UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
DEPARTAMENTO ACADÊMICO DE MATEMÁTICA

DAIANE PRISCILA SAMPAIO BUSSOLA

ESTUDO DO NÚMERO DE MAMOGRAFIAS E INCIDÊNCIA DE CÂNCER DE MAMA NO PARANÁ

TRABALHO DE CONCLUSÃO DE CURSO

CORNÉLIO PROCÓPIO

2017

DAIANE PRISCILA SAMPAIO BUSSOLA

ESTUDO DO NÚMERO DE MAMOGRAFIAS E INCIDÊNCIA DE CÂNCER DE MAMA NO PARANÁ

Trabalho de Conclusão de Curso apresentado ao Departamento Acadêmico de Matemática da Universidade Tecnológica Federal do Paraná como requisito parcial para obtenção do grau de "Licenciado em Matemática"

Orientador: Prof. Dr. Roberto Molina de Souza

CORNÉLIO PROCÓPIO

2017

Ministério da Educação
Universidade Tecnológica Federal do Paraná
Câmpus Cornélio Procópio
Diretoria de Graduação
Departamento de Matemática
Curso de Licenciatura em Matemática

FOLHA DE APROVAÇÃO

BANCA EXAMINADORA

_____

Roberto Molina de Souza
(orientador)

_____

Emílio Augusto Coelho Barros

_____

Glaucia Maria Bressan

"A Folha de Aprovação assinada encontra-se na Coordenação do Curso"

Aos meus pais, Sidnei e Madalena, aos meus irmãos, Matheus e Isabelle, que com muito carinho e apoio, não mediram esforços para que eu chegasse até esta etapa de minha vida.

AGRADECIMENTOS

Primeiramente a Deus por ser essencial em minha vida, autor do meu destino, meu guia, socorro presente na hora da angústia, por me guardar todos os dias, e não somente nestes anos como universitária.

A esta universidade, por todas as oportunidades, me proporcionando uma formação científica e cidadã.

Ao professor Roberto Molina de Souza por aceitar ser meu orientador, por todas as oportunidades que me proporcionou, pela orientação, apoio e confiança na pesquisa. Por ser um verdadeiro "pai acadêmico"durante nosso período de pesquisa juntos.

Aos professores da banca examinadora, pelas correções e contribuições para com este trabalho.

Ao professor André Martinez pela orientação no TCC 1 e pela valiosa contribuição que ajudou ao desenvolvimento desta pesquisa. À professora Elenice Weber pela orientação no TCC 2 e pela colaboração ao longo do curso.

Agradeço a todos os professores por me proporcionar o conhecimento não apenas racional, mas a manifestação de caráter no processo de formação profissional, por tanto que se dedicaram, não somente para ensinar, mas por terem me feito aprender.

Aos meus pais, Sidnei e Madalena, pelo amor, incentivo nas horas difíceis e apoio em todos os momentos. Aos meus irmãos, Matheus e Isabelle, que nos momentos de minha ausência dedicados ao estudo, sempre entenderam que o futuro é feito a partir da constante dedicação no presente!

Aos meus irmãos de quatro patas, Pluto e Lizzy, que sempre me receberam com amor incondicional, me alegrando até nos dias mais difíceis.

À minha avó Luzia e minha priminha Ana Lívia por todas as contribuições, amo vocês.

Meus sinceros agradecimentos aos amigos Anderson, Alisson, Carlos, Débora, Giovanna, Luiz Otávio, Paulo, Raphael e Renata, pelo companheirismo e cumplicidade, vocês continuarão presentes em minha vida sempre. Aos demais amigos que não citados, meus sinceros agradecimentos, saibam que são importantes para mim.

A todos que direta ou indiretamente fizeram parte da minha formação, o meu muito obrigada.

“Knowing is not enough; we must apply.
Willing is not enough; we must do.”

—Goethe

RESUMO

BUSSOLA, Daiane Priscila Sampaio. ESTUDO DO NÚMERO DE MAMOGRAFIAS E INCIDÊNCIA DE CÂNCER DE MAMA NO PARANÁ. 64 f. Trabalho de Conclusão de Curso – Departamento Acadêmico de Matemática, Universidade Tecnológica Federal do Paraná. Cornélio Procópio, 2017.

O câncer de mama, por ser um tipo de câncer que apresenta sinais visíveis quando em estágio avançado, tem sido objeto de estudo desde a antiguidade. Atualmente, nos anos 90, iniciou-se o movimento de conscientização e combate ao câncer de mama, chamado Outubro Rosa, com a fita rosa como símbolo, sendo até hoje uma ação mundial na luta contra este tipo câncer. Apesar dos esforços dos programas de prevenção e combate, o câncer de mama é o que mais acomete as mulheres no Brasil, excluindo o câncer de pele não melanoma. No sentido de contribuir no entendimento de algumas variáveis associadas ao câncer de mama, o objetivo deste trabalho é analisar a razão da quantidade de mamografias pelo número de mulheres realizadas nos municípios do Paraná e a razão entre o número de nódulos e o número de mamografias no período de junho de 2009 a junho de 2013, buscando possíveis relações com as variáveis explicativas grau de urbanização, PIB, renda média e taxa do número de médicos pela população feminina. Utilizando modelos de regressão linear múltiplos na presença de efeitos aleatórios sob o enfoque Bayesiano e frequenstista, observou-se evidências ao nível de $0,05$ de significância que quanto menor o PIB maior a razão do número de nódulos pelo número de exames encontrados em uma região.

Palavras-chave: Câncer de mama, Inferência Bayesiana, Modelos de Regressão

ABSTRACT

BUSSOLA, Daiane Priscila Sampaio. STUDY OF THE NUMBER OF MAMMOGRAMS AND INCIDENCE OF BREAST CANCER IN PARANÁ. 64 f. Trabalho de Conclusão de Curso – Departamento Acadêmico de Matemática, Universidade Tecnológica Federal do Paraná. Cornélio Procópio, 2017.

The breast cancer for being a type of cancer that show visible signs when in an advanced stage has been object of study since antiquity. In the $90's$, began the movement to raise awareness and fight against breast cancer, called Pink October, with the Pink Ribbon as symbol and is still a worldwide action in the fight against this cancer. Despite the efforts of prevention and combat programs, breast cancer is the type that most affects women in Brazil, excluding non-melanoma skin cancer. In order to contribute to the understanding of some variables associated with breast cancer, the objective of this study is to analyze the ratio between the number of mammograms and the number of women in the regions of Paraná and the ratio between the number of nodules and the number of mammograms in the period from June 2009 to June of 2013, seeking possible relations with the explanatory variables degree of urbanization, GDP, gross income and number of doctors. Using multiple linear regression models in the presence of random effects under the Bayesian and frequentist approach, it was observed evidence at the 0.05 level of significance that the lower the GDP the higher the ratio of the number of nodules by the number of tests found in a region.

Keywords: Breast cancer, Bayesian inference, Regression Model

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

INCA	Instituto Nacional do Câncer José Alencar Gomes da Silva
Fiocruz	Fundação Oswaldo Cruz
Inamps	Instituto Nacional de Assistência Médica da Previdência Social
PNCC	Programa Nacional de Controle do Câncer
PAISM	Programa de Assistência à Saúde da Mulher
SUS	Sistema Único de Saúde
SISMAMA	Sistema de Informação do Controle do Câncer de Mama
Iarc	International Agency for Research on Cancer
Sesa	Secretaria de Saúde do Paraná
DATASUS	Departamento de Informática do SUS
IPARDES	Instituto Paranaense de Desenvolvimento Econômico e Social
MMQ	Método dos Mínimos Quadrados
MCMC	Markov Chain Monte Carlo

SUMÁRIO

1 INTRODUÇÃO

O câncer de mama é um tipo de câncer que, quando em estágio avançado, apresenta sinais visíveis. Por esse motivo é uma enfermidade que foi registrada por médicos e conhecida da humanidade desde a antiguidade. Um breve histórico é apresentado abaixo segundo Mandal (2013).

Durante muitos anos foi tabu e vergonha falar sobre o câncer de mama, tanto que a detecção e diagnósticos eram raros. O advento de mulheres falando abertamente sobre a doença é um fenômeno recente. A partir dos anos 90, com o símbolo do câncer de mama, a fita rosa, iniciou-se a revolução no combate desse tipo de câncer.

Há mais de 3.500 anos, antigos egípcios descobriram e documentaram o câncer de mama que, para eles, era uma doença incurável. Em 460 a.C., Hipócrates a descreveu como uma doença humoral, ou seja, causada pelo desequilíbrio entre um dos quatro humores que compunham o corpo humano: sangue, fleuma, bile amarela e bile negra. Para ele, o câncer era causado pelo excesso de bile negra.

Em 200 d.C., o médico Cláudio Galeno sugeriu um tratamento que envolvesse ópio, óleo de rícino, alcaçuz, enxofre e pomadas. Nessa época, devido ao descobrimento tardio, o câncer afetava o corpo inteiro.

A crença de que o câncer era excesso de bile negra permaneceu até meados do século XVII, quando em 1680 médicos franceses começaram a contestar tal crença. A partir de então, surgiram muitas teorias sobre as possíveis causas da doença entre elas, a falta de filhos, a depressão, o sedentarismo e várias outras que relacionavam às questões da sexualidade feminina.

No final do século XVIII, o médico francês Henri Le Dran sugeriu a remoção cirúrgica do tumor para ajudar a tratar o câncer de mama. Esse fato levou à criação da mastectomia radical, no século XX. Essa cirurgia consiste em remover a mama, os músculos peitorais e a maioria dos linfonodos inferiores, médios e superiores. Hoje, este tipo de cirurgia é raramente realizada, por ser o procedimento mais agressivo ao corpo (MANDAL, 2014).

As teorias sobre o câncer avançaram no século XX. Mandal (2013) relata que em 1955, George Crile sugeriu que o câncer não era apenas localizado, mas que na verdade propagava-se por todo o corpo. Bernard Fisher descobriu a capacidade de metástase do câncer e, em 1976, publicou resultados de um método mais simples de cirurgia seguido por radiação e quimiotera-

pia, no qual notou a eficácia tanto quanto a mastectomia radical.

Para auxiliar na detecção do câncer de mama, em 1913 começaram a ser realizadas as radiografias das mamas. Mais tarde, em 1960 foi desenvolvido o mamógrafo, um aparelho que é um tipo especial de raio-X para as mamas. A mamografia é um exame capaz de identificar nódulos, mesmo antes de serem palpáveis, porém a confirmação de câncer de mama é realizada pela biópsia (INCA, 2015a). Desde 1976, a mamografia tem sido o método de escolha para detecção precoce (INCA; FIOCRUZ, 2014).

Com o avanço da medicina, em meados dos anos 90, menos de 10% das mulheres com câncer de mama passavam por mastectomia. Nessa época também surgiram novos tratamentos, como terapia hormonal e biológica. Atualmente, cientistas isolaram os genes causadores do câncer de mama, sendo eles: BRCA1, BRCA2 e ATM, para facilitar o tratamento precoce.

No Brasil, segundo dados do INCA e Fiocruz (2014), até a década de 70, a política pública para o controle do câncer de mama restringia-se a tratamentos e cirurgias realizados pela medicina previdenciária do Instituto Nacional de Assistência Médica da Previdência Social (Inamps). Em 1973, foi criado o Programa Nacional de Controle do Câncer (PNCC), iniciativa pioneira com foco em cânceres femininos, ofertando exames e ações de prevenção.

Devido à pressão e participação do movimento de mulheres, o Ministério da Saúde, em 1984, criou o Programa de Assistência à Saúde da Mulher (PAISM), esse programa incluía ações educativas para detecção precoce. Com o objetivo de ampliar a informação e prevenção dos cânceres femininos, foi lançado em 1987, em parceria do Ministério da Saúde e o Inamps, o programa Pró-Onco. No ano seguinte, foi criado o Sistema Único de Saúde (SUS), tornando-se mais abrangentes e em nível nacional as ações de controle (INCA; FIOCRUZ, 2014).

Nos anos 90, o movimento Outubro Rosa foi introduzido para estimular a participação da população na luta contra o câncer de mama, tendo como símbolo o laço cor de rosa. Nessa época, no Brasil, aconteceu o lançamento do Programa Viva Mulher, ação nacional organizada para o controle dos cânceres do colo do útero e mama.

A partir dos anos 2000, diversos materiais educativos para profissionais começaram a ser elaborados, como diretrizes e cartilhas. Também foram lançadas políticas direcionadas ao combate do câncer de mama. Entre elas estão a Política Nacional de Atenção Oncológica (2005) e o Pacto pela Saúde (2006). Em 2009, foi implantado o Sistema de Informação do Controle do Câncer de Mama (SISMAMA).

Ao longo dos últimos seis anos muitas campanhas e programas que influenciam a detecção precoce com vistas a reduzir a mortalidade por câncer de mama têm sido amplamente divulgados. A mais recente publicação é do livro sobre as “Diretrizes para detecção precoce do Câncer de Mama no Brasil” e também, a cartilha “Câncer de Mama: é preciso falar disso”,

ambos lançados pelo Ministério da Saúde. Dentre os novos programas, por meio de uma portaria foi instituído o Programa Nacional de Qualidade da Mamografia, para melhorar a qualidade dos exames realizados.

Apesar dos esforços dos programas de prevenção e combate, o câncer de mama é o tipo de câncer que mais acomete as mulheres no Brasil, excluindo o câncer de pele não melanoma (INCA, 2015b). De acordo com dados da *International Agency for Research on Cancer* (Iarc), o risco acumulado durante a vida de uma pessoa ter e morrer de câncer de mama no Brasil é de 6,4% e 1,6%, respectivamente.

No estado do Paraná, de acordo com dados da Secretaria de Saúde do Paraná (Sesa) no final dos anos 90 e início de 2000, foram registrados aproximadamente 768 casos de morte por câncer de mama. Desde então, os programas de prevenção e diagnóstico precoce se intensificaram.

Considerando o câncer de mama um problema de saúde pública no Brasil, paralelamente as campanhas de conscientização ocorridas em todo o mês de outubro, despertou-se a motivação em estudar este tema no trabalho de conclusão de curso. Além da contribuição a partir da revisão bibliográfica e metodológica, este trabalho apresenta uma aplicação com dados reais considerando casos de câncer de mama no estado do Paraná.

Rodrigues et al. (2015) estudar as interrelações entre a prevenção do câncer de mama e os fatores socioeconômicos, demográficos, comportamentais, regionais e de saúde na determinação da frequência temporal à busca por prevenção via realização de mamografias e exames de mama no Brasil. Achcar (2016) descrever a distribuição temporal da doença para os casos novos notificados nos municípios com mais de 200.000 habitantes no Estado de São Paulo (centros regionais), no período de janeiro de 2009 a dezembro de 2013, além de identificar algumas covariáveis socioeconômicas, demográficas e de saúde, que podem estar associadas com a doença e utilizar estas variáveis em diferentes modelos estatísticos.

Logo, no sentido de contribuir no entendimento de algumas variáveis associadas ao câncer de mama, o objetivo geral deste trabalho é analisar a razão do número de mamografias pelo número de mulhers realizadas nos municípios do Paraná e a razão entre o número de nódulos e o número de mamografias no período de junho de 2009 a junho de 2013, a partir de dados do Departamento de Informática do SUS - DATASUS, buscando possíveis relações com as covariáveis: grau de urbanização, PIB (Produto Interno Bruto), renda média e taxa do número de médicos pela população feminina.

Para atingir o objetivo geral, foram propostos os seguintes objetivos específicos:

- Uma breve revisão de literatura sobre o câncer de mama;
- Revisão de alguns conteúdos da teoria de regressão linear múltipla e inferência (frequen-

tista e Bayesiana);

- Construção de um conjunto de dados com o número de mamografias realizadas e casos de câncer no Paraná;
- Apresentação dos resultados utilizando métodos de inferência frequentista e Bayesiana nas aplicações;
- Interpretação e discussão dos resultados.

Este trabalho está organizado da seguinte maneira: no próximo Capítulo será apresentada a construção do conjunto de dados. No Capítulo 3, são apresentados os métodos estatísticos de inferência fequentista e Bayesiana. No Capítulo 4 são realizadas as análises dos dados segundo os modelos propostos. Finalmente, no Capítulo 5, encontra-se as conclusões e considerações finais.

2 BANCO DE DADOS

Para coletar os dados com a quantidade de mamografias realizadas e casos notificados de câncer de mama em mulheres em todo o estado do Paraná, mensalmente no período de junho/2009 a junho/2013, foram acessadas informações do banco de dados originários do DATASUS.

Os dados das covariáveis que indicam o nível sócio econômico das regiões do estado do Paraná, o grau de urbanização, o PIB real per capita, a renda média e o grau de urbanização, foram obtidos do *site* oficial do IPARDES (Instituto Paranaense de Desenvolvimento Econômico e Social).

Os dados coletados foram obtidos das fontes abaixo:

- SISMAMA

  Banco de dados integrado ao DATASUS, o SISMAMA é específico para buscas relacionadas ao câncer de mama em todo o Brasil. Nele, encontram-se as seguintes informações: quantidade de exames realizados, exames sem achados, exames contendo nódulo na mama esquerda ou direita, exames realizados em mulheres sem cirurgia anterior, exames com microcalcificação na mama direita ou esquerda e exames que diagnosticaram linfonodos nas axilas direita ou esquerda. Para este estudo serão considerados apenas a quantidade de exames realizados e os exames em que foram diagnosticados nódulos em uma das mamas.

- IPARDES - Instituto Paranaense de Desenvolvimento Econômico e Social

  O IPARDES disponibiliza um menu de pesquisas chamado ”Perfil avançado das regiões geográficas”, nele é possível selecionar a região de interesse e encontrar informações que possibilitam obter um perfil das regiões, como dados econômicos e demográficos provenientes de levantamentos feitos pelo último senso do estado.
  As covariáveis consideradas para este estudo são: população feminina, PIB real per capita, renda média e grau de urbanização.

Foram considerados para este estudo as regiões geográficas do Paraná, definidas pela Lei Estadual n°15.825/08 de 28/04/2008, disponibilizadas p elo IPARDES. Atualmente, são 10 regiões geográficas: Noroeste, Centro Ocidental, Norte Central, Norte Pioneiro, Centro Oriental, Oeste, Sudoeste, Centro-Sul, Sudeste e Metropolitana de Curitiba. No Apêndice A, estão detalhadas cada uma das regiões, com suas respectivas cidades.

3 MÉTODOS ESTATÍSTICOS

O delineamento deste estudo é considerado epidemiológico observacional analítico ecológico. É epidemiológico, pois estuda a distribuição e os determinantes das doenças ou condições relacionadas à saúde em populações especificadas; observacional analítico são os delineamentos para examinar a existência de associação entre uma exposição e uma doença ou condição relacionada à saúde. Um dos principais delineamentos de estudo analítico é o ecológico, em que possibilita examinar associações entre exposição e doença/condição relacionada na coletividade (LIMA-COSTA; BARRETO, 2003).

Para as aplicações serão utilizados modelos de regressão linear múltipla, que permitem verificar possíveis relações entre as variáveis respostas referente ao número de mamografias e números de nódulos, com as covariáveis demográficas PIB real per capita, renda média, grau de urbanização e taxa do número de médicos pela população feminina. Serão aplicados modelos estatísticos sob a metodologia frequentista e Bayesiana.

A seguir, será apresentada uma breve revisão de análise de regressão, as ideias gerais de inferência frequentista e Bayesiana e seleção de modelos sob estes dois enfoques.

3.1 ANÁLISE DE REGRESSÃO

A análise de regressão é uma técnica utilizada para verificar a existência de relação entre uma variável dependente com uma ou mais variáveis independentes. Consiste em buscar uma equação que possa explicar a alteração da variável dependente pela alteração dos níveis das variáveis independentes. Em qualquer sistema em que quantidades de variáveis podem mudar, é de grande interesse examinar os efeitos que algumas variáveis exercem (ou aparentam exercer) sobre a outra (DRAPER; SMITH, 1998).

Espera-se que possa existir algum tipo de relação entre as variáveis. Essa relação geralmente é expressada na forma de uma função de relação e em alguns casos essa função é muito complicada de entender ou descrever em termos simples. Nesses casos, pode-se aproximar a uma função mais simples matematicamente, como uma função polinomial, que contenha as variáveis apropriadas. Assim, na análise de regressão linear os dados são modelados utilizando

funções de previsão e os parâmetros são desconhecidos e estimados a partir dos dados. Esse modelo é chamado de modelo linear.

Esta relação pode ser analisada como um processo. Neste processo, os valores de $X_1, X_2, \ldots, X_p$ são chamados de Variáveis Regressoras[1] (*inputs*) e Y de Variável Resposta[2] (*output*) (DRAPER; SMITH, 1998).

Na Regressão Linear, quando o interesse é a relação de apenas uma variável de entrada com a variável de resposta, tem-se a chamada Regressão Linear Simples. Mas, quando utiliza-se mais de uma variável explanatória para predizer o comportamento de uma variável resposta, passa-se a nomeá-lo como modelo de regressão linear múltiplo (DRAPER; SMITH, 1998). Para esse trabalho, serão explorados os conceitos relacionados à Regressão Linear Múltipla.

Dada a vasta praticidade do método de Regressão Linear, as aplicações podem ser analisadas sob dois pontos de vista:

- Predição: Os modelos de regressão linear podem ser utilizados para fazer uma previsão do valor de y para um conjunto de dados observados de valores de Y e X.
- Regressão: Dadas variáveis ou covariáveis X_1,..., X_p, que podem estar relacionadas com a resposta ou variável dependente Y, a análise de regressão pode ser utilizada para estabelecer a magnitude da relação entre Y e $X_j, j = 1, ..., p$.

Matematicamente, regressão linear simples é o processo no qual estima-se os parâmetros (intercepto e coeficiente angular) de uma função $f(X)$. Estes parâmetros determinam as características da função que relaciona Y e X, que no modelo linear é representado pela reta de regressão. É essa reta que explica de forma geral a relação entre as variáveis. Isto implica que os valores observados de Y nem sempre serão iguais aos valores de $\hat{Y}$ estimados pela reta de regressão. Haverá sempre uma diferença, que é chamada de erro ou desvio. De modo geral, de acordo com (DRAPER; SMITH, 1998), essa relação de variáveis será representada da seguinte forma:

$$\text{Variável Resposta} = \text{Função Modelo} + \text{Erro}$$

Para ajustar os modelos de regressão, utiliza-se com frequência o Método dos Mínimos Quadrados (MMQ) que é um método matemático pelo qual se estima o intercepto e o coeficiente angular de uma reta de regressão. Ele definirá uma reta que minimizará a soma das distâncias ao quadrado entre os pontos plotados (X, Y) e a reta (X, Y'). Um exemplo deste método para o

[1] Também chamadas de variáveis de entrada, variáveis preditoras, X-variáveis ou variáveis independentes.
[2] Também chamada de variáveis de saída, Y-variáveis ou variáveis dependentes.

modelo de regressão linear simples é apresentado no Apêndice B.1.

A forma mais geral do modelo de regessão linear múltiplo para as variáveis $X_1, X_2, \ldots, X_k$, pode ser escrito como 1

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_p X_{pi} + \varepsilon_i \tag{1}$$

em que Y_i é o valor observado para a variável resposta Y na i-ésima unidade amostral; β_0 é a constante de regressão e representa a interseção da reta com o eixo de Y; $\beta_1, \beta_2, ..., \beta_p$ são os coeficientes de regressão, que representam a variação de Y em função da variação de uma unidade das covariáveis; $X_{1i}, X_{2i}, ..., X_{pi}$ são as covariáveis referentes a i-ésima unidade amostral com $i = 1, 2, \ldots, n$, para n indivíduos; ε_i é o erro associado ao modelo, que representa a distância entre os valores preditos pela reta e os valores observados.

Suponha que o modelo considerado seja escrito na forma 2

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{2}$$

em que, $\mathbf{Y}$ é um vetor de observação $(n \times 1)$, $\mathbf{X}$ é uma matriz de formas conhecidas $(n \times p)$, $\boldsymbol{\beta}$ é o vetor de parâmetros $(p \times 1)$, $\boldsymbol{\varepsilon}$ é o vetor de erros $(n \times 1)$. Assume-se que $E(\boldsymbol{\varepsilon_i}) = \mathbf{0}$ e $V(\boldsymbol{\varepsilon_i}) = \mathbf{I}\sigma^2$ (independência entre os elementos de $\boldsymbol{\varepsilon}$). Logo, na forma matricial o modelo pode ser escrito como 3

$$E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta} \tag{3}$$

A soma dos quadrados dos erros será dada por 4

$$\begin{aligned} \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} &= (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) \\ &= \boldsymbol{Y}'\boldsymbol{Y} - \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{Y} - \boldsymbol{Y}'\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} \\ &= \boldsymbol{Y}'\boldsymbol{Y} - 2\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{Y} + \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}. \end{aligned} \tag{4}$$

Isso acontece devido ao fato de que $\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{Y}$ é uma matriz 1×1, ou um escalar, que transpõe $(\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{Y})' = \boldsymbol{Y}'\boldsymbol{X}\boldsymbol{\beta}$ e deve ter o mesmo valor. A estimativa de mínimos quadrados de $\boldsymbol{\beta}$ é o valor $\mathbf{b}$, podendo ser determinado derivando a equação (4) em relação $\boldsymbol{\beta}$, igualando a equação da matriz resultante a zero e, ao mesmo tempo, substituindo $\boldsymbol{\beta}$ por $\mathbf{b}$. Isso fornece a equação 5

$$(\boldsymbol{X}'\boldsymbol{X})\boldsymbol{b} = \boldsymbol{X}'\boldsymbol{Y} \tag{5}$$

Para **X'X** não singular, sua inversa existe. Logo, a solução da equação 5 é dada por 6

$$\boldsymbol{b} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} \tag{6}$$

Esta solução tem as seguintes propriedades:

1. É uma estimativa de $\boldsymbol{\beta}$ que minimiza o erro da soma dos quadrados $\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}$ independente de quaisquer propriedades de distribuição dos erros;

2. Os elementos de $\mathbf{b}$ são funções lineares das observações de $Y_1, Y_2, \ldots, Y_n$ e fornecem estimativas imparciais dos elementos de $\boldsymbol{\beta}$ que possuem variância mínima, independente das propriedades de distribuição dos erros;

3. Se os erros são independentes, então $\mathbf{b}$ é um estimador de máxima verossimilhança de $\boldsymbol{\beta}$. Em termos de vetores, pode-se escrever $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \boldsymbol{I}\sigma^2)$, significando que $\boldsymbol{\varepsilon}$ segue uma distribuição normal n-dimensional multivariada com $E(\boldsymbol{\varepsilon}) = \mathbf{0}$, em que $\mathbf{0}$ denota o vetor contido apenas de zeros e mesmo comprimento de $\boldsymbol{\varepsilon}$ e $\boldsymbol{V}(\boldsymbol{\varepsilon}) = \mathbf{I}\sigma^2$; isto é, $\boldsymbol{\varepsilon}$ tem uma matriz de variância-covariância cujos elementos da diagonal, $V(\varepsilon_i), i = 1, 2, \ldots, n$, são todos σ^2 e elementos fora da diagonal, covariância $(\varepsilon_i, \varepsilon_j)$ para $i \neq j = 1, \ldots, n$, são todos zero. A função de verossimilhança para a amostra $Y_1, Y_2, \ldots, Y_n$ é definida, neste caso, como o produto 7

$$\prod_{i=1}^{n} \frac{1}{\sigma(2\pi)^{1/2}} e^{-\varepsilon_i^2/2\sigma^2} = \frac{1}{\sigma^n(2\pi)^{n/2}} e^{-\varepsilon'\varepsilon/2\sigma^2} \tag{7}$$

Assim, para um valor fixo de σ, maximizar a função de verossimilhança é equivalente a minimizar a quantidade $\varepsilon'\varepsilon$. Esse fato fornece uma justificativa para o uso do MMQ, porque em várias situações a suposição de que os erros são distribuídos normalmente é bastante sensata. Um exemplo da obtenção de estimadores de máxima verossimilhança é apresentado no Apêndice B.2.

Para a construção de intervalos com $100 \times (1-\alpha)\%$ de confiança $\boldsymbol{\beta}$ tem-se 8

$$\mathbf{b} \sim \mathbf{N}\left(\beta, \left(\mathbf{X}'\mathbf{X}\right)^{-1}\sigma^2\right) \tag{8}$$

Assim, para cada parâmetro separado, o intervalo de confiança é dado por 9

$$b_i \pm t_{\left(v, 1-\frac{\alpha}{2}\right)} \sqrt{V\left(b_i\right)} \tag{9}$$

em que $V(b_i) = (\mathbf{X}'\mathbf{X})^{-1}s^2$, sendo s^2 a estimativa da variância σ^2. $t_{(v,1-\alpha/2)}$ é o quantil da distribuição t de Student com parâmetro v graus de liberdade e nível de significância α. v é dado pelo número de observações menos o número de parâmetros estimados no modelo.

Para um nível de significância α, geralmente fixado em 5%, pode-se testar a hipótese de interesse $\beta_i = 0$ contra $\beta_i \neq 0$. Logo, deve-se verificar se o valor 0 está contido no intervalo de confiança dado em (9). Caso o valor 0 não esteja contido no intervalo $100 \times (1-\alpha)\%$ de

confiança, rejeita-se a hipótese de que $\beta_i = 0$ e conclui-se que existem evidências ao nível α de significância que a variável X influencia ou ajuda a explicar Y.

Para observações medidas ao longo do tempo, é usual introduzir uma componente aleatória para capturar a possível correlação que exista entre estas observações. Logo, o modelo linear de efeitos mistos geralmente é utilizado como uma extensão do modelo linear (MCLEAN et al., 1991).

Na forma matricial, o modelo linear de efeitos mistos pode ser escrito como 10

$$\boldsymbol{y} = \boldsymbol{XB} + \boldsymbol{ZU} + \boldsymbol{\varepsilon} \tag{10}$$

em que **y** é $m \times 1$ é o vetor coluna de resposta; **X** é uma matriz $m \times p$ com posto conhecido de $\mathbf{X} \leq m, p$; **B** é um vetor $p \times 1$ de efeitos fixos que são desconhecidos; **Z** é uma matriz $m \times q$; **U** é um vetor $q \times 1$ de efeitos aleatórios com $E(\mathbf{U}) = \mathbf{0}$; ε é um vetor $m \times 1$ com $E(\varepsilon) = 0$.

3.1.1 AKAIKE INFORMATION CRITERION (AIC)

Quando é necessário comparar mais de um modelo proposto para o mesmo conjunto de dados em termos de ajustes, o *Akaike Information Criterion (AIC)* (AKAIKE, 1974) é um critério de informação muito utilizado em seleção de modelos. Foi desenvolvido pelo estatístico Akaike em 1974, que utilizou a estimativa da informação de Kullback-Leibler, baseada na Função de Log-Verossimilhança (FLV) em seu ponto máximo, acrescida de uma penalidade associada ao número de parâmetros do modelo.

Desse modo, o *AIC* é definido como:

$$AIC = -2\sum_{i=1}^{n} lnL(\widehat{\mu}_i, y_i) + 2(p)$$

em que y_i é o i-ésimo valor da resposta e $\widehat{\mu}_i$ é a estimativa de y_i, quando se ajusta um modelo de parâmetros p por meio da maximização da FLV. O termo que se adiciona à FLV, chamado de função de penalidade, tem a finalidade de corrigir um viés proveniente da comparação de modelos de diferentes números de parâmetros. Entre vários modelos candidatos, deve ser escolhido aquele que apresentar o menor valor de *AIC*.

3.2 INFERÊNCIA BAYESIANA

Uma outra forma de se estimar os parâmetros de um modelo de regressão é utilizando a metodologia Bayesiana. A base da inferência Bayesiana é a fórmula de Bayes, que associa os dados com a informação *a priori*, obtendo-se a distribuição *a posteriori*. Esta, combina a

informação *a priori* com a informação dos dados. A partir da distribuição *a posteriori*, obtém-se as quatidades aleatórias relativas aos parâmetros de interesse (estimativas dos parâmetros).

3.2.1 TEOREMA DE BAYES

Antes de apresentar o Teorema de Bayes é necessário abordar o conceito da probabilidade condicional, que é um dos conceitos mais importantes da teoria de probabilidade. É utilizado quando se possui uma informação parcial a respeito do resultado de um experimento. As probabilidades condicionais podem ser utilizadas para computar mais facilmente as probabilidades desejadas.

Definição 3.2.1 (Probabilidade Condicional) *Seja* $(\omega, \mathbb{A}, P)$ *um espaço de probabilidade. Se* $B \in \mathbb{A}$ *e* $P(B) > 0$*, a probabilidade de A dado B é definida por 11*

$$P(A|B) = \frac{P(AB)}{P(B)}, \qquad A \in \mathbb{A}. \tag{11}$$

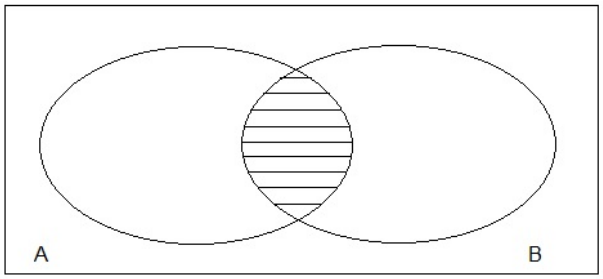


Figura 1: Diagrama de Venn para Probabilidade Condicional

Observe na Figura 1 que, se o evento B ocorrer, para que A ocorra, é necessário que a ocorrência real seja um ponto tanto em A quanto em B, isto é, ela deve estar em AB ($A \cap B$). Considerando que B ocorreu, tem-se que B se torna o novo, ou reduzido, espaço amostral. Logo, a probabilidade de que o evento AB ocorra será igual à probabilidade de AB relativa à probabilidade de B. A seguir, serão apresentados o Teorema da Probabilidade Composta e Teorema da Probabilidade Total (JAMES, 2010; CONTI, 2010).

Teorema 3.2.1 (Teorema da Probabilidade Composta) *Seja* $(\omega, \mathbb{A}, P)$ *um espaço de probabilidade. Então*

(i) $P(A \cap B) = P(A)P(B|A) = P(B)P(A|B), \forall\, A, B \in \mathbb{A}$,

(ii) $P(A_1 \cap A_2 \cap \cdots \cap A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2)\ldots$
$\ldots P(A_n|A_1 \cap \cdots \cap A_{n-1}), \ldots, A_n \in \mathbb{A}, \quad \forall\, n = 2, 3, \ldots$

Teorema 3.2.2 (Teorema da Probabilidade Total) *Se a sequência (finita ou enumerável) de eventos aleatórios* $A_1, A_2, \ldots$ *formar uma partição de* ω*, então*

$$P(B) = \sum_i P(A_i)P(B|A_i), \quad \forall\, B \in \mathbb{A}.$$

A partir dos teoremas 3.2.1 e 3.2.2, pode-se calcular a probabilidade de A_i dada a ocorrência de B:

$$P(A_i|B) = \frac{P(A_i \cap B)}{P(B)} \Rightarrow P(A_i|B) = \frac{P(A_i)P(B|A_i)}{\sum_j P(A_j)P(B|A_j)} \tag{12}$$

A expressão dada em 12 é a *fórmula de Bayes*. Esta é útil quando se conhece as probabilidades dos A_i e a probabilidade condicional de B dado A_i, mas não se conhece diretamente a probabilidade de B.

No contexto da Metodologia Bayesiana, a fórmula de Bayes pode ser interpretada da seguinte forma: antes de se conhecer o evento A_i, atribui-se uma probabilidade *a priori* para A_i, dada por $P(A_i)$. Essa probabilidade é condicionada a partir a ocorrência do evento B. A Figura 2, trás a representação do teorema de Bayes em um espaço amostral Ω.

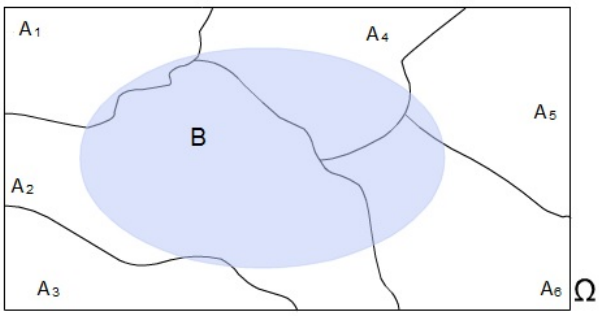


Figura 2: Representação do Teorema de Bayes

Assumindo uma amostra aleatória $y = (y_1, \ldots, y_n)$ independentes e identicamente distribuídos, suponha que o vetor aleatório Y possui uma distribuição conjunta dada pela função densidade de probabilidade $f(y|\theta)$, que também é conhecida como função de verossimilhança para θ. Uma vez que os dados foram observados e atribuídas distribuições *a priori* para θ, dada por $\pi(\theta)$, de 12 tem-se a distribuição *a posteriori* para θ_i dado y,

$$\pi(\theta_i|y) = \frac{f(y|\theta_i)\pi(\theta_i)}{\sum\limits_{j=i}^{k} f(y|\theta_j)\pi(\theta_j)}$$

Agora, suponha que o parâmetro θ assume valores contínuos num dado intervalo, considerando uma amostra aleatória $y = (y_1, \ldots, y_n)$, a distribuição *a posteriori* para θ dado y, pode ser escrita como

$$\pi(\theta|y) = \frac{f(y|\theta)\pi(\theta)}{\int f(y|\theta)\pi(\theta)d\theta}$$

em que a integral no denominador é definida no intervalo de variação de θ.

Em geral, não é necessário calcular a integral no denominador, de modo que a distribuição *a posteriori* é proporcional à distribuição *a priori* multiplicada pela verossimilhança, ou seja:

$$\pi(\theta \mid y) \propto f(y \mid \theta)\,\pi(\theta)$$

3.2.2 DISTRIBUIÇÃO *A PRIORI*

No enfoque Bayesiano, a distribuição *a priori* representa a escolha de uma distribuição de probabilidade e seus respectivos parâmetros (chamados de hiperparâmetros) com o que se conhece sobre os parâmetros a serem estimados, antes da avaliação dos dados. É necessário ter cautela ao definir a distribuição *a priori*, pois se a mesma não for definida corretamente, pode-se chegar a estimativas ruins. A distribuição *a priori* pode ser elicitada a partir de procedimentos subjetivos ou objetivos.

Uma distribuição *a priori* para um parâmetro θ pode ser descrita de várias formas: da opinião de um ou vários especialistas, a partir de estudos anteriores ou etapas anteriores do mesmo estudo ou de análises preliminares dos dados.

Na prática, quando não é possível obter informações para uma distribuição *a priori*, são utilizadas distribuições *a priori* aproximadamente não informativas, levando a resultados similares ao da inferência frequentista, pois utilizando uma distribuição *a priori* não informativa, a inferência baseia-se apenas na informação dos dados amostrais.

3.2.3 MEDIDAS RESUMO

Da distribuição *a posteriori* apresentada na equação (3.2.1), obtém-se as medidas resumo do parâmetro de interesse. Utilizando métodos de simulação, a partir da geração de valores aleatórios da distribuição do parâmetro de interesse, obtém-se um vetor de observações. Deste vetor, calcula-se as quantidades aleatórias pontuais, como a média ou a mediana, desvio padrão ou quartis e um intervalo de valores onde se espera que o parâmetro esteja contido, com um certo grau de credibilidade (intervalo de credibilidade).

Os intervalos de credibilidade são calculados de forma bastante intuitiva. Seja θ um

parâmetro unidimensional e sua distribuição condicional unimodal, um intervalo de credibilidade com probabilidade $(1-\alpha)$ é dado por (θ_*, θ^*) para

$$\int_{-\infty}^{\theta_*} \pi(\theta|y)d\theta = \frac{\alpha}{2}$$

e

$$\int_{\theta^*}^{\infty} \pi(\theta|y)d\theta = \frac{\alpha}{2}$$

O intervalo (θ_*, θ^*) é chamado de intervalo de credibilidade para θ com probabilidade $(1-\alpha)$, que pode ser comparado a ideia de intervalos de confiança da inferência frequentista.

Para um nível de significância α, geralmente fixado em 5%, pode-se testar a hipótese de interesse $\theta = a$ contra $\theta \neq a$. Logo, deve-se verificar se o valor a está contido no intervalo de credibilidade. Caso o valor a não esteja contido no intervalo $100 \times (1-\alpha)\%$ de confiança, rejeita-se a hipótese de que $\theta = a$.

3.2.4 MÉTODO DE SIMULAÇÃO

A palavra simulação se refere ao tratamento de um problema real por meio da reprodução em um ambiente, geralmente computacional, controlado pelo pesquisador (GAMERMAN; LOPES, 2006). Quando deseja-se resumir a informação descrita na distribuição *a posteriori*, dentre várias formas, pode-se utilizar os métodos baseados em simulação. Existem inúmeros métodos porém, para as aplicações deste trabalho, será utilizado o método de Monte Carlo via Cadeias de Markov (*Markov Chain Monte Carlo* - MCMC), que aplicado ao contexto Bayesiano resulta no Amostrador de Gibbs.

Uma cadeia de Markov é construída da seguinte maneira: suponha que o interesse seja gerar uma amostra de uma distribuição *a posteriori* $\pi(\theta|y)$, $\theta \in \mathbb{R}^k$. Logo, constrói-se uma cadeia de Markov, pois não é possível fazer diretamente a simulação. A cadeia possui espaço de estados no espaço paramétrico, que é o conjunto de todos os valores de θ, em que sua simulação é simples e a distribuição de equilíbrio seja dada por $\pi(\theta|y)$.

Se $\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(t)}$, é uma realização de uma cadeia, desse modo,

$$\theta^{(t)} \xrightarrow{t\to\infty} \theta \sim \pi(\theta|y)$$

Para estimar o valor esperado de g(θ) em relação a $\pi(\theta|y)$, ou seja,

$$E[g(\theta|y)] = \int g(\theta)\pi(\theta|y)d\theta$$

tem-se:

$$\frac{1}{t}\sum_{i=1}^{t} g(\theta^{(i)}) \xrightarrow{t\to\infty} E[g(\theta|y)] \qquad qc$$

em que "qc" significa convergência quase certa.

Na prática, $\theta^{(i)}$ pode estar correlacionado, mas pode-se considerar espaços adequados entre os $\theta^{(i)}$ gerados para garantir uma amostra aleatória de $\pi(\theta|y)$.

O Amostrador de Gibbs foi desenvolvido e tem sido aplicado principalmente no contexto de modelos estocásticos complexos, que envolvem um grande número de parâmetros. Nesses casos, a especificação direta da distribuição conjunta é inviável. Em vez disso, é especificado o conjunto completo de condicionais, geralmente assumindo que uma distribuição condicional depende apenas de algum subconjunto da "vizinhança" das variáveis (CHIB; GREENBERG, 1995; GELFAND; SMITH, 1990).

Por exemplo, suponha o interesse em obter inferências da distribuição a posteriori conjunta, $\pi(\theta|y), \theta = (\theta_1, \ldots, \theta_k)$. Para isso, simula-se quantidades aleatórias de distribuições condicionais completas $\pi(\theta_i|y, \theta_{(i)})$ que produzem uma cadeia de Markov, isto é, a cadeia irá sempre se mover para um novo valor, não existindo um mecanismo de aceitação-rejeição.

Se as distribuições condicionais forem completamente conhecidas, assume-se um conjunto inicial arbitrário de valores de $\theta_1^{(0)}, \theta_2^{(0)}, \ldots, \theta_k^{(0)}$ para o vetor de parâmetros θ. Assim, o algoritmo pode ser descrito como 13

$$(i) \quad Gerar \ \theta_1^{(1)} \ de \ \pi(\theta_1 \mid y, \theta_2^{(0)}, \ldots, \theta_k^{(0)}); \tag{13}$$

$$(ii) \quad Gerar \ \theta_2^{(1)} \ de \ \pi(\theta_2 \mid y, \theta_1^{(1)}, \theta_3^{(0)}, \ldots, \theta_k^{(0)});$$

$$(iii) \quad Gerar \ \theta_3^{(1)} \ de \ \pi(\theta_3 \mid y, \theta_1^{(1)}, \theta_2^{(1)}, \theta_4^{(0)}, \ldots, \theta_k^{(0)});$$

$$\vdots$$

$$(k) \quad Gerar \ \theta_k^{(1)} \ de \ \pi(\theta_k \mid y, \theta_1^{(1)}, \theta_2^{(1)}, \ldots, \theta_{k-1}^{(1)});$$

Assim, cada iteração será completa após k movimentos ao longo dos eixos coordenados das componentes de θ. Após a convergência, as resultantes formam uma amostra de $\pi(\theta)$.

Na geração de amostras de Gibbs deve-se considerar as primeiras iterações como período de aquecimento (*burn-in-samples*), que devem ser descartadas para eliminar o efeito de valores iniciais. Um exemplo com a aplicação do mesmo é apresentada no Apêdice B.3.

3.2.5 DEVIANCE INFORMATION CRITERION (DIC)

Quando é necessário comparar mais de um modelo proposto para o mesmo conjunto de dados em termos de ajustes, o *Deviance Information Criterion (DIC)* (SPIEGELHALTER et al., 2002) é um importante critério Bayesiano usado para a seleção de modelos. O *DIC* é similar ao *Akaike Information Criterion* apresentado na Seção 3.1.1. O *DIC* é dado por:

$$DIC = \widehat{D} + p_D$$

em que $\widehat{D}$ é a *deviance* (desvio) estimado *a posteriori* para os parâmetros de interesse e p_D é o número efetivo de parâmetros do modelo, dado por $p_D = D - \widehat{D}$, em que D é a média *a posteriori* da *deviance*. Quando dois valores de *DIC* são comparados, o menor valor de *DIC* implica no melhor modelo.

4 APLICAÇÃO

4.1 DADOS

Utilizando as bases de dados apresentadas na Seção 2, na Tabela 1 é apresentado um esboço do conjunto de dados utilizado para as análises. Nas Figuras 3 e 4 são apresentados as médias da razão do número de mamografias pelo tamanho da população feminina multiplicado por 10000 e as médias da razão do número de nódulos pelo número de mamografias multiplicado por 100, segundo as 10 regiões. Também são apresentados, nas figuras 5 e 6, gráficos de dispersão das variáveis grau de urbanização (GU), Produto Interno Bruto (PIB), renda média em salários mínimos (SM) e a taxa do número de médicos pela população feminina segundo as variáveis dependentes em estudo.

Tabela 1: Esboço do conjunto de dados.

Região	Tempo	$\frac{n^o\ Exames}{Pop.Fem.} \times 10000$	$\frac{n^o\ Nodulos}{n^o\ Exames} \times 100$	GU($\times 100$)	PIB($\times 10000$)	Renda(SM)	$\frac{n^o\ Médicos}{Pop.Fem.}$ ($\times 1000$)
Noroeste	06/2009	9,34	12,38	0,83	1,18	1,05	1,53
Noroeste	07/2009	36,68	13,16	0,83	1,18	1,05	1,53
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Noroeste	06/2013	1,54	3,85	0,80	1,35	1,02	1,41
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Met. Curitiba	06/2009	22,58	10,53	0,92	1,58	1,10	2,70
Met. Curitiba	07/2009	52,66	11,48	0,92	1,58	1,10	2,70
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Met. Curitiba	06/2013	34,35	13,88	0,92	2,51	1,10	3,19

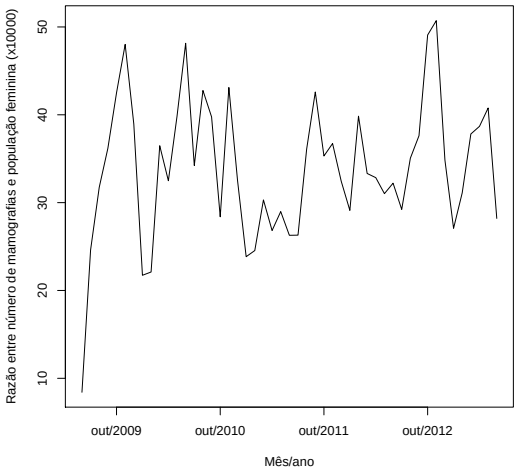


Figura 3: Média da razão do número de mamografias pelo tamanho da população feminina entre as regiões ao longo do período de estudo.

Observa-se na Figura 3 que nos meses de outubro de cada ano acontece um aumento nessas médias, sugerindo que a campanha Outubro Rosa no estado do Paraná são eficazes no período estudado. Na Figura 4, observa-se que a partir de um pico da razão do número de nódulos pelo número de exames no início de 2011, existe uma tendência de diminuição até o final do período de estudo.

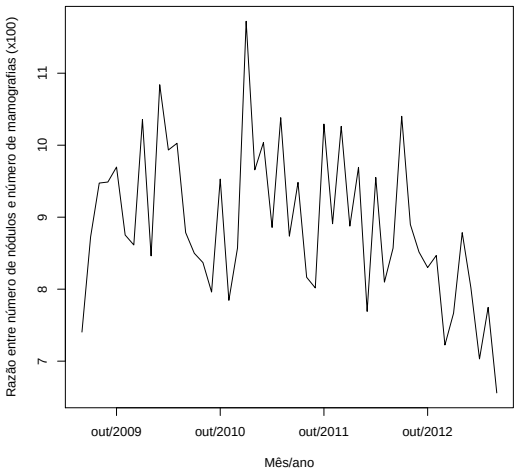


Figura 4: Média da razão do número de nódulos pelo número de exames entre as regiões ao longo do período de estudo.

As Figuras 5 e 6 tem o objetivo de apresentar as relações das variáveis independentes grau de urbanização, PIB, renda média e taxa do número de médicos pela população feminina com as variáveis dependentes em estudo. Os modelos propostos na próxima seção tem a função de quantificar estas relações lineares e, a partir de testes de hipóteses, evidenciar se estas

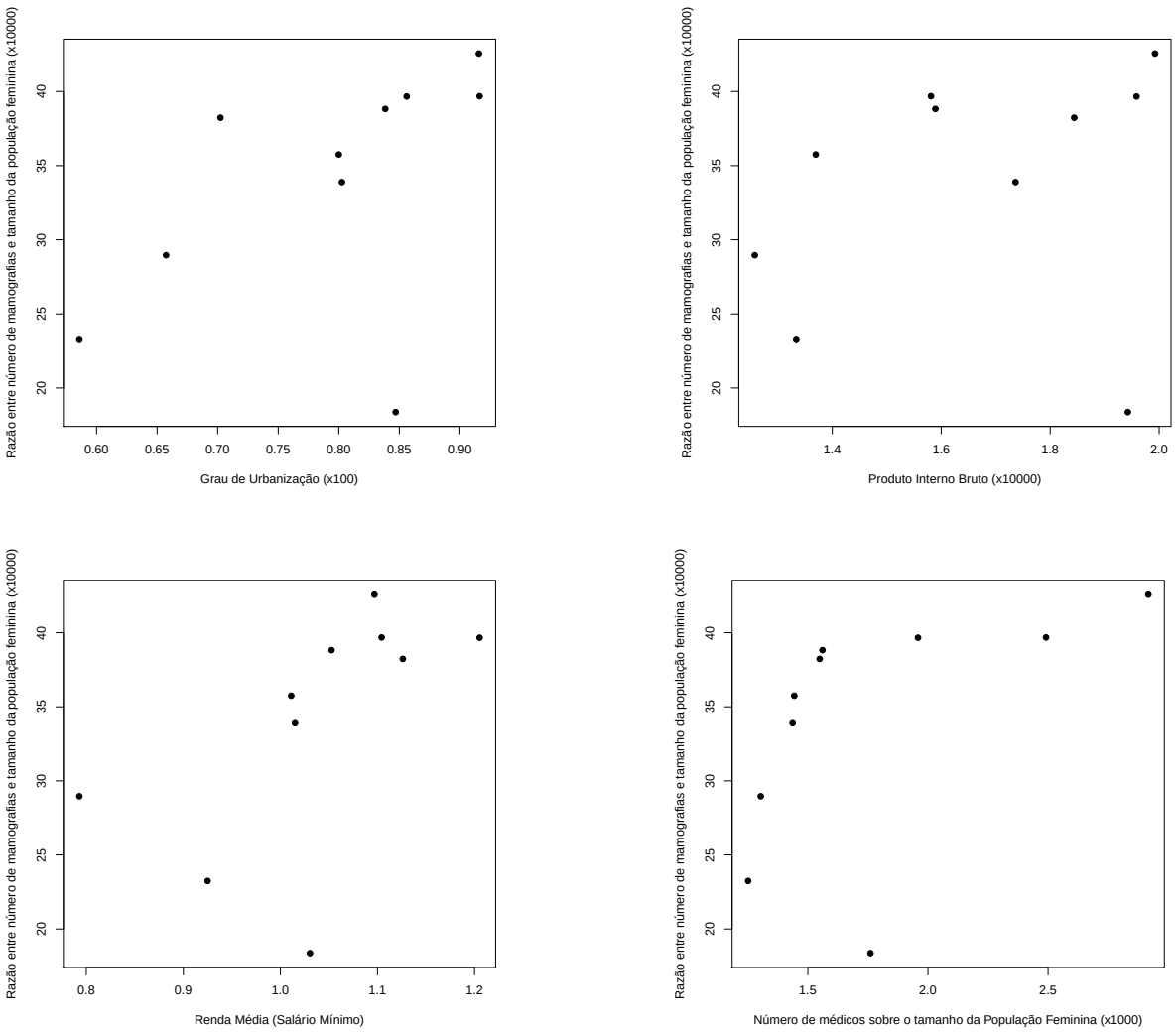


Figura 5: Gráfico de dispersão das variáveis independentes contra a média da razão do número de mamografias pelo tamanho da população feminina segundo a região.

relações são significativas ou não.

Na Figura 5, observa-se um ponto discrepante na relação linear da variável dependente com as variáveis independentes Grau de Urbanização e PIB. Este ponto representa a região Centro Oriental. Como na Figura 6, a partir da razão do número de nódulos com o número de mamografias este ponto deixou de ser discrepante, optou-se por manter esta região no estudo.

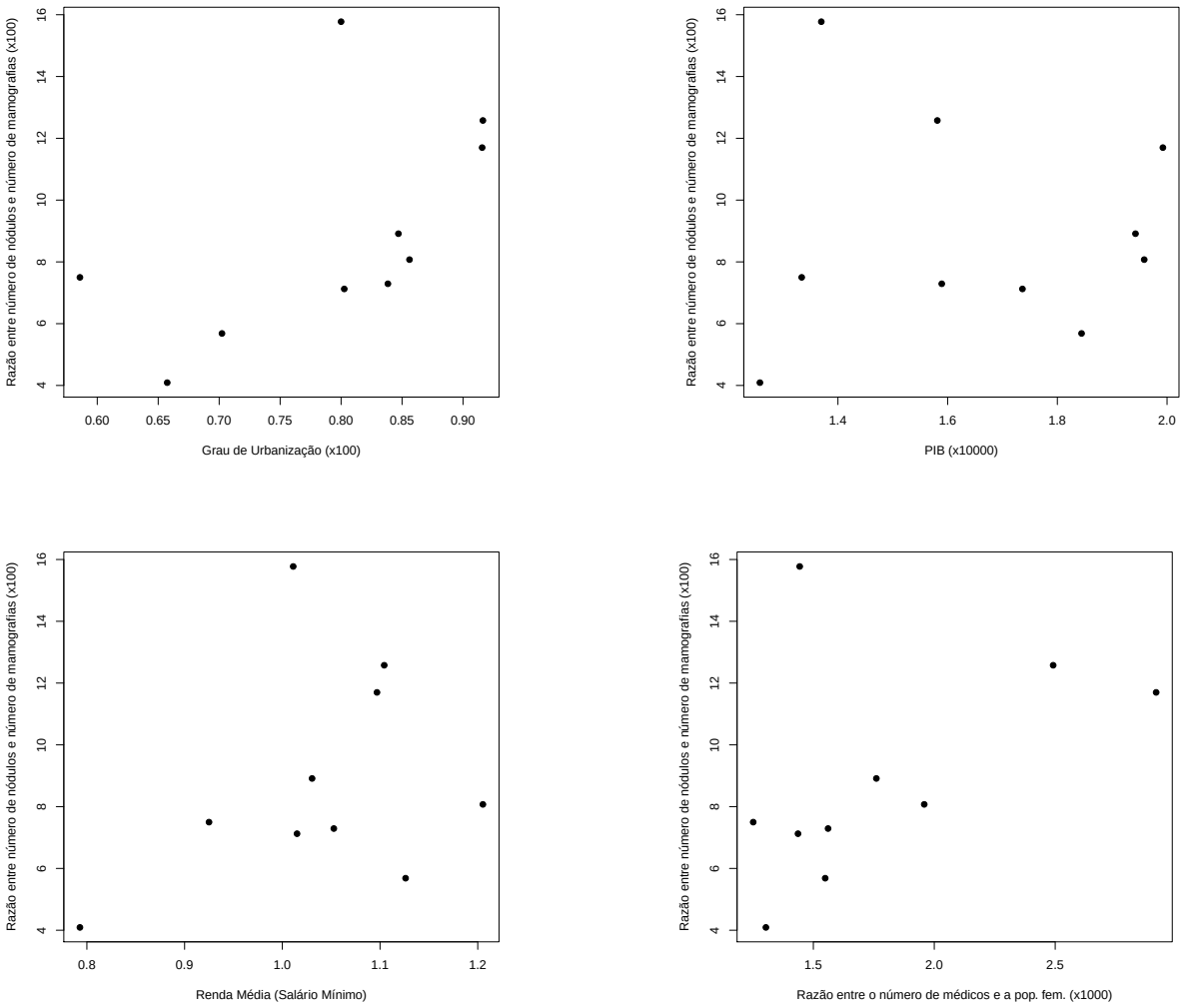


Figura 6: Gráfico de dispersão das variáveis independentes contra a média da razão do número de nódulos pelo número de mamografias segundo a região.

4.2 MODELOS

Para a análise do conjunto de dados em estudo, serão considerados 4 modelos distintos:

Modelo 1:

$$Y_{ijk} = \beta_{0k} + \sum_{l=1}^{4} \beta_{lk} X_{lij} + \mathrm{e}_{ijk} \tag{14}$$

em que $i = 1, \ldots, 10$ representam as regiões do Paraná; $j = 1, \ldots, 49$ representam o período; $k = 1$: razão do número de mamografrias pela população feminina e $k = 2$: razão do número de nódulos pelo número de mamografias, representam as variáveis dependentes; $\beta_{0k}, \beta_{1k}, \ldots, \beta_{4k}$ são os efeitos fixos associados às variáveis independentes, na região i e no período j:

- X_{1ij} representa o grau de urbanização;
- X_{2ij} representa o PIB;
- X_{3ij} representa a renda per capita;
- X_{4ij} representa a taxa de número de médicos pela população feminina.

Ainda, e_{ijk} são os erros associados a cada k variável dependente, assumindo distribuição Normal, independente identicamente distribuídas (i.i.d.), com média zero e variância σ_k^2.

Modelo 2:

$$Y_{ijk} = \beta_{0k} + \sum_{l=1}^{4} \beta_{lk} X_{lij} + \delta_{ik} + \mathrm{e}_{ijk} \tag{15}$$

em que $i = 1, \ldots, 10$ representam as regiões do Paraná; $j = 1, \ldots, 49$ representam o período e $k = 1, 2$ representam as variáveis dependentes; $\beta_{0k}, \beta_{1k}, \ldots, \beta_{4k}$ são os efeitos fixos associados às variáveis independentes, na região i e no período j, definidas no Modelo 1.

Ainda, tem-se que δ_{ik} representa o efeito das 10 regiões do estado do Paraná e assume distribuição Normal (i.i.d.) com média zero e variância $\sigma_{\omega k}^2$ e e_{ijk} são os erros associados a cada k variável dependente, assumindo distribuição Normal com média zero e variância σ_k^2.

Modelo 3:

$$Y_{ijk} = \beta_{0k} + \sum_{l=1}^{4} \beta_{lk} X_{lij} + \omega_{jk} + \mathrm{e}_{ijk} \tag{16}$$

em que $i = 1,\ldots,10$ representam as regiões do Paraná; $j = 1,\ldots,49$ representam o período e $k = 1,2$ representam as variáveis dependentes; $\beta_{0k},\beta_{1k},\ldots,\beta_{4k}$ são os efeitos fixos associados às variáveis independentes, na região i e no período j, definidas no Modelo 1.

Ainda, tem-se que ω_{jk} representa o efeito de tempo e assume distribuição Normal (i.i.d.) com média zero e variância $\sigma^2_{\delta k}$ e e_{ijk} são os erros associados a cada k variável dependente, assumindo distribuição Normal com média zero e variância σ^2_k.

Modelo 4:

$$Y_{ijk} = \beta_{0k} + \sum_{l=1}^{4} \beta_{lk} X_{lij} + \delta_{ik} + \omega_{jk} + \mathrm{e}_{ijk} \tag{17}$$

em que $i = 1,\ldots,10$ representam as regiões do Paraná; $j = 1,\ldots,49$ representam o período e $k = 1,2$ representam as variáveis dependentes; $\beta_{0k},\beta_{1k},\ldots,\beta_{4k}$ são os efeitos fixos associados às variáveis independentes, na região i e no período j, definidas no Modelo 1.

Ainda, tem-se que δ_{ik} assume distribuição Normal com média zero e variância $\sigma^2_{\delta k}$, ω_{jk} assume distribuição Normal com média zero e variância $\sigma^2_{\omega k}$ e e_{ijk} são os erros associados a cada k variável dependente, assumindo distribuição Normal (i.i.d.) com média zero e variância σ^2_k.

Na Tabela 2, é apresentado um resumo dos efeitos considerados nos modelos.

Tabela 2: Resumo dos modelos

Modelo	Efeito(s)
1	Sem efeito
2	Efeito de regiões
3	Efeito de tempo
4	Efeitos de regiões e tempo

4.3 ESTIMAÇÃO DOS PARÂMETROS

Para estimar os parâmetros pelo enfoque frequentista, utilizou-se a biblioteca *lme4* do Software R (R Core Team, 2017). Neste caso, é necessário apenas carregar a matriz dos dados e escrever o modelo escolhido. As principais linhas de comando estão disponíveis no Apêndice C.1.

Nas Tabelas 3 e 4 são apresentadas as estimativas dos parâmetros sob o enfoque frequentista, Intervalos de Confiança e o valor de AIC para cada modelo.

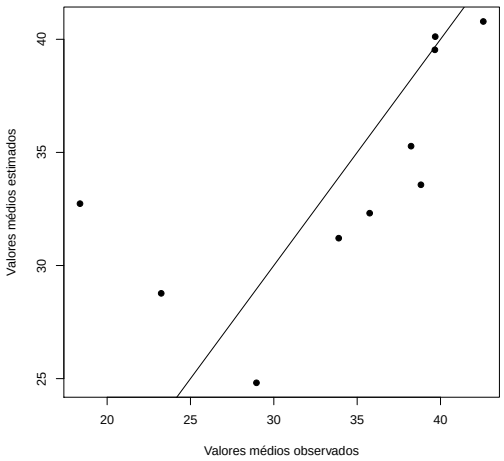
Como introduzido na Seção 3.1.1, a escolha do melhor modelo entre os propostos corresponde ao que apresenta o menor *AIC*. Para a variável número de mamografias sobre a

Tabela 3: Estimativas dos parâmetros dos modelos para $k = 1$ (frequentista)

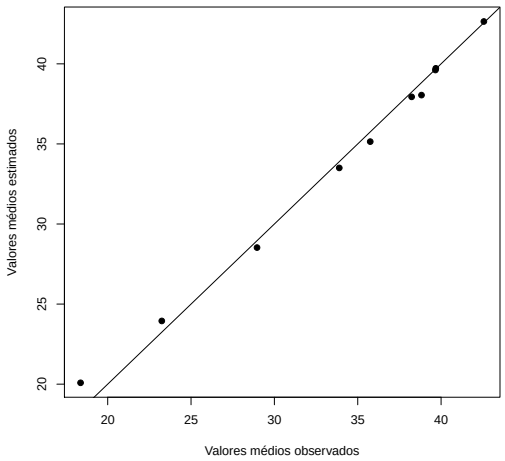
	Parâmetros	Estimativa	IC 95%	AIC
Modelo 1	β_0	$-3,72$	$(-21,69;14,24)$	$4268,60$
	β_1	$-1,03$	$(-28,83;26,77)$	
	β_2	$-3,21$	$(-8,16;1,75)$	
	β_3	$33,31$	$(11,42;55,19)$	
	β_4	$5,25$	$(0,13;10,36)$	
Modelo 2	β_0	$1,35$	$(-41,75;43,58)$	$4220,51$
	β_1	$-12,07$	$(-75,63;53,81)$	
	β_2	$0,10$	$(-5,50;5,21)$	
	β_3	$26,94$	$(-22,91;77,14)$	
	β_4	$7,95$	$(-3,77;18,80)$	
Modelo 3	β_0	$-4,37$	$(-21,45;12,78)$	$4235,56$
	β_1	$-0,38$	$(-26,77;25,94)$	
	β_2	$-5,56$	$(-11,79;0,54)$	
	β_3	$37,02$	$(15,28;58,49)$	
	β_4	$5,35$	$(0,49;10,19)$	
Modelo 4	β_0	$0,99$	$(-42,25;43,53)$	$4205,88$
	β_1	$-11,31$	$(-75,22;54,53)$	
	β_2	$0,24$	$(-7,51;6,88)$	
	β_3	$26,88$	$(-23,62;77,88)$	
	β_4	$7,70$	$(-3,90;18,57)$	

população feminina (Tabela 3), o modelo 4 apresenta melhor ajuste quando comparado aos demais. Ao nível de 5% de significância, nenhuma das covariáveis apresenta evidências de significância, pois o valor 0 está contido nos intervalos de confiança para os parâmetros deste modelo. Na Figura 7 são apresentados gráficos de ajustes dos valores médios observados segundo os valores estimados pelos modelos. Também observa-se um excelente ajuste para os modelos 2 e 4. A análise de resíduos dos modelos foi realizada a partir de gráficos de dispersão e quantil-quantil, verificando-se assim que a distribuição dos resíduos é independente e segue distribuição aproximadamente normal.

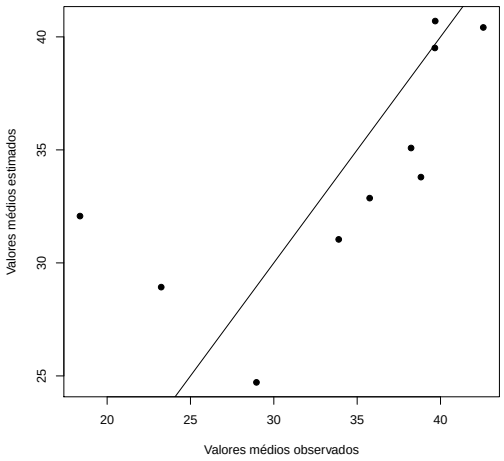
Para a variável número de nódulos sobre o número de mamografias (Tabela 4), o modelo 2 apresenta melhor ajuste quando comparado aos demais. Ao nível de 5% de significância, a covariável PIB apresenta evidências de significância com estimador negativo, o que sugere que quanto menor o PIB da região maior o número de nódulos. Na Figura 8 são apresentados gráficos de ajustes dos valores médios observados segundo os valores estimados pelos modelos. Também observa-se um excelente ajuste para os modelos 2 e 4.



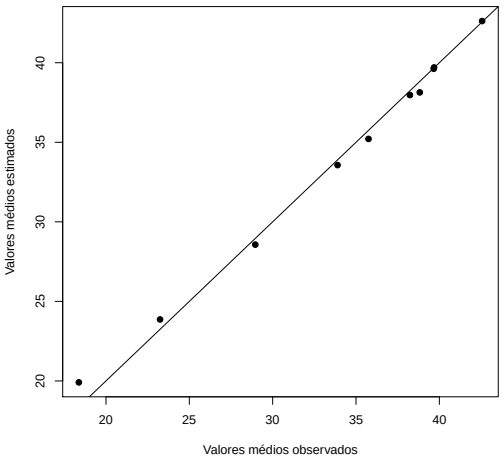
(a) Modelo 1



(b) Modelo 2

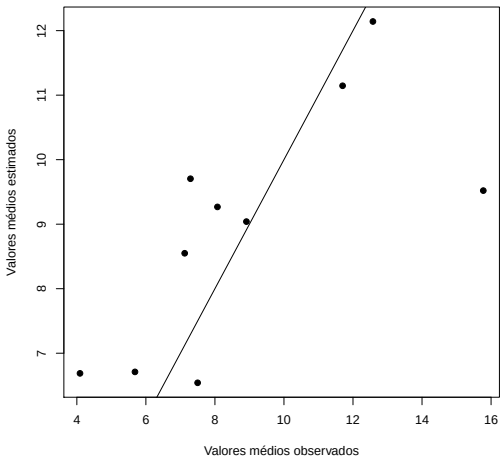


(c) Modelo 3

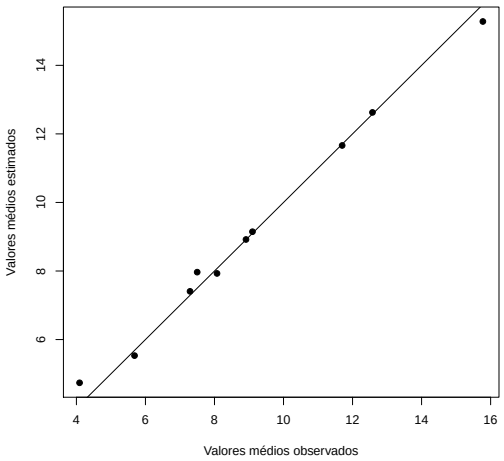


(d) Modelo 4

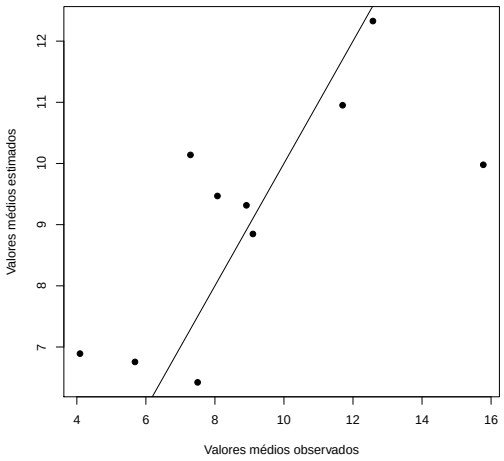
Figura 7: Gráfico de dispersão dos valores médios observados contra os valores estimados ($k = 1$).



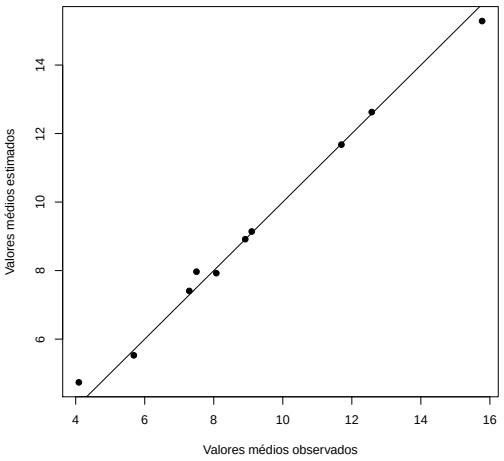
(a) Modelo 1



(b) Modelo 2



(c) Modelo 3



(d) Modelo 4

Figura 8: Gráfico de dispersão dos valores médios observados contra os valores estimados ($k = 2$).

Tabela 4: Estimativas dos parâmetros dos modelos para $k = 2$ (frequentista)

	Parâmetros	Estimativa	IC 95%	AIC
Modelo 1	β_0	$-3,46$	$-7,31;0,38$	$2710,53$
	β_1	$-16,75$	$10,80;22,70$	
	β_2	$-3,34$	$-4,41;-2,27$	
	β_3	$2,48$	$-2,24;7,20$	
	β_4	$1,17$	$0,07;2,26$	
Modelo 2	β_0	$-1,78$	$(-19,95;16,36)$	$2489,15$
	β_1	$13,73$	$(-11,81;39,40)$	
	β_2	$-1,49$	$(-2,49;-0,52)$	
	β_3	$-0,93$	$(-22,91;21,07)$	
	β_4	$1,82$	$(-1,67;5,17)$	
Modelo 3	β_0	$-3,46$	$(-7,29;0,36)$	$2706,45$
	β_1	$16,75$	$(10,83;22,66)$	
	β_2	$-3,34$	$(-4,40;-2,28)$	
	β_3	$2,48$	$(-2,21;7,17)$	
	β_4	$1,17$	$(0,08;2,25)$	
Modelo 4	β_0	$-1,78$	$(-19,95;16,36)$	$2491,15$
	β_1	$13,73$	$(-11,81;39,41)$	
	β_2	$-1,49$	$(-2,50;-0,51)$	
	β_3	$-0,93$	$(-22,91;21,07)$	
	β_4	$1,82$	$(-1,67;5,17)$	

4.4 MÉDIAS *A POSTERIORI*

A fim de obter as medidas resumo, das quantidades aleatórias geradas sob enfoque Bayesiano, utilizou-se o Software JAGS (PLUMMER, 2003) com o Software R como interface. Neste caso é necessário carregar a matriz dos dados, a entrada do modelo e escolha das distribuições *a priori*. Logo, para os 4 modelos propostos nesta aplicação, considerou-se as distribuições *a priori*, dadas por:

$$\begin{aligned}
\beta_{0k} &\sim N(a_{\beta_{0k}}, b^2_{\beta_{0k}}); \beta_{lk} \sim N(a_{\beta_{lk}}, b^2_{\beta_{lk}}) \\
\sigma_k^{-2} &\sim G(c_k, d_k) \\
\omega_{jk} &\sim N\left(0, \sigma^2_{\omega k}\right) \\
\delta_{ik} &\sim N\left(0, \sigma^2_{\delta k}\right)
\end{aligned}$$

em que $k = 1,2$ e $l = 1,2,3,4$. No segundo estágio da definição das distribuições *a priori* (GELMAN, 2006), tem-se $\sigma^{-2}_{\omega k} \sim G(e_k, f_k)$ e $\sigma^{-2}_{\delta k} \sim G(g_k, h_k)$. $N(a, b^2)$ que denota uma distribuição normal com média a e variância b^2; $G(c,d)$ denota uma distribuição gama com média c/d e variância c/d^2. A escolha dos hiperparâmetros que leve a distribuições *a priori* aproximadamente não informativas são dados por: $a = 0$; $b = 100$; $c = d = 0,001$; $e = f = g = h = 0,1$.

As principais linhas de comando estão disponíveis no Apêndice C.2.

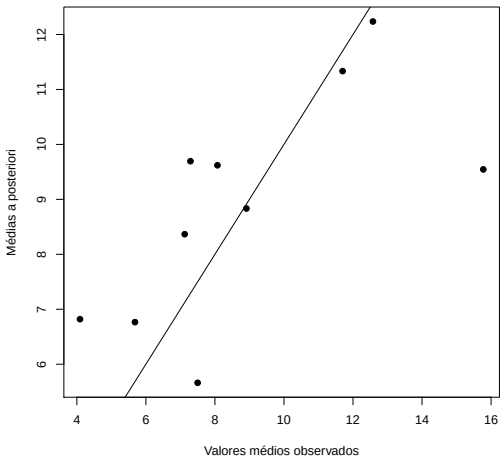
Nas Tabelas 5 e 6 são apresentadas as médias *a posteriori* das quantidades aleatórias sob o enfoque frequentista, Intervalos de Credibilidade IC_r e o valor de DIC para cada modelo.

Tabela 5: Média e 95% C_rI **para estimação dos parâmetros dos modelos para** $k = 1$

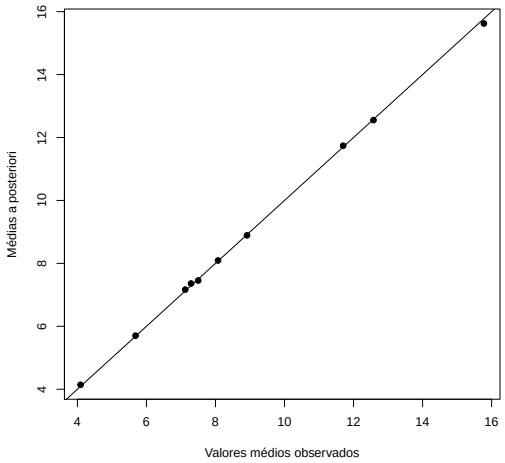
	Parâmetros	Média	95% C_rI	DIC
Modelo 1	β_0	-3.90	$(-22.16; 13.41)$	$4279,21$
	β_1	-2.30	$(-29.26; 24.37)$	
	β_2	-3.11	$(-8.33; 2.05)$	
	β_3	33.90	$(11.30; 57.66)$	
	β_4	5.42	$(0.75; 10.88)$	
Modelo 2	β_0	$1,79$	$(-53,43; 56,34)$	$4232,85$
	β_1	$-10,88$	$(-85,49; 65,10)$	
	β_2	$-0,14$	$(-5,12; 5,23)$	
	β_3	$25,26$	$(-36,84; 85,85)$	
	β_4	$8,14$	$(-4,57; 21,98)$	
Modelo 3	β_0	$-3,99$	$(-21,65; 14,05)$	$4255,99$
	β_1	$-1,25$	$(28,67; 26,77)$	
	β_2	$-5,68$	$(-12,17; 0,58)$	
	β_3	$37,05$	$(15,66; 57,97)$	
	β_4	$5,57$	$(0,85; 10,75)$	
Modelo 4	β_0	$2,53$	$(-49,16; 60,06)$	$4215,42$
	β_1	$-8,99$	$(-82,60; 64,75)$	
	β_2	$0,36$	$(-7,03; 7,97)$	
	β_3	$23,39$	$(-37,88; 78,61)$	
	β_4	$7,73$	$(-4.42; 21,79)$	

Como introduzido na Seção 3.2.5, a escolha do melhor modelo entre os propostos corresponde ao que apresenta o menor *DIC*. Para a variável número de mamografias sobre a população feminina (Tabela 5), o modelo 4 apresenta melhor ajuste quando comparado aos demais. Considerando os intervalos com 95% de credibilidade nenhuma das covariáveis apresenta evidências de significância. Na Figura 9 são apresentados gráficos de ajustes dos valores médios observados segundo os valores estimados pelos modelos. Também observa-se um excelente ajuste para os modelos 2 e 4.

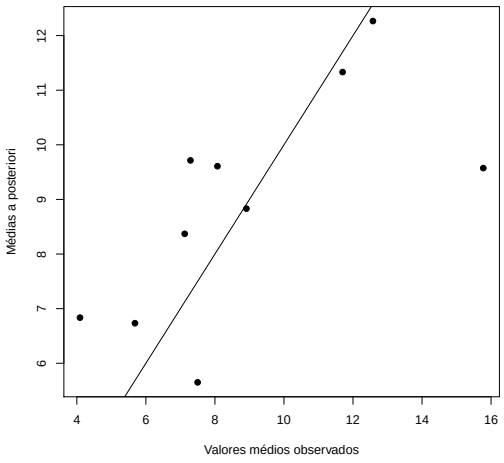
Para a variável número de nódulos sobre o número de mamografias (Tabela 6), o modelo 2 apresenta melhor ajuste quando comparado aos demais. Ao nível de 5% de significância, a covariável PIB apresenta evidências de significância com estimador negativo, o que sugere que quanto menor o PIB da região maior o número de nódulos. Na Figura 10 são apresentados gráficos de ajustes dos valores médios observados segundo os valores estimados pelos modelos. Também observa-se um excelente ajuste para os modelos 2 e 4.



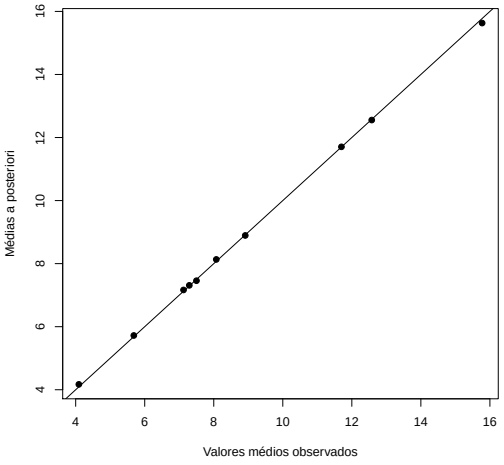
(a) Modelo 1



(b) Modelo 2



(c) Modelo 3

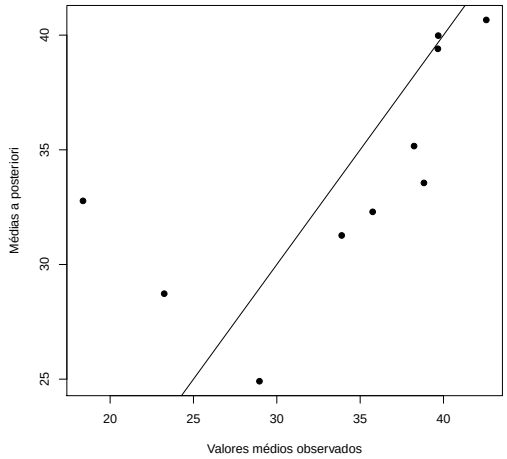


(d) Modelo 4

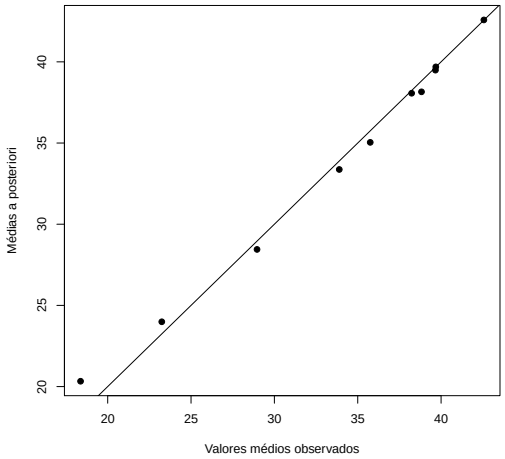
Figura 9: Gráfico de dispersão dos valores médios observados contra as médias *a posteriori* ($k = 1$).

Tabela 6: Média e 95% IC_r **para estimação dos parâmetros dos modelos para** $k = 2$

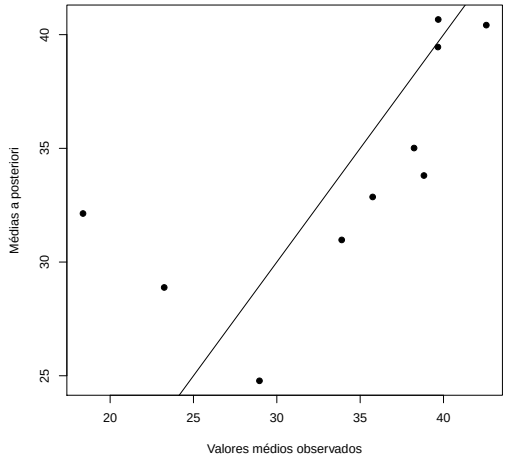
	Parâmetros	Média	95% C_rI	DIC
Modelo 1	β_0	$-3,93$	$(-9,70;1,66)$	$3064,72$
	β_1	$19,21$	$(10,96;27,58)$	
	β_2	$-3,49$	$(-4,96;-1,92)$	
	β_3	$2,28$	$(-4,67;9,24)$	
	β_4	$0,62$	$(-1,01;2,23)$	
Modelo 2	β_0	$-1,07$	$(-23,62;24,12)$	$2977,74$
	β_1	$13,21$	$(-25,93;46,74)$	
	β_2	$-2,59$	$(-3,88;-0,57)$	
	β_3	$-0,59$	$(-28,57;29,78)$	
	β_4	$2,21$	$(-2,66;7,84)$	
Modelo 3	β_0	$-3,99$	$(-9,54;1,36)$	$3065,26$
	β_1	$19,79$	$(10,95;29,01)$	
	β_2	$-3,53$	$(-5,09;-2,04)$	
	β_3	$2,13$	$(-4,24;8,73)$	
	β_4	$0,60$	$(-0,98;2,10)$	
Modelo 4	β_0	$-1,27$	$(-25,26;22,92)$	$2977,89$
	β_1	$13,22$	$(-21,75;43,55)$	
	β_2	$-2,19$	$(-4,05;-0,40)$	
	β_3	$-0,12$	$(-27,92;28,29)$	
	β_4	$2,07$	$(-3,03;7,09)$	



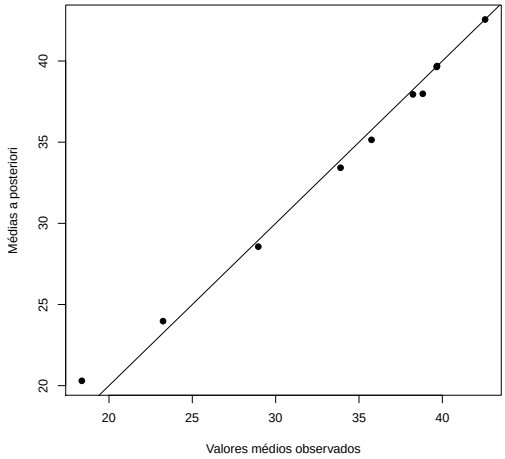
(a) Modelo 1



(b) Modelo 2



(c) Modelo 3



(d) Modelo 4

Figura 10: Gráfico de dispersão dos valores médios observados contra as médias *a posteriori* ($k = 2$).

5 CONCLUSÃO E CONSIDERAÇÕES FINAIS

O acesso a informação e o registro de dados tem possibilitado que, além da revisão de conceitos teóricos, ainda exista a possibilidade de aplicações em problemas do cotidiano. Neste trabalho foram apresentados os conceitos de regressão linear múltipla sob os enfoques frequentista e Bayesiano bem como uma aplicação destes conceitos, traduzidos em quatro modelos distintos para um conjunto de dados reais sobre o número de mamografias e nódulos em mulheres do estado do Paraná, no período de junho de 2009 a junho de 2013.

Nos modelos ajustados verificou-se que, para os dados em estudo, o efeito de tempo (Modelo 3) não trouxe ganho nos ajustes, porém o efeito aleatório de município (Modelo 2) foi fundamental para obter o melhor ajuste. Observe que, em todos os casos, o modelo 1 apresenta ajuste ruim que podem levar a conclusões erradas sobre a relação entre as variáveis. Assim, evidencia-se a necessidade do ajuste de modelos de efeitos mistos neste caso.

Nas Figuras 9 e 10, observa-se que a introdução de efeitos aleatórios proporciona melhores ajustes nestes modelos. Logo, os modelos 2 e 4 foram as melhores alternativas. A introdução do efeito de tempo não apresenta resultados relevantes nestes modelos, o que pode ser verificado nas Tabelas 3, 4, 5, 6. Os modelos 1 e 3, apresentam *AIC* e *DIC* parecidos e os modelos 2 e 4 também, o que retrata a não relevância desse efeito nos modelos.

Para os dois enfoques (frequentista e Bayesiano), a variável razão do número de mamografias pela população feminina apresenta o modelo 4 com melhor ajuste, por apresentar menor *AIC* e *DIC*. Para a variável razão do número de nódulos pelo número de mamografias, o melhor ajuste é o do modelo 2, que mesmo apresentando *AIC* e *DIC* com pouca diferença do modelo 4, é considerado um modelo parcimonioso.

No que diz respeito ao modelo 2, para a variável razão do número de nódulos pelo número de mamografias, β_2 foi considerado relevante do ponto de vista estatístico (ao nível de $0,05$ de significância), evidenciando assim que o PIB influencia nesta variável. O PIB é um dos indicadores que tem como objetivo principal mensurar a atividade econômica de uma determinada região. Como β_2 é negativo isto implica que quanto menor o PIB de uma região, maior é a razão entre número de nódulos encontrados número de mamografias realizados. É importante destacar que este estudo é observacional, ou seja, a relação encontrada não se trata de causa e efeito.

Os resultados deste trabalho foram apresentados no *II Latin American Statistical Computing Congress*, ocorrido na cidade de Valparaíso, no Chile, no período de 9 a 11 de março. O mesmo está submetido para publicação no *Chilean Journal of Statistics*.

Como perspectivas futuras deste trabalho, poderia-se considerar as distâncias entre as regiões, construindo-se assim uma análise espacial para o mesmo.

REFERÊNCIAS

ACHCAR, A. **Padrões temporais de casos de tuberculose em municípios do estado de São Paulo, no período de 2009 a 2013**. 2016. Dissertação (Mestrado em Saúde na Comunidade) - Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto.

AKAIKE, H. A new look at the statistical model identification. **IEEE Transactions on Automatic Control**, AC-19, n. 6, p. 716–723, 1974.

CHIB, S.; GREENBERG, E. Undestanding the metropolis-hastings algorithm. **The American Statistician**, v. 49, n. 4, p. 327–335, 1995.

CONTI, S. R. tradutor: A. R. D. **Probabilidade: um curso moderno com aplicaccões**. 8th. ed. Porto Alegre, BR: Bookman, 2010. 608 p.

DRAPER, N. R.; SMITH, H. **Applied Regression Analysis**. 3rd. ed. New York, US: J. Wiley and Sons, 1998.

GAMERMAN, D.; LOPES, H. F. **Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference**. London, GB: Chapman and Hall CRC, 2006.

GELFAND, A. E.; SMITH, A. F. M. Sampling-based approaches to calculating marginal densities. **Journal of the American Statistical Association**, v. 85, n. 410, p. 398–409, 1990.

GELMAN, A. Prior distributions for variance parameters in hierarchical models. **Bayesian Analysis**, v. 1, p. 1–19, 2006.

INCA. **Câncer de Mama: é preciso falar disso**. 2015.

INCA. **Diretrizes para detecção precoce do câncer de mama no Brasil**. 1st. ed. Rio de Janeiro: INCA, 2015. 171 p.

INCA; FIOCRUZ. **A mulher e o câncer de mama no Brasil**. 1st. ed. Rio de Janeiro: INCA: Coordenação Geral de Prevenção e Vigilância, 2014. 46 p.

IPARDES. `http://www.ipardes.gov.br/`. Acesso em 05/01/2017.

JAMES, B. R. **Probabilidade: um curso a nível intermediário**. 3rd. ed. Rio de Janeiro, BR: IMPA, 2010. 304 p.

LIMA-COSTA, M. F.; BARRETO, S. M. Tipos de estudos epidemiológicos: conceitos básicos e aplicações na área do envelhecimento. **Epidemiologia e Serviços de Saúde**, v. 12, n. 4, p. 189–201, 2003.

MANDAL, A. **History of Breast Cancer**. 2013. Disponível em: <http://www.news-medical.net/health/History-of-Breast-Cancer.aspx>. Acesso em: 27 de outubro de 2016.

MANDAL, A. **Types of Mastectomy**. 2014. Disponível em: <http://www.news-medical.net/health/Types-of-Mastectomy.aspx>. Acesso em: 27 de outubro de 2016.

MCLEAN, R. A.; SANDERS, W. L.; STROUP, W. W. A unified approach to mixed linear models. **The American Statistician**, v. 45, n. 1, p. 54–64, 1991.

PLUMMER, M. **JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling**. 2003.

R Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2017.

RODRIGUES, J. D.; CRUZ, M. S.; PAIXÃ£O, A. N. Uma anÃ¡lise da prevenÃ§Ã£o do cÃ¢ncer de mama no Brasil. **CiÃªncia & SaÃºde Coletiva**, v. 20, p. 3163 – 3176, 10 2015.

SPIEGELHALTER, D. J. et al. Bayesian measures of model complexity and fit. **Journal of the Royal Statistical Society: Series B (Statistical Methodology)**, Wiley Online Library, v. 64, n. 4, p. 583–639, 2002.

APÊNDICE A – REGIÕES DEMOGRÁFICAS DO PARANÁ

REGIÃO DEMOGRÁFICA/MUNICÍPIO

NOROESTE
Alto Paraíso
Alto Paraná
Alto Piquiri
Altônia
Amaporã
Brasilândia do Sul
Cafezal do Sul
Cianorte
Cidade Gaúcha
Cruzeiro do Oeste
Cruzeiro do Sul
Diamante do Norte
Douradina
Esperança Nova
Francisco Alves
Guairaçá
Guaporema
Icaraíma
Inajá
Indianópolis
Iporã
Itaúna do Sul
Ivaté
Japurá
Jardim Olinda
Jussara
Loanda
Maria Helena
Marilena
Mariluz
Mirador
Nova Aliança do Ivaí
Nova Londrina
Nova Olímpia
Paraíso do Norte
Paranacity
Paranapoema
Paranavaí
Perobal
Pérola
Planaltina do Paraná
Porto Rico
Querência do Norte
Rondon
Santa Cruz de Monte Castelo
Santa Isabel do Ivaí
Santa Mônica

REGIÃO DEMOGRÁFICA/MUNICÍPIO

Santo Antônio do Caiuá
São Carlos do Ivaí
São Jorge do Patrocínio
São Manoel do Paraná
São Pedro do Paraná
São Tomé
Tamboara
Tapejara
Tapira
Terra Rica
Tuneiras do Oeste
Umuarama
Xambrê

CENTRO OCIDENTAL
Altamira do Paraná
Araruna
Barbosa Ferraz
Boa Esperança
Campina da Lagoa
Campo Mourão
Corumbataí do Sul
Engenheiro Beltrão
Farol
Fênix
Goioerê
Iretama
Janiópolis
Juranda
Luiziana
Mamborê
Moreira Sales
Nova Cantu
Peabiru
Quarto Centenário
Quinta do Sol
Rancho Alegre D'Oeste
Roncador
Terra Boa
Ubiratã

NORTE CENTRAL
Alvorada do Sul
Ângulo
Apucarana
Arapongas
Arapuã
Ariranha do Ivaí

Astorga
Atalaia
Bela Vista do Paraíso
Bom Sucesso
Borrazópolis
Cafeara
Califórnia
Cambé
Cambira
Cândido de Abreu
Centenário do Sul
Colorado
Cruzmaltina
Doutor Camargo
Faxinal
Floraí
Floresta
Florestópolis
Flórida
Godoy Moreira
Grandes Rios
Guaraci
Ibiporã
Iguaraçu
Itaguajé
Itambé
Ivaiporã
Ivatuba
Jaguapitã
Jandaia do Sul
Jardim Alegre
Kaloré
Lidianópolis
Lobato
Londrina
Lunardelli
Lupionópolis
Mandaguaçu
Mandaguari
Manoel Ribas
Marialva
Marilândia do Sul
Maringá
Marumbi
Mauá da Serra
Miraselva
Munhoz de Melo
Nossa Senhora das Graças
Nova Esperança
Nova Tebas
Novo Itacolomi
Ourizona
Paiçandu
Pitangueiras
Porecatu
Prado Ferreira
Presidente Castelo Branco
Primeiro de Maio
Rio Bom
Rio Branco do Ivaí
Rolândia
Rosário do Ivaí
Sabáudia
Santa Fé
Santa Inês
Santo Inácio
São João do Ivaí
São Jorge do Ivaí
São Pedro do Ivaí
Sarandi
Sertanópolis
Tamarana
Uniflor

NORTE PIONEIRO

Abatiá
Andirá
Assaí
Bandeirantes
Barra do Jacaré
Cambará
Carlópolis
Congonhinhas
Conselheiro Mairinck
Cornélio Procópio
Curiúva
Figueira
Guapirama
Ibaiti
Itambaracá
Jaboti
Jacarezinho
Japira
Jataizinho
Joaquim Távora
Jundiaí do Sul
Leópolis
Nova América da Colina
Nova Fátima
Nova Santa Bárbara

Pinhalão
Quatiguá
Rancho Alegre
Ribeirão Claro
Ribeirão do Pinhal
Salto do Itararé
Santa Amélia
Santa Cecília do Pavão
Santa Mariana
Santana do Itararé
Santo Antônio da Platina
Santo Antônio do Paraíso
São Jerônimo da Serra
São José da Boa Vista
São Sebastião da Amoreira
Sapopema
Sertaneja
Siqueira Campos
Tomazina
Uraí
Wenceslau Braz

CENTRO ORIENTAL
Arapoti
Carambeí
Castro
Imbaú
Jaguariaíva
Ortigueira
Palmeira
Piraí do Sul
Ponta Grossa
Reserva
Sengés
Telêmaco Borba
Tibagi
Ventania

OESTE
Anahy
Assis Chateaubriand
Boa Vista da Aparecida
Braganey
Cafelândia
Campo Bonito
Capitão Leônidas Marques
Cascavel
Catanduvas
Céu Azul
Corbélia
Diamante D'Oeste
Diamante do Sul
Entre Rios do Oeste
Formosa do Oeste
Foz do Iguaçu
Guaíra
Guaraniaçu
Ibema
Iguatu
Iracema do Oeste
Itaipulândia
Jesuítas
Lindoeste
Marechal Cândido Rondon
Maripá
Matelândia
Medianeira
Mercedes
Missal
Nova Aurora
Nova Santa Rosa
Ouro Verde do Oeste
Palotina
Pato Bragado
Quatro Pontes
Ramilândia
Santa Helena
Santa Lúcia
Santa Tereza do Oeste
Santa Terezinha de Itaipu
São José das Palmeiras
São Miguel do Iguaçu
São Pedro do Iguaçu
Serranópolis do Iguaçu
Terra Roxa
Toledo
Três Barras do Paraná
Tupãssi
Vera Cruz do Oeste

SUDOESTE
Ampére
Barracão
Bela Vista da Caroba
Boa Esperança do Iguaçu
Bom Jesus do Sul
Bom Sucesso do Sul
Capanema
Chopinzinho
Clevelândia

Coronel Domingos Soares
Coronel Vivida
Cruzeiro do Iguaçu
Dois Vizinhos
Enéas Marques
Flor da Serra do Sul
Francisco Beltrão
Honório Serpa
Itapejara d'Oeste
Manfrinópolis
Mangueirinha
Mariópolis
Marmeleiro
Nova Esperança do Sudoeste
Nova Prata do Iguaçu
Palmas
Pato Branco
Pérola d'Oeste
Pinhal de São Bento
Planalto
Pranchita
Realeza
Salgado Filho
Salto do Lontra
Santa Izabel do Oeste
Santo Antônio do Sudoeste
São João
São Jorge d'Oeste
Saudade do Iguaçu
Sulina
Verê
Vitorino

CENTRO-SUL
Boa Ventura de São Roque
Campina do Simão
Candói
Cantagalo
Espigão Alto do Iguaçu
Foz do Jordão
Goioxim
Guarapuava
Inácio Martins
Laranjal
Laranjeiras do Sul
Marquinho
Mato Rico
Nova Laranjeiras
Palmital
Pinhão
Pitanga
Porto Barreiro
Quedas do Iguaçu
Reserva do Iguaçu
Rio Bonito do Iguaçu
Santa Maria do Oeste
Turvo
Virmond

SUDESTE
Antônio Olinto
Bituruna
Cruz Machado
Fernandes Pinheiro
General Carneiro
Guamiranga
Imbituva
Ipiranga
Irati
Ivaí
Mallet
Paula Freitas
Paulo Frontin
Porto Vitória
Prudentópolis
Rebouças
Rio Azul
São João do Triunfo
São Mateus do Sul
Teixeira Soares
União da Vitória

REGIÃO METROPOLITANA DE CURITIBA
Adrianópolis
Agudos do Sul
Almirante Tamandaré
Antonina
Araucária
Balsa Nova
Bocaiúva do Sul
Campina Grande do Sul
Campo do Tenente
Campo Largo
Campo Magro
Cerro Azul
Colombo
Contenda
Curitiba
Doutor Ulysses
Fazenda Rio Grande

Guaraqueçaba
Guaratuba
Itaperuçu
Lapa
Mandirituba
Matinhos
Morretes
Paranaguá
Piên
Pinhais
Piraquara
Pontal do Paraná
Porto Amazonas
Quatro Barras
Quitandinha
Rio Branco do Sul
Rio Negro
São José dos Pinhais
Tijucas do Sul
Tunas do Paraná

APÊNDICE B – REVISÃO DE ALGUNS PONTOS

B.1 MÉTODO DOS MÍNIMOS QUADRADOS

Da regressão linear simples tem-se $Y_i = \beta_0 + \beta_1 X_i + \varepsilon$. O método dos mínimos quadrados minimiza a soma dos quadrados dos resíduos, ou seja, $\sum_{i=1}^{n} \varepsilon_i^2$. A ideia desse método é que ao minimizar a soma do quadrado dos resíduos, encontra-se β_0 e β_1.

Substituindo ε_i por $Y_i - \beta_0 - \beta_1 X_i$, tem-se:

$$S = \sum_{i=1}^{n} \varepsilon^2 \;=\; \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

Derivando S em relação a β_0:

$$\frac{\partial S}{\partial \beta_0} = -2 \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i) \qquad (1)$$

Derivando S em relação a β_1:

$$\frac{\partial S}{\partial \beta_1} = -2 \sum_{i=1}^{n} X_i (Y_i - \beta_0 - \beta_1 X_i) \qquad (2)$$

Distribuindo e dividindo (1) por $2n$ tem-se:

$$\frac{-2\sum_{i=1}^{n} Y_i}{2n} + \frac{2\sum_{i=1}^{n} \beta_0}{2n} + \frac{2\sum_{i=1}^{n} \beta_1 X_i}{2n} = \frac{0}{2n}$$

$$\frac{-\sum_{i=1}^{n} Y_i}{n} + \frac{\sum_{i=1}^{n} \beta_0}{n} + \frac{\beta_1 \sum_{i=1}^{n} X_i}{n} = 0$$

Chamando $\bar{Y} = \dfrac{\sum_{i=1}^{n} Y_i}{n}$ e $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$, obtém-se,

$$\bar{Y} + \beta_0 + \beta_1 \bar{X} = 0$$
$$\beta_0 = \bar{Y} - \beta_1 \bar{X}$$

Agora, para encontrar o parâmetro β_1, substitui-se β_0 em (2):

$$-2\sum_{i=1}^{n} X_i(Y_i - \bar{Y} + \beta_1\bar{X} - \beta_1 X_i) = 0$$

$$\sum_{i=1}^{n} [X_i(Y_i - \bar{Y}) + \beta_1 X_i(\bar{X} - X_i)] = 0$$

$$\sum_{i=1}^{n} X_i(Y_i - \bar{Y}) + \beta_1 \sum_{i=1}^{n} X_i(\bar{X} - X_i) = 0$$

$$\beta_1 = \frac{\sum_{i=1}^{n} X_i(Y_i - \bar{Y})}{\sum_{i=1}^{n} X_i(\bar{X} - X_i)}$$

$$\beta_1 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (\bar{X} - X_i)^2}$$

B.2 ESTIMADOR DE MÁXIMA VEROSSIMILHANÇA

Estimar os parâmetros μ e σ^2 pelo método da Máxima Verossimilhança para distribuição normal.

$$\text{Seja,} f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{\frac{-1}{2\sigma^2}(y-\mu)^2\right\}$$

Então a Função de Verossimilhança é dada por,

$$\begin{aligned} L(\mu,\sigma^2) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{\frac{-1}{2\sigma^2}(y_i-\mu)^2\right\} \\ &= \prod_{i=1}^{n} (2\pi)^{-1/2}(\sigma^2)^{-1/2} exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2\right\} \\ &= \prod_{i=1}^{n} (2\pi)^{-1/2}(\sigma^2)^{-1/2} exp\left\{\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2\right\} \end{aligned}$$

Aplicando o produtório e o ln, tem-se

$$l(\mu,\sigma^2) = \frac{-n}{2}\ln(2\pi) - \frac{n}{2} ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2$$

Derivando em relação a σ^2,

$$\frac{\partial \ell(\mu,\sigma^2)}{\partial \sigma^2} = -\frac{1}{\sigma^2}\frac{n}{2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(y_i-\mu)^2$$
$$= -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(y_i-\mu)^2$$

Igualando a zero,

$$-\frac{n}{2\sigma^2} + \frac{1}{2}\sum_{i=1}^{n}\left(\frac{y_i-\mu}{\sigma^2}\right)^2 = 0 \;\Leftrightarrow\; -n + \sum_{i=1}^{n}\frac{(y_i-\mu)^2}{\sigma^2} = 0 \;\Leftrightarrow\; \widehat{\sigma}^2 = \frac{(n-1)}{n}s^2,$$

em que $s^2 = \dfrac{\sum_{i=1}^{n}(y_i-\mu)^2}{n-1}$,

Como

$$\frac{\partial^2 L(\mu,\sigma^2)}{\partial(\sigma^2)^2} = \frac{1}{(\sigma^2)^2}\left(\frac{n}{2} - \frac{(n-1)s^2}{\sigma^2}\right)$$

Avaliando a equação acima em $\widehat{\sigma}^2 = \frac{(n-1)s^2}{n}$, tem-se que

$$\frac{\partial^2 L(\mu,\sigma^2)}{\partial(\sigma^2)^2} = -\frac{n}{2}\frac{1}{\sigma^2} < 0$$

Então, o estimador de máxima verossimilhança para σ^2 é $\widehat{\sigma}^2 = \frac{(n-1)}{n}s^2$.

Agora, derivando em relação a μ, para estimar esse parâmetro.

$$\frac{\partial \ell(\mu,\sigma^2)}{\partial \mu} = -2\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)(-1)$$
$$= \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i-\mu) = \sum_{i=1}^{n}\left(\frac{y_i-\mu}{a}\right) \qquad (1)$$

Igualando (1) a zero, obtem-se

$$\sum_{i=1}^{n}\left(\frac{y_i-\mu}{a}\right) = 0$$
$$\frac{1}{a}\sum_{i=1}^{n}(y_i-\widehat{\mu}) = 0$$
$$n\widehat{\mu} = \sum_{i=1}^{n} y_i$$
$$\widehat{\mu} = \bar{y}$$

O possível estimador de Máxima Verossimilhança para a média é $\bar{y}$, então verifica-se se $\bar{y}$ é ponto de máximo, para isso, calcula-se,

$$\frac{\partial^2(\mu,\sigma^2)}{\partial\mu^2} = \frac{\partial^2}{\partial\mu^2}\left(\frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i-\mu)\right) = \frac{-n}{\sigma^2} < 0$$

Assim, $\bar{y}$ estimador de máxima verossimilhança para μ é $\widehat{\mu} = \bar{y}$

B.3 ALGORITMO GIBBS SAMPLING

Listing B.1: Código de um algoritmo Gibbs Sampling para um modelo de regressão linear simples

```
y <- c(0.10, 0.65, 0.30, 0.30, 0.28, 0.78, 0.28, 0.45)
x <- c(0.08, 0.17, 0.08, 0.30, 0.05, 0.18, 0.09, 0.45)

M <- 25000 #total de amostras geradas

sigma2 <- vector()
alfa <- vector()
beta <- vector()

#Chute inicial
sigma2[1] <- 1
alfa[1] <- 1
beta[1] <- 0

#Hiperparametros - Não informativos
n <- length(y)
a0 <- 1e+03
a1 <- 1e+03
b <- 0.01
d <- 0.01

for(m in 2:M)
  {
  e<- y-alfa[m-1]-beta[m-1]*x

  sigma2[m] <- 1/rgamma(1, (b+n/2), (d+0.5*sum(e^2)))

  mu0 <- y - beta[m-1]*x
  media_alfa <- ((a0^2)*sum(mu0)) / (sigma2[m] + (n*a0^2))
  var_alfa <- ((a0^2)*sigma2[m]) / (sigma2[m] + (n*a0^2))

  alfa[m] <- rnorm(1, media_alfa, sqrt(var_alfa))

  mu1 <- y - alfa[m]
```

```
  media_beta <- ((a1^2)*sum(x*mu1)) / (sigma2[m] + (a1^2)*sum(x^2))
  var_beta <- ((a1^2)*sigma2[m]) / (sigma2[m] + (a1^2)*sum(x^2))

  beta[m] <- rnorm(1, media_beta, sqrt(var_beta))
}

#########################################################################

plot(sigma2,type="l")
acf(sigma2)

plot(alfa,type="l")
acf(alfa)

plot(beta,type="l")
acf(beta)

#########################################################################
#########################################################################

bur <- 5000 #Burn-in

salto <- 10
S <- ((M-bur)/salto)

sigma2_n <- vector()
alfa_n <- vector()
beta_n <- vector()

for(s in 1:S)
  {
  sigma2_n[s] <- sigma2[salto*s]
  alfa_n[s] <- alfa[salto*s]
  beta_n[s] <- beta[salto*s]
}

plot(sigma2_n,type="l")
acf(sigma2_n)
```

```
hist(sigma2_n)

plot(alfa_n,type="l")
acf(alfa_n)
hist(alfa_n)

plot(beta_n,type="l")
acf(beta_n)
hist(beta_n)

############################################################

mean(sigma2_n)
sd(sigma2_n)

mean(alfa_n)
sd(alfa_n)

mean(beta_n)
sd(beta_n)

#Intervalos de Credibilidade

library(boa)

quantile(sigma2_n,c(0.025,0.975))
quantile(alfa_n,c(0.025,0.975))
quantile(beta_n,c(0.025,0.975))
```

APÊNDICE C – ALGORITMOS FREQUENTISTA E BAYESIANO

C.1 ALGORITMO FREQUENTISTA

Listing C.1: Código do modelo frequenstista

```
library(lme4) #Biblioteca para modelos mistos

dados<- read.csv("dados_1b.csv", sep=";", dec=",")

############################################################

####Número de exames para cada 10000 mulheres

y<- (dados$Exames/dados$Populacao.F)*10000 #Num. de mamografias
## pela população feminina
x1<- dados$GU/100 #Grau de urbanização
x2<- dados$PIB/10000 #PIB
x3<- dados$Renda/564 #renda média
x4<- (dados$Medicos/dados$Populacao.F)*1000 #Num. de médicos
## pela pop. feminina
w<- dados$Regiao #10 regiões
tp<- dados$Mes/dados$Ano #Razão entre mês e ano

############ Modelo 4
# Considera efeito do tempo e de regiões.

md4<- lmer(y ~ x1 + x2 + x3 + x4 + (1|w) + (1|tp), data = dados)
#md4 é o modelo misto

summary(md4) #exibe resumo dos resultados da função de ajuste
```

C.2 ALGORITMO BAYESIANO

Listing C.2: Código do modelo bayesiano

```
rm(list=ls())

library(R2jags)
library(mcmcplots)

dados<- read.csv("dados_1.csv", sep=";", dec=",")

############################################################

#Abordagem: razão cada 100 exames

y<- (dados$Nodulos/dados$Exames)*100
y[y=="NaN"]<- NA
x1<- dados$GU/100
x2<- dados$PIB/10000
x3<- dados$Renda/564
x4<- (dados$Medicos/dados$Populacao.F)*1000

dados_2<- cbind(dados,y,x1,x2,x3,x4)

medias_nodulo<- tapply(dados_2$y,dados_2$Regiao,mean,na.rm=TRUE)
medias_GU<- tapply(dados_2$x1,dados_2$Regiao,mean)
medias_PIB<- tapply(dados_2$x2,dados_2$Regiao,mean)
medias_Renda<- tapply(dados_2$x3,dados_2$Regiao,mean)
medias_Medicos<- tapply(dados_2$x4,dados_2$Regiao,mean)

set.seed(10012017)

ym<- matrix(dados_2$y,10,49,byrow=TRUE)
x1m<- matrix(dados_2$x1,10,49,byrow=TRUE)
x2m<- matrix(dados_2$x2,10,49,byrow=TRUE)
x3m<- matrix(dados_2$x3,10,49,byrow=TRUE)
x4m<- matrix(dados_2$x4,10,49,byrow=TRUE)

y<- ym
```

```
x1<- x1m
x2<- x2m
x3<- x3m
x4<- x4m

jags.data<- c("y","x1","x2","x3","x4")


## Modelo 4 que considera os efeitos de tempo e região

jags.par <- c("beta0","beta1","beta2","beta3","beta4",
              "tau.c","sigma.c","mu","tau.tau","sigma.tau")

estimativas<- jags(data = jags.data, parameters.to.save = jags.par,
model.file = "one_novo.txt",n.iter = 55000, n.burnin = 5000,
n.thin = 50, n.chain = 1)

#DIC by Spiegelhalter

estimativas2<- jags(data = jags.data, parameters.to.save = jags.par,
model.file = "one_novo.txt",n.iter = 55000, n.burnin = 5000,
n.thin = 50, n.chain = 2)

dev_pD <- dic.samples(estimativas2$model,  n.iter = 55000,
              n.burnin = 5000, n.thin = 50, type="pD")

DIC_pD<- c(sum(dev_pD$penalty),sum(dev_pD$penalty)+
              + sum(dev_pD$deviance))

#####################################################################
#Arquivo "one_novo.txt":

# model
# {
# for(i in 1:10)
# {
# for(j in 1:49)
# {
```

```
# y[i,j] ~ dnorm(mu[i,j],tau.c)                                          74
# mu[i,j] <- beta0 + beta1*x1[i,j] + beta2*x2[i,j] + beta3*x3[i,j] +     75
#                          + beta4*x4[i,j] + tau[j] + w[i]               76
# }                                                                      77
# }                                                                      78
                                                                         79
# for(j in 1:49)                                                         80
# {                                                                      81
# tau[j] ~ dnorm(0,tau.tau)                                              82
# }                                                                      83
                                                                         84
# for(i in 1:10)                                                         85
# {                                                                      86
# w[i] ~ dnorm(0,tau.w)                                                  87
# }                                                                      88
                                                                         89
# tau.w ~ dgamma(0.1,0.1)                                                90
# sigma.w<- 1/sqrt(tau.w)                                                91
                                                                         92
# beta0 ~ dnorm(0,0.00001)                                               93
# beta1 ~ dnorm(0,0.0001)                                                94
# beta2 ~ dnorm(0,0.0001)                                                95
# beta3 ~ dnorm(0,0.0001)                                                96
# beta4 ~ dnorm(0,0.0001)                                                97
                                                                         98
# tau.c ~ dgamma(0.001,0.001)                                            99
# sigma.c<- 1/sqrt(tau.c)                                                100
                                                                         101
# tau.tau ~ dgamma(0.1,0.1)                                              102
# sigma.tau<- 1/sqrt(tau.tau)                                            103
# }                                                                      104
```