Análise Estatística de fatores socioeconômicos e geográficos associados à Mortalidade Infantil

Arthur Henrique da Rocha Hintz[1] **e Beatriz Woos Buffon**[1]

[1]Curso de Estatística, Universidade Federal de Santa Maria, Santa Maria, Rio Grande do Sul, Brazil.

Autores Contribuintes: beatriz.buffon@acad.ufsm.br, arthur.hintz@acad.ufsm.br

Resumo

O conjunto de dados InfMort é composto por 399 observações e 11 variáveis que simulam uma análise de mortalidade infantil baseada em dados do estado do Paraná, Brasil, no ano de 2010. Esse banco de dados foi criado com o objetivo de demonstrar aplicações de variáveis-resposta binomiais em estudos de saúde pública. As variáveis incluem informações socioeconômicas, como índices de desenvolvimento, pobreza, analfabetismo, além de dados sobre saúde e geolocalização dos municípios. A análise deste conjunto permite explorar fatores que influenciam a mortalidade infantil para estudar relações entre variáveis e propor soluções para problemas sociais.

Palavras-chaves: Análise Socioeconômica; Binomial; Mortalidade Infantil;

1 Introdução

A mortalidade infantil é um dos indicadores mais relevantes para avaliar as condições de saúde e desenvolvimento de uma população. No Brasil, as desigualdades regionais e socioeconômicas influenciam diretamente as taxas de mortalidade infantil, tornando este um problema complexo e multifacetado. Identificar os fatores associados a essas taxas é essencial para orientar políticas públicas que busquem a redução das desigualdades e a melhoria da qualidade de vida.

O conjunto de dados InfMort foi desenvolvido para simular cenários que exploram as causas e os determinantes da mortalidade infantil, com base em informações de municípios do Paraná, Brasil, em 2010. A problemática central reside em entender como essas variáveis interagem para influenciar a mortalidade infantil. Por exemplo, regiões com maior pobreza e menor cobertura do Programa Saúde da Família podem apresentar taxas mais elevadas de mortalidade. Além disso, fatores como nível educacional e desenvolvimento econômico são potenciais determinantes a serem investigados. O objetivo desta análise é explorar essas relações, identificando os fatores mais relevantes e como eles

afetam diretamente a mortalidade infantil. Espera-se que os resultados forneçam subsídios para intervenções mais eficazes no combate às desigualdades regionais e socioeconômicas, contribuindo para a melhoria da saúde infantil.

2 Descrição dos Dados

O banco de dados, inicialmente, possui 399 observações e 11 variáveis, as quais podem ser descritas:

1. `x`: Localização Longitudinal da cidade;
2. `y`: Localização Latitudinal da cidade;
3. `dead`: Número de mortalidade infantil;
4. `bornalive`: Número de bebês nascidos vivos;
5. `IFDM`: Índice FIRJAN de desenvolvimento da cidade;
6. `illit`: Índice de analfabetismo;
7. `lGDP`: Logaritmo do produto nacional bruto;
8. `cli`: Proporção de crianças que vivem num agregado familiar com metade do salário mínimo;
9. `lpop`: Logaritmo do número de habitantes;
10. `PSF`: Proporção abordada pelo Bolsa Família;
11. `poor`: Proporção de indivíduos com baixa renda per capita.

Algumas observações aleatórias do banco de dados podem ser visualizadas na Tabela 1 a seguir:

x	y	dead	bornalive	IFDM	illit	lGDP	cli	lpop	PSF	poor
-50.87454	-22.93413	0	62	0.7487	8.5	10.099	43.17	8.687	11	7.96
-50.08508	-23.29604	9	676	0.7446	9.8	9.332	42.62	10.645	920	8.20
-53.90928	-24.79339	0	76	0.6928	10.8	9.455	39.84	8.639	814	5.51
-51.07151	-23.05693	2	172	0.7585	7.3	9.805	30.73	9.687	11	3.18
-50.16239	-23.73344	2	59	0.6649	12.0	9.134	50.38	8.476	11	10.34

Tabela 1: Amostra de cinco registros do conjunto de dados InfMort.

As análises dos dados referem-se a um processo crítico em relação as variáveis que influenciam na mortalidade infantil. Dessa forma, na Tabela 2 foi vericado medidas de tendência central, medidas de disperção e as relações entre as variáveis. Sendo assim, as variáveis são todas do tipo numéricas, não há valores faltantes nos dados e é possível verificar um outlier com 231 mortes em um local específico.

skim_variable	n_missing	mean	sd	p0	p25	p50	p75	p100
x	0	-51.87	1.44	-54.48	-53.06	-52.05	-50.81	-48.37
y	0	-24.39	1.06	-26.47	-25.35	-24.27	-23.46	-22.57
dead	0	4.61	14.49	0.00	0.00	2.00	3.50	231.00
bornalive	0	381.08	1439.59	20.00	64.50	119.00	246.00	25345.00
IFDM	0	0.70	0.06	0.55	0.67	0.70	0.73	0.80
illit	0	10.20	3.92	1.50	7.30	10.10	12.70	19.00
IGDP	0	9.50	0.38	8.68	9.25	9.47	9.70	11.55
cli	0	44.47	13.80	10.22	34.64	44.32	54.20	78.45
lpop	0	9.29	1.06	7.27	8.55	9.13	9.79	14.43
PSF	0	539.43	656.94	1.00	11.00	11.00	1095.50	1990.00
poor	0	10.27	7.21	0.84	5.18	8.43	13.39	38.11

Tabela 2: Análise descritiva do conjunto de dados `InfMort`.

Os gráficos de dispersão da Figura 2 permitem verificar a relação entre todas as covariáveis com a variável resposta.

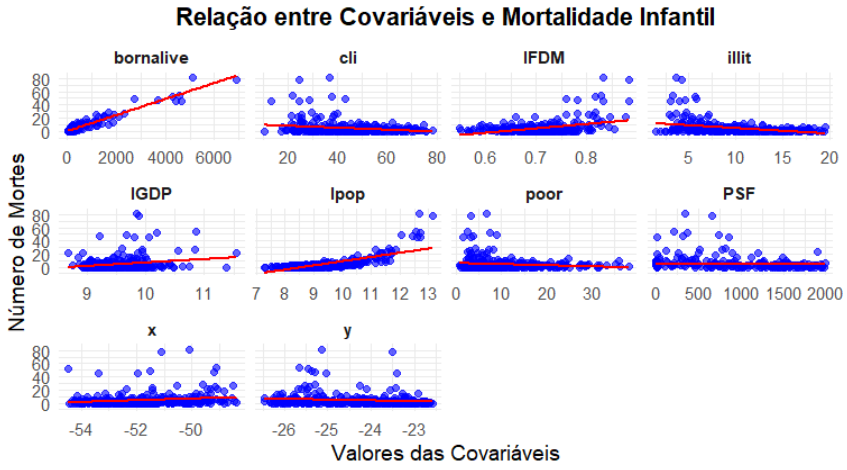

Figura 1: Relação entre Covariáveis e Mortalidade Infantil

2.1 Interpretação das Covariáveis Exibidas

`bornalive` **(Número de nascidos vivos):** Existe uma relação positiva clara entre o número de nascidos vivos e a mortalidade infantil. Cidades com mais nascidos vivos apresentam mais mortes infantis. Isso é esperado, pois um maior número de nascimentos aumenta o risco absoluto de mortalidade infantil.

`cli` **(Proporção de crianças em famílias de baixa renda):** Os pontos são dispersos, e a linha de tendência está quase horizontal, indicando ausência de relação clara. A proporção de crianças vivendo em famílias de baixa renda não parece influenciar diretamente o número de mortes infantis, pelo menos de forma linear.

`IFDM` **(Índice FIRJAN de Desenvolvimento Municipal):** Existe uma relação ligeiramente positiva, sugerindo que municípios mais desenvolvidos (maior IFDM) têm

mais mortes infantis. Cidades mais desenvolvidas podem apresentar maior mortalidade infantil absoluta. Isso pode ser explicado por fatores como:

- Maior população e número absoluto de nascidos vivos, levando a mais mortes;
- Melhor qualidade de registros nas cidades desenvolvidas, refletindo números mais precisos.

`illit` **(Taxa de analfabetismo):** Existe uma relação **negativa clara**, com menos mortalidade infantil em cidades com menor analfabetismo. Altas taxas de analfabetismo estão associadas a maiores números de mortes infantis. Isso é consistente com o impacto negativo do baixo nível educacional sobre a saúde pública e a mortalidade infantil.

`lGDP` **(Logaritmo do PIB per capita):** A relação é levemente positiva, indicando que cidades com maior PIB per capita apresentam mais mortes infantis. Essa relação também pode ser influenciada pelo tamanho populacional, já que municípios economicamente mais fortes tendem a ser mais populosos. Além disso, o PIB per capita pode refletir desigualdades internas que influenciam a mortalidade infantil.

`lpop` **(Logaritmo do número de habitantes):** Existe uma relação positiva clara entre o tamanho da população e o número de mortes infantis. Cidades mais populosas têm mais mortes infantis em números absolutos, o que é esperado. Isso não significa necessariamente que a taxa de mortalidade infantil seja maior em cidades grandes, mas apenas que há mais mortes em termos absolutos devido ao maior número de nascidos vivos.

`poor` **(Proporção de indivíduos com baixa renda per capita):** A relação é muito fraca e dispersa, mas com uma tendência ligeiramente positiva. Municípios com maior proporção de indivíduos de baixa renda podem ter uma leve associação com maior mortalidade infantil, mas a relação não é forte. Fatores socioeconômicos podem desempenhar um papel aqui, mas outras variáveis como saúde pública e educação podem interferir.

`PSF` **(Cobertura pelo Programa Saúde da Família):** Existe uma relação negativa clara, indicando que uma maior cobertura do programa está associada a menor mortalidade infantil. O PSF parece cumprir seu papel na redução da mortalidade infantil. Municípios com maior cobertura do programa possuem melhores indicadores de saúde para crianças.

`x` **e** `y` **(Coordenadas geográficas):** Não há uma relação clara entre a localização longitudinal (`x`) ou latitudinal (`y`) e o número de mortes infantis. A posição geográfica das cidades não parece ser um fator relevante para a mortalidade infantil, mas pode haver fatores regionais não capturados diretamente pelas coordenadas (ex.: características socioeconômicas ou culturais).

Na Figura 3 do gráfico de "Dispersão da Mortalidade Infantil por Índice" é possível verificar que muitos locais não apresentam mortalidade infantil, o que é evidente pela concentração de pontos na linha 0 do eixo vertical. Isso sugere que algumas localidades

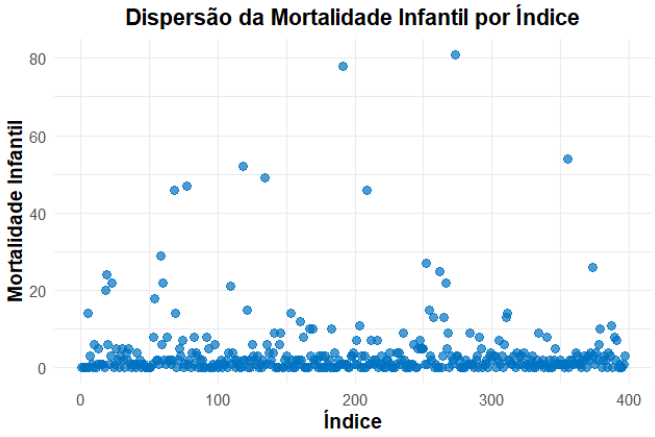

Figura 2: Dispersão da Mortalidade Infantil por Índice

possuem boas condições de saúde, como acesso a serviços materno-infantis, programas sociais eficazes e melhores condições socioeconômicas. Além disso, pode estar relacionado a localidades com populações pequenas, onde o número absoluto de óbitos tende a ser menor. No entanto, o gráfico também destaca a existência de locais com mortalidade significativamente alta, evidenciando desigualdades que merecem uma análise mais detalhada.

O gráfico de correlação entre as variáveis da Figura 4 pode ser interpretado a partir do tamanho dos círculos e da cor. Quanto maior o círculo e mais escura é a cor mais forte é a correlação, também, se puxar mais para o azul é positiva, se for vermelha é negativa.

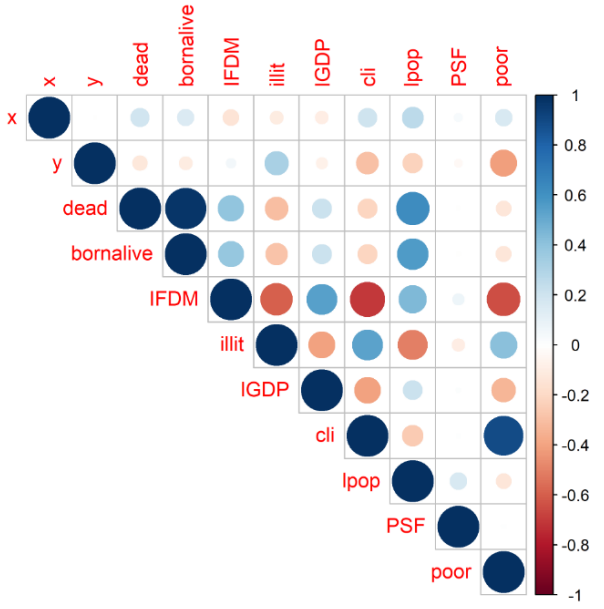

Figura 3: Correlação entre as variáveis

Com base na matriz de correlação apresentada e na descrição das variáveis, aqui está uma interpretação dos principais padrões que podem ser observados:

1. **Relação entre `dead` e `bornalive`:** Existe uma correlação positiva forte entre o número de mortes infantis (`dead`) e o número de bebês nascidos vivos (`bornalive`). Isso é esperado, pois onde há mais nascimentos, também tende a haver mais óbitos infantis em termos absolutos.

2. **Impacto do desenvolvimento (`IFDM`):** `IFDM` apresenta correlações negativas com variáveis como `illit` (índice de analfabetismo) e `poor` (proporção de indivíduos com baixa renda per capita). Isso faz sentido, pois um maior desenvolvimento está associado a melhores condições educacionais e econômicas. Por outro lado, `IFDM` possui correlações positivas com `lGDP` (logaritmo do produto interno bruto) e `lpop` (logaritmo do número de habitantes), indicando que cidades mais desenvolvidas tendem a ter economias maiores e populações mais expressivas.

3. **Pobreza e condições de vida (`poor`, `cli`, `PSF`):** Há uma correlação positiva moderada a forte entre `poor` e `cli` (proporção de crianças em famílias com baixa renda), indicando que essas variáveis estão relacionadas com desigualdade social. `PSF` (cobertura do Bolsa Família) apresenta correlações positivas com indicadores de pobreza, como `poor` e `cli`. Isso reflete que o programa está mais presente em regiões com maior vulnerabilidade.

4. **Educação e mortalidade infantil:** `illit` (índice de analfabetismo) tem uma correlação positiva com `dead`, sugerindo que taxas mais altas de analfabetismo podem estar associadas a maiores níveis de mortalidade infantil. Além disso, `illit` se correlaciona negativamente com `IFDM`, indicando que o analfabetismo tende a ser menor em cidades mais desenvolvidas.

5. **Tamanho da população (`lpop`):** `lpop` apresenta correlação positiva com `lGDP`, indicando que cidades com populações maiores tendem a ter economias mais robustas. No entanto, a relação de `lpop` com indicadores sociais como `poor` e `cli` parece ser menos evidente (possivelmente neutra ou fraca).

6. **Localização geográfica (`x`, `y`):** As variáveis de localização longitudinal (`x`) e latitudinal (`y`) apresentam correlações muito fracas com as outras variáveis, sugerindo que o local geográfico pode não ser diretamente relacionado aos indicadores analisados.

3 Ajuste dos Dados

Foi removido o número de mortalidade igual a 231, pois distoava do resto dos dados e a observação 52 também se mostrou influente no modelo.

4 Modelo Ajustado

A variável aleatória Y, representando a quantidade de mortes infantis, é assumida como seguindo uma Distribuição Binomial Negativa, a qual pertence à família exponencial canônica com parâmetro de dispersão. Essa escolha é apropriada dado que o espaço amostral de Y é $\{0, 1, 2, \ldots, n\}$ e o parâmetro de média $\mu > 0$.

A função de ligação logarítmica, definida como $\eta = \log(\mu)$, foi utilizada em função da natureza discreta e positiva dos dados, bem como do conjunto de valores possíveis da variável de interesse. Entre as covariáveis disponíveis no banco de dados, apenas quatro se mostraram significativas no modelo final. Assim, o modelo é expresso como:

$$\eta = \log(\mu) = \sum_{j=1}^{4} X_{ji}\beta_j.$$

Embora a distribuição de Poisson também tenha sido considerada, ela não apresentou um ajuste satisfatório aos dados, reforçando a escolha da Binomial Negativa para modelar os dados.

A estimação do parâmetro μ foi realizada por meio da função `glm.nb` que emprega um processo iterativo até alcançar a convergência para o modelo ajustado. Assim, o modelo pode ser visualizado:

Estimate	**Std. Error**	**z value**	**Pr($>\|z\|$)**	
(Intercept)	-6.8445318	1.4360743	-4.766140	0.0000019
x	0.0594268	0.0215809	2.753674	0.0058930
y	-0.0801179	0.0327035	-2.449825	0.0142926
IFDM	-2.3582629	0.6754348	-3.491474	0.0004804
lpop	1.1007013	0.0395759	27.812385	0.0000000

Tabela 3: Modelo do MLG

O modelo pode ser escrito na forma:

$$\log(\hat{\mu}) = -6.8 + 0.059X_1 - 0.08X_2 - 2.354X_3 + 1.1X_4,$$

em que:

$$X_1 = x, \quad x \in [-54.483, -48.367]$$

$$X_2 = y, \quad x \in [-26.472, -22.573]$$

$$X_3 = IFDM, \quad x \in [0.550, 0.883]$$

$$X_4 = lpop, \quad x \in [7.268, 13.144]$$

5 Resultados

As principais variáveis que estão associadas ao número de mortalidade infantil são as coordenadas da cidade, o índice de desenvolvimento da cidade e a quantidade de habitantes.

O AIC do modelo foi de 1357.007.

5.1 Interpretação dos Betas

- β_1: Significa que a cada 1 grau da longitude da cidade, aumenta 1.1051709 no número médio de óbitos, ou seja, quanto mais para o leste do Paraná, o número de mortalidade infantil tende a ser maior.
- β_2: Significa que a cada 1 grau da latitude da cidade, aumenta 0.9048374 no número médio de óbitos, ou seja, quanto mais para o norte do Paraná, o número de mortalidade infantil tende a aumentar.
- β_3: Significa que quanto maior o IFDM da cidade, aumenta 0.090718 no número médio de óbitos.
- β_4: A cada 1 unidade no log da população aumenta, em média, 3.004166 no número de óbitos infantis da cidade.

5.2 Análise de Diagnóstico

A análise de diagnóstico é uma etapa essencial na modelagem estatística para avaliar se o modelo ajustado é adequado para os dados. Isso envolve verificar se os pressupostos do modelo foram atendidos, se as variáveis escolhidas são relevantes e se há evidências de especificação incorreta ou outros problemas estruturais. No contexto apresentado, foram realizados testes específicos para validar o modelo.

5.3 Desvio

Devido ao p-valor igual a 0.080889, maior que o nível de significância de $\alpha = 5\%$, não há evidências estatísticas suficientes para rejeitar o modelo ajustado. Ou seja, o modelo é adequado para os dados. O modelo nulo (apenas a constante, sem variáveis explicativas) teve um desvio de 2669.374. O modelo ajustado (com as variáveis explicativas) apresentou um desvio de 431.786. Essa redução substancial no desvio indica que o modelo ajustado conseguiu explicar grande parte da variabilidade dos dados, mostrando que as variáveis incluídas têm relevância no ajuste.

5.4 Teste RESET

De acordo com o teste RESET, a função de ligação está adequada e outras especificações estão corretas, devido ao p-valor = 0.644, maior que 5% de significância.

5.5 Coeficiente de Determinação Generalizado

- McFadden: 0.299362
- Cox and Snell (ML): 0.764803
- Nagelkerke (Cragg and Uhler): 0.770931

O R^2 de Nagelkerke é dado por 0.770931, ou seja, significa que 77.09% da variabilidade da mortalidade infantil pode ser explicada pelas variáveis incluídas no modelo.

5.6 Resíduos Deviance

A figura apresenta um gráfico de resíduos padronizados em função do índice das observações, utilizado para diagnosticar a adequação do modelo ajustado. Os resíduos estão distribuídos de forma aleatória ao longo do gráfico, sem apresentar um padrão ou tendência evidente. Isso é um indicativo de que o modelo está bem especificado e que as variáveis explicativas incluídas capturam bem a estrutura dos dados.

A maioria dos resíduos está dentro do intervalo entre -2 e 2, que corresponde a 95% da probabilidade da distribuição normal padrão. Isso sugere que os resíduos seguem uma distribuição aproximadamente normal, como esperado em modelos bem ajustados. Alguns poucos resíduos ultrapassam o limite de ± 2, mas isso não é incomum em conjuntos de dados reais.

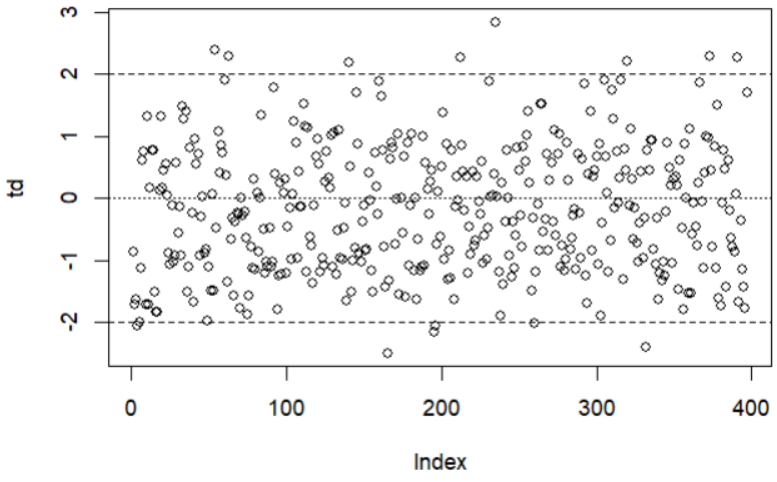

Figura 4: Gráfico dos Resíduos

5.7 Envelope Simulado

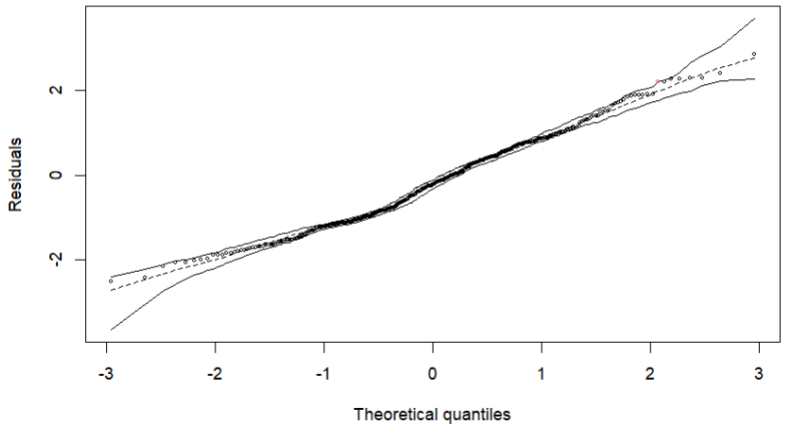

Figura 5: Gráfico do Envelope Simulado

A maior parte dos pontos segue uma linha reta próxima da diagonal, o que indica que os resíduos estão aproximadamente normalmente distribuídos. Isso suporta a suposição de normalidade dos resíduos, essencial para modelos lineares. Pequenos desvios nos extremos podem ser investigados mais detalhadamente se forem relevantes para o objetivo do estudo. Este comportamento é comum em dados reais e não é necessariamente problemático, em que até 5% dos pontos podem estar fora dos limites (linhas tracejadas), o que é aceitável e esperado. Assim, o gráfico sugere uma boa adequação do modelo. A distribuição dos resíduos é consistente com a distribuição Binomial Negativa assumida, o que valida a escolha da distribuição para o modelo ajustado.

5.8 Alavancagem

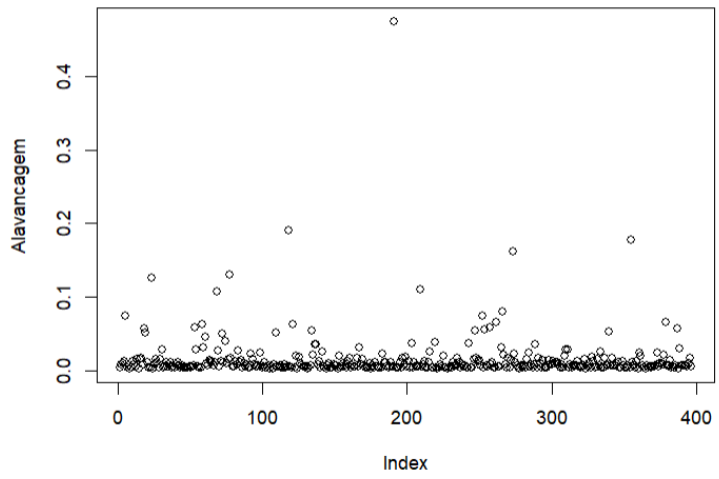

Figura 6: Gráfico da Alavancagem

A observação 191, não apresentou ser um ponto influênte no modelo.

5.9 Distância de Cook

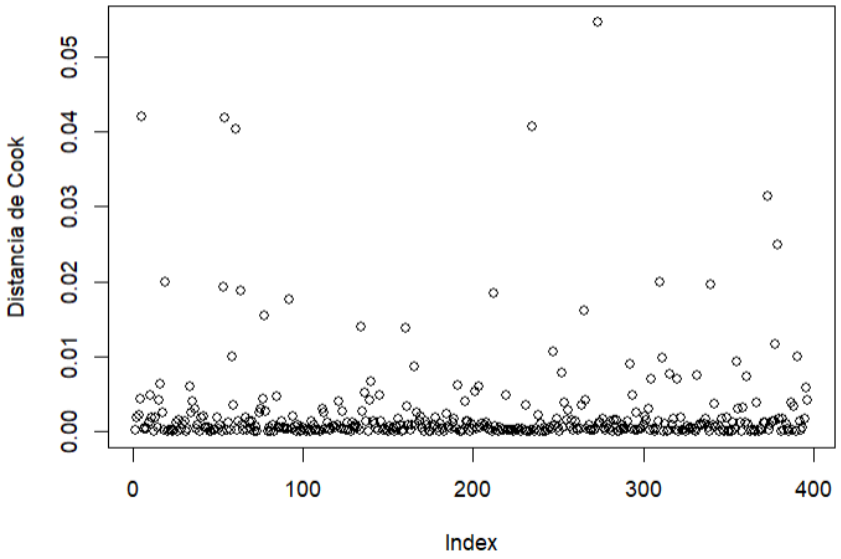

Figura 7: Gráfico da Distância de Cook

Os pontos mais discrepantes identificados no gráfico da Distância de Cook foram analisados, e sua remoção não resultou em alterações significativas nos coeficientes da regressão.

5.10 DFFITS

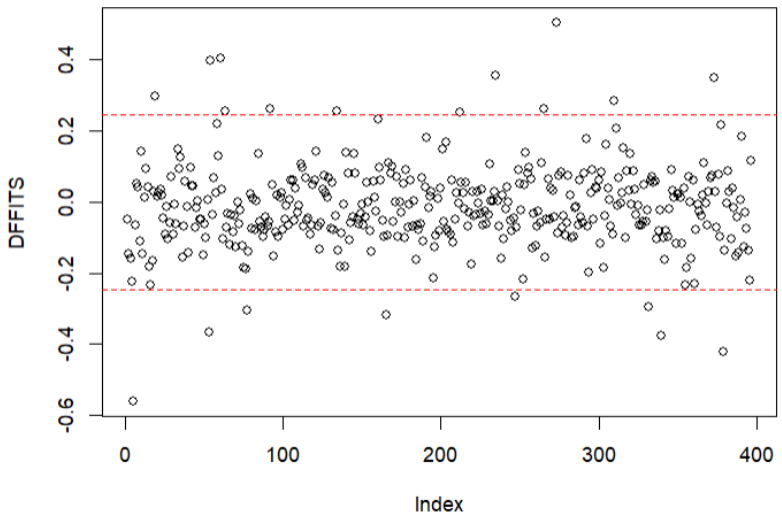

Figura 8: Gráfico dos DFFITS

De acordo com o gráfico, nota-se que não tem nenhuma observação com impacto desproporcional no modelo ajustado.

5.11 Predição

O gráfico a seguir mostra a relação entre os valores reais da quantidade de mortes, com os valores preditos a partir do modelo ajustado.

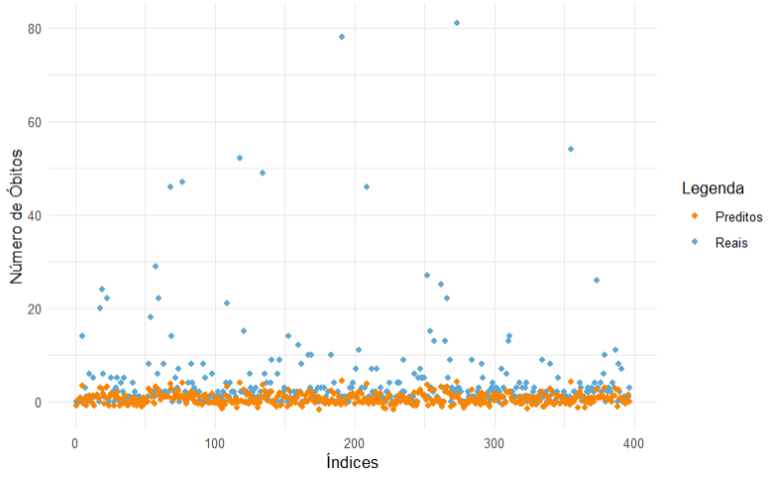

Figura 9: Gráfico de Predição da Mortalidade Infantil

6 Conclusão

A partir do modelo ajustado para a análise da mortalidade infantil, observou-se que as principais variáveis associadas ao número de óbitos infantis são as coordenadas geográficas da cidade, o índice de desenvolvimento da cidade (IFDM) e a quantidade de habitantes (log-população). Os coeficientes de regressão indicam que tanto a longitude quanto a latitude da cidade influenciam significativamente a mortalidade infantil, com um aumento observado à medida que as cidades se localizam mais para o leste e norte do Paraná. Além disso, a variável IFDM mostrou uma associação negativa com o número de óbitos, enquanto a população apresentou uma relação positiva, indicando que cidades com maior população têm, em média, mais mortes infantis. O AIC do modelo ajustado foi de 1357.007, e o valor de R^2 de Nagelkerke indicou que cerca de 77,09% da variabilidade da mortalidade infantil foi explicada pelas variáveis selecionadas, o que sugere um bom poder explicativo do modelo. O teste de diagnóstico mostrou que os resíduos estão distribuídos aleatoriamente e seguem uma distribuição aproximadamente normal, o que valida a adequação do modelo para os dados. Além disso, as análises de alavancagem e distância de Cook indicaram que não há pontos influentes que possam comprometer os resultados. O modelo mostrou-se robusto e bem ajustado, sendo capaz de explicar uma parte significativa da variabilidade observada nos dados. A análise de predição também indicou uma

boa correspondência entre os valores reais e preditos, corroborando a qualidade do ajuste. Com isso, é possível concluir que o modelo proposto é adequado para entender os determinantes da mortalidade infantil no Paraná, com destaque para as variáveis geográficas e demográficas.