Proceeding Series of the Brazilian Society of Computational and Applied Mathematics, v. 11, n. 1, 2025.

Trabalho apresentado no XLIII CNMAC, Centro de Convenções do Armação Resort - Porto de Galinhas - PE, 2024

Proceeding Series of the Brazilian Society of Computational and Applied Mathematics

Uso de Regressão Linear Múltipla para Previsão da Atividade Inibitória da Enzima N-miristoiltransferase de *Leishmania donovani*

Soraya de O. Bandeira[1], Carlos M. R. Sant'Anna [2], Marcelo D. Cruz[3]
PPGMMC/UFRRJ, Seropédica, RJ
Igor C. A. Lima[4]
IME/UERJ, Rio de Janeiro, RJ

Resumo. A leishmaniose é uma doença negligenciada (DN) e representa um desafio mundial. Os medicamentos usados no tratamento contra leishmaniose são administrados com base na inibição de outras doenças e tem efeitos colaterais. O campo da modelagem molecular tem se mostrado eficiente para o desenvolvimento de novos medicamentos, quanto a recurso e tempo. Este trabalho tem como objetivo aplicar o método QSAR (Quantitative Structure Activity) para a modelagem molecular de estruturas sintetizadas e promissoras no combate a *Leishmania donovani*, em caráter inibitório do protozoário. O modelo proposto atendeu aos requisitos da área de QSAR, devendo ser, desta forma, expandido para novos estudos com bases maiores.

Palavras-chave. QSAR, Aprendizado de máquina, Leishmaniose, *Leishmania donovani*

1 Introdução

A leishmaniose é uma patologia que ocorre principalmente em países tropicais, causada por protozoários do gênero *Leishmania*, neste trabalho escolheu-se o parasita *Leishmania donovani* . A leishmaniose é um desafio mundial, pois esta incluída no grupo de doenças negligenciadas (DNs), que afetam 15 % das pessoas no mundo e há uma dificuldade na confecção de novos fármacos para os tratamentos [14]. A descoberta de moléculas bioativas que possam combater a leishmaniose é um processo caro e demorado, novas estratégias são continuamente buscadas para otimizar esse processo. As estratégias como triagem virtual (VS) e ferramentas computacionais em pesquisa in sílico mostrou-se uma alternativa viável no entendimento na pesquisas de novos fármacos oferecendo redução do tempo e de recursos financeiros. Considera-se que existe uma relação entre as propriedades de uma molécula, sua estrutura química e sua atividade biológica, então, buscam-se estabelecer relações matemáticas simples para descrever e prever a atividade de um conjunto de inibidores semelhantes [6]. O campo da modelagem molecular tem se mostrado eficiente para o desenvolvimento de novos medicamentos Este trabalho tem como objetivo aplicar o método QSAR (Quantitative Structure Activity) para a modelagem molecular de estruturas sintetizadas e promissoras no combate a Leishmania donovani, juntamente com uma análise de métodos lineares de aprendizado de máquina em caráter inibitório do protozoário.

[1]sorayaquimica@gmail.com
[2]cmauricio.rsantanna@gmail.com
[3]madibcruz@gmail.com
[4]ilima@uerj.br

No presente trabalho, foi empregada a metodologia de regressão linear múltipla e, ainda, a avaliação da estratégia de seleção de variáveis, na predição de atividade inibitória da enzima N-miristoiltransferase da *Leishmania donovani*, em uma base de dados com 42 variáveis preditoras e 21 amostras. A escolha do melhor modelo se baseará naquele que tiver o menor RMSE, desde que os critérios de QSAR sejam atendidos.

2 Metodologia

2.1 Base de dados

Baseada na patente depositada por [10] , foram extraídas 21 moléculas originalmente sintetizadas com diferentes grupos funcionais (tais como éter, amina e éster) se alternando nas posições substituintes. As amostras continham valores de IC_{50} (concentração necessária para inibir 50% da atividade da enzima N-miristoiltransferase da *L. donovani*.

A partir dos dados de IC_{50} e de representação molecular, foram empregados os *softwares* Spartan'20 [18] e GOLD versão 2022.3.0 [2] para a etapa de modelagem molecular, com a obtenção de diversos descritores químicos para análise de QSAR.

2.2 Análise de dados

Foi utilizado o *software* R [15] para manipulação e análise de dados, junto com os pacotes readxl [20], stats [16], caret [8], lmtest [22], olsrr [7] e ggplot2 [19].

Com a base de dados definida, foi empregada a distância de Mahalanobis para identificar se haviam outliers multivariados na base de dados, após a remoção das variáveis preditoras que tenham as maiores correlações lineares (adotou-se o valor de corte $r > |\pm 0,75|$) [9]. Desta forma, a base de dados reduziu, de 41 variáveis para 23 variáveis candidatas.

A seguir, foi construída uma matriz de correlações com as variáveis preditoras e a variável dependente pIC50, exemplificado em um correlograma na figura 1:

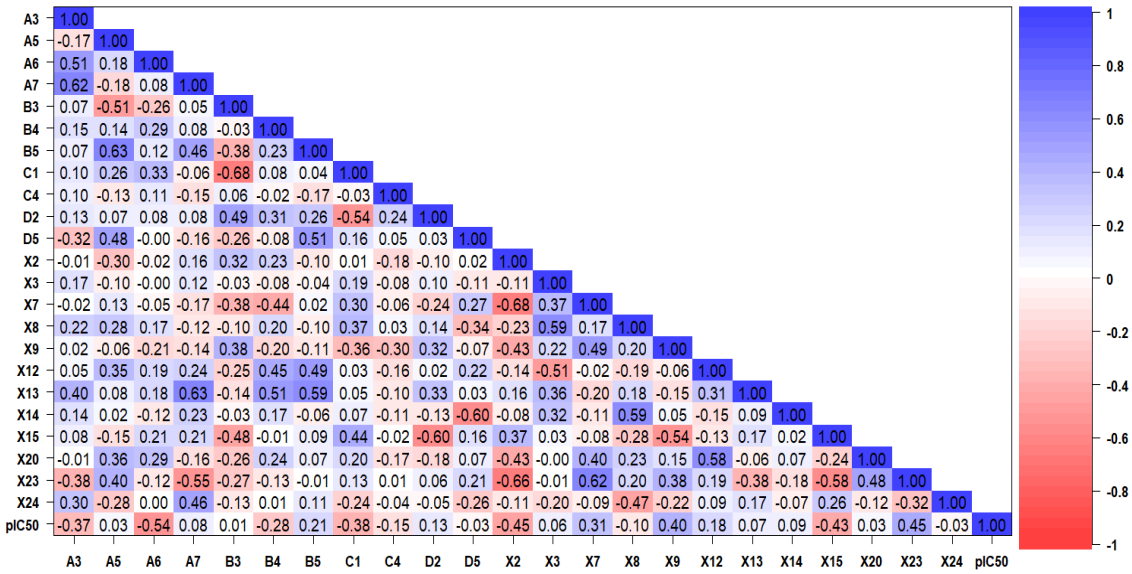

Figura 1: Matriz de correlações lineares. Fonte: [1], 2024.

Baseado neste resultado, deu-se prosseguimento à exclusão de variáveis cujo valor do coeficiente de correlação entre a variável preditora e a variável dependente fosse $r < |\pm 0,3|$ [3]. Foram excluídas as variáveis A4, A5, A7, B4, B5, C2, C4, D2, D3, D5, X5, X8, X12, X13, X14, X20 e X24, por apresentarem uma correlação desprezível. A base de dados passou para oito variáveis independentes (A3, A6, C1, X2, X7, X9, X15 e X23) e uma variável dependente (pIC_{50}).

A seguir, foi empregada a distância de Mahalanobis, para identificar possíveis outliers multivariados. A figura 2 mostra o resultado. A linha tracejada vermelha é o valor limite, obtido através da raiz quadrada de uma distribuição qui-quadrado, com nível de significância $\alpha = 0,975$ e graus de liberdade $gl = 8$ [5]:

$$\text{Valor crítico} = \sqrt{\chi^2_{(\alpha=0,975\ ;\ gl=8)}} = 4,19 \tag{1}$$

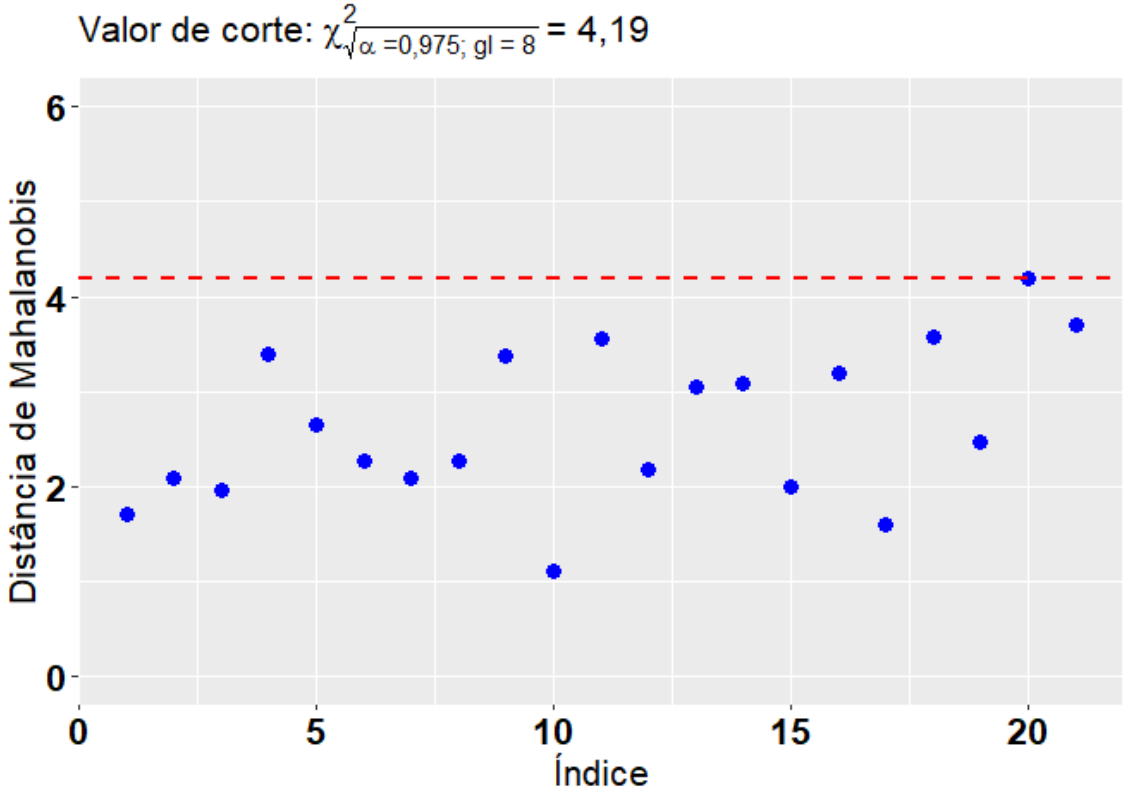

Figura 2: Distância de Mahalanobis. Fonte: [1], 2023.

Após a remoção da amostra número 20, com um valor calculado de distância $d \cong 4,1972$, o banco de dados foi avaliado em questão da multicolinearidade, através do fator de inflação da variância (2) de cada variável [12].

$$VIF_k = \frac{1}{1 - r_k^2} \tag{2}$$

Das variáveis iniciais (A3, A6, C1, X2, X7, X9, X15 e X23), foram excluídas as variáveis X9 e X23. Desta forma, as variáveis utilizadas são A3, A6, C1, X2, X7 e X15.

A seguir, o banco de dados foi particionado em dois, na proporção 70%/30% (etapas calibração e validação, respectivamente). Foi empregado também a validação cruzada *leave-one-out*, LOO.

Para a modelagem preditiva, foi empregada a regressão com melhores subconjuntos. Para a escolha do modelo candidato, foram consideradas de forma conjunta diversas métricas, taís como os critérios de informação de Akaike (3) e Bayesiano (4):

$$AIC = \left[\frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{n}\right] \times e^{2k/n} \tag{3}$$

$$BIC = \left[\frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{n}\right] \times n^{k/n} \tag{4}$$

onde y_i é o valor observado, $\hat{y}_i$ é o valor previsto pelo modelo, n é o tamanho da amostra, k é o número de variáveis independentes, e T é o número de parâmetros, incluindo o intercepto.

Para avaliar os pressupostos da regressão, foram empregados os testes de Shapiro-Wilk-Royston [21], para avaliar a normalidade dos resíduos; de Durbin-Watson [11], para verificar a ausência de autocorrelação serial; e de Breusch-Pagan , para avaliar a homocedasticidade [13] . Na capacidade preditiva, foi avaliado também a raiz quadrada do erro médio, (5).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2} \tag{5}$$

Na área de QSAR, é utilizado na construção e validação de modelos preditivos o coeficiente de correlação de validação cruzada (6) [17], obtido após a partição dos conjuntos em treino e teste e usado na etapa de validação cruzada, .

$$Q^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_{(i)}\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)} \tag{6}$$

onde $\hat{y}_{(i)}$ é a i-ésima amostra prevista no modelo, sem que a mesma estivesse presente no momento da construção. Já $\bar{y}$ é a média das amostras da etapa de calibração.

Outra métrica importante na validação de modelos QSAR é a obtenção do coeficiente de determinação permutado médio corrigido (7) [17].

$${}^{c}r_p^2 = r \times \sqrt{r^2 - r_p^2} \tag{7}$$

onde r_p^2 é o coeficiente de determinação permutado médio. Este coeficiente é a média de n coeficientes de determinação de cada modelo cuja variável dependente Y foi permutada. Neste trabalho, foram realizadas 1000 permutações em Y, para obtenção do r_p^2.

Conforme [4] e [3] indicam, na validação cruzada LOO, as seguintes relações devem ser preservadas: $r^2 > Q^2$ e RMSEC $<$ RMSEP. Uma condição aceitável é $Q^2 > 0,5$ e $r^2 > 0,6$. Eles alertam que se a diferença entre o coeficiente de determinação e o de validação cruzada for maior que intervalo de 0,2 a 0,3, ou seja, $r^2 - Q^2 > (0,2 \text{ a } 0,3)$, o modelo tem sobreajuste.

3 Resultados

Com a base de dados particionada, foi empregada a regressão por melhores subconjuntos, para encontrar os melhores modelos preditos, dentre todos $2^K - 1$ modelos possíveis, onde K é o número de variáveis independentes. Cinco modelos candidatos foram selecionados, conforme mostrados na

tabela 1. Os coeficientes de determinação variaram de 0,39 a 0,76, mostrando que mesmo com todas as variáveis, o modelo só conseguiria explicar 76% da variabilidade na variável dependente pIC_{50}.

Tabela 1: Figuras de mérito das regressões por melhores subconjuntos.

Modelo	r^2	r^2_{aj}	r^2_{pred}	Cp	AIC	BIC
A3	0,39	0,35	0,23	10,77	28,81	31,13
A3 + X2	0,60	0,54	0,43	4,99	24,12	27,21
A3 + C1 + X7	0,70	0,63	0,54	3,18	21,43	25,30
A3 + C1 + X7 + X15	0,74	0,64	0,51	3,91	21,51	26,14
A3 + A6 + C1 + X7 + X15	0,75	0,62	0,45	5,45	22,75	28,16
A3 + A6 + C1 + X2 + X7 + X15	0,76	0,60	0,19	7,00	23,96	30,14

O modelo candidato, baseado nos critérios AIC, BIC e Cp, foi aquele com três variáveis, apresentado na equação (8). Este modelo apresentou um $r^2 = 0,7022$, um coeficiente de determinação ajustado $r^2_{aj} = 0,6277$, um erro-padrão da regressão $\hat{\sigma} = 0,3994$; um teste F de significância global $F = 9,431.$ que excede o valor crítico $F(1 - 0,05;\ \nu_1 = 3;\ \nu_2 = 12) = 3,49$, o que indica que a regressão é significativa.

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1(A3) + \hat{\beta}_2(C1) + \hat{\beta}_3(X7) \tag{8}$$

A tabela 2 mostra os resultados de cada coeficiente $\hat{\beta}_j$ obtido; do erro-padrão dos coeficientes, $\text{ep}\left(\hat{\beta}_j\right)$; do teste t de significância (com $\alpha = 0,05$) e o valor-p, $Pr\,(>|t|)$.

Tabela 2: Inferência sobre o modelo de regressão selecionado.

	$\hat{\beta}_j$	$\text{ep}\left(\hat{\beta}_j\right)$	teste t	$Pr\,(>\|t\|)$
Intercepto	6,109719	1,990828	3,0690	0,0097
A3	-0,4767	0,1394	-3,4180	0,0051
C1	-0,0637	0,0243	-2,6240	0,0222
X7	0,0162	0,0052	3,1090	0,0090

Através do valor-p obtido, observa-se que todos os coeficientes são significativos, pois são menores que o nível de significância adotado ($\alpha = 0,05$).

Para a normalidade de resíduos, foi empregado o teste de Shapiro-Wilk com modificação de Royston [114]. Sob a hipótese nula $H_0 : e_i \sim N(0, \sigma^2)$ e a hipótese alternativa $H_1 : c.c.$, a estatística do teste foi $W = 0,91883$, com valor-p $= 0,1615$. Logo, este pressuposto não foi violado.

Para a homocedasticidade, foi empregado o teste studentizado de Breusch-Pagan [13][109]. A estatística do teste foi $BP = 0,63149$, com três graus de liberdade e valor-p $= 0,8892$. Portanto, os resíduos são homocedásticos, ou seja, apresentam variância constante.

Para a ausência de autocorrelação serial, foi empregado o teste Durbin-Watson [11]. A estatística do teste foi $DW = 1,582$, com valor-p $= 0,1551$. Entretanto, o teste de Durbin-Watson apresenta duas regiões inconclusivas, e apenas a guiagem pelo valor-p pode, neste caso, não fornecer a decisão correta. Para um modelo de regressão com três variáveis independentes (k = 3) e $\alpha = 0,05$, temos o limite inferior da região inconclusiva $d_{Inf} = 0,86$ e o limite superior da região inconclusiva $d_{Sup} = 1,73$. Desta forma, não podemos concluir sobre se este pressuposto foi violado ou não [11] .

Além das métricas avaliadas, em específico da área de QSAR, vale ressaltar que o coeficiente de determinação, da etapa de calibração, é maior que o coeficiente de determinação de validação cruzada $r^2_{cal} = 0,7022 > Q^2 = 0,5373$; bem como a raiz quadrada do erro médio de calibração é menor que a da etapa de validação, RMSEC $= 0,3459 <$ RMSEP $= 0,8633$. Por fim, a diferença entre o r^2 e o Q^2 ($r^2_{cal} - Q^2 = 0,1649$) foi menor que o intervalo de $0,2 - 0,3$, o que demonstra que o modelo não tem sobreajuste [3]. O modelo obtido passou nos critérios $Q^2 > 0,5$ e $r^2_{cal} > 0,6$. De acordo com [5], os modelos com ${}^c r^2_p > 0,5$ são considerados estatisticamente robustos, indicando que o modelo desenvolvido não foi obtido meramente ao acaso.

A figura 3 mostra o gráfico de dispersão com os valores observados da atividade inibitória pIC_{50} na abscissa e os valores preditos pelo modelo na ordenada. As amostras utilizadas no conjunto de calibração estão em azul e as amostras alocadas para o conjunto teste, em vermelho.

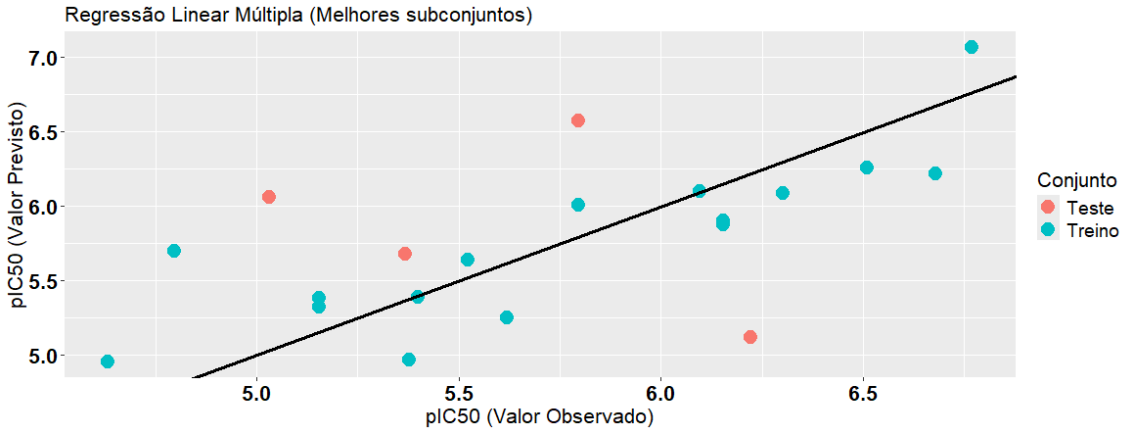

Figura 3: Valores observados versus valores preditos da atividade inibitória pIC_{50}. Fonte: [1], 2024.

O gráfico mostra um bom ajuste, para dados experimentais a partir de uma base de dados pública. Como era de se esperar, há uma alta variabilidade no conjunto teste. Isto se deve a variações presentes nos substituintes.

4 Considerações Finais

Este trabalho teve como objetivo avaliar o uso de regressão por melhores subconjuntos na predição de valores de atividade inibitória pIC_{50} da enzima N-miristoiltransferase em *Leishmania donovani*.

A importância deste trabalho reside no fato desta ser uma doença negligenciada, que afeta milhares de brasileiros e cujos tratamentos atuais apresentam efeitos colaterais nocivos aos pacientes.

O modelo apresentado obteve um bom desempenho e atendeu aos requisitos do modelo de mínimos quadrados, tais como normalidade dos resíduos e homocedasticidade, bem como critérios específicos da área de QSAR, como o coeficiente de determinação de validação cruzada maior que 0,5.

Sugere-se que este trabalho seja ampliado para uma base de dados maior, com a investigação também de outras variáveis preditoras, bem como o estudo detalhado do mecanismo de interação entre a molécula e o sítio ativo.

DOI: 10.5540/03.2025.011.01.0509

Referências

[1] S. O. Bandeira. "O uso de algoritmos de Aprendizado de Máquina para a busca de modelos de previsão da atividade Inibitoria da enzima N miristoiltransferase de Leishmania". Dissertação de Mestrado. Universidade Federal Rural do Rio de Janeiro, Rio de Janeiro, 2024.

[2] Cambridge Crystallographic Data Centre. Online. Acessado em 26 jul. 2023, <https://www.ccdc.cam.ac.uk/solutions/software/gold/>.

[3] M. M. C. Ferreira e R. Kiralj. "Métodos quimiométricos em relações quantitativas estrutura - atividade (QSAR)". Em: **Jornal Química Medicinal Métodos e Fundamentos em Planejamento de Fármacos** 28(5) (2019), pp. 387–454.

[4] R. S. Ferreira, G. Oliva e A. Andricopulo. "D. Integração das técnicas de triagem virtual e triagem biológica automatizada em alta escala oportunidades e desafios e PD de fármacos". Em: **Química Nova** V. 34 N. 10 (2011), pp. 1770–1778.

[5] K. Filzmoser P.; Varmuza. Online. Acessado em 26 jul. 2023, <https://CRAN.R-project.org/package=chemometrics>.

[6] L. S. Garcia. "Estudos computacionais de potenciais inibidores da enzima N- Miristoiltransferase de Plasmodium falciparum e Leishmania donovani." Tese de doutorado. Universidade Federal de Lavras, Minas Gerais, 2017.

[7] A. Hebbali. Online. Acessado em 26 jul. 2023, <https://CRAN.R-project.org/package=olsrr>.

[8] M. Kuhn. Online. Acessado em 26 jul. 2023, <https://doi.org/10.18637/jss.v028.i05>.

[9] M. Kuhn. "Building Predictive Models in R Using the caret Package". Em: **Journal of Statistical Software** 28(5) (2008), pp. 1–26. DOI: doi.org/10.18637/jss.v028.i05.

[10] R. Leatherbarrow, E. Tate, Z. Yu e M. Rackham. "Nobel Compounds and their use in therapy". Tese de doutorado. Reino Unido: World Intellectual Property rganization, 2013.

[11] D. M. Levine, D. F. Stephan e K. A. Szabat. 2016.

[12] D. C. Montgomery, E. A. Peck e G. G. Vining. "Introduction to linear regression analysis". Em: **Hoboken: John Wiley Sons** 5 (2012), p. 168.

[13] R. Pedace. Online. Acessado em 25 dez. 2023, <https://www.academia.edu/17535082/Econometrics_for_Dummies_by_Roberto_Pedace_John_Wiley_and_Sons_Inc_Hoboken_NJ_2013_pp_xvi_342>.

[14] S. S. Santos, R. V. ARAÚJO, J. GIAROLLA, O. E. SEOUD e E. I. Ferreira. "Searching for drugs for Chagas disease, leishmaniasis and schistosomiasis". Em: **Journal International of Antimicrobial Agents** 4 (2020), pp. 2–18. DOI: 10.1016/j.ijantimicag.2020.105906.

[15] R Core Team Foundation for Statistical Computing. Online. Acessado em 2023 e 2024, <https://www.R-project.org/>.

[16] R Core Team. Online. Acessado em 26 jul. 2023, <https://www.R-project.org/>.

[17] R. E. Walpole, R. H. Myers, S. L. Myers e K. Ye. "Probability Statistics for Engineers Scientists". Em: **Journal Pearson Education** 9 (2012).

[18] Wavefunction. Online. Acessado em 26 jul. 2023, <https://www.wavefun.com/>.

[19] H. Wickham. Online. Acessado em 26 jul. 2023, <https://CRAN.R-project.org/>.

[20] H. Wickham e J. Bryan. Online. Acessado em 26 jul. 2023, <https://CRAN.R-project.org/package=readxl/>.

[21] C. Zaiontz. Online. Acessado em 25 dez. 2023, <https://real-statistics.com/tests-normality-and-symmetry/statistical-tests-normality-symmetry/shapiro-wilk-expanded-test/>.

[22] A. Zeileis e T. Hothorn. Online. Acessado em 26 jul. 2023, <https://CRAN.R-project.org/doc/Rnews/1>.

