Carolina Torres Bichara

Mensurando a influência de fatores socioeconômicos na taxa de incidência de Tuberculose entre os municípios do Rio de Janeiro

Niterói - RJ, Brasil

15 de dezembro de 2022

Universidade Federal Fluminense

Carolina Torres Bichara

Mensurando a influência de fatores socioeconômicos na taxa de incidência de Tuberculose entre os municípios do Rio de Janeiro

Trabalho de Conclusão de Curso

Monografia apresentada para obtenção do grau de Bacharel em Estatística pela Universidade Federal Fluminense.

Orientador(a): Profa. Dra. Ana Maria Lima de Farias

Niterói - RJ, Brasil

15 de dezembro de 2022

Universidade Federal Fluminense

Carolina Torres Bichara

Mensurando a influência de fatores socioeconômicos na taxa de incidência de Tuberculose entre os municípios do Rio de Janeiro

Monografia de Projeto Final de Graduação sob o título *"Mensurando a influência de fatores socioeconômicos na taxa de incidência de Tuberculose entre os municípios do Rio de Janeiro"*, defendida por Carolina Torres Bichara em 15 de dezembro de 2022, na cidade de Niterói, no Estado do Rio de Janeiro, pela banca examinadora constituída pelos professores:

Profa. Dra. Ana Maria Lima de Farias
Departamento de Estatística - UFF

Prof. Dr. Jony Arrais Pinto Junior
Departamento de Estatística - UFF

Profa. Dra. Jessica Quintanilha Kubrusly
Departamento de Estatística - UFF

Niterói, 15 de dezembro de 2022

Resumo

A Tuberculose é uma das mais antigas doenças infecciosas e tramissíveis da humanidade causada principalmente pelo bacilo aeróbico *Mycobacterium tuberculosis*. No Brasil são notificados aproximadamente 100 mil casos novos e ocorrem cerca de 4,5 mil mortes por ano em decorrência da Tuberculose e, em especial, o Rio de Janeiro se destaca entre os três primeiros estados com maior taxa de incidência da doença nos últimos anos. O conhecimento da dinâmica da doença é imprescindível para subsidiar estratégias de vigilância e controle da mesma, desta forma, este trabalho tem como objetivo identificar associações entre a Taxa de Incidência de Tuberculose e indicadores socioeconômicos nos municípios do estado do Rio de Janeiro através de um Modelo de Regressão Linear Múltiplo. Primeiramente foi desenvolvida uma análise exploratória, via mapas coropléticos, para entender a distribuição da variável resposta e das variáveis explicativas no espaço geográfico. Posteriormente, constatou-se através do modelo que a a cada uma unidade ou ponto percentual da densidade intradomiciliar, do percentual de ocupados com Ensino Médio completo ou da taxa de HIV existe um aumento na taxa de incidência de Tuberculose.O modelo final obteve um $R^2 = 0,6001$.

Palavras-chave: Tuberculose. Taxa de incidência. Indicadores Sociais. Regressão Linear Múltipla.

Agradecimentos

Primeiramente, agradeço a Deus pelo dom da vida.

Agradeço a minha mãe pelo amor, suporte e incentivo, incondicionais, em todos os momentos da vida.

Agradeço a minha família. A vitória sempre será nossa.

Agradeço a coordenadora e minha orientadora Ana Maria de Lima Farias pelo apoio em momentos difíceis, pela paciência e pelo aprendizado ao longo deste trabalho.

Agradeço ao professor Jony Arrais pelos ensinamentos tanto em sala de aula quanto durante a construção deste trabalho.

Agradeço ao corpo docente da Universidade Federal Fluminense.

Agradeço aos muitos amigos, em especial do corpo estudantil do Instituto de Matemática e Estatística, que me acompanharam nesse processo de aprendizado e evolução tanto na vida acadêmica quanto na vida pessoal.

Um agradecimento especial ao Lucas Ribeiro pelas horas livres dele me apoiando na escrita deste trabalho.

Agradeço a todos que em algum momento deixaram uma palavra de incentivo e um aprendizado por essa jornada.

Agradeço a Atlética do Instituto de Matemática e Estatística da Universidade Federal Fluminense pelos muitos amigos, pelas boas risadas, pelas muitas medalhas e pela construção profissional que me abriu portas no mercado de trabalho.

Sumário

Lista de Figuras

Lista de Tabelas

1 Introdução p. 11

1.1 Histórico da doença . . . . . p. 11

2 Materiais e Métodos p. 15

2.1 Área de estudo . . . . . p. 15

2.2 Banco de dados . . . . . p. 16

2.2.1 Casos confirmados de tuberculose . . . . . p. 16

2.2.2 Taxa de incidência de tuberculose . . . . . p. 16

2.2.3 Dados socioeconômicos . . . . . p. 16

2.2.3.1 Renda . . . . . p. 17

2.2.3.2 Condições da moradia . . . . . p. 17

2.2.3.3 Educação . . . . . p. 17

2.2.3.4 Desenvolvimento humano . . . . . p. 17

2.2.3.5 Bens de consumo . . . . . p. 18

2.2.3.6 Saúde . . . . . p. 18

2.3 Métodos . . . . . p. 18

2.3.1 Mapas Coropléticos . . . . . p. 18

2.3.2 Regressão Linear Múltipla . . . . . p. 19

2.3.2.1 Modelo Geral . . . . . p. 19

2.3.2.2 Forma Matricial . . . . . p. 19
2.3.2.3 Estimador para $\underline{\beta}$ . . . . . p. 21
2.3.2.4 Distribuição Amostral de $\hat{\underline{\beta}}$ . . . . . p. 21
2.3.3 Teste de Wald . . . . . p. 22
2.3.4 Seleção de variáveis . . . . . p. 23
2.3.4.1 Critério de Informação de Akaike . . . . . p. 24
2.3.5 Análise de resíduos . . . . . p. 24
2.3.5.1 Linearidade . . . . . p. 25
2.3.5.2 Normalidade . . . . . p. 26
2.3.5.3 Homocedasticidade . . . . . p. 26
2.3.5.4 Multicolinearidade . . . . . p. 27
2.3.5.5 Coeficiente de Determinação . . . . . p. 27
2.3.5.6 Coeficiente de Determinação Ajustado . . . . . p. 27

3 Análise dos Resultados p. 29
3.1 Análise Exploratória . . . . . p. 29
3.2 Modelos de Regressão Linear . . . . . p. 34
3.2.1 Modelo 1 . . . . . p. 34
3.2.1.1 Análise de Resíduos - Modelo 1 . . . . . p. 35
3.2.2 Modelo 2 . . . . . p. 38
3.2.3 Modelo 3 . . . . . p. 39
3.2.4 Análise de Resíduos . . . . . p. 40

4 Conclusão p. 44

Referências p. 46

Apêndice 1 – Gráficos de Dispersão p. 47

Lista de Figuras

1 Mapa do Estado do Rio de Janeiro. . . . . p. 15

2 Padrão esperado quando há linearidade nos resíduos. . . . . p. 25

3 Padrão esperado quando não há linearidade nos resíduos. . . . . p. 25

4 Mapa da Taxa de Incidência da Tuberculose por 100 mil habitantes nos municípios do Rio de Janeiro no ano de 2019. . . . . p. 29

5 Mapa da Taxa de Incidência da Tuberculose por região do Rio de Janeiro no ano de 2019. . . . . p. 30

6 Gráfico Boxplot por região. . . . . p. 31

7 Gráfico de dispersão da taxa de incidência de Tuberculose por densidade intradomiciliar. . . . . p. 31

8 Gráfico de dispersão da taxa de incidência de Tuberculose por percentual de ocupados com ensino médio completo. . . . . p. 32

9 Gráfico de dispersão da taxa de incidência de Tuberculose por percentual de domicílios com automóvel de uso particular. . . . . p. 32

10 Gráfico de dispersão da taxa de incidência de Tuberculose por taxa de incidência de HIV. . . . . p. 33

11 Gráfico de correlação . . . . . p. 34

12 Relação entre os resíduos, a variável resposta e o valor esperado . . . . p. 36

13 Gráfico QQ-Plot . . . . . p. 37

14 Suposição de homocedasticidade . . . . . p. 37

15 Gráfico da Distância de Cook . . . . . p. 38

16 Gráfico de Box-Cox . . . . . p. 39

17 Relação entre os resíduos, a variável resposta e o valor esperado . . . . p. 41

18 Suposição de homocedasticidade . . . . . p. 42

19 Gráfico QQ-Plot . . . . . p. 42

20 Gráfico da Distância de Cook . . . . . p. 43

21 Gráficos de dispersãos entre a taxa de incidência de Tuberculose e percentual de chefes de domicílio com renda até um salário mínimo e rendimento mensal domiciliar per capita. . . . . p. 47

22 Gráficos de dispersãos entre a taxa de incidência de Tuberculose e percentual de domicílios conectados a abastecimento de água e percentual de pessoas em domicílios urbanos com coleta de lixo. . . . . p. 48

23 Gráficos de dispersãos entre a taxa de incidência de Tuberculose e a taxa de analfabetismo entre pessoas de 15 anos ou mais, o IDH-M e a taxa de envelhecimento da população. . . . . p. 49

Lista de Tabelas

1 Estudos sobre a Tuberculose e indicadores socioeconômicos entre os anos de 2000 e 2020. . . . . p. 13

2 λ e suas respectivas transformações . . . . . p. 26

3 Estimativas dos parâmetros do Modelo 1 . . . . . p. 35

4 Estimativas dos parâmetros do Modelo 3 . . . . . p. 40

5 Estimativas dos parâmetros do Modelo 1 . . . . . p. 50

6 Estimativas dos parâmetros do Modelo 2 . . . . . p. 51

1 Introdução

A Tuberculose é uma das mais antigas doenças infecciosas e tramissíveis da humanidade causada principalmente pelo bacilo aeróbico *Mycobacterium tuberculosis*, podendo ser causada também, embora mais raramente, por outras espécies de agentes como a Mycobacterium bovis, M. africanum e M. microti FIOCRUZ (2022).

Ainda de acordo com a FIOCRUZ (2022), a doença afeta principalmente os pulmões mas pode atingir outros órgãos do corpo como rins, meninges e ossos. A mesma tem como principais sintomas emagrecimento acentuado, tosse com ou sem secreção por mais de três semanas, febre baixa geralmente à tarde, sudorese noturna, cansaço excessivo, falta de apetite, palidez e rouquidão.

A transmissão da tuberculose acontece por via respiratória, pela eliminação de aerossóis produzidos pela tosse, fala ou espirro de uma pessoa com tuberculose ativa, sem tratamento; e a inalação de aerossóis por um indivíduo suscetível. O tratamento à base de antibióticos tem duração de seis meses a um ano e está disponível gratuitamnete no Sistema Único de Saúde no Brasil.

1.1 Histórico da doença

A doença que acomete a humanidade desde a antiguidade, continua sendo um dos principais problemas de saúde e uma das principais causas de morte no mundo.

Segundo a WHO (2022), 10 milhões de pessoas são infectadas por Tuberculose ao ano e 1,5 milhões morrem da doença ao ano no mundo. Estima-se que um quarto da população esteja infectada, apesar da maioria das pessoas não desenvolver a doença.

O relatório global da WHO (2022), do ano de 2022, indica que o Brasil está entre os 30 países com as maiores taxas de incidência da doença. O país registra aproximadamente 100 mil casos novos e cerca de 4,5 mil mortes por ano em decorrência da enfermidade.

No país, dois estados destacam-se com maior coeficiente de incidência, são eles: Amazonas com mais de 70 casos/100 mil habitantes e Rio de Janeiro com mais de 65 casos/100 mil habitantes, assim como suas capitais que também apresentaram os maiores coeficientes de seus respectivos estados, segundo dados do banco nacional do SINAN do MS (2019).

Ainda de acordo com o boletim epidemiológico do MS (2019), a taxa de incidência no Rio de Janeiro é aproximadamente duas vezes superior a do Brasil e vem aumentando desde 2015. Enquando no Brasil houve um aumento de 2 casos/100 mil habitantes, no Rio de Janeiro o aumento foi de 6 casos/100 mil habitantes no período de 2013 a 2018.

No Brasil, Vicentin (2002) analisou a participação governamental no combate à doença e mostrou que as diferentes orientações da estratégia política e econômica condicionaram o papel do Estado como sendo o principal determinante no processo de controle da tuberculose.

A busca da relação entre componentes da vida social e a morbi-mortalidade é um caminho frequentemente trilhado pelos estudos epidemiológicos. Embora ainda haja dúvidas sobre o modo como se dá a relação entre o social e o biológico, cresce a necessidade de identificar indicadores específicos de condições de vida que se associem ao desenvolvimento de determinadas doenças. Assim, é bastante apropriado o estudo da associação entre indicadores pertencentes à esfera social e o desenvolvimento de determinadas doenças, como a Tuberculose. Da mesma forma, é relevante a quantificação da força dessas associações mediante técnicas estatísticas.

Para este trabalho foi realizada uma revisão de literatura. Dentre os 7 artigos analisados, 7 utilizaram como técnica estatística Modelo de Regressão Linear, 6 possuem como área de estudo regiões dentro do Brasil e 4 deles tem como a incidência de Tuberculose como variável dependente. Na Tabela 1 a seguir estão dispostos os artigos com destaque para os indicadores que apresentaram associação com a incidência ou a mortalidade de Tuberculose.

Tabela 1: Estudos sobre a Tuberculose e indicadores socioeconômicos entre os anos de 2000 e 2020.

Autores	**Ano**	**Região**	**Variável dependente**	**Indicadores associados**
Silva et al.	2011	Nordeste, Brasil	Mortalidade	Taxa de HIV.
Paiva et al.	2019	Pará, Brasil	Incidência	Recebimento de benefício social do governo, renda, escolaridade e sexo.
Fasca et al.	2008	Rio de Janeiro, Brasil	Incidência	Proporção de pobres, índice de Gini, log da densidade de pobres e incidência média de HIV.
Vincentin et al.	2002	Rio de Janeiro, Rio de Janeiro, Brasil	Mortalidade	Índice de Robin Hood, razão de renda entre os 10% mais ricos e os 40% mais pobres e a propoção de chefes de famílias com renda média entre um e dois salários mínios.
Angelo et. al.	2008	Juiz de Fora, Minas Gerais, Brasil	Incidência	Densidade de pobres, incidência de AIDS e valor médio do imposto territorial urbano.
Cerqueira et al	2017	Lisboa e Porto, Portugal	Incidência	Taxa de HIV, taxa de atribuição de rendimento social de inserção e taxa da população estrangeira residente.
Ximenes et. al.	2009	Brasil	Incidência	Domicílios com mais de 4 moradores, percentual de pessoas alfabetizadas, percentual de pessoas empregadas e percentual de pessoas com refrigerador, máquina de lavar, computador e ar-condicionado.

O objetivo deste trabalho é verificar se existe associação entre a taxa de incidência de Tuberculose e fatores socioeconômicos nos municípios do Rio de Janeiro e comparar se os resultados encontrados corroboram com os resultados dos estudos utilizados como referência.

No Capítulo 2, de Materiais e Métodos, será apresentado o material de estudo, bem como o conceito do modelo e os métodos usados neste trabalho. Já no Capítulo 3, de Análise dos Resultados, será apresentada a análise exploratória do conjunto de dados utilizado, e logo depois, todas as análises e resultados obtidos para esses dados. Por fim, no Capítulo 4 de Conclusão, serão reforçados todos os resultados relevantes necessários para atingir o objetivo central deste trabalho.

2 Materiais e Métodos

Neste capítulo serão apresentadas as metodologias que foram implementadas para a realização das análises. Materiais e Métodos está dividido em 3 Seções, sendo a área de estudo e suas características, o banco de dados utilizado e o embasamento teórico estatístico para modelagem dos dados.

2.1 Área de estudo

A área de estudo é composta pelos municípios do Estado do Rio de Janeiro que se encontra localizado na região Sudeste do Brasil. De acordo com o Instituto Brasileiro de Geografia e Estatística, no ano de 2010, o estado compreendia uma área territorial de 43.780,20 km^2 subdividida em 92 municípios, uma população de 15.989.929 habitantes e uma densidade demográfica de 365,23 hab/km^2, assumindo respectivamente a terceira e a segunda colocação dos estados mais populosos e povoados do país.

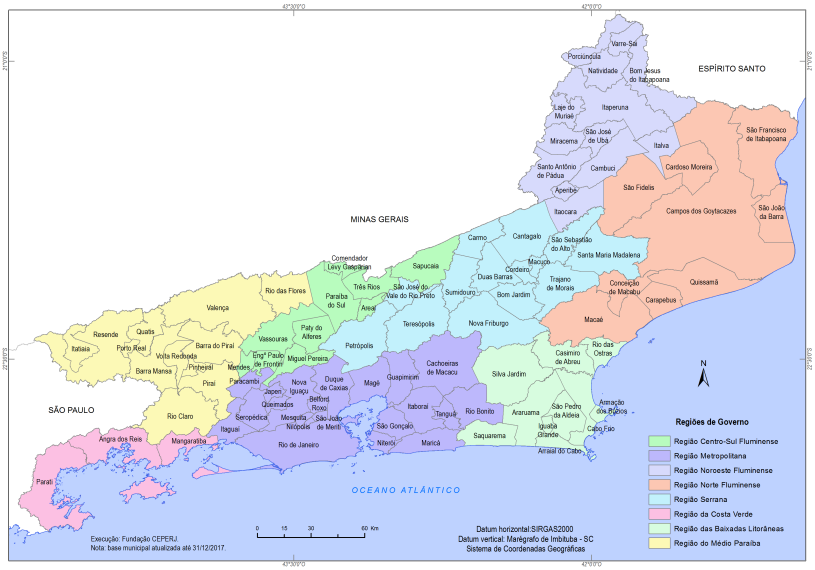

Figura 1: Mapa do Estado do Rio de Janeiro.

2.2 Banco de dados

Nesta seção serão apresentados os dados utilizados, suas devidas fontes e a construção de variáveis adequadas para este estudo. Durante todo o trabalho, toda estruturação dos dados e todas as análises foram realizadas por meio do *software* Microsoft Excel na versão 2016 e do *software* estatístico R na versão 4.1.2 e utilizando a IDE RStudio na versão 1.3.1.

2.2.1 Casos confirmados de tuberculose

Os dados sobre os casos de tuberculose foram obtidos através do Departamento de Informática do Sistema Único de Saúde (DATASUS), disponibilizados pelo Ministério da Saúde por meio do Sistema de Informação de Agravos de Notificação. Para este estudo, foram utilizadas como unidade de observação o ano do diagnóstico, o município de residência e os casos confirmados da doença quando a variável tipo de entrada era caso novo, não sabe ou pós óbito.

2.2.2 Taxa de incidência de tuberculose

A Taxa de Incidência de Tuberculose é um coeficiente de morbidade. Neste caso, a taxa de incidência de tuberculose foi calculada para cada município em estudo pela fórmula:

$$\text{Taxa de incidência do município} = \frac{\text{Número de casos de tuberculose no município}}{\text{Tamanho da população residente no município}} \times 100.000$$

O tamanho populacional de cada município foi obtido pelo Censo Demográfico de 2010 realizado pelo IBGE.

2.2.3 Dados socioeconômicos

Os dados socioeconômicos e demográficos foram obtidos a partir do Censo Demográfico de 2010. O Censo Demográfico tem por objetivo contar os habitantes do território nacional, identificar suas características e revelar como vivem os brasileiros, produzindo informações imprescindíveis para a definição de políticas públicas e a tomada de decisões de investimentos da iniciativa privada ou de qualquer nível de governo. E também cons-

tituem a única fonte de referência sobre a situação de vida da população nos municípios e em seus recortes internos, como distritos, bairros e localidades, rurais ou urbanas, cujas realidades dependem de seus resultados para serem conhecidas e terem seus dados atualizados.

Para a seleção dos indicadores socioeconômicos levou-se em consideração a importância na determinação das condições de vida e produção da tuberculose, sendo utilizado como base o estudo de revisão dos artigos científicos mencionado no Capítulo de Introdução com temas relacionados a incidência e a mortalidade por tuberculose e a disponibilidade de informação no Censo 2010. Sob este aspecto, foram construídos os seguintes indicadores:

2.2.3.1 Renda

- Renda domiciliar per capita: rendimento mensal domiciliar per capita nominal.
- Proporção de pobres: percentual de chefes de domicílio com renda até um salário mínimo.
- Densidade intradomiciliar: percentual da população que vive em domicílios com densidade superior a duas pessoas por dormitório.

2.2.3.2 Condições da moradia

- Abastecimento de água: percentual de domicílios conectados a rede geral de abastecimento de água e com canalização interna em pelo menos um cômodo.
- Coleta de lixo: percentual de pessoas em domicílios urbanos com coleta de lixo.

2.2.3.3 Educação

- Analfabetismo: taxa de analfabetismo entre pessoas de 15 anos ou mais de idade.
- Ocupados com ensino médio completo: percentual dos ocupados com ensino médio completo.

2.2.3.4 Desenvolvimento humano

- IDH-M: o índice de desenvolvimento humano é uma medida composta de indicadores de três dimensões do desenvolvimento humano: longevidade, educação e renda.

O índice apresenta variação de 0 a 1, sendo quanto mais próximo de 1, maior o desenvolvimento humano.

2.2.3.5 Bens de consumo

- Automóvel de uso particular: percentual de domicílios com automóvel de uso particular.

2.2.3.6 Saúde

- Envelhecimento: taxa de envelhecimento da população.
- HIV: taxa de incidência de HIV por município por 100 mil habitantes.

2.3 Métodos

Nesta seção serão apresentados as ferramentas preliminares usadas na análise exploratória dos dados e os principais conceitos envolvidos na utilização de Modelos de Regressão Linear Múltiplo.

2.3.1 Mapas Coropléticos

Um mapa coroplético é uma técnica com o objetivo de retratar uma superfície estatística por meio de uma área gráfica. Cada unidade de interesse se destaca de forma a representar diferentes magnitudes de um determinado atributo, ou seja, proporcionalmente ao nível da medida da variável estatística que está sendo retradada no mapa. Os mapas coropléticos são elaborados com dados quantitativos e qualitativos. Para os dados quantitativos a legenda é dividida em classes conforme os métodos de classificação utilizados, intervalos iguais, quantis, médias, quebras-máximas, quebras-naturais, fisher-jenks e etc. e regras próprias de utilização da variável visual.

As variáveis visuais mais utilizadas em mapas coropléticos são os valores de cor, variando sua intensidade conforme a seqüência de valores apresentados nas classes estabelecidas.

2.3.2 Regressão Linear Múltipla

2.3.2.1 Modelo Geral

O Modelo de Regressão Linear Múltiplo é uma ferramenta que define uma relação estatística linear entre uma variável resposta Y e $p-1$ variáveis independentes: $x_1, x_2, \ldots, x_{p-1}$. A suposição fundamental desse modelo é que a média da distribuição da variável resposta varia de forma linear com as variáveis $x_1, x_2, \ldots, x_{p-1}$. A equação do modelo geral é dada por:

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \ldots + \beta_p X_{p-1,i} + \varepsilon_i, \tag{2.1}$$

- Y_i é o valor da variável resposta na i-ésima observação;
- $X_{i,j}$ é o valor da j-ésima variável independente na i-ésima observação;
- ε_i é o erro aleatório para a i-ésima observação e $\varepsilon_i \sim N(0, \sigma^2)$;
- β's são os parâmetros do modelo. Onde :

$$E(Y) = \beta_0 + \sum_{i=1}^{p-1} \beta_i X_i$$

O Modelo de Regressão Linear Múltiplo, assim como o Simples, requer algumas pressuposições. São elas:

- Linearidade do modelo;
- A distribuição condicional de Y dado X é normal;
- Homocedasticidade da variável resposta Y dado o conjunto de variáveis independentes X.

2.3.2.2 Forma Matricial

O desenvolvimento da teoria de regressão linear múltipla é facilitado ao considerar sua forma matricial. De forma geral temos:

$$Y_i = \beta_0+\beta_1 X_{i,1}+\ldots+\beta_p X_{i,p-1}+\varepsilon_i \;\Rightarrow\; \begin{cases} Y_1 = \beta_0 + \beta_1 X_{1,1} + \ldots + \beta_p X_{1,p-1} + \varepsilon_1 \\ Y_2 = \beta_0 + \beta_1 X_{2,1} + \ldots + \beta_p X_{2,p-1} + \varepsilon_2 \\ \vdots \\ Y_n = \beta_0 + \beta_1 X_{n,1} + \ldots + \beta_p X_{n,p-1} + \varepsilon_n \end{cases} \tag{2.2}$$

Abrindo em forma matricial os elementos dessa igualdade em vetores e matrizes:

$$\underline{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, X = \begin{bmatrix} 1 & X_{1,1} & X_{1,2} & \ldots & X_{1,p-1} \\ 1 & X_{2,1} & X_{2,2} & \ldots & X_{2,p-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n,1} & X_{n,2} & \ldots & X_{n,p-1} \end{bmatrix}, \underline{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix}, \underline{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \tag{2.3}$$

Podemos escrever:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{1,1} & X_{1,2} & \ldots & X_{1,p-1} \\ 1 & X_{2,1} & X_{2,2} & \ldots & X_{2,p-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n,1} & X_{n,2} & \ldots & X_{n,p-1} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \tag{2.4}$$

Ou seja, o modelo geral também pode ser definido como:

$$\underline{Y} = \mathbf{X}\underline{\beta} + \underline{\varepsilon}, \qquad \underline{\varepsilon} \sim N_n(\underline{\mu} = 0, \Sigma = \sigma^2 I) \tag{2.5}$$

Onde:

- $\underline{Y}$ é o vetor aleatório da variável resposta Y com dimensão n;
- $\mathbf{X}$ é a matriz das variáveis respostas X's, com dimensão $n \times p$;
- $\underline{\beta}$ é o vetor dos parâmetros $\beta's$, com dimensão p;
- $\underline{\varepsilon}$ é o vetor do erro aleatório, com dimensão n;
- n é o tamanho da amostra;

2.3.2.3 Estimador para $\underline{\beta}$

O estimador para $\underline{\beta}$ por mínimos quadrados é aquele que minimiza a soma dos quadrados dos erros ε_i. A soma dos quadrados dos erros pode ser definida por:

$$\begin{aligned} SQE = \sum_{i=1}^{n} \varepsilon_i^2 = \underline{\varepsilon}^T \underline{\varepsilon} &= (\underline{Y} - X\underline{\beta})^T (\underline{Y} - X\underline{\beta}) \\ &= (\underline{Y}^T - X^T \underline{\beta}^T)(\underline{Y} - X\underline{\beta}) \\ &= \underline{Y}^T \underline{Y} - \underline{Y}^T X\underline{\beta} - \underline{\beta}^T X^T \underline{Y} + \underline{\beta}^T X^T X\underline{\beta} \\ &= \underline{Y}^T \underline{Y} - 2\underline{\beta}^T X^T \underline{Y} + \underline{\beta}^T X^T X\underline{\beta} \end{aligned} \tag{2.6}$$

Para encontrar o estimador de $\underline{\beta}$ pelo método de mínimos quadrados é necessário derivar SQE em relação a β e igualar a zero.

$$\frac{\partial SQE}{\partial \underline{\beta}} = -2X^T \underline{Y} + 2X^T X\underline{\beta} = 0 \tag{2.7}$$

Assim, temos:

$$\hat{\underline{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \underline{Y}, \tag{2.8}$$

em que a matriz $(\mathbf{X}^T\mathbf{X})^{-1}$ existe caso nenhuma das colunas de $\mathbf{X}$ seja combinação linear das demais.

2.3.2.4 Distribuição Amostral de $\hat{\underline{\beta}}$

O estimador $\hat{\underline{\beta}}$ tem as seguintes propriedades:

- $\hat{\underline{\beta}}$ minimiza a soma de quadrados dos resíduos independente de qualquer propriedade quanto à distribuição dos erros;
- Os elementos de $\hat{\underline{\beta}}$, que são combinações lineares das observações $Y_1, ..., Y_n$, são estimadores não viesados dos elementos de $\underline{\beta}$ e possuem a menor variância dentre todos os estimadores não viesados que são combinação linear dos $Y_i's$, independentes das propriedades da distribuição dos erros;
- Sob a suposição de que os erros são independentes e normalmente distribuídos com média zero e variância constante σ^2, $\hat{\underline{\beta}}$ também é o estimador de máxima verossimilhança para $\underline{\beta}$.

Sendo assim, temos que:

$$E(\underline{\hat{\beta}}) = \underline{\beta} \tag{2.9}$$

e a matriz de covariância de $\underline{\hat{\beta}}$ é dada por:

$$\begin{aligned}
\Sigma_{\underline{\hat{\beta}}} = V(\underline{\hat{\beta}}) &= V((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\underline{Y}) \\
&= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T V(\underline{Y})[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T]^T \\
&= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T V(\underline{Y})\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \\
&= (\mathbf{X}^T\mathbf{X})^{-1}(\sigma^2 I)\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \\
&= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \\
&= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}
\end{aligned} \tag{2.10}$$

Portanto,

$$\underline{\hat{\beta}} \sim N_p(\underline{\beta}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}). \tag{2.11}$$

2.3.3 Teste de Wald

O teste de Wald permite avaliar se cada parâmetro individualmente tem efeito estatisticamente significativo no modelo, a um nível de significância α.

As hipóteses do teste são:

$$\begin{cases} H_0: & \beta_j = 0 \\ H_1: & \beta_j \neq 0 \end{cases}$$

onde, β_j é o efeito da j-ésima variável preditora.

Estatística de teste é definida por:

$$Z = \frac{\hat{\beta}_j}{\sqrt{V\hat{A}R(\hat{\beta}_j)}}$$

Sob H_0, $Z \sim N(0,1)$.

Região Crítica (RC): $RC = \left\{z \in \mathbb{R} | |z| \leq z_{\frac{\alpha}{2}}\right\}$

Após encontrar o valor observado na estatística de teste, é importante saber que decisão tomar. Se $z_{obs} \in RC$ rejeita-se H_0, a um nível de significância α, ou seja, há evidências de que existe uma associação estatisticamente significativa X_j e a chance em favor do evento de interesse. Caso contrário, não rejeita-se H_0 a um nível α, e assim afirmar que há evidências de que não existe uma associação estatisticamente significativa X_j e a chance em favor do evento de interesse.

2.3.4 Seleção de variáveis

Segundo Kutner et al. (2004), a identificação de bons subconjuntos de variáveis preditoras potencialmente úteis para serem incluídas no modelo de regressão final constitui um dos problemas mais difíceis na Análise de Regressão.

Em contrapartida, se todas as variáveis preditoras estão incluídas no modelo linear múltiplo final, o mesmo poderá ficar super-ajustado aos dados. Desse modo, este modelo dificilmente produzirá boas previsões e a variância dos estimadores dos parâmetros será maior do que a de modelos mais simples. Nesse sentido, é fundamental utilizar uma medida e/ou um algoritmo para seleção de variáveis que possuam um bom desempenho e realizem essa tarefa em um tempo relativamente baixo em comparação com a seleção de um modelo ajustando-se todos os possíveis.

Neste trabalho, será utilizado o método de seleção *Stepwise*, que consiste em adicionar e remover iterativamente covariáveis do modelo. Este método dispõe das três abordagens detalhadas abaixo:

- *Forward stepwise*
  - O modelo de regressão inicia apenas com o intercepto;
  - As demais variáveis candidatas são incluídas uma a uma;
  - Se a nova adição foi estatisticamnete significante, mantenha a variável; caso contrário, retire a variável, volte ao modelo anterior e pare o algoritmo;
  - Os passos 2 e 3 se repetem até que a adição de qualquer nova variável não seja estatisticamente significante, isto é, a adição de uma nova variável não reduz o AIC do modelo.
- *Backward stepwise*
  - O modelo de regressão inicia com todas as variáveis candidatas;

  - Se houver alguma variável cujo coeficiente é estatisticamente não significativo, elimine a variável que tenha menor nível de significância no modelo (maior p-valor); caso contrário, esse é o modelo final;
  - O passos 2 se repete até atingir um modelo no qual todas as variáveis são estatisticamente significantes, isto é, quando a retirada de uma variável explicativa não gera uma redução no AIC.

- *Forward-backward stepwise*
  - Trata-se de uma combinação das seleções do tipo *forward* e *backwards*;
  - Os passos *forward* e *backwards* são intercalados, de forma a adicionarmos variáveis que sejam significativas e retirarmos variáveis que não sejam estatisticamente significativas;
  - O algoritmo para quando não for mais possível adicionar variáveis novas que sejam estatisticamente significantes, ou retirar variáveis incluídas que forem estatisticamente não significantes.

2.3.4.1 Critério de Informação de Akaike

Uma medida utilizada na comparação de modelos é o Critério de Informação de Akaike (AIC), definido por:

$$AIC = 2p - 2log(\hat{L}),$$

sendo p o número de parâmetros estimados e L a função de máxima verossimilhança.

Quanto menor a soma dos resíduos ao quadrados e quanto menor a quantidade de variáveis preditivas p, menor o valor do AIC. Como busca-se um modelo com a menor quantidade de variáveis preditoras possíveis e com valor baixo para a soma dos resíduos ao quadrados, opta-se pelo modelo com menor AIC.

2.3.5 Análise de resíduos

A Análise de Resíduos é uma técnica que auxilia na avaliação da adequabilidade do modelo de regressão. Os pressupostos de linearidade, normalidade, homocedaticidade e multicolinearidade são validados para que os resultados encontrados sejam confiáveis.

2.3.5.1 Linearidade

O diagnóstico de linearidade na Regressão Linear Múltipla será feito em relação a cada variável preditiva X_k. Para verificar a suposição de que a média de Y varia de forma linear com X_k plota-se em um gráfico os valores de $X_{i,k}$ versus os resíduos padronizados ($e_i{}^*$). O padrão esperado é o apresentado na Figura 2. Se o gráfico não tiver o padrão esperado, isto é, ele se parece com a Figura 3, será diagnosticada a não-linearidade de $E[Y]$ com a variável X_k.

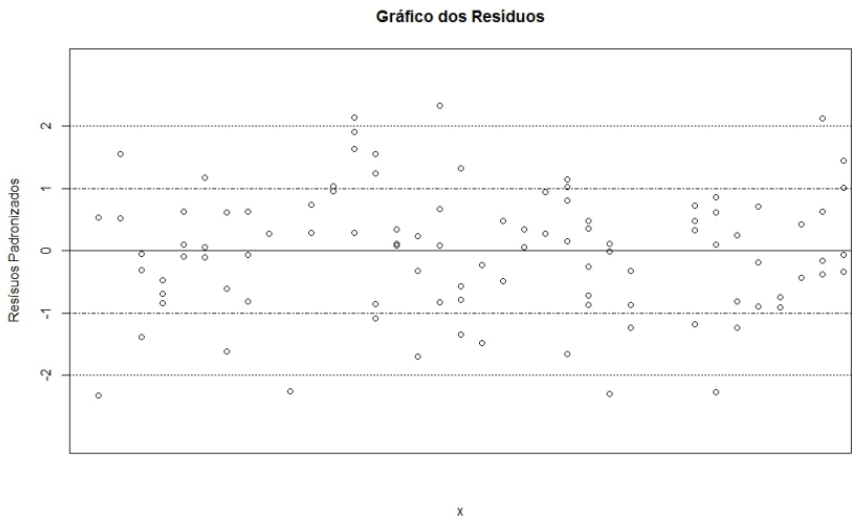

Figura 2: Padrão esperado quando há linearidade nos resíduos.

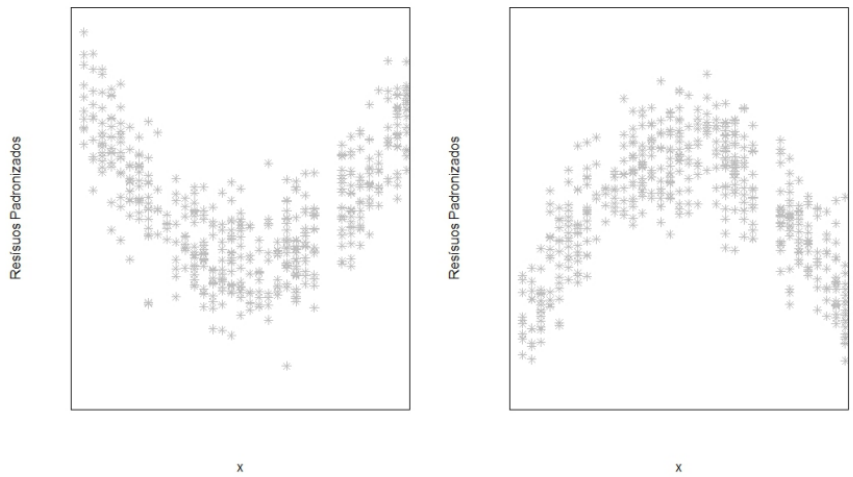

Figura 3: Padrão esperado quando não há linearidade nos resíduos.

2.3.5.2 Normalidade

Para detectar normalidade dos erros vamos verificar se os resíduos são aparentemente normais. Para verificar se a amostra de resíduos segue uma distribuição normal construímos um gráfico QQplot dos resíduos padronizados, que plota os quantis amostrais versus os quantis teóricos. O padrão esperado para que se valide normalidade é que os pontos estejam próximos a reta identidade.

Para auxiliar no diagnóstico, podemos checar se a amostra de resíduos padronizados segue distribuição Normal através do teste de Shapiro-Wilk.

- Transformação Box-Cox

  Esse método produz uma transformação da variável resposta Y tal que as pressuposições de homocedasticidade e normalidade dos resíduos sejam simultaneamente satisfeitas (COUTO et al., 2009). A família de transformações Box-Cox (BOX; COX, 1964) é amplamente utilizada, pois permite identificar a melhor transformação, com base na utilização de um valor λ que maximize o estimador de máxima verossimilhança e minimize o resíduo (CHUNG; PEARN; YANG, 2007). Box & Cox definiram a seguinte família de transformações:

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

  A tabela a seguir apresenta alguns valores usuais de λ e suas respectivas transformações.

Tabela 2: λ e suas respectivas transformações

λ	**Transformação**
-1	$\frac{1}{Y}$
-0,5	$\frac{1}{\sqrt{Y}}$
0	$\ln Y$
0,5	$\sqrt{Y}$
2	Y^2

2.3.5.3 Homocedasticidade

A homocedasticidade dos dados pode ser verificada através do gráfico dos valores ajustados *versus* os resíduos estudentizados. Para atender a hipótese de homocedasticidade

espera-se visualizar pontos distribuídos de forma aleatória, porém caso o gráfico apresente padrões nos pontos, temos que a hipótese de homocedasticidade não é satisfeita.

2.3.5.4 Multicolinearidade

A multicolinearidade é um problema observado constantemente ao se estimar um modelo de regressão. Para o diagnóstico da multicolinearidade, o Fator da Inflação da Variância (VIF) é uma medida estatística que pode ser utilizada. Essa medida é da forma:

$$VIF_j = \frac{1}{1 - R_j^2}, \qquad j = 1, 2, ..., p - 1. \tag{2.12}$$

onde R_j^2 é o coeficiente de determinação do modelo de regressão j.

Com isso, quanto mais próximos de 1 os valores de VIF_j forem, menor a preocupação com a multicolinearidade e seus efeitos. Além disso, valores de $VIF_j > 10$ são considerados indicadores de sérios problemas de multicolinearidade e dessa forma, as variáveis que apresentarem isto devem ser retiradas do modelo.

2.3.5.5 Coeficiente de Determinação

Este coeficiente denotado por R^2 serve para avaliar o quão bem o modelo ajustado utilizado representa os dados observados, ou seja, este coeficiente avalia como a proporção da variação total da variável resposta Y_b é explicado pelo modelo ajustado. É dado por:

$$R^2 = \frac{SQR_{eg}}{SQT}$$

onde SQR_{eg} é a soma dos quadrados da regressão e SQT é a soma dos quadrados totais.

Este coeficiente varia de entre 0 e 1, de forma a ser representado em porcentagem, ou seja, $0\% \leq R^2 \leq 100\%$. Então quanto mais próximo de 1 (100%) estiver o R^2, melhor será o ajuste do modelo.

2.3.5.6 Coeficiente de Determinação Ajustado

Embora o R^2 aumente ao adicionar uma nova covariável ao modelo, isso não significa que este novo modelo é melhor que o anterior. Sendo assim, para corrigir isto, é definido

o **coeficiente de determinação ajustado**, denotado por R_a^2, dado por:

$$R_a^2 = 1 - \left(\frac{n-1}{n-p}\right)(1 - R^2)$$

em que $R_a^2 < R^2$, menos quando $R^2 = 1$. Além disso, este coeficiente também varia de 0 a 1 e é representado em porcentagem da forma $0\% \leq R_a^2 \leq 100\%$.

3 Análise dos Resultados

Neste capítulo serão apresentados uma análise exploratória e em seguida os resultados dos modelos realizados buscando a possível relação entre a taxa de incidência de tuberculose com as características sociodemograficas descritas no Capítulo de Materiais e Métodos. Todas as análises e gráficos gerados neste trabalho foram realizados através do *software* estatístico R na versão 4.1.2 e utilizando a IDE RStudio na versão 1.3.1.

3.1 Análise Exploratória

Inicialmente foi feita uma análise exploratória para melhor compreensão dos dados. Uma técnica amplamente utilizada em estatística espacial para descrever dados aglomerados por área são os mapas coropléticos. Assim, a Figura 4 ilustra a distribuição da taxa de Incidência de Tuberculose entre os municípios do estado do Rio de Janeiro e nota-se que os municípios que apresentam as maiores taxas de incidência da doença são Japeri, Nova Iguaçu, Mesquita, Magé, Volta Redonda e Rio de Janeiro.

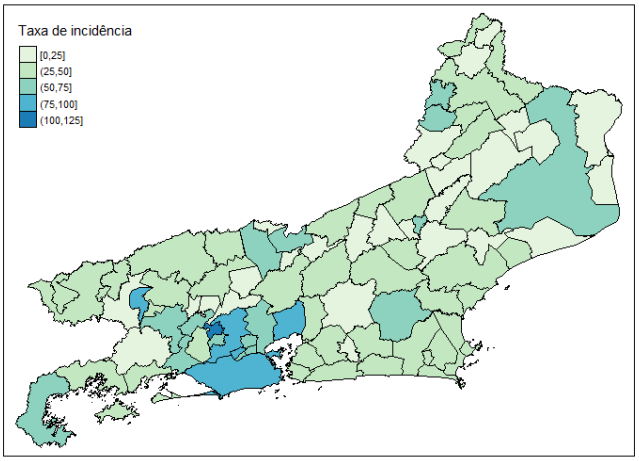

Figura 4: Mapa da Taxa de Incidência da Tuberculose por 100 mil habitantes nos municípios do Rio de Janeiro no ano de 2019.

Analisando o mapa seccionado pelas 8 regiões de governo do estado do Rio de Janeiro, conforme a Figura 5, observou-se que 5 dos 6 municípios com as maiores taxas de incidência estão localizados na região Metropolitana.

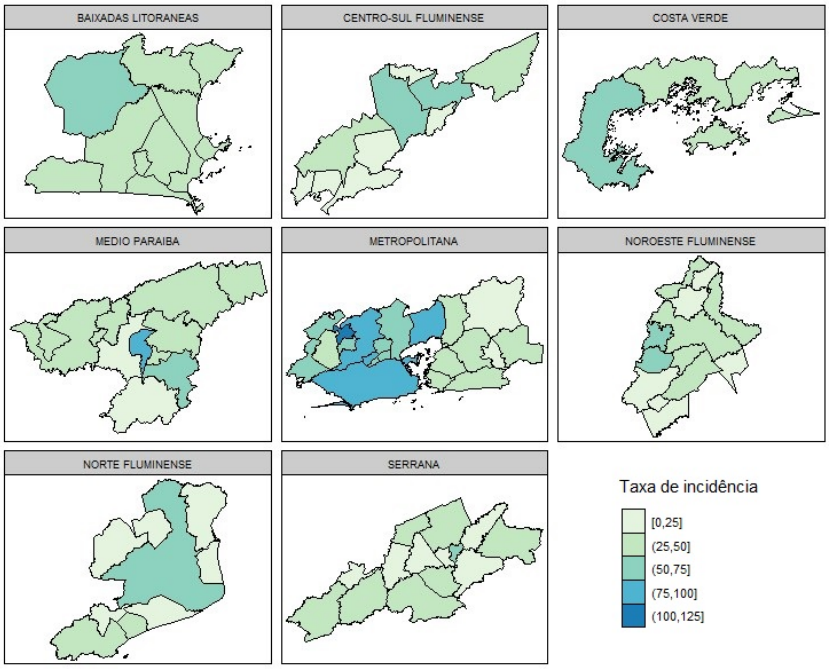

Figura 5: Mapa da Taxa de Incidência da Tuberculose por região do Rio de Janeiro no ano de 2019.

Em seguida, investigou-se a relação de cada variável explicativa com a variável resposta através de gráficos de dispersão. A seguir estão dispostas as Figuras com as variáveis que se mostraram mais relevantes em um primeiro momento na análise exploratória, as demais Figuras estão dispostas no 1.

Através de um gráfico de Boxplot das 8 regiões de governo do estado do Rio de Janeiro, foi possível verificar evidências de comportamento distinto da taxa de incidência de Tuberculose entre as regiões, com destque para a região Metropolitana que apresentou mediana superior as demais.

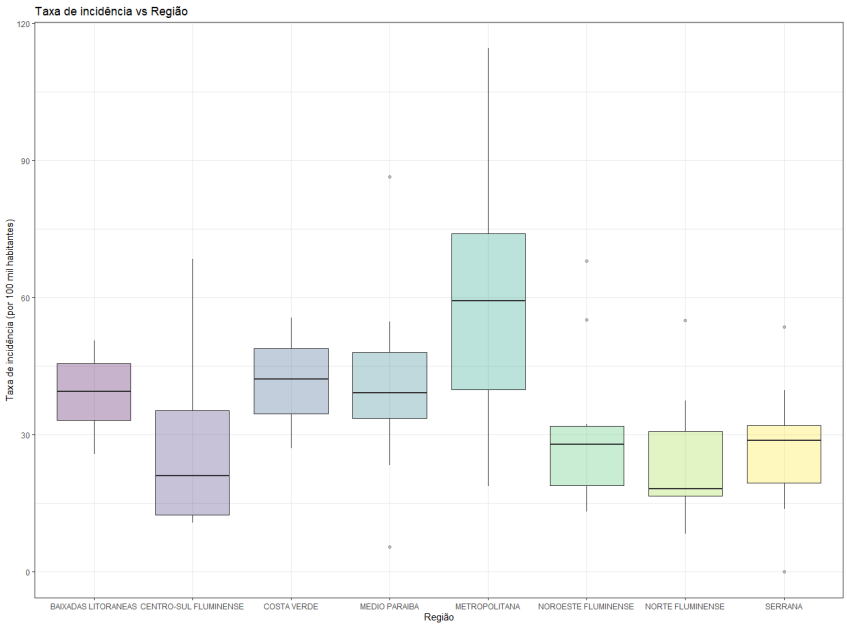

Figura 6: Gráfico Boxplot por região.

Na Figura 7 verificou-se que a densidade intradomiciliar aparenta variar linearmente com a taxa de incidência de Tuberculose de forma que quanto maior a densidade intradomiciliar maior a taxa de incidência de Tuberculose.

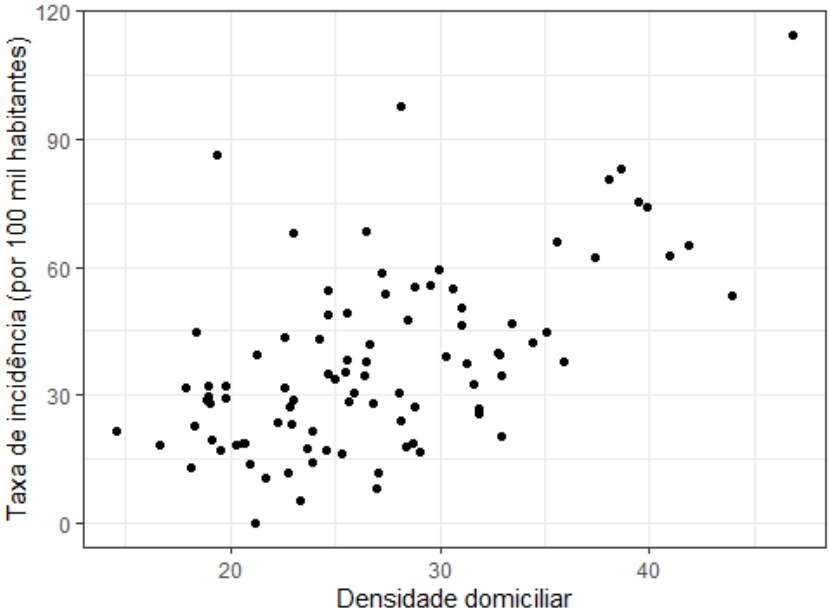

Figura 7: Gráfico de dispersão da taxa de incidência de Tuberculose por densidade intradomiciliar.

Já na Figura 8, o percentual de ocupados com ensino médio completo evidencia uma leve tendência linear crescente com a taxa de incidência da doença.

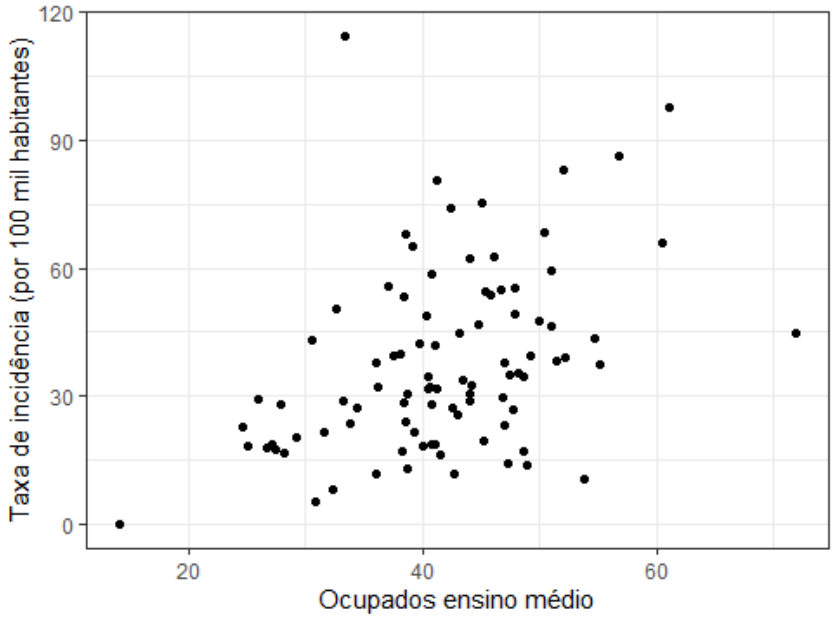

Figura 8: Gráfico de dispersão da taxa de incidência de Tuberculose por percentual de ocupados com ensino médio completo.

Verificou-se graficamente que quanto maior o percentual de domicílios com automóvel de uso particular menor a taxa de incidência de Tuberculose, conforme podemos observar na Figura 9.

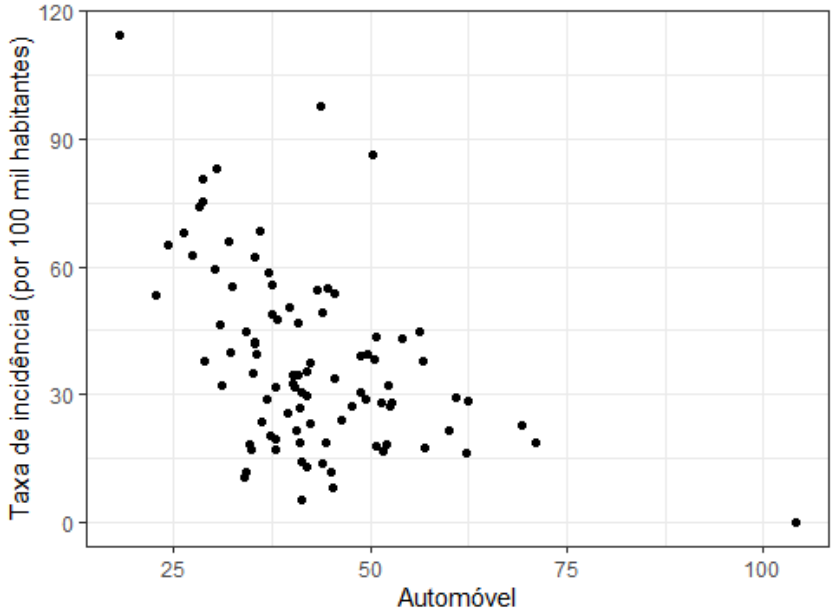

Figura 9: Gráfico de dispersão da taxa de incidência de Tuberculose por percentual de domicílios com automóvel de uso particular.

Conforme esperado, graficamente verificamos evidências de que a taxa de HIV varia linearmente com a taxa de Tuberculose, de forma que quanto maior a taxa de HIV maior a incidência de Tuberculose.

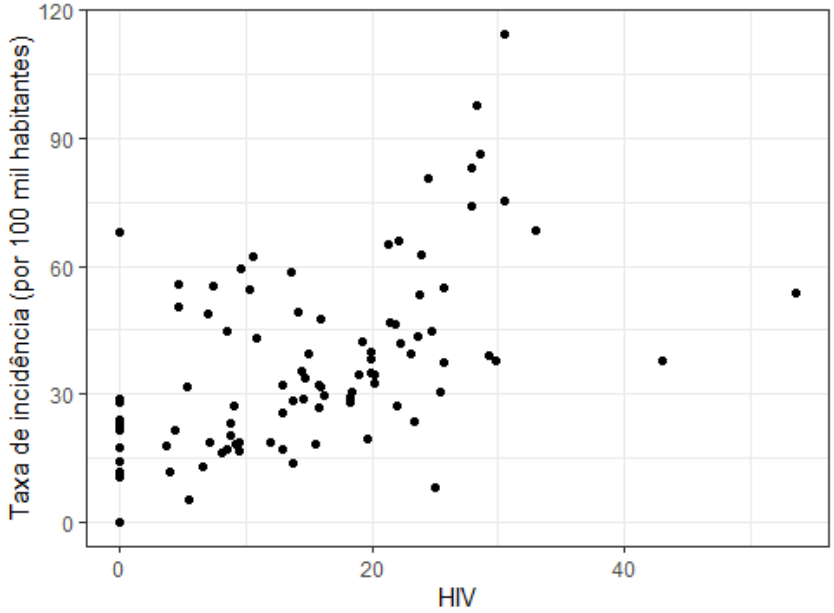

Figura 10: Gráfico de dispersão da taxa de incidência de Tuberculose por taxa de incidência de HIV.

Por fim, construiu-se uma matriz de correlação para avaliar se existiam covariáveis altamente relacionadas. Através da Figura 11 verificou-se que as variáveis renda per capita, analfabetismo e IDHM apresentavam alta correlação com outras variáveis, logo, foram retiradas da base para ajuste do primeiro modelo que será descrito na seção a seguir.

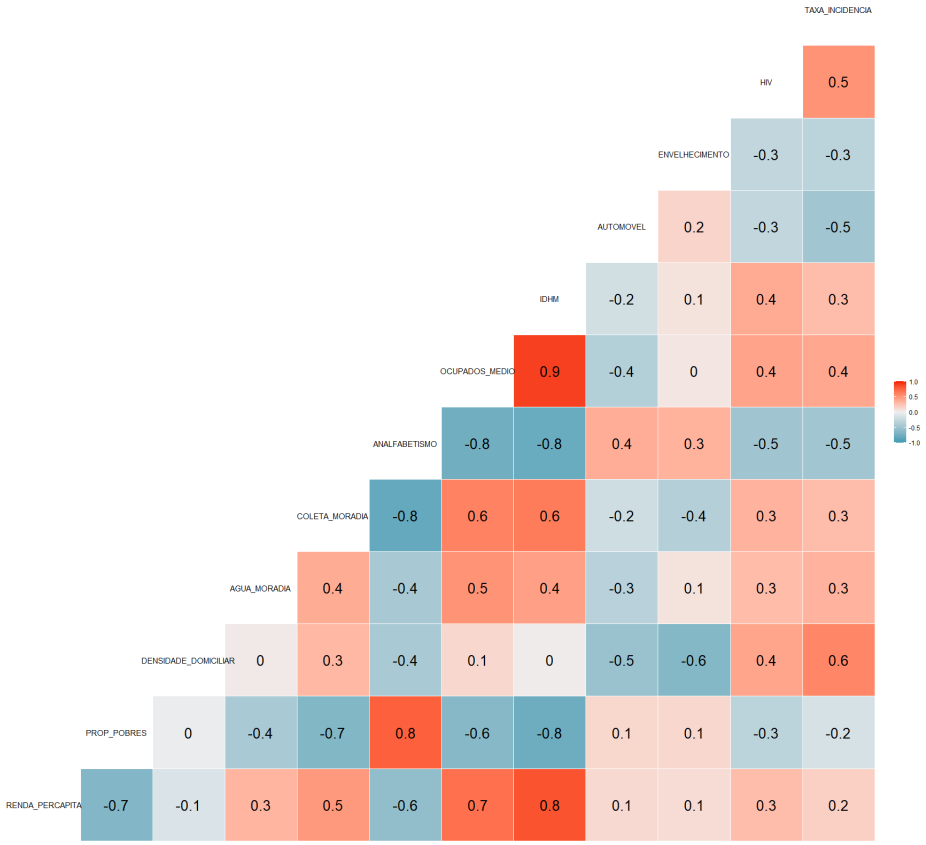

Figura 11: Gráfico de correlação

3.2 Modelos de Regressão Linear

Nesta seção serão apresentados os resultados dos Modelos de Regressão Linear Múltipla ajustados para os dados deste trabalho. Adotou-se um nível de significância de 5% para avaliação dos testes realizados neste trabalho.

3.2.1 Modelo 1

Primeiramente, realizou-se um ajuste com o modelo completo, contendo todas as variáveis explicativas descritas em 2.2, exceto renda per capita, analfabetismo e IDH-M que apresentaram alta correlação com outras covariáveis conforme verificado anteriormente.

A covariável região foi adicionada ao modelo com objetivo de avaliar se, comparativamente, existem diferenças significantes na taxa de incidência de Tuberculose quando o indivíduo pertence a regiões diferentes. No Modelo 1, a região de referência é Baixadas

Tabela 3: Estimativas dos parâmetros do Modelo 1

Parâmetros	**Estimativas**	**P-valor**
Intercepto	23.62	0.60319
Região (Centro-Sul)	-1.41	0.84760
Região (Costa Verde)	4.15	0.67665
Região (Médio Paraíba)	6.21	0.39126
Região (Metropolitana)	8.80	0.15331
Região (Noroeste)	5.13	0.56390
Região (Norte)	-10.22	0.17652
Região (Serrana)	9.09	0.23688
Proporção de pobres	1.34	0.05166
Densidade intradomiciliar	1.06	0.06270
Abastecimento de água	0.18	0.10050
Coleta de lixo	-0.24	0.52195
Ocupados c/ Ensino Médio	0.84	0.00856*
Automóvel particular	-0.14	0.47397
Envelhecimento	-1.35	0.36129
HIV	0.46	0.01981*

Litorâneas.

De acordo com a tabela 3 com as estimativas dos parâmetros do Modelo 1, somente as variáveis Ocupados c/ Ensino Médio e Taxa de HIV se mostraram siginificativas. O Modelo 1 apresenta evidências de que quanto maior a proporção de Ocupados c/ Ensino Médio ou quanto maior a taxa de HIV, maior será a taxa de incidência de Tuberculose.

3.2.1.1 Análise de Resíduos - Modelo 1

Inicialmente foi realizada a análise gráfica das relações entre a variável resposta Y e os valores esperados $\hat{Y}$ e os resíduos. Em ambos os gráficos os resíduos não apresentam comportamento padrão, se mostrando distribuídos aleatoriamente, e não foram observados outliers.

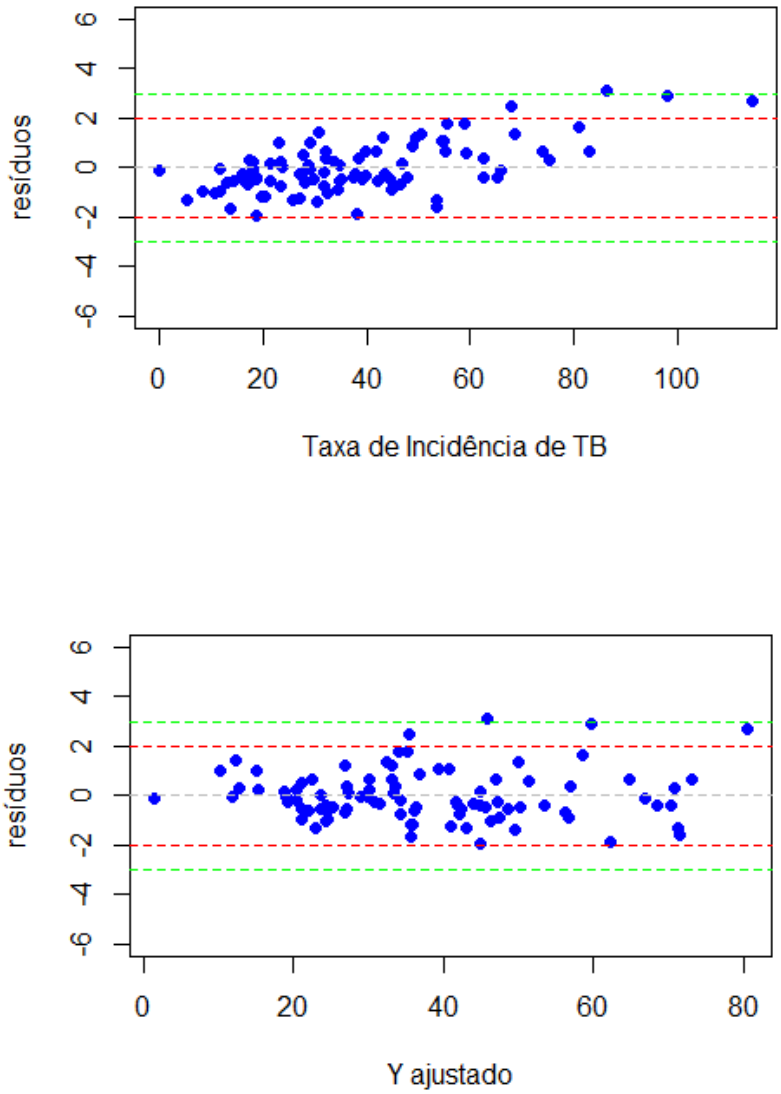

Figura 12: Relação entre os resíduos, a variável resposta e o valor esperado

O pressuposto de Normalidade não foi atendido de acordo com o resultado do teste de *Shapiro-Wilk* ($p - valor = 0,00264$). Graficamente, verifica-se através do *QQ-plot* apresentado na Figura 13 que uma cauda pesada pode estar afetando o ajuste.

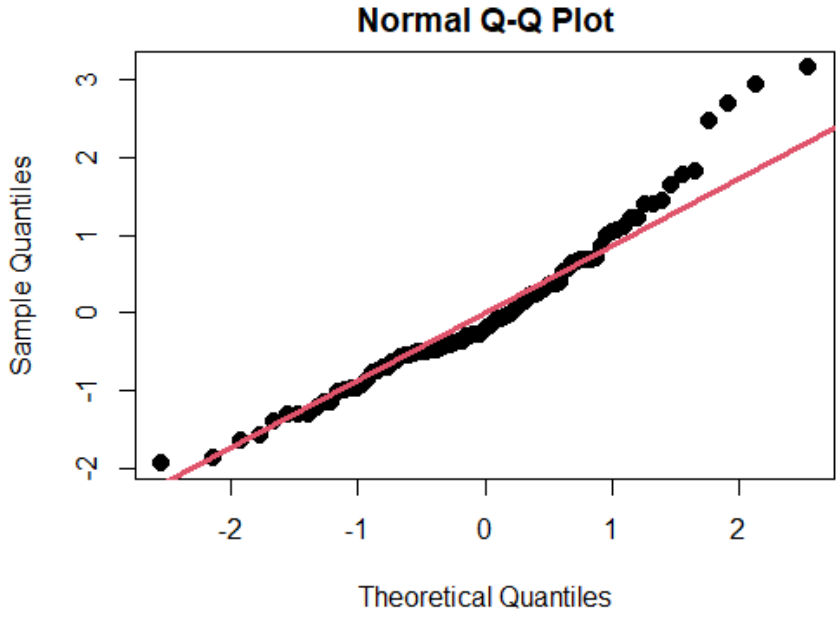

Figura 13: Gráfico QQ-Plot

A análise gráfica apresentada na Figura 14 e o teste de *Breusch-Pagan* corroboraram com evidências de homecedasticidade dos resíduos ($p - valor = 0,6027$).

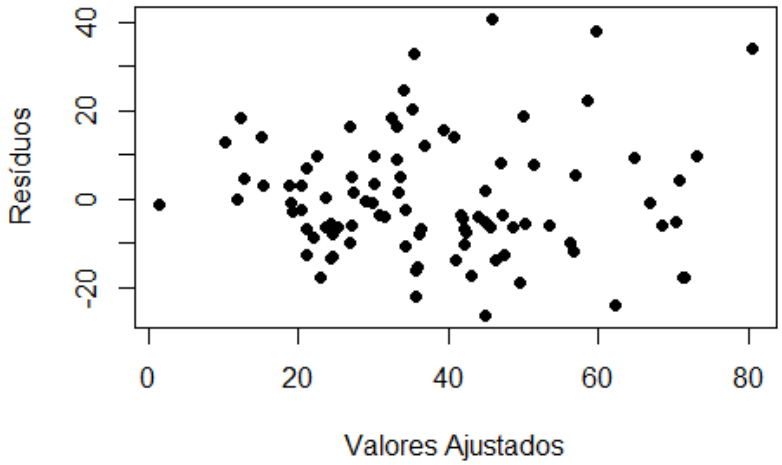

Figura 14: Suposição de homocedasticidade

Na Figura 15, que apresenta a Distância de Cook, observa-se que não foram identificados pontos influentes.

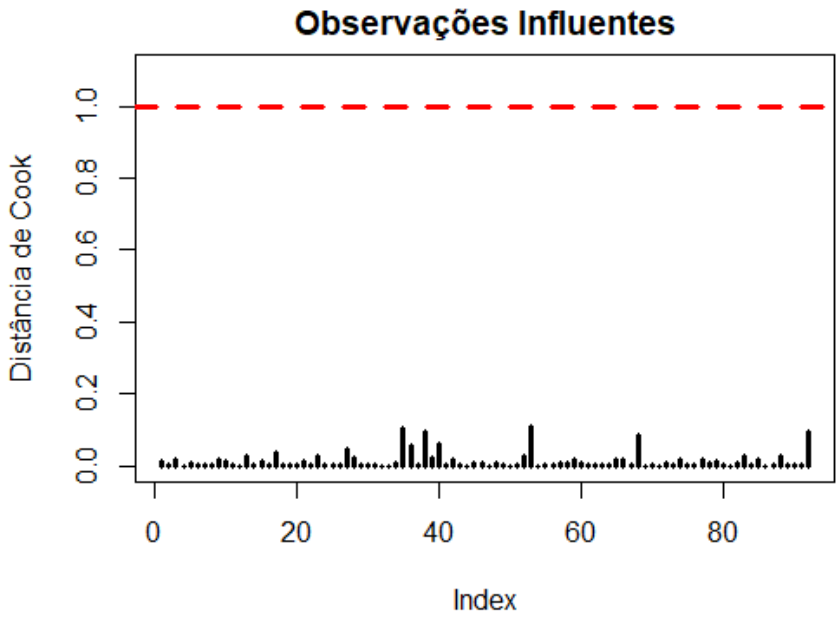

Figura 15: Gráfico da Distância de Cook

O Fator da Inflação da Variância (VIF) apresentou evidências que os dados são não correlacionados, onde nenhuma covariável apresentou VIF > 7.

Desta forma, o Modelo 1 atendeu os pressupostos com exceção da suposição de Normalidade e apresentou um R^2 onde explica 52,49% dos dados.

3.2.2 Modelo 2

No intuito de verificar se uma transformação na variável resposta solucionaria o problema de não-normalidade, foi aplicada a função *boxcox* do pacote *MASS*. A Figura 16 indicou $\lambda = 0,5$ e portanto foi aplicada a transformação da raiz quadrada na variável resposta Taxa de Incidência de Tuberculose.

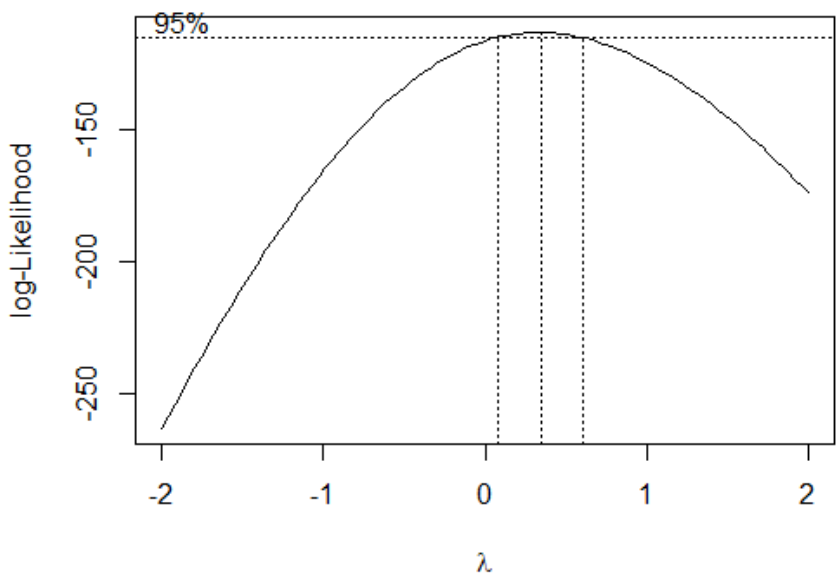

Figura 16: Gráfico de Box-Cox

O Modelo 2, que é o modelo completo com a variável resposta transformada, atendeu aos pressupostos de linearidade, normalidade e homocedastidade, entretanto possuía variáveis pouco significativas como pode ser observado no 2.

3.2.3 Modelo 3

Com o objetivo de ajustar um terceiro modelo mais parcimonioso, aplicou-se o método de seleção Stepwise, através da função *stepAIC* do pacote *MASS*. Assim, o método indicou a seleção de 7 covariáveis, das 12 originais, para compor o modelo.

O Modelo 3 é composto pelas variáveis explicativas região, proporção de pobres, densidade intradomiciliar, ocupados com Ensino Médio completo, e taxa de incidência de HIV e pela variável resposta transformada raiz quadrada da taxa de incidência de tuberculose. As estimativas dos parâmetros e os valores p seguem na Tabela 4

De acordo com a Tabela 4, verificou-se que as regiões não se mostraram significativas ao comparadas com a categoria de referência Baixadas Litorâneas. A variável proporção de pobres também não se mostrou estatisticamente significativa ao nível de significância de 5%. Apesar disso, ambas as variáveis foram mantidas neste modelo final.

Analisando alguns coeficientes do modelos, interpreta-se que:

- A estimativa da média da raiz da taxa de Incidência de Tuberculose é de 2.73 quando

Tabela 4: Estimativas dos parâmetros do Modelo 3

Parâmetros	**Estimativas**	**P-valor**
Intercepto	2.73	0.08522
Região (Centro-Sul)	-0.29	0.61014
Região (Costa Verde)	0.49	0.53595
Região (Médio Paraíba)	0.69	0.20945
Região (Metropolitana)	0.34	0.47736
Região (Noroeste)	0.84	0.19524
Região (Norte)	-0.96	0.11457
Região (Serrana)	0.57	0.341202
Proporção de pobres	0.09	0.05678
Densidade intradomiciliar	0.12	0.00062*
Ocupados c/ Ensino Médio	0.08	0.00018*
HIV	0.04	0.00491*

as demais covariáveis assumem valor zero.

- As variáveis Densidade Intradomiciliar e Ocupados c/ Ensino Médio tem uma relação ascendente com a raiz da taxa de Incidência de Tuberculose. Assim, a cada ponto percentual de aumento de cada uma dessas variáveis, aumenta-se a raiz da taxa, respectivamente, em 0,12 e 0,08 unidades.

- Por fim, o efeito esperado de HIV também é de aumento na raiz da taxa de Incidência de Tuberculose.

3.2.4 Análise de Resíduos

Inicialmente foi realizada a análise gráfica das relações entre a variável resposta $\sqrt{Y}$ e os valores esperados $\hat{Y}$ e os resíduos. Em ambos os gráficos os resíduos não apresentam comportamento padrão, se mostrando distribuídos aleatoriamente, e não foram observados outliers.

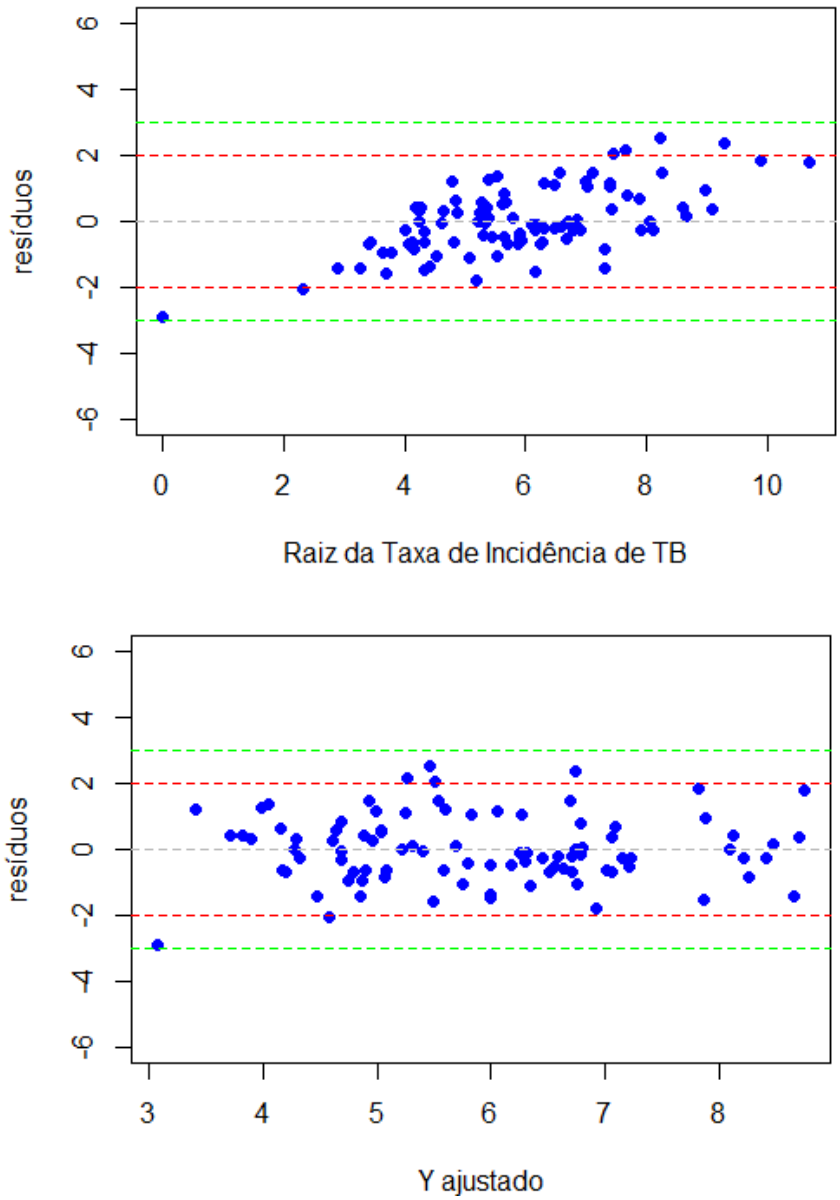

Figura 17: Relação entre os resíduos, a variável resposta e o valor esperado

A análise gráfica apresentada na Figura 18 e o teste de *Breusch-Pagan* corroboraram com evidências de homecedasticidade dos resíduos ($p - valor = 0,6970$).

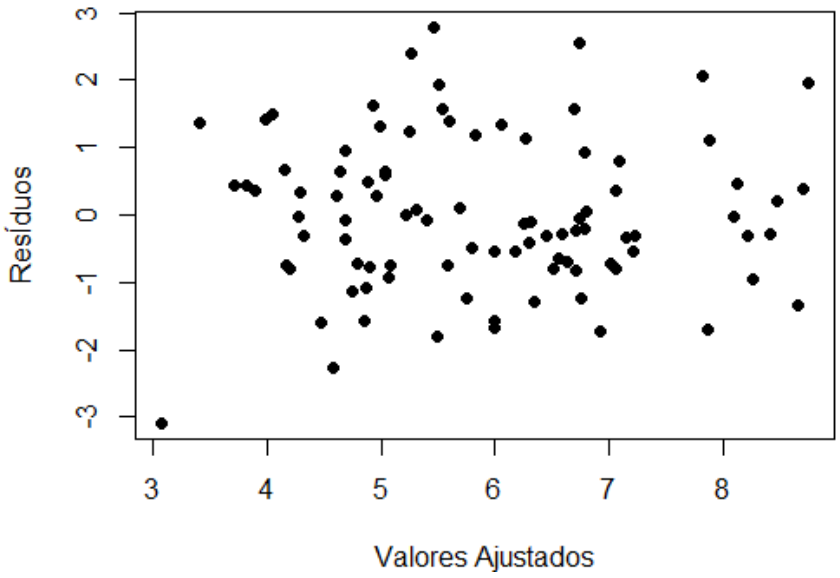

Figura 18: Suposição de homocedasticidade

O pressuposto de Normalidade foi atendido de acordo com o resultado do teste de *Shapiro-Wilk* ($p - valor = 0,6705$) e com a análise gráfica do *QQ-plot* apresentado na Figura ??

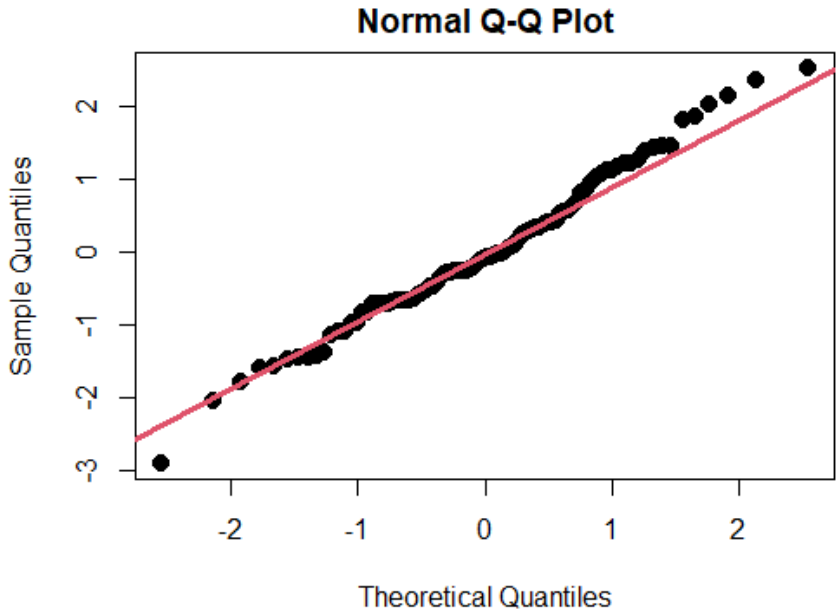

Figura 19: Gráfico QQ-Plot

Na Figura 20, que apresenta a Distância de Cook, observa-se que não foram identificados pontos influentes.

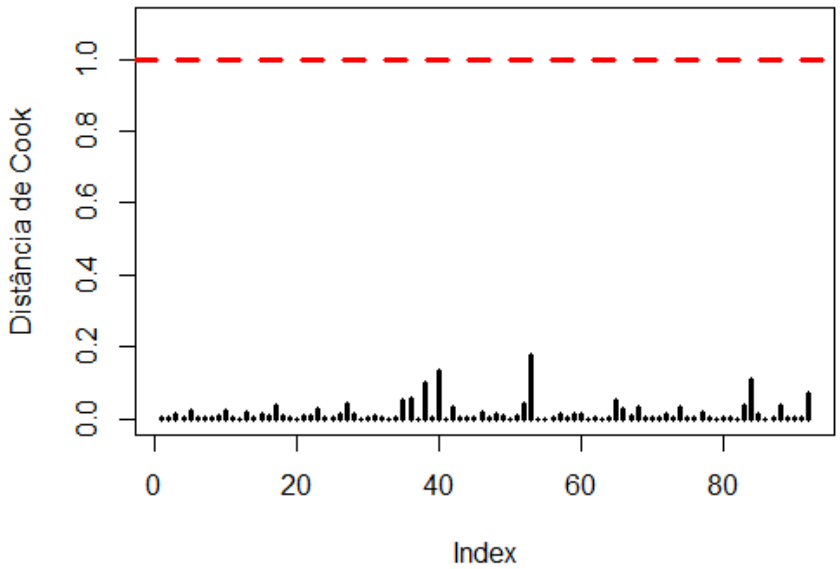

Figura 20: Gráfico da Distância de Cook

O Fator da Inflação da Variância (VIF) apresentou evidências que os dados são não correlacionados, onde nenhuma covariável apresentou VIF > 5.

Desta forma, o Modelo Final atendeu os pressupostos necessários e apresentou um R^2 relativamente adequado onde explica 60,01% dos dados.

4 Conclusão

A Tuberculose é uma das mais antigas doenças infecciosas e tramissíveis da humanidade. No Brasil são notificados aproximadamente 100 mil casos novos e ocorrem cerca de 4,5 mil mortes por ano em decorrência da Tuberculose e, em especial, o Rio de Janeiro se destaca entre os três primeiros estados com maior taxa de incidência da doença nos últimos anos.

O presente trabalho teve como objetivo criar um modelo para inferir sobre a existência de relação entre a taxa de incidência de Tuberculose e determinados fatores socioeconômicos, tomando como área de estudo os 92 municípios do Rio de Janeiro.

O banco de dados, composto por informações extraídas do DATASUS e do IBGE (Censo 2010), após depurado, passou por uma análise exploratória para verificar o comportamento da variável resposta do ponto de vista espacial e a possível relação das variáveis explicativas com a taxa de incidência de Tuberculose. Nesta etapa, a região Metropolitana se destacou por ter a maior mediana dentre as demais, assim como por ter 5 dos 6 municípios com as maiores taxas do Estado, são eles Japeri, Nova Iguaçu, Mesquita, Volta redonda e Rio de Janeiro. Com relação as variáveis explicativas, a densidade intradomiciliar, o percentual de ocupados com Ensino Médio completo e a taxa de incidência de HIV apresentaram evidências de variação linear positiva com a taxa de incidência de Tuberculose, já o percentual do domicílios com automóvel de uso particular apresentou evidência de relação linear negativa. Construiu-se uma matriz de correlação para avaliar covariáveis altamente relacionadas, desta forma foram retiradas do banco de dados as variáveis renda per capita, analfabetismo e IDHM.

O Modelo de Reressão Linear Múltiplo, que melhor se ajustou os dados, adotou uma transformação da raiz quadrada na variável resposta para que se adequasse ao pressuposto de Normalidade. Utilizou-se do método de seleção Stepwise para que se identifica-se as variáveis mais relevantes para o modelo, são elas: região, proporção de pobres, densidade intradomiciliar, ocupados com Ensino Médio completo e taxa de HIV. O modelo final

apresentou um $R^2 = 0,6001$.

De acordo com os resultados encontrados no modelo final, a distinção de região não se mostrou significativa quando comparada com a categoria de referência Baixadas Litorâneas. Estimou-se que a média da raiz da taxa de incidência de Tiberculose é 2,73 quando as demais covariáveis assumem valor zero. Identificou-se evidências de que a densidade intradomiciliar, o percentual de ocupados com Ensino Médio completo e a taxa de HIV tem uma relação positiva com a taxa de incidência de Tuberculose. Ou seja, a cada uma unidade ou ponto percentual dessas variáveis existe um aumento na taxa de incidência de Tuberculose. E que a relação entre o percentual de domícilios com automóvel particular e a taxa de incidência de Tuberculose é de decrescimento.

Os resultados encontrados neste trabalho, de que existe relação entre variáveis sociodemograficas e a taxa de incidência de Tuberculose, corroboram com resultados de outros pesquisadores como Fasca (2011), Paiva (2019), Fasca (2008), Angelo (2008), Cerqueira (2017) e Ximenes (2009) como visto na Tabela 1.

De forma geral, a regressão linear clássica mostrou-se ser uma boa ferramenta estatística na análise dos dados propostos neste trabalho. Mostraram-se as possíveis relações das variáveis estudadas com a Tuberculose. Certamente que outros modelos poderiam ser testados como o MLG ou o Beta. No entanto eles não foram alvo desse trabalho, ficando esta sugestão para trabalhos posteriores.

O modelo proposto neste trabalho possibilita investigar fatores importantes para entender melhor os condicionantes da incidência por Tuberculose. Estas análises podem nortear políticas de prevenção no que se diz respeito à incidência por Tuberculose.

Referências

ANGELO, J. R. (re) produção do espaço urbano de juiz de fora - mg e distribuição espacial da tuberculose. 2008.

BOX, G.; COX, D. An analysis of transformations. journal of the royal society 26: 211-252. 1964.

CERQUEIRA, S. S. C. Modelação da taxa de incidência de tuberculose nas Áreas metropolitanas de lisboa e porto. 2017.

CHUNG, S.; PEARN, W.; YANG, Y. A comparison of two methods for transforming con-normal manucturing data. international journal of advanced manufacturing tecnhnology. 2007.

COUTO, M. et al. Transformação de dados em experimentos com abobrinha italiana em ambiente protegido. ciência rual 39: 1701-1707. 2009.

FASCA, S. F. Tuberculose e condições de vida: uma análise do estado do rio de janeiro, 2000 a 2002. 2008.

FASCA, S. F. Tuberculose e condições de vida: uma análise do estado do rio de janeiro, 2000 a 2002. 2011.

FIOCRUZ. *Tuberculose*. 2022. ⟨https://portal.fiocruz.br/doenca/tuberculose⟩. Acessado em 10/08/2022.

IBGE. *Censo Demográfico*. 2022. ⟨https://ces.ibge.gov.br/apresentacao/portarias/200-comite-de-estatisticas-sociais/base-de-dados/1146-censo-demografico.html⟩. Acessado em 10/06/2022.

MS. Brasil livre da tuberculose: evolução dos cenários epidemiológicos e operacionais da doença. v.50, n.9. 2019.

PAIVA, B. L. Modelo preditivo de determinantes socioeconômicos da tuberculose em população indígena do estado do pará, brasil. 2019.

VICENTIN, G. Evolução da mortalidade por tuberculose no município do rio de janeiro. 2002.

WHO. *Tuberculose*. 2022. ⟨https://www.who.int/health-topics/tuberculosis⟩. Acessado em 10/08/2022.

XIMENES, R. A. de A. Is it better to be rich in a poor area or poor in a rich area? a multilevel analysis of a case-control study of social determinants of tuberculosis. 2009.

APÊNDICE 1 – Gráficos de Dispersão

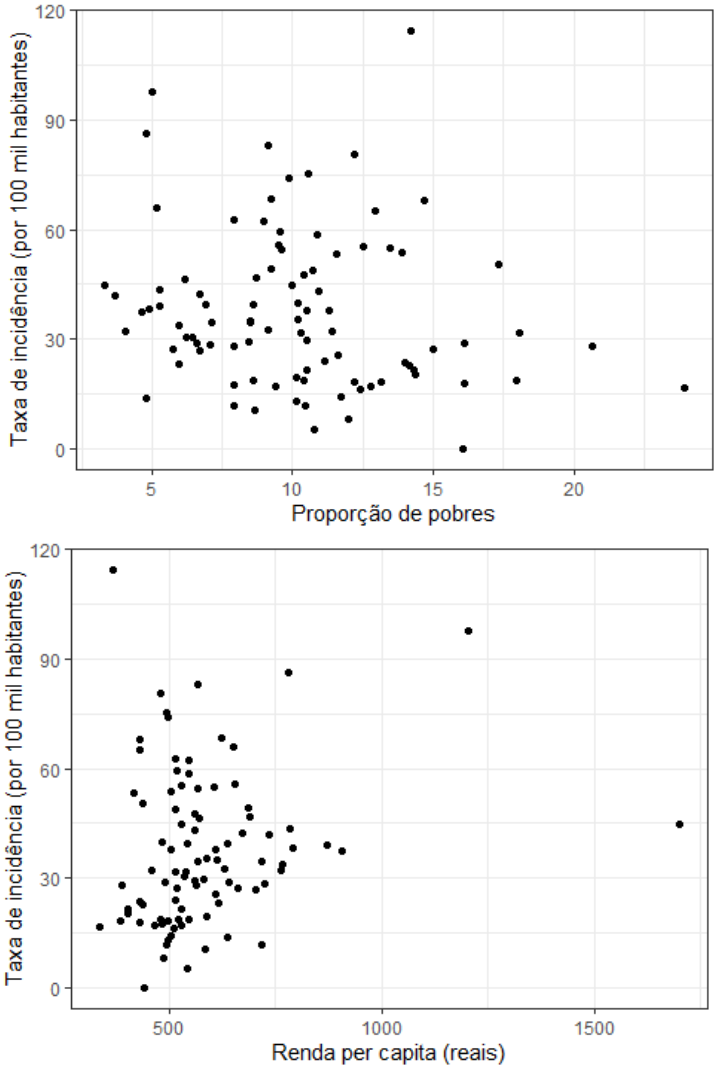

Figura 21: Gráficos de dispersãos entre a taxa de incidência de Tuberculose e percentual de chefes de domicílio com renda até um salário mínimo e rendimento mensal domiciliar per capita.

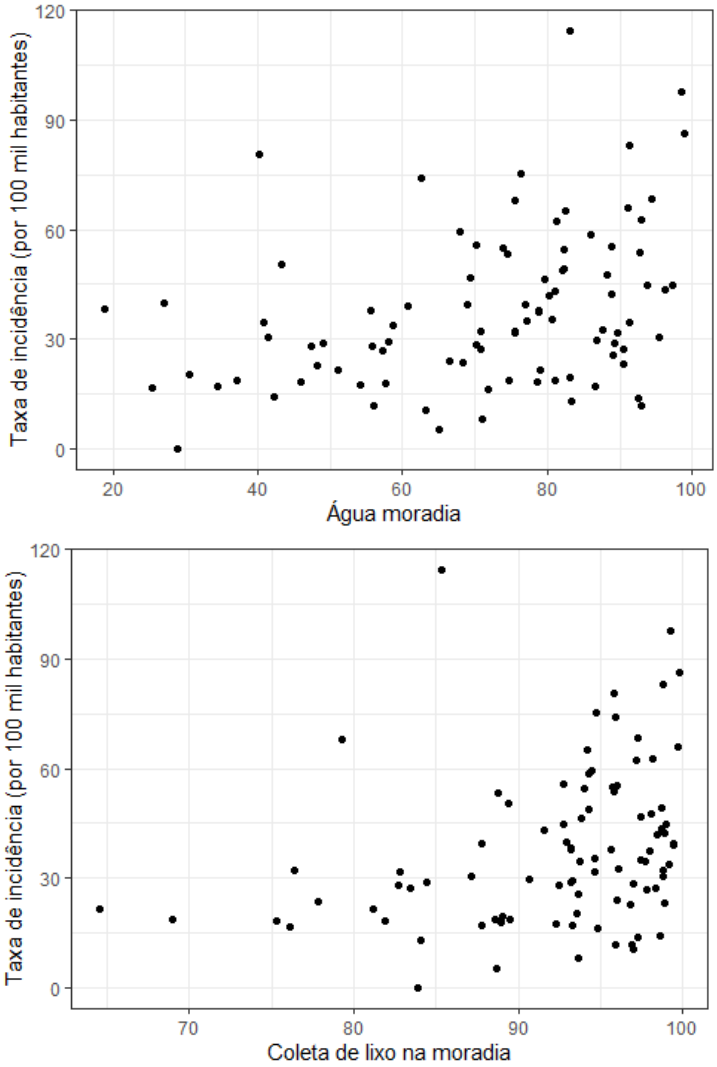

Figura 22: Gráficos de dispersãos entre a taxa de incidência de Tuberculose e percentual de domicílios conectados a abastecimento de água e percentual de pessoas em domicílios urbanos com coleta de lixo.

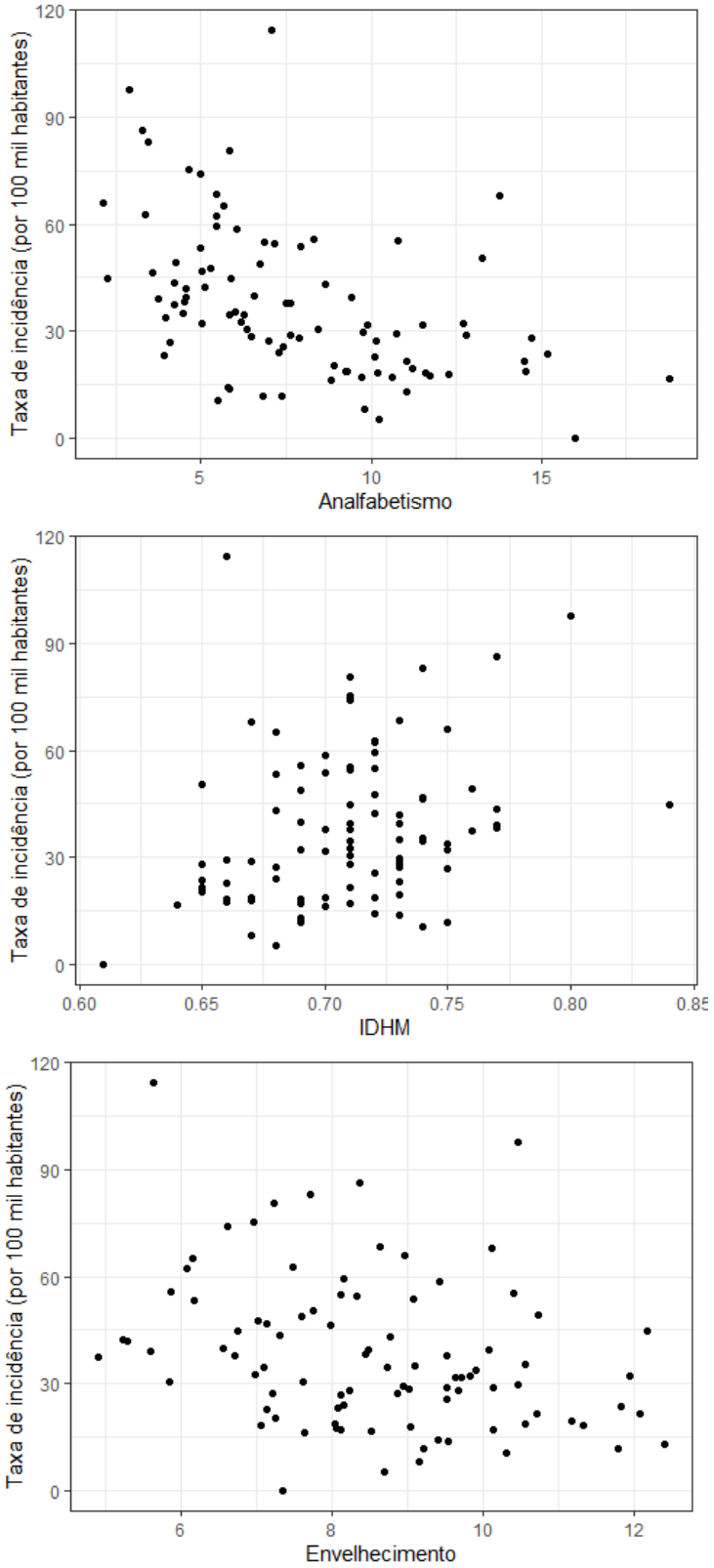

Figura 23: Gráficos de dispersãos entre a taxa de incidência de Tuberculose e a taxa de analfabetismo entre pessoas de 15 anos ou mais, o IDH-M e a taxa de envelhecimento da população.

APÊNDICE 2 – Resultados dos Modelos

Tabela 5: Estimativas dos parâmetros do Modelo 1

Parâmetros	**Estimativa**	**P-valor**
Intercepto	60,18034	0,6317
Região (Centro-Sul)	-0,33571	0,9654
Região (Costa Verde)	1,73061	0,8634
Região (Médio Paraíba)	7,99491	0,2935
Região (Metropolitana)	7,09732	0,2637
Região (Noroete)	10,10189	0,28
Região (Norte)	-6,85965	0,3823
Região Serrana	12,10265	0,1336
Renda per capita	0,03082	0,129
Proporção de pobres	1,94548	0,0287
Densidade intradomiciliar	1,09612	0,0678
Abastecimento de água	0,15375	0,1793
Coleta de lixo	-0,42463	0,3085
Analfabetismo	-1,68508	0,3118
Ocupados c/ Ensino Médio	0,56991	0,2775
IDHM	-85,48259	0,633
Automóvel particular	-0,27567	0,2434
Envelhecimento	-1,33296	0,3927
HIV	0,41871	0,0365

Tabela 6: Estimativas dos parâmetros do Modelo 2

Parâmetros	**Estimativa**	**P-valor**
Intercepto	6,275656	0,5336
Região (Centro-Sul)	-0,325064	0,6012
Região (Costa Verde)	0,086205	0,915
Região (Médio Paraíba)	0,392605	0,5193
Região (Metropolitana)	0,242351	0,6331
Região (Noroete)	0,453024	0,5449
Região (Norte)	-0,804183	0,2031
Região Serrana	0,908633	0,16
Renda per capita	0,002132	0,1898
Proporção de pobres	0,175789	0,0142
Densidade intradomiciliar	0,058989	0,2177
Abastecimento de água	0,010669	0,2448
Coleta de lixo	-0,039623	0,2369
Analfabetismo	-0,135789	0,3097
Ocupados c/ Ensino Médio	0,045414	0,2808
IDHM	-0,804802	0,9553
Automóvel particular	-0,041126	0,032
Envelhecimento	-0,159564	0,2037
HIV	0,035836	0,026