Proceeding Series of the Brazilian Society of Computational and Applied Mathematics, v. 12, n. 1, 2026.

Trabalho apresentado no XLIV CNMAC, Fundação Getulio Vargas - Rio de Janeiro - RJ, 2025.

Proceeding Series of the Brazilian Society of Computational and Applied Mathematics

Análise de Regressão Linear Múltipla para Identificar Fatores Socioeconômicos e Sanitários Associados à Dispersão da Covid-19 no Brasil

Pedro H. A. G. Moura[1], Regina S. Lanzillotti[2]
PPG-CompMat, IME–UERJ, Rio de Janeiro, RJ

Resumo. Este estudo utiliza a regressão linear múltipla para analisar os fatores que influenciaram a disseminação e a letalidade da Covid-19 no Brasil. Foram consideradas variáveis como renda per capita, população, taxa de analfabetismo e acesso à água e esgoto, avaliando-se seus impactos nos acometimentos e óbitos pela doença em cada Unidade Federativa (UF) e diferenciando os efeitos entre áreas urbanas e rurais. Os modelos foram ajustados utilizando a técnica dos Mínimos Quadrados Ordinários (OLS – *Ordinary Least Squares*) por meio da biblioteca `statsmodels` da linguagem Python. Os resultados indicam que a população foi o fator mais relevante para explicar a propagação do vírus, tanto em áreas urbanas quanto rurais. Além disso, a taxa de analfabetismo e a renda per capita apresentaram impacto significativo em algumas das análises, enquanto as variáveis de saneamento mostraram menor influência na maioria das aplicações.

Palavras-chave. regressão linear, Covid-19, análise inferencial estatística

1 Introdução

A pandemia de Covid-19 evidenciou disparidades na saúde pública, associadas a fatores como saneamento, escolaridade e renda. Modelos estatísticos, como a regressão linear múltipla, têm sido amplamente utilizados para quantificar essas relações e orientar políticas públicas.

Diversos estudos apontaram a influência de fatores socioeconômicos na disseminação da Covid-19. Aquino [[1]] relacionou precariedades de saneamento a maiores taxas de casos e óbitos. Nascimento [1[10]], por meio de equações estruturais, identificou que melhores condições de saneamento reduzem a incidência, embora o efeito da renda varie conforme o contexto. Martins [[8]] também observou maior incidência em áreas com infraestrutura básica deficiente.

A aplicação de modelos inferenciais avançados permite identificar e quantificar a influência de múltiplos fatores socioeconômicos na evolução da pandemia.

Diante desse cenário, este trabalho tem como objetivo aplicar a regressão linear múltipla para identificar os fatores socioeconômicos e sanitários mais associados à propagação e à mortalidade da Covid-19 no Brasil, diferenciando os efeitos entre áreas urbanas e rurais.

Para o desenvolvimento deste artigo, foram utilizados dados do Ministério da Saúde [11], nos períodos de 27/03/2020 a 04/01/2025; do Instituto Brasileiro de Geografia e Estatística (IBGE), referentes ao Censo Demográfico de 2023 [5, 6]; e do Painel Saneamento Brasil do Instituto Trata Brasil [2], do ano de 2021. Através da análise de regressão linear múltipla, este artigo busca verificar se as variáveis independentes do modelo eram significativamente preditivas com base nos resultados segundo Análise de Variância (ANOVA).

[1]pedro.avellar@pos.ime.uerj.br
[2]reginalanzillotti@ime.uerj.br

2 Metodologia

Este estudo investiga se as cinco variáveis independentes do modelo regressivo - renda per capita, população, taxa de analfabetismo e acesso à água e esgoto - que poderiam possuir um impacto significativo de acordo com os parâmetros regressivos e validação segundo a Análise de Variância (ANOVA).

A ANOVA é uma ferramenta estatística amplamente utilizada para avaliar a significância de modelos de regressão. O teste ANOVA de regressão divide a variação total dos dados em duas componentes principais: a variação explicada pelo modelo (SQReg) e a variação residual (SQRes). A estatística F, obtida a partir dessas componentes, permite testar a hipótese nula de que todos os coeficientes de regressão (exceto o intercepto) são iguais a zero, indicando se as variáveis preditoras contribuem significativamente para explicar a variável dependente [4, 9].

Matematicamente, a estatística F é calculada como:

$$F = \frac{\text{SQReg}/\text{GLReg}}{\text{SQRes}/\text{GLRes}} \tag{1}$$

onde SQReg e SQRes são as somas de quadrados da regressão e do resíduo, respectivamente, e GLReg e GLRes representam seus graus de liberdade correspondentes. Se o p-valor associado à estatística F for inferior a um nível de significância pré-definido (geralmente 0,05), rejeita-se a hipótese nula, sugerindo que pelo menos uma das variáveis independentes tem efeito significativo sobre a variável resposta [3, 7].

Esse método é amplamente utilizado em pesquisas para validar os modelos preditivos e garantir a robustez das inferências estatísticas [13]. Para identificar outliers, aplicamos a distância de Mahalanobis [12].

A equação geral do modelo é dada por:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_n X_n + \varepsilon, \tag{2}$$

onde Y é a variável dependente (resposta); X_i são as variáveis independentes (preditoras); β_0 é o intercepto da regressão; β_i são os coeficientes de regressão e ε representa o erro aleatório.

O ajuste dos modelos foi realizado através da técnica dos Mínimos Quadrados Ordinários (OLS – *Ordinary Least Squares*) por meio da biblioteca `statsmodels` da linguagem Python. Esta abordagem permite estimar os coeficientes dos modelos com base na minimização da soma dos quadrados dos resíduos, sendo amplamente empregada em análises estatísticas devido à sua simplicidade e eficiência computacional. Os indicadores de ajuste, como R^2 , R^2 ajustado, valores dos coeficientes, intervalos de confiança e p-valores, foram obtidos diretamente a partir da saída dos modelos estimados.

3 Resultados

Nesta seção, são apresentadas as estimativas dos casos de acomentimento e óbitos por Covid-19, obtidos a partir da aplicação do modelo de regressão linear múltipla segundo os fatores socioeconômicos e sanitários associados à propagação da Covid-19 no Brasil, referente aos casos de acometimento tanto na área rural quanto urbana.

As Tabelas 1 e 2 referem-se aos casos de acometimento por Covid-19, enquanto as Tabelas 3 e 4 apresentam os dados relacionados aos óbitos causados pela doença.

DOI: 10.5540/03.2026.012.01.0250

Tabela 1: Regressão Linear Múltipla dos Acometimentos por Covid-19

Especificação	Coeficientes			Intervalo de Confiança Coeficientes	
	β	Erro Padrão	p-valor	0.025	0.975
Área Rural					
Renda Per Capita	81.9065	56.202	**0.160**	-34.972	198.785
População Rural	0.2371	0.016	0.000	0.203	0.271
Analfabetos	-1.7821	0.646	0.012	-3.125	-0.439
População Rural Sem Água	0.0285	0.242	**0.907**	-0.474	0.531
População Rural Sem Esgoto	0.1657	0.250	**0.515**	-0.355	0.686
R^2**: 0.971**					
R^2 **Ajustado: 0.964**					
R**: 0.985**					
Teste F: 140.1					
Prob(TesteF): $\mathbf{2.17}e^{-15}$					
Área Urbana					
Renda Per Capita	125.5435	30.699	0.000	61.878	189.209
População Urbana	0.1189	0.010	0.000	0.098	0.140
Analfabetos	-0.5796	0.367	**0.128**	-1.340	0.181
População Urbana Sem Água	-0.0412	0.073	**0.577**	-0.192	0.110
População Urbana Sem Esgoto	0.0349	0.026	**0.185**	-0.018	0.088
R^2**: 0.966**					
R^2 **Ajustado: 0.958**					
R**: 0.982**					
Teste F: 124.6					
Prob(TesteF): $\mathbf{2.24}e^{-15}$					

A aplicação do modelo linear regressivo apresentou um bom ajuste, com coeficiente de adequação $R^2 = 0.971$ para a área rural e $R^2 = 0.966$ para a área urbana, correspondendo a coeficientes de correlação múltipla de 0.985 e 0.982, respectivamente. Dessa maneira, a regressão múltipla para o modelo regressivo dos casos de acometimento de Covid-19 torna-se:

(a) na área rural,

$$\begin{aligned}\textbf{Acometimentos de Covid} = {} & 81.9065\ \text{Renda Per Capita} + 0.2371\ \text{População Rural} \\ & - 1.7821\ \text{Analfabetos} + 0.0285\ \text{População Rural sem Água} \\ & + 0.1657\ \text{População Rural sem Esgoto}\end{aligned} \tag{3}$$

(b) na área urbana,

$$\begin{aligned}\textbf{Acometimentos de Covid} = {} & 125.5435\ \text{Renda Per Capita} + 0.1189\ \text{População Urbana} \\ & - 0.5796\ \text{Analfabetos} - 0.0285\ \text{População Urbana sem Água} \\ & + 0.0349\ \text{População Urbana sem Esgoto}\end{aligned} \tag{4}$$

Na área rural, apenas a população rural e o analfabetismo foram estatisticamente significantes com níveis de significância de 0.000 e 0.012. Na área urbana, a significância do modelo é explicada pela renda per capita e a população urbana, com p-valores 0.000 em ambas as variáveis.

Dessa forma, foi tratada a regressão linear com apenas as variáveis estatisticamente significantes com a remoção das variáveis, renda per capita e saneamento na área rural e, analfabetos e saneamento na área urbana. Os coeficientes de determinação encontrados foram $R^2 = 0.962$ e $R^2 = 0.961$, respectivamentes nestas localidades, consequentemente, os coeficientes de correlação múltipla apresentaram pequena variação, passando para 0.981 e 0.980, indicando que as associações praticamente não sofreram alterações, Tabela 2.

Tabela 2: Regressão Linear Múltipla dos Acometimentos de Covid-19 Removidas as Variáveis Não Significantes

Especificação	Coeficientes			Intervalo de Confiança Coeficientes	
	β	Erro Padrão	p-valor	0.025	0.975
Área Rural					
População Rural	0.2545	0.015	0.000	0.224	0.285
Analfabetos	-1.1427	0.258	0.000	-1.676	-0.610
R^2**: 0.962**					
R^2 **Ajustado: 0.958**					
R**: 0.981**					
Teste F: 300.6					
Prob(TesteF): $\mathbf{1.02}e^{-17}$					
Área Urbana					
Renda Per Capita	137.3102	24.318	0.000	87.226	187.394
População Urbana	0.1074	0.007	0.000	0.093	0.122
R^2**: 0.961**					
R^2 **Ajustado: 0.958**					
R**: 0.980**					
Teste F: 306.4					
Prob(TesteF): $\mathbf{2.61}e^{-18}$					

Os casos de acometimento de Covid-19 tornaram-se:

(a) na área rural,

$$\textbf{Acometimentos de Covid} = 0.2545 \text{ População Rural } - 1.1427 \text{ Analfabetos} \quad (5)$$

(b) na área urbana:

$$\textbf{Acometimentos de Covid} = 137.3102 \text{ Renda Per Capita } + 0.1074 \text{ População Urbana} \quad (6)$$

De maneira similar, foi reaplicado o modelo regressivo multivariado para avaliar que variáveis tiveram maior impacto na ocorrência de óbitos por Covid-19 e os resultados estão na Tabela 3.

Os coeficientes de determinação foram $R^2 = 0.995$ e $R^2 = 0.993$ para a área rural e urbana, respectivamente, indicando correlações múltiplas de 0.997 e 0.996. Os resultados mostraram que na área rural, apenas a população rural possui significância para o modelo, mas a renda per

capita e analfabetismo apresentaram p-valores 0.066 e 0.079 e foram considerados que faziam parte do modelo. Na área urbana, tanto a população quanto o analfabetismo foram estatisticamente significativas, com p-valores de 0.000 e 0.007, respectivamente. Assim, as equações regressivas assumiram as feições:

(a) na área rural,

$$\begin{aligned}\textbf{Óbitos por Covid} = {} & 0.7996 \text{ Renda Per Capita } + 0.0045 \text{ População Rural} \\ & - 0.0084 \text{ Analfabetos } + 0.00329 \text{ População Rural sem Água} \\ & - 0.0030 \text{ População Rural sem Esgoto}\end{aligned} \tag{7}$$

(b) na área urbana:

$$\begin{aligned}\textbf{Óbitos por Covid} = {} & 0.1246 \text{ Renda Per Capita } + 0.0043 \text{ População Urbana} \\ & - 0.0142 \text{ Analfabetos } - 5.671e^{-5} \text{ População Urbana sem Água} \\ & + 0.0004 \text{ População Urbana sem Esgoto}\end{aligned} \tag{8}$$

Tabela 3: Regressão Linear Múltipla dos Óbitos por Covid-19

Especificação	Coeficientes			Intervalo de Confiança Coeficientes	
	β	Erro Padrão	p-valor	0.025	0.975
Área Rural					
Renda Per Capita	0.7669	0.396	0.066	-0.056	1.590
População Rural	0.0045	0.000	0.000	0.004	0.005
Analfabetos	-0.0084	0.005	0.079	-0.018	0.001
População Rural Sem Água	-0.00329	0.002	**0.100**	-0.006	0.001
População Rural Sem Esgoto	-0.0030	0.002	**0.102**	-0.007	0.001
R^2**: 0.995**					
R^2 **Ajustado: 0.994**					
R**: 0.997**					
Teste F: 800.1					
Prob(TesteF): $\mathbf{3.28}e^{-23}$					
Área Urbana					
Renda Per Capita	0.1246	0.399	**0.758**	-0.703	0.952
População Urbana	0.0043	0.000	0.000	0.004	0.005
Analfabetos	-0.0142	0.005	0.007	-0.024	-0.004
População Urbana Sem Água	$-5.671e^{-5}$	0.001	**0.953**	-0.002	0.002
População Urbana Sem Esgoto	0.0004	0.000	**0.290**	0.000	0.001
R^2**: 0.993**					
R^2 **Ajustado: 0.991**					
R**: 0.996**					
Teste F: 626.6					
Prob(TesteF): $6.06e^{-23}$					

Foi aplicada a regressão para avaliar os resultados após a retirada das variáveis de saneamento básico de ambas as regressões e da renda per capita da área urbana, Tabela 4. Após a remoção

DOI: 10.5540/03.2026.012.01.0250

das variáveis de menor significância, os modelos apresentam coeficientes de adequação e correlação bastante similares aos da Tabela 3. Contudo, a renda per capita perde a signifância na área rural, pois o p-valor igualou-se a 0.982, o que a levou a ser removida nesta área geográfica. Ressalta-se que os coeficientes das demais variáveis permaneceram os mesmos, bem como os coeficientes de adequação e correlação, indicando que esta variável não impactava no modelo. Por outro lado, o modelo da área urbana mostrou um bom ajuste, não havendo necessidade de reaplicá-lo.

Tabela 4: Regressão Linear Múltipla dos Óbitos por Covid-19 Final

Especificação	Coeficientes			Intervalo de Confiança Coeficientes	
	β	Erro Padrão	p-valor	0.025	0.975
Área Rural					
População Rural	0.0046	0.000	0.000	0.004	0.005
Analfabetos	-0.0257	0.002	0.000	-0.030	-0.021
R^2**: 0.990**					
R^2 **Ajustado: 0.989**					
R**: 0.995**					
Teste F: 1196					
Prob(TesteF): 9.20e^{-25}					
Área Urbana					
População Urbana	0.0043	0.000	0.000	0.004	0.005
Analfabetos	-0.0093	0.003	0.008	-0.016	-0.003
R^2**: 0.992**					
R^2 **Ajustado: 0.992**					
R**: 0.996**					
Teste F: 1591					
Prob(TesteF): 4.44e^{-27}					

Assim, as regressões lineares tornaram-se:

(a) na área rural,

$$\textbf{Óbitos por Covid} = 0.0046 \text{ População Rural } - 0.0257 \text{ Analfabetismo} \tag{9}$$

(b) na área urbana:

$$\textbf{Óbitos por Covid} = 0.0043 \text{ População Rural } - 0.0093 \text{ Analfabetismo} \tag{10}$$

4 Considerações Finais

Os resultados obtidos com a aplicação do modelo de regressão linear múltipla indicam que a população, tanto em áreas urbanas quanto rurais, foi a variável com maior impacto na disseminação e letalidade da Covid-19. A análise revelou que, na área rural, a taxa de analfabetismo e a renda per capita apresentaram influência significativa, enquanto na área urbana, a taxa de analfabetismo manteve-se relevante, mas a renda demonstrou baixa significância estatística para explicar os óbitos por Covid-19.

DOI: 10.5540/03.2026.012.01.0250

No que diz respeito às variáveis de saneamento básico, a falta de acesso à água e ao esgoto mostrou baixa relevância estatística na maioria dos modelos, com exceção da população sem acesso ao esgoto na área rural, que teve impacto mais expressivo nos óbitos por Covid-19.

Apesar do bom ajuste dos modelos, há limitações. A base de dados utilizada conta com um número relativamente pequeno de observações (27 para os modelos urbanos e 26 para os rurais), além de um conjunto restrito de variáveis explicativas.

Dessa forma, estudos futuros podem ampliar essa abordagem por meio da inclusão de um conjunto mais abrangente de variáveis e de uma base de dados atualizada, permitindo uma análise mais aprofundada sobre os fatores socioeconômicos e sanitários que influenciam a propagação e a letalidade da Covid-19 no Brasil.

Referências

[1] D. S. Aquino. "Influência do acesso a saneamento básico na incidência e na mortalidade por COVID-19: análise de regressão linear múltipla nos estados brasileiros". Em: **Revista Thema** 18 (2020). https://doi.org/10.15536/thema.V18.Especial.2020.319-331.1798, pp. 319–331.

[2] Instituto Trata Brasil. **Painel Saneamento Brasil**. Acessado em 17/10/2024, https://www.painelsaneamento.org.br/explore/indicador.

[3] A. Field. **Discovering Statistics Using IBM SPSS Statistics**. 4th. SAGE Publications, 2013. ISBN: 978-1446249178.

[4] D. N. Gujarati. **Econometria Básica**. 4ª. Elsevier, 2006. ISBN: 85-352-1664-6.

[5] IBGE. **IBGE divulga rendimento domiciliar per capita 2023 para Brasil e unidades da federação**. Acessado em 17/10/2024, https://agenciadenoticias.ibge.gov.br/agencia-sala-de-imprensa/2013-agencia-de-noticias/releases/39262-ibge-divulga-rendimento-domiciliar-per-capita-2023-para-brasil-e-unidades-da-federacao.

[6] IBGE. **Painel Nacional por Amostra de Domicílios Contínua**. Acessado em 17/10/2024, https://painel.ibge.gov.br/pnadc/.

[7] M. H. Kutner, C. J. Nachtsheim, J. Neter e W. Li. **Applied Linear Statistical Models**. 5th. McGraw-Hill, 2005. ISBN: 978-0073012281.

[8] A. S. Martins e outros. "Concessão privatista do saneamento e a incidência da Covid-19 em favelas do Rio de Janeiro". Em: **Saúde em Debate** 45.spe2 (2021), pp. 82–91. ISSN: 2358-2898. DOI: 10.1590/0103-11042021E206.

[9] D. C. Montgomery, E. A. Peck e G. G. Vining. **Introduction to Linear Regression Analysis**. 5th. John Wiley & Sons, 2012. ISBN: 978-1-119-59821-0.

[10] E. S. Nascimento, F. M. C Carvalho e E Gomes. "Relação entre fatores socioeconômicos e a pandemia da covid-19". Em: **Revista de Saúde Pública** 33.1 (2024). https://doi.org/10.1590/S0104-12902024220248pt, e220248. ISSN: 1984-0470.

[11] Ministério da Saúde. **COVID-19 no Brasil**. Acessado em 23/01/2025, https://infoms.saude.gov.br/extensions/covid-19_html/covid-19_html.html.

[12] B.G. Tabachnick e L.S. Fidell. **Using Multivariate Statistics**. Always learning. https://books.google.com.br/books?id=ucj1ygAACAAJ. Pearson Education, 2013. ISBN: 9780205849574.

[13] J. M. Wooldridge. **Introductory Econometrics: A Modern Approach**. 6th. Cengage Learning, 2016. ISBN: 978-1305270107.

