-0,504	1,015	214
$\alpha_0 = 150$	149,980	1,014	263
$\alpha_1 = 1,2$	1,199	1,014	253
$\alpha_{11} = -0,3$	-0,297	1,014	249
$\phi = 0,336$	0,341	1,012	242
$\theta = 0,5$	0,507	1,012	242

\*Média da distribuição a posteriori.

A Figura 10 apresenta a comparação entre a densidade dos valores observados  $y$  com os dados replicados da distribuição preditiva  $y_{rep}$  para os casos *FP2*. Foram replicadas 1500 densidades preditivas para cada modelo, sendo que, na Figura 10, apenas 200 são apresentadas para melhor visualização. Pode-se notar que os dois modelos estão bem ajustados aos dados.

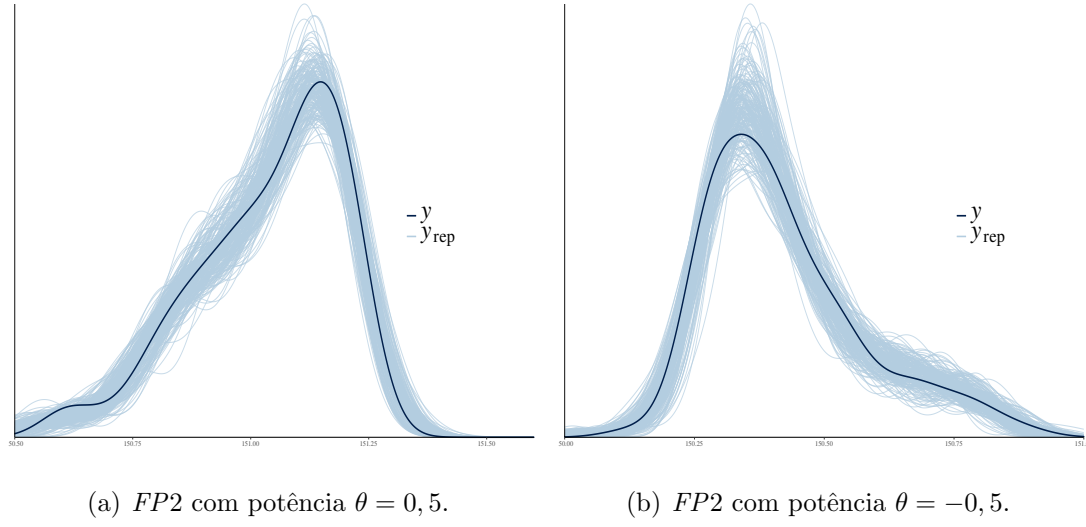


Figura 10: Comparação dos dados observados  $y$  com os dados replicados da distribuição preditiva  $y_{rep}$ .

### 4.2.3 Aplicação para o conjunto de dados IMC

Nesta subseção utiliza-se novamente os dados IMC para ajustar o modelo de primeira ordem, segundo a abordagem proposta.

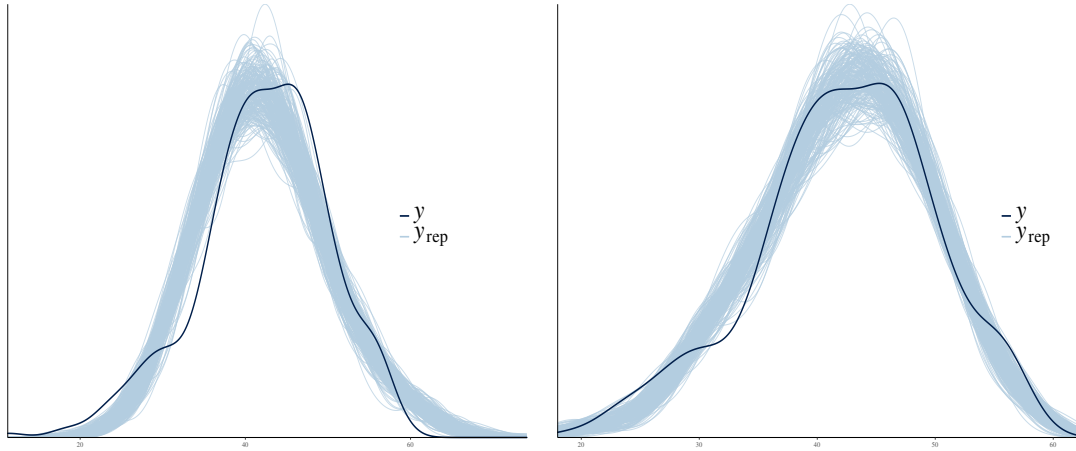
Usando distribuições a priori vagas para os coeficientes de regressão e  $\Phi \sim N(0, 3)$  obteve-se as estimativas dos parâmetros do modelo proposto apresentados na Tabela 9. Observa-se que houve convergência do ajuste do modelo, visto que o  $\hat{R}$  é próximo de 1 para todas as estimativas, além disso, o traço desta ajuste é estacionário (ver anexo 3).

Tabela 9: Estimativas dos parâmetros sob a abordagem proposta, para o conjunto de dados IMC, utilizando a técnica *HMC*

Parâmetros	Estimativas*	$\hat{R}$	$n_{eff}$
$\beta_0$	78,820	1,004	634
$\beta_1$	-81,667	1,001	1044
$\Phi$	-0,509	1,002	646
$\theta$	-0,748	1,002	646

\* Média da distribuição a posteriori.

No entanto, conforme apresentado na Figura 11 (a) que mostra a densidade dos valores observados  $y$  com 200 dados replicados da distribuição preditiva  $y_{rep}$ , verifica-se que, apesar da convergência, com o uso de prioris não informativas os ajustes foram apenas razoáveis. Apenas quando fornecido prioris bastante informativas (Figura 11 (b)) obteve-se resultados mais similares à abordagem clássica.



(a) Resultado com uso de prioris vagas. (b) Resultado com uso de prioris informativas.

Figura 11: Comparação dos dados observados  $y$  com os dados replicados da distribuição preditiva  $y_{rep}$ , dados IMC.

#### 4.2.4 Aplicação para o conjunto de dados aranha

O diagrama de dispersão para o peso e o comprimento corporal de aranhas fêmeas está apresentado na Figura 12, indicando uma relação não linear entre as variáveis. A tendência não linear parece poder ser captada por um *FP1*, porém, nota-se razoável heterogeneidade de variâncias dos pesos conforme o comprimento aumenta. Esse problema pode ser remediado utilizando a família de transformações Box-Cox (Box & Cox, 1964). Assim, o modelo é

$$\frac{y^{(\lambda)-1}}{\lambda} = \beta_0 + \beta_1 x^\theta + \varepsilon \quad (43)$$

tal que  $y^{(\lambda)} = \log(y)$  se  $\lambda = 0$  e  $y^{(\lambda)} = y^\lambda$  se  $\lambda \neq 0$ . O código para o ajuste do modelo *FP1* segundo a abordagem proposta foi então adaptada para acomodar a transformação em  $y$ .

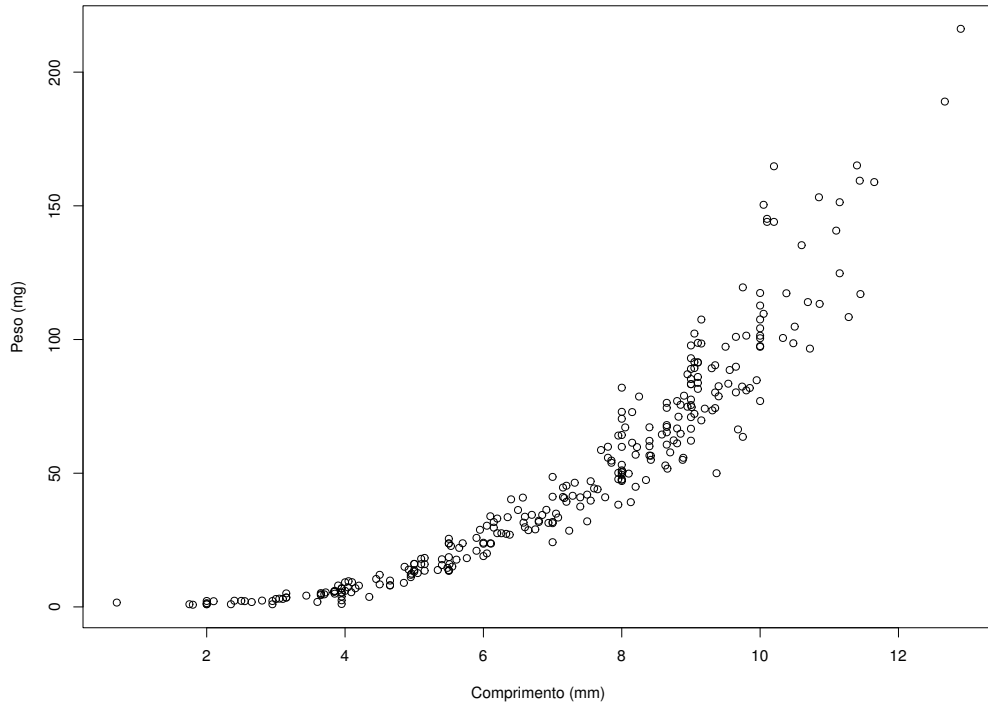


Figura 12: Diagrama de dispersão do comprimento (mm) e peso (mg) de aranhas.

A Tabela 10 apresenta as estimativas dos parâmetros do modelo proposto. Para estes dados, utilizou-se como priori uma distribuição normal  $\beta_0 \sim N(50; 5^2)$  para o coeficiente linear, uma distribuição normal  $\beta_1 \sim N(2; 1^2)$  para o coeficiente angular e, para a constante  $\lambda$  uma distribuição normal  $\lambda \sim N(0; 2^2)$ . Observa-se que houve convergência do ajuste do modelo, visto que o  $\hat{R}$  é próximo de 1 para todas as estimativas, bem como, os traços dos parâmetros apresentam-se de forma estacionária, conforme anexo 3.

Tabela 10: Estimativas dos parâmetros sob a abordagem proposta, para o conjunto de dados aranha, utilizando a técnica *HMC*

Parâmetros	Estimativas*	$\hat{R}$	$n_{eff}$
$\beta_0$	-2,72	1,000	432
$\beta_1$	1,39	1,001	335
$\lambda$	0,326	1,001	330
$\Phi$	0,69	1,001	280
$\theta$	0,995	1,001	282

\*Média da distribuição a posteriori.

A Figura 13 apresenta a comparação entre a densidade dos valores observados  $y$  com 200 dados replicados da distribuição preditiva  $y_{rep}$ . Nela pode-se verificar que o modelo está bem ajustado aos dados.

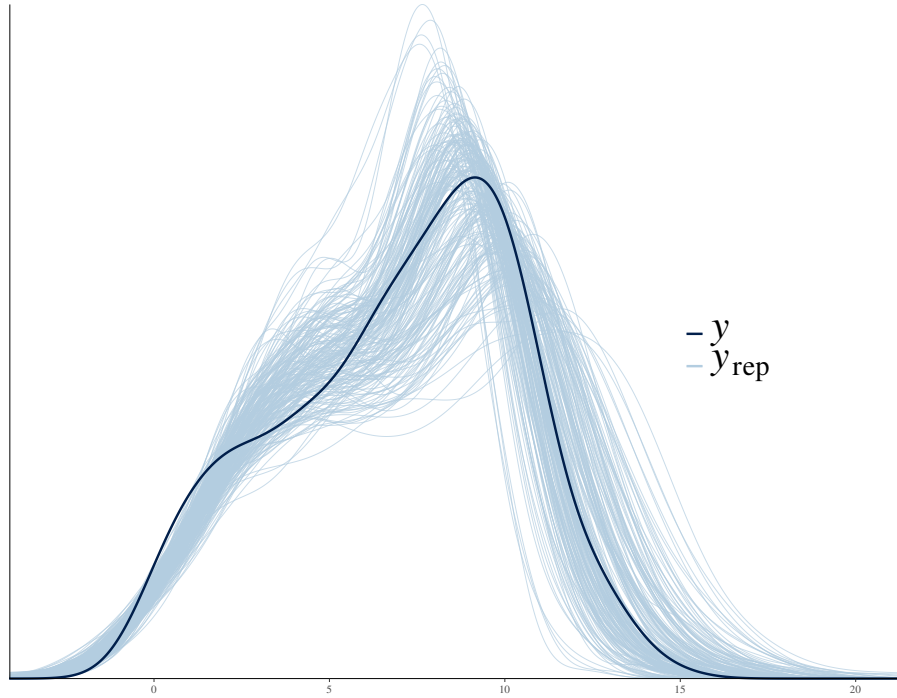


Figura 13: Comparação dos dados observados  $y$  com os dados replicados da distribuição preditiva  $y_{rep}$ , para o conjunto de dados aranhas.

Neste capítulo, fez-se uma comparação dos métodos tradicionais, sob a abordagem clássica e bayesiana para  $FPs$ , em que verificou-se pequenas diferenças nas escolhas das covariáveis, bem como nas potências do  $FP$ . Nas simulações para avaliar o percentual de acerto no grau do  $FP$ , a abordagem clássica apresentou melhores resultados quando a função subjacente é um  $FP1$ , enquanto que a abordagem bayesiana foi melhor quando a função subjacente é um  $FP2$ .

Nos estudos de simulação para avaliar o modelo proposto nesta tese, os resultados foram interessantes tanto para simulação de um  $FP1$  quanto para  $FP2$ , sendo que, a técnica  $HMC$  foi melhor que a  $MCMC$ . Para aplicação do modelo proposto nos dados  $IMC$  os resultados foram bons apenas utilizando-se prioris menos difusas, o mesmo vale para os dados aranha.



## 5 CONCLUSÕES

Este trabalho teve como objetivo explorar a técnica de modelagem via polinômios fracionários sob as abordagens clássica e bayesiana, bem como, propor um modelo *FP2* modificado com ajuste via regressão não linear bayesiana. Para tanto, utilizaram-se conjuntos de dados da literatura univariados e multivariados para comparação das técnicas usuais. Além disso, tendo como principal motivação as diferenças na seleção das covariáveis, bem como nas potências estimadas em cada técnica, fez-se um estudo de simulação para uma covariável, com o intuito de observar a proporção de acertos da verdadeira função subjacente. O ajuste via metodologia bayesiana do modelo polinômio fracionário modificado proposto foi avaliado a partir de dados simulados e dados reais.

Nas aplicações em dados da literatura, verificou-se resultados semelhantes entre as técnicas, havendo pequenas diferenças entre os valores das potências estimadas, nas covariáveis selecionadas, bem como nas estimativas dos coeficientes de regressão. Tais diferenças se devem ao fato de que, na abordagem clássica os coeficientes de regressão são obtidos a partir do *MLE*, ao passo que, na abordagem bayesiana, as estimativas dos coeficientes são a moda da distribuição a posteriori, além de que a abordagem bayesiana de Bové & Held (2011) sinalizou resultar em modelos mais complicados.

Tanto a abordagem clássica quanto a bayesiana apresentaram grande proporção de acerto. Entretanto, quando a função subjacente foi *FP1*, apesar de ambas abordagens terem resultados acima de 84%, a técnica clássica apresentou melho-

res resultados. Por outro lado, quando a função subjacente foi *FP2*, a abordagem bayesiana apresentou resultados bem melhores que a clássica.

Quanto ao modelo proposto, nas simulações de *FP* de ordem 1, com potências  $\theta = -2$ ,  $\theta = -1$  e  $\theta = -0,5$ , ambos métodos *HMC* e *MCMC* apresentaram excelentes estimativas. Para  $\theta = 0,5$ ,  $\theta = 1$  e  $\theta = 2$ , houve convergência somente nos métodos *HMC*. Nos resultados simulados para *FP2*, só houve êxito nos métodos de simulação Hamiltonianos, apresentando ótimas estimativas, bem como bom ajuste aos dados, comparados com a distribuição preditiva. Na aplicação da abordagem bayesiana proposta a dados reais, os ajustes apresentarem maiores desafios. Para o conjunto de dados IMC, os resultados foram apenas razoáveis, utilizando-se prioris não-informativas, indicando a necessidade de pesquisas mais detalhadas sobre a elicitación de prioris para esse tipo de modelo, visto que, utilizando-se prioris informativas, os resultados foram muito bons e mais similares à abordagem clássica. Uma provável complexidade é a falta de independência entre os parâmetros potência e o coeficiente de regressão. Já para o conjunto de dados aranha, obteve-se excelentes resultados com a implementação do método proposto com a transformação de Box-Cox.

A abordagem proposta nesta tese tem grande potencial para ser aplicada numa grande variedade de conjunto de dados, devido à sua flexibilidade. É recomendado que sejam feitos estudos de simulação mais amplos englobando outras curvaturas e variando-se  $\sigma^2$  além das que foram usadas aqui. Todavia, este modelo tem a vantagem de ser parcimonioso, visto que, diferentemente dos modelos clássico e bayesiano para *FP* propostos por Royston & Altman (1994) e Bové & Held (2011), respectivamente, ele precisa estimar uma única potência por regressora para ajustar um *FP2*. Além disso, o método de ajuste para o modelo proposto permite maior flexibilidade na estimação das potências, isto é, pode acomodar mais curvaturas, visto que, pode-se ter qualquer valor dentro do intervalo  $(-3,3)$  que também pode ser ampliado, pois este não está restrito ao conjunto  $S$  como nas técnicas usuais. Em

termos práticos, esta pode ser a maior vantagem do modelo proposto.

## Anexos

### Anexo 1

- Descrição das variáveis para os dados sobre câncer (Royston & Sauerbrei, 2008).

x1	age	Age (years)
x2	svi	Seminal vessel invasion (Y/ N)
x3	pgg45	Percentage Gleason score 4 or 5
x4	cavol	Cancer volume (ml)
x5	weight	Prostate weight (g)
x6	bph	Amount of benign prostatic hyperplasia (g)
x7	cp	Amount of capsular penetration (g)
x8	lpsa	log PSA concentration *VARIÁVEL RESPOSTA

- Descrição das variáveis para os dados sobre ozônio (Bové & Held, 2011).

y	Maximum 1-hour average ozone level [ppm] Upland, CA
x0	dayOfYear
x1	Month
x2	Day of month
x3	Day of week

```

x4          500 millibar pressure height [m]
x5          Wind speed [mph]
x6          Relative humidity [%]
x7          Temperature [F] Sandberg
x8          Inversion base height [feet]
x9          Pressure gradient [mm Hg] from LAX to Daggett, CA
x10         Visibility [miles]
x11         Inversion base temperature [F]
x12         Temperature [F] El Monte

```

- Função para simulação de modelos *mfp* e *bfp* para verificar a porcentagem de acertos no grau do *FP* segundo a abordagem de ajuste de modelo quando a função subjacente é um *FP1* e um *FP2*.

```

simulacao <- function(n_sim, n_reptreino,n,meanx,sigx,sigy,R2,fp)
{
  mfp_npot = matrix(0,n_sim,n_reptreino)
  mfp_pot = matrix(0,n_sim,2*n_reptreino)
  bfp_npot = matrix(0,n_sim,n_reptreino)
  bfp_pot = matrix(0,n_sim,2*n_reptreino)

  beta = (R2/(1-R2))^0.5

  for (i in 1:n_sim) {

    if(fp==1) {
      x = rlnorm(n,meanx,sigx) ### covariavel para PF1
      y = rnorm(n,beta*x,sigy)
    }
    else {

```

```

x = rnorm(n,meanx,sigx) ### covariavel para PF2
y = rnorm(n,beta*((x-meanx)^2),sigy)
}

dados = as.data.frame(cbind(y,x))
n_teste = ceiling(n*0.10)
for (j in 1:n_reptreino){
  dados_sample = sample(1:nrow(dados), n_teste, replace = FALSE)
  dados_teste = dados[dados_sample,]
  dados_treino = dados[-dados_sample,]

  ### modelo baysiano ###
  m1_bfb = BayesMfp(y ~
    bfp(x, max = 2),
    data = dados_treino,
    nModels = 3000,
    method = "exhaustive",
    chainlength = 2e+5)
  resumo1 = summary(m1_bfb[1])
  post = getPosteriorParms(m1_bfb)[]
  n_pot_bfp = length(resumo1$summaryMat[,1])-1 ### numero de potencias
  # pot_bfp = m1_bfb$'1'[1] ### potencias p1 e/ou p2
  pot_bfp = resumo1$dataframe[7] ### potencias p1 e/ou p2
  bfp_npot[i,] = n_pot_bfp
  # bfp_pot[i,] = pot_bfp

  # modelo classico ###
  m2_mfp = mfp(y ~
    fp (x, df=4, select = 0.05),
    family = gaussian,
    data = dados_treino)
  resumo2 = summary(m2_mfp)

```

```

n_pot_mfp = (length(resumo2$coefficients[,1])-1)
pot_mfp = m2_mfp$powers ### potencias p1 e/ou p2
mfp_npot[i,] = n_pot_mfp
mfp_pot[i,] = pot_mfp
}
}
list(mfp_pot=mfp_pot,mfp_npot=mfp_npot,bfp_npot=bfp_npot)
}

```

## Anexo 2

### Máxima densidade a posteriori - MAP

De acordo com Lesaffre & Lawson (2012), em princípio, a seleção de variável Bayesiana (e modelo) é relativamente simples e envolve apenas o cálculo das probabilidades do modelo posterior. Seja  $p(M_m)$  a probabilidade a priori para o modelo  $M_m$ , então sua probabilidade posterior é

$$p(M_m|\mathbf{x}) = \frac{p(\mathbf{x}|M_m)p(M_m)}{\sum_{j=1}^K p(\mathbf{x}|M_j)p(M_j)}, \quad (44)$$

com

$$p(\mathbf{x}|M_m) = \int p(\mathbf{x}|\boldsymbol{\theta}_m, M_m)p(\boldsymbol{\theta}_m|M_m)d\boldsymbol{\theta}_m, \quad (45)$$

que é a distribuição marginal do  $m$ -ésimo modelo em média sobre os possíveis valores dos parâmetros do modelo. O princípio Bayesiano determina a escolha do modelo com maior probabilidade posterior, que é o modelo máximo a posteriori (*MAP*). Uma expressão alternativa da probabilidade posterior é

$$p(M_m|\mathbf{x}) = \frac{BF_{mk}p(M_m)}{\sum_{j=1} BF_{jk}p(M_j)}, \quad (46)$$

em que  $BF_{jk} = p(\mathbf{x}|M_j)/p(\mathbf{x}|M_k)$  é o fator de Bayes comparando o modelo  $M_j$  com o modelo  $M_k$ .

## Ponderação Bayesiana de modelos - BMA

Ghosh et al. (2006) comentam que, na abordagem Bayesiana pode-se estar interessado em encontrar um modelo específico que faça o melhor trabalho de previsão. Ou, pode-se desejar apenas um valor previsto, não um modelo específico. Em qualquer um dos casos, a função de perda mais popular é a perda de erro de predição quadrático, isto é, o quadrado da diferença entre o valor predito/estimado e o valor a ser previsto/estimado. O melhor preditor/estimador acaba por ser a estimativa da média da ponderação Bayesiana de modelos e o melhor modelo preditivo é aquele que minimiza a perda preditiva posterior esperada.

Suponha que calculamos as probabilidades posteriores



$$p(M_k|x) = \frac{p(x|M_m)p(M_k)}{\sum_{j=1}^k p(x|M_j)p(M_j)}, \quad (47)$$

de todos os modelos  $M_1, \dots, M_k$ . Segundo Held & Bové (2014) escolhendo o *MAP* como o melhor modelo, descartar os demais modelos pode não ser o mais adequado a se fazer. Muitas vezes, uma quantidade desconhecida, digamos,  $\lambda = h(\boldsymbol{\theta})$ , é o objetivo de interesse, e afirmações de incerteza sobre  $\lambda$  são o objetivo final da análise estatística. Nesse caso, podemos simplesmente calcular a distribuição marginal posterior de  $\lambda$  que é uma mistura discreta de distribuições posteriores específicas do modelo  $K$  com pesos dados pelas probabilidades do modelo posterior:

$$f(\lambda|x) = \sum_{j=1}^k p(\lambda|x, M_m)p(M_k), \quad (48)$$

esta é uma ponderação Bayesiana de modelos (*BMA*), que leva em conta toda a incerteza sobre o mesmo.

## Erro quadrático médio de predição

Para Lesaffre & Lawson (2012), em essência, uma função de perda (erro quadrado ou de outra forma) poderia ser configurada entre uma observação e seu valor previsto. Para obter este valor previsto, uma simulação do modelo ajustado é usada, na verdade, a partir da distribuição preditiva a posteriori (*posterior predictive distribution* - *PPD*), gerando  $\tilde{y}_i$ . Este valor é comparado ao valor observado através da função de perda preditiva a posteriori (*posterior predictive loss* - *PPL*). Para perda de erro quadrático temos

$$MSPE = \sum_{i=1}^n (y_i - \tilde{y}_i)^2, \quad (49)$$

em que  $MSPE$  é o erro quadrático médio de predição (*mean-square predictive error*);  $y_i$  é o valor observado.

Como no caso do erro quadrático médio (*mean squared errors - MSE*), em um contexto *MCMC*, deve-se tomar uma decisão sobre como estimar o valor preditivo. Geralmente, uma medida é calculada em cada iteração do amostrador e, em seguida, é calculada a média da amostra. Assim, a forma final seria nesse caso

$$MSPE = \frac{1}{K} \sum_{k=1}^K MSPE_k, \quad (50)$$

com  $MSPE_k$  calculando a partir de (49) na  $k$ -ésima iteração do processo *MCMC*. O  $RMSPE$  é a raiz do  $MSPE$ .

## Amostragem do modelo posterior

Bové & Held (2011) comentam que, para uma única covariável, o número de possíveis modelos *FP* com graus de liberdade  $m = 0, 1, 2, 3$  é  $d(m) = 1, 8, 36$  e  $120$ , respectivamente. Se, por exemplo, tivermos o grau do *FP* máximo  $m_{max} = 2$  e  $k = 5$  covariáveis, existem  $(1 + 8 + 36)^5 = 184.528.125$  diferentes modelos. Isso muitas vezes torna um cálculo exaustivo de todas as probabilidades do modelo posterior  $f(\boldsymbol{\theta}|\mathcal{D})$  para todo  $\boldsymbol{\theta} \in \boldsymbol{\Theta}$  via (35), (28) e (34), inviável. Ao invés de usar estratégias como procedimentos *BE* para este fim, os autores tomam amostras da distribuição a posteriori  $f(\boldsymbol{\theta}|\mathcal{D})$  via *MCMC*.

A distribuição proposta  $q(\boldsymbol{\theta}'|\boldsymbol{\theta})$  é formada por quatro diferentes tipos de movimentos, que definem como passar do modelo atual  $\boldsymbol{\theta}$  para  $\boldsymbol{\theta}'$ :

**BIRTH** - Selecione aleatoriamente uma das covariáveis com  $FP$  com  $m_i < m_{max}$  graus de liberdade. Adicione uma potência para  $\mathbf{p}_i$  aleatoriamente de  $S$ .

**DEATH** - Selecione aleatoriamente uma das covariáveis com  $FP$  com  $m_i > 0$  graus de liberdade. Remova um potência escolhida aleatoriamente de  $\mathbf{p}_i$ .

**MOVE** - Selecione aleatoriamente uma das covariáveis com  $FP$  com  $m_i > 0$  graus de liberdade. Remova um potência escolhida aleatoriamente de  $\mathbf{p}_i$ , então, retire aleatoriamente uma potência de  $S$  e adicione-a a  $\mathbf{p}_i$ .

**SWITCH** - Selecione aleatoriamente uma das covariáveis com vetor de potência  $\mathbf{p}_i$  não vazio. Selecione aleatoriamente uma das outras covariáveis com o vetor de potências  $\mathbf{p}_j$ . Alterne os vetores de potência  $\mathbf{p}_i$  e  $\mathbf{p}_j$ .

Observe que o movimento *SWITCH* é sensível apenas para  $k > 1$  covariáveis, mas para  $k = 1$  todos os modelos podem ser facilmente avaliados sem qualquer amostragem do modelo. O movimento *SWITCH* foi projetado para rastrear com eficiência modelos com alta probabilidade posterior mesmo em situações em que as covariáveis são quase colineares. Cada proposta começa com a escolha probabilística de um dos tipos de movimento, com as quatro probabilidades  $b_{p_\theta}$ ,  $d_{p_\theta}$ ,  $m_{p_\theta}$  e  $s_{p_\theta}$  dependendo da dimensão atual de  $p_\theta$  de todo vetor de parâmetros  $\boldsymbol{\theta}$ :

$$\begin{aligned} b_{p_\theta} &= 1, \quad d_{p_\theta} = m_{p_\theta} = s_{p_\theta} = 0, & \text{se } p_\theta; \\ b_{p_\theta} &= d_{p_\theta} = m_{p_\theta} = s_{p_\theta} = \frac{1}{4}, & \text{se } 0 < p_\theta < p_{max}; \\ b_{p_\theta} &= 0, \quad d_{p_\theta} = m_{p_\theta} = s_{p_\theta} = \frac{1}{4}, & \text{se } p_\theta = p_{max}, \end{aligned}$$

em que o valor  $p_{max} := \min\{n - 3 - a, k \times m_{max}\}$  leva em conta que mais do que  $n - 3 - a$  potências tornariam impróprias as distribuições posteriores no modelo (ver Liang et al., 2008).

A proposta do novo modelo  $\boldsymbol{\theta}'$  é aceito com probabilidade

$$\alpha(\boldsymbol{\theta}'|\boldsymbol{\theta}) = \min \left\{ 1, \frac{f(\mathcal{D}|\boldsymbol{\theta}')}{f(\mathcal{D}|\boldsymbol{\theta})} \frac{f(\boldsymbol{\theta}')}{f(\boldsymbol{\theta})} \frac{q(\boldsymbol{\theta}|\boldsymbol{\theta}')}{q(\boldsymbol{\theta}'|\boldsymbol{\theta})} \right\},$$

que é a probabilidade usual de aceitação de *Metropolis-Hastings*; em caso de rejeição, a cadeia permanece no modelo anterior  $\boldsymbol{\theta}$ . As únicas partes de  $\alpha(\boldsymbol{\theta}'|\boldsymbol{\theta})$  que ainda precisam ser calculadas é a razão anterior  $\frac{f(\boldsymbol{\theta}')}{f(\boldsymbol{\theta})}$  e a razão proposta  $\frac{q(\mathcal{D}|\boldsymbol{\theta}')}{q(\mathcal{D}|\boldsymbol{\theta})}$  porque o fator de Bayes  $\frac{f(\mathcal{D}|\boldsymbol{\theta}')}{f(\mathcal{D}|\boldsymbol{\theta})}$  é conhecido de (35). Ambas as razões, anterior e proposta, dependem do tipo de movimento proposto.

## Transformação de Box-Cox

Em diversas situações práticas, a suposição de normalidade dos dados não é válida. Um procedimento comum para que os dados se aproximem de uma distribuição normal, é a transformação de Box-Cox. Segundo Box & Cox (1964), busca-se um valor de  $\lambda$  que maximize a verossimilhança, tal que os dados passem a ter uma distribuição normal, a partir do seguinte procedimento:

$$y^{(\lambda)} = \begin{cases} \frac{y^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ \log y, & \lambda = 0, \end{cases}$$

em que assume-se que para alguns  $\lambda$  conhecido, o vetor  $\mathbf{y}^{(\lambda)} = (y_1^{(\lambda)}, \dots, y_n^{(\lambda)})$  de observações transformadas satisfaz o modelo linear

$$E(\mathbf{y}^{(\lambda)}) = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

que é ajustado via *OLS*.

## **Anexo 3**

### **Traços dos modelos da seção 4.2.1**

Na subseção 4.2.1 apresentou-se um estudo de dados simulados de modelos *FP1* para avaliar o modelo proposto nesta tese. As Figuras 14 e 15 apresentam os traços para as simulações da equação (40), em que, pode-se notar que as séries são estacionárias para todos os parâmetros, indicando convergência do ajuste.

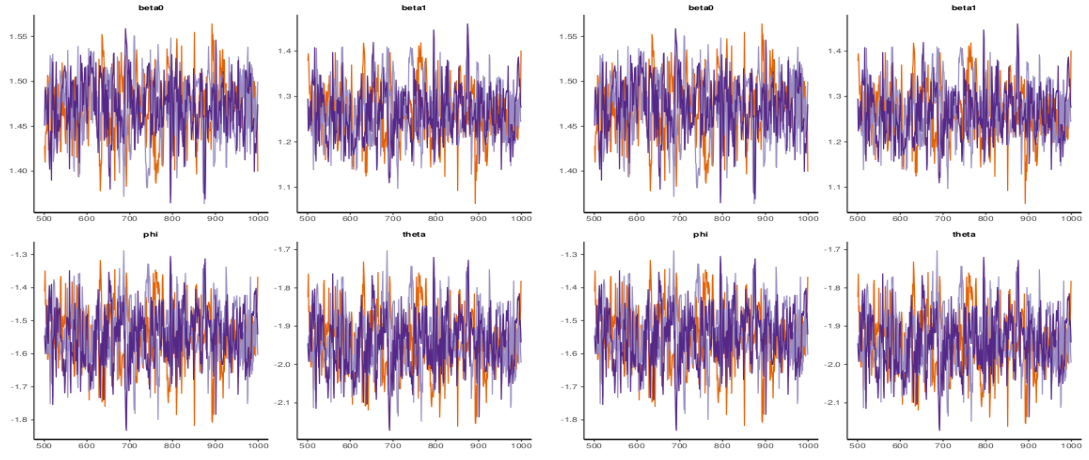
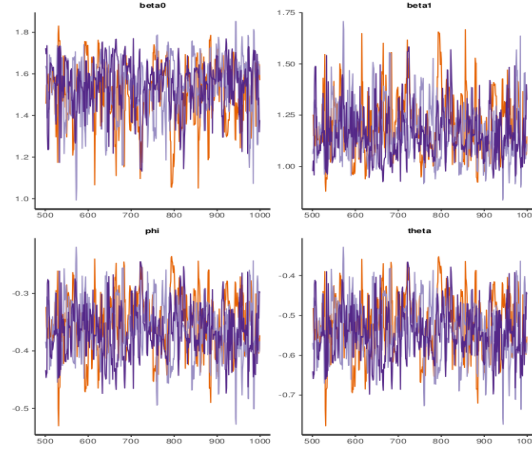
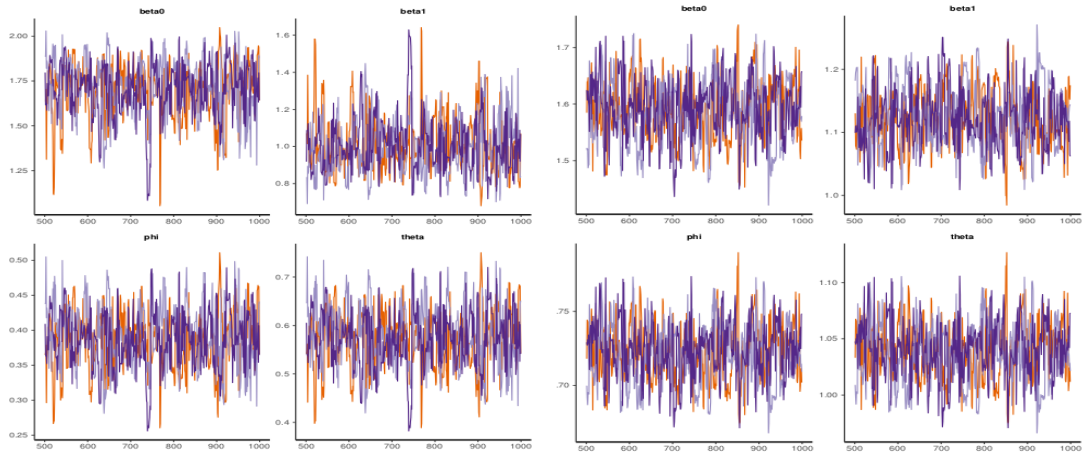
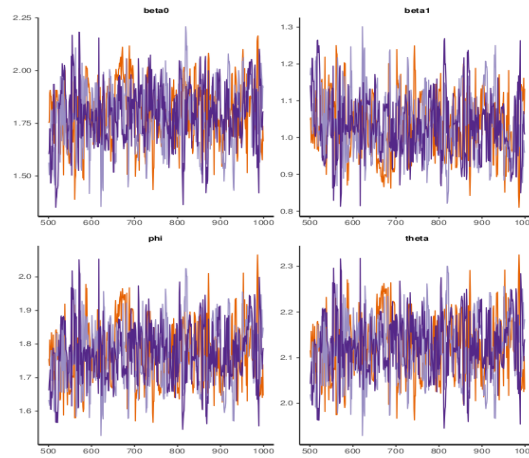
(a)  $\theta = -2$ .(b)  $\theta = -1$ .(c)  $\theta = -0.5$ .

Figura 14: Traço dos dados simulados sob modelos *FP1* com diferentes potências.



(a)  $\theta = 0,5$ .

(b)  $\theta = 1$ .



(c)  $\theta = 2$ .

Figura 15: Traço dos dados simulados sob modelos *FP1* com diferentes potências.

## Traços dos modelos da seção 4.2.2

Na subseção 4.2.2 apresentou-se um estudo de dados simulados de modelos *FP2* para avaliar o modelo proposto nesta tese. A Figura 16 apresenta os traços para

as simulações das equações (41) e (42), em que, pode-se verificar que as séries são estacionárias para todos os parâmetros, indicando convergência do ajuste.

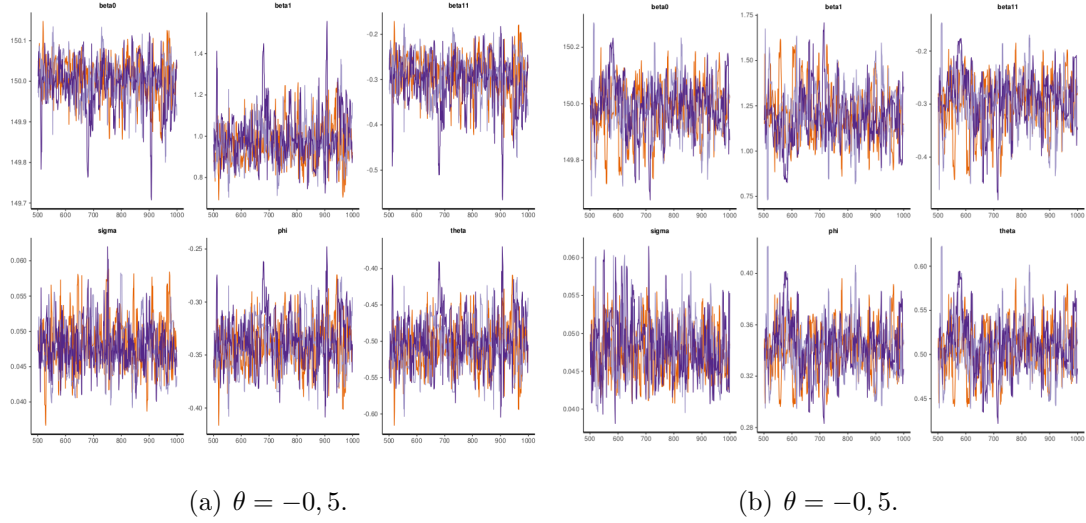
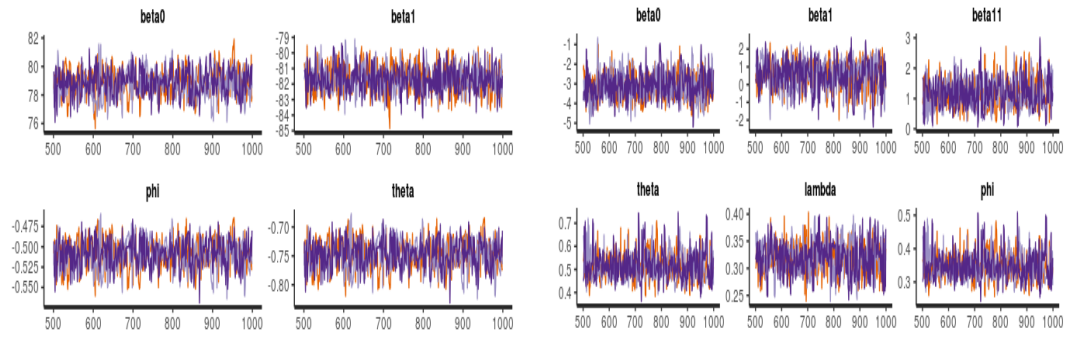


Figura 16: Traço dos dados simulados sob modelos *FP2* com diferentes potências.

## Traços dos modelos IMC e aranha

Nas subseções 4.2.3 e 4.2.4 apresentou-se a aplicação do modelo proposto nos dados IMC e aranha, respectivamente. A Figura 17 apresenta os traços desta aplicação, em quem pode-se notar que as séries são estacionárias para todos os parâmetros, indicando convergência dos ajustes.





(a) Traço para os dados IMC.

(b) Traço para os dados aranha.

Figura 17: Traço dos dados IMC e aranha.

## REFERÊNCIAS BIBLIOGRÁFICAS

ORIGINAL POR GARETH AMBLER E MODIFICADO POR AXEL BENNER.

**mfp: Multivariable Fractional Polynomials**, 2015. R package version 1.5.2.

BATES, D. M.; WATTS, D. G. **Nonlinear regression analysis and its applications**. Wiley, 1988. Wiley series in probability and mathematical statistics. Applied probability and statistics.

BOVÉ, D. S.; HELD, L. Bayesian fractional polynomials. **Statistics and Computing**, v.21, n.3, p.309–324, 2011.

BOX, G. E.; COX, D. R. An analysis of transformations. **Journal of the Royal Statistical Society. Series B (Methodological)**, p.211–252, 1964.

BREIMAN, L. E. O.; FRIEDMAN, J. H. Estimating Optimal Transformations for Multiple Regression and Correlation. , n.September, p.580–598, 1985.

DENISON, D.; HOLMES, C.; MALLICK, B.; SMITH, A. **Bayesian Methods for Nonlinear Classification and Regression**. , 2002.

DRAPER, N. R.; SMITH, H. **Applied regression analysis**. John Wiley & Sons, 2014. 326v.

GELMAN, A.; CARLIN, J. B.; STERN, H. S.; DUNSON, D. B.; VEHTARI, A.; RUBIN, D. B. **Bayesian Data Analysis**. 3. ed. Chapman and Hall/CRC, 2014. Chapman and Hall/CRC Texts in Statistical Science.

GELMAN, A.; HILL, J. **Data analysis using regression and multilevel/hierarchical models**. Cambridge university press, 2006.

GELMAN, A.; JAKULIN, A.; PITTAU, M. G.; SU, Y.-S.; ET AL. A weakly informative default prior distribution for logistic and other regression models. **The Annals of Applied Statistics**, v.2, n.4, p.1360–1383, 2008.

GHOSH, J. K.; DELAMPADY, M.; SAMANTA, A. **An Introduction to Bayesian Analysis. Theory and Methods.** , 2006. 356p.

GORDON, R. A. **Regression Analysis for the Social Sciences.** New York, NY, EUA: Routledge, 2015.

GUJARATI, D. N.; POTER, D. C. **Econometria Básica.** New York, NY, EUA: The McGraw-Hill, 2011. 918p.

HELD, L.; BOVÉ, D. S. **Applied Statistical Inference: Likelihood and Bayes.** 1. ed. Springer-Verlag Berlin Heidelberg, 2014.

HOFFMAN, M. D.; GELMAN, A. The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. v.15, p.1593–1623, 2014.

KORNER-NIEVERGELT, F.; ROTH, T.; VON FELTEN, S.; GUÉLAT, J.; ALMASI, B.; KORNER-NIEVERGELT, P. **Bayesian Data Analysis in Ecology Using Linear Models with R, BUGS, and Stan.** 1. ed. Academic Press, 2015.

LESAFFRE, E.; LAWSON, A. B. **Bayesian biostatistics.** John Wiley & Sons, 2012.

LIANG, F.; PAULO, R.; MOLINA, G.; CLYDE, M. A.; BERGER, J. O. Mixtures of g priors for Bayesian variable selection. **Journal of the American Statistical Association**, v.103, n.481, p.410–423, 2008.

LUKE, A.; DURAZO-ARVIZU, R.; ROTIMI, C.; PREWITT, T. E.; FORRESTER, T.; WILKS, R.; OGUNBIYI, O. J.; SCHOELLER, D. A.; MCGEE, D.; COOPER, R. S. Relation between body mass index and body fat in black population samples from Nigeria, Jamaica, and the United States. **American journal of epidemiology**, v.145, n.7, p.620–628, 1997.

MCELREATH, R. **Statistical Rethinking: A Bayesian Course with Examples in R and Stan**. Chapman and Hall/CRC, 2015. Chapman and Hall/CRC Texts in Statistical Science.

MONTGOMERY, D. C.; PECK, E. A.; VINING, G. G. **Introduction to linear regression analysis**. Canada: Wiley, 2012. 504p.

PAULINO, C. D.; TURKMAN, M. A. A.; V, B. **Estatística Bayesiana**. Lisboa: Fundação Calouste Gulbenkian, 2003.

PLUMMER, M. JAGS Version 3.3. 0 user manual. **International Agency for Research on Cancer, Lyon, France**, 2012.

PLUMMER, M. **rjags: Bayesian Graphical Models using MCMC**, 2018. R package version 4-8.

R CORE TEAM. **R: A Language and Environment for Statistical Computing**. R Foundation for Statistical Computing, Vienna, Austria, 2018.

RAWLINGS, J. O.; PANTULA, S. G.; DICKEY, D. A. **Applied Regression Analysis: A Research Tool, Second Edition**. , 1998. 1v.

ROBERT, C. P.; GEORGE, C. **Use R ! introducing monte carlo methods with R**. , 2010. 283p.

ROYSTON, P.; ALTMAN, D. G. Regression Using Fractional Polynomials of Continuous Covariates: Parsimonious Parametric Modelling. **Journal of the Royal Statistical Society**, v.43, n.3, p.429–467, 1994.

ROYSTON, P.; SAUERBREI, W. **Multivariable Model-Building: A Pragmatic Approach to Regression Analysis based on Fractional Polynomials for Modelling Continuous Variables**. England: Wiley, 2008. 1–303p.

SAUERBREI, W.; ROYSTON, P. 3 The Multivariable Fractional Polynomial Approach , with Thoughts about Opportunities and Challenges in Big Data. p.36–54, 2017.

SHKEDY, Z.; AERTS, M.; MOLENBERGHS, G.; BEUTELS, P.; VAN DAMME, P. Modelling age-dependent force of infection from prevalence data using fractional polynomials. **Statistics in Medicine**, v.25, n.9, p.1577–1591, 2006.

STAMEY, T. A.; KABALIN, J. N.; MCNEAL, J. E.; JOHNSTONE, I. M.; FREIHA, F.; REDWINE, E. A.; YANG, N. Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate. II. Radical prostatectomy treated patients. **The Journal of urology**, v.141, n.5, p.1076–1083, 1989.

STAN DEVELOPMENT TEAM. Stan Modeling Language Users Guide and Reference Manual. p.1–488, 2017.

STAN DEVELOPMENT TEAM. RStan: the R interface to Stan, 2018. R package version 2.18.2.

STROPA, A. A.; TRINCA, L. A. Maturity evaluation in brown spider females: testing two models. **Journal of Health and Environmental Sciences**, v.1, p.3–11, 2005.