Correlating Socioeconomic Factors and Syphilis Incidence: A Case Study in Brazil

Matheus Tavares[1], Elias Mendes[1], Diego Dias[2(✉)], Elisa Tuler[1], Rodolfo Villaça[2], Marcelo Guimarães[3], and Leonardo Rocha[1]

[1] Universidade Federal de São João del Rei (UFSJ), São João del Rei, Brazil
matheus93t@aluno.ufsj.edu.br, {etuler,lcrocha}@ufsj.edu.br
[2] Universidade Federal do Espírito Santo (UFES), Vitória, Brazil
{diego.dias,rodolfo.villaca}@ufes.br
[3] Universidade Federal de São Paulo (UNIFESP), São Paulo, Brazil
mendeselias2014@aluno.ufsj.edu.br

Abstract. Syphilis is a growing public health concern in Brazil, disproportionately affecting vulnerable populations. Socioeconomic factors, such as education, income, and access to healthcare, significantly influence its incidence. This study compares six statistical methods—Pearson and Spearman correlations, univariate and multivariate linear regression, logarithmic regression, and Poisson regression—to analyze these relationships using data from Minas Gerais, Brazil. The results highlight strong associations between the incidence of syphilis and both educational and health indicators, emphasizing the importance of preventive policies. Methodological comparisons reveal significant discrepancies, with some approaches misestimating the impact of variables due to multicollinearity or data distribution effects. These findings underscore the importance of selecting appropriate statistical models to ensure accurate epidemiological interpretations and support the development of effective public health strategies.

Keywords: Statistical methods · Syphilis · Correlation analysis · Regression models · Data Modeling

1 Introduction

Syphilis is a chronic bacterial infection caused by *Treponema pallidum*, which progresses through distinct clinical stages and can result in severe complications if left untreated [26]. In Brazil, the disease has emerged as a growing public health concern, particularly due to the rising incidence observed in recent years, disproportionately affecting vulnerable populations and diverse social groups. Socioeconomic factors, including educational attainment, income levels, access to healthcare services, and regional disparities, play a critical role in the transmission and persistence of syphilis. A comprehensive understanding of these

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
O. Gervasi et al. (Eds.): ICCSA 2025, LNCS 15649, pp. 65–80, 2025.
https://doi.org/10.1007/978-3-031-96997-3_5

relationships is essential for developing more effective prevention and control strategies [6].

In this context, computational approaches applied to epidemiology provide valuable tools for analyzing the distribution of syphilis within populations. Statistical methods and machine learning techniques facilitate the identification of patterns, the prediction of trends, and the assessment of the influence of socioeconomic variables on disease incidence. However, the heterogeneity of methodological approaches employed in the literature [6,13,15,19,22] raises concerns about the consistency and comparability of findings. Each statistical method is based on distinct assumptions, offering specific advantages and limitations, and the inappropriate selection of analytical techniques may lead to misinterpretations, ultimately compromising the effectiveness of intervention strategies [23].

To address this issue, the present study aims to compare the application of different statistical methods in analyzing the correlation between syphilis incidence and socioeconomic factors (i.e., education, health, and income). Specifically, we evaluate Pearson correlation [3], Spearman correlation [9], univariate linear regression [1], multivariate linear regression [1], logarithmic regression [2], and Poisson regression [11]. Our analysis focuses on data from selected municipalities in Minas Gerais, one of Brazil's largest and most populous states, characterized by significant regional disparities in socioeconomic development. The study examines the occurrence of gestational syphilis cases [8], using data obtained from the Minas Gerais Transparency Portal[1], the João Pinheiro Foundation[2], and the Minas Gerais Open Data Portal[3]. By assessing the adequacy and robustness of each method, this research seeks to contribute to the standardization of epidemiological analyses and to enhance the understanding of the interrelationships between socioeconomic and epidemiological variables.

While detailed results are presented later in this work, preliminary analyses indicate that educational indicators exhibit the strongest associations with syphilis incidence, underscoring their importance in guiding public policies. Health-related variables, particularly those linked to mortality and healthcare access, also show consistent patterns, while economically disadvantaged populations remain the most vulnerable.

Moreover, methodological divergences observed across models emphasize the importance of selecting appropriate techniques. Certain methods may produce biased coefficients due to data distribution characteristics or multicollinearity, reinforcing the need for a comparative and critical approach.

In sum, our findings highlight the importance of choosing appropriate methodologies to ensure reliable interpretations of the dataset. Moreover, no single strategy consistently captures all relevant correlations, making it essential to consider multiple statistical approaches.

[1] https://www.dados.mg.gov.br/vi/dataset/dados_sifilis_ses/resource/c1570999-cbbf-4441-ac53-acd229c36c9f?view_id=e598cfd0-ec9f-48e0-a371-ce1b035d4f91.
[2] https://imrs.fjp.mg.gov.br/Consultas.
[3] https://dados.mg.gov.br/.

2 Related Work

The relationship between disease incidence and socioeconomic factors has been extensively investigated in the literature, particularly in the context of sexually transmitted infections (STIs). Various methodological approaches have been employed to analyze incidence rates and associated risk factors. In general, studies indicate that conditions such as low income, limited education, and restricted access to healthcare services are directly linked to an increase in STI cases [14,15,24,25].

Although multiple studies have identified these conditions as relevant factors, no absolute consensus exists regarding all the determinants that influence STI incidence. Discrepancies arise due to variations in the databases analyzed, selection criteria, and geographic scope of the studies. Additionally, the choice of methodological approaches plays a crucial role in shaping research findings. Various statistical techniques have been used to explore these relationships. For instance, Barbosa et al. [13] investigated risk factors associated with vertical transmission of syphilis in Belo Horizonte, Brazil, employing multivariate logistic regression. Their findings identified low maternal education and inadequate prenatal care as key determinants. In contrast, [19] examined the incidence of congenital syphilis in Recife, Brazil, employing Spearman's correlation to assess the relationship between disease incidence and social inequalities. Their results demonstrated a strong association between higher syphilis incidence and neighborhoods with a lower Human Development Index (HDI). However, neither study explicitly justified the choice of statistical models used for analysis.

Furthermore, spatiotemporal analyses have been applied to understand the distribution of syphilis over time. For example, [22] employed Poisson regression to model syphilis dynamics in São Paulo, Brazil, highlighting seasonal patterns and the impact of public health policies on disease control. Meanwhile, [15] utilized logarithmic regression to assess social inequalities in congenital syphilis incidence in Rio de Janeiro, reinforcing the necessity of targeted interventions for vulnerable populations. These studies, similar to previous ones, applied distinct methodological approaches, often without a clear rationale for their selection. This raises important questions: Which methods are most appropriate for analyzing the relationship between syphilis and socioeconomic factors? Is there a universally optimal approach, or does the choice of method depend on specific epidemiological characteristics? What criteria should guide the selection of an appropriate analytical method?

The diversity of statistical techniques used in the literature reflects the inherent complexity of syphilis epidemiology but also poses challenges to the comparability of results. Different methodological choices can lead to divergent conclusions, complicating the development of consistent public health strategies [23]. In light of this, the present study aims to conduct a comparative analysis of various statistical methodologies used to assess the correlation between syphilis incidence and socioeconomic indicators. The objective is to evaluate the robustness and applicability of each method, contributing to a more standardized and evidence-based approach in epidemiological research.

3 Evaluation Methodology

In this section, we detail the evaluation methodology used to compare statistical and machine learning methods applied in the analysis of the correlation between socioeconomic factors and syphilis incidence.

3.1 Evaluated Methods

Based on the studies presented in the previous section, we consider two correlation methods and four regression methods, as detailed below:

1. **Pearson Correlation:** measures the strength and direction of a linear relationship between two quantitative variables [3]. The coefficient r is:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{1}$$

2. **Spearman Correlation:** assesses the strength of a monotonic relationship based on ranked values [9]. The coefficient ρ is:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \tag{2}$$

   where d_i is the rank difference and n is the number of observations.
3. **Univariate Linear Regression:** models the relationship between a dependent variable Y and one independent variable X [1]:

$$Y = \beta_0 + \beta_1 X + \epsilon \tag{3}$$

4. **Multivariate Linear Regression:** extends to multiple independent variables:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon \tag{4}$$

5. **Logarithmic Regression:** fits data where Y depends on the natural logarithm of X [2]:

$$Y = \beta_0 + \beta_1 \ln(X) + \epsilon \tag{5}$$

6. **Poisson Regression:** models count data where the dependent variable represents the number of events [11]:

$$\log(\lambda) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p \tag{6}$$

   where λ is the expected event rate.

3.2 Analysed Datasets

Our study examines the occurrence of syphilis cases [8] occurring in municipalities of Minas Gerais, one of Brazil's largest and most populous states, characterized by significant regional disparities in socioeconomic development. Specifically, we considered data related to gestational syphilis. This disease, when left untreated, can lead to several complications for the fetus, including: miscarriage, stillbirth, prematurity, low birth weight, and congenital syphilis itself, which can lead to neurological, bone, and skin problems in the baby. It is essential to emphasize the importance of early screening, diagnosis, and treatment of syphilis during pregnancy in order to prevent such consequences and ensure the health of the fetus. These were obtained from Minas Gerais Open Data Portal[4]. In our study, we consider the attributes listed in Table 1.

Table 1. Gestational Syphilis Dataset Attributes

Attribute	Description
_id	Unique Identifier
ID_AGRAVO	Code of the agravo
DT_NOTIFIC	Date of Notification
NU_ANO	Year of Notification
ID_MUNICIP	Municipality
ID_REGIONA	Region
DT_DIAG	Date of Diagnosis
DT_NASC	Date of Birth
NU_IDADE_N	Age in days/months
CS_SEXO	Sex
CS_RACA	Race/Color
ID_MN_RESI	Municipality of Residence
ID_RG_RESI	Region of Residence
ANTSIFIL_N	History of Syphilis at Birth
TRA_ESQUEM	Treatment Regimen
LABC_IGG	IgG Laboratory Test
EVOLUCAO	Case Evolution

The socioeconomic information were obtained from Minas Gerais Transparency Portal[5] and the João Pinheiro Foundation[6]. We selected five cities with

[4] https://www.dados.mg.gov.br/vi/dataset/dados_sifilis_ses/resource/c1570999-cbbf-4441-ac53-acd229c36c9f?view_id=e598cfd0-ec9f-48e0-a371-ce1b035d4f91.

[5] https://www.dados.mg.gov.br/vi/dataset/dados_sifilis_ses/resource/c1570999-cbbf-4441-ac53-acd229c36c9f?view_id=e598cfd0-ec9f-48e0-a371-ce1b035d4f91.

[6] https://imrs.fjp.mg.gov.br/Consultas.

the highest incidence of syphilis cases in the state of Minas Gerais, one of the largest and most populous states in Brazil. These cities belong to different regions with varying levels of socioeconomic development (i.e., Belo Horizonte, Juiz de Fora, Ipatinga, Governador Valadares and Montes Claros). We considered the period from 2010 to 2023 and all the available attributes, which correspond to different social indicators: Income, with a sub-indicator of "Management"; Education with sub-indicators "Access" and "Management"; and Health with the sub-indicators "Management", "Hospital medical care" and "Primary care".

Regarding preprocessing strategies, a check for missing data was performed, identifying empty columns and incomplete years. Indicators that presented three or more missing values throughout the period were removed from the study; this amount was chosen because it is equivalent to approximately 20% of the missing data. Although there is no strict consensus in the literature on the exact percentage of missing data that compromises the analysis, some guidelines suggest missing rates greater than 20% can introduce significant biases [4,10,12,21]. At the end of this stage, we were left with a robust set of indicators: 18 Education, 22 Health, and 13 Income indicators, ensuring the comprehensive scope of our study. Finally, we combined the two data collections (syphilis incidence data and socioeconomic data) into a single collection. Finally, we encoded the categorical variables using LabelEncoder [16] and normalized the data using the MinMaxScaler [17] or StandardScaler [18] techniques, ensuring that all indicators were on the same scale and could be compared correctly.

3.3 Experimental Procedures

Our analyses were structured into two lines. First, our goal was to evaluate the quality of the regressions performed by each of the methods. We did this by comparing the function generated with the real data reported in the data collection. The objective was to compare these regression strategies, identifying their limitations in the specific scenario under consideration. To facilitate this comparison, we used the Mean Squared Error (MSE) and the Coefficient of Determination (R^2), two widely used metrics in the evaluation of regression models due to their simplicity and ability to measure the quality of the fits [5].

1. **Mean Squared Error (MSE):** calculates the average of the squares of the differences between predicted and observed values, penalizing more significant errors more severely. MSE helps highlight large discrepancies between predictions and observations due to the quadratic penalization of errors. The smaller the MSE, the better the model fit:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{7}$$

2. **Coefficient of Determination ($\mathbf{R^2}$):** indicates the proportion of the total variability of the dependent variable that the regression model explains. The R^2 does not provide information about the magnitude of the errors but rather about the proportion of the variability the model explains. It is calculated as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{8}$$

where $\bar{y}$ is the mean of the observed values. The value of R^2 ranges from 0 to 1, with values close to 1 indicating that the model explains most of the variability in the data.

In our second analysis, we compared the indicators indicated by each of the strategies analyzed, a crucial step in our research. We applied the correction strategies (Pearson and Spearman) separately and evaluated the five indicators that were associated with the incidence of gestational syphilis, as well as the coefficients related to each of these indicators. We then applied the regression strategies (Univariate and Multivariate Linear Regression, Logarithmic Regression, and Poisson Regression) and evaluated the weights they returned that were associated with the indicators returned by the correlation methods. This comparison was performed separately for each attribute group (educational, health, and income), underscoring the significance of our research in evaluating the complementarity of the strategies considered.

4 Discussion of Results

As mentioned in the previous section, we focused our first analysis on evaluating the predictive performance of the regression models, comparing the predicted values with the actual data. These analyses were performed separately for each city, considering both metrics, R^2 and MSE. In this first evaluation, we considered all the indicators analyzed (health, education, and income), with 80% of the data randomly selected to perform the progression and the evaluation of the result being done on the remaining 20% of the collection. An important observation is that for R^2, a Univariate Linear Regression Linear, the calculation is done for each indicator individually, and, subsequently, the average is calculated. The results can be observed in Tables 2, 3, 4, 5 and 6

Table 2. Results for Belo Horizonte

Metric	Regression Method			
	Logarithmic	Linear Univariate	Linear Multivariate	Poisson
MSE	0.00	0.033	0.00	0.00
R^2	1.00	0.685	1.00	1.00

Our first observation is that some algorithms demonstrate 100% accuracy (R^2) and zero error (MSE). This is properly interpreted as data overfitting, which occurs when a model, in this case a regression model, cannot generalize the data's behavior. It ends up self-adjusting to the data in such a way that it

Table 3. Results for Juiz de Fora

Metric	Regression Method			
	Logarithmic	Linear Univariate	Linear Multivariate	Poisson
MSE	0.00	0.026	0.00	0.00
R^2	1.00	0.722	1.00	1.00

Table 4. Results for Ipatinga

Metric	Regression Method			
	Logarithmic	Linear Univariate	Linear Multivariate	Poisson
MSE	0.00	0.032	0.00	0.00
R^2	1.00	0.70	1.00	1.00

Table 5. Results for Gorvenador Valadares

Metric	Regression Method			
	Logarithmic	Linear Univariate	Linear Multivariate	Poisson
MSE	0.00	0.044	0.00	0.00
R^2	1.00	0.58	1.00	1.00

Table 6. Results for Montes Claros

Metric	Regression Method			
	Logarithmic	Linear Univariate	Linear Multivariate	Poisson
MSE	0.00	0.042	0.00	0.00
R^2	1.00	0.59	1.00	1.00

cannot make effective predictions for new data. This phenomenon occurs when the model memorizes the specific patterns of the training set, including possible noise, instead of capturing the general trend. As a consequence, its performance on unseen data may be significantly lower, making it unsuitable for practical applications. This reinforces the need for adjustments, such as regularization, collection of more data, or the adoption of more robust models to avoid this problem.

Another hypothesis that can explain this issue is Multicollinearity. [7], which occurs when two or more independent variables in a statistical model are highly correlated, making it difficult to determine the individual impact of each one on the dependent variable. This problem mainly affects regression models [20], as it reduces the stability of coefficient estimates and can lead to distorted interpretations of the results. When working with social indicators, multicollinearity can arise naturally, as many of these are interconnected. For example, variables such as per capita income, level of education, and employment rate tend to present

Table 7. Health Indicators

Indicator	Health
Indicator 1	Lung cancer mortality rate
Indicator 2	Breast cancer mortality rate
Indicator 3	Number of doctors per 1000 inhabitants
Indicator 4	Proportional mortality of the elderly population
Indicator 5	Crude mortality rate

Table 8. Education Indicators

Indicator	Education
Indicator 1	School attendance rate for children aged 0 to 3 years old
Indicator 2	School attendance rate for children aged 4 and 5 years old
Indicator 3	Net enrollment rate for elementary school
Indicator 4	Per capita spending on education activities
Indicator 5	Total number of schools offering complete regular elementary school

Table 9. Income Indicators

Indicator	Income
Indicator 1	Disabled people benefiting from BPC
Indicator 2	Total number of BPC beneficiaries
Indicator 3	Elderly people benefiting from BPC
Indicator 4	Per capita income in the formal sector
Indicator 5	Total BPC transfers

strong correlations with each other. Likewise, health indicators, such as infant mortality rate, access to basic sanitation, and life expectancy, often have significant statistical relationships. This problem can generate unstable coefficients, making it challenging to identify which factors really influence the response variable, which highlights the need always to consider different models to perform analyses and thus avoid hasty and perhaps erroneous conclusions.

In our second analysis, which focuses on analyzing the main indicators indicated by each method as the most relevant to explain the incidence of gestational syphilis, for reasons of space, we will consider only the data related to the city of Belo Horizonte, capital of Minas Gerais. As mentioned in Sect. 3.3, we first applied the Pearson and Spearman correlation methods. In this case, we separated our analyses for each group of indicators (education, health and income). In all three scenarios, the five main indicators returned by the two correlations are the same and can be observed in Tables 7, 8 and 9.

Table 10. Correlation Between Indicators from Different Strategies (Health)

Indicator	Regression Method					
	Pearson	Spearman	Logarithmic	Linear Uni.	Linear Multi.	Poisson
Indicador 1	0.856944	0.859341	0.63	0.90	0.44	5.48
Indicador 2	0.829384	0.828571	0.40	0.87	0.32	8.69
Indicador 3	0.828375	0.815385	0.34	0.87	0.23	4.15
Indicador 4	0.827313	0.810989	0.28	0.87	0.19	–
Indicador 5	0.660386	0.677711	-0.62	0.62	0.13	−7.28

Table 11. Correlation Between Indicators from Different Strategies (Education)

Indicator	Regression Method					
	Pearson	Spearman	Logarithmic	Linear Uni.	Linear Multi.	Poisson
Indicador 1	0.907816	0.912088	0.26	0.95	0.18	–
Indicador 2	0.889390	0.894505	0.30	0.93	0.24	–
Indicador 3	0.760739	0.752476	0.55	0.76	0.29	11.60
Indicador 4	0.760739	0.589410	–	0.62	–	−15.83
Indicador 5	-0.792058	-0.792298	–	0.75	–	20.12

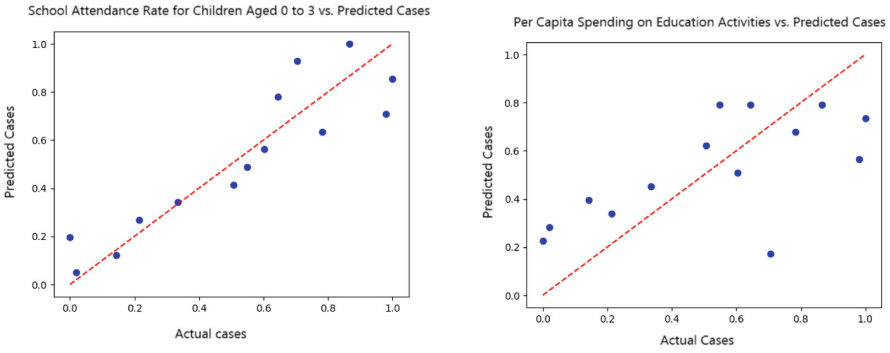
In our subsequent analysis, we evaluated the correlation between the main indicators indicated by the Pearson and Spearman correlations and the score assigned to these indicators by the regression methods. Again, we separated our analyses for each group of indicators (health, education and income) and the results can be seen in Tables 10, 11 and 12. In the cells indicated to "–", the regression method was not able to recover the respective indicator.

Table 12. Correlation Between Indicators from Different Strategies (Income)

Indicator	Regression Method					
	Pearson	Spearman	Logarithmic	Linear Uni.	Linear Multi.	Poisson
Indicador 1	0.887070	0.912088	0.38	1.01	0.30	–
Indicador 2	0.868712	0.829670	0.24	0.97	0.19	–
Indicador 3	0.848366	0.818681	–	0.92	–	–
Indicador 4	0.832164	0.811882	–	0.83	–	3.78
Indicador 5	0.828375	0.815385	–	0.87	–	3.93

Analyzing the results, our first observation is that Univariate Linear Regression follows a similar behavior to Pearson and Spearman, with high correction values for the same indicators indicated by these correlation methods, for the

three sets of indicators. If only the MSE and R^2 results were considered, Univariate Linear Regression would have been wrongly discarded. Another positive aspect of this regression model is that we can perform an evaluation of how a given variable behaves when describing the incidence of syphilis. To exemplify this analysis, we analyzed two indicators related to education, presented in Fig. 1.



(a) School attendance rate for children aged 0 to 3 years old

(b) Per capita Spending on education activities.

Fig. 1. Individualized Indicator Analysis.

In the images above, we can see how both indicators are behaving in relation to the increase/decrease in cases of gestational syphilis. Contrasting the results of Fig. 1 (a) with the error metrics, we have an R^2 of 0.82 and an MSE of 0.02. In other words, it can reproduce 82% of syphilis cases well, and misses 20%. For Fig. 1 (b), we have an R^2 of 0.42 and an MSE of 0.06. This means that it can abstract only 42% of cases well. This analysis is essential due to reliability. The other algorithms end up analyzing everything together, preventing a more precise analysis, and allowing analysis errors. In addition, we saw that Univariate Linear Regression was the only one that brought coherent values in an analysis.

Another regression model that, despite having suffered from multicollinearity, also showed some positive aspects, was the Poisson. It also has reliability support. In our analyses, when suggesting correlations, it already attributes confidence to these correlations. Showing a robust model for an analysis in our context. Furthermore, it presents itself as a good model for modeling countable variables, which is our case, case count. For the other cases, multicollinearity between the indicators prevented deeper analyses of the indicators from being performed. Again, if one of these models had been chosen for the analyses, without a prior study of how they work and how the data are distributed, it could have led to distorted analyses and erroneous conclusions.

4.1 Qualitative Discussion of Evaluated Methods

Statistical Methods. Pearson's correlation measures the linear relationship between two variables, assuming that the data follows a normal distribution. This method is sensitive to outliers, as it relies on mean and variance. However, it is suitable when the relationship between variables is both proportional and linear. In contrast, Spearman's correlation assesses monotonic relationships, which do not necessarily have to be linear. Since it is based on rank rather than absolute values, it does not require normally distributed data and is less affected by outliers. This makes Spearman's correlation particularly useful for detecting non-linear but monotonic relationships, such as exponential or logarithmic trends.

Our analysis revealed that the data exhibits highly complex, predominantly non-linear relationships. When applying Spearman's correlation, we identified strong associations that Pearson's correlation failed to detect. This suggests that Pearson's method may overlook important patterns. In contrast, Spearman's correlation not only confirmed the relationships detected by Pearson but also expanded the set of identified correlations with greater accuracy. Notably, it was the only method to reveal negative correlations, indicating variables inversely associated with syphilis incidence. Given these findings, we conclude that Spearman's correlation is the most appropriate method for analyzing the dataset used in this study, as it provides a more comprehensive and precise understanding of variable relationships. Considering that this research aims to inform social policies to mitigate syphilis cases, the use of reliable methods such as Spearman's correlation is both fundamental and necessary.

Regression Models. Logarithmic regression employs a logarithmic transformation to model nonlinear relationships. The coefficients are interpreted in terms of percentage changes, making this method particularly useful for variables exhibiting large-scale variation. However, it may introduce distortions when applied to datasets with many values close to zero, and it may not adequately capture non-log-linear relationships.

Univariate linear regression examines the relationship between each independent variable and the dependent variable separately. This approach can reveal strong associations that might otherwise be masked in multivariate regression. However, it does not account for interactions between variables, potentially leading to misleading conclusions by overestimating or underestimating their individual importance.

Multivariate linear regression is particularly suitable when analyzing categorical dependent variables (e.g., presence or absence of an event), allowing for the interpretation of coefficients as changes in the probability of the event occurring. Nevertheless, its application to continuous variables can be less intuitive. Additionally, it assumes independence of observations, which poses a challenge when autocorrelation exists between municipalities or regions, as observed in our dataset. Poisson regression is well-suited for modeling count data, employing a probabilistic approach to estimate uncertainty and highlight statistically significant relationships. Moreover, it allows the incorporation of prior knowledge into

the modeling process. However, the interpretation of its coefficients may be less intuitive compared to classical regression methods.

Given the complexity of the relationships within our dataset, relying on a single statistical approach may be insufficient. A combination of methods may be required to fully understand the correlations between syphilis incidence and socioeconomic indicators. Each method presents distinct advantages and limitations, necessitating careful selection based on the specific research objectives. Logarithmic regression is valuable when interpreting marginal effects and percentage-based relationships. Univariate linear regression is useful for identifying initial patterns before developing a more comprehensive model. If the dependent variable is categorical (e.g., presence or absence of syphilis), multivariate linear regression may be more appropriate. Meanwhile, Poisson regression is advantageous when analyzing count data, such as the number of syphilis cases. In our analysis, univariate linear regression emerged as the most suitable method for examining the relationship between syphilis incidence and socioeconomic indicators. Unlike other models, it demonstrated a higher explanatory power, offering deeper insights into individual predictors while avoiding issues related to multicollinearity and model interpretability. Furthermore, its ability to analyze each indicator separately enhances the robustness and clarity of the findings, enabling a more precise evaluation of the factors influencing syphilis incidence. In contrast, other models exhibited limitations such as sensitivity to multicollinearity, reduced explainability, and challenges in handling datasets with a high number of indicators.

4.2 Main Results

Overall, the results demonstrate that the logarithmic, multivariate linear, and Poisson regression models achieved perfect predictive performance across all evaluated cities, with an R^2 of 1.00 and MSE of 0.00. While at first glance this may suggest excellent model performance, such outcomes strongly indicate overfitting, as the models may have memorized the training data rather than generalized underlying patterns. In contrast, the univariate linear regression model showed significantly lower R^2 values, ranging from 0.58 to 0.72 depending on the city, and MSE between 0.026 and 0.044, suggesting a more realistic fit. These findings highlight the importance of carefully evaluating model complexity and addressing potential multicollinearity or data leakage to ensure that predictions are generalizable to unseen data.

5 Conclusion and Future Works

This study aimed to compare different statistical methods for analyzing the relationship between the incidence of gestational syphilis and socioeconomic indicators, specifically health, education, and income. By employing Pearson and Spearman correlation, as well as univariate and multivariate linear regression,

logarithmic regression, and Poisson regression models, we obtained a more comprehensive understanding of the interactions among the analyzed variables.

The findings underscore the critical influence of methodological choices on epidemiological analyses. While Pearson correlation exhibited limitations in capturing non-linear relationships, Spearman correlation proved to be a robust approach for detecting monotonic associations. Univariate regression also demonstrated strong explanatory power in modeling socioeconomic indicators. The comparative analysis of municipalities in Minas Gerais with high incidence rates of gestational syphilis revealed distinct patterns across different regional contexts, reinforcing the necessity of segmenting indicators into categories such as health, education, and income. These insights provide a solid statistical foundation for the formulation of targeted and more effective public health policies.

Given these findings, further research is warranted. A logical next step is to extend this analysis to other regions of Brazil, applying diverse statistical approaches and regression models to refine our understanding of the socioeconomic dynamics driving the spread of sexually transmitted infections. This expanded investigation will contribute to the development of more precise and data-driven strategies for the prevention and control of gestational syphilis at a national level.

Acknowledgment. The authors acknowledge partial funding from the Coordination for the Improvement of Higher Education Personnel - Brazil (CAPES) - Funding Code 001, the National Council for Scientific and Technological Development (CNPq), and the Research Support Foundation of the State of Minas Gerais (FAPEMIG). We also thank the Research and Innovation Support Foundation of Espírito Santo (FAPES) for the resources provided through the PROAPEM project (368/2022 - P: 2022-NGKM5) and the PDPG project (129/2021 - P: 2021-GL60J), which were fundamental to the completion of this work.

References

1. de Administração Pública (ENAP), E.N.: Introdução aos modelos de regressão linear. ENAP (2020). https://repositorio.enap.gov.br/bitstream/1/4788/1/Livro_Regress%C3%A3o%20Linear.pdf
2. Bellégo, C., Benatia, D., Pape, L.: Dealing with logs and zeros in regression models. arXiv preprint arXiv:2203.11820 (2022). https://arxiv.org/abs/2203.11820
3. Benesty, J., Chen, J., Huang, Y.: Pearson correlation coefficient. In: Noise Reduction in Speech Processing, pp. 1–4. Springer (2009). https://doi.org/10.1007/978-3-642-00296-0_5
4. Bennett, D.A.: How can i deal with missing data in my study? Aust. N. Z. J. Public Health **25**(5), 464–469 (2001). https://doi.org/10.1111/j.1467-842X.2001.tb00294.x, https://pubmed.ncbi.nlm.nih.gov/11688629/
5. Chicco, D., Warrens, M.J., Jurman, G.: The coefficient of determination r-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. PeerJ Comput. Sci. **7**, e623 (2021). https://doi.org/10.7717/peerj-cs.623

6. Domingues, R., do C. Leal, M.: Incidência de sífilis congênita e fatores associados à transmissão vertical da sífilis: dados do estudo nascer no brasil. Cadernos de Saúde Pública **32**(6), e0008241 (2016). https://doi.org/10.1590/0102-311X00082415
7. Farrar, D.E., Glauber, R.R.: Multicollinearity in regression analysis: the problem revisited. Rev. Econ. Stat. **49**(1), 92–107 (1967). https://doi.org/10.2307/1937887
8. Favero, M.L.D.C., Ribas, K.A.W., Costa, M.C.D., Bonafé, S.M.: Sífilis congênita e gestacional: notificação e assistência pré-natal. Arch. Health Sci. **26**(1), 2–8 (2019). https://doi.org/10.17696/2318-3691.26.1.2019.1137
9. Hauke, J., Kossowski, T.: Comparison of values of Pearson's and spearman's correlation coefficients on the same sets of data. Quaestiones Geographicae **30**(2), 87–93 (2011). https://doi.org/10.2478/v10117-011-0021-1
10. Heymans, M.W., Twisk, J.W.R.: Handling missing data in clinical research. J. Clin. Epidemiol. **151**, 185–188 (2022). https://doi.org/10.1016/j.jclinepi.2022.08.016, https://www.sciencedirect.com/science/article/pii/S0895435622002506, epub 2022 Sep 21
11. Hilbe, J.M.: Negative binomial regression. Cambridge University Press, pp. 40–58 (2011). https://doi.org/10.1017/CBO9780511973420
12. Jakobsen, J.C., Jesper Kjaer, M.G., Gluud, C.: When and how should multiple imputation be used for handling missing data in Randomised clinical trials – a practical guide with flowcharts. BMC Med. Res. Methodol. **17**(1), 162 (2017). https://doi.org/10.1186/s12874-017-0442-1, https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-017-0442-1
13. Lima, M., Santos, R., Barbosa, G., Ribeiro, G.: Incidência e fatores de risco para sífilis congênita em belo horizonte, minas gerais, 2001–2008. Ciência e Saúde Coletiva **18**(2), 499–506 (2013). https://doi.org/10.1590/S1413-81232013000200023
14. Medeiros, J., Yamamura, M., da Silva, Z., Domingues, C., Waldman, E., Chiaravalloti-Neto, F.: Spatiotemporal dynamics of syphilis in pregnant women and congenital syphilis in the state of são paulo, brazil. Sci. Rep. **12**(1), 585 (2022). https://doi.org/10.1038/s41598-021-04530-y
15. Oliveira, R., Silva, A., Gonçalves, M.: Incidência de sífilis congênita em adolescentes: um estudo epidemiológico na região sudeste do brasil. Revista Paulista de Pediatria **39**(1), 1–8 (2021). https://doi.org/10.1590/1984-0462/2021/39/0101
16. Pedregosa, F., et al.: Scikit-learn: Machine learning in python - label encoding (2011). https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html
17. Pedregosa, F., et al.: Scikit-learn: Machine learning in python - minmaxscaler (2011). https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html
18. Pedregosa, F., et al.: Scikit-learn: Machine learning in python - standardscaler (2011). https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html
19. Ramos, R., Carneiro, G., Oliveira, A., Cunha, T., Ramos, V.: Incidence of congenital syphilis according to inequalities and living conditions in the city of Recife, Pernambuco, Brazil. Revista Brasileira de Saúde Materno Infantil **21**(1), 1–10 (2021). https://doi.org/10.1590/1806-93042021000100002
20. Shrestha, N.: Detecting multicollinearity in regression analysis. Am. J. Appl. Math. Stat. **8**(2), 39–42 (2020)
21. da Silva, M.A., Bustamante-Teixeira, M.T.: Uso da imputação múltipla de dados faltantes: uma simulação utilizando dados epidemiológicos. Cad. Saude Publica **25**(2), 268–276 (2009)

22. da Silva, P., Almeida, J., Santos, M.: A Bayesian approach to understand congenital syphilis distribution: a spatiotemporal analysis in Brazil. Sci. Rep. **11**, 1–10 (2021). https://doi.org/10.1038/s41598-021-04530-0
23. Singh, D.D., Singh, A., Kumar, A., Singh, R.: Selection of appropriate statistical methods for data analysis. Ann. Card, Anaesth. **22**(3), 297–301 (2019). https://doi.org/10.4103/aca.ACA_248_18, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6639881/
24. Soares, K., Prado, T.N., Zandonade, E., Moreira-Silva, S.F., Miranda, A.E.: Análise espacial da sífilis em gestantes e sífilis congênita no estado do espírito santo, 2011–2018. Epidemiologia e Serviços de Saúde **29**(1), e2018193 (2020). https://doi.org/10.5123/S1679-49742020000100018
25. Souza, V., Mendes, A., Costa, L.: Desigualdades na incidência de sífilis congênita no Brasil: uma abordagem espacial. Revista de Saúde Pública **56**(1), 1–12 (2022). https://doi.org/10.11606/S1518-8787.2022056003800
26. Tsuboi, M., et al.: Prevalence of syphilis among men who have sex with men: a global systematic review and meta-analysis from 2000-20. Lancet Global Health **9**(8), e1110–e1118 (2021). https://doi.org/10.1016/S2214-109X(21)00221-7, epub 2021 Jul 8