

**UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO
CAMPUS GOIABEIRAS**

Gabriella Ferreira Demarque

**ANÁLISE DE INCOMPLETITUDE EM DADOS PÚBLICOS DA
ÁREA DE SAÚDE MATERNA DO BRASIL**

**VITÓRIA
2021**

GABRIELLA FERREIRA DEMARQUE

**ANÁLISE DE INCOMPLETITUDE EM DADOS PÚBLICOS DA
ÁREA DE SAÚDE MATERNA DO BRASIL**

**Trabalho de Conclusão de Curso sub-
metido à Universidade Federal do Es-
pírito Santo, como requisito necessário
para obtenção do grau de Bacharel em
Estatística**

Vitória, outubro de 2021

UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO

GABRIELLA FERREIRA DEMARQUE

Orientador(a): Profa. Dra. Agatha
Sacramento Rodrigues
Universidade Federal do Espírito Santo -
UFES

Profa. Dra. Nátnaly Adriana Jiménez Monroy
Universidade Federal do Espírito Santo -
UFES

Me. Cristiane de Freitas Paganoti
Faculdade de Medicina da Universidade de
São Paulo - FMUSP

Vitória, 04 de outubro de 2021

*Aos meus pais, José Carlos Demarque e Rúbia Mara Ferreira Demarque, e à minha avó,
Nazilda Moreira Ferreira, dedico este trabalho com muito amor e carinho.*

Agradecimentos

A Deus, pela minha vida, por todo o amparo, proteção, força e por me ajudar a entrentar todos os obstáculos ao longo do curso.

Aos meus pais, José Carlos e Rúbia Mara, e meu irmão, Matheus, que são o meu porto seguro, sempre me incentivaram, me ajudaram e nunca deixaram eu abaixar a cabeça nos momentos mais difíceis da minha vida pessoal e acadêmica. Obrigada pelo amor incondicional, sem vocês eu não teria conseguido.

À minha avó, Nazilda, pelo amor incondicional, por cuidar de mim e sempre me incentivar a estudar.

À toda minha família Ferreira e Demarque, em especial meus tios Natan e Rosane, que são como meus segundos pais, e meu primo Vitor, que sempre estiveram ao meu lado.

Ao meu namorado, Vyctor, por todo amor, paciência, conselhos, incentivos, compreensão, dedicação e por estar sempre ao meu lado desde o início da minha caminhada na Ufes.

Aos meus amigos, Fernanda, Elayne e Gabriel, por toda a parceria, conselhos e por termos conseguido vencer juntos todos os desafios da graduação. Em especial, a Fernanda pela parceria desde o primeiro dia, ajuda em todos os momentos e apoio imensurável que me deu ao longo dos anos.

Às minhas grandes amigas, Thais e Victoria, pela amizade sincera, conselhos, compreensão, tanto amor, carinho e lealdade.

A todos os professores do Departamento de Estatística da Ufes que fizeram parte da minha caminhada e me ajudaram ao longo do curso, em especial aos professores Agatha, Alessandro, Bartolomeu, Bruno, Fabio, Luciana, Nátaly e Patrick. Muito obrigada por todos os ensinamentos, vocês foram essenciais na minha caminhada.

À minha orientadora e amiga, Agatha, por toda dedicação, conselhos, orientações, conversas, pela paciência, pela ajuda e por sempre acreditar em mim.

A todos aqueles que contribuíram, de alguma forma, para a realização deste trabalho.

Aos participantes da banca examinadora.

“Que todos os nossos esforços estejam sempre focados no desafio à impossibilidade. Todas as grandes conquistas humanas vieram daquilo que parecia impossível.”
(Charles Chaplin)

Resumo

Os modelos de regressão beta podem ser utilizados para modelar proporções que estão limitadas no intervalo $(0, 1)$. Para os casos em que a variável resposta assume valores no intervalo $[0, 1]$, o modelo de regressão beta inflacionado em zero pode ser adotado. Este estudo objetiva analisar e explicar a incompletude (proporção de dados faltantes) para prematuridade, em níveis estadual e municipal, por meio de indicadores socioeconômicos: Índice de Desenvolvimento Humano Municipal (IDHM), IDHM Educação, IDHM Longevidade, IDHM Renda, Índice de Gini, Taxa de analfabetismo, Esperança de Vida, Taxa de água e de esgoto. A informação sobre prematuridade é obtida no Sistema de Informações sobre Nascidos Vivos (SINASC), assim como a informação da incompletude de outras variáveis obstétricas consideradas. Além disso, também o objetivo consiste em avaliar a correlação da incompletude para prematuridade com a incompletude de outras variáveis obstétricas (tipo de gravidez, tipo de parto, anomalias congênitas e número de consultas de pré-natal). Na análise em nível estadual, o modelo mais adequado foi o de regressão beta com dispersão fixa e encontramos que os indicadores IDHM e Taxa de Água apresentaram influência negativa na incompletude para prematuridade, ou seja, estados que possuem baixos valores de IDHM e Taxa de Água tendem a apresentar maiores índices de incompletude. Para a análise em nível municipal, o modelo mais adequado foi o de regressão beta inflacionado em zero os indicadores IDHM Educação, IDHM Renda e Nascimentos apresentam relação negativa com a incompletude para prematuridade esperada e para a probabilidade do município ter incompletude igual a zero, os indicadores IDHM Longevidade e IDHM Renda apresentam relação positiva e as variáveis Água/esgoto e Nascimentos apresentam relação negativa.

Palavras-chave: Coeficiente de Correlação de Postos de Spearman, Incompletude, Indicadores Socioeconômicos, Modelo de regressão beta, Modelo de regressão beta inflacionado em zero, Qualidade dos dados, Saúde Materno-Infantil, SINASC.

Abstract

Beta regression models can be used to model proportions that are bounded in the range $(0, 1)$. For cases where the response variable takes values in the range $[0, 1]$, the zero-inflated beta regression model can be adopted. This study aims to analyze and explain the incompleteness (proportion of missing data) for prematurity, at state and municipal levels, through socioeconomic indicators: Municipal Human Development Index (IDHM), IDHM Education, IDHM Longevity, IDHM Income, Gini Index, Illiteracy Rate, Life Expectancy, Water and Sewage Rate. Information on prematurity is obtained from the Information System on Live Births (SINASC), as well as information on incompleteness of other considered obstetric variables. In addition, the objective is also to evaluate the correlation of incompleteness for prematurity with the incompleteness of other obstetric variables (type of pregnancy, type of delivery, congenital anomalies and number of prenatal visits). In the analysis at the state level, the most suitable model was the beta regression with fixed dispersion and we found that the IDHM and Water Rate indicators had a negative influence on incompleteness for prematurity, that is, states that have low values of IDHM and Water Rate tend to have higher incompleteness rates. For the analysis at the municipal level, the most suitable model was the zero-inflated beta regression model. The indicators IDHM Education, IDHM Income and Births show a negative relationship with incompleteness for expected prematurity and for the probability of the municipality having incompleteness equal to zero, the IDHM Longevity and IDHM Income indicators show a positive relationship and the variables Water/Sewage and Births show a negative relationship.

Keywords: Spearman Rank Correlation Coefficient, Incompleteness, Socioeconomic Indicators, Beta regression model, Beta regression model inflated to zero, Data quality, Maternal and Child Health, SINASC.

Listas de ilustrações

Figura 1 – Curva de Lorenz.	33
Figura 2 – Representações do correlograma.	35
Figura 3 – Exemplo de gráfico de envelope.	43
Figura 4 – Mapa de calor para incompletude para prematuridade.	52
Figura 5 – Mapa de calor para incompletude para tipo de gravidez.	52
Figura 6 – Mapa de calor para incompletude para tipo de parto.	53
Figura 7 – Mapa de calor para incompletude para anomalias congênitas.	54
Figura 8 – Mapa de calor para incompletude para consultas de pré-natal.	54
Figura 9 – Correlograma das variáveis de incompletude do SINASC - nível estadual. “Prematuridade” indica a incompletude para prematuridade, “Gravidez” indica a incompletude para tipo de gravidez, “Parto” indica a incom- pletude para tipo de parto, “Consultas” indica a incompletude para o número de consultas de pré-natal e “Anomalias” indica a incompletude para anomalias congênitas.	56
Figura 10 – Mapa dos estados para esperança de vida ao nascer.	57
Figura 11 – Mapa dos estados para taxa de analfabetismo.	58
Figura 12 – Mapa dos estados para Índice de Gini.	58
Figura 13 – Mapa dos estados para IDHM.	58
Figura 14 – Mapa dos estados para IDHM Educação.	59
Figura 15 – Mapa dos estados para IDHM Longevidade.	59
Figura 16 – Mapa dos estados para IDHM Renda.	60
Figura 17 – Mapa dos estados para o percentual de acesso ao abastecimento de água.	60
Figura 18 – Mapa dos estados para o percentual de acesso ao esgotamento sanitário.	61
Figura 19 – Correlograma dos indicadores socioeconômicos (2017 ou 2018) e incom- pletude para prematuridade (ano 2019) - nível estadual. “Prematuridade” indica a incompletude para prematuridade.	62
Figura 20 – Mapa dos municípios para incompletude para prematuridade.	63
Figura 21 – Mapa dos municípios para incompletude para tipo de gravidez.	64
Figura 22 – Mapa dos municípios para incompletude para tipo de parto.	64
Figura 23 – Mapa dos municípios para incompletude para anomalias congênitas.	65
Figura 24 – Mapa dos municípios para incompletude para consultas de pré-natal.	66

Figura 25 – Correlograma das incompletude das variáveis do SINASC - nível municipal. “Prematuridade” indica a incompletude para prematuridade, “Gravidez” indica a incompletude para tipo de gravidez, “Parto” indica a incompletude para tipo de parto, “Consultas” indica a incompletude para o número de consultas de pré-natal e “Anomalias” indica a incompletude para anomalias congênitas.	67
Figura 26 – Mapa dos municípios para taxa de analfabetismo.	68
Figura 27 – Mapa dos municípios para taxa de água e esgoto.	68
Figura 28 – Mapa dos municípios para esperança de vida ao nascer.	69
Figura 29 – Mapa dos municípios para Índice de Gini.	69
Figura 30 – Mapa dos municípios para IDHM.	70
Figura 31 – Mapa dos municípios para IDHM Educação.	70
Figura 32 – Mapa dos municípios para IDHM Longevidade.	71
Figura 33 – Mapa dos municípios para IDHM Renda.	71
Figura 34 – Correlograma dos indicadores socioeconômicos (Censo 2010) e incompletude para prematuridade (ano 2019) - nível municipal. “Prematuridade” indica a incompletude para prematuridade.	72
Figura 35 – Gráficos de diagnóstico do modelo final - <i>fit2</i>	76
Figura 36 – Gráficos de dependência parcial para estados.	79
Figura 37 – Gráfico de dependência parcial com as duas variáveis explicativas (IDHM) e (Água) para estados.	79
Figura 38 – Envelope do modelo final considerando a transformação para variável de interesse para dados municipais.	81
Figura 39 – Análise dos resíduos para municípios.	85
Figura 40 – Gráficos de envelope do modelo final para municípios - <i>fit7</i>	85
Figura 41 – Gráficos de dependência parcial para municípios.	88
Figura 42 – Gráfico de dependência parcial com as duas variáveis explicativas (IDHM Educação) e (IDHM Renda) para municípios - Número de nascimentos fixada em 802.	89

Listas de tabelas

Tabela 1 – Funções de ligação. $\Phi(\cdot)$ é a função de distribuição acumulada de uma variável aleatória normal padrão.	37
Tabela 2 – Incompletude das variáveis para o Brasil entre 2016 e 2019.	55
Tabela 3 – Estimativas dos parâmetros do modelo completo para o primeiro caminho - ajuste para estados (<i>fit1</i>).	74
Tabela 4 – Estimativas dos parâmetros do modelo reduzido para o segundo caminho - ajuste para estados (<i>fit2</i>).	74
Tabela 5 – Estimativas dos parâmetros do modelo completo para o primeiro caminho - ajuste para estados (<i>fit11</i>).	75
Tabela 6 – Estimativas dos parâmetros do modelo reduzido para o segundo caminho - ajuste para estados (<i>fit12</i>).	75
Tabela 7 – Resultado dos ajustes para estados.	75
Tabela 8 – Resultados inferenciais sem as potenciais observações influentes.	77
Tabela 9 – Valor de $\hat{\mu}$ obtido pela Equação (5.1) a depender do valor do IDHM e Água.	78
Tabela 10 – Estimativas dos parâmetros do modelo reduzido para o primeiro caminho - ajuste para municípios (<i>fit1</i>).	82
Tabela 11 – Estimativas dos parâmetros do modelo reduzido para o segundo caminho - ajuste para municípios (<i>fit3</i>).	82
Tabela 12 – Estimativas dos parâmetros do modelo reduzido para o terceiro caminho - ajuste para municípios (<i>fit5</i>).	83
Tabela 13 – Estimativas dos parâmetros do modelo reduzido para o quarto caminho - ajuste para municípios (<i>fit7</i>).	84
Tabela 14 – Resultado dos ajustes para municípios.	84
Tabela 15 – Valor de $\hat{\mu}$ obtido pela Equação (5.2) a depender do valor IDHM Educação, IDHM Renda e do número de nascimentos.	87

Listas de abreviaturas e siglas

- AIC: Critério de Akaike;
- BIZ: Beta Inflacionada em Zero;
- CNES: Cadastro Nacional dos Estabelecimentos de Saúde;
- DATASUS: Departamento de Informática do SUS;
- DN: Declaração de Nascido Vivo;
- Fiocruz: Fundação Oswaldo Cruz;
- IBGE: Instituto Brasileiro de Geografia e Estatística;
- IDH: Índice de Desenvolvimento Humano;
- IDHM: Índice de Desenvolvimento Humano Municipal;
- IGP: Idade Gestacional do Parto;
- MLG: Modelos Lineares Generalizados;
- OMS: Organização Mundial da Saúde;
- ONU: Organização das Nações Unidas;
- PBDA: Portal Brasileiro de Dados Abertos;
- PCDaS: Plataforma de Ciência de Dados aplicada à Saúde;
- PIB: Produto Interno Bruto;
- PNAD Contínua: Pesquisa Nacional por Amostra de Domicílios Contínua;
- PNUD: Programa das Nações Unidas para o Desenvolvimento;
- PROADESS: Plataforma de Avaliação do Desempenho do Sistema de Saúde;
- RBIZ: Regressão Beta Inflacionado em Zero;
- SIM: Sistema de Informações sobre Mortalidade;
- SINASC: Sistema de Informações sobre Nascidos Vivos;
- SIS: Sistema de Informações em Saúde;

SNIS: Sistema Nacional de Informações sobre Saneamento;

SUS: Sistema Único de Saúde;

UF: Unidade Federativa do Brasil.

Sumário

1	INTRODUÇÃO	23
1.1	Objetivos	27
1.1.1	Objetivo Geral	27
1.1.2	Objetivos Específicos	27
1.2	Organização	27
2	METODOLOGIA	29
2.1	Qualidade dos dados	29
2.2	O banco de dados	30
2.2.1	Variáveis obstétricas	30
2.2.2	Variáveis socioeconômicas	31
2.2.2.1	IDHM, IDHM Educação, IDHM Longevidade e IDHM Renda	31
2.2.2.2	Taxa de Analfabetismo	32
2.2.2.3	Taxa de Água	32
2.2.2.4	Taxa de Esgoto	33
2.2.2.5	Índice de Gini	33
2.2.2.6	Esperança de vida	34
2.2.2.7	Taxa de Água e Esgoto	34
2.3	Análise de Correlação	34
2.4	Modelos de Regressão	35
2.4.1	Modelo Beta	36
2.4.2	Modelo de Regressão Beta	37
2.4.3	Modelo de Regressão Beta com dispersão variável	39
2.4.4	Testes da razão de verossimilhanças para heteroscedasticidade	40
2.4.5	Análise de diagnóstico e critério de Akaike	40
2.4.5.1	Análise de resíduos	41
2.4.5.2	Análise de influência	42
2.4.5.2.1	Alavanca generalizada	42
2.4.5.2.2	Distância de Cook	42
2.4.5.3	Gráfico envelope	43
2.4.5.4	Critério de Akaike - AIC	43
2.4.6	Régressão Beta Inflacionado em Zero	44
2.4.6.1	Modelo Beta Inflacionado	44
2.4.6.2	Modelo Beta Inflacionado em zero	44
2.4.6.3	Modelo de Regressão Beta Inflacionado em Zero	47

2.4.6.4	Análise diagnóstico	48
2.4.6.4.1	Resíduo Quantil Aleatorizado	48
2.4.6.2	Envelope Simulado	48
3	ANÁLISE EXPLORATÓRIA PARA DADOS ESTADUAIS	51
3.1	Análise das variáveis de incompletude	51
3.2	Análise das variáveis socieconômicas	55
4	ANÁLISE EXPLORATÓRIA PARA DADOS MUNICIPAIS	63
4.1	Análise das variáveis de incompletude	63
4.2	Análise das variáveis socieconômicas	66
5	MODELAGEM VIA REGRESSÃO BETA	73
5.1	Modelagem em nível estadual	73
5.1.1	Análise de diagnóstico	75
5.1.1.1	Interpretação do modelo	77
5.2	Modelagem em nível municipal	80
5.2.1	Análise de diagnóstico	84
5.2.1.1	Interpretação do modelo	85
6	CONSIDERAÇÕES FINAIS	91
 Referências		93
 APÊNDICES		101
APÊNDICE A – ANÁLISE DESCRIPTIVA - UF		103
APÊNDICE B – ANÁLISE DESCRIPTIVA - MUNICÍPIOS		111
APÊNDICE C – MODELAGEM - UF		117
APÊNDICE D – MODELAGEM - MUNICÍPIOS		121

1 Introdução

O acesso à informação está previsto no artigo 5º inciso XXXIII da Constituição Federal (Martins, 2014). Nesse contexto, podemos citar o Portal Brasileiro de Dados Abertos (PBDA), que foi construído pelo governo para centralizar a busca e o acesso dos dados e informações públicas (Moreira *et al.*, 2017), e o Sistema de Informações em Saúde (SIS). O SIS foi criado pelo Ministério da Saúde na década de 1990, com o intuito de criar uma estrutura capaz de gerar e transformar dados em informações (da Silva, 2016). Para isto, existem profissionais responsáveis por cada etapa: seleção, coleta, classificação, armazenamento, análise, divulgação e recuperação dos dados.

Ferraz *et al.* (2009) e Franco *et al.* (2012) colocam que o Departamento de Informática do Sistema Único de Saúde (DATASUS) desempenha um importante papel no processo de informação na saúde. Além disso, ele é responsável por manter a disposição todos os SIS em uso no Brasil e nele encontramos informações sobre Indicadores de Saúde, Assistência à Saúde, Epidemiológica e Morbidade, Rede Assistencial (Cadastro Nacional dos Estabelecimentos de Saúde - CNES), Estatísticas Vitais (Sistema de Informações sobre Nascidos Vivos - SINASC, Sistema de Informações sobre Mortalidade - SIM e Câncer - sítio do Instituto Nacional de Câncer), Demográficas e Socioeconômicas e Saúde Suplementar.

Jorge *et al.* (2007) destacam, entre todos o SIS, o SINASC, que tem por objetivo reunir informações relativas aos nascimentos de nascidos vivos ocorridos em todo o território nacional. Além disso, reúne informações sobre fatores importantes, do ponto de vista epidemiológico, e relacionados ao nascimento, a mãe, a gestação e ao parto. O SINASC tem como documento padronizado para a coleta dos dados a Declaração de Nascido Vivo (DN), documento obrigatório para o registro da criança no cartório.

A DN, padronizada pelo Ministério da Saúde, possui 52 campos, entre os quais podem ser destacados: informações da mãe (tais como: idade, raça/cor, estado civil, escolaridade, ocupação, número de filhos vivos e mortos, município de residência e etc), informações sobre o recém-nascido (tais como: peso, raça/cor, apgar no primeiro e no quinto minuto, anomalias genéticas e etc), informações sobre a gestação (tais como: número de consultas de pré-natal, idade gestacional, tipo de gravidez, tipo de parto, data do nascimento e etc), e informações sobre o local de ocorrência do parto (Cunha *et al.*, 2017).

Dessa forma, o SINASC permite a construção de indicadores que subsidiam o planejamento, a gestão e a avaliação de políticas e ações de vigilância e atenção à saúde materno-infantil, além de processar dados demográficos e epidemiológicos sobre a mãe e o recém-nascido (Gabriel *et al.*, 2014; Ministério da Saúde, 2019).

A magnitude do SINASC e sua importância para a saúde pública em particular tem despertado a necessidade de avaliação das suas informações, seja do ponto de vista quantitativo (cobertura do sistema), seja do qualitativo (confiabilidade das informações), de forma que os indicadores calculados refletem realmente o perfil da população (Theme Filha *et al.*, 2004).

Segundo Assunção (2012), durante a análise de bancos de dados é comum se deparar com variáveis que apresentam dados faltantes (*missings*), que podem ter diferentes origens: problema no armazenamento nos dados, não preenchimento cadastral, campos preenchidos de forma errada, entre outros. Como coloca Little e Rubin (2019) esta questão pode causar inúmeros problemas nas análises e cada vez mais pesquisadores têm estudado métodos estatísticos para tentar solucioná-la, como métodos de imputação de dados.

Assim como a maioria dos bancos de dados públicos, ao trabalhar com o SINASC, é preciso lidar com os dados faltantes. Como coloca Carvalho (2017), a presença de dados faltantes contribuem para a redução da precisão dos resultados e por isso é tão importante desenvolver estudos sobre esse assunto. No entanto, antes de fazer qualquer tipo de tratamento nos dados faltantes, é preciso tentar entender o motivo pelo qual eles surgiram, assim como avaliar a qualidade dos dados em questão a partir de indicadores de qualidade de dados, como, por exemplo, a incompletude.

Atualmente, existem vários estudos sobre a avaliação das informações do SINASC por meio dos indicadores de qualidade de dados. Dentre eles, destacam-se Silva *et al.* (2001), Theme Filha *et al.* (2004), Silva *et al.* (2013), Guimarães *et al.* (2013) e Gabriel *et al.* (2014). Os indicadores utilizados nesses estudos foram: confiabilidade, completude (proporção de dados preenchidos) e concordância.

Nesse estudo, trabalhamos com o indicador de incompletude, definido como a proporção de informação ignorada da variável de interesse. Vamos analisar as variáveis de incompletude para idade gestacional do parto (indicando quando há prematuridade para idade gestacional do parto menor que 37 semanas), tipo de gravidez (única, dupla ou tripla), tipo de parto (vaginal ou cesáreo), número de consultas de pré-natal e anomalias congênitas (considerando apenas o caso em que há algum tipo de malformação congênita).

A incompletude da idade gestacional do parto (IGP), também chamada ao longo desse trabalho como incompletude sobre prematuridade, é a principal variável de incompletude de interesse nesse estudo, e temos o interesse em avaliar a sua associação com alguns indicadores socieconômicos. Em outras palavras, queremos avaliar se estados e municípios com altos valores de incompletude de IGP também são aqueles com os piores indicadores socieconômicos.

Segundo a Organização Mundial da Saúde (OMS), cerca de 15 milhões de bebês nascem prematuramente todos os anos no mundo (WHO, 2012). Além disso, como mostram

Lajos *et al.* (2014), as taxas de partos prematuros vêm aumentando ao longo dos anos, na maioria dos países que possuem informações confiáveis. No Brasil, a câmara dos deputados aprovou uma proposta em 2021 que prevê ações para enfrentamento do parto prematuro (<https://www.camara.leg.br/noticias/777047-comissao-aprova-proposta-que-preve-acoes-para-enfrentamento-do-parto-prematuro>). A ocorrência do parto prematuro e seus fatores de risco tem sido alvo de muitos estudos, dada a importância do tema. Podemos destacar os seguintes estudos: Ramos e Cuman (2009), Bittar e Zugaib (2009), Howson *et al.* (2013), Lajos *et al.* (2014), Passini Jr *et al.* (2014), Walani (2020), Varella (2021).

São os indicadores socieconômicos considerados: o Índice de Desenvolvimento Humano Municipal (IDHM) e suas três vertentes (Educação, Longevidade e Renda), o Índice de Gini, a Taxa de Analfabetismo, a Taxa de Água (percentual da população urbana residente em domicílios ligados à rede de abastecimento de água), a Taxa de Esgoto (percentual da população urbana residente em domicílios ligados à rede de esgotamento sanitário) e Esperança de vida. Tanto para a análise dos estados quanto para a análise dos municípios serão utilizados os mesmos indicadores, com exceção dos indicadores Taxa de Água e Taxa de Esgoto que, para os municípios, são calculados em uma só variável, denominada de Taxa de água e de esgoto, que indica o percentual de pessoas em domicílios com abastecimento de água e esgotamento sanitário inadequados, sendo assim, um indicador de não acesso (diferente do indicador considerado para os estados).

O IDHM é uma medida que leva em consideração três dimensões do desenvolvimento humano: longevidade (mede a expectativa de vida da população), educação (mede o acesso ao conhecimento) e a renda (mede o padrão de vida). A Organização Mundial da Saúde (OMS) (WHO, 2013) diz que a importância do IDHM está relacionada com alguns fatores, tais como: 1) contraponto ao Produto Interno Bruto (PIB) - visa o desenvolvimento centrado nas pessoas e não se limita ao crescimento econômico; e, 2) comparação entre município e estímulo a melhoria - estimula as autoridades a implementarem políticas públicas que visam priorizar a qualidade de vida dos municípios. Além disso, caracteriza as faixas de desenvolvimento humano como sendo: muito baixo ($0 \leftarrow 0,49$), baixo ($0,50 \leftarrow 0,59$), médio ($0,60 \leftarrow 0,69$), alto ($0,70 \leftarrow 0,79$) e muito alto ($0,8 \leftarrow 1,0$). O Brasil obteve IDHM de 0,778 em 2017.

O Índice de Gini é um medidor de desigualdade de dados, sendo muito utilizado para medir a desigualdade de renda. Segundo Esparza *et al.* (2020), é a medida de desigualdade mais usada no mundo, além de ser uma valiosa ferramenta estatística. O Brasil registrou um índice de 0,533 em 2017. Também varia de 0 a 1, sendo que quanto mais próximo de 0, menor é a desigualdade social.

O indicador do analfabetismo diz respeito ao percentual de pessoas com 15 e mais anos de idade que não sabem ler nem escrever. Braga e Mazzeu (2017) revelam que o

Brasil possui uma elevada taxa de analfabetismo, que estão ligadas a situações de pobreza, exclusão e baixo desenvolvimento econômico. Para esse indicador, o Brasil obteve em 2017 uma taxa de 7%, aproximadamente, de pessoas com mais de 15 anos analfabetas.

Os indicadores de água e esgoto dos estados representam, respectivamente, o percentual da população urbana residente em domicílios ligados à rede de abastecimento de água e percentual da população urbana residente em domicílios ligados à rede de esgotamento sanitário, com base no Sistema Nacional de Informações sobre Saneamento (SNIS). O Manual de Saneamento Básico (Brasil, 2016) mostra como a falta de saneamento básico afeta negativamente a saúde pública, além de estar ligada a altas taxas de mortalidade infantil. Sabendo disso, valores altos para essas variáveis indicam o quanto aquele local precisa melhorar. Já o indicador Taxa de água e de esgoto, referente às informações municipais, refere-se ao percentual de pessoas em domicílios com abastecimento de água e esgotamento sanitário inadequados. Para esses indicadores, o Brasil obteve taxas de 85,8% (água) e 66% (esgoto) em 2017.

Por último, a esperança de vida ao nascer representa número médio de anos de vida esperados para um recém-nascido, mantido o padrão de mortalidade existente na população residente. Segundo Fernandes Guerra e Fígoli (2013), a esperança de vida é um importante indicador sociodemográfico que expressa o número médio de anos a ser vivido por uma determinada população a partir de determinada idade. Com isso, podemos dizer que quanto maior o valor dessa variável, melhor é a condição de vida daquele local. O Brasil obteve, em 2018, uma média de 72,8 de anos de vida.

Em relação aos indicadores socioeconômicos em nível estadual, é válido ressaltar que consideramos os dados da Pesquisa Nacional por Amostra de Domicílios Contínua (PNAD Contínua) dos anos de 2016 e 2017, informações mais recentes disponíveis (PNAD, 2018). O ano mais recente disponível com informações para os indicadores socioeconômicos em nível municipal é o do último Censo Demográfico, realizado em 2010.

A hipótese a ser considerada é sobre a existência de associação entre a porcentagem de incompletude para prematuridade e os indicadores socioeconômicos. Nesse sentido, há o interesse em estudar a proporção de dados faltantes para prematuridade com relação às variáveis socioeconômicas, por meio de um modelo de regressão. Como a proporção desses dados está limitada no intervalo (0, 1), o modelo de regressão beta pode ser considerado. Ao longo deste trabalho, os temas aqui serão abordados com maiores detalhes. No que segue, os objetivos geral e específicos deste trabalho serão descritos.

1.1 Objetivos

1.1.1 Objetivo Geral

Estudar a proporção de dados faltantes sobre prematuridade com relação às variáveis socioeconômicas, em níveis estadual e municipal, por meio de ajustes do modelo de regressão beta.

1.1.2 Objetivos Específicos

- Avaliar a incompletude das variáveis obstétricas do SINASC de interesse (tipo de gravidez, tipo de parto, anomalias congênitas e número de consultas de pré-natal), em níveis estadual e municipal para os anos de 2016, 2017, 2018 e 2019;
- Analisar os indicadores socioeconômicos em níveis estadual e municipal, fazendo recortes daqueles estados e municípios que apresentam os melhores e menores índices socioeconômicos.

1.2 Organização

O presente trabalho será dividido da seguinte forma:

- Capítulo 2: **Metodologia.** Aqui serão abordados os métodos considerados neste trabalho, tais como: a coleta dos dados, as variáveis escolhidas do SINASC e o cálculo da incompletude para cada uma delas, os indicadores socioeconômicos considerados, os conceitos da análise de correlação e a modelagem via regressão beta (modelo de regressão beta, modelo de regressão beta com dispersão variável e modelo de regressão beta inflacionado em zero).
- Capítulo 3: **Análise exploratória dos dados estaduais.** Neste capítulo serão abordadas as análises exploratórias dos dados dos estados, mostrando o comportamento das variáveis de incompletude e dos indicadores socioeconômicos ao longo dos anos e identificando quais são os melhores e piores estados com relação a esses indicadores. Além disso, também serão apresentadas as análises de correlação da incompletude para prematuridade, tanto com as demais variáveis de incompletude obstétricas, quanto com os indicadores socioeconômicos.
- Capítulo 4: **Análise exploratória dos dados municipais.** Aqui serão abordadas as análises exploratórias dos dados dos municípios, apresentando o comportamento das variáveis de incompletude e socioeconômicas, identificando quais são os melhores e piores municípios em relação a esses indicadores. Além disso, também serão apresentadas as análises de correlação da incompletude para prematuridade, tanto

com as demais variáveis de incompletude obstétricas, quanto com os indicadores socioeconômicos.

- Capítulo 5: **Modelagem via regressão beta.** Este capítulo é dividido em duas partes: modelagem para estados e modelagem para municípios. Para o dois casos, serão abordados os processos de escolha das variáveis explicativas no modelo, análises de diagnóstico e interpretações dos modelos.
- Capítulo 6: **Considerações finais.** Serão abordados neste capítulo os principais resultados das análises feitas.

2 Metodologia

Nesse capítulo, são descritos todos os métodos considerados neste trabalho. Na Seção 2.1, serão apresentadas as métricas de qualidade dos dados. Já a Seção 2.2 é dedicada para as bases de dados consideradas e onde elas foram obtidas, as variáveis escolhidas do SINASC e suas respectivas fórmulas para o cálculo da incompletude e também os indicadores socioeconômicos considerados. Na Seção 2.3 são apresentados os métodos de análise de correlação considerados. Por último, fechando a parte da metodologia, a Seção 2.4 mostrará a parte da modelagem via modelos de regressão beta.

Todo desenvolvimento computacional deste trabalho é realizado no *software R* (R Core Team, 2020) e todo código utilizado está disponível no Apêndice deste trabalho.

2.1 Qualidade dos dados

Sob a ótica de Merino *et al.* (2016), ao trabalhar com grandes volumes de dados, podemos enfrentar alguns desafios, como: qualidade dos dados, caracterização adequada dos dados, interpretação correta dos resultados e etc. Segundo ele, o processo de análise de qualidade dos dados é o principal aspecto na avaliação de conjuntos de dados, pois garante que eles estejam adequados aos fins para os quais foram originalmente destinados. Em outras palavras, a qualidade está diretamente ligada à confiabilidade dos dados.

Segundo Lima (2010), para avaliar a qualidade dos dados, existem alguns indicadores de qualidade, sendo eles: completude, conformidade, acurácia, consistência e temporalidade. Por completude, entende-se como a taxa de preenchimento de um determinado campo. A conformidade avalia o quanto os dados estão de acordo com os padrões descrito no dicionário. Já a acurácia avalia o quão próximo ou distante uma observação está da realidade, sendo que, quanto maior a acurácia, melhor o resultado. A consistência está relacionada à vários aspectos, tais como frequência de preenchimento, duplicidade ou falta de registros, presença de outliers e etc. Por último, a temporalidade diz respeito ao tempo entre os eventos e ao armazenamento de informações históricas de uma determinada variável, por exemplo.

Além disso, Guimarães *et al.* (2014) ressaltam que alguns fatores podem comprometer a qualidade dos dados de um sistema de informações como, por exemplo, o SINASC. Dessa forma, podemos destacar a subnotificação e a presença de dados faltantes (ignorado ou não preenchido). Assim, para saber como está a qualidade dos dados de um determinado sistema, é preciso conhecer alguns indicadores de qualidade.

A completude de preenchimento de um banco de dados é avaliada a partir do comple-

mentar da incompletude, ou seja, da porcentagem de dados faltantes (campos em brancos ou ignorados), como coloca o relatório técnico do Ministério da Saúde (Ministério da Saúde, 2019). Romero e Cunha (2006) e Gabriel *et al.* (2014) criaram uma classificação para avaliar uma base de dados com base neste indicador. Eles consideram como excelente menos de 5% de dados não preenchidos; bom de 5% a 10%; regular de 10% a 20%; ruim de 20% a 50% e muito ruim acima de 50%.

Neste trabalho, vamos considerar a incompletude para avaliar a qualidade das informações obtidas para algumas variáveis de interesse do SINASC.

2.2 O banco de dados

2.2.1 Variáveis obstétricas

Os dados do SINASC foram extraídos pela Plataforma de Ciência de Dados aplicada à Saúde (PCDaS) da Fundação Oswaldo Cruz (Fiocruz), disponível em <https://bigdata-metadados.icict.fiocruz.br/organization/pcdas>.

Sabendo que a variável “Número de nascimentos” corresponde ao número de nascimentos por ano e por estado ou por município, a depender da granularidade da análise, as incompletudes das variáveis do SINASC consideradas são calculadas da seguinte maneira:

- Incompletude para prematuridade:** A variável que identifica a idade gestacional do parto no SINASC é GESTACAO, com os seguintes valores: 9 - Ignorado, 1 - Menos de 22 semanas, 2 - 22 a 27 semanas, 3 - 28 a 31 semanas, 4 - 32 a 36 semanas, 5 - 37 a 41 semanas, 6 - 42 semanas e mais. A incompletude para prematuridade (ou porcentagem de dados faltantes para prematuridade) é calculada da seguinte maneira:

$$\text{Incompletude para prematuridade} = \frac{\text{Número de casos 9}}{\text{Número de nascimentos}}.$$

- Incompletude para tipo de gravidez:** A variável que identifica o tipo de gravidez no SINASC é GRAVIDEZ, com os seguintes valores: 9 - Ignorado, 1 - Única, 2 - Dupla, 3 - Tripla e mais. A incompletude para tipo de gravidez (ou porcentagem de dados faltantes para tipo de gravidez) é calculada da seguinte maneira:

$$\text{Incompletude para tipo de gravidez} = \frac{\text{Número de casos 9}}{\text{Número de nascimentos}}.$$

- Incompletude para tipo de parto:** A variável que identifica o tipo de parto no SINASC é PARTO, com os seguintes valores: 9 - Ignorado, 1 - Vaginal, 2 - Cesáreo.

A incompletude para tipo de parto (ou porcentagem de dados faltantes para tipo de parto) é calculada da seguinte maneira:

$$\text{Incompletude para tipo de parto} = \frac{\text{Numero de casos 9}}{\text{Numero de nascimentos}}.$$

4. **Incompletude para consultas de pré-natal:** A variável que identifica o número de consultas de pré-natal no SINASC é CONSULTAS, com os seguintes valores: 1 - Nenhuma consulta, 2 - de 1 a 3 consultas, 3 - de 4 a 6 consultas, 4 - 7 e mais consultas, 9 - Ignorado. A incompletude para consultas de pré-natal (ou porcentagem de dados faltantes para consultas de pré-natal) é calculada da seguinte maneira:

$$\text{Incompletude para consultas de pre natal} = \frac{\text{Numero de casos 9}}{\text{Numero de nascimentos}}.$$

5. **Incompletude para anomalias congênitas:** A variável que identifica anomalias congênitas no SINASC é IDANOMAL, com os seguintes valores: 1 - Sim, 2 - Não, 9 - Ignorado. A incompletude para anomalias (ou porcentagem de dados faltantes para anomalias) é calculada da seguinte maneira:

$$\text{Incompletude para anomalias} = \frac{\text{Numero de casos 9}}{\text{Numero de nascimentos}}.$$

2.2.2 Variáveis socioeconômicas

Os dados socioeconômicos estaduais foram obtidos pela plataforma do Atlas do Desenvolvimento Humano no Brasil, disponível em <http://www.atlasbrasil.org.br> e pela plataforma de Metodologia de Avaliação do Desempenho do Sistema de Saúde (PROADESS) que pertence à Fiocruz, disponível em <http://www.proadess.icict.fiocruz.br>.

Já os dados socioeconômicos municipais, que possuem informações referentes aos Censos Demográficos de 2010, foram retirados da plataforma Base dos Dados, disponível em <https://basedosdados.org/dataset/mundo-onu-adh>.

Quanto aos indicadores socioeconômicos que serão utilizados para os estados, listamos: IDHM e três suas vertentes: Educação, Longevidade e Renda, Índice de Gini, Taxa de Analfabetismo, Taxa de Água, Taxa de Esgoto e Esperança de vida. No que segue, serão mostrados o cálculo e a interpretação de cada indicador escolhido.

2.2.2.1 IDHM, IDHM Educação, IDHM Longevidade e IDHM Renda

Segundo a WHO (WHO, 2013), o IDHM é baseado no Índice de Desenvolvimento Humano (IDH), porém ele ajusta-se à realidade brasileira (a partir dos dados do Censo Demográfico) e às características dos municípios. O cálculo é feito a partir de três dimensões do desenvolvimento humano: educação (que mede o acesso ao conhecimento), a longevidade

(que mede a expectativa de vida da população) e a renda (que mede o padrão de vida). Todas as informações são retiradas dos Censos Demográficos do IBGE.

O IDHM Educação (acesso ao conhecimento) é calculado por meio da média aritmética de dois indicadores: a escolaridade da população adulta (percentual de pessoas com 18 anos ou mais com ensino fundamental completo), que tem peso 1 e o fluxo escolar da população jovem (média aritmética do percentual de jovens de 11 a 13 anos frequentando o ensino fundamental, do percentual de jovens de 15 a 17 anos com ensino fundamental completo e do percentual de jovens de 18 a 20 anos com ensino médio completo), que tem peso 2.

O IDHM Longevidade (vida longa e saudável) é medido pela expectativa de vida ao nascer, que mede o número médio de anos de vida que um recém nascido vai viver em um determinado município. O IDHM Renda (padrão de vida) é medido a partir da renda municipal *per capita*, isto é, a renda média do município. O cálculo é feito a partir da soma da renda de todos os residentes dividida pelo número total de pessoas que moram no município. Por fim, podemos definir o cálculo do IDHM da seguinte forma:

$$\text{IDHM} = \sqrt[3]{(\text{IDHM Educação} \times \text{IDHM Longevidade} \times \text{IDHM Renda})}$$

Os valores do IDHM variam de 0 a 1, sendo que quanto mais próximo de 1, mais desenvolvido é o local. Além disso, são consideradas 5 faixas de desenvolvimento humano, sendo elas: muito baixo ($0 \leftarrow 0,49$), baixo ($0,50 \leftarrow 0,59$), médio ($0,60 \leftarrow 0,69$), alto ($0,70 \leftarrow 0,79$) e muito alto ($0,8 \leftarrow 1,0$).

2.2.2.2 Taxa de Analfabetismo

Segundo a Organização Pan-americana de Saúde (2002), a Taxa de Analfabetismo corresponde ao número de pessoas residentes de 15 anos ou mais de idade que não sabem ler nem escrever. Ele pode ser calculado pela razão: número de pessoas residentes de 15 anos ou mais de idade que não sabem ler nem escrever dividido pelo total residente desta faixa etária multiplicado por 100. Nesse sentido, podemos dizer que quanto menor for o valor da variável, mais desenvolvido é o local em questão.

2.2.2.3 Taxa de Água

A Taxa de Água mede a cobertura de serviços de abastecimento adequado de água à população e é calculada a partir da razão entre a população residente em domicílios particulares permanentes servidos por rede geral de abastecimento de água dividido pela população total residente em domicílios particulares permanentes multiplicada por 100 (Organização Pan-americana de Saúde, 2002). Para esse indicador, quanto maior for o valor, mais desenvolvido é o local em questão.

2.2.2.4 Taxa de Esgoto

A Taxa de Esgoto mede o percentual da população urbana residente em domicílios ligados à rede de esgotamento sanitário. Ela é calculada pela divisão entre população residente em domicílios particulares permanentes servidos por rede coletora ou fossa séptica no domicílio e a população total residente em domicílios particulares permanentes, multiplicada por 100 (Organização Pan-americana de Saúde, 2002). Da mesma forma que o indicador taxa de Água, quanto maior for o valor da variável, mais desenvolvido é o local em questão.

2.2.2.5 Índice de Gini

O Índice de Gini é um indicador de desigualdade muito usado para medir a desigualdade de renda de um local. Seu cálculo é feito com base na Curva de Lorenz, apresentada na Figura 1. Ela indica o quanto o percentual acumulado de renda (eixo Y) varia de acordo com percentual acumulado da população (eixo X). Nishi (2010) mostra que a área entre a reta vermelha e a curva preta é denominada de área de concentração. Se não houvesse concentração, a linha cinza sobreporia a linha azul e resultaria numa distribuição igualitária de renda.

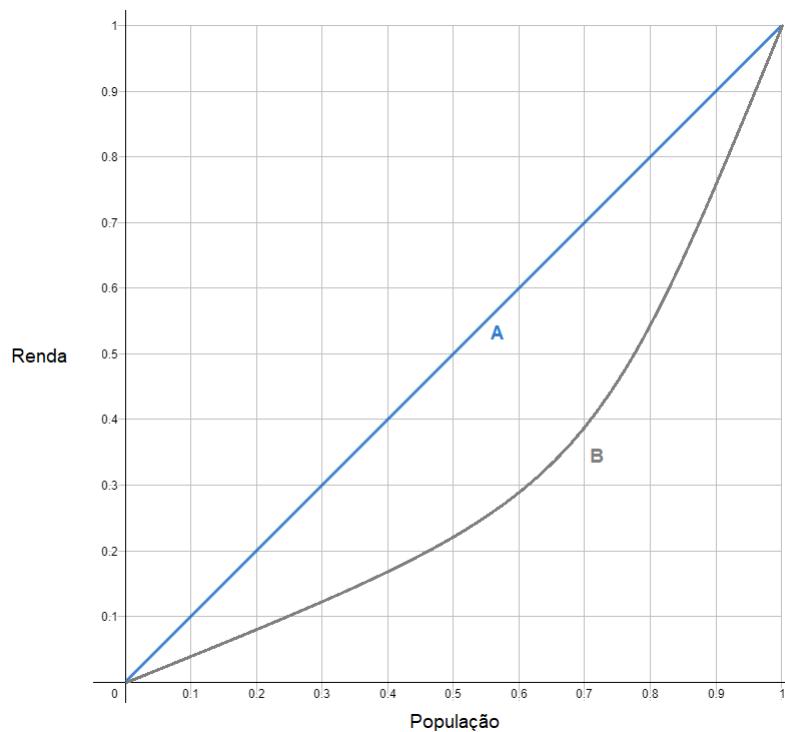


Figura 1 – Curva de Lorenz.

A partir do gráfico mostrado, podemos ver que o Índice de Gini é a razão $G = \frac{A}{A+B}$.

Esse indicador pode ser calculado da seguinte forma:

$$G = 1 - \sum_{i=1}^n (Y_i + Y_{i-1})(X_i - X_{i-1}),$$

em que G é o coeficiente de Gini, X é a proporção acumulada da população e Y é a proporção acumulada da renda. Ele varia de 0 a 1, sendo que quanto mais próximo de 0, menor é a desigualdade de renda do local em questão.

2.2.2.6 Esperança de vida

Como coloca a Organização Pan-americana de Saúde (2002), a esperança de vida corresponde ao número médio de anos de vida esperados para um recém-nascido, mantido o padrão de mortalidade existente na população residente, em determinado espaço geográfico, no ano considerado. Seu cálculo é feito a partir da razão entre o número correspondente a uma geração inicial de nascimentos dividido pelo tempo cumulativo vivido por essa mesma geração. Além disso, temos que quanto maior for o valor da variável, mais desenvolvido é o local em questão.

2.2.2.7 Taxa de Água e Esgoto

Quanto aos indicadores socioeconômicos que serão considerados para os municípios, todos os indicadores descritos acima também serão utilizados e são calculados da mesma forma, com exceção das variáveis Água e Esgoto, que para os municípios são calculadas juntas e correspondem ao percentual de pessoas em domicílios com abastecimento de água e esgotamento sanitário inadequados.

A Taxa de Água e Esgoto corresponde a razão entre as pessoas que vivem em domicílios cujo abastecimento de água não provém de rede geral e cujo esgotamento sanitário não é realizado por rede coletora de esgoto ou fossa séptica e a população total residente em domicílios particulares permanentes multiplicado por 100. São considerados apenas os domicílios particulares permanentes. Sendo assim, podemos dizer que quanto menor for o valor da variável Taxa de água e esgoto, mais desenvolvido é o município.

2.3 Análise de Correlação

A análise de correlação mede o grau de dependência entre duas variáveis. O coeficiente de correlação ρ varia entre -1 e 1 , sendo que quanto mais próximo dos extremos, mais forte é a evidência de que há correlação entre as duas variáveis em questão. Segundo Johnson e Bhattacharyya (2019), quando ρ estiver próximo de 1 , significa que há uma correlação forte e positiva (relação direta entre as variáveis), quando ρ estiver próximo de -1 , indica que há correlação forte e negativa (relação inversa entre as variáveis) e, por

fim, quando ρ estiver próximo de 0, há indícios de correlação fraca ou que não há relação linear entre as variáveis. Existem alguns coeficientes de correlação, mas aqui abordaremos dois: de Pearson e Spearman.

Origuela (2018) mostra que o coeficiente de correlação de Pearson é uma medida que avalia a relação linear entre duas variáveis contínuas. Além disso, ele tem uma hipótese de que as duas variáveis analisadas vêm de uma distribuição Normal ou pelo menos simétrica e assume relação linear entre as variáveis. Em cenários em que essas suposições são falhas, o mais indicado é usar o coeficiente de correlação de postos de Spearman, pelo fato dele não fazer essas suposições.

Sendo assim, vamos utilizar a correlação de Spearman, que é bastante usada na análise de dados. Este coeficiente é uma medida não paramétrica da correlação de postos, ou seja, pela posição dos dados dispostos na forma ordenada, das variáveis x e y , cuja fórmula é dada por:

$$\rho = 1 - \frac{6 \sum_{i=1}^n d_i^2}{(n^3 - n)},$$

em que n é o número de observações e d_i , com $i = 1, \dots, n$, é a diferença entre cada posição de x e y .

Friendly (2002) apresenta uma forma mais clara de visualização para as correlações entre as variáveis, sendo denominado de correlograma. Ele nada mais é do que um gráfico colorido que representa a matriz de correlação, sendo que quanto mais escura for a cor, maior é o indício de que há correlação entre as variáveis estudadas. O correlograma pode ser representado de várias formas, como mostra a Figura 2.

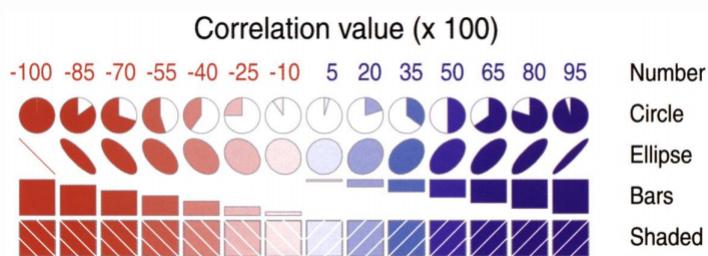


Figura 2 – Representações do correlograma.

Fonte: Friendly (2002)

2.4 Modelos de Regressão

A análise de modelos de regressão é uma importante área da Estatística que permite investigar, explicar e modelar a relação entre uma variável resposta (de interesse) e um conjunto de outras variáveis, chamadas de explicativas, covariáveis ou variáveis independentes (Guimarães, 2008). Essas variáveis são nomeadas assim porque, de certa forma, podem explicar o comportamento da variável de interesse em questão. Entre os

principais usos da análise de regressão, podemos citar: descrição dos dados, estimação dos parâmetros, predição e estimativa de eventos futuros.

Quando a variável que se deseja modelar é uma proporção (porcentagem), ou seja, que está limitada no intervalo $(0, 1)$, devemos levar em conta o suporte do modelo de regressão a ser utilizado. Existem diversos tipos de modelos de regressão que são utilizados para modelar a proporção de alguma variável no intervalo $(0, 1)$. Kieschnick e McCullough (2003) mostraram sete tipos de modelos diferentes para o caso em questão, são eles: modelo normal linear, modelo logito, modelo normal censurado, modelo normal não linear, modelo de quasi-verossimilhança e modelo de regressão que utiliza a distribuição beta. Após mostrarem algumas limitações dos modelos propostos, eles concluíram que o mais adequado seria o modelo de regressão que utiliza a distribuição beta para a variável resposta, que utiliza apenas a função de ligação logito para a esperança condicional (Ospina, 2007).

Dentro desse contexto, Ferrari e Cribari-Neto (2004) propõem um modelo de regressão beta que utiliza uma reparametrização da distribuição beta, para modelar a média da resposta junto com um parâmetro de precisão. Podemos dizer que “a distribuição beta é bastante flexível em problemas nos quais a variável resposta está limitada no intervalo $(0, 1)$, como taxas, proporções e etc” (Diniz e Melo, 2019)[p. 10].

2.4.1 Modelo Beta

A distribuição beta é uma distribuição de probabilidade contínua de parâmetros α e β , cuja função densidade é dada por:

$$f(y; \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1} (1-y)^{\beta-1}, \quad 0 < y < 1, \text{ e } \alpha, \beta > 0, \quad (2.1)$$

em que $\Gamma(\cdot)$ é a função gama, expressa por: $\Gamma(\alpha, \beta) = \int_0^1 t^{\alpha-1} (1-t)^{\beta-1} dt$.

Usamos a notação $Y \sim Beta(\alpha, \beta)$ quando Y é uma variável aleatória que possui distribuição beta de parâmetros α e β . Além disso, temos que a esperança e a variância da densidade beta são dadas, respectivamente, por:

$$E(Y) = \frac{\alpha\beta}{\alpha + \beta}, \quad Var(Y) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}.$$

Ferrari e Cribari-Neto (2004) especificam em seu artigo uma reparametrização da distribuição beta, considerando $\mu = \alpha/(\alpha + \beta)$ e $\phi = \alpha + \beta$. O motivo para a reparametrização pode ser explicado porque queremos uma estrutura de regressão para modelar a média da resposta junto com um parâmetro de precisão. Ou seja, podemos reescrever os parâmetros α e β como $\alpha = \mu\phi$ e $\beta = \phi(1 - \mu)$. Dessa forma, substituindo os novos parâmetros na distribuição beta na Equação (2.1), obtemos a seguinte função de densidade

$$f(y; \mu, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1} (1-y)^{(1-\mu)\phi-1}, \quad 0 < y < 1, \quad (2.2)$$

em que $0 < \mu < 1$ e $\phi > 0$. Como mostrado por Diniz e Melo (2019), a esperança e a variância podem ser reescritas como

$$E(Y) = \mu, \quad Var(Y) = \frac{V(\mu)}{1 + \phi} = \frac{\mu(1 - \mu)}{1 + \phi}, \quad (2.3)$$

em que $V(\mu) = \mu(1 - \mu)$ é a função de variância. Nessa reparametrização, μ é a média de Y e ϕ é o parâmetro de precisão, uma vez que, quanto maior ele for, menor será a variância de Y . Para mais detalhes sobre a reparametrização da beta, consultar (Ferrari e Cribari-Neto, 2004), (Bayer, 2011), (Galarza, 2014), (de Oliveira, 2004), (Ospina, 2007) e (Silva, 2020).

Os procedimentos de modelagem e inferência para a regressão beta são semelhantes aos dos Modelos Lineares Generalizados (MLG), que podem ser vistos em (Cordeiro e Demétrio, 2008) e (Paula, 2004), com exceção ao fato de que a distribuição da variável resposta (distribuição beta reparametrizada, dada pela Equação (2.2)) não faz parte da família exponencial.

2.4.2 Modelo de Regressão Beta

Conforme foi mostrado em Bayer (2011), Ospina (2007) e Ferrari e Cribari-Neto (2004), temos Y_1, \dots, Y_n variáveis aleatórias independentes, em que cada Y_t , com $t = 1, \dots, n$, segue uma distribuição da mesma forma que a Equação (2.2), com média μ_t e o parâmetro de precisão ϕ . O modelo de regressão beta pode ser obtido assumindo que a média de Y_t pode ser escrita como

$$g(\mu_t) = \sum_{i=1}^k x_{ti} \beta_i = \eta_t, \quad (2.4)$$

em que $\beta = (\beta_1, \dots, \beta_k)^\top$ é o vetor dos parâmetros de regressão desconhecidos ($\beta \in \mathbb{R}^k$), x_{t1}, \dots, x_{tk} é o vetor com as observações de k variáveis independentes (covariáveis ou variáveis explicativas) e conhecidas. Vale ressaltar que, quando tiver o intercepto (β_0) no modelo, $x_{t1} = 1$, para todo $t = 1, \dots, n$.

A função de ligação, representada por $g(\cdot)$, é uma função que vincula a média ao preditor linear, além de ser uma função monótona e duas vezes diferenciável. Existem algumas escolhas para as funções de ligações $g(\cdot)$, entre elas: logit, probit, complementar log-log, cauchy e log-log. As fórmulas das funções citadas anteriormente estão descritas na Tabela 1.

Função de ligação	$g(\cdot)$
Logit	$g(\mu) = \log \{\mu/(1 - \mu)\}$
Probit	$g(\mu) = \Phi^{-1}(\mu)$
Complementar log-log	$g(\mu) = \log(1 - \log(1 - \mu))$
Cauchy	$g(\mu) = \tan[\pi(\mu - 0, 5)]$
Log-log	$g(\mu) = -\log \{-\log(\mu)\}$

Tabela 1 – Funções de ligação. $\Phi(\cdot)$ é a função de distribuição acumulada de uma variável aleatória normal padrão.

Suponha que temos uma amostra y_1, \dots, y_n de Y_1, \dots, Y_n , em que a função de densidade para cada observação é dada por $f(y_t; \theta)$ (Equação 2.2), com $\theta = (\beta, \phi)^\top$. A função de verossimilhança para θ é dada por

$$L(\theta; y) = \prod_{t=1}^n f(y_t; \theta),$$

com $y = (y_1, \dots, y_n)^\top$.

O logaritmo da função de verossimilhança é dado por

$$l(\theta) = l(\theta; y) = \sum_{t=1}^n l_t(\mu_t, \phi), \quad (2.5)$$

onde

$$\begin{aligned} l_t(\mu_t, \phi) &= \log \Gamma(\phi) - \log \Gamma(\mu_t \phi) - \log \Gamma((1 - \mu_t)\phi) + (\mu_t \phi - 1) \log y_t \\ &\quad + \{(1 - \mu_t)\phi - 1\} \log(1 - y_t), \end{aligned}$$

em que $\mu_t = g^{-1}(\eta_t)$.

Para encontrar os estimadores de máxima verossimilhança derivamos o logaritmo da função de verossimilhança (Equação 2.5) em relação aos parâmetros (função escore) e igualamos a zero.

Seja $y_t^* = \log \{y_t/(1 - y_t)\}$ e $\mu_t^* = \psi(\mu_t \phi) - \psi((1 - \mu_t)\phi)$, em que $\psi(\cdot)$ é a função digama, ou seja, $\psi(z) = d \log \Gamma(z)/dz$, com $z > 0$. A função escore é dada por $U_\theta(\theta) = (U_\beta(\beta, \phi), U_\phi(\beta, \phi))$ e assim, temos que

$$\begin{aligned} U_\beta(\beta, \phi) &= \frac{dl(\theta)}{d\beta} = \phi X^\top T(y^* - \mu^*) \\ U_\phi(\beta, \phi) &= \frac{dl(\theta)}{d\phi} = \sum_{t=1}^n \mu_t(y_t^* - \mu_t^*) + \log(1 - y_t) - \psi((1 - \mu_t)\phi) + \psi(\phi), \end{aligned}$$

em que X é uma matriz $n \times k$, cuja t -ésima linha é x_t^\top , $T = \text{diag}\{1/g'(\mu_1), \dots, 1/g'(\mu_n)\}$, em que g' é a derivada de primeira ordem de μ_t , $y^* = (y_1^*, \dots, y_n^*)^\top$ e $\mu^* = (\mu_1^*, \dots, \mu_n^*)^\top$.

A matriz de informação de Fisher é formada pelas derivadas de segunda ordem do logaritmo da função de verossimilhança em relação aos parâmetros β e ϕ . Assim, como mostrado em Ferrari e Cribari-Neto (2004), a matriz de informação de Fisher de β e ϕ é dada por

$$K = K(\beta, \phi) = \begin{pmatrix} K_{\beta\beta} & K_{\beta\phi} \\ K_{\phi\beta} & K_{\phi\phi} \end{pmatrix}. \quad (2.6)$$

Análogo ao que Ospina (2007) fez, para a Equação (2.6), temos que $K_{\beta\beta} = \phi X^\top W X$, em que $W = \text{diag}\{w_1, \dots, w_n\}$, com

$$w_t = \phi \{\psi'(\mu_t \phi) + \psi'((1 - \mu_t)\phi)\} \frac{1}{\{g'(\mu_t)\}^2}.$$

De forma semelhante, temos $K_{\beta\phi} = K_{\phi\beta}^\top = X^\top T c$, em que $c = (c_1, \dots, c_n)^\top$, com

$$c_t = \phi \{ \psi'(\mu_t \phi) \mu_t - \psi'((1 - \mu_t) \phi)(1 - \mu_t) \}, \quad t = 1, \dots, n,$$

sabendo que $\psi'(\cdot)$ é a função trigama.

Por fim, temos que $K_{\phi\phi} = \text{tr}(D)$, em que $D = \text{diag}\{d_1, \dots, d_n\}$, com

$$d_t = \psi'(\mu_t \phi) \mu_t^2 + \psi'((1 - \mu_t) \phi)(1 - \mu_t)^2 - \psi'(\phi).$$

Sob as condições de regularidade usuais para estimativa de máxima verossimilhança, quando o tamanho da amostra é grande,

$$\begin{pmatrix} \hat{\beta} \\ \hat{\phi} \end{pmatrix} \sim N_k \left(\begin{pmatrix} \beta \\ \phi \end{pmatrix}, K^{-1} \right),$$

em que $\hat{\beta}$ e $\hat{\phi}$ são os estimadores de máxima verossimilhança de β e ϕ , respectivamente, ou seja, é o valor $\hat{\theta}$ de θ tal que $U(\theta) = 0$, com $\hat{\theta} = (\hat{\beta}, \hat{\phi})$.

As estimativas de máxima verossimilhança de θ não podem ser obtidas de maneira analítica e métodos numéricos são considerados. Consideraremos os métodos BFGS e Fisher Escoring (Cribari-Neto e Zeileis, 2010), implementados no pacote *betareg* (Zeileis *et al.*, 2016) no *software R*.

2.4.3 Modelo de Regressão Beta com dispersão variável

O modelo de regressão beta, proposto por Ferrari e Cribari-Neto (2004) e descrito na Subseção 2.4.2, assume que o parâmetro de precisão ϕ é constante para todas as observações. Entretanto, quando usamos modelos em que a dispersão é constante, quando na verdade ela é variável, podemos gerar perdas consideráveis para o modelo (Ospina, 2007). Por este motivo, Ospina (2007) propôs um modelo de regressão beta com dispersão variável.

Vale ressaltar que, por mais que ϕ seja constante para todas as observações, as variâncias não serão as mesmas, visto que elas variam de acordo com as médias desconhecidas (Equação 2.3). Para mais informações, consultar Bayer (2011), Paolino (2001) e Smithson e Verkuilen (2006).

Continuamos considerando que Y_1, \dots, Y_n são variáveis aleatórias independentes, em que Y_t tem função densidade apresentada na Equação (2.2) com média μ_t , definida na Equação (2.4). Diferente do considerado na Subseção 2.4.2, o parâmetro de precisão agora é dado por ϕ_t , com $t = 1, \dots, n$.

Assim, podemos escrever ϕ_t como

$$h(\phi_t) = \sum_{j=1}^q z_{tj} \gamma_j = v_t, \tag{2.7}$$

em que $\gamma = (\gamma_1, \dots, \gamma_q)^\top$ é um vetor de parâmetros desconhecidos, z_{t1}, \dots, z_{tq} são as observações de q variáveis independentes ($q < n$) fixas e conhecidas e $h(\cdot)$ é uma função de ligação estritamente monótona e duas vezes diferenciável.

Para estimarmos o vetor de parâmetros $\theta = (\beta, \gamma)$, vamos utilizar o método de máxima verossimilhança. Sendo assim, o logaritmo da função de verossimilhança é dado por

$$l(\theta) = l(\beta, \gamma) = \sum_{t=1}^n l_t(\mu_t, \phi_t), \quad (2.8)$$

em que

$$\begin{aligned} l_t(\mu_t, \phi_t) = & \log \Gamma(\phi_t) - \log \Gamma(\mu_t \phi_t) - \log \Gamma((1 - \mu_t) \phi_t) + (\mu_t \phi_t - 1) \log y_t + \\ & + \{(1 - \mu_t) \phi_t - 1\} \log(1 - y_t), \end{aligned}$$

em que $\mu_t = g^{-1}(\eta_t)$ e $\phi_t = h^{-1}(v_t)$, com $\eta_t = \sum_{i=1}^k x_{ti} \beta_i$ e $v_t = \sum_{j=1}^q z_{tj} \gamma_j$.

O desenvolvimento para obtenção da função escore e da matriz de informação de Fisher, além de como obter as estimativas de máxima verossimilhança de θ , é realizado de maneira análoga ao apresentado na Subseção 2.4.2, considerando $\phi_t = h^{-1}(v_t)$ no lugar de ϕ . Maiores detalhes podem ser vistos em (Ospina, 2007).

2.4.4 Testes da razão de verossimilhanças para heteroscedasticidade

Quando consideramos um modelo de regressão beta em que o parâmetro de precisão ϕ_t , com $t = 1, \dots, n$, depende das observações e de parâmetros desconhecidos, é interessante realizar testes para saber quando a hipótese de dispersão fixa é violada (Ospina, 2007).

A hipótese nula de homoscedasticidade dada por $\mathbf{H}_0 : \phi_1 = \dots = \phi_n = \phi$, que é equivalente a $\mathbf{H}_0 : \gamma_{(q-1)} = 0$, em que $\gamma_{(q-1)} = (\gamma_2, \dots, \gamma_q)^\top$, na Equação (2.7), considerando que $z_{t1} = 1$ para $t = 1, \dots, n$.

A estatística da razão de verossimilhanças (RV) é dada por

$$RV = 2 \left\{ l(\hat{\beta}, \hat{\gamma}) - l(\tilde{\beta}, \tilde{\gamma}) \right\},$$

em que $l(\beta, \gamma)$ é a função de log-verossimilhança apresentada na Equação (2.8), $\hat{\beta}$ e $\hat{\gamma}$ são os estimadores de máxima verossimilhança de β e γ , respectivamente, e $(\tilde{\beta}^\top, \tilde{\gamma}^\top)^\top$ é o estimador de máxima verossimilhança restrito de $(\beta^\top, \gamma^\top)^\top$, sob a hipótese nula. Sob condições usuais de regularidade, RV converge em distribuição para $\chi_{(q-1)}^2$ (Ospina, 2007).

2.4.5 Análise de diagnóstico e critério de Akaike

Como colocado por Ospina (2007), a validação de um modelo é essencial para a análise de regressão, visto que ela avalia a qualidade do ajuste. A análise de diagnóstico tem

como objetivo analisar os desvios entre as observações e os valores ajustados do modelo, além de verificar sua influência na análise. Ela vem da análise dos resíduos, com o intuito de identificar valores atípicos (*outliers*), suas influências e verificar a adequação da distribuição da variável resposta. As análises gráficas também são extremamente importantes para a análise de diagnóstico, principalmente o uso de envelopes simulados, proposto por Atkinson (1981), que avalia a qualidade do ajuste do modelo.

2.4.5.1 Análise de resíduos

Resíduo é uma medida usada para verificar diferenças entre os valores reais e ajustados. Como o intuito é verificar essas discrepâncias, faz sentido pensar que os resíduos sejam baseados na diferença $y_t - \widehat{E}(Y_t)$, com $t = 1, \dots, n$. Assim, o resíduo padronizado, proposto por Ferrari e Cribari-Neto (2004), é dado por:

$$r_t = \frac{y_t - \hat{\mu}_t}{\sqrt{\widehat{Var}(y_t)}},$$

em que $\hat{\mu}_t = g^{-1}(x_t^\top \hat{\beta})$ e $\widehat{Var}(y_t) = \hat{\mu}_t(1 - \hat{\mu}_t)/(1 + \hat{\phi})$.

Ospina (2007) propôs novos resíduos baseados no processo iterativo Scoring de Fisher para β quando ϕ é fixo. Esse resíduo é chamado de resíduo ponderado e é dado por

$$r_t^* = \frac{y_t^* - \hat{\mu}_t^*}{\phi \sqrt{v_t}}, \quad (2.9)$$

em que $v_t = \widehat{Var}(Y_t^*)$ e $Y_t^* = \log\{Y_t/(1 - Y_t)\}$. Toda a demonstração deste processo pode ser vista em Ospina (2007) e Espinheira *et al.* (2008).

Espinheira *et al.* (2008) definem o resíduo ponderado padronizado 1 como sendo

$$r_t^p = \phi^{1/2} r_t^* = \frac{y_t^* - \hat{\mu}_t^*}{\sqrt{v_t}}, \quad (2.10)$$

em que r_t^* é o resíduo da Equação (2.9).

Além disso, Espinheira *et al.* (2008) colocaram uma outra alternativa para o resíduo ponderado, porém agora baseando-se na variância de z , em que $z = \hat{\eta} + \hat{W}^{-1}\hat{T}(y^* - \hat{\mu}^*)$. Assim, o resíduo ponderado padronizado 2 pode ser definido como

$$r_t^{pp} = \frac{r_t^*}{\sqrt{\phi^{-1}(1 - h_{tt}^*)}} = \frac{r_t^p}{\sqrt{(1 - h_{tt}^*)}} = \frac{y_t^* - \hat{\mu}_t^*}{\sqrt{v_t(1 - h_{tt}^*)}}, \quad (2.11)$$

em que h_{tt}^* é o t-ésimo elemento da diagonal da matriz H . Para mais detalhes, consultar (Ospina, 2007)[p. 12] e (Espinheira *et al.*, 2008)[p. 409].

Espinheira *et al.* (2008) sugerem, dentre os resíduos padronizados e ponderados existentes, que seja utilizado o resíduo ponderado padronizado 2, uma vez que seus resultados da simulação indicam que a distribuição deste resíduo é mais aproximada pela

distribuição normal padrão do que pelo resíduo ordinário padronizado. Por esse motivo, o resíduo ponderado padronizado 2 será considerado neste trabalho.

Se o gráfico de r_t^p versus o índice das observações (t) não mostrar nenhuma observação fora padrão detectável, significa dizer que o modelo esteja ajustado. Além disso, se o gráfico de r_t^p versus valores ajustados ($\hat{\eta}_t$) mostrar alguma tendência, pode ser um problema de escolha incorreta de função de ligação (Ferrari e Cribari-Neto, 2004).

Os resíduos para modelos de regressão beta dispersão variável são obtidos de maneira análoga e podem ser vistos em (Ferrari *et al.*, 2011).

2.4.5.2 Análise de influência

Quando um modelo é ajustado, aspectos importantes de um modelo podem ser dominados por uma única observação. Uma etapa a ser considerada na análise diagnóstico é a avaliação de pontos influentes, ou seja, pontos que inferem de maneira individual nos resultados inferenciais. No que segue, definimos brevemente a alavanca generalizada e a distância de Cook para o modelo de regressão beta com dispersão fixa. Para o modelo de regressão beta com dispersão variável, o Capítulo 5 da tese de Ospina (2007) é uma ótima referência.

2.4.5.2.1 Alavanca generalizada

Ospina (2007) mostra que a matriz de alavanca generalizada para o estimador de máxima verossimilhança de $\theta = (\beta, \phi)$ é dada por

$$GL_\theta = D_\theta \left(-\frac{\partial^2 l(\theta)}{\partial \theta \partial \theta^\top} \right)^{-1} \frac{\partial^2 l(\theta)}{\partial \theta \partial y^\top},$$

avaliado em $\hat{\theta}$, em que $D_\theta = \partial \mu / \partial \theta^\top$. Para mais detalhes, consultar (Ferrari e Cribari-Neto, 2004).

2.4.5.2.2 Distância de Cook

Uma medida da influência de cada observação nas estimativas do parâmetro de regressão é a distância de Cook (Cook, 1977; Ferrari e Cribari-Neto, 2004). Ela mede o impacto de uma única observação nas estimativas dos coeficientes do modelo de regressão. Uma aproximação para a distância de Cook é dada por:

$$C_t = \frac{h_{tt} r_t^2}{k(1 - h_{tt})^2},$$

em que r_t é o resíduo padronizado, h_{tt} é o elemento na t -ésima linha e t -ésima coluna da matriz chapéu H (Ferrari e Cribari-Neto, 2004) e k é o número de covariáveis do modelo.

2.4.5.3 Gráfico envelope

Quando não conhecemos a distribuição dos resíduos, gráficos semi-normais com o envelope são técnicas de diagnóstico bastante úteis (Atkinson, 1981). A ideia principal do envelope é adicionar bandas de confiança, que são comumente usadas para identificar se os resíduos estão de acordo com o modelo proposto. Em outras palavras, ele verifica se o modelo ajustado se adequa bem aos dados. Um gráfico de envelope que possui uma proporção considerável de pontos (observações) fora das bandas de confiança mostra evidências de que o modelo ajustado não é adequado aos dados do estudo. Ferrari e Cribari-Neto (2004) colocam que as observações que ficam fora dos limites do envelope precisam ser investigadas de uma forma mais aprofundada.

Para ilustrar um gráfico de envelope bem ajustado, vamos pegar um exemplo exposto por Cribari-Neto e Zeileis (2010) com os dados do rendimento de gasolina de Prater (1956): proporção de petróleo bruto convertido em gasolina explicada pela temperatura (em graus Fahrenheit) na qual toda a gasolina evaporou em determinado lote (indicado pelo nível de cinza). A Figura 3 abaixo mostra o ajuste de um modelo em que todas as observações encontram-se dentro das bandas de confiança (envelope). Por este motivo, podemos dizer que há evidências de que o modelo ajustado em questão é adequado para os dados.

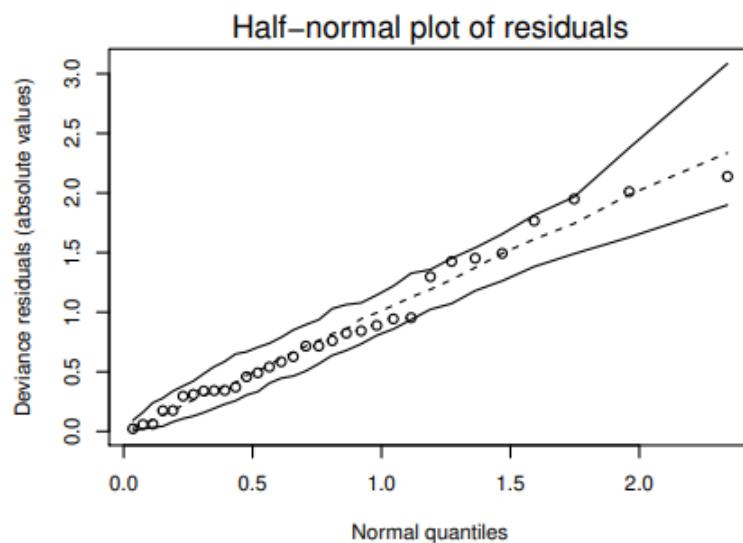


Figura 3 – Exemplo de gráfico de envelope.

2.4.5.4 Critério de Akaike - AIC

Akaike (1974) propôs um critério para seleção de modelos, denominado de critério de Akaike (AIC), definido por

$$AIC = -2 \log[L(\hat{\theta})] + 2k,$$

em que $L(\hat{\theta})$ é a função de verossimilhança estimada, $\hat{\theta}$ é o vetor de parâmetros e k é o número de parâmetros. Além disso, podemos dizer que o modelo mais adequado para os dados do estudo é aquele que apresentar o maior AIC em módulo.

2.4.6 Regressão Beta Inflacionado em Zero

2.4.6.1 Modelo Beta Inflacionado

Como vimos anteriormente, o modelo de regressão beta é bastante útil quando queremos modelar taxas, razões ou proporções. No entanto, se a variável resposta apresentar zeros e/ou uns (intervalos $[0, 1]$, $[0, 1)$ ou $(0, 1]$), este modelo recebe o nome de inflacionado pelo fato de que “a massa de probabilidade em zero e/ou um excede o que é permitido pela distribuição beta” (Ospina Martinez, 2008)[p. 5]. Dessa forma, Diniz e Melo (2019) discutem que o interesse é construir uma estrutura de probabilidade para esses pontos de massa, sendo possível misturar uma distribuição discreta (distribuição Bernoulli), que permita valores 0 e 1, com uma distribuição contínua (distribuição Beta) para o restante das observações (que estão no intervalo $(0, 1)$).

Ospina e Ferrari (2010) propõem o modelo de regressão beta inflacionado em zero e em um. Como o próprio nome do modelo indica, esse modelo permite o cenário que apresenta observações zeros e uns. No que segue, vamos apresentar um caso especial do modelo de Ospina e Ferrari (2010), uma vez que nos dados que motivam esse trabalho só apresentam inflação em zero. Desta forma, apresentamos no que segue o modelo de regressão beta inflacionado em zero.

2.4.6.2 Modelo Beta Inflacionado em zero

Para o caso em que os dados estão no intervalo $[0, 1)$, é importante adicionar à distribuição beta um ponto de massa em zero. Para isto, Ospina Martinez (2008) propôs considerar que a parte contínua dos dados será modelado pela distribuição beta, dada pela Equação (2.2), e o componente discreto, isto é, o ponto de massa, será modelado por meio de uma distribuição degenerada em 0.

Sendo assim, a função de distribuição acumulada da mistura de distribuições é escrita da seguinte forma

$$BI_0(y; \alpha, \mu, \phi) = \alpha \mathbb{I}_{[0,1]}(y) + (1 - \alpha)F(y; \alpha, \mu, \phi), \quad (2.12)$$

em que $\mathbb{I}_A(y)$ é a função indicadora, que assume valor 1 se $y \in A$ e 0 se $y \notin A$, a função $F(y; \mu, \phi)$ é a função acumulada beta $B(\mu, \phi)$ e $\alpha \in (0, 1)$ é o parâmetro de mistura.

Como mostrado em Liberal Pereira (2010), a função densidade de probabilidade

de Y é dada por:

$$bi_0(y; \alpha, \mu, \phi) = \begin{cases} \alpha, & \text{se } y = 0 \\ (1 - \alpha)f(y; \mu, \phi), & \text{se } y \in (0, 1), \end{cases} \quad (2.13)$$

em que $f(y; \mu, \phi)$ é a função de densidade beta, apresentada na Equação (2.2).

Neste caso, temos que a distribuição dada pela Equação (2.12) é chamada de distribuição beta inflacionada no ponto zero (BIZ), ou seja, $Y \sim BIZ(\alpha, \mu, \phi)$ e $\alpha = P(Y = 0)$ (Ospina Martinez, 2008). Além disso, temos que a média e a variância da distribuição BIZ são dadas, respectivamente, por:

$$E(Y) = (1 - \alpha)\mu \quad \text{e} \quad Var(Y) = (1 - \alpha)\frac{V(\mu)}{\phi + 1} + \alpha(1 - \alpha)\mu^2,$$

em que $V(\mu)$ é a função da variância e ϕ é o parâmetro de precisão.

Seja y_1, \dots, y_n uma amostra com n observações de $Y \sim BIZ(\alpha, \mu, \phi)$, podemos calcular a função de verossimilhança de $\theta = (\alpha, \mu, \phi)^\top$ da seguinte maneira:

$$L(\theta) = \prod_{t=1}^n bi_c(y_t; \alpha, \mu, \phi) = L_1(\alpha)L_2(\mu, \phi),$$

em que

$$\begin{aligned} L_1(\alpha) &= \prod_{t=1}^n \alpha^{\mathbb{I}_{\{0\}}(y_t)}(1 - \alpha)^{1 - \mathbb{I}_{\{0\}}(y_t)} = \alpha^{T_1}(1 - \alpha)^{n - T_1}, \\ L_2(\mu, \phi) &= \prod_{t=1}^n f(y_t; \mu, \phi)^{1 - \mathbb{I}_{\{0\}}(y_t)}, \end{aligned}$$

com $T_1 = \sum_{t=1}^n \mathbb{I}_{\{0\}}(y_t)$. O logaritmo da função de verossimilhança de θ pode ser escrito como:

$$l(\theta) = \log[L(\theta)] = l_1(\alpha) + l_2(\mu, \phi),$$

em que

$$\begin{aligned} l_1(\alpha) &= T_1 \log(\alpha) + (n - T_1) \log(1 - \alpha), \\ l_2(\mu, \phi) &= (n - T_1) \log \left\{ \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1 - \mu)\phi)} \right\} + T_2(\mu\phi - 1) + T_3((1 - \mu)\phi - 1), \end{aligned}$$

com $T_2 = \sum_{t:y_t \in (0,1)} \log y_t$ e $T_3 = \sum_{t:y_t \in (0,1)} \log(1 - y_t)$.

Para encontrarmos o vetor escore $U(\alpha, \mu, \phi)$ basta derivar $l(\theta)$ em relação a cada um dos parâmetros desconhecidos. Assim, temos que

$$U(\alpha, \mu, \phi) = (U_\alpha(\alpha), U_\mu(\mu, \phi), U_\phi(\mu, \phi)) = \left(\frac{dl(\theta)}{d\alpha}, \frac{dl(\theta)}{d\mu}, \frac{dl(\theta)}{d\phi} \right),$$

sabendo que

$$\begin{aligned} U_\alpha(\alpha) &= \frac{T_1}{\alpha} - \frac{(n - T_1)}{1 - \alpha}, \\ U_\mu(\mu, \phi) &= \phi \{ (n - T_1)[\psi((1 - \mu)\phi) - \psi(\mu\phi)] + T_2 - T_3 \}, \\ U_\phi(\mu, \phi) &= (n - T_1)[\psi(\phi) - \mu\psi(\mu\phi) - (1 - \mu)\psi((1 - \mu)\phi)] + \mu T_2 - (1 - \mu)T_3, \end{aligned}$$

e $\psi(\cdot)$ é a função digama.

A matriz de informação de Fisher é dada por

$$K(\theta) = \begin{pmatrix} K_{\alpha\alpha} & 0 & 0 \\ 0 & K_{\mu\mu} & K_{\mu\phi} \\ 0 & K_{\phi\mu} & K_{\phi\phi} \end{pmatrix},$$

em que

$$\begin{aligned} K_{\alpha\alpha} &= \frac{1}{\alpha(1 - \alpha)}, \\ K_{\mu\mu} &= (1 - \alpha)\phi^2 \{ \psi'(\mu\phi) + \psi'((1 - \mu)\phi) \}, \\ K_{\mu\phi} = K_{\phi\mu} &= (1 - \alpha)\phi \{ \psi'(\mu\phi)\mu - \psi'((1 - \mu)\phi)(1 - \mu) \}, \\ K_{\phi\phi} &= (1 - \alpha) \{ \mu^2\psi'(\mu\phi) + (1 - \mu)^2\psi'((1 - \mu)\phi) - \psi'(\phi) \}. \end{aligned}$$

Para encontrar os estimadores de máxima verossimilhança de θ , precisamos encontrar as soluções de $U(\alpha, \mu, \phi) = 0$, digamos $\hat{\theta} = (\hat{\alpha}, \hat{\mu}, \hat{\phi})$. As soluções não podem ser obtidas de maneira analítica e métodos numéricos são considerados. Consideraremos o método RS, Rigby and Stasinopoulos (Stasinopoulos *et al.*, 2007), por meio da função *gamlss* (pacote *gamlss*) do *software R*.

Além disso, sob condições de regularidade, a distribuição assintótica do estimador de máxima verossimilhança de θ é dada por:

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{\mathcal{D}} \mathcal{N}_3(0, K(\theta)^{-1}),$$

em que $\hat{\theta} = (\hat{\alpha}, \hat{\mu}, \hat{\phi})$ é o estimador de máxima verossimilhança de $\theta = (\alpha, \mu, \phi)$, $\xrightarrow{\mathcal{D}}$ significa dizer que converge em distribuição, \mathcal{N}_3 é a distribuição normal trivariada e $K(\theta)$ é a matriz de informação de Fisher para a distribuição beta inflacionada em zero.

Uma vez encontrada a distribuição assintótica dos estimadores de máxima verossimilhança, intervalos de confiança para os parâmetros podem ser obtidos. Para mais detalhes sobre a distribuição beta inflacionada em zero $[0, 1]$, a distribuição beta inflacionada em um $(0, 1]$ e a distribuição beta inflacionada em zero e um $[0, 1]$, consultar (Ospina e Ferrari, 2010), (Liberal Pereira, 2010) e (Diniz e Melo, 2019).

2.4.6.3 Modelo de Regressão Beta Inflacionado em Zero

O modelo de regressão beta inflacionado em zero (RBIZ) é definido por (2.13) e pelos componentes sistemáticos:

$$g(\mu_t) = \sum_{j=1}^p x_{tj} \beta_j = x_t^\top \beta,$$

$$h(\alpha_t) = \sum_{j=1}^k z_{tj} \gamma_j = z_t^\top \gamma,$$

em que $\beta = (\beta_1, \dots, \beta_p)^\top$ e $\gamma = (\gamma_1, \dots, \gamma_k)^\top$ são os parâmetros a serem estimados e, ainda, $x_t = (x_{t1}, \dots, x_{tp})^\top$ e $z_t = (z_{t1}, \dots, z_{tk})^\top$ são as variáveis explicativas com dimensões p e k , respectivamente, e $g(\cdot)$ e $h(\cdot)$ são as funções de ligação.

Consideramos aqui as seguintes funções de ligação:

$$g(\mu_t) = \log \left(\frac{\mu_t}{1 - \mu_t} \right),$$

$$h(\alpha_t) = \log \left(\frac{\alpha_t}{1 - \alpha_t} \right).$$

Consequentemente, temos que

$$\alpha_t = P(Y_t = 0) = \frac{\exp(z_t^\top \gamma)}{1 + \exp(z_t^\top \gamma)},$$

$$1 - \alpha_t = P(Y_t \in (0, 1)) = \frac{1}{1 + \exp(z_t^\top \gamma)},$$

$$\mu_t = \frac{\exp(x_t^\top \beta)}{1 + \exp(x_t^\top \beta)}, \quad \text{com } t = 1, \dots, n. \quad (2.14)$$

Nesse caso, o modelo é chamado de modelo de regressão logístico beta inflacionado em zero (Diniz e Melo, 2019). O vetor escore e matriz de informação de Fisher de $\theta = (\beta, \gamma, \phi)$ são obtidas de maneira análoga ao que apresentado na Subseção 2.4.6.2 e mais detalhes podem ser vistos em (Diniz e Melo, 2019).

Ao considerar uma variável aleatória com distribuição inflacionada em zero definida na Equação (2.13), podemos considerar uma reparametrização e vê-lo como um modelo aditivo generalizado para locação, forma e escala, que são os modelos conhecidos como GAMLSS (Stasinopoulos *et al.*, 2007). Ao considerar a reparametrização $\mu = \mu$, $\sigma = \frac{1}{\phi+1}$ e $\nu = \frac{\alpha}{1-\alpha}$, podemos utilizar a função BEINF0() do pacote *gamlss* do R para ajustar o modelo RBIZ.

Consideramos aqui as seguintes funções de ligação:

$$g_1(\mu_t) = \log \left(\frac{\mu_t}{1 - \mu_t} \right) = x_t^\top \beta,$$

$$g_2(\nu_t) = \log(\nu_t) = z_t^\top \gamma,$$

$$g_3(\sigma_t) = \log \left(\frac{\sigma_t}{1 - \sigma_t} \right) = w_t^\top \lambda, \quad (2.15)$$

com $\lambda = (\lambda_1, \dots, \lambda_l)^\top$ e $w_t = (w_{t1}, \dots, w_{tl})^\top$, para $t = 1, \dots, n$. Vale ressaltar que definimos anteriormente o modelo RBIZ considerando o parâmetro de precisão ϕ fixo. No entanto, para ϕ variável, como considerado em (2.15), a inferência é dada de forma análoga e todo desenvolvimento pode ser visto em (Simas *et al.*, 2010). Assim, temos que μ_t e α_t continuam como apresentadas em (2.14), para $t = 1, \dots, n$.

2.4.6.4 Análise diagnóstico

2.4.6.4.1 Resíduo Quantil Aleatorizado

Para o caso de modelos mostrado na Subseção 2.4.6.3, podemos considerar o resíduo quantil aleatorizado dado por

$$r_t^q = \Phi^{-1}(u_t), \quad t = 1, \dots, n,$$

em que u_t é uma variável aleatória uniforme no intervalo $(a_t, b_t]$, sendo a_t e b_t limites da função densidade da distribuição BIZ quando $y \uparrow y_t$ (Diniz e Melo, 2019)[p.24]. De forma semelhante ao apresentado na Subseção 2.4.5.1, temos que para um modelo ajustado, o gráfico de r_t^q versus o índice das observações não deve mostrar nenhuma observação fora do padrão detectável (Diniz e Melo, 2019). Além disso, se o gráfico de r_t^q versus valores ajustados apresentar alguma tendência, pode indicar um problema de escolha incorreta de função de ligação.

2.4.6.4.2 Envelope Simulado

Como vimos anteriormente, Atkinson (1981) propôs o uso de envelopes simulados para avaliar a qualidade do ajuste do modelo. Para o caso do modelo de regressão beta inflacionado em zero, vamos utilizar o pacote do R *hnp*, proposto por Moral *et al.* (2017). Como apresentado em Diniz e Melo (2019), o método consiste em obter e ordenar os valores absolutos de um modelo diagnóstico versus as estatísticas de ordem esperadas de uma distribuição meio-normal, que é dado por

$$\Phi^{-1} \left(\frac{i + n - \frac{1}{8}}{2n + \frac{1}{2}} \right).$$

Como visto em Diniz e Melo (2019)[p. 25], para obter o gráfico do envelope simulado consistem em 5 passos. São eles:

1. Ajustar o modelo;
2. Obter e ordenar as medidas de diagnóstico de interesse;
3. Simular 99 ou mais variáveis resposta usando a mesma matriz de modelos, distribuição de erros e estimativas de parâmetros;

4. Ajustar o modelo a cada variável resposta simulada e extrair e ordenar o diagnóstico do modelo;
5. Calcular os percentis desejados dos valores de diagnóstico simulados em cada valor da estatística de ordem esperada e utilizá-los para formar o envelope.

Para avaliar a qualidade do ajuste através do envelope, temos que quanto mais observações estiverem fora das bandas de confiança, existem maiores evidências de que o modelo ajustado não é adequado aos dados.

3 Análise exploratória para dados estatísticos

Este capítulo será dividido em duas partes: a Seção 3.1 é dedicada à análise das variáveis de incompletude das variáveis consideradas do SINASC (descrita na Seção 2.2) dos anos de 2016 a 2019 e a Seção 3.2 é dedicada à análise das variáveis socioeconômicas de 2016 a 2017 (últimos anos disponíveis e também 2018 para esperança de vida ao nascer). Neste capítulo, todas as análises são realizadas em nível estadual, ou seja, a unidade de análise é a Unidade Federativa do Brasil (UF).

3.1 Análise das variáveis de incompletude

Nesta seção, vamos analisar o comportamento da incompletude das variáveis obstétricas ao longo do tempo, de 2016 a 2019.

A Figura 4 nos mostra a distribuição da incompletude para prematuridade em relação aos estados nos quatro anos disponíveis. Podemos perceber que as maiores porcentagens no ano de 2016 encontram-se na região Norte, com os estados do Pará e de Rondônia. Para os anos de 2017 e 2018, os maiores índices continuam na região Norte, mas agora com os estados do Amapá e também de Rondônia. Para o ano de 2019, observamos que as incompletudes para prematuridade diminuíram bastante em todo o Brasil. Ao observar as regiões, notamos um comportamento bem uniforme na região Sul e Sudeste, com a incompletude baixa para todos os estados que compõem essas regiões. O que mais nos chama a atenção é o estado do Amapá, que contém o maior percentual de dados faltantes para prematuridade em dois anos consecutivos, 2017 e 2018.

A distribuição da incompletude para tipo de gravidez é apresentada na Figura 5. Claramente, podemos perceber que o estado do Ceará foi o que obteve os maiores índices de incompletude para essa variável nos três primeiros anos do estudo. No entanto, este estado teve uma queda considerável para o ano de 2019, mudando completamente o comportamento que vinha apresentando. De maneira geral, as regiões possuem comportamentos semelhantes entre os estados, com exceção do Ceará nos anos de 2016 a 2018, com baixos índices de dados faltantes.

A análise de incompletude para tipo de parto encontra-se na Figura 6. Nitidamente, podemos perceber que o estado do Acre foi o que obteve os maiores índices de incompletude para essa variável nos dois primeiros anos do estudo. Para os dois últimos anos, observamos uma queda para este estado. Em contrapartida, os estados de Roraima e do Mato Grosso do Sul tiveram os menores percentuais para os quatro anos de estudo. Mas no geral, as

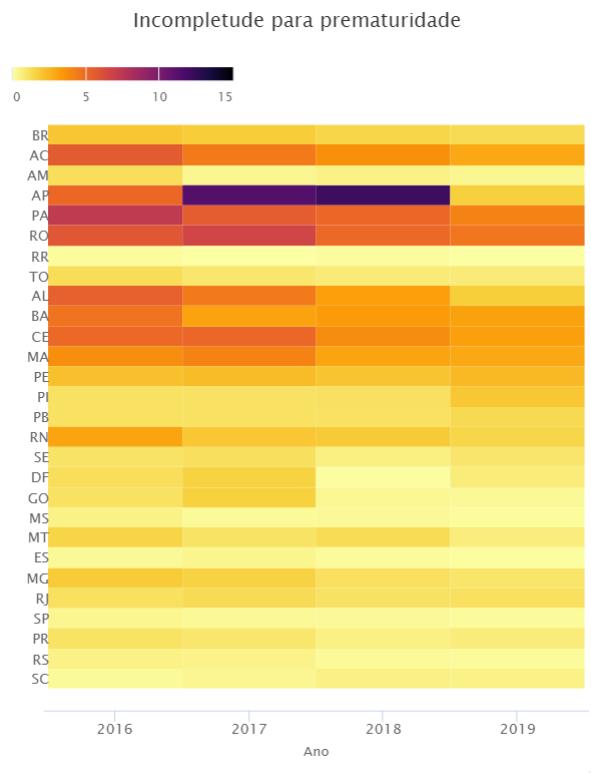


Figura 4 – Mapa de calor para incompletude para prematuridade.

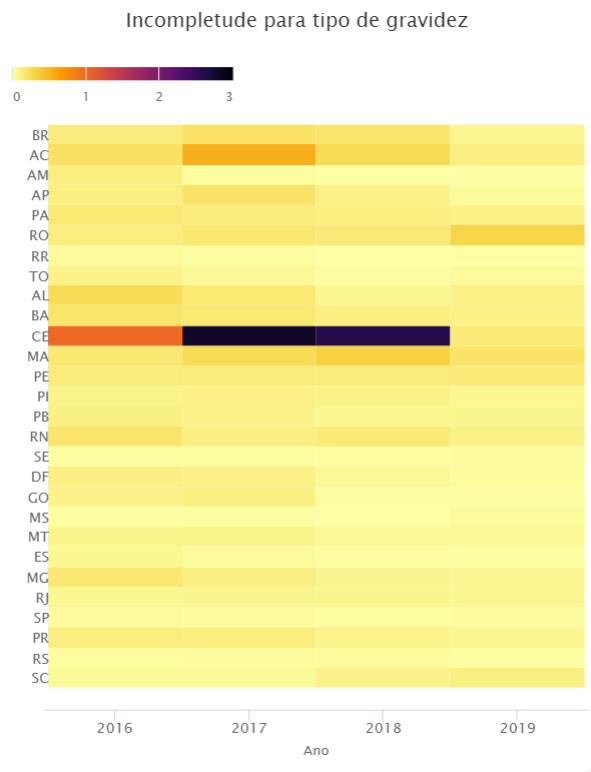


Figura 5 – Mapa de calor para incompletude para tipo de gravidez.

regiões se comportam de maneira uniforme.

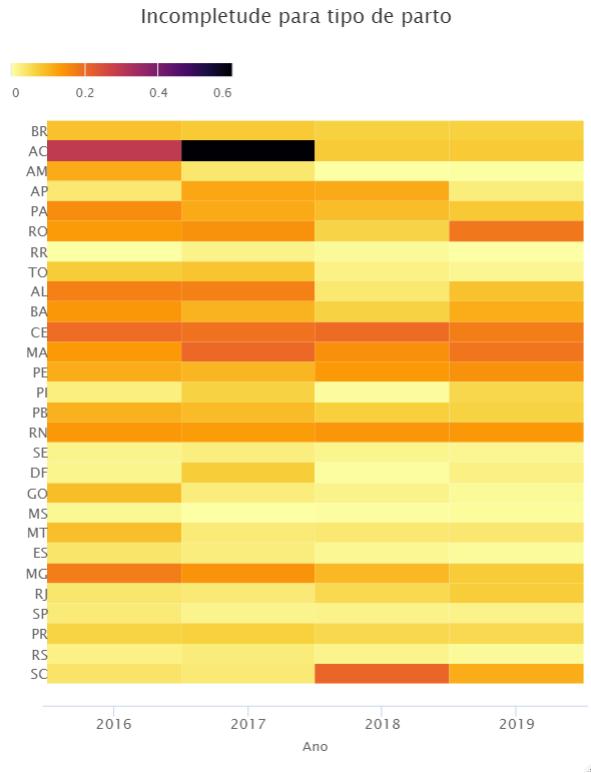


Figura 6 – Mapa de calor para incompletude para tipo de parto.

Em relação à incompletude para a variável de anomalias congênitas, temos a Figura 7. O que mais chama a atenção ao analisarmos o gráfico é o estado do Distrito Federal, que apresenta os maiores índices de incompletude para esta variável para os quatro anos de estudo, disparadamente, estando na faixa entre 24% e 25%. Outro estado que possui altos índices é o de Goiás. Em contrapartida, o percentual para os demais estados do Brasil é relativamente baixo.

Uma variável de extrema importância para avaliar a atenção à saúde da mulher é o número de consultas de pré-natal. Os maiores índices de incompletude para esta variável encontram-se nos estados do Rio Grande do Norte e do Rio de Janeiro, para os quatro anos, e no estado da Paraíba, para o ano de 2017. No geral, analisando a Figura 8, vemos que os estados se comportam de maneira semelhante, com baixa incompletude para essa variável.

Pela Tabela 2, podemos observar a incompletude das variáveis obstétricas consideradas para o Brasil ao longo dos anos. Percebemos que a variável anomalias congênitas foi a que obteve os maiores índices de dados faltantes, para todos os anos de estudo, seguido pela incompletude para prematuridade. Em contrapartida, a incompletude para tipo de parto foi a que obteve os menores valores, para os quatro anos de estudo.

Na Figura 9 apresentamos a análise de correlação de Spearman, apresentado na Seção 2.3, entre as variáveis de incompletude nos anos de 2016 a 2019. Podemos

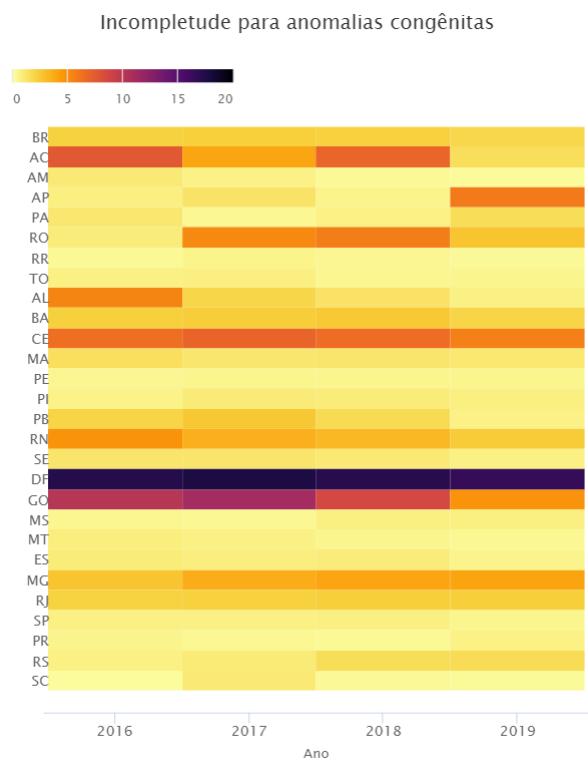


Figura 7 – Mapa de calor para incompletude para anomalias congênitas.

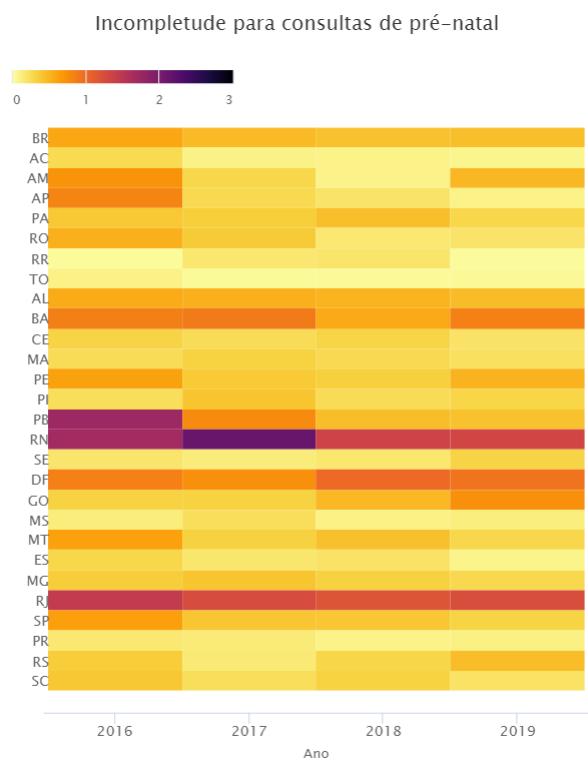


Figura 8 – Mapa de calor para incompletude para consultas de pré-natal.

Variáveis SINASC	Ano			
	2016	2017	2018	2019
Incompletude para prematuridade	1,97	1,74	1,45	1,27
Incompletude para tipo de gravidez	0,13	0,21	0,18	0,07
Incompletude para tipo de parto	0,09	0,07	0,06	0,07
Incompletude para consultas	0,59	0,47	0,42	0,44
Incompletude para anomalias	2,14	2,24	2,18	1,88

Tabela 2 – Incompletude das variáveis para o Brasil entre 2016 e 2019.

observar que as correlações não mudam muito de um ano para o outro, apresentando comportamentos muito semelhantes. De forma geral, encontramos correlações positivas para todas as variáveis do SINASC. Isto implica que quanto maior a incompletude de uma variável, maior será a incompletude da outra variável do SINASC.

Ao considerar a correlação da incompletude para prematuridade, variável de maior interesse neste trabalho, com as demais variáveis, podemos observar que as maiores correlações são com a incompletude para tipo de gravidez e com a incompletude para tipo de parto.

3.2 Análise das variáveis socieconômicas

Um indicador bastante importante para avaliar as condições de vida de uma população é esperança de vida ao nascer, que é medida pela média de anos vividos. Analisando a Figura 10, podemos perceber os maiores índices para esta variável encontram-se nos estados de Santa Catarina, Distrito Federal, Rio Grande do Sul e São Paulo, para os três anos de estudo. Já os menores índices encontram-se nos estados do Alagoas e Maranhão, para os três anos de estudo. O Brasil obteve os valores médios de 75, 72, 75, 99 e 76 anos, para os anos de 2016, 2017 e 2018, respectivamente. No geral, vemos que as regiões se comportam de maneira semelhante, com valores da esperança de vida ao nascer acima de 70,4.

Ao analisar a Figura 11 da distribuição da Taxa de Analfabetismo nos anos 2016 e 2017, podemos destacar, entre os menores valores, os estados do Distrito Federal, com 2,6%, Rio de Janeiro, com 2,7% e São Paulo e Santa Catarina, com 2,6%, para o ano de 2016. Para 2017, temos Rio de Janeiro, com 2,7% e Santa Catarina, com 2,8%. Em relação aos maiores valores, podemos destacar o estado da Paraíba, com 19,4%, para o ano de 2016. Para 2017, destacam-se os estados do Alagoas, com 22,2%, Piauí, com 20,6% e Paraíba, com 20,4%.

Como dito anteriormente, o Índice de Gini indica a desigualdade de renda, ou seja, quanto mais próximo de 1, mais desigual é o lugar. Dessa forma, na Figura 12 podemos

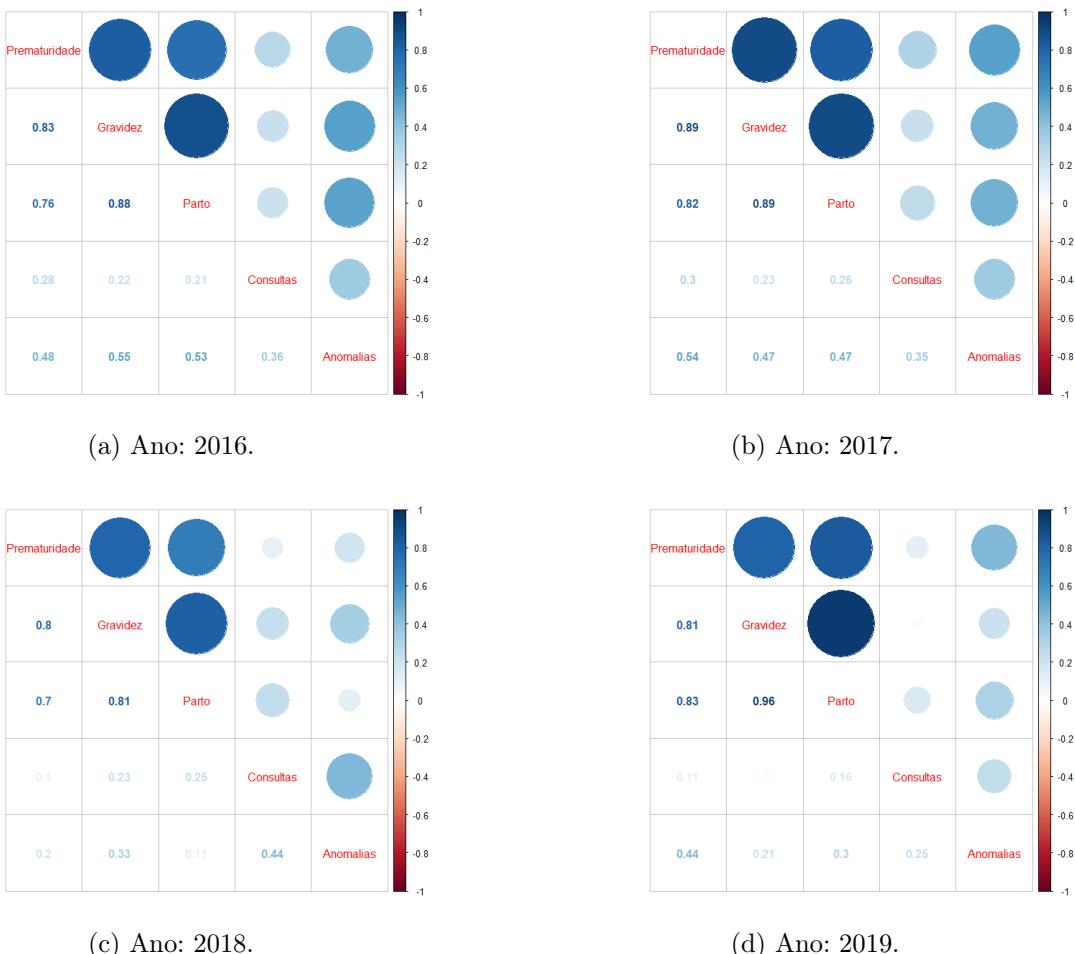


Figura 9 – Correlograma das variáveis de incompletude do SINASC - nível estadual. “Prematuridade” indica a incompletude para prematuridade, “Gravidez” indica a incompletude para tipo de gravidez, “Parto” indica a incompletude para tipo de parto, “Consultas” indica a incompletude para o número de consultas de pré-natal e “Anomalias” indica a incompletude para anomalias congênitas.

observar que o estado que possui menor índice é o de Santa Catarina ($Gini = 0,42$) e os que possuem maiores índices em 2017 são Amazonas, Bahia e Distrito Federal ($Gini = 0,6$). O Brasil obteve índice igual a 0,55 para os dois anos e a maioria dos estados varia entre 0,5 e 0,59.

Quando falamos de IDHM, logo pensamos em lugares com maiores condições de vida para morar. Analisando o mapa de calor na Figura 13, podemos destacar os estados do Distrito Federal ($IDHM = 0,85$) e São Paulo ($IDHM = 0,83$) que obtiveram os maiores índices nos dois anos consecutivos. Em contrapartida, os estados de Alagoas ($IDHM = 0,68$) e Maranhão ($IDHM = 0,68$) foram os que alcançaram o IDHM mais baixo do país. De maneira geral, quando comparamos os IDHM's das regiões com o IDHM do Brasil nos mesmos anos ($IDHM = 0,78$), que é classificado como alto, elas estão bem próximas da marca do país. Além disso, nos chama a atenção o fato de o Distrito Federal possuir

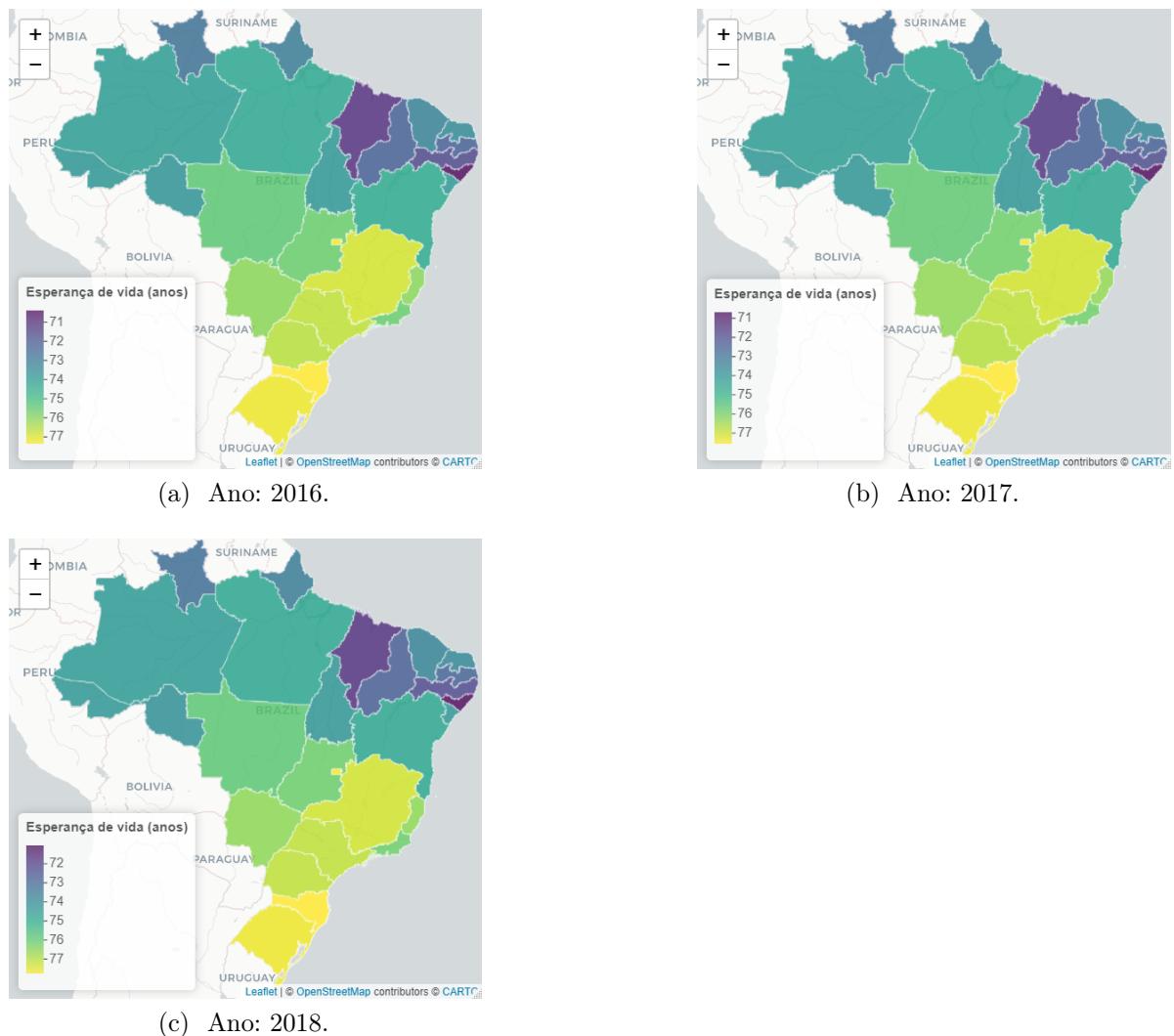


Figura 10 – Mapa dos estados para esperança de vida ao nascer.

o maior IDHM do país e, quando falamos de desigualdade de renda, também possuir o maior Índice de Gini.

Semelhante ao IDHM, temos a Figura 14 abaixo para o IDHM Educação. Pelo mapa, podemos destacar os estados de São Paulo (IDHM Educação = 0,84, em 2016 e IDHM Educação = 0,85, em 2017) e Distrito Federal (IDHM Educação = 0,82, em 2016 e IDHM Educação = 0,80) que obtiveram os maiores índices nos dois anos consecutivos. Em relação aos menores índices, estão os estados de Sergipe (IDHM Educação = 0,63, em 2016 e IDHM Educação = 0,64, em 2017) e Alagoas (IDHM Educação = 0,64 para os dois anos).

Apresentamos a Figura 15 referente ao IDHM Longevidade. Analisando os mapas, podemos destacar o estado do Distrito Federal (IDHM Longevidade = 0,89 para os dois anos) que obteve o maior índice nos dois anos consecutivos. Em relação aos menores índices, destaca-se o estado do Maranhão (IDHM Longevidade = 0,76 para os dois anos).

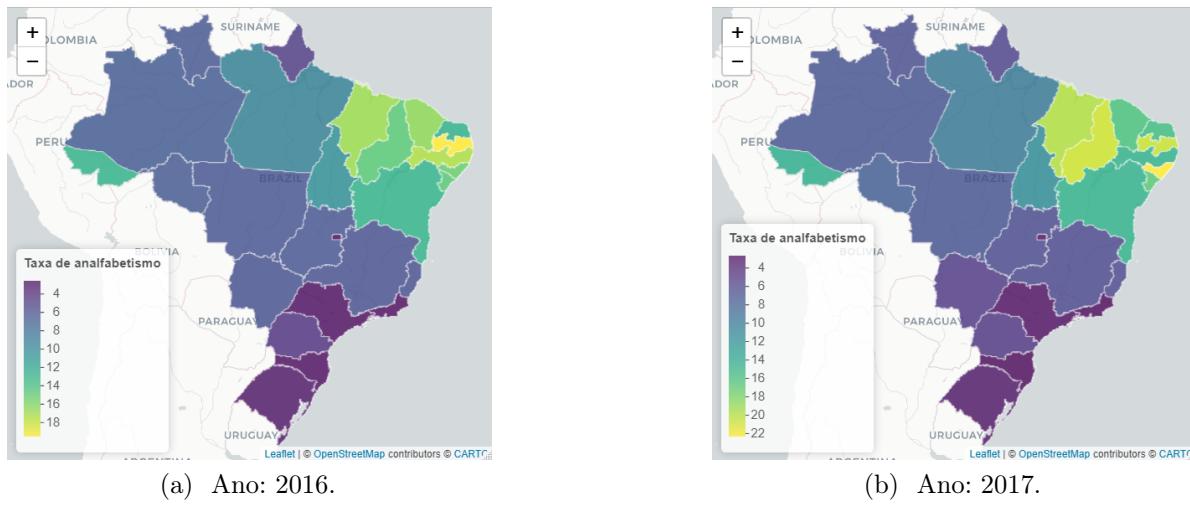


Figura 11 – Mapa dos estados para taxa de analfabetismo.

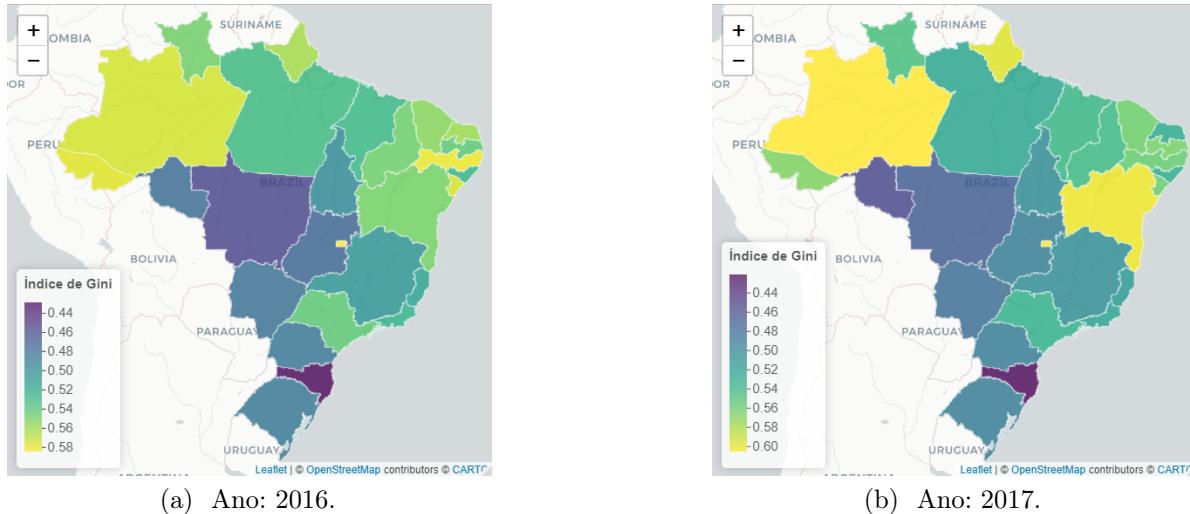


Figura 12 – Mapa dos estados para Índice de Gini.

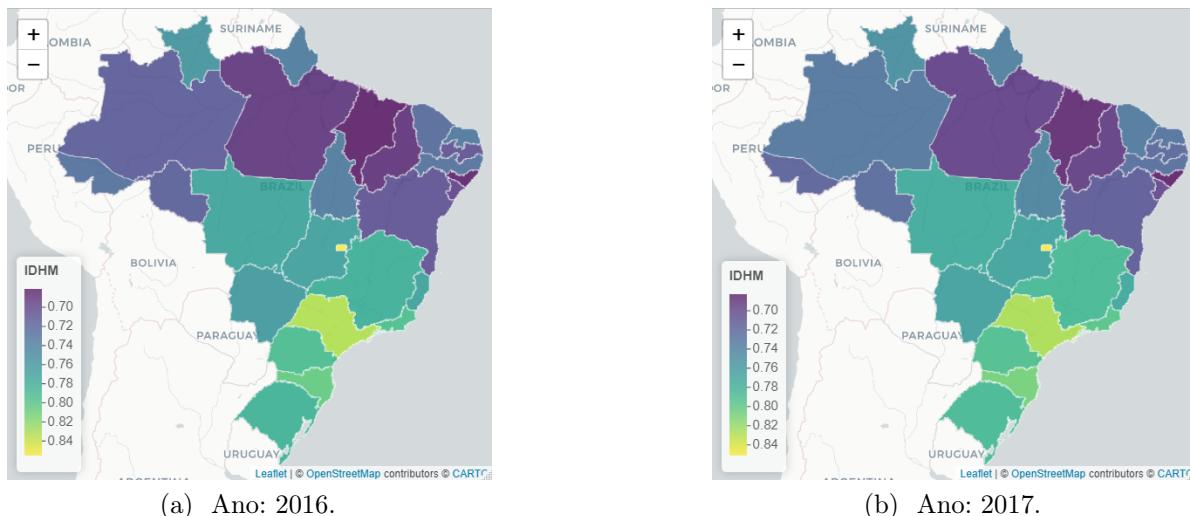
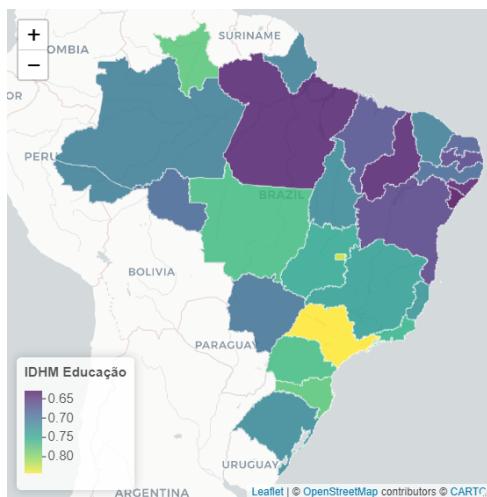
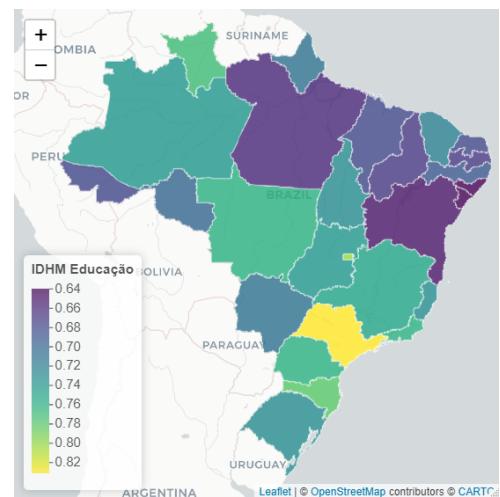


Figura 13 – Mapa dos estados para IDHM.

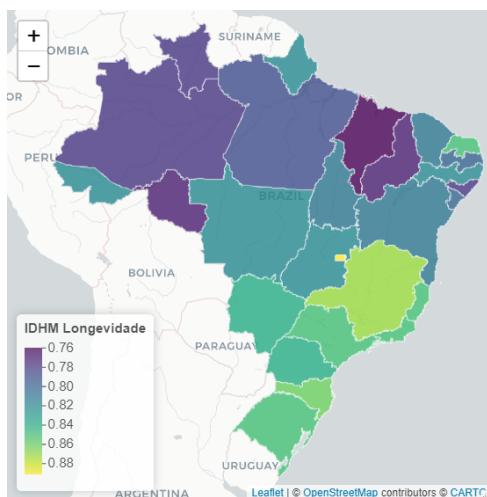


(a) Ano: 2016.

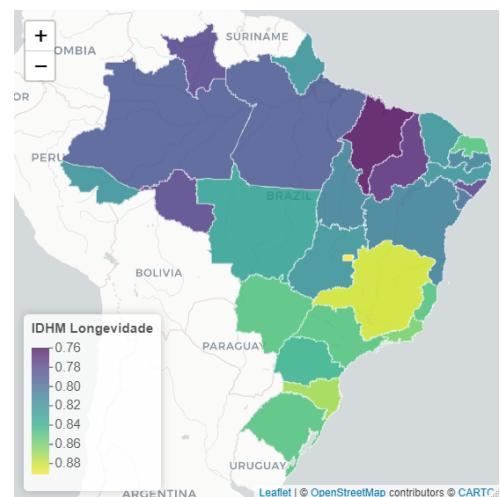


(b) Ano: 2017.

Figura 14 – Mapa dos estados para IDHM Educação.



(a) Ano: 2016.



(b) Ano: 2017.

Figura 15 – Mapa dos estados para IDHM Longevidade.

A Figura 16 apresenta os mapas em relação ao IDHM Renda. Podemos observar que os mesmos estados citados para o IDHM Longevidade são destaques aqui também. O estado do Distrito Federal (IDHM Renda = 0,85, para 2016 e IDHM Renda = 0,86, para 2017) foi o que obteve o maior índice nos dois anos consecutivos. E o estado do Maranhão (IDHM Longevidade = 0,76, para os dois anos) obteve o menor índice nos dois anos.

Como sabemos, a variável água mede o percentual de pessoas que possuem abastecimento de água adequado, ou seja, quanto mais próximo de 100%, melhor é o lugar. Dessa forma, na Figura 17 podemos observar que os estados que possuem os maiores índices são Roraima (Água = 99,69 em 2016 e em 2017) e Distrito Federal (Água = 99,06 em 2016 e Água = 98,71), considerados como os melhores estados com relação a esta variável. Em contrapartida, o que possui o menor índice é o estado do Amapá (Água = 38,46 em 2016

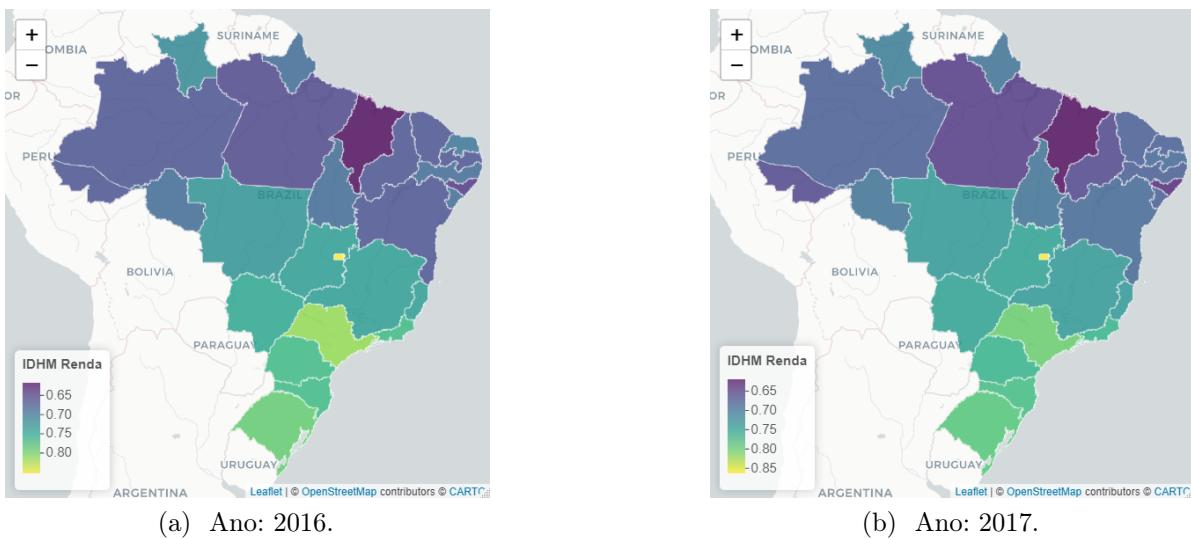


Figura 16 – Mapa dos estados para IDHM Renda.

e Água = 40,44 em 2017).

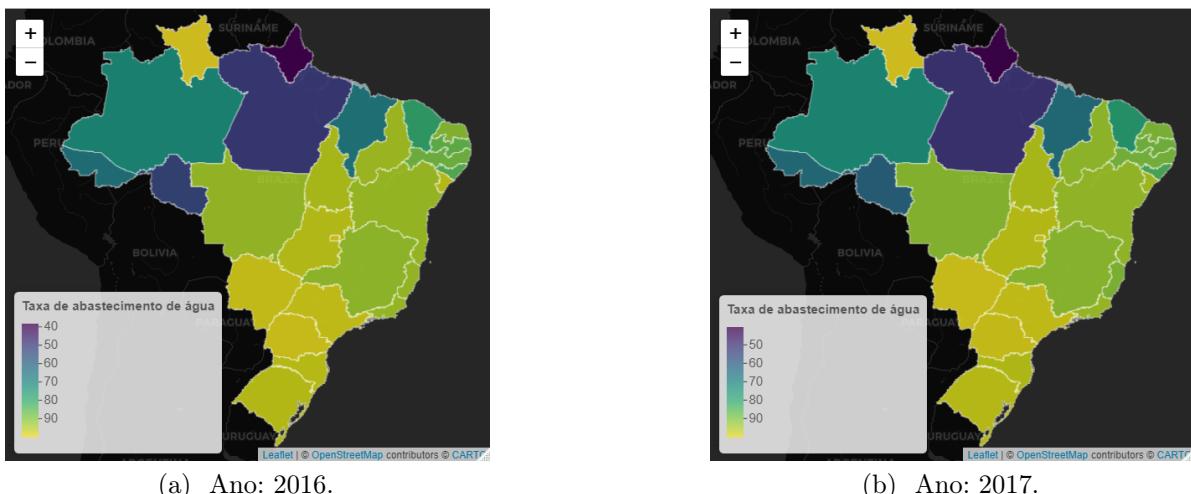


Figura 17 – Mapa dos estados para o percentual de acesso ao abastecimento de água.

Ao observar a Figura 18, podemos observar que o estado que possui a maior Taxa de Esgoto é o de São Paulo (Esgoto = 98,03 em 2016 e Esgoto = 97,94 em 2017), podendo ser considerado como o melhor estado nesse indicador. Em contrapartida, os que possuem menores índices são os estados de Rondônia (Esgoto = 5,34 em 2016 e Esgoto = 5,95 em 2017) e Amapá (Esgoto = 6,36 em 2016 e Esgoto = 7,36 em 2017).

Como visto nas análises acima, as variáveis socieconômicas pouco mudaram nos dois anos disponíveis (três para esperança de vida). Por esse motivo, para o objetivo principal deste trabalho de correlacionar a incompletude para prematuridade com as variáveis socioeconômicas, será considerada a incompletude para prematuridade de 2019 e os dados socioeconômicos mais recentes disponíveis em nível estadual, ou seja, o ano

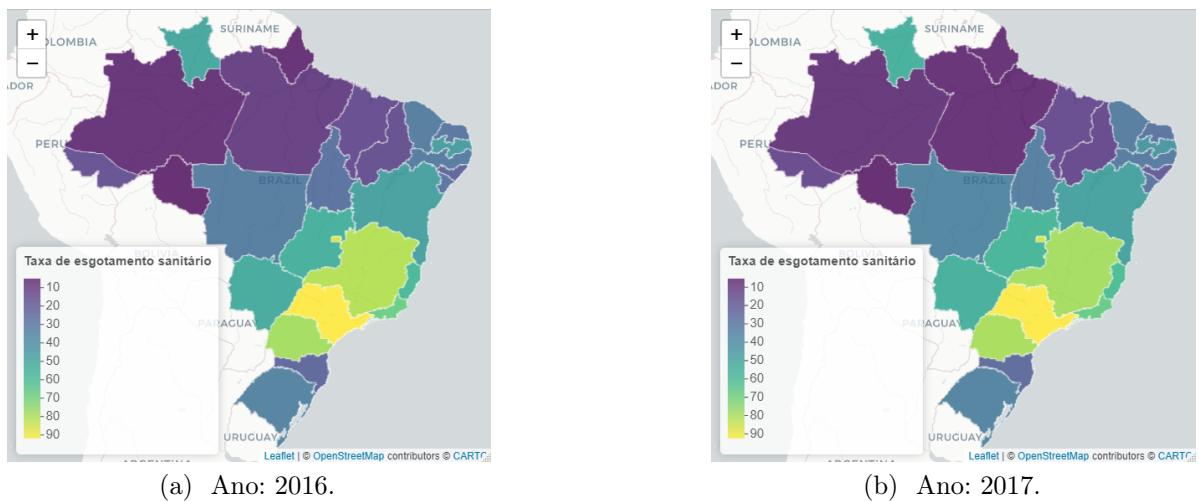


Figura 18 – Mapa dos estados para o percentual de acesso ao esgotamento sanitário.

de 2018 para esperança de vida ao nascer e o ano de 2017 para os demais indicadores socieconômicos.

Na Figura 19 apresentamos o correlograma entre as variáveis socieconômicas do ano de 2017 (ano de 2018 para esperança de vida) e a incompletude de prematuridade do ano de 2019. Os coeficientes de correlação de Spearman entre a incompletude para prematuridade e os indicadores IDHM, IDHM Educação, IDHM Longevidade, IDHM Renda, Taxa de Água, Taxa de Esgoto e Esperança de vida são negativos, variando entre $\rho = -0,44$ (para IDHM Longevidade) e $\rho = -0,72$ (para Taxa de Água). Isto significa que, quanto mais alto os valores dessas variáveis, menor será a incompletude para prematuridade (relação inversa).

Em contrapartida, para as variáveis Índice de Gini e Taxa de Analfabetismo encontramos coeficientes de correlação e Spearman positivos. Assim, quanto mais alto os valores dessas variáveis, mais alto será também a incompletude para prematuridade.

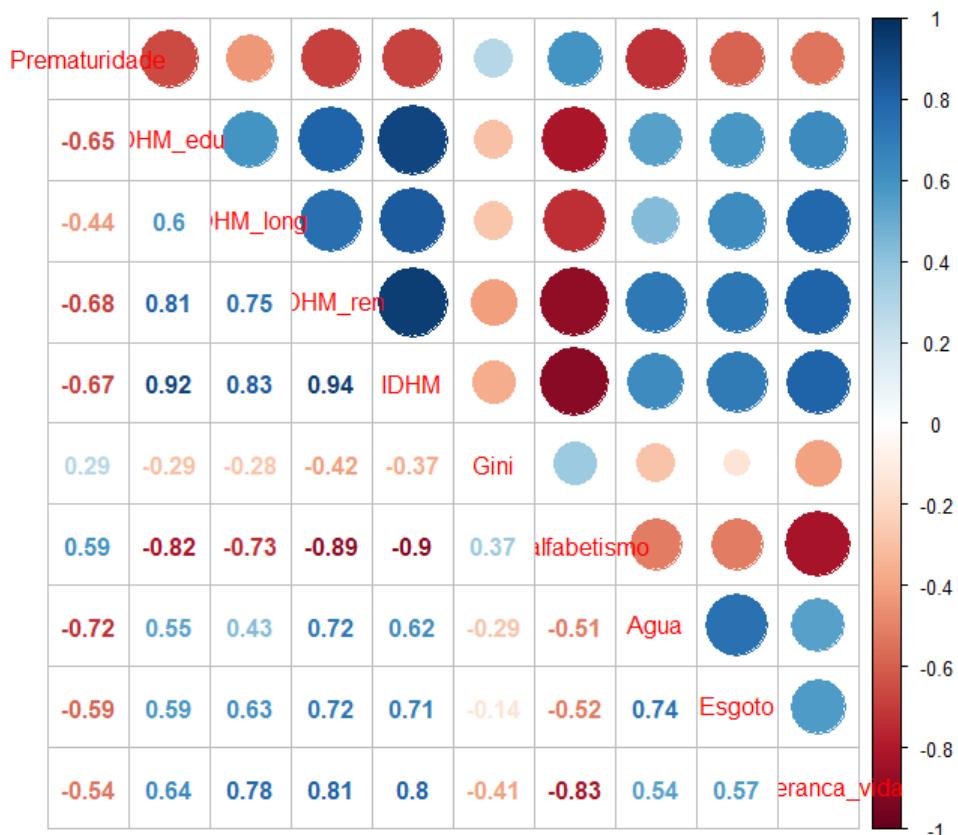


Figura 19 – Correlograma dos indicadores socioeconômicos (2017 ou 2018) e incompletude para prematuridade (ano 2019) - nível estadual. “Prematuridade” indica a incompletude para prematuridade.

4 Análise exploratória para dados municipais

Este capítulo será dividido em duas partes: a Seção 4.1 é dedicada à análise das variáveis de incompletude das variáveis consideradas do SINASC (descritas na Seção 2.2) e a Seção 4.2 é dedicada à análise das variáveis socioeconômicas para o ano de 2010 (ano mais recente com essas informações, referente ao Censo Demográfico).

As análises nesse capítulo são realizadas em nível municipal, ou seja, a unidade de análise é o município.

4.1 Análise das variáveis de incompletude

No que segue, apresentamos os mapas das variáveis de incompletude do SINASC do ano de 2019.

A Figura 20 nos mostra a distribuição da incompletude para prematuridade em relação aos municípios no ano de 2019. Podemos perceber que a grande maioria dos municípios possuem um percentual baixo de incompletude para prematuridade, 3225 municípios possuem taxa de incompletude menor do que 0,1%. Em relação aos maiores valores, podemos destacar os municípios com mais de 70% de incompletude para a variável de interesse, são eles: Iguáí-BA, com 90,69%, Ipirá-BA, com 81,56%, José Gonçalves De Minas-MG, com 77,14%, Trajano De Moraes-RJ, com 74,76%, Turmalina-MG, com 74% e Veredinha-MG, com 71,66%.

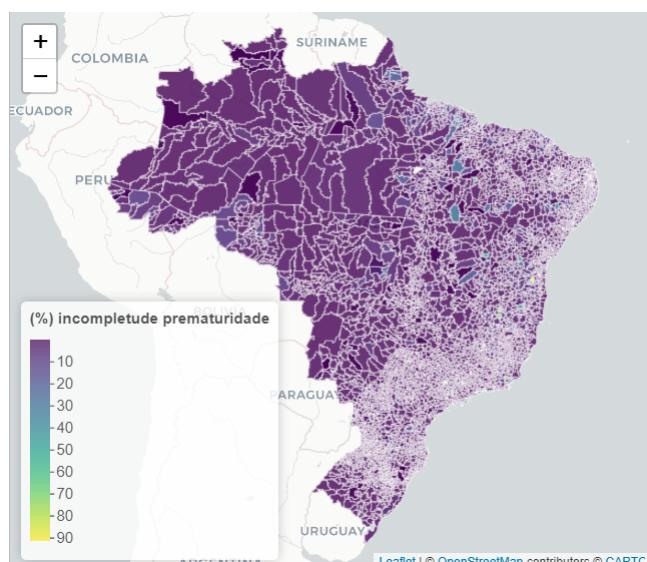


Figura 20 – Mapa dos municípios para incompletude para prematuridade.

Podemos avaliar a distribuição da variável de porcentagem de incompletude para tipo de gravidez a partir da Figura 21. De forma semelhante ao gráfico anterior, podemos perceber que a grande maioria dos municípios possuem um percentual baixo de incompletude para tipo de gravidez, 5437 municípios possuem taxa de incompletude menor do que 0,1%. Em relação ao maior valor, podemos destacar o município de Senador José Bento-MG, que possui uma taxa de incompletude para tipo de gravidez igual a 9,09%.

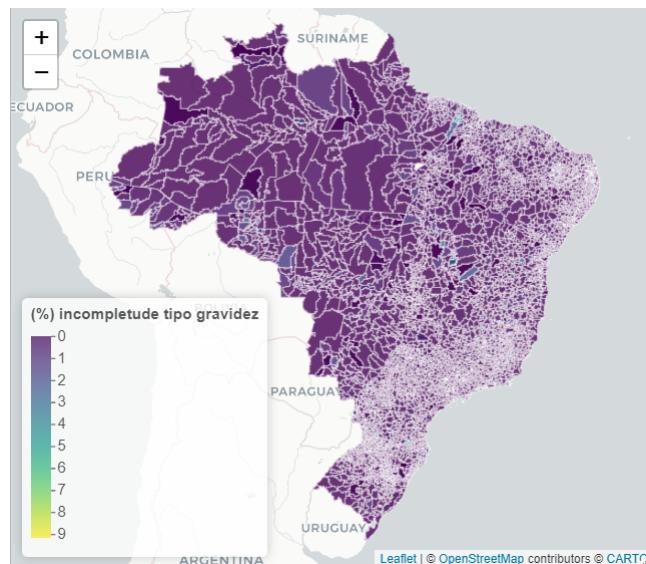


Figura 21 – Mapa dos municípios para incompletude para tipo de gravidez.

A distribuição da incompletude para tipo de parto encontra-se na Figura 22. A grande maioria dos municípios apresenta baixos valores para essa variável, 5501 municípios possuem incompletude menor do que 0,1%. Em relação ao maior valor, podemos destacar o município de Leme Do Prado-MG, com um percentual de incompletude para tipo de parto igual a 8,10%.

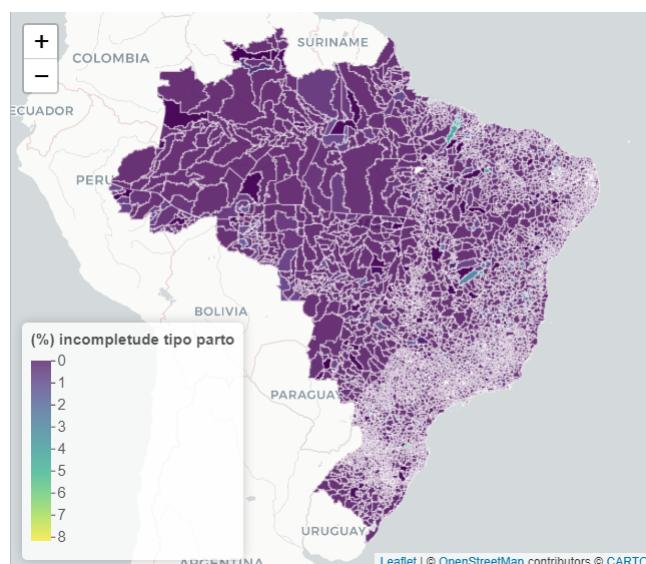


Figura 22 – Mapa dos municípios para incompletude para tipo de parto.

Em relação à incompletude para a variável de anomalias congênitas, temos a Figura 23. Aqui, também temos a maioria dos municípios com baixos valores para essa variável, 3522 municípios com incompletude menor do que 0,1%. O que difere este dos outros gráficos apresentados é a escala, que aqui varia de 0% até 90%. Fazendo um filtro, podemos destacar os municípios com mais de 90% de incompletude para a variável de interesse, são eles: Barão De Monte Alto-MG, com 97,43%, Patrocínio Do Muriaé-MG, com 96,66%, Muriaé-MG, com 96,13%, Nova Independência-SP, com 95%, Miradouro-MG, com 93%, Bela Vista De Minas-MG, com 91,81%, João Monlevade-MG, com 91,30%, Rio Piracicaba-MG, com 90,60%, Nova Era-MG, com 90,57%, Paracatu-MG, com 90,33% e Antônio Prado De Minas-MG, com 90%. Além disso, há 47 municípios com incompletude maior do que 50%. Para estes municípios, poderíamos classificá-los como “muito ruim”, de acordo com o proposto por Gabriel *et al.* (2014) e Romero e Cunha (2006).

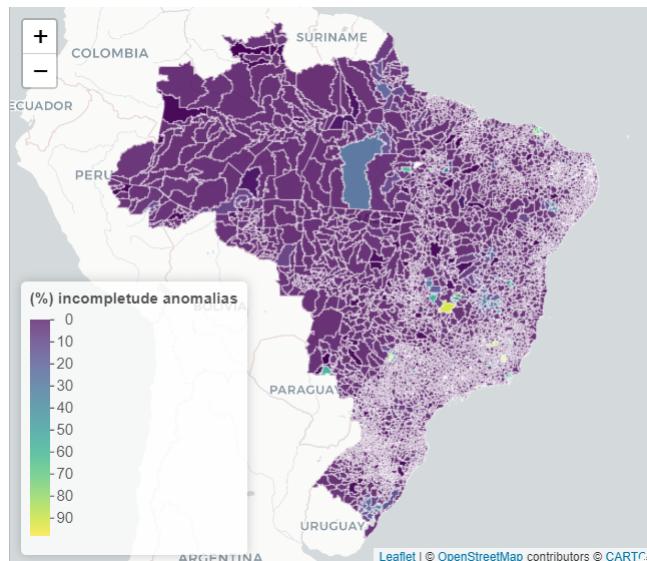


Figura 23 – Mapa dos municípios para incompletude para anomalias congênitas.

Pela Figura 24, observamos que a grande maioria dos municípios possui baixos valores para a incompletude para número de consultas de pré-natal, 4691 municípios possuem incompletude menor do que 0,1%. Os maiores índices para esta variável encontram-se nos seguintes municípios: General Carneiro-PR e General Carneiro-MT, com 26,27%, Arroio Do Padre-RS, com 15%, Santo Antônio Do Leste-MT, com 12,12%, Morro Redondo-RS, com 10,41 e Lagoa Santa-MG, com 10%.

Na Figura 25 apresentamos a análise de correlação de Spearman entre as variáveis de incompletude nos anos de 2016 a 2019. Podemos observar que as correlações não mudam muito de um ano para o outro, apresentando comportamentos muito semelhantes. De forma geral, encontramos correlações positivas para todas as variáveis do SINASC. Isto implica que quanto maior a incompletude de uma variável, maior será a incompletude da outra variável do SINASC. Ao considerar a correlação da incompletude para prematuridade com

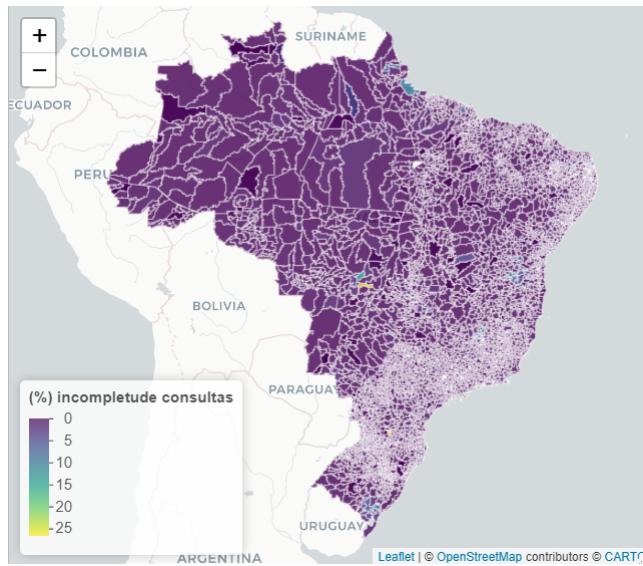


Figura 24 – Mapa dos municípios para incompletude para consultas de pré-natal.

as demais variáveis, podemos observar que as maiores correlações são com a incompletude para tipo de gravidez e com a incompletude para tipo de parto.

4.2 Análise das variáveis socieconômicas

Na Figura 26 apresentamos o mapa dos municípios para o indicador de Taxa de Analfabetismo do Censo de 2010. Percebemos que a grande maioria dos municípios possui taxa de analfabetismo entre 5% e 25% (3740 municípios). Dentre os menores valores, podemos destacar o município de Feliz-RS, que possui taxa de analfabetismo igual a 0,95%. Se filtrarmos uma taxa acima de 40%, encontramos 31 municípios. Em relação aos maiores valores, destaca-se o município de Alagoinha do Piauí-PI, com uma taxa igual a 44,40%.

Para representar o indicador Taxa de água e esgoto, temos a Figura 27. Podemos perceber que a maioria dos municípios possui taxa de água e esgoto entre 0% e 10%, sendo 435 municípios com valor de 0%. Há 14 municípios com mais de 70% dos domicílios sem acesso à abastecimento de água e esgotamento sanitário, com destaque para o município de Chaves-PA, com 85,36% para essa variável.

A maioria dos municípios possui valores para esperança de vida entre 70 e 74 anos (Figura 28). Entre os maiores valores, podemos destacar os municípios de Blumenau-SC e Brusque-SC, ambos com média de 78,64 anos. Em contrapartida, os municípios que possuem os menores valores são: Cacimbas-PB e Roteiro-AL, com média de 65,30 anos.

Ao analisar a Figura 29, percebemos que a maioria dos municípios possui valor do índice de Gini entre 0,45 e 0,6 (aproximadamente 4033 municípios). Dentre os municípios com os menores índices, podemos destacar: Santa Maria do Herval-RS, com Gini = 0,30, São Vandelino-RS e Vale Real-RS, com Gini = 0,29 e Botuvera-SC e São José do Hortêncio-



Figura 25 – Correlograma das incompletude das variáveis do SINASC - nível municipal. “Prematuridade” indica a incompletude para prematuridade, “Gravidez” indica a incompletude para tipo de gravidez, “Parto” indica a incompletude para tipo de parto, “Consultas” indica a incompletude para o número de consultas de pré-natal e “Anomalias” indica a incompletude para anomalias congênitas.

RS, com Gini = 0,28. Em contrapartida, em relação aos municípios com os maiores índices, destacam-se: Itamarati-AM e São Gabriel da Cachoeira-AM, com Gini = 0,80, Isaías Coelho-PI, com Gini = 0,79 e Alto Parnaíba-MA, Jequitibá-MG, Santa Rosa dos Purus-AC e Uiramutã-RR, com Gini = 0,78.

Para o IDHM (Figura 30), a maioria (3631) dos municípios possui indicador entre 0,50 e 0,70. Dentre os valores extremos, temos que 26 municípios possuem IDHM muito baixo ($0 \leftarrow 0,49$) e 45 municípios possuem IDHM muito alto ($0,80 \leftarrow 1,0$). Podemos destacar os municípios de São Caetano do Sul-SP, com IDHM = 0,86, Águas de São Pedro-SP, com IDHM = 0,854, Fernando Falcão-MA, com IDHM = 0,44 e Melgaço-PA, com IDHM = 0,42.

O mapa do IDHM Educação pode ser visto na Figura 31. Percebemos que a maioria

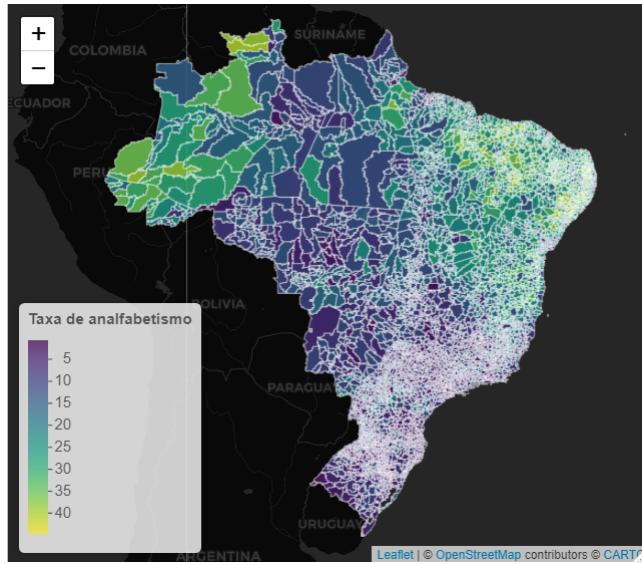


Figura 26 – Mapa dos municípios para taxa de analfabetismo.

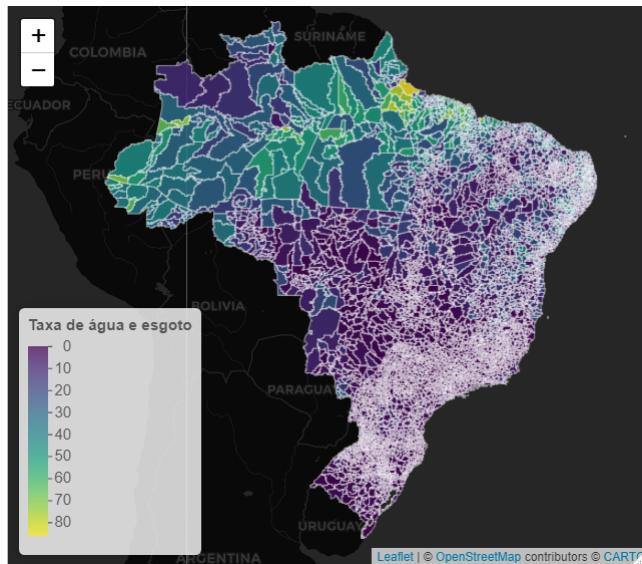


Figura 27 – Mapa dos municípios para taxa de água e esgoto.

dos municípios possui IDHM Educação entre 0,40 e 0,60. Entre os valores extremos, podemos destacar os municípios de Águas de São Pedro-SP, com IDHM Educação = 0,83, São Caetano do Sul-SP, com IDHM Educação = 0,81, Chaves-PA, com IDHM Educação = 0,23 e Melgaço-PA, com IDHM Educação = 0,21.

A Figura 32 é referente ao mapa para IDHM Longevidade. Podemos ver que a escala varia de 0,7 a 0,85, ou seja, todos os municípios possuem valores altos para o IDHM Longevidade. Entre os valores extremos, destacam-se os seguintes municípios: Balneário Camboriú-SC, Blumenau-SC, Brusque-SC e Rio do Sul-SC, com IDHM Longevidade = 0,89, e Cacimbas-PB e Roteiro-AL, com IDHM Longevidade = 0,67.

Por fim, temos a Figura 33 referente ao IDHM Renda. Observamos que a escala varia de 0,40 a 0,8, a maioria dos municípios encontra-se na faixa de 0,55 e 0,7, aproximadamente

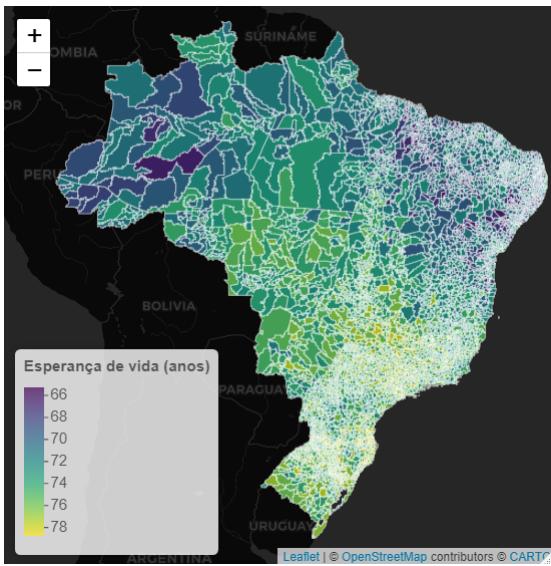


Figura 28 – Mapa dos municípios para esperança de vida ao nascer.

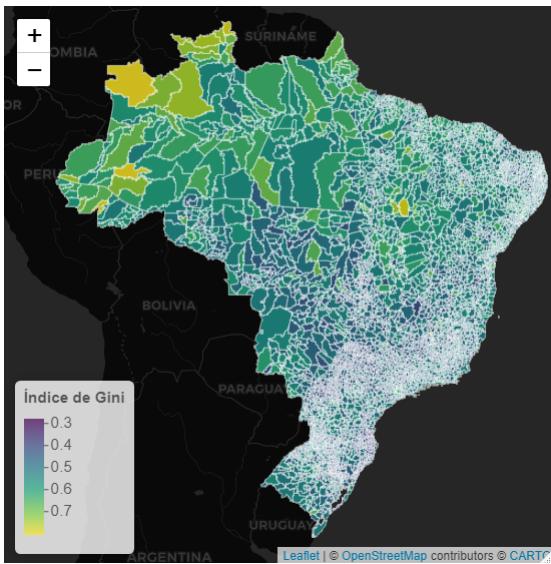


Figura 29 – Mapa dos municípios para Índice de Gini.

3123. Entre os valores extremos, destacam-se os seguintes municípios: São Caetano do Sul-SP, com IDHM Renda = 0,89, Niterói-RJ, com IDHM = 0,85, Belágua-MA e Fernando Falcão-MA, com IDHM Renda = 0,42 e Marajá do Sena, com IDHM Renda = 0,40.

Na Figura 34 apresentamos o correograma entre as variáveis socieconômicas do ano de 2010 e a incompletude para prematuridade do ano de 2019. Os coeficientes de correlação de Spearman entre a incompletude para prematuridade e os indicadores IDHM, IDHM Educação, IDHM Longevidade, IDHM Renda e Esperança de vida são negativos, variando entre $\rho = -0,28$ (para IDHM Educação) e $\rho = -0,33$ (para IDHM Longevidade, IDHM Renda e Esperança de vida). Isto significa que, quanto mais alto os valores dessas variáveis, menor será a incompletude para prematuridade.

Em contrapartida, para as variáveis Índice de Gini, Taxa de Analfabetismo e Taxa

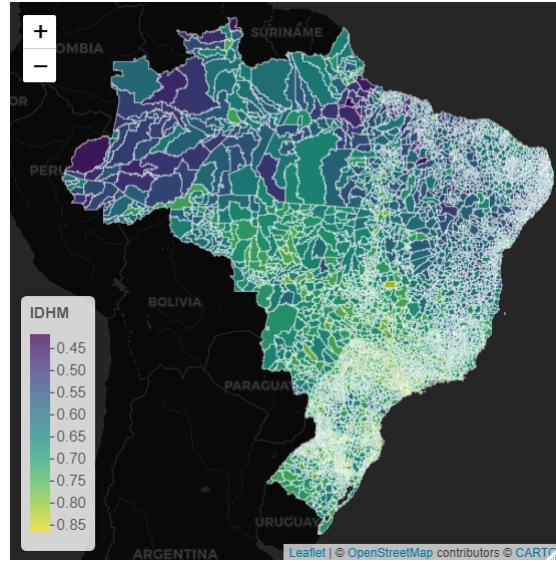


Figura 30 – Mapa dos municípios para IDHM.

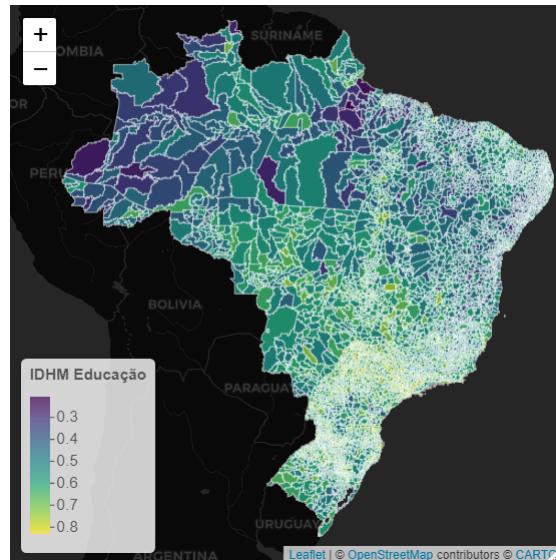


Figura 31 – Mapa dos municípios para IDHM Educação.

de água e esgoto encontramos coeficientes de correlação de Spearman positivos. Assim, quanto mais alto os valores dessas variáveis, mais alto será também a incompletude para prematuridade.

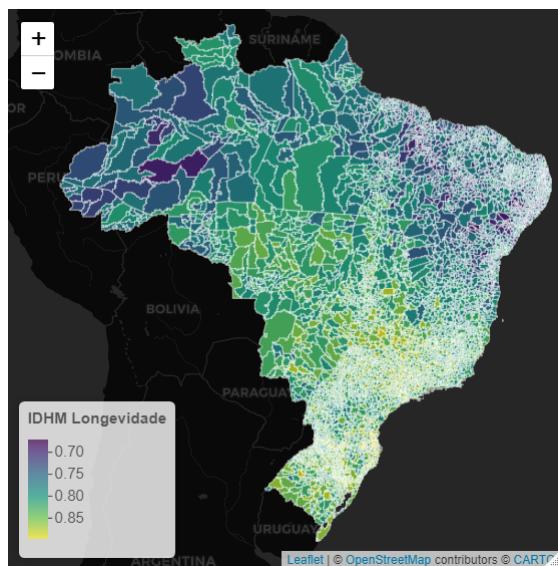


Figura 32 – Mapa dos municípios para IDHM Longevidade.

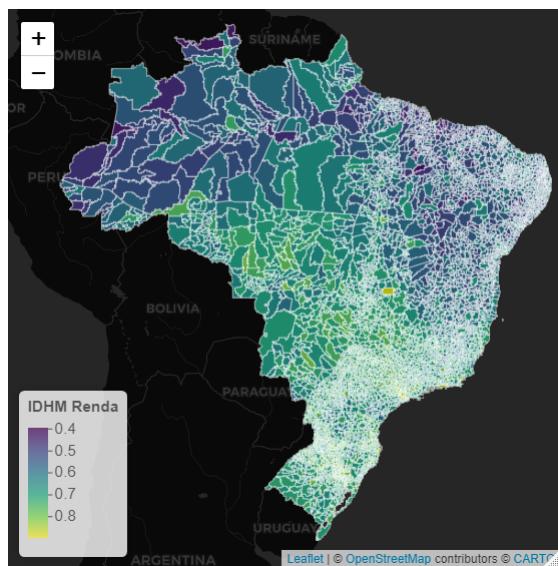


Figura 33 – Mapa dos municípios para IDHM Renda.

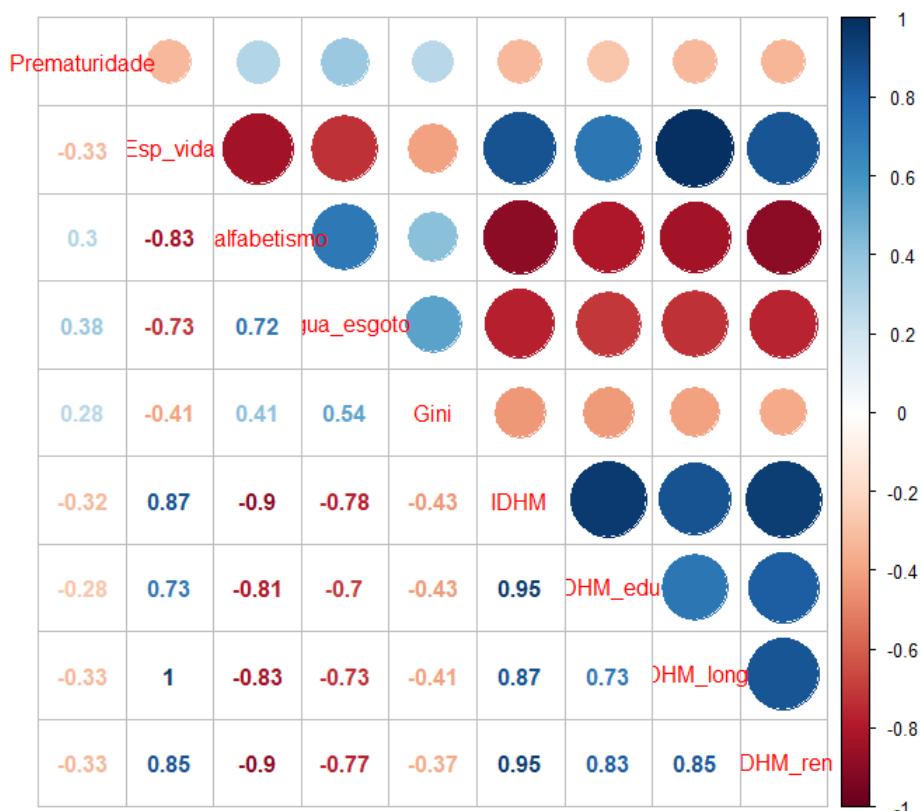


Figura 34 – Correlograma dos indicadores socioeconômicos (Censo 2010) e incompletude para prematuridade (ano 2019) - nível municipal. “Prematuridade” indica a incompletude para prematuridade.

5 Modelagem via Regressão Beta

Este capítulo será dividido em duas partes: a Seção 5.1 conterá os ajustes do modelo de regressão em nível estadual e a Seção 5.2 é dedicada aos ajustes do modelo de regressão em nível municipal.

5.1 Modelagem em nível estadual

Nesta seção, ajustamos modelos de regressão beta com dispersão fixa e variável, apresentados nas Subseções 2.4.2 e 2.4.3, para modelar a incompletude para prematuridade (ano de 2019) considerando os indicadores socioeconômicos do ano de 2017 (para esperança de vida o ano é 2018) como variáveis explicativas.

Como temos apenas $n = 27$ observações para o ajuste, faremos uma pré-seleção das variáveis socioeconômicas com o auxílio do correlograma apresentado na Figura 19. Quando a correlação entre duas variáveis socioeconômicas for maior que 0,80, será selecionada aquela com a maior correlação com a incompletude para prematuridade. Além disso, também só serão pré-selecionados os indicadores com correlação maior que 0,4 com a incompletude para prematuridade. Assim, são selecionadas para o ajuste as variáveis IDHM, IDHM Renda, IDHM Longevidade, Taxa de Água, Taxa de Esgoto e Esperança de vida.

Como IDHM é uma função de IDHM Renda e Longevidade, vamos seguir dois caminhos de ajuste. No primeiro caminho, vamos modelar a incompletude para prematuridade como uma função do IDHM, Taxa de Água, Taxa de Esgoto e Esperança de vida. Para o segundo caminho, o interesse é modelar a incompletude para prematuridade como uma função do IDHM Renda, IDHM Longevidade, Taxa de Água, Taxa de Esgoto e Esperança de vida.

Para cada caminho, foram consideradas inicialmente cinco funções de ligação: logit, probito, complementar log-log, cauchy e log-log. No entanto, nenhum modelo convergiu quando utilizamos a função de ligação Cauchy. Como coloca Fernandes Guerra e Fígoli (2013)[p. 11]: “se o modelo não convergir, os coeficientes não são confiáveis. Um dos principais fatores que explicam a não conversão do modelo é a insuficiência de casos em relação ao número de variáveis independentes incluídas no modelo”. Por esse motivo, a função de ligação Cauchy não foi considerada.

Em cada combinação de caminho e função de ligação, o modelo saturado (completo) é ajustado para verificar quais variáveis são importantes para a explicação da variabilidade da incompletude para prematuridade. É considerado um nível de significância de 5% e,

assim, as variáveis serão retiradas sequencialmente se o p-valor associado for maior que 5%, resultando no modelo final (só com as variáveis com p-valor < 0,05).

Para cada cenário, foram ajustados modelos de regressão beta com dispersão fixa e com dispersão variável, realizando o teste da razão de verossimilhanças para heteroscedasticidade (Subseção 2.4.4) para avaliar se o modelo com dispersão variável é mais indicado. O modelo com dispersão variável não foi considerado o melhor modelo para nenhuma das funções de ligação utilizadas.

Fixado o caminho e com o processo de modelagem descrito acima, cada uma das quatro funções de ligação apresentou o seu modelo final e o “modelo vencedor” (modelo escolhido dentre os melhores modelos de cada função de ligação) é aquele com o maior AIC em módulo. Todos os ajustes realizados podem ser vistos com detalhes em <https://rpubs.com/gabi-demarque/812724>. No que segue, apresentamos apenas os resultados do modelo vencedor para o primeiro caminho e para o segundo caminho de modelagem.

Para o primeiro caminho, o melhor ajuste foi obtido quando considerada a função de ligação logito, em que o ajuste completo é mostrado na Tabela 3 (*fit1*). Podemos observar que, além da precisão, apenas as covariáveis IDHM e Água foram significantes para o modelo, ao nível de significância $\alpha = 0,05$. Após retirar as variáveis Esgoto e Esperança de vida do modelo, na Tabela 4 está o ajuste do modelo final (*fit2*) para o primeiro caminho.

Coeficiente	Estimativa	Erro padrão	z valor	$\text{Pr}(> z)$
Intercepto	4,762	6,361	0,749	0,454
IDHM	-15,691	6,307	-2,488	0,013
Água	-0,024	0,010	-2,505	0,012
Esgoto	0,007	0,010	0,654	0,513
Esperança de vida	-0,058	0,113	0,509	0,611
Precisão	159,500	46,650	3,419	< 0,001

Tabela 3 – Estimativas dos parâmetros do modelo completo para o primeiro caminho - ajuste para estados (*fit1*).

Coeficiente	Estimativa	Erro padrão	z valor	$\text{Pr}(> z)$
Intercepto	5,758	2,364	2,436	0,0149
IDHM	-11,346	3,462	-3,277	0,001
Água	-0,020	0,007	-2,857	0,004
Precisão	155,570	45,590	3,412	< 0,001

Tabela 4 – Estimativas dos parâmetros do modelo reduzido para o segundo caminho - ajuste para estados (*fit2*).

De forma semelhante, prosseguimos para o segundo caminho. O melhor ajuste foi obtido pela função de ligação complementar log-log e o ajuste completo é mostrado na

Tabela 5 (*fit11*). Após a retirada do modelo das variáveis Esgoto, Esperança de vida e IDHM Longevidade, na Tabela 6 está o ajuste do modelo final (*fit12*).

Coeficiente	Estimativa	Erro padrão	<i>z</i> valor	Pr(> <i>z</i>)
Intercepto	0,811	6,961	0,116	0,907
IDHM Renda	-12,290	5,293	-2,322	0,020
IDHM Longevidade	8,361	6,252	1,337	0,1811
Água	-0,018	0,010	-1,843	0,065
Esgoto	0,000	0,010	0,030	0,976
Esperança de vida	-0,0249	0,125	-0,199	0,842
Precisão	156,800	45,900	3,416	< 0,001

Tabela 5 – Estimativas dos parâmetros do modelo completo para o primeiro caminho - ajuste para estados (*fit11*).

Coeficiente	Estimativa	Erro padrão	<i>z</i> valor	Pr(> <i>z</i>)
Intercepto	2,941	1,734	1,697	0,010
IDHM Renda	-8,078	2,797	-2,888	0,004
Água	-0,019	0,008	-2,515	0,012
Precisão	145,460	42,710	3,406	< 0,001

Tabela 6 – Estimativas dos parâmetros do modelo reduzido para o segundo caminho - ajuste para estados (*fit12*).

Na Tabela 7 estão os valores de AIC dos modelos obtidos pelos dois caminhos de modelagem. Pelo critério de Akaike, o melhor modelo para explicar a incompletude para prematuridade é o modelo *fit2* e este é o modelo escolhido. No que segue, apresentamos a análise diagnóstico do ajuste deste modelo.

Ajuste	Função de ligação	AIC
<i>fit1</i>	logito	-183,54
<i>fit2</i>	logito	-186,94
<i>fit11</i>	complementar log-log	-181,39
<i>fit12</i>	complementar log-log	-185,57

Tabela 7 – Resultado dos ajustes para estados.

5.1.1 Análise de diagnóstico

Para avaliar a qualidade do ajuste do modelo (ajuste *fit2*), vamos fazer a análise de diagnóstico, apresentada na Figura 35. Pelo gráfico dos resíduos versus o índice das observações, podemos perceber que há duas observações com resíduo menor que -2, são elas: observações #3 (resíduo de -2,267) e #22 (resíduo de -2,796). A observação #3 possui incompletude para a prematuridade igual a 0,0032, IDHM = 0,733 e Água = 73,97%.

Já a observação #22 possui o menor valor para incompletude para a prematuridade, que corresponde a 0,0003, IDHM = 0,752 e Água = 99,69% (maior valor da variável).

Além disso, pelo gráfico da distância de Cook, percebemos que a observação #4 representa a maior distância de Cook ($C_4 = 0,594$). A observação #4 possui incompletude para a prematuridade igual a 0,017, IDHM de 0,74 e Água de 40,44% (menor valor da variável). Já o gráfico de alavancagem generalizada versus valores preditos (*Generalized leverage vs predicted values*), vemos que há duas observações distantes das demais, são elas: observação #4 ($GL_4 = 0,655$) e observação #22 ($GL_{22} = 0,387$).

Pelo gráfico de envelope, podemos ver que o modelo *fit2* obteve um ótimo ajuste, visto que todas as observações encontram-se dentro das bandas de confiança.

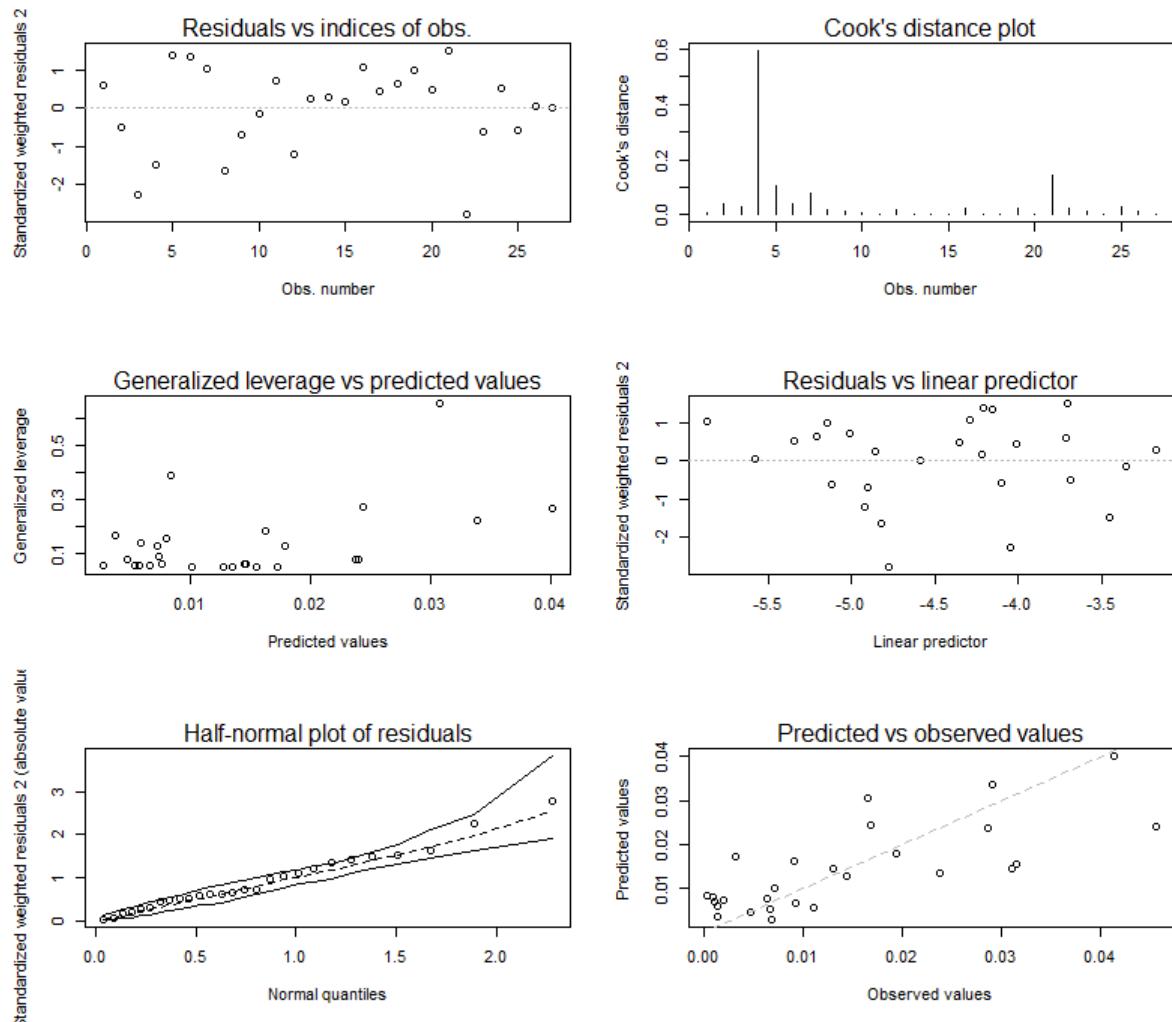


Figura 35 – Gráficos de diagnóstico do modelo final - *fit2*.

Para avaliar o impacto que as observações anteriormente destacadas causam nos resultados inferenciais, vamos realizar dois novos ajustes. No primeiro retiramos as observações #3 e #22 e no segundo, o modelo *fit2* é reajustado sem a observação #4. Os resultados podem ser vistos na Tabela 8. Podemos observar que, apesar de alterar as

estimativas pontuais dos parâmetros e o valor p, a retirada das observações identificadas como potenciais influentes não altera a inferência.

Ajuste	Parâmetros	β_0	β_1	β_2	ϕ
Modelo final	Estimativa p-valor	5,758 0,015	-11,346 0,001	-0,0202 0,004	155,572 $< 0,001$
Sem os casos 3 e 22	Estimativa p-valor	5,902 0,005	-11,613 0,000	-0,019 0,003	207,895 $< 0,001$
Sem o caso 4	Estimativa p-valor	4,584 0,057	-8,619 0,021	-0,030 0,001	160,526 $< 0,001$

Tabela 8 – Resultados inferenciais sem as potenciais observações influentes.

5.1.1.1 Interpretação do modelo

Uma vez que o modelo escolhido é aquele com função de ligação logito e ao considerar as estimativas apresentadas na Tabela 4, temos que

$$g(\hat{\mu}_t) = \log \left(\frac{\hat{\mu}_t}{1 - \hat{\mu}_t} \right) = 5,76 - 11,35 \times x_{1t} - 0,02 \times x_{2t}$$

$$\hat{\mu}_t = \frac{\exp(5,76 - 11,35 \times x_{1t} - 0,02 \times x_{2t})}{1 + \exp(5,76 - 11,35 \times x_{1t} - 0,02 \times x_{2t})}, \quad (5.1)$$

em que x_1 representa o IDHM e x_2 representa a Taxa de Água.

O valor esperado estimado da incompletude para prematuridade é uma função do IDHM e da Água, ou seja, para cada valor diferente desses indicadores, o valor esperado estimado da incompletude de prematuridade muda. Para exemplificar essa relação, na Tabela 9 apresentamos o valor esperado da incompletude (μ) estimado para cada combinação de algumas medidas-resumo de IDHM e Água.

A Figura 36a nos mostra o gráfico de IDHM versus $\hat{\mu}$ quando fixada variável Água em seu valor médio (85,71). Podemos perceber que a relação entre IDHM e $\hat{\mu}$ é inversa, ou seja, a medida que o valor do IDHM aumenta, o valor de $\hat{\mu}$ diminui, fixado o valor de Água. Um estado com IDHM de 0,7, por exemplo, apresenta um aumento de 75% no valor esperado estimado de incompletude que um estado com IDHM de 0,75, quando fixamos a variável Taxa de Água.

Da mesma forma, a Figura 36b apresenta o gráfico da Água versus $\hat{\mu}$ quando fixamos a variável IDHM em seu valor médio (0,75) e também percebemos uma relação inversa dessa variável com a variável resposta. Ainda, um estado com Taxa de Água de 70%, por exemplo, apresenta um aumento de 22% no valor esperado estimado de incompletude que um estado com Taxa de Água de 80%, fixado o IDHM.

Por fim, a Figura 37 nos mostra a relação conjunta entre as duas variáveis explicativas e $\hat{\mu}$. Olhando para as cores da escala do gráfico, podemos observar alguns pontos:

Medida (IDHM)	Medida (Água)	IDHM	Água	$\hat{\mu}$
Mínimo	Mínimo	0,68	40,44	0,059
Mínimo	1º quartil	0,68	81,27	0,027
Mínimo	3º quartil	0,68	96,41	0,020
Mínimo	Máximo	0,68	99,69	0,019
Mínimo	Mediana	0,68	91,82	0,022
Mínimo	Média	0,68	85,71	0,025
1º quartil	Mínimo	0,72	40,44	0,038
1º quartil	1º quartil	0,72	81,27	0,017
1º quartil	3º quartil	0,72	96,41	0,013
1º quartil	Máximo	0,72	99,69	0,012
1º quartil	Mediana	0,72	91,82	0,014
1º quartil	Média	0,72	85,71	0,016
3º quartil	Mínimo	0,78	40,44	0,020
3º quartil	1º quartil	0,78	81,27	0,009
3º quartil	3º quartil	0,78	96,41	0,007
3º quartil	Máximo	0,78	99,69	0,006
3º quartil	Mediana	0,78	91,82	0,007
3º quartil	Média	0,78	85,71	0,008
Máximo	Mínimo	0,85	40,44	0,009
Máximo	1º quartil	0,85	81,27	0,004
Máximo	3º quartil	0,85	96,41	0,003
Máximo	Máximo	0,85	99,69	0,003
Máximo	Mediana	0,85	91,82	0,003
Máximo	Média	0,85	85,71	0,004
Mediana	Mínimo	0,74	40,44	0,031
Mediana	1º quartil	0,74	81,27	0,014
Mediana	3º quartil	0,74	96,41	0,010
Mediana	Máximo	0,74	99,69	0,010
Mediana	Mediana	0,74	91,82	0,011
Mediana	Média	0,74	85,71	0,013
Média	Mínimo	0,75	40,44	0,028
Média	1º quartil	0,75	81,27	0,012
Média	3º quartil	0,75	96,41	0,009
Média	Máximo	0,75	99,69	0,009
Média	Mediana	0,75	91,82	0,010
Média	Média	0,75	85,71	0,011

Tabela 9 – Valor de $\hat{\mu}$ obtido pela Equação (5.1) a depender do valor do IDHM e Água.

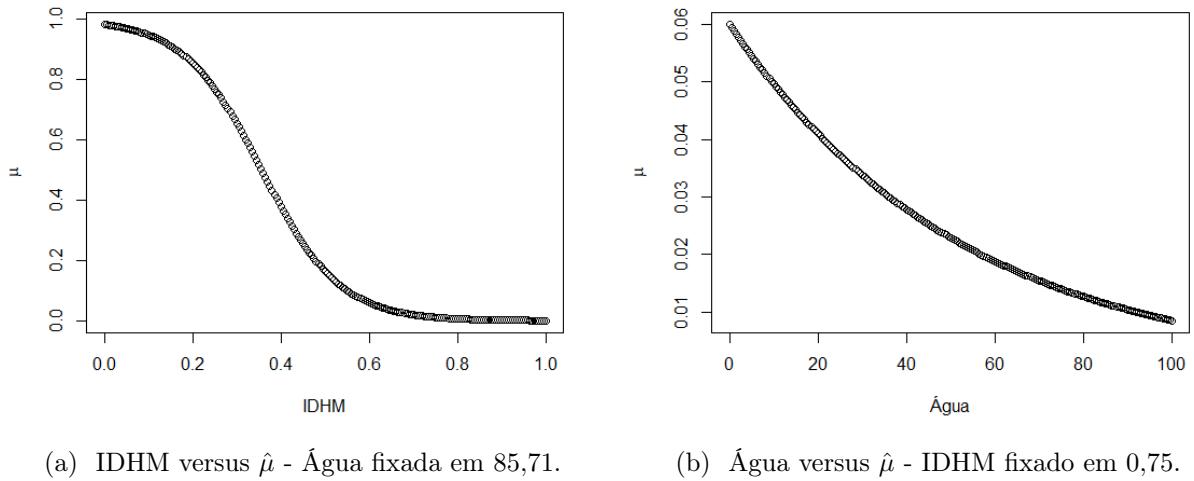


Figura 36 – Gráficos de dependência parcial para estados.

1) para valores de IDHM menores que, aproximadamente 0,18, a incompletude esperada é quase 1, independentemente do valor da Água; 2) para valores de IDHM maiores que 0,7, aproximadamente, a incompletude esperada é muito próxima de zero, principalmente para valores da Água maiores que 40%; e, 3) para valores de IDHM entre 0,18 e 0,7, aproximadamente, os valores da incompletude esperada variam entre 0,4 e 0,7, a depender do valor da Água, sendo este o intervalo de maior interação entre as duas variáveis.

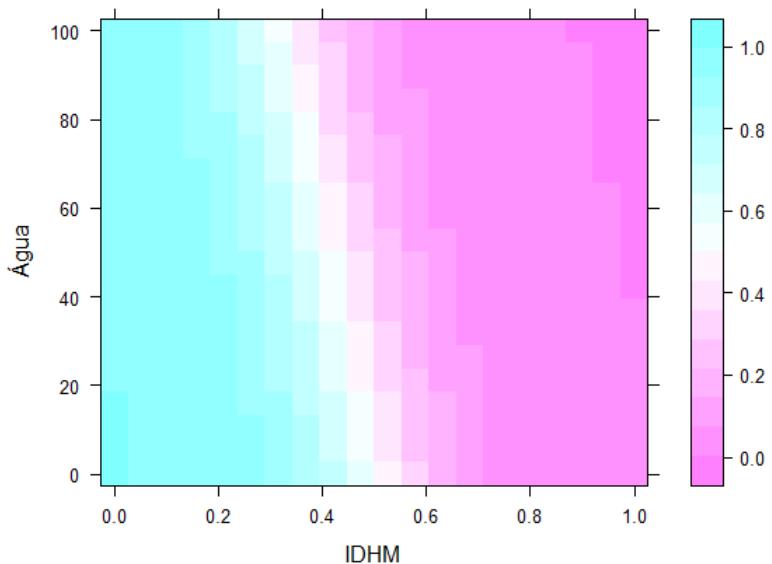


Figura 37 – Gráfico de dependência parcial com as duas variáveis explicativas (IDHM) e (Água) para estados.

5.2 Modelagem em nível municipal

Nesta seção, ajustamos modelo de regressão beta, apresentado na Seção 2.4, para modelar a incompletude para prematuridade (ano de 2019) considerando os indicadores socioeconômicos do Censo Demográfico de 2010 como variáveis explicativas.

Para evitar problemas de multicolinearidade, quando a correlação entre duas variáveis socieconómicas for maior que 0,80, será selecionada aquela que possuir a maior correlação com a incompletude para prematuridade. Faremos uma pré-seleção das variáveis socieconómicas com o auxílio do correlograma apresentado na Figura 34. Assim, são selecionadas para o ajuste as variáveis IDHM, IDHM Educação, IDHM Longevidade, IDHM Renda, Gini, Taxa de Água e Esgoto.

Nesta análise, selecionamos os municípios com número de nascimentos maior ou igual a 100, pelo fato de que muitos municípios têm valores pequenos para o total de nascimentos em um ano (Baroni *et al.*, 2021). Sendo assim, vamos trabalhar com $n = 3407$ observações.

Seguindo o mesmo processo de modelagem para os estados, começamos a modelagem para os municípios utilizando o modelo de regressão beta com dispersão fixa e com dispersão variável. No entanto, 35,34% (1204 observações) dos municípios possuem incompletude para prematuridade igual a zero. Para contornar esse problema e obter uma variável que assume valores no intervalo aberto (0, 1), consideramos a seguinte transformação na variável resposta (Espinheira *et al.*, 2008):

$$y^* = \frac{y(n - 1) + 0,5}{n},$$

em que y é a incompletude para prematuridade e n é o tamanho da amostra.

Entretanto, mesmo considerando a transformação para a variável de interesse, não conseguimos obter bons ajustes. O gráfico de envelope para o ajuste dos municípios citado anteriormente pode ser visto na Figura 38. Claramente, podemos observar que a maioria das observações encontram-se fora dos limites do envelope. Sendo assim, podemos dizer que há indícios de afastamento da suposição do modelo de regressão beta para a variável resposta. Os ajustes de todos os modelos podem ser vistos em <https://rpubs.com/gabidemarque/812741>.

No que segue, adotamos a modelagem que adequadamente considera a presença de zeros: modelo de regressão beta inflacionado em zero, descrito na Subseção 2.4.6.2. Por este motivo, utilizamos a função *gamlss*, do pacote *gamlss* do R para o ajuste do modelo RBIZ. Também neste pacote estão implementadas as técnicas de diagnóstico que consideramos nesse trabalho. Neste ajuste, não temos interesse em fazer inferência sobre o parâmetro σ e ele só será utilizado para “controlar” a dispersão.

Como IDHM é uma função do IDHM Educação, IDHM Longevidade e IDHM

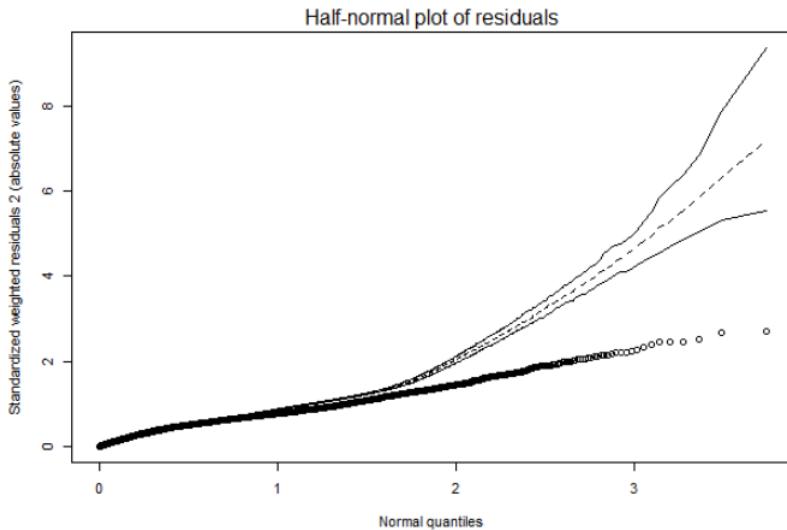


Figura 38 – Envelope do modelo final considerando a transformação para variável de interesse para dados municipais.

Renda, vamos seguir quatro caminhos de ajuste. No primeiro caminho, vamos modelar a incompletude para prematuridade como uma função do IDHM, Gini e Taxa de Água e Esgoto. Para o segundo caminho, o interesse é modelar a incompletude para prematuridade como uma função do IDHM Educação, IDHM Longevidade, IDHM Renda, Gini e Taxa de Água e Esgoto. Para o terceiro caminho, consideramos como variáveis explicativas IDHM, Gini, Taxa de Água e Esgoto e Nascimentos. Por fim, para o quarto caminho, a incompletude para prematuridade é função do IDHM Educação, IDHM Longevidade, IDHM Renda, Gini e Taxa de Água e Esgoto e Nascimentos. As variáveis definidas nos quatro caminho são consideradas para μ , σ e ν .

Em cada combinação de caminho, o modelo saturado (completo) é ajustado para verificar quais variáveis são importantes para a explicação da variabilidade da incompletude para prematuridade. É considerado um nível de significância de 5% e, assim, as variáveis serão retiradas sequencialmente se o p-valor associado for maior que 5%, resultando no modelo final (só com as variáveis com p-valor $< 0,05$).

No que segue, apresentamos apenas os resultados do modelo vencedor para o primeiro, segundo, terceiro e quarto caminhos.

Para o primeiro caminho, o melhor ajuste foi obtido retirando algumas variáveis o modelo completo, são elas: 1) em relação ao parâmetro μ , Gini e Água/esgoto não foram significantes para o modelo, ao nível de significância $\alpha = 0,05$; e, 2) para os parâmetros ν e σ , todas as variáveis foram significantes para o modelo. Após retirar as variáveis Gini e Água/esgoto do modelo para μ , na Tabela 10 está o ajuste do modelo final (*fit1*) para o primeiro caminho.

De forma semelhante, prosseguimos para o segundo caminho, em que o melhor

Coeficiente	Estimativa	Erro padrão	t valor	Pr(> t)
μ				
Intercepto	0,213	0,230	0,924	0,356
IDHM	-5,943	0,360	-16,495	< 0,001
σ				
Intercepto	-0,139	0,288	-0,483	0,684
IDHM	-2,916	0,353	-8,262	< 0,001
Gini	1,169	0,287	4,078	< 0,001
Água/esgoto	-0,004	0,001	-2,739	0,005
ν				
Intercepto	-0,335	0,639	-0,524	0,629
IDHM	2,411	0,708	3,406	< 0,001
Gini	-3,168	0,691	-4,582	< 0,001
Água/esgoto	-0,027	0,005	-5,749	0,006

Tabela 10 – Estimativas dos parâmetros do modelo reduzido para o primeiro caminho - ajuste para municípios (*fit1*).

ajuste foi obtido retirando algumas variáveis o modelo completo, são elas: 1) em relação ao parâmetro μ , as variáveis IDHM Longevidade, Gini e Água/esgoto não foram significantes para o modelo, ao nível de significância $\alpha = 0,05$; 2) para o parâmetro σ , apenas a variável Água/esgoto não foi significante; e, 3) para o parâmetro ν , todas as variáveis foram significantes para o modelo. Após retirar as variáveis não significantes, na Tabela 11 está o ajuste do modelo final para o segundo caminho (*fit3*).

Coeficiente	Estimativa	Erro padrão	t valor	Pr(> t)
μ				
Intercepto	-0,496	0,201	-2,470	0,0136
IDHM Educação	-2,096	0,448	-4,673	< 0,001
IDHM Renda	-3,167	0,526	-6,026	< 0,001
σ				
Intercepto	-1,692	0,429	-3,945	< 0,001
IDHM Educação	-1,834	0,377	-4,864	< 0,001
IDHM Longevidade	1,649	0,629	2,624	0,009
IDHM Renda	-0,928	0,522	-1,778	0,075
Gini	0,958	0,285	3,358	< 0,001
ν				
Intercepto	-3,299	1,069	-3,086	0,002
IDHM Educação	-2,865	0,755	-3,795	< 0,001
IDHM Longevidade	4,836	1,621	2,982	0,003
IDHM Renda	3,717	1,161	3,201	0,001
Gini	-3,494	0,702	-4,978	< 0,001
Água/esgoto	-0,021	0,005	-4,497	< 0,001

Tabela 11 – Estimativas dos parâmetros do modelo reduzido para o segundo caminho - ajuste para municípios (*fit3*).

Além das variáveis socieconômicas, consideramos também a variável de número de nascimentos (Nascimentos) para todos os parâmetros, seguindo o que faz Diniz e Melo (2019). Para o terceiro caminho, o melhor ajuste foi obtido retirando algumas variáveis do modelo completo, são elas: 1) para o parâmetro μ , as variáveis Gini e Água/esgoto não foram significantes para o modelo; 2) para o parâmetro σ , apenas a variável Nascimentos foi significante; e, 3) para o parâmetro ν , apenas a variável Gini não foi significante. Na Tabela 12 está o ajuste do modelo final para o terceiro caminho (*fit5*).

Coeficiente	Estimativa	Erro padrão	t valor	Pr($> t $)
μ				
Intercepto	-1,494	0,155	-9,622	< 0,001
IDHM	-3,154	0,247	-12,772	< 0,001
Nascimentos	-0,000	0,000	-4,508	< 0,001
σ				
Intercepto	-1,400	0,022	-64,946	< 0,001
Nascimentos	-0,000	0,000	-3,887	< 0,001
ν				
Intercepto	-4,603	0,542	-8,492	< 0,001
IDHM	6,921	0,783	8,837	< 0,001
Água/esgoto	-0,021	0,005	-4,491	< 0,001
Nascimentos	-0,001	0,000	-9,918	< 0,001

Tabela 12 – Estimativas dos parâmetros do modelo reduzido para o terceiro caminho - ajuste para municípios (*fit5*).

Por fim, para o quarto caminho, o melhor ajuste foi obtido retirando algumas variáveis do modelo completo, são elas: 1) para o parâmetro μ , as variáveis IDHM Longevidade, Gini e Água/Esgoto não foram significantes para o modelo; 2) para o parâmetro σ , as variáveis Água/esgoto e Nascimentos não foram significantes para o modelo; e, 3) por fim, para o parâmetro ν , apenas as variáveis IDHM Educação e Gini não foram significantes. Aqui, para modelar o parâmetro σ fizemos de duas formas: (i) atribuímos apenas a variável Nascimentos; e, (ii) acrescentamos a variável Nascimentos, junto com todas as outras covariáveis. Entretanto, ao comparar o AIC dos ajustes para σ , observamos que o ajuste (ii) foi o mais adequado. Por este motivo, utilizamos o ajuste em (ii) e na Tabela 13 está o ajuste do modelo final para o quarto caminho (*fit7*).

Na Tabela 14 estão os valores de AIC obtidos de cada um dos quatro modelos apresentados. Pelo critério de Akaike, o melhor modelo para explicar a incompleitude para prematuridade é o modelo *fit7* e este é o modelo escolhido. No que segue, apresentamos a análise diagnóstico do ajuste deste modelo.

Coeficiente	Estimativa	Erro padrão	t valor	Pr(> t)
μ				
Intercepto	-0,569	0,202	-2,814	0,005
IDHM Educação	-2,049	0,449	-4,568	< 0,001
IDHM Renda	-3,065	0,525	-5,833	< 0,001
Nascimentos	-0,000	0,000	-2,151	0,032
σ				
Intercepto	-1,615	0,424	-3,811	< 0,001
IDHM Educação	-1,864	0,376	-4,954	< 0,001
IDHM Longevidade	1,733	0,621	2,791	0,005
IDHM Renda	-1,031	0,518	-1,990	0,047
Gini	0,832	0,288	2,891	0,004
ν				
Intercepto	-6,438	0,981	-6,563	< 0,001
IDHM Longevidade	3,723	1,643	2,265	0,024
IDHM Renda	5,234	1,020	5,132	< 0,001
Água/esgoto	-0,016	0,005	-3,433	< 0,001
Nascimentos	-0,001	0,000	-9,941	< 0,001

Tabela 13 – Estimativas dos parâmetros do modelo reduzido para o quarto caminho - ajuste para municípios (*fit7*).

Ajuste	AIC
<i>fit1</i>	-7632,122
<i>fit3</i>	-7677,244
<i>fit5</i>	-7727,272
<i>fit7</i>	-7849,130

Tabela 14 – Resultado dos ajustes para municípios.

5.2.1 Análise de diagnóstico

Para avaliar a qualidade do ajuste para o modelo final (*fit7*), vamos fazer a análise de diagnóstico por meio dos gráficos do resíduo quantil aleatorizado e do envelope (Subseção 2.4.6.4). Com eles, conseguimos verificar se a escolha do modelo é adequada, detectar observações influentes e possíveis erros na escolha de funções de ligação. Pela Figura 39, temos a análise dos resíduos quantis aleatorizados.

Para o gráfico do Resíduo Quantil Aleatorizado vs Índice das observações, colocamos um limite para detectar a presença de observações influentes, sendo elas maiores que 2 ou menores que -2. Com isso, analisando a Figura 39a, podemos observar que há algumas observações que ultrapassam o limite estipulado para os resíduos. Por este motivo, retiramos essas observações e fizemos um novo ajuste do modelo final. Entretanto, a retirada desses resíduos elevados não mudou a inferência para μ e ν , então vamos continuar considerando o modelo final como sendo o *fit7*. Já pela análise da Figura 39b, não detectamos nenhuma tendência nos dados que indique possível erro de escolha da função de ligação.

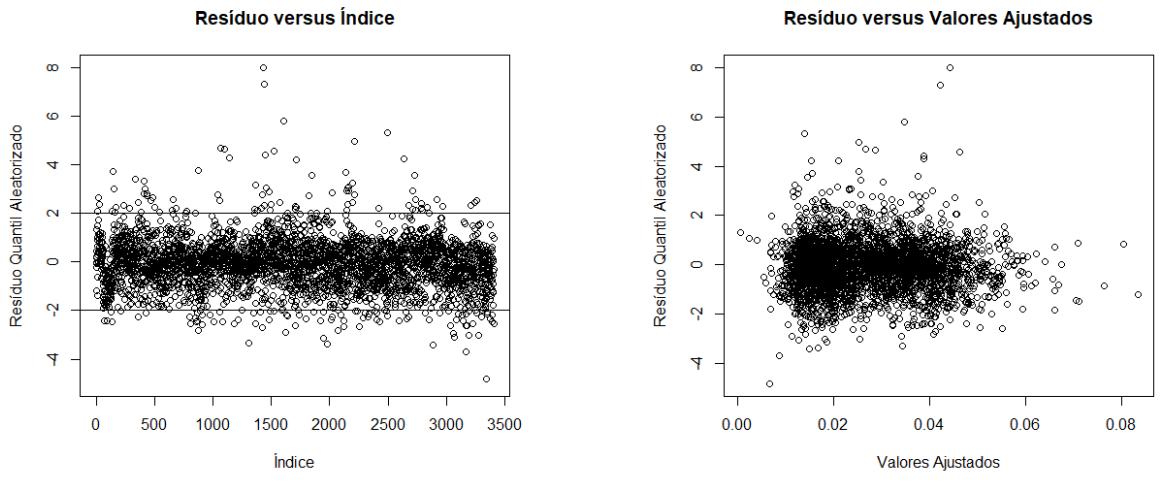


Figura 39 – Análise dos resíduos para municípios.

Por fim, o gráfico de envelope para o modelo final está na Figura 40. Podemos observar que a maioria das observações estão distribuídas de maneira uniforme dentro das bandas de confiança. Dessa forma, concluímos que não há indícios de afastamento da suposição do modelo de regressão beta inflacionado em zero para a variável de interesse.

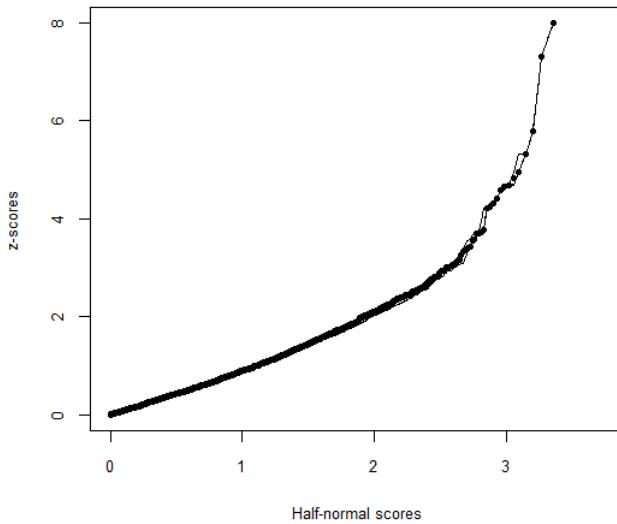


Figura 40 – Gráficos de envelope do modelo final para municípios - *fit7*.

5.2.1.1 Interpretação do modelo

Como descrito na Subseção 2.4.6.3, a probabilidade de ter incompletude igual a zero é $\alpha = P(Y = 0)$. Ainda, pela reparametrização utilizada, temos que $\nu = \alpha/(1 - \alpha)$

e $\log(\nu) = z^\top \gamma$ (Equação 2.15). Assim, temos que $\alpha = \exp(z^\top \gamma) / [1 + \exp(z^\top \gamma)]$. Desta maneira, ao considerar as estimativas dos parâmetros na Tabela 13, podemos escrever a probabilidade estimada de incompletude zero como:

$$\hat{\alpha}_t = \frac{\exp(-6,438 + 3,723 \times z_{1t} + 5,234 \times z_{2t} - 0,016 \times z_{3t} - 0,001 \times z_{4t})}{1 + \exp(-6,438 + 3,723 \times z_{1t} + 5,234 \times z_{2t} - 0,016 \times z_{3t} - 0,001 \times z_{4t})},$$

em que z_1 representa IDHM Longevidade, z_2 representa IDHM Renda, z_3 representa Água/esgoto e z_4 representa Nascimentos.

Podemos interpretar a razão de chances de ter incompletude zero de forma direta ao usar as estimativas dos parâmetros γ . Assim, a cada aumento de 0,1 no IDHM Longevidade, a chance de incompletude zero aumenta 45% ($\exp(3,723 \times 0,1) = 1,45$), fixados IDHM Renda, Água/esgoto e Nascimentos. A cada aumento de 0,1 no IDHM Renda, a chance de incompletude zero aumenta 69% ($\exp(5,234 \times 0,1) = 1,69$), fixados IDHM Longevidade, Água/esgoto e Nascimentos. A cada decréscimo de 10 unidades de Água/esgoto, a chance de incompletude zero aumenta em 17% ($\exp[-0,016 \times (-10)] = 1,17$), fixados IDHM Renda, Longevidade e Nascimentos. Por fim, temos que a cada decréscimo de 100 nascidos no município, a chance de incompletude zero aumenta 10% ($\exp[-0,001 \times (-100)] = 1,10$), fixados IDHM Renda, Longevidade e Água/esgoto.

No que segue, interpretamos a incompletude esperada condicional a estar no intervalo aberto $(0, 1)$. Uma vez que o modelo final (*fit7*) é aquele em que o parâmetro μ tem como função de ligação a logit e ao considerar as estimativas apresentadas na Tabela 13, temos que

$$g(\hat{\mu}_t) = \log\left(\frac{\hat{\mu}_t}{1 - \hat{\mu}_t}\right)$$

$$\hat{\mu}_t = \frac{g(\hat{\mu}_t)}{1 + \exp(g(\hat{\mu}_t))} = \frac{-0,568594 - 2,049015 \times x_{1t} - 3,064540 \times x_{2t} - 0,000019 \times x_{3t}}{\exp(-0,568594 - 2,049015 \times x_{1t} - 3,064540 \times x_{2t} - 0,000019 \times x_{3t}) + 1}, \quad (5.2)$$

em que x_1 representa o IDHM Educação, x_2 representa o IDHM Renda e x_3 representa o número de nascimentos, com $t = 1, \dots, n$.

O valor esperado estimado da incompletude para prematuridade é uma função do IDHM Educação, IDHM Renda e do número de nascimentos, ou seja, para cada valor diferente desses indicadores, o valor esperado estimado da incompletude de prematuridade muda. Para exemplificar essa relação, na Tabela 15 apresentamos o valor de μ estimado para cada combinação de algumas medidas-resumo de IDHM Educação, IDHM Renda e Nascimentos.

A Figura 41a nos mostra o gráfico do IDHM Educação versus $\hat{\mu}$ quando fixadas as variáveis IDHM Renda e Nascimentos em seus respectivos valores médios, 0,64 e 802.

M. IDHM-E	M. IDHM-R	M. Nasc.	IDHM Edu.	IDHM Ren.	Nasc.	$\hat{\mu}$
Mínimo	Mínimo	Mínimo	0,207	0,417	100	0,093
Mínimo	Máximo	Mínimo	0,207	0,891	100	0,007
Mínimo	Média	Mínimo	0,207	0,640	100	0,049
Máximo	Mínimo	Mínimo	0,811	0,417	100	0,029
Máximo	Máximo	Mínimo	0,811	0,891	100	0,007
Máximo	Média	Mínimo	0,811	0,640	100	0,015
Média	Mínimo	Mínimo	0,557	0,417	100	0,048
Média	Máximo	Mínimo	0,557	0,891	100	0,012
Média	Média	Mínimo	0,557	0,640	100	0,025
Mínimo	Mínimo	Máximo	0,207	0,417	158587	0,005
Mínimo	Máximo	Máximo	0,207	0,891	158587	0,001
Mínimo	Média	Máximo	0,207	0,640	158587	0,003
Máximo	Mínimo	Máximo	0,811	0,417	158587	0,001
Máximo	Máximo	Máximo	0,811	0,891	158587	0,000
Máximo	Média	Máximo	0,811	0,640	158587	0,001
Média	Mínimo	Máximo	0,557	0,417	158587	0,002
Média	Máximo	Máximo	0,557	0,891	158587	0,001
Média	Média	Máximo	0,557	0,640	158587	0,001
Mínimo	Mínimo	Média	0,207	0,417	802	0,092
Mínimo	Máximo	Média	0,207	0,891	802	0,023
Mínimo	Média	Média	0,207	0,640	802	0,049
Máximo	Mínimo	Média	0,811	0,417	802	0,029
Máximo	Máximo	Média	0,811	0,891	802	0,007
Máximo	Média	Média	0,811	0,640	802	0,015
Média	Mínimo	Média	0,557	0,417	802	0,047
Média	Máximo	Média	0,557	0,891	802	0,012
Média	Média	Média	0,557	0,640	802	0,025

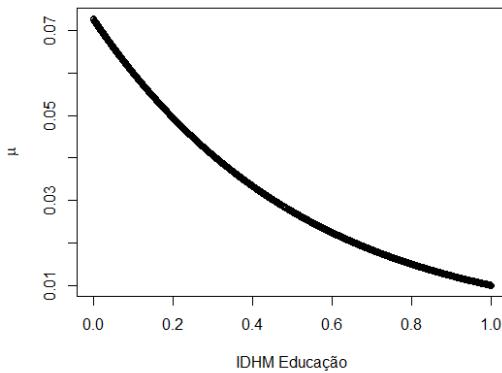
Tabela 15 – Valor de $\hat{\mu}$ obtido pela Equação (5.2) a depender do valor IDHM Educação, IDHM Renda e do número de nascimentos.

Podemos perceber que a relação entre IDHM Educação e $\hat{\mu}$ é inversa, ou seja, a medida que o valor do IDHM Educação aumenta, o valor de $\hat{\mu}$ diminui. Um município com IDHM Educação de 0,7, por exemplo, apresenta um aumento de, aproximadamente, 11% no valor esperado estimado de incompletude que um município com IDHM Educação de 0,75, quando IDHM Renda e Nascimentos estão fixadas.

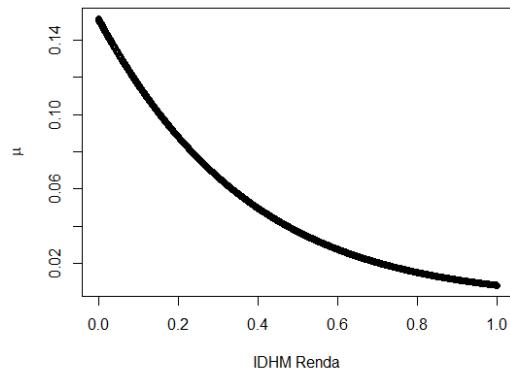
Da mesma forma, a Figura 41b apresenta o gráfico do IDHM Renda versus $\hat{\mu}$ quando fixadas as variáveis IDHM Educação e Nascimentos em seus respectivos valores médios, 0,56 e 802. Também percebemos uma relação inversa, com as demais variáveis fixas. Ainda, um município com IDHM Renda de 0,7, por exemplo, apresenta um aumento de, aproximadamente, 35% no valor esperado estimado de incompletude que um município com IDHM Renda de 0,8, quando IDHM Educação e Nascimentos estão fixadas.

Além disso, a Figura 41c nos mostra o gráfico dos Nascimentos versus $\hat{\mu}$ quando

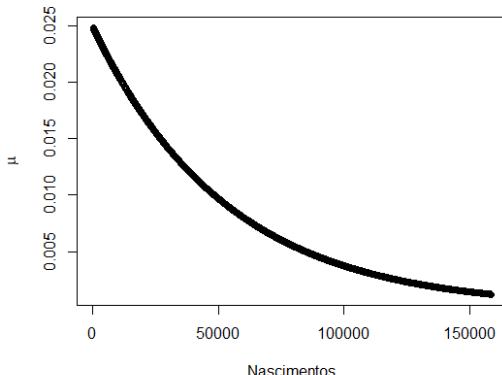
fixadas as variáveis IDHM Educação e IDHM Renda em seus respectivos valores médios, 0,56 e 0,64. Aqui também encontramos uma relação inversa com a incompletude esperada, fixadas as demais variáveis explicativas. Um município com número de nascimentos igual a 1000, por exemplo, apresenta um aumento de, aproximadamente, 8% no valor esperado estimado de incompletude que um município com número de nascimentos igual a 5000, quando IDHM Educação e IDHM Renda estão fixadas.



- (a) IDHM Educação vs $\hat{\mu}_t$ - IDHM Renda e Nascimentos fixados em 0,64 e 802, respectivamente.



- (b) IDHM Renda vs $\hat{\mu}_t$ - IDHM Educação e Nascimentos fixados em 0,56 e 802, respectivamente.



- (c) Nascimentos vs $\hat{\mu}_t$ - IDHM Educação e IDHM Renda fixados em 0,56 e 0,64, respectivamente.

Figura 41 – Gráficos de dependência parcial para municípios.

Por fim, a Figura 42 nos mostra a relação conjunta entre as duas variáveis de IDHM e $\hat{\mu}$, quando fixada a variável Nascimentos em seu valor médio (802). Olhando para as cores da escala do gráfico, podemos observar alguns pontos: 1) para valores do IDHM Educação e do IDHM Renda menores que, aproximadamente, 0,2, a incompletude esperada fica em torno de 0,25 e 0,4; 2) para valores do IDHM Educação e do IDHM Renda maiores que 0,6, aproximadamente, a incompletude esperada é próxima de zero; e 3) para as demais

combinações dos valores de IDHM Educação e Renda, a incompletude esperada varia entre 0,05 e 0,25, aproximadamente.

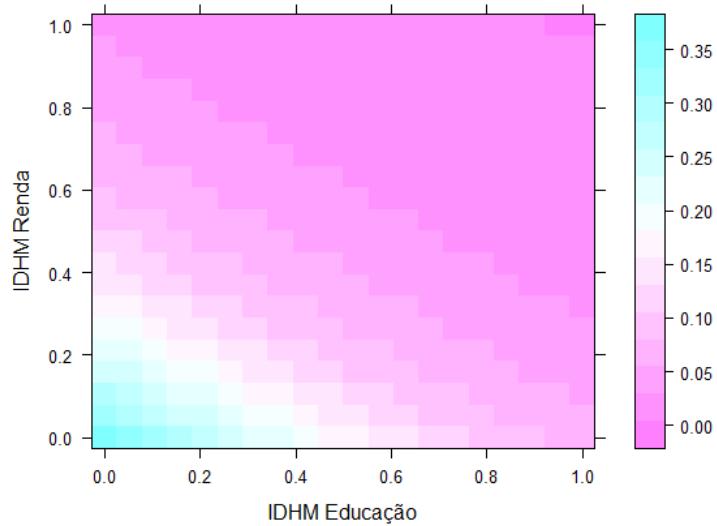


Figura 42 – Gráfico de dependência parcial com as duas variáveis explicativas (IDHM Educação) e (IDHM Renda) para municípios - Número de nascimentos fixada em 802.

6 Considerações Finais

O presente trabalho teve como objetivo analisar a qualidade dos dados de algumas variáveis do SINASC, através do indicador de qualidade incompletude, e estudar indicadores socioeconômicos com potenciais associações com a incompletude para prematuridade, variável de incompletude de maior interesse neste estudo, em níveis estadual e municipal.

Para o ajuste em nível estadual e ao considerar que a incompletude para prematuridade está limitada no intervalo $(0, 1)$, o modelo de regressão beta com dispersão fixa (Subseção 2.4.2) e dispersão variável (Subseção 2.4.3) foram considerados. Para os dados em nível municipal, as observações estavam no intervalo $[0, 1]$ e, por este motivo, o modelo de regressão beta inflacionado em zero (Subseção 2.4.6.2) foi adotado. O desenvolvimento computacional foi realizado com o auxílio dos pacotes *betareg* e *gamlss* do *software R*, nos quais estão implementados a estimação dos parâmetros e as análises de diagnóstico de todos os modelos considerados.

Nos Capítulos 3 e 4, foram analisadas as incompletudes das variáveis obstétricas ao longo do tempo (2016 a 2019) e avaliada a correlação entre elas para estados e municípios, além de mostrar a distribuição do Brasil com relação a essas variáveis. Por meio dessas análises, foi possível notar que, para o Brasil, a variável indicadora de anomalias congênitas apresenta os maiores índices de incompletude, seguida pela variável que informa sobre prematuridade. A variável tipo de parto apresenta os menores índices de incompletude, independentemente do ano em consideração. Além disso, como esperado, as correlações entre as incompletudes das variáveis do SINASC são positivas, com uma maior intensidade para os dados dos estados.

Ainda nesses capítulos, foram feitas análises com relação aos dados socieconômicos: IDHM (e as três vertentes: Longevidade, Escolaridade e Renda), Gini, Taxa de Analfabetismo, Esperança de Vida, Taxa de Água e de Esgoto. Nesta análise, os estados e municípios foram avaliados, identificando os que possuem os melhores e os piores indicadores. Na análise dos estados, foi possível realizar uma comparação entre os anos de 2016 e 2017 (e 2018 para Esperança de vida) e foi observado que os estados pouco mudaram de um ano para outro.

No Capítulo 5, os ajustes dos modelos para dados estaduais e municipais foram descritos. Em relação aos estados (Seção 5.1), vimos que o modelo de regressão beta com dispersão fixa se ajustou bem aos dados. Os modelos com dispersão variável não resultaram em bons ajustes. No modelo final do ajuste para estados, observamos que apenas as variáveis IDHM e Água foram significantes para o modelo. As duas variáveis apresentam relação negativa com a incompletude esperada e foi possível perceber que, independentemente

do valor da Água, para os menores valores do IDHM ($< 0,18$, aproximadamente), a incompletude esperada é bem próxima de 1. Em contrapartida, também independente do valor da Água, para os maiores valores do IDHM ($> 0,7$), a incompletude esperada é muito próxima de zero, principalmente para valores da Água maiores que 40%. Para valores de IDHM entre 0,18 e 0,7, aproximadamente, os valores da incompletude esperada variam entre 0,4 e 0,7, a depender do valor da Água.

Já em relação aos municípios (Seção 5.2), começamos a modelagem da mesma forma que para estados (modelos de regressão beta com dispersão fixa e variável). Observamos que há 35,34% (1204 observações) dos municípios com incompletude zero e, por isso, foi primeiramente realizada uma transformação para obter uma variável que assume valores no intervalo aberto (0, 1). No entanto, o ajuste não ficou bom, visto que a maioria dos pontos estavam fora das bandas de confiança do gráfico envelope. Por fim, adequadamente utilizamos a abordagem do modelo de regressão beta inflacionado em zero (Subseção 2.4.6.2) e o modelo final apresentou ótimo ajuste.

Para a incompletude para prematuridade em nível municipal, foram significativas as variáveis IDHM Longevidade, IDHM Renda, Água/esgoto e Nascimentos para a probabilidade de incompletude ser igual a zero, em que as variáveis de IDHM apresentam uma relação positiva com a probabilidade de incompletude zero e as demais variáveis significativas apresentam uma relação inversa. Já as variáveis IDHM Educação, IDHM Renda e Nascimentos foram significativas para a incompletude esperada condicional a estar no intervalo (0, 1), em que todas elas apresentam relação negativa com a incompletude esperada.

Com esse trabalho, esperamos ter contribuído para as discussões sobre a qualidade dos dados do SINASC e sua relação com as características socieconômicas de estados e municípios brasileiros.

Referências

- Akaike(1974) Hirotugu Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723. Citado na pág. Citado na página [43](#).
- Assunção(2012) Fernando Assunção. *Estratégias para tratamento de variáveis com dados faltantes durante o desenvolvimento de modelos preditivos*. Tese de Doutorado, Universidade de São Paulo. Citado na pág. Citado na página [24](#).
- Atkinson(1981) Anthony C Atkinson. Two graphical displays for outlying and influential observations in regression. *Biometrika*, 68(1):13–20. Citado na pág. Citado 3 vezes nas páginas [41](#), [43](#) e [48](#).
- Baroni *et al.*(2021) Lais Baroni, Rebecca Salles, Samella Salles, Marcel Pedroso, Jefferson Lima, Igor Morais, Lucas Carraro, Raphael de Freitas Saldanha, Carlos Sousa, Carlos Cardoso *et al.* Neonatal mortality rates in brazilian municipalities: from 1996 to 2017. *BMC Research Notes*, 14(1):1–3. Citado na pág. Citado na página [80](#).
- Bayer(2011) Fábio Mariano Bayer. *Modelagem e inferência em regressão beta*. Tese de Doutorado, tese de doutorado, Universidade Federal de Pernambuco, Recife. Citado na pág. Citado 2 vezes nas páginas [37](#) e [39](#).
- Bittar e Zugaib(2009) Roberto Eduardo Bittar e Marcelo Zugaib. Indicadores de risco para o parto prematuro. *Revista Brasileira de Ginecologia e Obstetrícia*, 31:203–209. Citado na pág. Citado na página [25](#).
- Braga e Mazzeu(2017) Ana Carolina Braga e Francisco José Carvalho Mazzeu. O analfabetismo no brasil: lições da história. *Revista on line de Política e Gestão Educacional*, páginas 24–46. Citado na pág. Citado na página [25](#).
- Brasil(2016) Trata Brasil. Manual do saneamento básico: Entendendo o saneamento básico ambiental no brasil e sua importância socioeconômica, 2012, 2016. Citado na pág. Citado na página [26](#).
- Carvalho(2017) Melissa Mello Carvalho. Dados faltantes em análises: uma revisão sobre métodos estatísticos flexíveis a incompletude. Em *II Simpósio de Métodos Numéricos em Engenharia*. Citado na pág. Citado na página [24](#).
- Cook(1977) R Dennis Cook. Detection of influential observation in linear regression. *Technometrics*, 19(1):15–18. Citado na pág. Citado na página [42](#).
- Cordeiro e Demétrio(2008) Gauss Moutinho Cordeiro e Clarice GB Demétrio. Modelos lineares generalizados e extensões. *Piracicaba: USP*. Citado na pág. Citado na página [37](#).

Cribari-Neto e Zeileis(2010) Francisco Cribari-Neto e Achim Zeileis. Beta regression in r. *Journal of statistical software*, 34(1):1–24. Citado na pág. Citado 2 vezes nas páginas 39 e 43.

Cunha *et al.*(2017) Elenice Machado da Cunha, José Muniz da Costa Vargens *et al.*

da Silva(2016) Luciana Bezerra da Silva. Sistemas de informações em saúde como ferramenta para gestão do sus. *Saúde e Desenvolvimento*, 8(5). Citado na pág. Citado na página 23.

de Oliveira(2004) Marcos Santos de Oliveira. *Um Modelo de Regressão Beta: teoria e aplicações*. Tese de Doutorado, Instituto de Matemática e Estatística da Universidade de São Paulo, 12/04/2004. Citado na pág. Citado na página 37.

Diniz e Melo(2019) Jean Sabino Magalhães Diniz e Daniel Leite Martins Melo. Modelo de regressão beta inflacionado em zero e um: uma aplicação à proporção de mulheres nas empresas. Citado na pág. Citado 7 vezes nas páginas 36, 37, 44, 46, 47, 48 e 83.

Esparza *et al.*(2020) Luz Judith Rodríguez Esparza, Dolly Anabel Ortiz Lazcano, Julio César Macías Ponce e Octavio Martín Maza Díaz Cortés. Bilateral gini index: Application for regional studies and international comparisons. *RBEST: Revista Brasileira de Economia Social e do Trabalho*, 2:e020010–e020010. Citado na pág. Citado na página 25.

Espinheira *et al.*(2008) Patrícia L Espinheira, Silvia LP Ferrari e Francisco Cribari-Neto. On beta regression residuals. *Journal of Applied Statistics*, 35(4):407–419. Citado na pág. Citado 2 vezes nas páginas 41 e 80.

Fernandes Guerra e Fígoli(2013) Francismara Fernandes Guerra e Moema Bueno Gonçalves Fígoli. Esperança de vida e sua relação com indicadores de longevidade: um estudo demográfico para o brasil, 1980-2050. *Revista Brasileira de Estudos de População*, 30: S85–S102. Citado na pág. Citado 2 vezes nas páginas 26 e 73.

Ferrari e Cribari-Neto(2004) Silvia Ferrari e Francisco Cribari-Neto. Beta regression for modelling rates and proportions. *Journal of applied statistics*, 31(7):799–815. Citado na pág. Citado 7 vezes nas páginas 36, 37, 38, 39, 41, 42 e 43.

Ferrari *et al.*(2011) Silvia LP Ferrari, Patricia L Espinheira e Francisco Cribari-Neto. Diagnostic tools in beta regression with varying dispersion. *Statistica Neerlandica*, 65 (3):337–351. Citado na pág. Citado na página 42.

Ferraz *et al.*(2009) Lygia Helena Valle da Costa Ferraz *et al.*O SUS, o DATASUS e a informação em saúde: uma proposta de gestão participativa. Tese de Doutorado. Citado na pág. Citado na página 23.

Franco *et al.*(2012) Joel Levi Ferreira Franco *et al.* Sistemas de informação. Citado na pág.
Citado na página [23](#).

Friendly(2002) Michael Friendly. Corrgrams: Exploratory displays for correlation matrices.
The American Statistician, 56(4):316–324. Citado na pág. Citado na página [35](#).

Gabriel *et al.*(2014) Guilherme Paiva Gabriel, Letícia Chiquetto, André Moreno Morcillo,
Maria do Carmo Ferreira, Ivan Gilberto M Bazan, Luísa Dias Daolio, Jéssica J Rocha
Lemos e Emília de Faria Carniel. Avaliação das informações das declarações de nascidos
vivos do sistema de informação sobre nascidos vivos (sinasc) em campinas, são paulo,
2009. *Revista Paulista de Pediatria*, 32(3):183–188. Citado na pág. Citado 4 vezes nas
páginas [23](#), [24](#), [30](#) e [65](#).

Galarza(2014) Christian E Galarza. Regressão beta e aplicações. Citado na pág. Citado na
página [37](#).

Guimarães *et al.*(2013) Eliete Albano de Azevedo Guimarães, Zulmira Maria de Araújo
Hartz, Antônio Ignácio de Loyola Filho, Antônio José de Meira e Zélia Maria Profeta da
Luz. Avaliação da implantação do sistema de informação sobre nascidos vivos em
municípios de minas gerais, brasil. *Cadernos de Saúde Pública*, 29:2105–2118. Citado na
pág. Citado na página [24](#).

Guimarães *et al.*(2014) Eliete Albano de Azevedo Guimarães, Rose Ferraz Carmo, Antônio
Ignácio de Loyola, Antônio José de Meira e Zélia Maria Profeta Luz. O contexto
organizacional do sistema de informações sobre nascidos vivos segundo profissionais de
saúde do nível municipal. *Revista Brasileira de Saúde Materno Infantil*, 14:165–172.
Citado na pág. Citado na página [29](#).

Guimarães(2008) Paulo Ricardo Bittencourt Guimarães. Métodos quantitativos estatísticos.
Citado na pág. Citado na página [35](#).

Howson *et al.*(2013) Christopher P Howson, Mary V Kinney, Lori McDougall e Joy E
Lawn. Born too soon: preterm birth matters. *Reproductive health*, 10(1):1–9. Citado na
pág. Citado na página [25](#).

Johnson e Bhattacharyya(2019) Richard A Johnson e Gouri K Bhattacharyya. *Statistics:
principles and methods*. John Wiley & Sons. Citado na pág. Citado na página [34](#).

Jorge *et al.*(2007) Maria Helena Prado deMello Jorge, Ruy Laurenti e Sabina Léa Davidson
Gotlieb. Quality analysis of brazilian vital statistics: the experience of implementing
the sim and sinasc systems. *Ciência & Saúde Coletiva*, 12(3):643. Citado na pág. Citado
na página [23](#).

Kieschnick e McCullough(2003) Robert Kieschnick e Bruce D McCullough. Regression analysis of variates observed on (0, 1): percentages, proportions and fractions. *Statistical modelling*, 3(3):193–213. Citado na pág. Citado na página 36.

Lajos *et al.*(2014) Giuliane Jesus Lajos *et al.* Estudo multicêntrico de investigação em prematuridade no brasil: implementação, correlação intraclasse e fatores associados à prematuridade espontânea= multicenter study on preterm birth in brazil: implementation, intracluster correlation and associated factors to spontaneous preterm birth. Citado na pág. Citado na página 25.

Liberal Pereira(2010) Tarciana Liberal Pereira. Regressão beta inflacionada: inferência e aplicações. Citado na pág. Citado 2 vezes nas páginas 44 e 46.

Lima(2010) Claudia Risso de Araujo Lima. Gestão da qualidade dos dados e informações dos sistemas de informação em saúde: subsídios para a construção de uma metodologia adequada ao brasil. Citado na pág. Citado na página 29.

Little e Rubin(2019) Roderick JA Little e Donald B Rubin. *Statistical analysis with missing data*. John Wiley & Sons. Citado na pág. Citado na página 24.

Martins(2014) Ricardo Marcondes Martins. Direito fundamental de acesso à informação. *A&C-Revista de Direito Administrativo & Constitucional*, 14(56):127–146. Citado na pág. Citado na página 23.

Merino *et al.*(2016) Jorge Merino, Ismael Caballero, Bibiano Rivas, Manuel Serrano e Mario Piattini. A data quality in use model for big data. *Future Generation Computer Systems*, 63:123–130. Citado na pág. Citado na página 29.

Ministério da Saúde(2019) BRASIL. Ministério da Saúde. Saúde brasil 2019: uma análise da situação de saúde com enfoque nas doenças imunopreveníveis e na imunização, 2019. Citado na pág. Citado 2 vezes nas páginas 23 e 30.

Moral *et al.*(2017) Rafael A Moral, John Hinde e Clarice GB Demétrio. Half-normal plots and overdispersed models in r: the hnp package. *Journal of Statistical Software*, 81(1): 1–23. Citado na pág. Citado na página 48.

Moreira *et al.*(2017) Fábio Mosso Moreira, Ricardo César Gonçalves Sant'Ana, Plácida Leopoldina Ventura Amorim da Costa Santos e Zaira Regina Zafalon. Metadados para descrição de datasets e recursos informacionais do “portal brasileiro de dados abertos”. *Perspectivas em Ciência da Informação*, 22:158–185. Citado na pág. Citado na página 23.

Nishi(2010) Lisandro Fin Nishi. Coeficiente de gini: uma medida de distribuição de renda. *Florianópolis: Universidade do Estado de Santa Catarina*. Citado na pág. Citado na página 33.

- Organização Pan-americana de Saúde(2002) Ministério da Saúde Organização Pan-americana de Saúde. *Indicadores básicos para a saúde no Brasil: conceitos e aplicações*. Brasil. Ministério da Saúde. Citado na pág. Citado 3 vezes nas páginas 32, 33 e 34.
- Origuela(2018) Letícia Aparecida Origuela. *Estudo da influência de eventos sobre a estrutura do mercado brasileiro de ações a partir de redes ponderadas por correlações de Pearson, Spearman e Kendall*. Tese de Doutorado, Universidade de São Paulo. Citado na pág. Citado na página 35.
- Ospina(2007) Patricia Leone Espinheira Ospina. *Regressão beta*. Tese de Doutorado, Universidade de São Paulo. Citado na pág. Citado 7 vezes nas páginas 36, 37, 38, 39, 40, 41 e 42.
- Ospina e Ferrari(2010) Raydonal Ospina e Silvia LP Ferrari. Inflated beta distributions. *Statistical papers*, 51(1):111–126. Citado na pág. Citado 2 vezes nas páginas 44 e 46.
- Ospina Martinez(2008) Raydonal Ospina Martinez. *Modelos de regressão beta inflacionados*. Tese de Doutorado, Universidade de São Paulo. Citado na pág. Citado 2 vezes nas páginas 44 e 45.
- Paolino(2001) Philip Paolino. Maximum likelihood estimation of models with beta-distributed dependent variables. *Political Analysis*, 9(4):325–346. Citado na pág. Citado na página 39.
- Passini Jr *et al.*(2014) Renato Passini Jr, Jose G Cecatti, Giuliane J Lajos, Ricardo P Tedesco, Marcelo L Nomura, Tabata Z Dias, Samira M Haddad, Patricia M Rehder, Rodolfo C Pacagnella, Maria L Costa *et al.* Brazilian multicentre study on preterm birth (emip): prevalence and factors associated with spontaneous preterm birth. *PLoS one*, 9(10):e109069. Citado na pág. Citado na página 25.
- Paula(2004) Gilberto Alvarenga Paula. *Modelos de regressão: com apoio computacional*. IME-USP São Paulo. Citado na pág. Citado na página 37.
- PNAD(2018) PNAD. *Pesquisa Nacional por Amostra de Domicílios Contínua*, 2018. Citado na pág. Citado na página 26.
- Prater(1956) NH Prater. Estimate gasoline yields from crudes. *Petroleum Refiner*, 35(5): 236–238. Citado na pág. Citado na página 43.
- R Core Team(2020) R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020. URL <https://www.R-project.org/>. Citado na pág. Citado na página 29.

Ramos e Cuman(2009) Helena Ângela de Camargo Ramos e Roberto Kenji Nakamura Cuman. Fatores de risco para prematuridade: pesquisa documental. *Escola Anna Nery*, 13:297–304. Citado na pág. Citado na página 25.

Romero e Cunha(2006) Dalia E Romero e Cynthia Braga da Cunha. Avaliação da qualidade das variáveis sócio-econômicas e demográficas dos óbitos de crianças menores de um ano registrados no sistema de informações sobre mortalidade do brasil (1996/2001). *Cadernos de Saúde Pública*, 22:673–681. Citado na pág. Citado 2 vezes nas páginas 30 e 65.

Silva *et al.*(2001) AA Silva, Valdinar Sousa Ribeiro, AF Borba Jr, Liberata Campos Coimbra e Raimundo Antonio da Silva. Evaluation of data quality from the information system on live births in 1997-1998. *Revista de Saúde Pública*, 35(6):508–514. Citado na pág. Citado na página 24.

Silva(2020) Erika Rayanne Fernandes da Silva. Modelo de regressão beta modal. Dissertação de Mestrado, Brasil. Citado na pág. Citado na página 37.

Silva *et al.*(2013) Ricarly Soares da Silva, Conceição Maria de Oliveira, Daniela Karina da Silva Ferreira e Cristine Vieira do Bonfim. Avaliação da completitude das variáveis do sistema de informações sobre nascidos vivos (sinasc) nos estados da região nordeste do brasil, 2000 e 2009. *Epidemiologia e Serviços de Saúde*, 22(2):347–352. Citado na pág. Citado na página 24.

Simas *et al.*(2010) Alexandre B Simas, Wagner Barreto-Souza e Andréa V Rocha. Improved estimators for a general class of beta regression models. *Computational Statistics & Data Analysis*, 54(2):348–366. Citado na pág. Citado na página 48.

Smithson e Verkuilen(2006) Michael Smithson e Jay Verkuilen. A better lemon squeeizer? maximum-likelihood regression with beta-distributed dependent variables. *Psychological methods*, 11(1):54. Citado na pág. Citado na página 39.

Stasinopoulos *et al.*(2007) D Mikis Stasinopoulos, Robert A Rigby *et al.* Generalized additive models for location scale and shape (gamlss) in r. *Journal of Statistical Software*, 23(7):1–46. Citado na pág. Citado 2 vezes nas páginas 46 e 47.

Theme Filha *et al.*(2004) Mariza Miranda Theme Filha, Silvana Granado Nogueira da Gama, Cynthia Braga da Cunha e Maria do Carmo Leal. Confiabilidade do sistema de informações sobre nascidos vivos hospitalares no município do rio de janeiro, 1999-2001. *Cadernos de Saúde Pública*, 20:S83–S91. Citado na pág. Citado na página 24.

Varella(2021) Portal Drauzio Varella. Bebês prematuros: tudo o que você precisa saber, 2021. URL <https://drauziovarella.uol.com.br/pediatrica/>

- [bebés-prematuros-tudo-o-que-voce-precisa-saber/](#). Acessado em Junho 21, 2021.
Citado na pág. Citado na página 25.
- Walani(2020) Salimah R Walani. Global burden of preterm birth. *International Journal of Gynecology & Obstetrics*, 150(1):31–33. Citado na pág. Citado na página 25.
- WHO(2012) World Health Organization WHO. Born too soon: the global action report on preterm birth. Citado na pág. Citado na página 24.
- WHO(2013) World Health Organization WHO. Programa das nações unidas para o desenvolvimento. atlas do desenvolvimento humano no brasil. *World Health Organization*.
Citado na pág. Citado 2 vezes nas páginas 25 e 31.
- Zeileis *et al.*(2016) Achim Zeileis, Francisco Cribari-Neto, Bettina Gruen, Ioannis Kosmidis, Alexandre B Simas, Andrea V Rocha e Maintainer Achim Zeileis. Package ‘betareg’. *R package*, 3:2. Citado na pág. Citado na página 39.

Apêndices

APÊNDICE A – Análise descritiva - UF

```

library(dplyr)          # operador pipe (manipulação de banco de dados)
library(ggplot2)         # gráficos
library(GGally)          # análise descritiva
library(summarytools)    # análise descritiva
library(modelsummary)    # análise descritiva
library(Hmisc)           # matriz de correlação com valor p
library(corrplot)        # gráfico correlograma
library(rgdal)            # ler arquivo ORG
library(readxl)           # ler arquivo xlsx
library(htmltools)        # tabelas descritivas
library(highcharter)     # fazer gráficos interativos (mapas de calor)
library(stringr)          # trabalha com strings
library(mapview)          # salvar mapas estáticos
library(htmlwidgets)      # salvar mapas dinâmicos
library(leaflet)          # mapas
library(writexl)          # exportar tabelas em xlsx
library(geobr)             # informações do IBGE

# Lendo os dados - UF
dados <- read_excel("dados_uf.xlsx")

## Criando as linhas com os dados do Brasil (soma de todos os nascimentos)
dado <- dados %>%
  group_by(Ano) %>%
  summarise(BR = sum(Nascimentos))

#-----#
# Análise de correlação
df <- read_excel("dados_UF.xlsx")

## Análise para os indicadores obstétricos
### Ano de 2016
df1 <- df %>%
  filter(Ano == 2016)

```

```
names(df1)[c(4, 5, 6, 7, 8, 17)] <- c("Prematuridade", "Gravidez", "Parto", "Consultas",
                                         "Anomalias", "Esp_vida")

dados_select <- select(df1, -UF, - Ano, -Nascimentos, -IDHM, -IDHM_edu, -IDHM_long,
                        -IDHM_ren, -Gini, -Analfabetismo, -Agua, -Esgoto, -Esp_vida)

# rcorr(as.matrix(dados_select1), type = "spearman")$r

corplot.mixed(cor(dados_select, use = "pairwise",
                   method = "spearman"))

#-----

### Ano de 2017
df2 <- df %>%
  filter(Ano == 2017)

names(df2)[c(4, 5, 6, 7, 8, 17)] <- c("Prematuridade", "Gravidez", "Parto", "Consultas",
                                         "Anomalias", "Esp_vida")

dados_select1 <- select(df2, -UF, - Ano, -Nascimentos, -IDHM, -IDHM_edu, -IDHM_long,
                        -IDHM_ren, -Gini, -Analfabetismo, -Agua, -Esgoto, -Esp_vida)

corplot.mixed(cor(dados_select1, use = "pairwise",
                   method = "spearman"))

#-----

### Ano de 2018
df3 <- df %>%
  filter(Ano == 2018)

names(df3)[c(4, 5, 6, 7, 8, 17)] <- c("Prematuridade", "Gravidez", "Parto", "Consultas",
                                         "Anomalias", "Esp_vida")

dados_select2 <- select(df3, -UF, - Ano, -Nascimentos, -IDHM, -IDHM_edu, -IDHM_long,
                        -IDHM_ren, -Gini, -Analfabetismo, -Agua, -Esgoto, -Esp_vida)
```

```

corrplot.mixed(cor(dados_select2, use = "pairwise",
                    method = "spearman"))

#-----

### Ano de 2019
df4 <- df %>%
  filter(Ano == 2019)

names(df4)[c(4, 5, 6, 7, 8)] <- c("Prematuridade", "Gravidez", "Parto", "Consultas",
                                    "Anomalias")

dados_select3 <- select(df4, -UF, - Ano, -Nascimentos, -IDHM, -IDHM_edu, -IDHM_long
                        -IDHM_ren, -Gini, -Analfabetismo, -Agua, -Esgoto, -Esperanc

corrplot.mixed(cor(dados_select3, use = "pairwise",
                    method = "spearman"))

#-----

## Análise para os indicadores socioeconômicos
### Ano de 2019
df4 <- df %>%
  filter(Ano == 2019)

names(df4)[c(4, 5, 6, 7, 8)] <- c("Prematuridade", "Gravidez", "Parto", "Consultas",
                                    "Anomalias")

dados_select4 <- select(df4, -UF, - Ano, -Nascimentos, -Gravidez, -Parto, -Consulta
                        -Anomalias)

corrplot.mixed(cor(dados_select4, use = "pairwise",
                    method = "spearman"))

#####
dados_comp <- read_excel("dados1_UF.xlsx")

# Mapas de calor

```

```
## Transformando os dados para fator
dados_comp$UF <- as.factor(dados_comp$UF) # dados_comp = dados1_UF (tem o Brasil)

## Alterando a ordem dos estados (separando por região)
dados_comp$UF <- factor(dados_comp$UF,
                         levels = c("BR",
                                   "AC", "AM", "AP", "PA", "RO", "RR", "TO",
                                   "AL", "BA", "CE", "MA", "PE", "PI", "PB", "RN", "SE",
                                   "DF", "GO", "MS", "MT",
                                   "ES", "MG", "RJ", "SP",
                                   "PR", "RS", "SC"))

#----- #

## Mapa da incompletude para prematuridade
fntltp <- JS("function(){
  return this.point.x + ' ' + this.series.yAxis.categories[this.point.y] + ':<br>' +
  Highcharts.numberFormat(this.point.value, 2);
}")

premat <- hchart(
  dados_comp,
  "heatmap",
  hcAxes(
    x = Ano,
    y = UF,
    value = incomp_premat
  )
) %>%
  hc_colorAxis(
    stops = color_stops(10, viridisLite::inferno(10, direction = -1))
  ) %>%
  hc_yAxis(
    title = list(text = ""),
    reversed = TRUE,
    offset = -20,
    tickLength = 0,
```

```

gridLineWidth = 0,
minorGridLineWidth = 0,
labels = list(style = list(fontSize = "13px"))
) %>%

hc_xAxis(
  title = list(text = "Ano"),
  gridLineWidth = 0,
  offset = 20,
  labels = list(style = list(fontSize = "13px"))
) %>%

hc_tooltip(
  formatter = fntltp
) %>%

#hc_xAxis(
#  plotLines = list(plotline)) %>%
hc_title(
  text = "Taxa de incompletude para prematuridade"
) %>%

#hc_subtitle(
#  text = "Número de casos por 100,000 pessoas"
#) %>%

hc_legend(
  layout = "horizontal",
  verticalAlign = "top",
  align = "left",
  valueDecimals = 0
) %>%

hc_size(height = 1000)

premat

### Salvar versão dinâmica do gráfico
htmlwidgets::saveWidget(premat, file = 'prematuridade.html')

```

```
### Salvar versão estática do gráfico
mapshot(premat, file = "prematuridade.png")
#####
# Mapas do Brasil
## Lendo os dados
map1 <- read_excel("dados_UF.xlsx")

## Renomeando as UF's (deixar igual ao arquivo shapefile)
map2 <- map1 %>%
  mutate(UF = forcats::fct_recode(UF,
    "ACRE" = "AC",
    "ALAGOAS" = "AL",
    "AMAZONAS" = "AM",
    "AMAPA" = "AP",
    "BAHIA" = "BA",
    "CEARA" = "CE",
    "DISTRITO FEDERAL" = "DF",
    "ESPIRITO SANTO" = "ES",
    "GOIAS" = "GO",
    "MARANHAO" = "MA",
    "MINAS GERAIS" = "MG",
    "MATO GROSSO DO SUL" = "MS",
    "MATO GROSSO" = "MT",
    "PARA" = "PA",
    "PARAIBA" = "PB",
    "PERNAMBUCO" = "PE",
    "PIAUI" = "PI",
    "PARANA" = "PR",
    "RIO DE JANEIRO" = "RJ",
    "RIO GRANDE DO NORTE" = "RN",
    "RONDONIA" = "RO",
    "RORAIMA" = "RR",
    "RIO GRANDE DO SUL" = "RS",
    "SANTA CATARINA" = "SC",
    "SERGIPE" = "SE",
    "SAO PAULO" = "SP"
  ))
```

```

    "TOCANTINS"           = "TO"))
#-----#
## IDHM 2016
### Filtrando o ano de análise
map3 <- map2 %>%
  filter(Ano == 2016)

### Carregando os dados do arquivo shapefile
estados <- readOGR("data/UFEBRASIL.shp")

estados@data$UF <-
  iconv(estados@data$NM_ESTADO, from = 'UTF-8', to = 'ASCII//TRANSLIT') %>%
  toupper()

estados@data <- estados@data %>%
  left_join(map3)

es_ll <- spTransform(estados, CRS("+init=epsg:4326"))

### Customizando a paleta de cores
mypalette_l <- colorNumeric(palette = "viridis",
                             domain = es_ll@data>IDHM,
                             na.color = "transparent")

### Customizando a legenda
labels <- paste("<p>", "Estado: ", es_ll@data$UF, "</p>",
                "<p>", "IDHM: ",
                es_ll@data>IDHM, "</p>",
                sep = "") %>%
  lapply(htmltools::HTML)

### Fazendo o gráfico
mapa_idhm <- leaflet() %>%
  setView(lng = -56.0949, lat = -15.5989, zoom = 4) %>%
  addProviderTiles(providers$CartoDB.Positron) %>%
  addPolygons(data = estados,
              fillColor = ~mypalette_l(IDHM),

```

```
weight = 1,  
smoothFactor = 0.5,  
color = "white",  
fillOpacity = 0.8,  
highlight = highlightOptions(  
    weight = 5,  
    color = "#666",  
    fillOpacity = 0.7,  
    bringToFront = TRUE),  
    label = labels) %>%  
addLegend(data = es_ll,  
    pal = mypalette_l,  
    values = ~ IDHM,  
    opacity = 0.7,  
    title = "IDHM",  
    position = "bottomleft")
```

mapa_idhm

```
### Salva a versão estática do mapa  
mapshot(mapa_idhm, file = "idhm2016.png")  
  
### Salva a versão dinâmica do mapa  
saveWidget(mapa_idhm, file = "idhm2016.html")
```

APÊNDICE B – Análise descritiva - Municípios

```

library(dplyr)          # operador pipe (manipulação de banco de dados)
library(ggplot2)         # gráficos
library(GGally)          # análise descritiva
library(summarytools)    # análise descritiva
library(modelsummary)    # análise descritiva
library(Hmisc)           # matriz de correlação com valor p
library(corrplot)        # gráfico correlograma
library(rgdal)           # ler arquivo ORG
library(readxl)          # ler arquivo xlsx
library(htmltools)        # tabelas descritivas
library(highcharter)     # fazer gráficos interativos (mapas de calor)
library(stringr)          # trabalha com strings
library(mapview)          # salvar mapas estáticos
library(htmlwidgets)      # salvar mapas dinâmicos
library(leaflet)          # mapas
library(writexl)          # exportar tabelas em xlsx
library(geobr)            # informações do IBGE

# Juntando as informações pra análise descritiva
completo <- read_excel("Dados_muni_wide.xlsx")

completo <- completo %>%
  filter(Ano >= 2016) %>%
  mutate(Incomp_pemat = round(faltante_pemat/(Nascimentos)*100, 4),
         Incomp_grav = round(faltante_tipo_gesta/(Nascimentos)*100, 4),
         Incomp_parto = round(faltante_tipo_parto/(Nascimentos)*100, 4),
         Incomp_consulta = round(faltante_consulta/(Nascimentos)*100, 4),
         Incomp_anomalias = round(faltante_anomalia/(Nascimentos)*100, 4)) %>%
  dplyr::select(UF, Municipio, Codigo, Ano, Nascimentos, Incomp_pemat, Incomp_grav,
                Incomp_parto, Incomp_consulta, Incomp_anomalias)

## Exporta a tabela para arquivo csv
# write.table(completo, "dados2_muni.csv")

```

```

write_xlsx(completo, "dados_muni-sinasc.xlsx")

#-----

# Análise de correlação

df <- read_excel("dados_muni-sinasc.xlsx")

## Variáveis obstétricas
### Ano 2019
df <- df %>%
  filter(Ano == 2019)

names(df)[c(6, 7, 8, 9, 10)] <- c("Prematuridade", "Gravidez", "Parto", "Consultas",
                                    "Anomalias")

dados_select <- select(df, -UF, -Municipio, -Codigo, -Ano, -Nascimentos)

# rcorr(as.matrix(dados_select), type = "spearman")$r

corrplot.mixed(cor(dados_select, use = "pairwise",
                    method = "spearman")) # corr_muni2019-sinasc

## Variáveis socioeconômicas
df1 <- read_excel("dados_muni-descr2019.xlsx")

names(df1)[c(4, 5)] <- c("Prematuridade", "Esp_vida")

dados_select <- select(df1, -UF, -Municipio, -Codigo)

# rcorr(as.matrix(dados_select), type = "spearman")$r

corrplot.mixed(cor(dados_select, use = "pairwise",
                    method = "spearman")) # corr_muni2019-indic

#####
# Mapas de calor dos municípios
## Variáveis obstétricas

```

```

#### Incompletude prematuridade 2019
##### Filtrando o ano de análise
dados <- read_excel("dados_muni-sinasc.xlsx")

dados <- dados %>%
  filter(Ano == 2019)

municipios <- read_municipality(code_muni = "all", year = 2019)

municipios$Municipio <-
  iconv(municipios$name_muni, from = 'UTF-8', to = 'ASCII//TRANSLIT') %>%
  toupper()

municipios <- municipios %>%
  left_join(dados)

write_xlsx(municipios, "dados_muni-mapa.xlsx")

##### Customizando a paleta de cores
mypalette_l <- colorNumeric(palette = "viridis",
                               domain = municipios$Incomp_premat,
                               na.color = "transparent")

##### Customizando a legenda
labels <- paste("<p>", "Municipio: ", municipios$Municipio, "</p>",
                "<p>", "UF: ", municipios$UF, "</p>",
                "<p>", "(%) incompletude prematuridade: ",
                municipios$Incomp_premat, "</p>",
                sep = "") %>%
  lapply(htmltools::HTML)

##### Fazendo o gráfico
incomp_premat <- leaflet() %>%
  setView(lng = -56.0949, lat = -15.5989, zoom = 4) %>%
  addProviderTiles(providers$CartoDB.Positron) %>%
  addPolygons(data = municipios,
              fillColor = ~mypalette_l(Incomp_premat),
              weight = 1,
              smoothFactor = 0.5,

```

```

color = "white",
fillOpacity = 0.8,
highlight = highlightOptions(
  weight = 5,
  color = "#666",
  fillOpacity = 0.7,
  bringToFront = TRUE),
label = labels) %>%
addLegend(data = municipios,
  pal = mypalette_l,
  values = ~ Incomp_premat,
  opacity = 0.7,
  title = "(%) incompletude prematuridade",
  position = "bottomleft")

incomp_premat

##### Salva a versão estática do mapa
mapshot(incomp_premat, file = "incomp_premat2019.png")

##### Salva a versão dinâmica do mapa
saveWidget(incomp_premat, file = "incomp_premat2019.html")

#-----

### IDHM 2010

##### Customizando a paleta de cores
mypalette_l <- colorNumeric(palette = "viridis",
  domain = municipios$IDHM,
  na.color = "transparent")

##### Customizando a legenda
labels <- paste("<p>", "Municipio: ", municipios$Municipio, "</p>",
  "<p>", "UF: ", municipios$UF, "</p>",
  "<p>", "IDHM: ",
  municipios$IDHM, "</p>",
  sep = "") %>%
lapply(htmltools::HTML)

```

```
#### Fazendo o gráfico
idhm <- leaflet() %>%
  setView(lng = -56.0949, lat = -15.5989, zoom = 4) %>%
  addProviderTiles(providers$CartoDB.DarkMatter) %>%
  addPolygons(data = municipios,
    fillColor = ~mypalette_l(IDHM),
    weight = 1,
    smoothFactor = 0.5,
    color = "white",
    fillOpacity = 0.8,
    highlight = highlightOptions(
      weight = 5,
      color = "#666",
      fillOpacity = 0.7,
      bringToFront = TRUE),
    label = labels) %>%
  addLegend(data = municipios,
    pal = mypalette_l,
    values = ~ IDHM,
    opacity = 0.7,
    title = "IDHM",
    position = "bottomleft")
```

idhm

```
#### Salva a versão estática do mapa
mapshot(idhm, file = "idhm2010.png")
```

```
#### Salva a versão dinâmica do mapa
saveWidget(idhm, file = "idhm2010.html")
```


APÊNDICE C – Modelagem - UF

```

# Carregando os pacotes necessários
library(readxl) # ler arquivo xlsx
library(dplyr) # operador pipe (manipulação de banco de dados)
library(ggplot2) # gráficos
library(corrplot) # gráfico correlograma
library(betareg) # modelagem regressão beta
library(lmtest) # teste da razão de verossimilhança
library(lattice) # gráficos de dependência

# Carregando os dados
dados <- read_excel("dados_UF_modelo.xlsx")

# Colocando a região Sudeste como categoria de referência
dados1 <- dados %>%
  mutate(Regiao = factor(Regiao,
                        levels = c("Sudeste", "Sul", "Nordeste", "Norte", "Centro-Oeste")))

names(dados1)[c(5, 6, 7, 8, 9, 18)] <- c("Prematuridade", "Gravidez", "Parto",
                                         "Consultas", "Anomalias", "Esp_vida")

## Função de ligação logit
### Com IDHM
#### Modelo com dispersão fixa
fit1 <- betareg(Prematuridade ~ IDHM + Agua + Esgoto + Esp_vida, data = dados1, link = "logit")
summary(fit1)

fit2 <- betareg(Prematuridade ~ IDHM + Agua, data = dados1, link = "logit")
summary(fit2)
AIC(fit1, fit2)

#### Modelos com dispersão variável
fit2.1 <- betareg(Prematuridade ~ IDHM + Agua | IDHM, data = dados1, link = "logit")
summary(fit2.1)
fit2.2 <- betareg(Prematuridade ~ IDHM + Agua | Agua, data = dados1, link = "logit")
summary(fit2.2)

```

```

fit2.3 <- betareg(Prematuridade ~ IDHM + Agua | IDHM + Agua, data = dados1, link = "logit")
summary(fit2.3)

##### Teste da Razão de Verossimilhança (TRV): dispersão fixa vs dispersão variável
lrtest(fit2, fit2.1)
lrtest(fit2, fit2.2)
lrtest(fit2, fit2.3)

AIC(fit2, fit2.1, fit2.2, fit2.3)

### Com IDHM renda, Longevidade e Renda
##### Modelo com dispersão fixa
fit3 <- betareg(Prematuridade ~ IDHM_ren + IDHM_long + Agua + Esgoto + Esp_vida,
                 data = dados1, link = "logit")
summary(fit3)

fit4 <- betareg(Prematuridade ~ IDHM_ren + Agua, data = dados1, link = "logit")
summary(fit4)
AIC(fit3, fit4)

##### Modelos com dispersão variável
fit4.1 <- betareg(Prematuridade ~ IDHM_ren + Agua | IDHM_ren, data = dados1, link = "logit")
summary(fit4.1)
fit4.2 <- betareg(Prematuridade ~ IDHM_ren + Agua | Agua, data = dados1, link = "logit")
summary(fit4.2)
fit4.3 <- betareg(Prematuridade ~ IDHM_ren + Agua | IDHM_ren + Agua, data = dados1, link = "logit")
summary(fit4.3)

##### Teste da Razão de Verossimilhança (TRV): dispersão fixa vs dispersão variável
lrtest(fit4, fit4.1)
lrtest(fit4, fit4.2)
lrtest(fit4, fit4.3)

AIC(fit4, fit4.1, fit4.2, fit4.3)

# Análise de diagnóstico
suppressWarnings(RNGversion("3.5.0"))
set.seed(123)
par(mfrow = c(3, 2))

```

```

plot(fit2, which = 1:6, type = "sweighted2")

#-----
# Gráficos de dependência
Agua <- dados$Agua
IDHM <- dados$IDHM

# função que calcula mu como função de idhm e agua
mu_calculo <- function(idhm, agua){
  out <- exp(5.76 - 11.35 * idhm - 0.02 * agua)/(1 + exp(5.76 - 11.35 * idhm - 0.02
  return(out)
}

# 1º gráfico
idhm <- seq(0, 1, by = 0.005)
plot(idhm, mu_calculo(idhm, agua = 85.71), xlab = "IDHM", ylab = expression(mu))

#-----
# 2º gráfico
# Fixa o IDHM no valor médio (0.75) e faz o gráfico de mu versus água
agua <- seq(0, 100, by = 0.5)
plot(agua, mu_calculo(idhm = 0.75, agua), xlab = "Água", ylab = expression(mu))

#-----
# 3º gráfico: gráfico de calor

idhm_valor <- seq(0, 1, length.out = 20) # idhm
agua_valor <- seq(0, 100, length.out = 20) # água

data <- expand.grid(IDHM = idhm_valor, Agua = agua_valor)
data$mu <- mu_calculo(idhm = data$IDHM, agua = data$Agua) #mu que vai calcular como

levelplot(mu ~ IDHM * Agua, data = data, xlab = "IDHM", ylab = "Água", main = "")

```


APÊNDICE D – Modelagem - Municípios

```

library(readxl) # ler arquivo xlsx
library(dplyr) # operador pipe (manipulação de banco de dados)
library(writexl) # exportar tabelas em xlsx
library(gamlss) # modelagem inflated beta
library(hnp) # gráficos de envelope
library(lattice) # gráficos de dependência
set.seed(1)

# Carregando os dados
dados_pemat <- read_excel("prematuridade_muni_wide.xlsx")

indic_2010_muni <- read_excel("indic_censo_2010_muni.xlsx")
View(indic_2010_muni)

indic_2010_muni <- indic_2010_muni %>%
  rename(Codigo = Codmun6)

options(scipen=999)

# Filtrando os dados: nascidos >= 100
dados_pemat_2019 <- dados_pemat %>%
  filter(Ano == 2019)

dados_pemat_2019 <- dados_pemat_2019 %>%
  mutate(incomp_pemat = (faltante_pemat/total_nascidos))

dados_pemat_2019 <- dados_pemat_2019 %>%
  dplyr::select(UF, Municipio, Codigo, total_nascidos, incomp_pemat) %>%
  filter(total_nascidos >= 100)

# Juntando os dados
dados <- left_join(dados_pemat_2019, indic_2010_muni, by = "Codigo")
dados1 <- na.omit(dados)

```

```
# Exportando os dados
write_xlsx(dados1, "dados_muni-modelo.xlsx")

#-----

# Seleção dos modelos
## Modelo 1: com IDHM, GINI e AGUA_ESGOTO
aj0 <- gamlss(incomp_pemat ~ IDHM + GINI + AGUA_ESGOTO,
               nu.formula = ~ IDHM + GINI + AGUA_ESGOTO,
               sigma.formula = ~ IDHM + GINI + AGUA_ESGOTO,
               family = BEINFO , data = dados1)
summary(aj0)

aj1 <- gamlss(incomp_pemat ~ IDHM,
               nu.formula = ~ IDHM + GINI + AGUA_ESGOTO,
               sigma.formula = ~ IDHM + GINI + AGUA_ESGOTO,
               family = BEINFO , data = dados1)
summary(aj1)

## Comparando os ajustes para o modelo 1
AIC(aj0, aj1)

#-----

## Modelo 2: com IDHM_E, IDHM_L, IDHM_R, GINI e AGUA_ESGOTO
aj2 <- gamlss(incomp_pemat ~ IDHM_E + IDHM_L + IDHM_R + GINI + AGUA_ESGOTO,
               nu.formula = ~ IDHM_E + IDHM_L + IDHM_R + GINI + AGUA_ESGOTO,
               sigma.formula = ~ IDHM_E + IDHM_L + IDHM_R + GINI + AGUA_ESGOTO,
               family = BEINFO , data = dados1)
summary(aj2)

aj3 <- gamlss(incomp_pemat ~ IDHM_E + IDHM_R,
               nu.formula = ~ IDHM_E + IDHM_L + IDHM_R + GINI + AGUA_ESGOTO,
               sigma.formula = ~ IDHM_E + IDHM_L + IDHM_R + GINI,
               family = BEINFO , data = dados1)
summary(aj3)
```

```

aj4 <- gammelss(incomp_premat ~ IDHM_E + IDHM_R,
                 nu.formula = ~ IDHM_E + IDHM_L + IDHM_R + GINI + AGUA_ESGOTO,
                 sigma.formula = ~ IDHM_E + IDHM_L + GINI,
                 family = BEINFO, data = dados1)
summary(aj4)

## Comparando os ajustes para o modelo 2
AIC(aj2, aj3, aj4)

#-----
## Modelo 3: com IDHM, GINI e AGUA_ESGOTO, total_nascidos
aj0_n <- gammelss(incomp_premat ~ IDHM + GINI + AGUA_ESGOTO + total_nascidos,
                   nu.formula = ~ IDHM + GINI + AGUA_ESGOTO + total_nascidos,
                   sigma.formula = ~ total_nascidos,
                   family = BEINFO, data = dados1)
summary(aj0_n)

aj1_n <- gammelss(incomp_premat ~ IDHM + total_nascidos,
                   nu.formula = ~ IDHM + AGUA_ESGOTO + total_nascidos,
                   sigma.formula = ~ total_nascidos,
                   family = BEINFO, data = dados1)
summary(aj1_n)

## Comparando os ajustes para o modelo 3
AIC(aj0_n, aj1_n)

#-----
## Modelo 4: com IDHM_E, IDHM_L, IDHM_R, GINI e AGUA_ESGOTO, total_nascidos
aj2_n <- gammelss(incomp_premat ~ IDHM_E + IDHM_L + IDHM_R + GINI + AGUA_ESGOTO + total_nascidos,
                   nu.formula = ~ IDHM_E + IDHM_L + IDHM_R + GINI + AGUA_ESGOTO + total_nascidos,
                   sigma.formula = ~ IDHM_E + IDHM_L + IDHM_R + GINI + AGUA_ESGOTO + total_nascidos,
                   family = BEINFO, data = dados1)
summary(aj2_n)

aj3_n <- gammelss(incomp_premat ~ IDHM_E + IDHM_R + total_nascidos,
                   nu.formula = ~ IDHM_L + IDHM_R + AGUA_ESGOTO + total_nascidos,
                   sigma.formula = ~ IDHM_E + IDHM_L + IDHM_R + GINI,

```

```

family = BEINFO, data = dados1)
summary(aj3_n)

## Comparando os ajustes para o modelo 4
AIC(aj2_n, aj3_n)

## Critérios de informação (comparando todos os modelos feitos acima)
AIC(aj1, aj7, aj0_n, aj1_n, aj2_n, aj3_n)

#-----
## Análise de diagnóstico
### Para o melhor ajuste - aj3_n
dados1$id <- 1:dim(dados1)[1]

res <- aj3_n$residuals
residuo <- data.frame(id = dados1$id, res = res )
residuo  <- residuo[residuo$res < Inf, ]

plot(res, ylim = c(-5 ,8), xlab = "Índice", ylab = "Resíduo Quantil Aleatorizado",
     main = "Resíduo versus Índice")

# Refazendo o ajuste sem as observações com resíduos >2 e <-2 pra
# ver se muda o ajuste
resid1 <- residuo %>%
  dplyr::filter(res > -2 & res < 2)

dados2 <- left_join(resid1, dados1, by = "id")

aj3_n1 <- gammLSS(incomp_premat ~ IDHM_E + IDHM_R + total_nascidos,
                  nu.formula = ~ IDHM_L + IDHM_R + AGUA_ESGOTO + total_nascidos,
                  sigma.formula = ~ IDHM_E + IDHM_L + IDHM_R + GINI,
                  family = BEINFO , data = dados2)
summary(aj3_n1)

ajustado <- fitted(aj3_n)
plot(ajustado, res, xlab = "Valores Ajustados", ylab = "Resíduo Quantil Aleatorizado",
     main = "Resíduo versus Valores Ajustados")

```

```

# Envelope

d.fun <- function(obj) resid (obj)
s.fun <- function(n, obj) rBEINFO(n, obj$mu.fv, obj$sigma.fv, obj$nu.fv)
f.fun <- function(y.) {
  gamlss(incomp_premat ~ IDHM_E + IDHM_R + total_nascidos,
         nu.formula = ~ IDHM_L + IDHM_R + AGUA_ESGOTO + total_nascidos,
         sigma.formula = ~ IDHM_E + IDHM_L + IDHM_R + GINI,
         family = BEINFO , data = dados1)
}

hnp(aj3_n, newclass = TRUE , diagfun = d.fun, simfun = s.fun,
    fitfun = f.fun, xlab = "Half-normal scores", ylab = "z-scores ",
    main = "", pch = 20, cex = 1, cex.lab = .8, cex.axis = .8)

#-----
# Gráficos de dependência

IDHM_E <- dados$IDHM_E
IDHM_R <- dados$IDHM_R
Nascimentos <- dados$total_nascidos

# função que calcula mu como função de idhm_e, idhm_r, nascimentos
mu_calculo <- function(idhm_e, idhm_r, nascimentos){
  out <- exp(-0.568593854 - 2.049015472 * idhm_e - 3.064539675 * idhm_r - 0.0000191
  (1 + exp(-0.568593854 - 2.049015472 * idhm_e - 3.064539675 * idhm_r - 0.0000191
  return(out)
}

#-----
# 1) Fixa IDHM Educação e IDHM Renda no valor médio. Grafico de mu versus valores de
idhm_r <- seq(0, 1, by = 0.00029352)
nascimentos <- seq(100, 158587, by = 46.52)

plot(nascimentos, mu_calculo(idhm_e = 0.5567, idhm_r = 0.640, nascimentos), xlab =

#-----
# 2) Fixa IDHM Renda e Nascimentos no valor médio. Gráfico de mu versus valores de
idhm_e <- seq(0, 1, by = 0.00029352)

```

```
plot(idhm_e, mu_calculo(idhm_e, idhm_r = 0.640, nascimentos = 801.9), xlab = "IDHM Educação",  
#-----  
# 3) Fixa IDHM Educação e Nascimentos no valor médio. Gráfico de mu versus valores de IDHM  
idhm_r <- seq(0, 1, by = 0.00029352)  
  
plot(idhm_r, mu_calculo(idhm_e = 0.5567, idhm_r, nascimentos = 801.9), xlab = "IDHM Renda",  
#-----  
# 4) Fixa Nascimentos na média. Gráfico de calor, em que: X: IDHM Educação, Y: IDHM Renda  
nascimentos <- 801.9 # média  
idhme_valor <- seq(0, 1, length.out = 20) # idhm educação  
idhmr_valor <- seq(0, 1, length.out = 20) # idhm renda  
  
data <- expand.grid(IDHMe = idhme_valor, IDHMr = idhmr_valor)  
data$mu <- mu_calculo(idhm_e = data$IDHMe, idhm_r = data$IDHMr, nascimentos) #mu que vai para o heatmap  
  
levelplot(mu ~ IDHMe * IDHMr, data = data, xlab = "IDHM Educação", ylab = "IDHM Renda",
```