Examining the Relationship Between Socioeconomic Factors and HIV Prevalence in the United States

Principal Investigators: [Jiahe Ling, Meishu Zhao, Yani Sun] ([jling9@wisc.edu])

Introduction

According to the World Health Organization, there are an estimated 1.2 million people with HPV in the United States as of 2022. In the absence of a cure, the government must strategize effectively to address this ongoing public health crisis. In this project, we explore the relationship between various socioeconomic factors and HIV prevalence rates in the United States. Identifying the social and economic factors that may affect the spread and management of HIV is the main goal of our study. The HIV dataset contains information on prevalence, diagnosis, and death rates at the county level. We analyzed this dataset in conjunction with Social Vulnerability Index (SVI) data from the 2016-2020 American Community Survey (ACS).

Through our analysis, we identified five important variables in the SVI dataset that were strongly correlated with HIV prevalence. These variables are the percentage of Black/African Americans, the overall percentage of minorities, the percentage of households in which housing costs are a burden on income, the percentage of single-parent households with children under the age of 18, and the percentage of households without a car. These findings suggest that communities with higher percentages in these categories tend to have increased HIV prevalence rates. This insight is crucial for developing targeted public health strategies and interventions, particularly in areas where these socioeconomic factors are more prevalent.

Analysis

Data Preparation

We used two key datasets to conduct our research: the 2020 CDC/ATSDR Social Vulnerability Index (SVI) and HIV prevalence data, both at the county level. The SVI dataset includes data on four key themes: Socioeconomic Status, Household Characteristics, Racial & Ethnic Minority Status, Housing Type and transportation, and an overall vulnerability ranking. To ensure the reliability and accuracy of our analysis, we removed duplicates and addressed missing values by replacing them with the most common value. After data cleaning, we have around 3100 US counties in the data with over 150 socioeconomic variables.

Time Series Analysis of HIV Statistics

We initially conducted a time series analysis focusing on the trends of HIV prevalence rates among different racial groups over the years. This analysis not only highlighted the general trend of HIV prevalence but also underscored the need to develop models better targeting vulnerable populations, revealing disparities in how HIV affects diverse racial groups. Furthermore, this analysis revealed an increasing trend in HIV diagnoses and deaths after 2020, coinciding with the COVID-19 pandemic, suggesting a growing public health concern post-COVID. The time series plots provided a clear indication of the evolving nature of the HIV epidemic, reinforcing the importance of timely and effective interventions.

Figure 1: Time Series Plot of HIV Data by Case for Difference Race

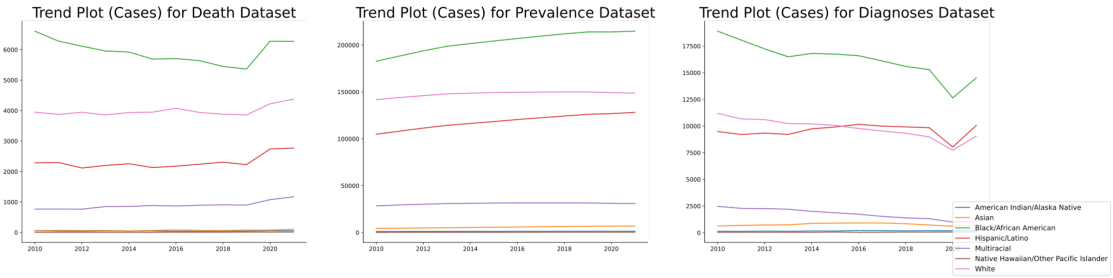


Figure 2: Time Series Plot of HIV Data by Rate for Difference Race

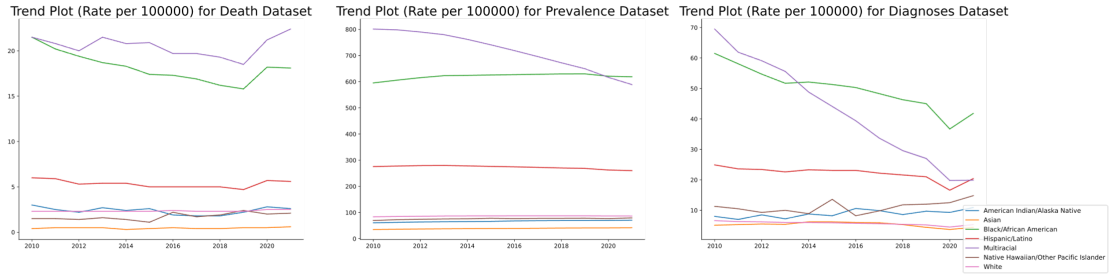


Figure 3: Time Series Plot of Change of HIV Data by Case for Difference Race

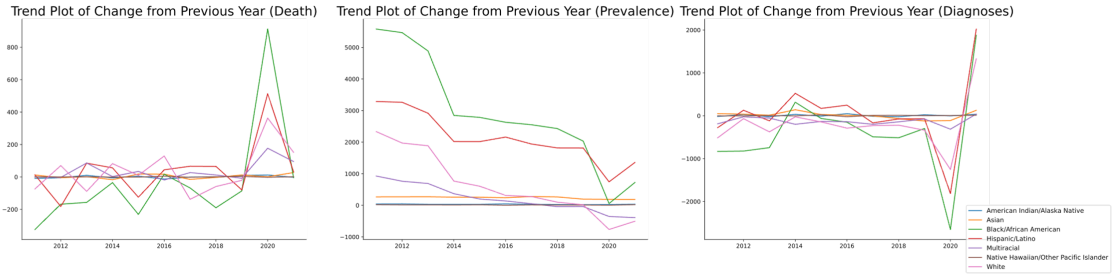
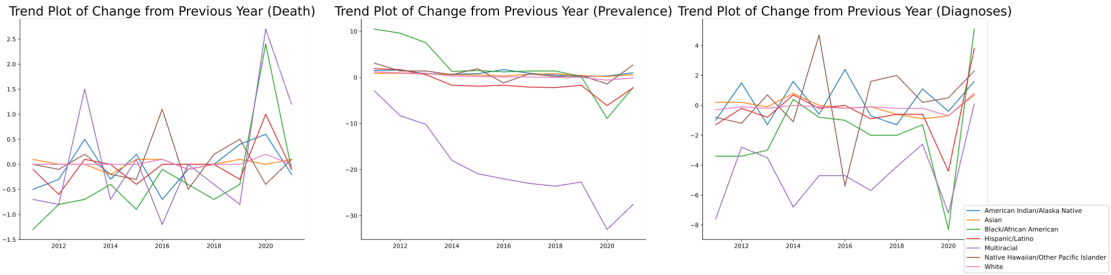


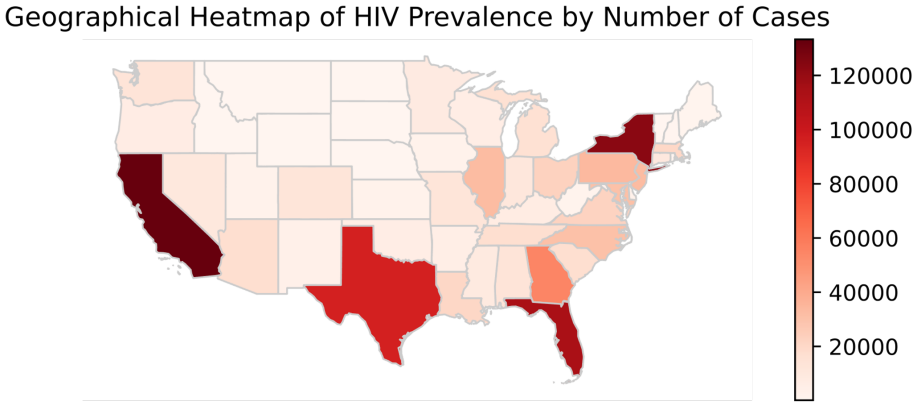
Figure 4: Time Series Plot of Change of HIV Data by Rate for Difference Race



Geographic Heat Mapping

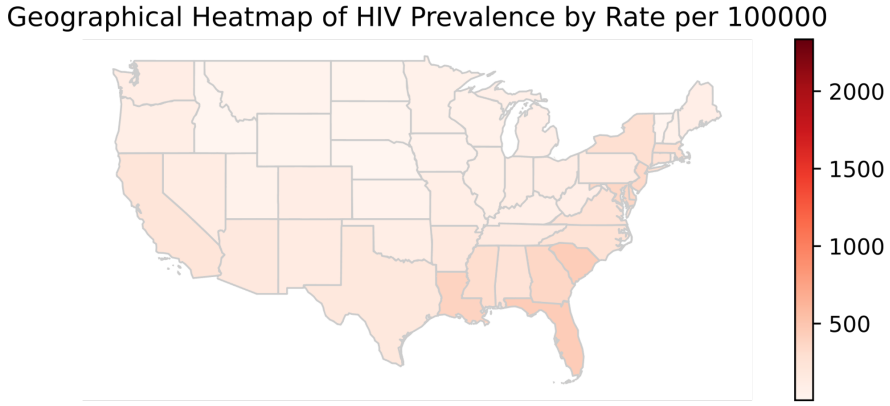
Following the time series analysis, we utilized a geographic heat map to visualize HIV prevalence rates across different states in the United States. This heat map of HIV cases (Figure 5) highlighted regional disparities in HIV prevalence, emphasizing the need for a localized understanding and intervention. However, after standardizing the HIV cases by converting them to rates per 100,000 individuals, as shown in Figure 6, the disparities appeared less significant. Nonetheless, we observed a marginally higher rate in the southern part of the U.S. The darker regions on the map indicate areas where targeted public health strategies might be most urgently required.

Figure 5: Heatmap of HIV Prevalence By Case



Notes: The states indicated by darker red colors represent a larger number of cases. However, this map is not adjusted for the population size of each state.

Figure 6: Heatmap of HIV Prevalence By Rate

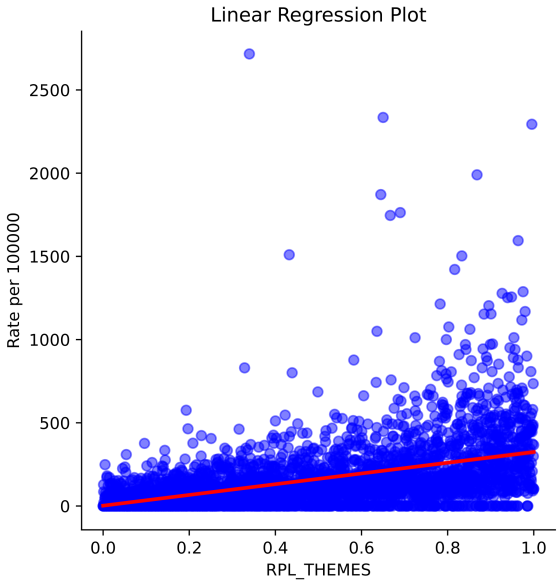


Notes: Southern states, such as California, Texas, and Florida, exhibit slightly higher HIV prevalence rates.

Initial Regression Analysis

The time series analysis and heat map demonstrated the necessity of a model precisely targeting HIV-risk populations. Our subsequent, more focused investigations identified specific variables that might impact HIV prevalence. We began by examining the relationship between the overall vulnerability index (RPL_THEMES) from the SVI dataset and HIV prevalence rates. Interestingly, this analysis did not yield statistically significant results. However, a positive trend was observed, suggesting a potential association between social vulnerability factors and HIV prevalence. This finding led us to consider the need for a more refined or comprehensive model to capture this relationship more accurately.

Figure 7: HIV Prevalence vs. SVI Overall Vulnerability Ranking



Notes: The RPL_THEMES represents the overall vulnerability ranking for U.S. counties, where a higher score (approaching 1) indicates greater vulnerability. As this vulnerability score increases, a corresponding rise in HIV prevalence rates is observed.

Correlation Analysis

In our analysis, we used the CDC's categorization of SVI numerical variables: "E_" for estimates, "EP_" for percentage estimates, "EPL_" for percentile percentages, and "F_" for binary flags. We constructed correlation tables for each group, highlighting in red those variables with a high correlation with the HIV prevalence rate per 100,000. Through our correlation matrices, we identified five significant continuous variables that showed the strongest correlation (over 0.35) with the HIV prevalence rate. These included factors such as the percentage of Black/African Americans, the overall percentage of minorities, and the percentage of households without a car. Interestingly, our analysis also revealed that the top categorical variables were essentially continuous variables converted into categorical forms, thus guiding our focus towards these key continuous variables for further analysis.

Figure 8: Correlation matrix for E (estimates)

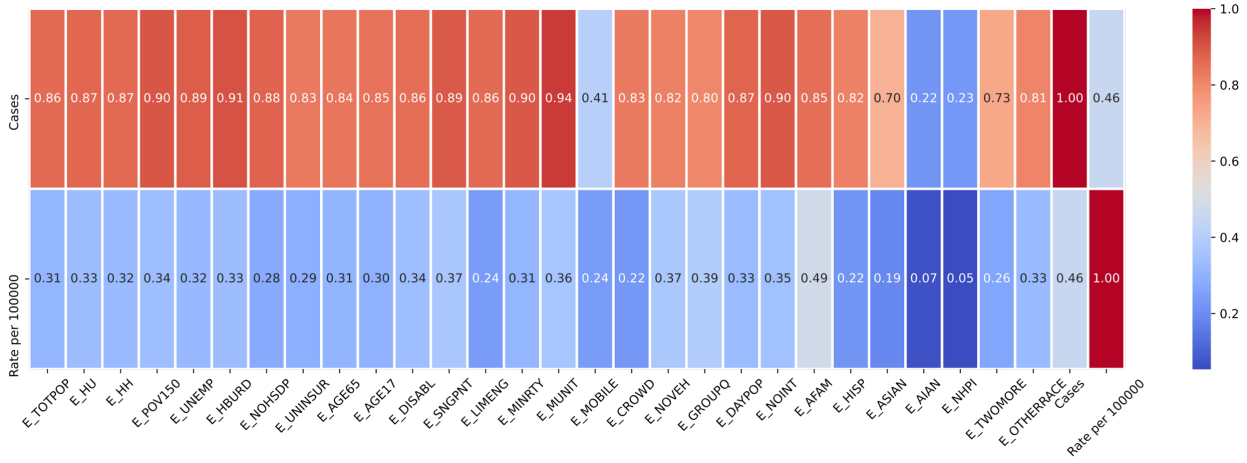


Figure 9: Correlation matrix for EP (percentage estimates)

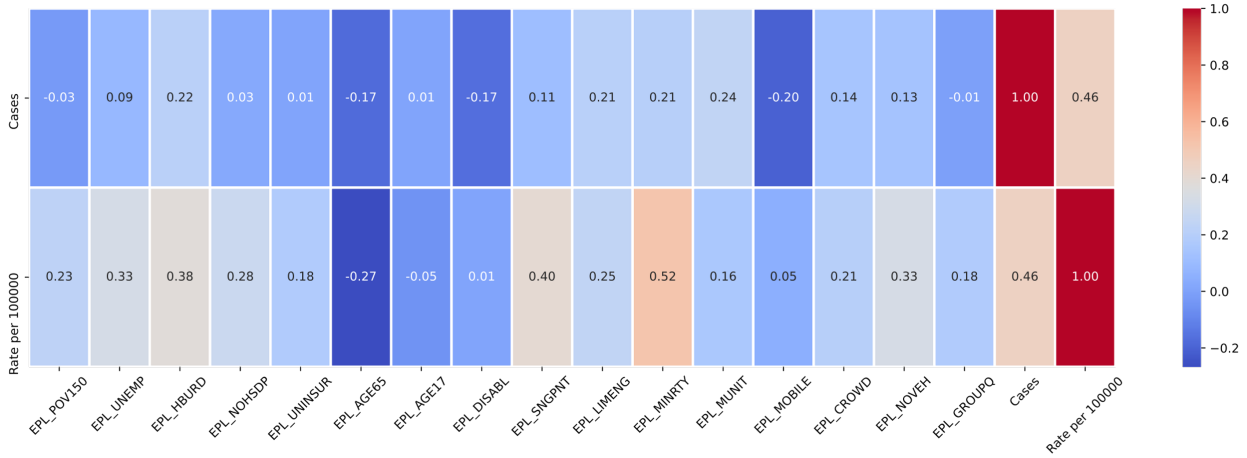
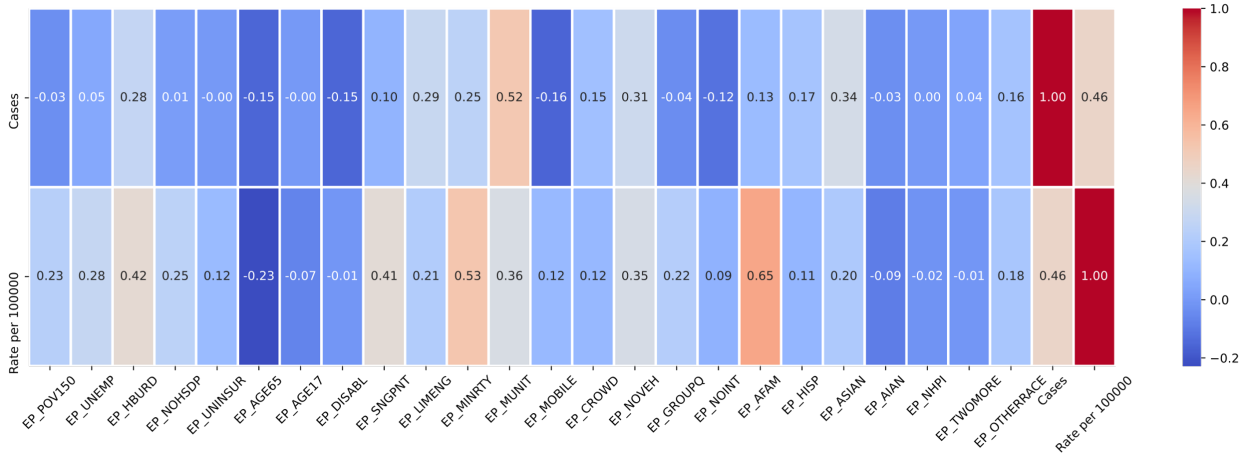


Figure 10: Correlation matrix for EPL (percentile percentages)



Note: For the neatness of the picture, the meaning of the names of the variables is not shown in the figure. We will explain the important variables in the subsequent regression modeling section.

Regression Modeling

The results from our OLS regression model provide information of how specific SVI variables relate to the HIV prevalence rate per 100,000 individuals. The model's R-squared value of 0.560 suggests that about 56% of the variability in HIV prevalence rates can be explained by these variables.

Interpretation of Variables:

EP_AFAM (Percent of African Americans): Coefficient: 7.1090, P-value: < 0.001

> Each unit increase in the percentage of African Americans in a population is associated with an increase of 7.109 cases of HIV per 100,000 individuals. This strong positive correlation indicates that African American communities may be disproportionately affected by HIV, highlighting an area for targeted public health interventions.

EP_MINRTY (Percent of Minorities): Coefficient: 1.4263, P-value: < 0.001

> A one-unit increase in the percentage of minority populations is correlated with an increase of 1.4263 HIV cases per 100,000 individuals. This suggests that minority communities have a higher vulnerability to HIV prevalence.

EP_HBURD (Housing Burden): Coefficient: 5.1352, P-value: < 0.001

> For one percent increase in housing burden, there is an associated increase of 5.1352 HIV cases per 100,000 individuals.

EP_SNGPNT (Single Parent Households): Coefficient: 0.5818, P-value: 0.660

> Although the coefficient suggests a slight increase in HIV prevalence with more single-parent households, the high p-value indicates that this finding is not statistically significant.

E_NOVEH (Households without Vehicles): Coefficient: 0.0029, P-value: < 0.001

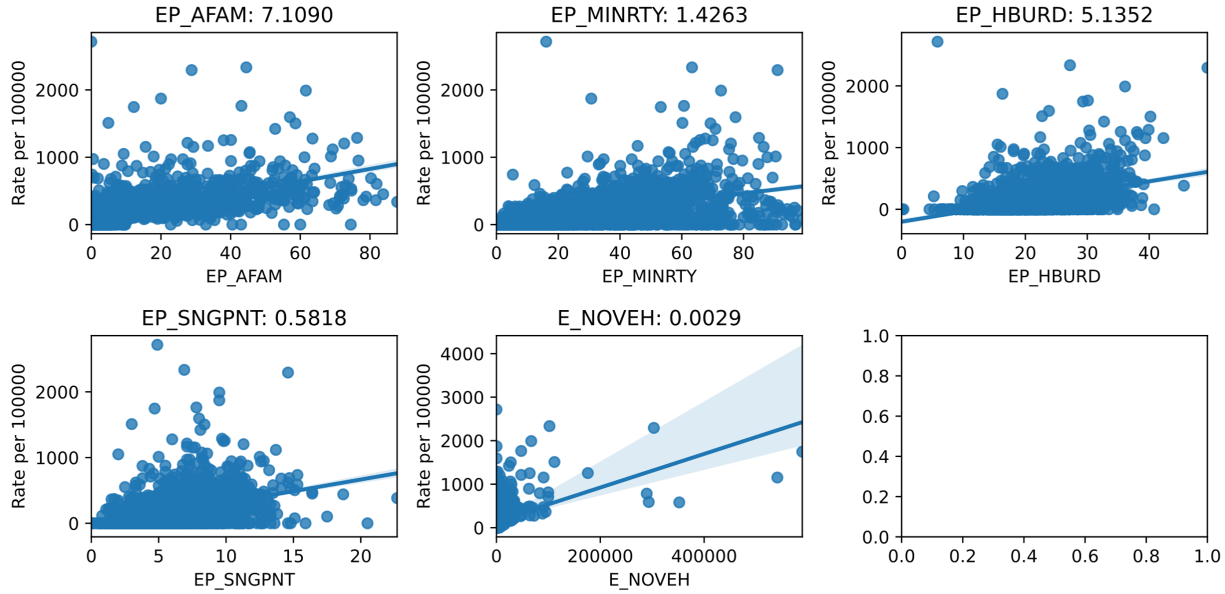
> For one unit increase in the percentage of households without vehicles, there is an associated increase of 0.0029 HIV cases per 100,000 individuals. This is an interesting variable and tell us accessibility to transportation tools may associate with HIV prevalence.

Table 1: Summary statistics

Variable	Full Name	Coefficient	Std. Error	Significance
EP_AFAM	Percent of African Americans	7.1090	0.220	Significant
EP_MINRTY	Percent of Minorities	1.4263	0.165	Significant
EP_HBURD	Housing Burden	5.1352	0.516	Significant
EP_SNGPNT	Single Parent Households	0.5818	1.322	Not Significant
E_NOVEH	Households without Vehicles	0.0029	0.000	Significant

The findings in our OLS regression model uncovered the disproportionate impact of HIV on African American and minority communities and highlight the role of housing and transportation. While the relationship with single-parent households requires further investigation due to its statistical insignificance, the overall model provides a valuable tool for public health planning and targeted interventions, emphasizing areas that require the most attention.

Figure 11: Correlation matrix for EPL (percentile percentages)



Conclusions and directions for future research

Our study has uncovered a significant relationship between certain socioeconomic factors and HIV prevalence rates in the United States, highlighting a clear correlation between higher percentages of vulnerable populations, as indicated by the SVI, and increased HIV prevalence. This finding is crucial for public health authorities in formulating targeted interventions, particularly in communities where these vulnerabilities are more pronounced. For future research, there is a need to broaden the study's scope by including additional socioeconomic factors, which could reveal further influences on HIV prevalence. Adopting a longitudinal approach to examine changes over time would provide deeper insights into how these relationships evolve. Moreover, exploring more complex models that capture the multifaceted nature of socioeconomic factors and their interplay with public health outcomes represents a valuable direction for advancing our understanding and response to the HIV epidemic.

Table 2: OLS Summary

OLS Regression Results

Dep. Variable:	Rate per 100000	R-squared:	0.560
Model:	OLS	Adj. R-squared:	0.559
Method:	Least Squares	F-statistic:	798.6
Date:	Thu, 14 Dec 2023	Prob (F-statistic):	0.00
Time:	18:08:29	Log-Likelihood:	-19905.
No. Observations:	3143	AIC:	3.982e+04
Df Residuals:	3137	BIC:	3.986e+04
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-61.6664	11.730	-5.257	0.000	-84.666	-38.667
EP_AFAM	7.1090	0.220	32.271	0.000	6.677	7.541
EP_MINRTY	1.4263	0.165	8.668	0.000	1.104	1.749
EP_HBURD	5.1352	0.516	9.953	0.000	4.124	6.147
EP_SNGPNT	0.5818	1.322	0.440	0.660	-2.010	3.174
E_NOVEH	0.0029	0.000	22.380	0.000	0.003	0.003

Omnibus:	3428.401	Durbin-Watson:	1.695
Prob(Omnibus):	0.000	Jarque-Bera (JB):	716553.466
Skew:	5.097	Prob(JB):	0.00
Kurtosis:	76.265	Cond. No.	9.70e+04

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 9.7e+04. This might indicate that there are
strong multicollinearity or other numerical problems.

Citation

"HIV Data and Statistics." *World Health Organization*, World Health Organization, www.who.int/teams/global-hiv-hepatitis-and-stis-programmes/hiv/strategic-information/hiv-data-and-statistics. Accessed 9 Nov. 2023.