Universidade de Brasília
Departamento de Estatística

Regressão quantílica na análise da letalidade da COVID-19

Gabriel Tormin Alves

Monografia apresentada ao Departamento de Estatística da Universidade de Brasília como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Brasília
2022

Gabriel Tormin Alves

Regressão quantílica na análise da letalidade da COVID-19

Orientador(a): Prof. Dr. Helton Saulo Bezerra dos Santos

Monografia apresentada ao Departamento de Estatística da Universidade de Brasília como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Brasília
2022

Dedico este trabalho à minha família e meus amigos, sem o seu suporte não conseguiria ter passado por tantos desafios.

Agradecimentos

- Agradeço a Deus, se cheguei até aqui foi porque o Senhor me sustentou.
- Agradeço ao departamento e aos professores, por me ajudarem e me guiarem na minha formação acadêmica e profissional.

Resumo

Nesse trabalho um modelo de regressão quantílica logística para dados limitados é utilizado para analisar os determinantes da taxa de letalidade da COVID-19 nos municípios do estado de São Paulo. São considerados três períodos (ondas) da pandemia ocorridos entre fevereiro de 2020 e março de 2022. O uso de uma metodologia que leva em consideração dados limitados é especialmente importante no contexto de regressão quantílica, uma vez que resultados mais confiáveis são obtidos com métodos que restringem a inferência dentro de um intervalo limitado.

Palavras-chaves: Regressão Quantílica; COVID-19; Taxa de letalidade.

Lista de Tabelas

Lista de Figuras

Sumário

1 Introdução

A pandemia do novo coronavírus (COVID-19) é um tema de discussão para muitos estudos e análises. O comportamento do vírus e de pacientes que contraem a doença se tornam foco de muitas pesquisas que visam entender o retrato da situação para que no futuro decisões sejam tomadas corroboradas por dados e evidências.

Esse trabalho foca em uma medida de risco de mortalidade devido à COVID-19 particularmente importante que serve para monitorar a gravidade da pandemia, que é a taxa de letalidade. Essa taxa é calculada como a razão entre o número de mortes confirmadas pelo número de casos confirmados; ver detalhes em ⟨https://ourworldindata.org/mortality-risk-covid⟩.

Dados de taxas e proporções são usualmente modelados na literatura utilizando regressão beta (FERRARI; CRIBARI-NETO, 2004). No entanto, esse modelo é baseado na média e pode ser inadequado se a variável dependente observada segue uma distribuição assimétrica, que é o caso dos dados da taxa de letalidade. Além disso, a regressão beta não permite fornecer uma imagem mais abrangente do efeito das variáveis explicativas sobre a variável dependente ao longo do espectro da variável dependente; o que é possível através da regressão quantílica introduzida por Koenker e Bassett Jr (1978), ver também Davino, Furno e Vistocco (2014). Todavia, quando a variável dependente é limitada, como é o caso da taxa de letalidade (limitada entre 0 e 1), podem ocorrer problemas quando métodos estatísticos tradicionais são utilizados. Em geral, usar métodos que restringem a inferência dentro do intervalo limitado leva a conclusões mais confiáveis; ver Bottai, Cai e McKeown (2010).

Nesse contexto, esse trabalho visa utilizar um modelo de regressão quantílica logística para dados limitados, introduzido por Bottai, Cai e McKeown (2010), para a modelagem da taxa de letalidade da COVID-19 de municípios do estado de São Paulo. As análises serão realizadas para três períodos (ondas) da pandemia ocorridos entre fevereiro de 2020 e março de 2022. Desta maneira, com uma abordagem quantílica, será feito um estudo dos fatores determinantes da letalidade.

O restante desse trabalho está estruturado da seguinte forma. Na Seção 2, o modelo de regressão quantílica logística é brevemente apresentado. Nessa seção, breves detalhes de estimação dos parâmetros são também apresentados. Na Seção 3, é apresentada uma discussão preliminar sobre os dados utilizados. A Seção 4 serão mostrados os resultados de estimação. E, por fim, a Seção 5 apresenta as conclusões finais

2 Regressão quantílica logística

2.1 O modelo

Considere a variável resposta y, uma variável contínua e pertencente ao intervalo $[y_{min}; y_{max}]$, e um conjunto de s covariáveis, $X = \{x_1, ..., x_s\}$. Seja $Q_y(p)$ o p-ésimo quantil de y dado o conjunto de covariáveis, em que p é uma proporção entre zero e um.

Assumindo que para qualquer quantil p existe um conjunto de parâmetros, $\beta_p = \{\beta_{p,0}, \beta_{p,1}, ..., \beta_{p,s}\}$, e uma função não-decrescente h do intervalo $[y_{min}; y_{max}]$ para a linha real, pode-se escrever o modelo da seguinte maneira

$$h\{Q_y(p)\} = \beta_{p,0} + \beta_{p,1}x_1 + ... + \beta_{p,s}x_s.$$

Considerando uma variável resposta limitada no intervalo unitário, o que se assemelha a uma probabilidade em vários aspectos, tem-se que uma escolha razoável para a função h é o link logístico (BOTTAI; CAI; MCKEOWN, 2010)

$$h(y) = \text{logit}(y) = \log\left(\frac{y - y_{min}}{y_{max} - y}\right).$$

Em consequência,

$$Q_y(p) = \frac{\exp(\beta_{p,0} + \beta_{p,1}x_1 + ... + \beta_{p,s}x_s)y_{max} + y_{min}}{\exp(\beta_{p,0} + \beta_{p,1}x_1 + ... + \beta_{p,s}x_s) + 1}.$$

O modelo Regressivo Quantílico possui uma certa flexibilidade em relação a heterocedasticidade, já que as diferentes estimativas dos coeficientes se adaptam ao longo do espectro quantílico. Multicolinearidade deve ser evitada por afetar a capacidade do ajuste do modelo aos dados; ver Davino, Romano e Vistocco (2022) para um estudo sobre como multicolinearidade e homocedasticidade afeta o modelo. A aditividade dos efeitos é assumida.

2.2 Estimação dos parâmetros

Em regressão linear as estimativas são obtidas através da minimização da função de erro quadrática médio. Como na regressão quantílica o foco é no p-ésimo quantil, uma outra função de perda precisa ser utilizada. Conforme Koenker e Machado (1999),

a função de perda deve ser

$$l_p(u) = \begin{cases} up, & \text{se } u \geq 0, \\ u(p-1), & \text{se } u < 0. \end{cases}$$

Logo, para obter o conjunto $\hat{\beta}$ de estimativas dos parâmetros, é necessário resolver o seguinte problema de minimização:

$$\min_{\beta \in \mathbb{R}^{q+1}} \sum_{i=1}^{n} l_p[y_i - (\beta_{p,0} + \beta_{p,1}x_1 + ... + \beta_{p,s}x_s)].$$

O conjunto $\hat{\beta}$ de estimativas é obtido através da função `rq()` do pacote `quantreg` em conjunto com pacote `rms`, ambos do `R`, que proporciona ferramentas que facilitam a estimação. O método padrão "br"foi utilizado para obter o ajuste; esse método é descrito em detalhes em Koenker e d'Orey (1987), Koenker e d'Orey (1994).

3 Descrição e análise exploratória dos dados

3.1 Conjunto de dados

O banco de dados construído para a regressão foi montado com base em três fontes:

- Os dados relativos aos casos de COVID-19 são fornecidos pela Secretaria de Estado da Saúde de São Paulo (SES-SP), a partir dos microdados estão incluídas informações sobre fatores de risco do paciente, e se o caso veio a óbito - disponível em ⟨https://www.saopaulo.sp.gov.br/planosp/simi/dados-abertos/⟩;
- Os dados da Campanha Nacional de Vacinação contra COVID-19 estão disponíveis no portal do dados abertos do Ministério da Saúde, o openDataSUS, a partir deles é possível calcular o número de doses e qual dose foi aplicada - disponível em ⟨https://opendatasus.saude.gov.br⟩;
- Informações sócio-demográficas, como população e tamanho do território do município foram adquiridas a partir do Instituto Brasileiro de Geografia e Estatística (IBGE).

Como a situação epidemiológica é sensível a problemas e medidas do momento em que se passa foi optado por estratificar os dados em três momentos que coincidem com as ondas da pandemia, ilustrados na Figura 1:

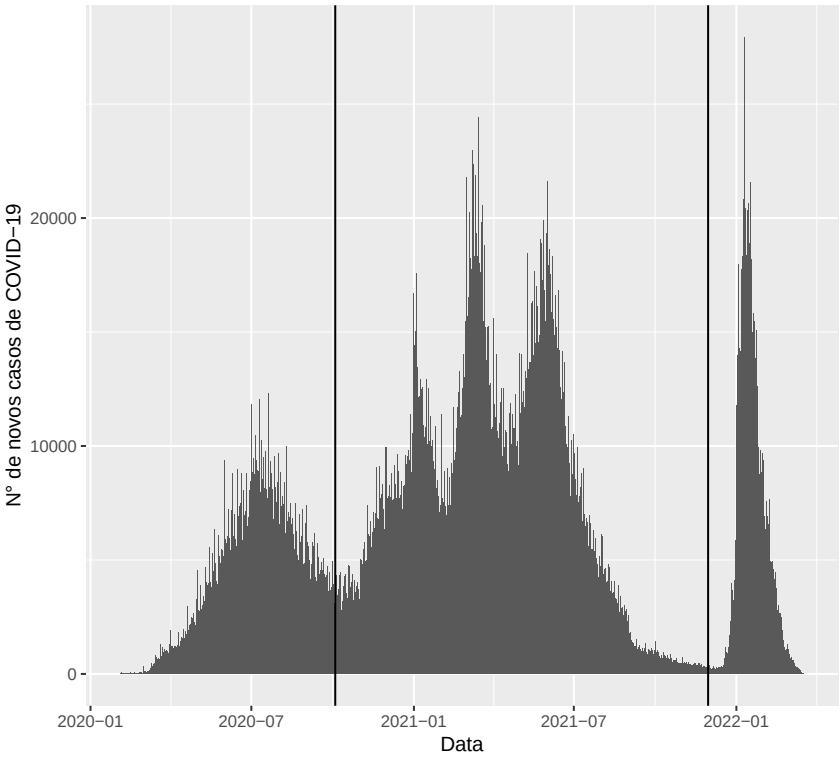
- **Período 1** - De 04/02/2020 a 04/10/2020;
- **Período 2** - De 05/10/2020 a 30/11/2021;
- **Período 3** - De 01/12/2021 a 17/03/2022.

Como variável resposta temos:

- `let`: taxa de letalidade da COVID-19 (número de mortes dividido pelo número total de casos) para o período levando em conta o município de origem da notificação. Na Figura 2 temos uma representação gráfica da taxa de letalidade em cada município do Estado de São Paulo, gerado a partir dos dados fornecidos pelo SES-SP;

As seguintes covariáveis são consideradas:

Figura 1: Gráfico de novos casos de COVID-19



Fonte: Elaborado pelo autor.

- `Risco`: A porcentagem de casos de COVID-19 com alguma complicação devida a fatores de risco (Asma, Diabetes, Cardiopatia, Doença Hematológica, Hepática, Neurológica, entre outros), gerada a partir dos microdados da SES-SP, levando em conta o município de origem da notificação;

- `Idade_mediana`: A idade mediana (em anos) dos pacientes com COVID-19 no município durante cada período. Essa informação foi extraída dos microdados do SES-SP;

- `densidade2021`: Densidade populacional do município (em pessoas por km^2), calculada a partir da projeção da população de 2021, de acordo com o IBGE. Esta variável se mantém constante em todos os períodos;

- `IDHM`: Índice de desenvolvimento humano municipal (IDHM) em 2020 fornecido pelo IBGE. É um índice que varia entre 0 e 1, sendo quanto mais próximo de 1 maior o desenvolvimento humano. Esta variável também se mantém constante em todos os períodos;

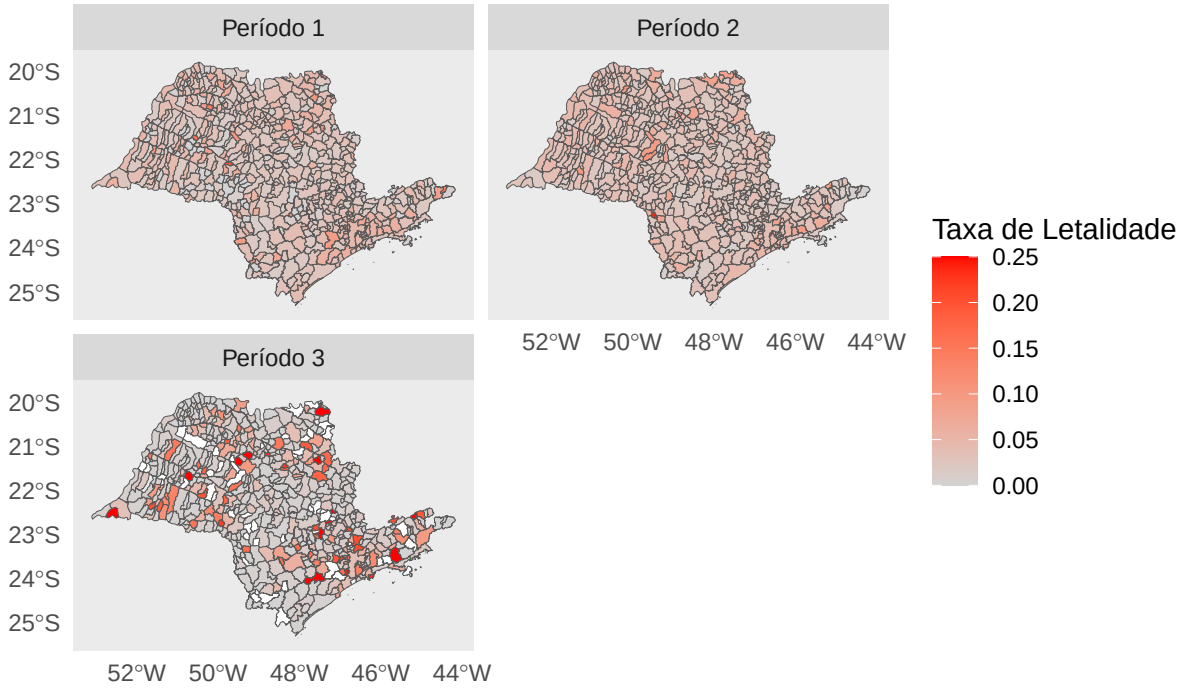
- `dose1`: porcentagem normalizada pelo método min-max da população municipal com pelo menos uma dose da campanha de vacinação contra o COVID-19 sem contar

as doses únicas, nos intervalos, calculado a partir dos dados do openDataSUS e das projeções populacionais do IBGE para 2021;

- `dose2`: porcentagem normalizada pelo método min-max da população municipal totalmente vacinada de acordo com o antigo plano vacinal (duas doses ou dose única) contra a COVID-19 dentro de cada intervalo, calculado a partir dos dados do openDataSUS e das projeções populacionais do IBGE para 2021;

- `reforco`: porcentagem normalizada pelo método min-max da população local vacinada com a primeira dose de reforço contra a COVID-19, calculado a partir dos dados do openDataSUS e das projeções populacionais do IBGE para 2021. Esta variável só é utilizada no último período analisado;

Figura 2: Gráfico da taxa de letalidade (`let`) do COVID-19 no Estado de São Paulo.



Fonte: Elaborado pelo autor.

3.2 Análise exploratória dos dados

Inicialmente será analisada a variável resposta `let`. A Figura 3 apresenta os histogramas para as taxas de letalidade dos municípios nos três períodos estudados. Consoante com a Tabela 1, que apresenta estatísticas descritivas, podemos observar que o

período 3 possui valores mais dispersos, ao passo que os períodos 1 e 2 possuem desvio padrão (DP) mais próximos de zero.

Na Tabela 1 são encontrados os coeficientes de variação (CV), assimetria (CA) e excesso de curtose (CC). De uma maneira geral a taxa de letalidade se distribui de maneira assimétrica com cauda pesada. Ressalta-se que o tamanho da amostra no período 3 é menor devido a alguns municípios não terem apresentado casos.

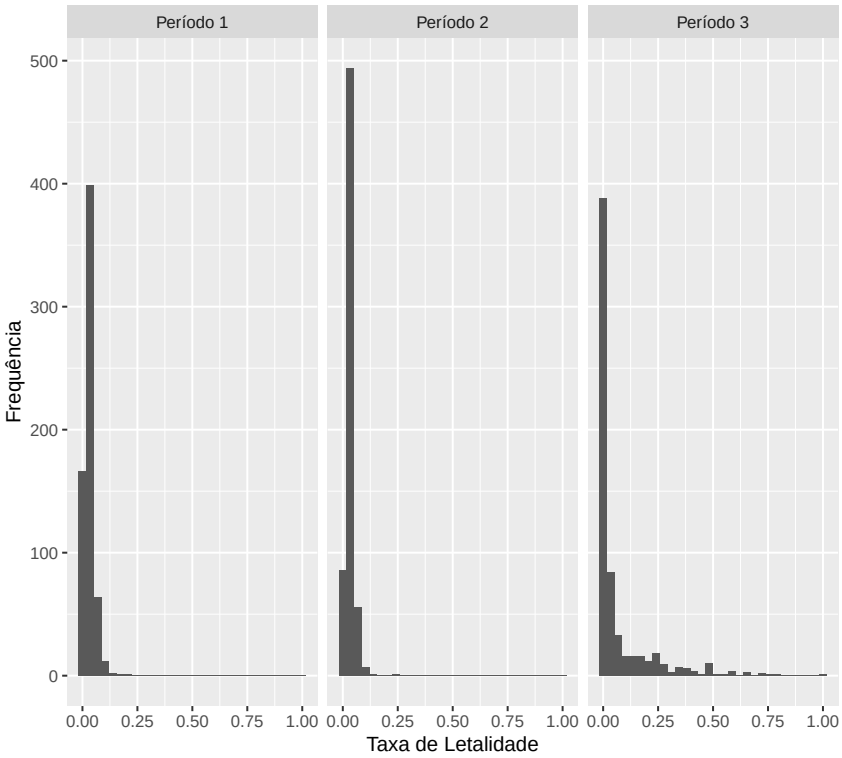
Levando em conta que possui a menor mediana e um CV de 196.382%, o terceiro período tem a maior dispersão de dados; ver a Figura 3. Este comportamento será ainda discutido com o contexto das demais covariáveis, porém é necessário explicitar que o número de pacientes de COVID-19 diminuiu no estado como um todo, o que permite a presença de uma maior variância nos dados.

Tabela 1: Estatísticas sumárias para as taxas de letalidade observadas por período.

Período	média	mediana	DP	CV	CA	CC	min.	max.	n
Período 1	0.030	0.027	0.022	73.234	2.047	12.144	0.000	0.200	645
Período 2	0.032	0.029	0.018	56.493	3.455	31.623	0.002	0.239	645
Período 3	0.070	0.010	0.138	196.382	2.917	12.582	0.000	1.000	636

Fonte: Elaborado pelo autor.

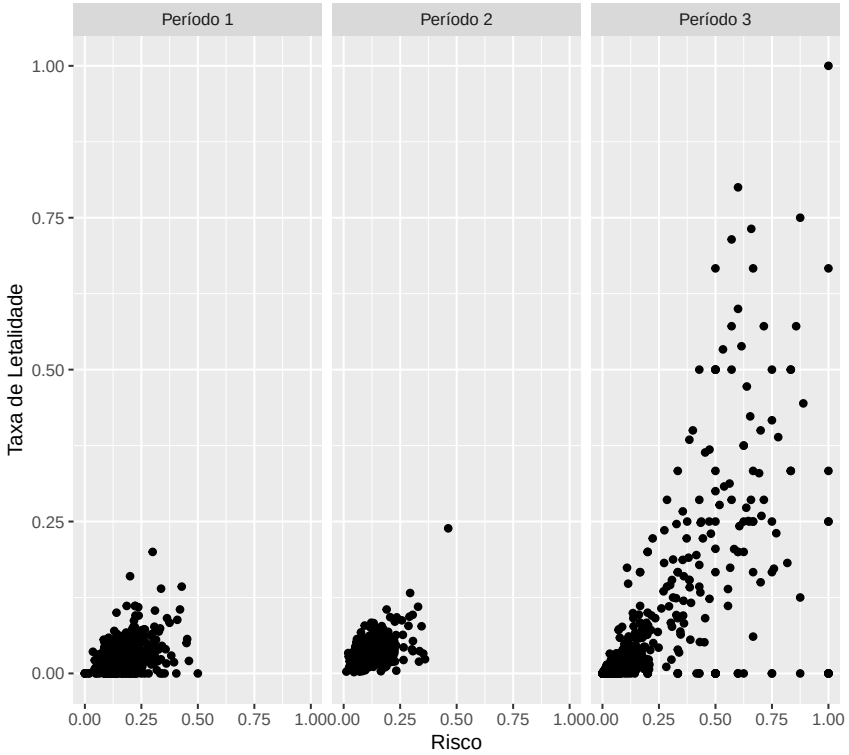
Figura 3: Histograma para as taxas de letalidade observadas.



Fonte: Secretaria de Estado da Saúde de São Paulo (SES-SP).

A Figura 4 apresenta um gráfico de dispersão para os três períodos da taxa de letalidade versus `Risco`. Dessa figura, observa-se uma possível relação positiva entre a proporção de pacientes em risco e a taxa de letalidade. Observa-se também que no terceiro há a dispersão da taxa de letalidade aumenta a medida que o risco aumenta.

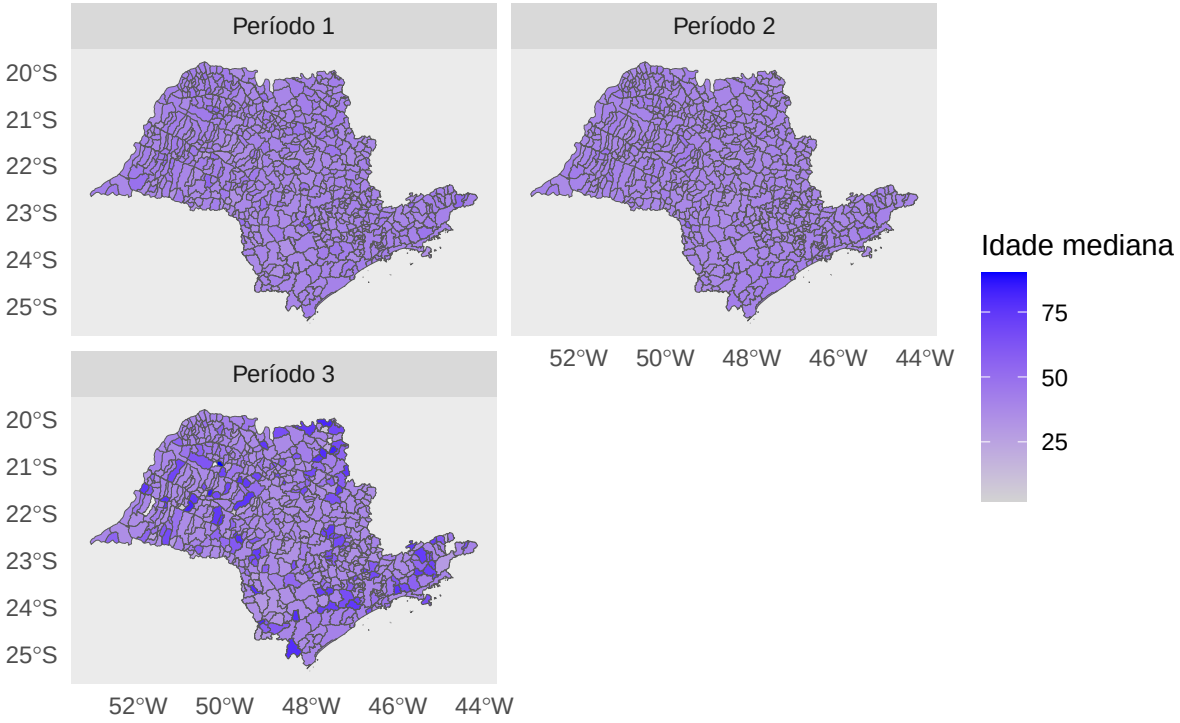
Figura 4: Gráfico de dispersão da taxa de letalidade x `Risco`



Fonte: Elaborado pelo autor.

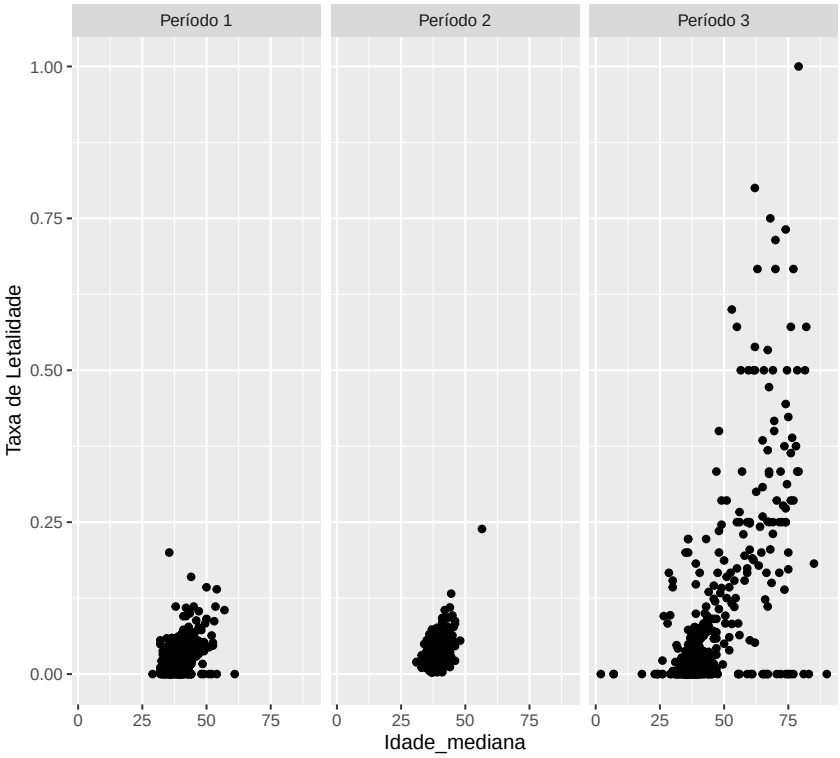
A Figura 5 apresenta o mapa da idade mediana dos pacientes com COVID-19 dos municípios no Estado de São Paulo. A idade mediana dos pacientes de COVID-19 tem o maior número de ocorrência por volta de 39 anos, que se aproxima da idade mediana de 35.7 do Estado de São Paulo de acordo com a Fundação Sistema Estadual de Análise de Dados; ver SEADE (2022). A idade mediana aparenta ter relação positiva com a taxa de letalidade; ver Figura 6. Tal comportamento pode influenciar na presença de taxas de letalidade maiores no Período 3.

Figura 5: Gráfico da idade mediana dos pacientes com COVID-19 no estado de São Paulo.



Fonte: Elaborado pelo autor.

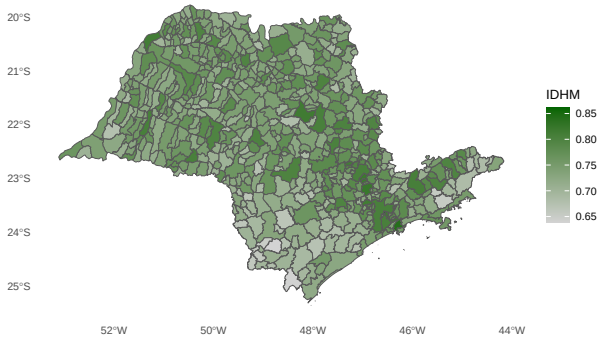
Figura 6: Gráfico de dispersão da taxa de letalidade versus a idade mediana dos pacientes



Fonte: Elaborado pelo autor.

As Figuras 7 e 8 apresentam os mapas com o IDHM e densidade populacional dos municípios do Estado de São Paulo. Dessas figuras, é possível observar que a capital do estado e o seu entorno tem a maior densidade populacional, alguns destes coincidem com os municípios de maior taxa de letalidade na Figura 2. Porém, uma inspeção na Figura 9, não indica nenhum comportamento explícito das covariáveis com a variável resposta.

Figura 7: Gráfico do IDHM dos municípios do Estado de São Paulo.

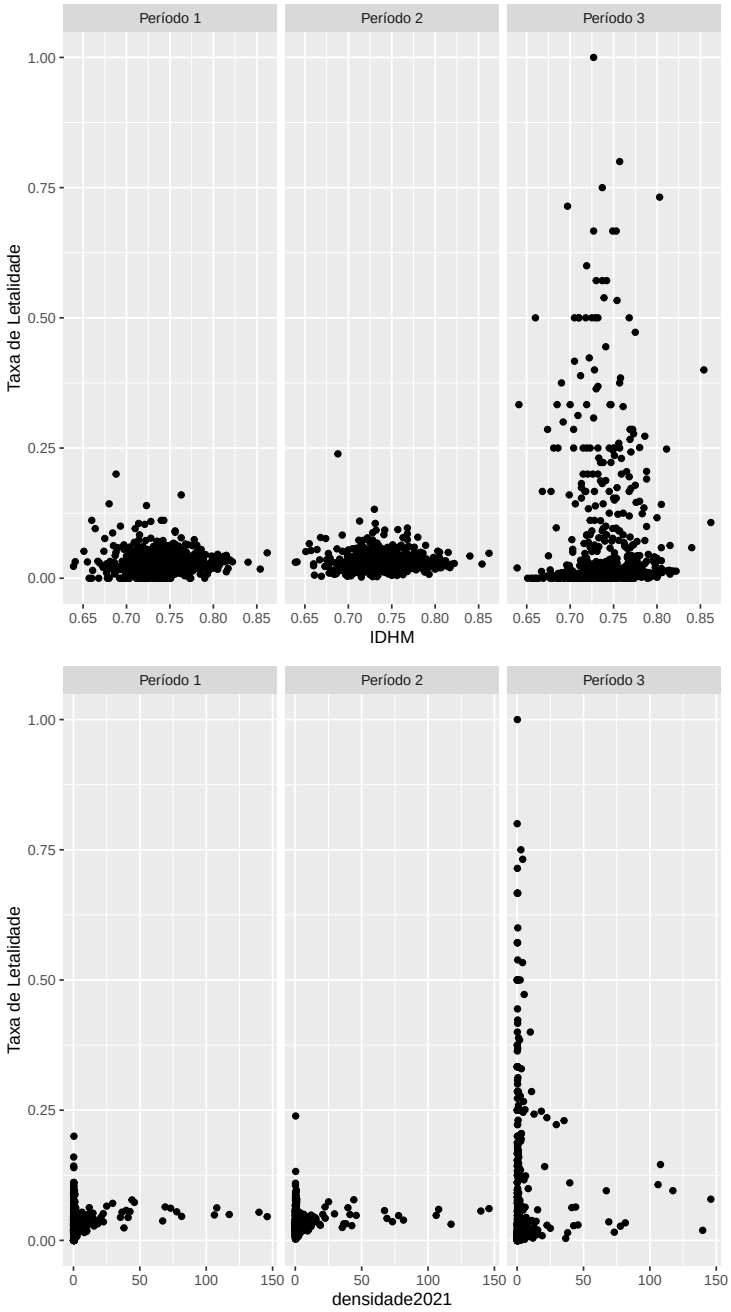


Fonte: Elaborado pelo autor.

Figura 8: Gráfico da densidade populacional dos municípios do Estado de São Paulo.

Fonte: Elaborado pelo autor.

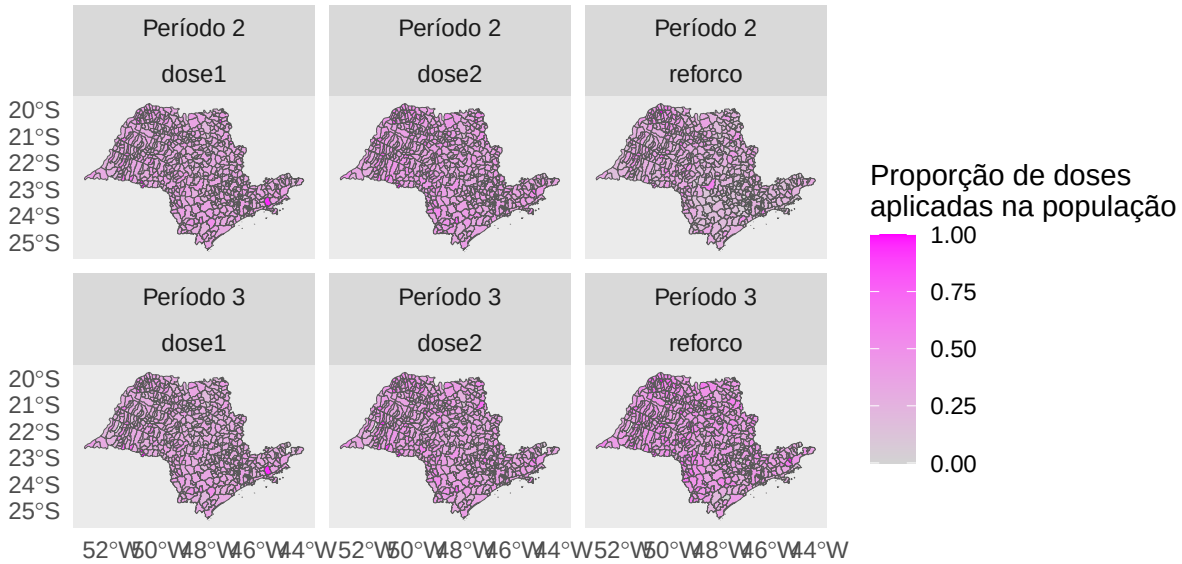
Figura 9: Gráficos de Dispersão do IDHM e da densidade populacional versus a taxa de letalidade dos municípios do Estado de São Paulo.



Fonte: Elaborado pelo autor.

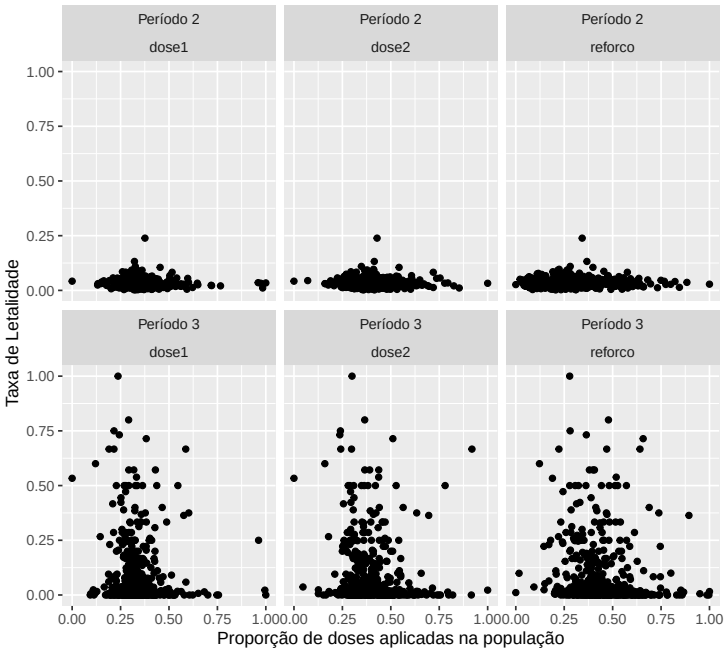
As Figura 10 apresentam os mapas com as doses dos municípios do Estado de São Paulo, e a Figura 11 apresenta os gráficos da dispersão das doses versus a taxa de letalidade. As covariáveis ligadas a aplicação das doses da vacina para COVID-19 foram normalizadas por causa do número de doses em alguns casos estar superando a estimativa da população resultando em proporções maiores que 1.

Figura 10: Gráfico das doses dos municípios do Estado de São Paulo.



Fonte: Elaborado pelo autor.

Figura 11: Gráficos de dispersão das doses versus a taxa de letalidade dos municípios do Estado de São Paulo.



Fonte: Elaborado pelo autor.

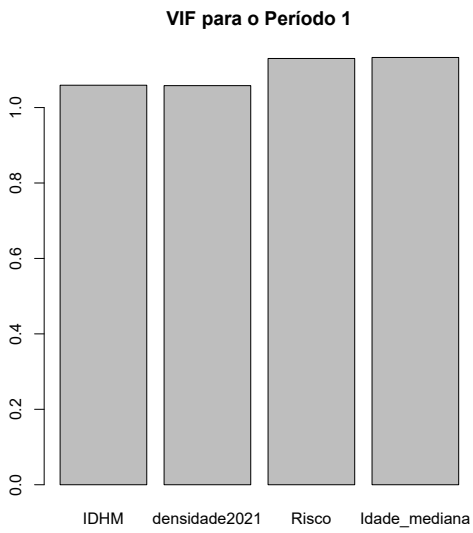
4 Resultados de estimação

Nessa seção, os dados da taxa de letalidade da COVID-19 dos municípios são modelados através do modelo de regressão quantílica logística para dados limitados, introduzido por Bottai, Cai e McKeown (2010) e apresentado na Seção 2. Todos os códigos elaborados para tratamento e modelagem estão em `R` (R Core Team, 2021) e disponíveis em: ⟨https://github.com/Torm198/RegressaoQuantilicaCovid⟩.

4.1 Resultados para Período 1

Como o Período 1 precede o inicio da campanha nacional de vacinação, o modelo não considera nenhuma das variáveis ligadas a esta, já que nenhuma aplicação de dose foi registrada nesse período. Portanto, o modelo escolhido para este período leva conta apenas as variáveis `IDHM`, `densidade2021`, `Risco` e `Idade_mediana`. A Figura 12 apresenta o fator de inflação da variância (VIF) para cada covariável considerada. Como nenhum VIF apresentou um valor muito alto, podemos prosseguir com a escolha do modelo.

Figura 12: Valores VIF para variáveis escolhidas para o modelo do Período 1.



Fonte: Elaborado pelo autor.

Para o modelo final, foram removidas covariáveis que não foram estatisticamente significativas em nenhum valor de quantil $q \in \{0.01, \ldots, 0.99\}$. Temos, então o seguinte

modelo final:

$$Q_{y,\text{logit}}(p) = \beta_{p,0} + \beta_{p,1}\texttt{IDHM} + \beta_{p,2}\texttt{densidade2021} + \beta_{p,3}\texttt{Risco} + \beta_{p,4}\texttt{Idade\_mediana}. \quad (4.1.1)$$

As Figuras 13 e 14 mostram as estimativas dos parâmetros (e p-valores) do modelo de regressão quantílica logística deste período (4.1.1) para $q \in \{0.01, \ldots, 0.99\}$. A Figura 13 também mostra o intervalo de confiança de 95% dos parâmetros, obtido através da estimação do erro padrão de cada pelo método *bootstrap*, e assumindo quantil normal.Observa-se que existem dinâmicas assimétricas com estimativas mudando em q e alguns parâmetros mudando seus sinais. Nota-se também que:

- Os resultados revelam que em grande parte do espectro quantílico o efeito de todas as covariáveis aumenta a taxa de letalidade. Em especial `IDHM` que tem o maior efeito para quantis mais baixos, provavelmente porque municípios mais desenvolvidos receberam mais pacientes de COVID-19 o que leva a uma taxa de letalidade "mínima"mais alta, existe a possibilidade também desse efeito ser positivo devido a variância do coeficiente na região. Quando em quantis mais altos esta variável começa a ter um efeito negativo sob a taxa de letalidade indicando que, municípios com IDHM maior possuem melhor capacidade para lidar com os casos de COVID-19;

- O coeficiente `IDHM` para o quartil q=0.05 é tem o valor estimado aproximadamente de 0.246, em termos de razão de chances ele pode ser interpretado da seguinte maneira: $(\exp(0.246) - 1) * 100 = 28$, ou seja, para cada incremento no IDHM, aumenta-se em 28% as chances de letalidade neste período;

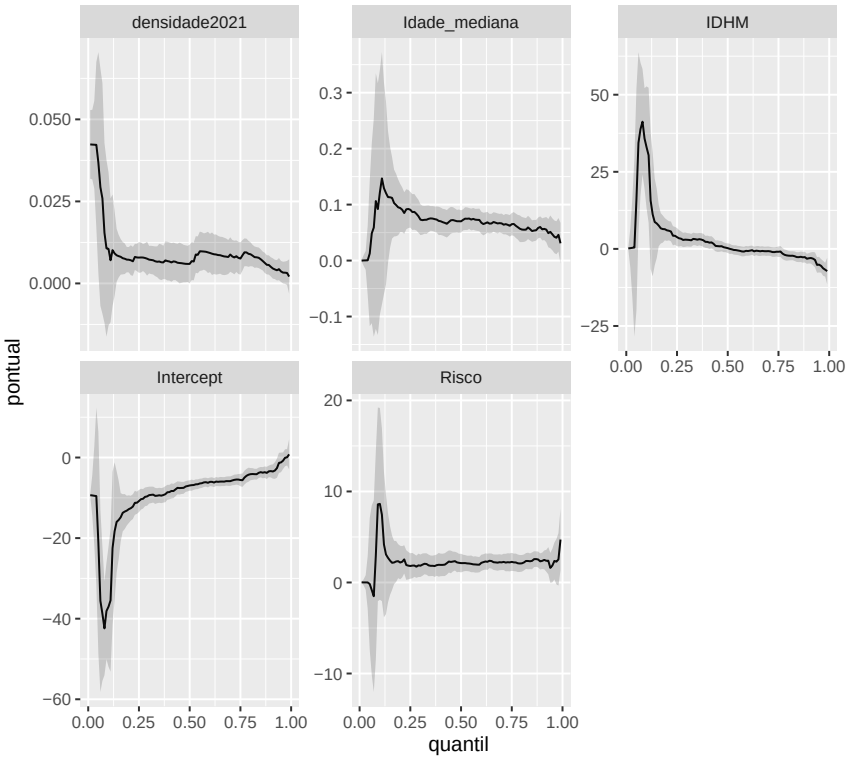
- Um incremento unitário em `Risco` no quartil q=0.95 é esperado que aumente aproximadamente em 1.859 unidades no logito da variável resposta;

- O coeficiente da variável explicativa `Risco` assumiu valor negativo por um breve momento nos quartis iniciais, possivelmente indicando algum tipo de resposta inicial que permite pacientes de risco terem uma chance de sobrevivência maior, contudo ao longo do espectro o coeficiente assume valores positivos demonstrando o comportamento esperado de maior o número de fatores de risco maior a chance de fatalidade;

- Foram observadas nas demais covariáveis um comportamento esperado, a idade mediana dos pacientes e a densidade populacional do município tendo efeito positivo para taxa de letalidade, ou seja, para cada incremento nelas maior a taxa;

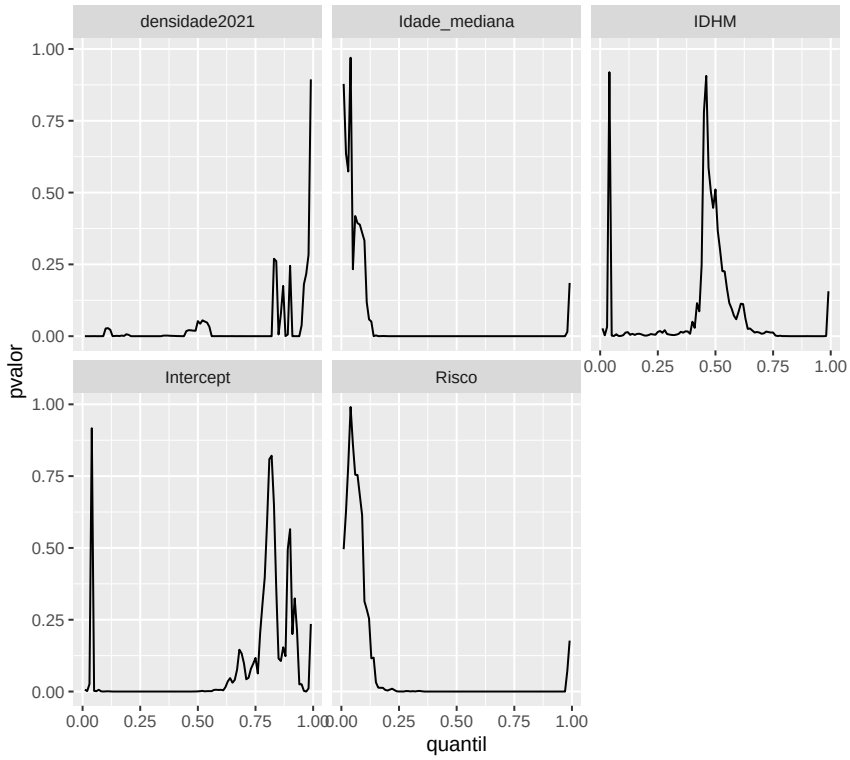
- Uma análise do p-valor indica que os coeficientes das variáveis de IDHM, risco e idade mediana possuem melhor capacidade preditiva para quantis maiores, enquanto a variável de densidade tem melhor estimação para quantis menores. O modelo como um todo aparenta ter melhor capacidade estimativa para quartis perto de 0.5;

Figura 13: Estimativas dos coeficientes com intervalo de confiança de 95% ao longo dos quantis q para o Período 1.



Fonte: Elaborado pelo autor.

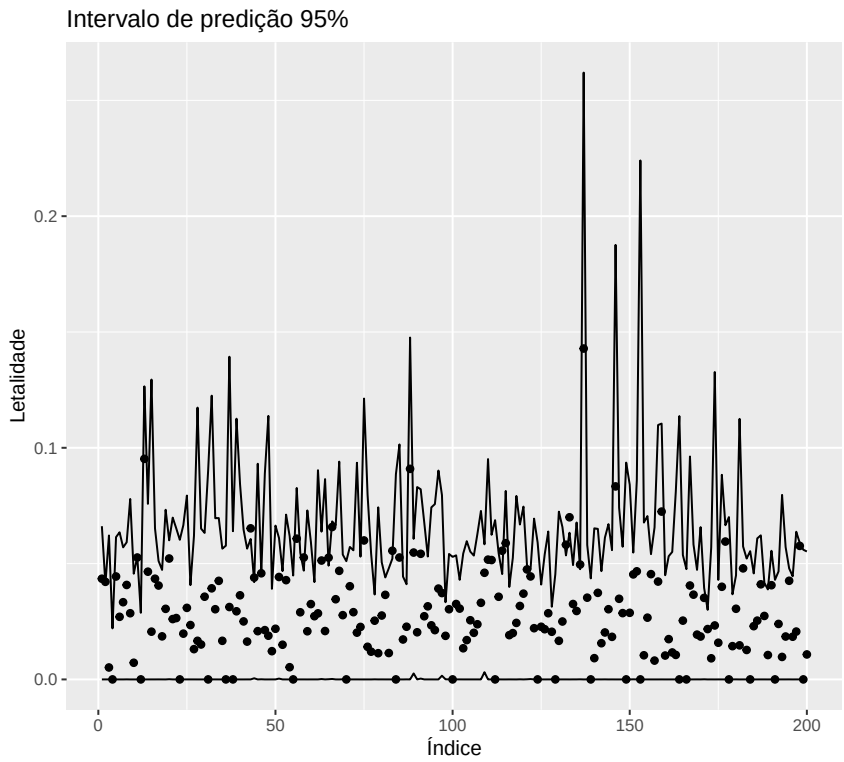
Figura 14: p-valor estimado dos coeficientes ao longo dos quantis q para o Período 1.



Fonte: Elaborado pelo autor.

A Figura 15 apresenta o intervalo de predição de 95% dos modelo para o Período 1. A construção desta é feita da seguinte forma: Para cada valor de $q \in \{0.025, 0.975\}$, os parâmetros do modelo foram estimados e as respectivas estimativas foram utilizadas para obter as previsões de 0.025 e 0.975. As previsões foram realizadas em 200 observações sorteadas do banco de dados, o restante foi utilizado para obter os coeficientes. A Figura 15 mostra que 89.5% das observações estão dentro dos limites do intervalo de predição, muito próximo ao valor nominal.

Figura 15: Intervalos de previsão de 95% para o Período 1.

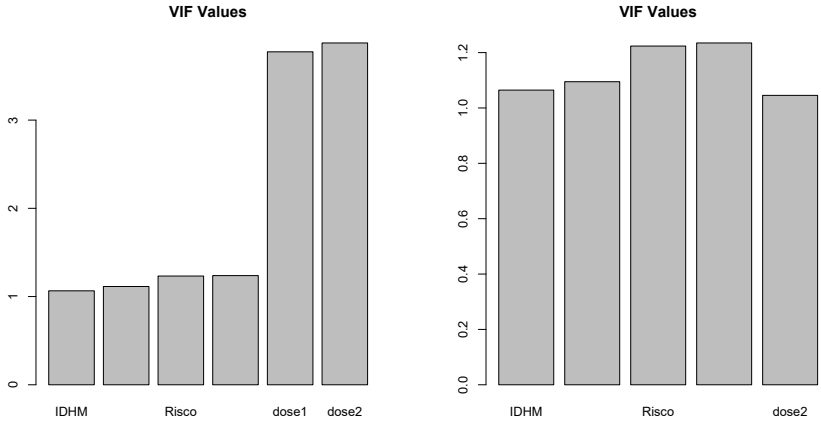


Fonte: Elaborado pelo autor.

4.2 Resultados para Período 2

No segundo período a campanha de vacinação já estava acontecendo logo podemos adicionar as variáveis deste contexto no modelo, como a segunda dose nesta época era considerada o esquema vacinal completo ela irá ser usada no modelo. Ao adicionar primeira dose acaba sendo redundante para o modelo por ter uma alta correlação com a variável da segunda dose resultando num VIF maior para as duas, como ilustrado na Figura 16, ao ser retirada observa se uma melhora geral.

Figura 16: Comparação dos valores VIF para inserção e remoção de `dose1` do Período 2.

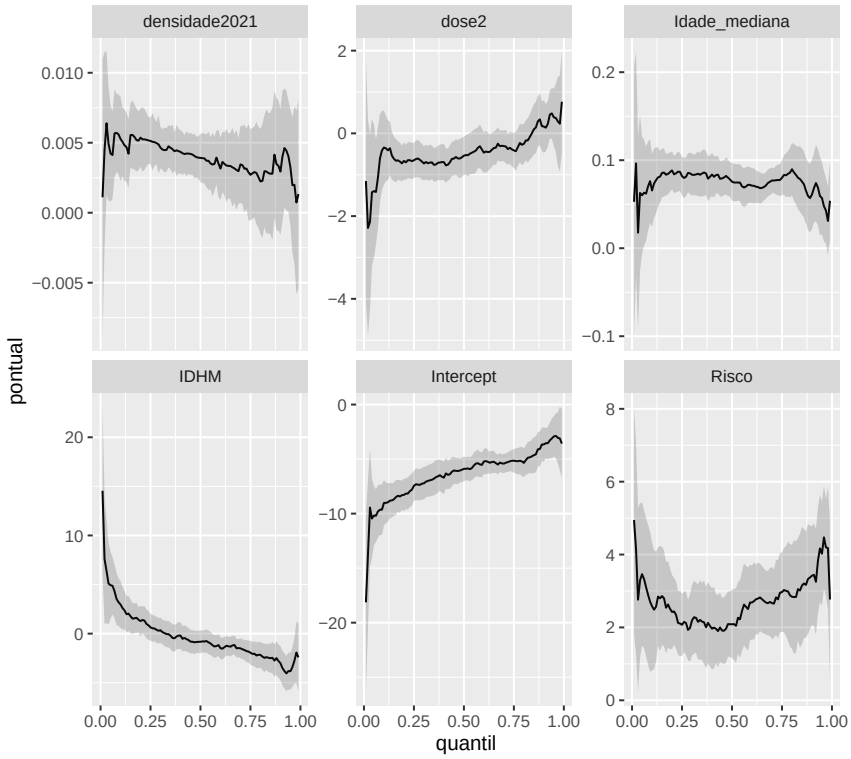


Fonte: Elaborado pelo autor.

Portanto o seguinte modelo foi escolhido para a estimação:

$$Q_{y,\text{logit}}(p) = \beta_{p,0} + \beta_{p,1}\texttt{IDHM} + \beta_{p,2}\texttt{densidade2021} + \beta_{p,3}\texttt{Risco} + \beta_{p,4}\texttt{Idade\_mediana} + \beta_{p,5}\texttt{dose2} \tag{4.2.1}$$

Figura 17: Estimativas dos coeficientes com IC de 95% ao longo dos quantis q para o Período 2.



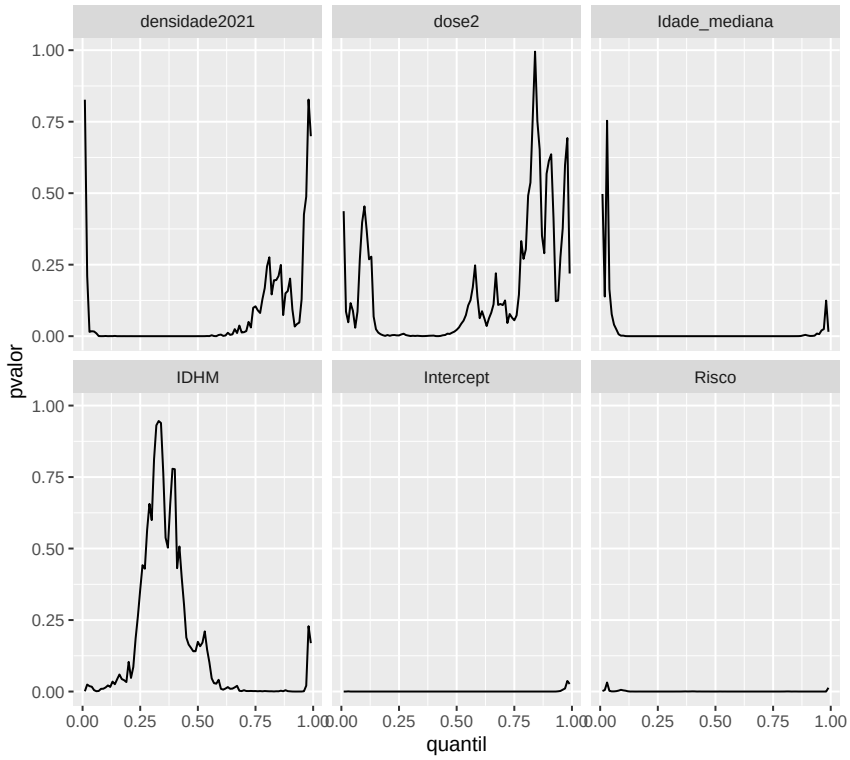
Fonte: Elaborado pelo autor.

Os resultados da estimação dos coeficientes (Figura 17) e suas respectivas estimativas do p-valor (Figura 18) indicam que:

- Como no período anterior o intercepto e os coeficientes atrelados ao IDHM, idade mediana apresentam comportamento similar;
- coeficiente da densidade populacional neste contexto apresenta um característica negativa na extrema direita do espectro que pode ser explicado tanto pela variância, sendo indicado um baixo poder preditivo nessa região, quanto pelo fato das regiões mais densas no estado não apresentarem letalidade mais alta, desta forma apresentando um limite superior menor;
- `dose2` apresenta um comportamento esperado para a vacinação, com um coeficiente agindo de maneira negativa sob a taxa de letalidade em boa parte do espectro. O comportamento positivo para taxa de letalidade pode ser explicado pela possibilidade dos casos e óbitos onde há alta mortalidade serem de pacientes não vacinados em áreas com taxas de aplicação altas, porém sem ainda completar o esquema vacinal de uma parcela da população, assim grande parte das notificações de casos de COVID-19 podem ser de um grupo mais sensível comparado ao devidamente imunizado. Há também uma outra possível explicação: a variância nessa parte do espectro quantílico, que é mais alta, considerando também que a Figura 18 indica um alto p-valor para essa região;
- `dose2` no quantil q=0.05 com um aumento unitário diminui o logito da taxa de letalidade em -1.395 unidades aproximadamente. Visto que o exponencial desse valor é igual aproximadamente a 0.248, a interpretação pela razão de chance é uma redução de 24.8% na chance de letalidade a cada unidade em dose2;
- Um incremento unitário na idade mediana do município, no quantil q=0.95, causa um aumento na logito da taxa de letalidade de 0.057 unidades aproximadamente;
- A análise do p-valor ao longo do espectro quantílico indica uma melhor capacidade preditiva do modelo para prever o limite inferior num intervalo de predição considerando que todos os coeficientes obtiveram um valor significativo quando próximo ao quartil q=0.05. Contudo o modelo apresenta menor capacidade preditiva para os limites superiores visto que boa parte das covariáveis demonstraram p-valor significativo nos quartis de 0.1 e 0.9;

Assim como no Período 1 foi feito uma análise da performance de predição do modelo 4.2.1 para um intervalo de 95% na Figura 19. Mais uma vez foram sorteadas

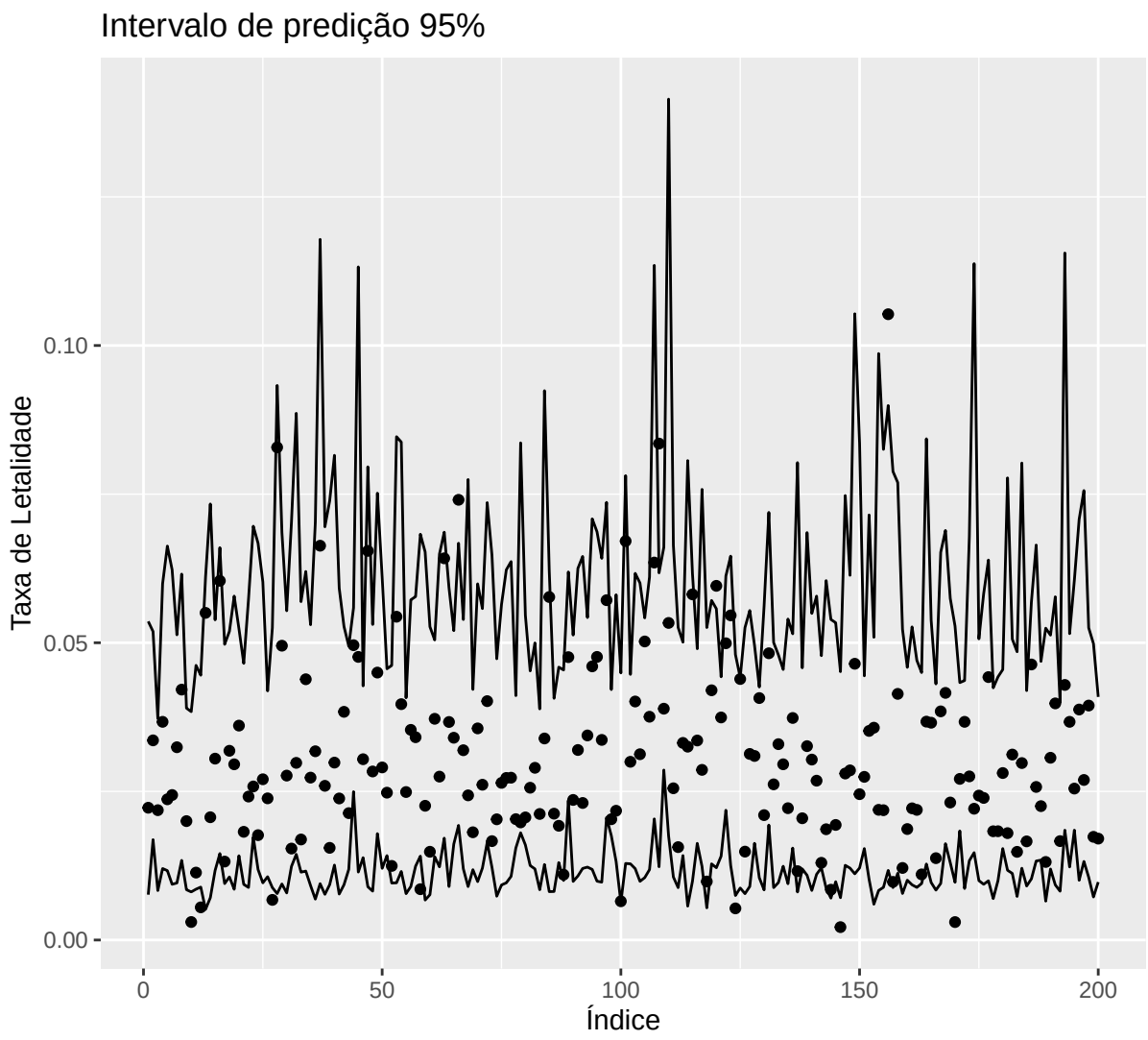
Figura 18: p-valor estimado dos coeficientes ao longo dos quantis q para o Período 2.



Fonte: Elaborado pelo autor.

45 observações do banco de dados para compor o banco para a validação. A Figura 19 mostra o desempenho do modelo para o segundo período 4.2.1, obteve 94.5% das observações dentro intervalo, valor muito próximo do esperado do intervalo.

Figura 19: Performance do modelo no banco de validação do Período 2.

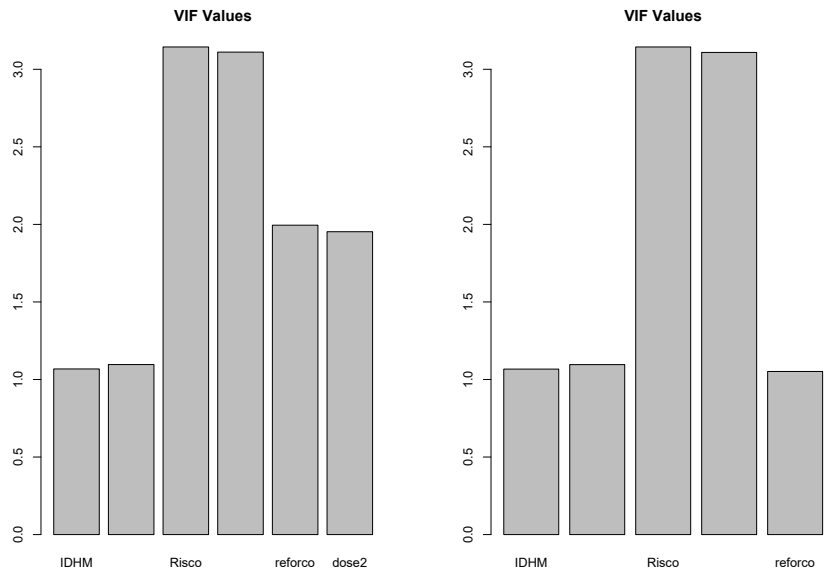


Fonte: Elaborado pelo autor.

4.3 Resultados para Período 3

Para modelo final foi optado por retirar a covariável `dose2` em favor da `reforco`, por minimizar em um modo geral o VIF e a redundância, como a Figura 20 demonstra, além de ser mais relevante para o contexto deste período. Apesar dos VIF maiores as variáveis de risco e idade mediana serão mantidas por terem apresentado grande consistência nos períodos anteriores.

Figura 20: Comparação dos valores VIF incluindo e removendo `dose2` do Período 3.



Fonte: Elaborado pelo autor.

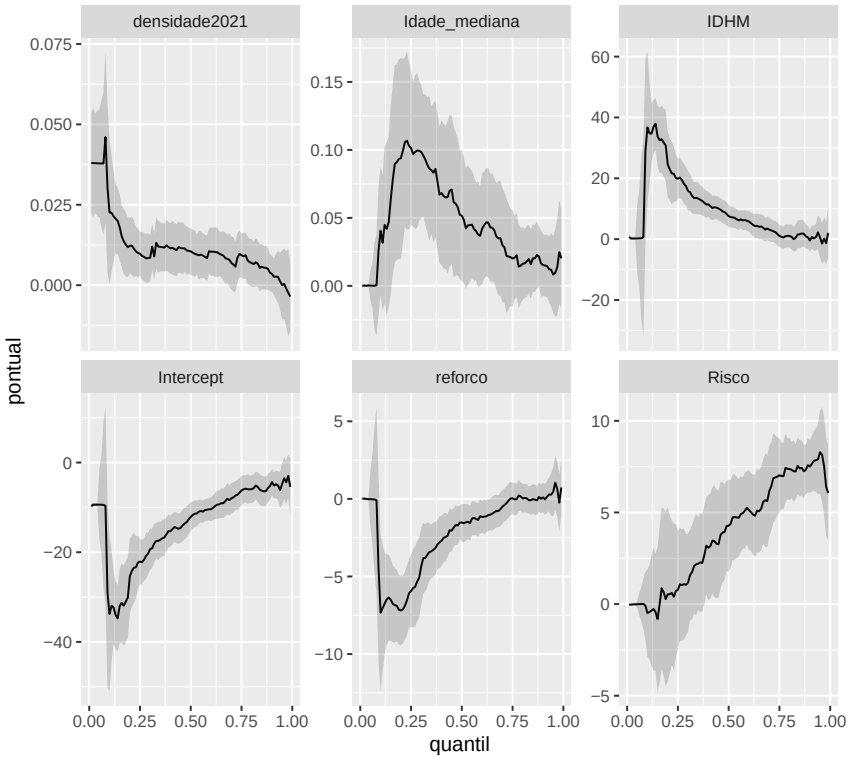
Portanto o modelo escolhido será:

$$Q_{y,\text{logit}}(p) = \beta_{p,0} + \beta_{p,1}\texttt{IDHM} + \beta_{p,2}\texttt{densidade2021} + + \beta_{p,3}\texttt{Risco} + \beta_{p,4}\texttt{Idade\_mediana} + \beta_{p,5}\texttt{reforco} \tag{4.3.1}$$

Nas Figuras 21 e 22

- Considerando a alta variância nos dados devido ao número menor de casos de COVID-19 no banco de dados para este período, todos os intervalos de predição dos coeficiente estão maiores;
- Mesmo assim algumas variáveis apresentaram um comportamento esperado como `reforco` tendo um efeito negativo à taxa de letalidade, e as variáveis `Risco` e `idade_mediana` apresentarem um efeito positivo em relação a variável resposta;
- O intercepto, o IDHM e a densidade populacional demonstraram desenvolvimento similar aos casos passados. No caso do intercepto possuir um valor negativo no início do espectro e aumentar o seu valor mais ao final, já o IDHM e a densidade populacional mostrarem efeitos positivos no início e decrescerem ao longo dos quartis, para densidade chegar a obter efeito negativo;
- O aumento de cada unidade em `reforco` em q=0.1 causa uma variação no logito da taxa de letalidade de -7.33 unidades, aproximadamente. A interpretação pela razão

Figura 21: Estimativas dos coeficientes com IC de 95% ao longo dos quantis q para o Período 3.



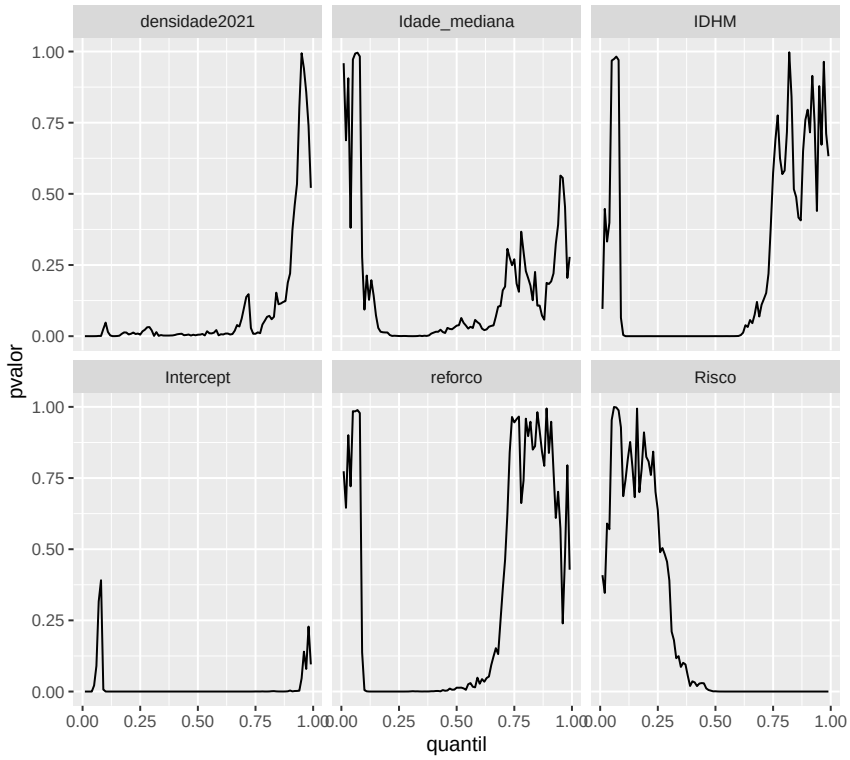
Fonte: Elaborado pelo autor.

de chances desse coeficiente é uma redução de aproximadamente 0.07% na chance de letalidade;

- Para `Risco`, no quartil q=0.9, o aumento unitário provoca um aumento aproximado de 7.5 no logito da taxa de letalidade;

- O p-valor nesse caso indica que o modelo possui baixa capacidade preditiva para quartis mais extremos, boa parte das variáveis tem perda abaixo de q=0.1 e acima de q=0.9 a um nível não desejável. Esse fenômeno pode ser explicado pela dispersão maior nos dados já comentada anteriormente. Apesar disso para quartis mais centrais, perto de q=0.5 ainda o modelo mantém sua capacidade preditiva;

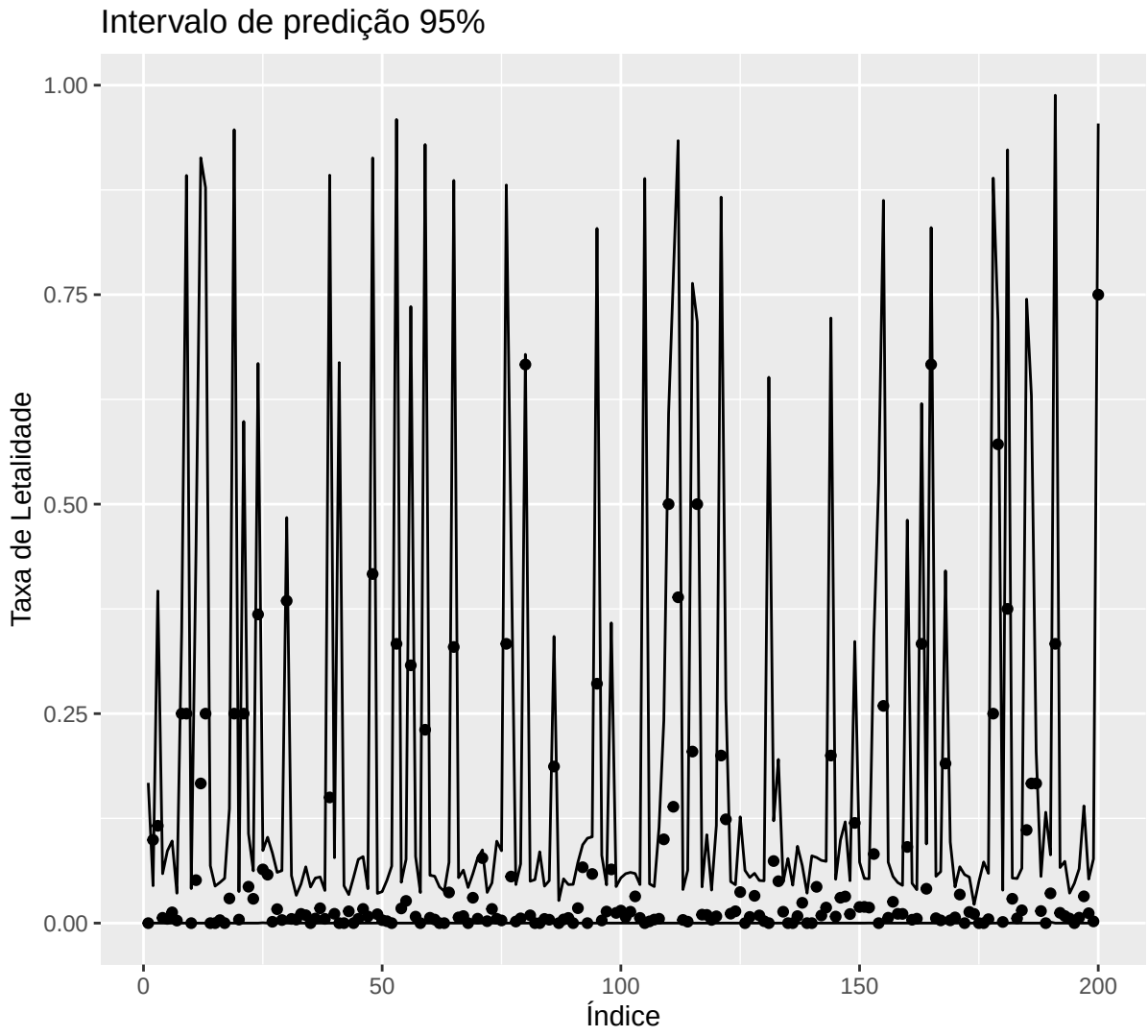
Figura 22: p-valor estimado dos coeficientes ao longo dos quantis q para o Período 3.



Fonte: Elaborado pelo autor.

Finalmente, como nas Figuras 15 e 19, a Figura 23 mostra a capacidade preditiva a partir de um banco criado a partir do sorteio de 200 observações. O modelo 4.3.1 teve uma performance de 97.5% das variáveis contidas dentro do intervalo de 95%. A alta variabilidade deste período pode ser observada nos intervalos estimados de maior amplitude.

Figura 23: Performance do modelo no banco de validação do Período 3.



Fonte: Elaborado pelo autor.

5 Conclusão

Este trabalho aplicou o modelo de regressão quantílica logística no contexto da COVID-19 no contexto dos municípios de São Paulo. Apesar de problemas com os dados para o terceiro período os resultados para os demais foram satisfatórios a ponto de permitir uma previsão, suficientemente precisa, para a taxa de letalidade dado o contexto inserido em São Paulo.

Os intervalos obtidos mostram também um retrato de possíveis situações para determinadas condições, permitindo uma análise preditiva, indicando o pior e o melhor caso possível para cada contexto.

Os resultados atuais indicam possibilidade de outros estudos, considerando que foram obtidos antes do término da pandemia. Há espaço para analises longitudinais mais robustas em relação a volatilidade da variável estudada, e também a alternativa do aumento da escala do projeto, sendo factível a inserção de outros estados brasileiros na análise. O tópico da COVID-19 inspira uma série de trabalhos.

Referências

BOTTAI, M.; CAI, B.; MCKEOWN, R. E. Logistic quantile regression for bounded outcomes. *Statistics in Medicine*, v. 29, n. 2, p. 309–317, 2010.

DAVINO, C.; FURNO, M.; VISTOCCO, D. *Quantile Regression*. Chichester, UK: Wiley, 2014.

DAVINO, C.; ROMANO, R.; VISTOCCO, D. Handling multicollinearity in quantile regression through the use of principal component regression. *METRON*, Springer Science and Business Media LLC, fev. 2022. Disponível em: ⟨https://doi.org/10.1007/s40300-022-00230-3⟩.

FERRARI, S.; CRIBARI-NETO, F. Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, v. 31, p. 799–815, 2004.

KOENKER, R.; Bassett Jr, G. Regression quantiles. *Econometrica*, v. 46, p. 33–50, 1978.

KOENKER, R.; D'OREY, V. Computing regression quantiles. *Applied Statistics*, v. 36, p. 383–393, 1987.

KOENKER, R.; D'OREY, V. Computing regression quantiles. *Applied Statistics*, v. 43, p. 410–414, 1994.

KOENKER, R.; MACHADO, J. A. F. Goodness of fit and related inference processes for quantile regression. *Journal of the American Statistical Association*, Taylor & Francis, v. 94, n. 448, p. 1296–1310, 1999. Disponível em: ⟨https://www.tandfonline.com/doi/abs/10.1080/01621459.1999.10473882⟩.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2021. Disponível em: ⟨https://www.R-project.org/⟩.

SEADE. População paulista cresceu 20% em 20 anos. 2022. Disponível em: ⟨https://www.seade.gov.br/populacao-paulista-cresceu-20-em-20-anos/⟩.

Apêndice

Tabela 2: Estimativas pontuais dos coeficientes em alguns quantis para os três períodos

Período	Coeficiente	0.05	0.25	0.5	0.75	0.95
Período 1	Intercepto	-21.18	-10.80	-6.94	-5.58	-1.21
	IDHM	15.98	3.81	0.25	-0.93	-5.11
	densidade2021	0.04	0.01	0.01	0.01	0.00
	Risco	-0.17	1.81	2.13	2.26	1.86
	Idade_mediana	0.01	0.09	0.07	0.07	0.05
Período 2	Intercepto	-10.17	-7.45	-5.89	-5.15	-2.92
	IDHM	4.94	0.64	-0.83	-1.94	-3.83
	densidade2021	0.00	0.01	0.00	0.00	0.00
	Risco	3.46	2.07	2.09	2.97	4.02
	Idade_mediana	0.06	0.08	0.08	0.08	0.06
	dose2	-1.39	-0.65	-0.54	-0.41	0.39
Período 3	Intercepto	-9.39	-22.08	-12.10	-6.22	-4.77
	IDHM	0.23	19.85	7.59	1.07	0.51
	densidade2021	0.04	0.01	0.01	0.01	0.00
	Risco	-0.01	0.78	4.27	7.03	8.29
	Idade_mediana	0.00	0.10	0.05	0.02	0.01
	reforco	-0.02	-5.94	-1.52	0.04	0.45

Fonte: Elaborado pelo autor.