Multiple Regression Analysis on Survival Rates

Introduction

This analysis focuses on understanding the factors influencing cancer survival rates using a multiple regression model. The dependent variable in this study is the survival rate, and various socioeconomic, educational, racial, and geographic variables were considered as predictors. Additionally, a multicollinearity analysis was conducted to ensure the robustness of the regression model.

Methodology

A multiple regression analysis was performed with survival rate as the dependent variable. The independent variables included in the initial model were:

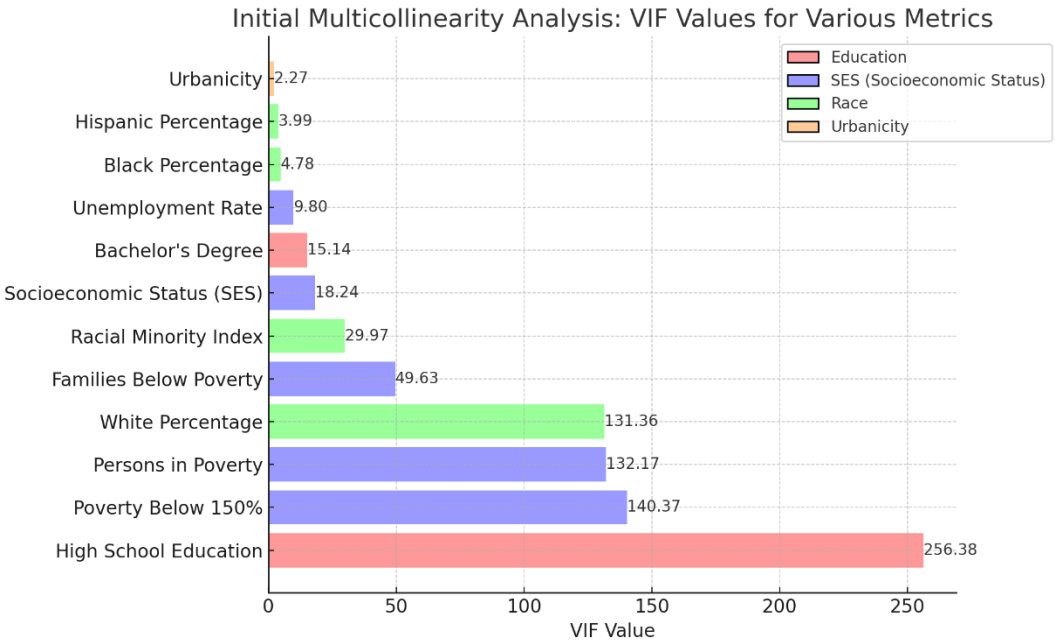
Education	Poverty	Demographics	Urbanization
High School Education	Poverty Below 150%	White Percentage	Urbanicity
Bachelor's Degree	Persons in Poverty	Black Percentage	
	Families Below Poverty	Hispanic Percentage	
	Unemployment Rate	Racial Minority Index	

A multicollinearity analysis was conducted using Variance Inflation Factors (VIF) to identify highly correlated variables, which could potentially distort the regression coefficients.

Results

1. Initial Multicollinearity Analysis

Figure 1: *Initial Variance Inflation Factors*

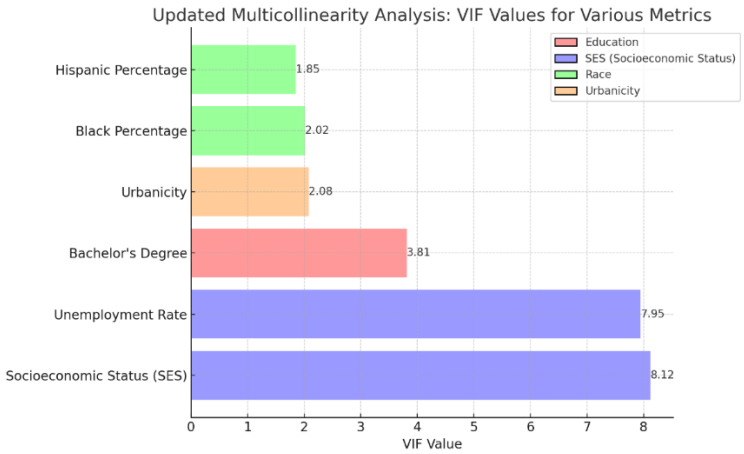


These initial VIF values indicated significant multicollinearity, particularly among education and poverty-related variables.

2. Updated Multicollinearity Analysis

After feature selection to address multicollinearity, the following variables were retained:

Figure 2: *Updated Variance Inflation Factors*



All updated VIF values are below the common threshold of ten, indicating acceptable levels of multicollinearity.

3. Model Performance

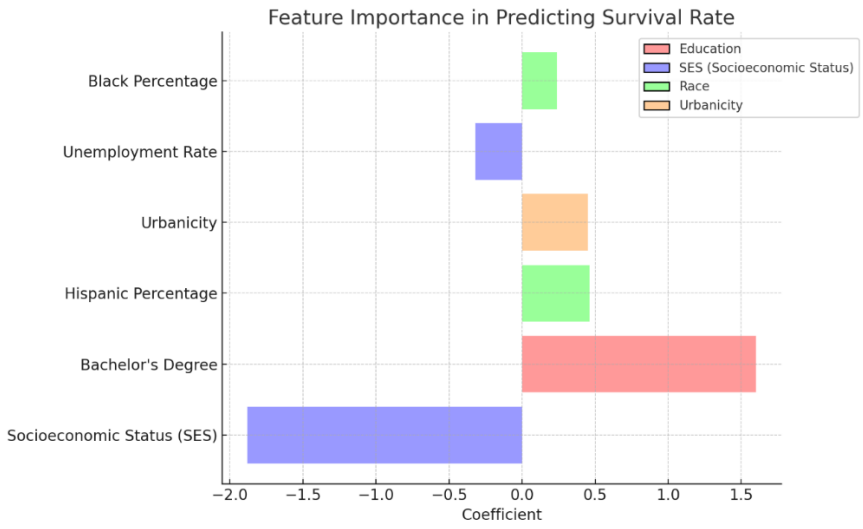
Mean Squared Error (MSE): 23.75

R-squared Score: 0.2915

The R-squared score of 0.2915 suggests that approximately 29.15% of the variance in the survival rate is explained by the selected features. This indicates that while the model has some predictive power, other key factors influencing survival rates may not have been included in the analysis.

4. Feature Coefficients

Figure 3: *Feature Coefficients for Survival Rate Prediction*



Key Findings and Interpretations

1. **SES (Socioeconomic Status)**: The negative coefficient for SES is surprising and suggests a counterintuitive relationship with survival rates. This finding warrants further investigation to understand whether this result is due to the way SES is calculated or other confounding factors.

2. **Bachelor's Degree**: A strong positive relationship with survival rates supports the well-established link between higher education levels and better health outcomes.

3. **Hispanic and Black Percentages**: The positive relationships between these racial demographics and survival rates are intriguing. These findings might relate to the "Hispanic Paradox," where Hispanic populations often experience better health outcomes despite lower socioeconomic status. Further exploration of underlying factors is necessary.

4. **Urbanicity**: The positive coefficient for urbanicity suggests that urban areas, due to better access to healthcare facilities, tend to have higher survival rates.

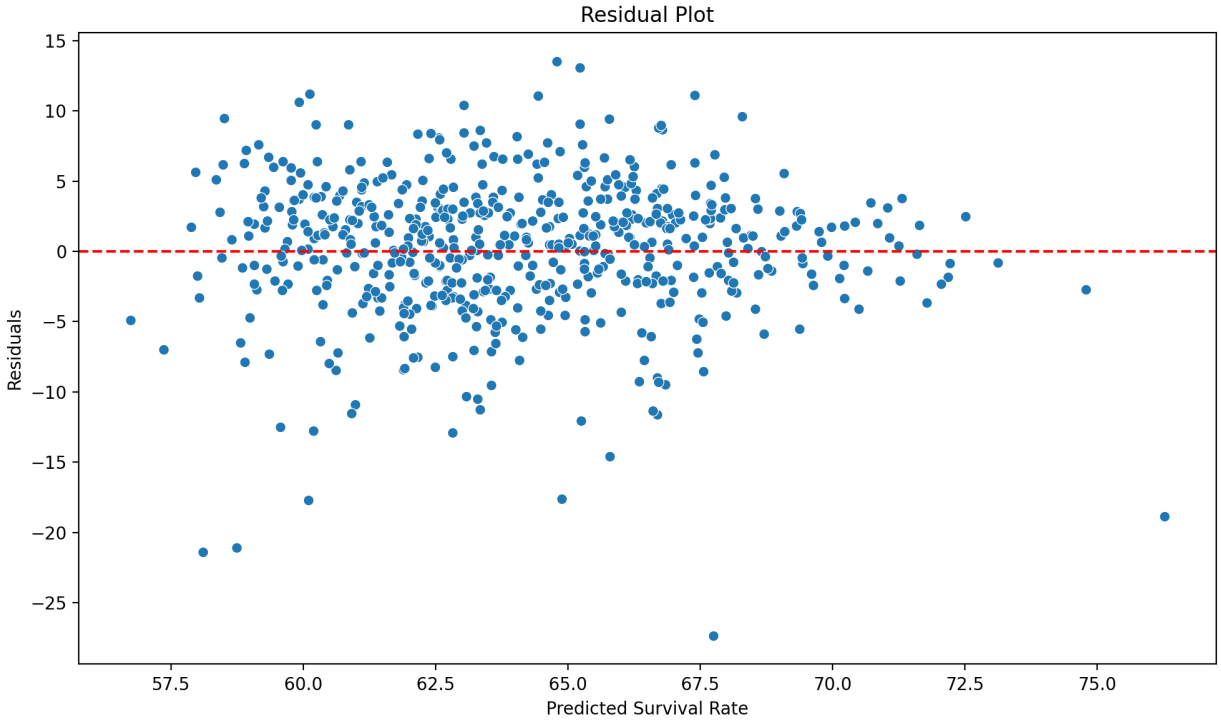
5. **Unemployment Rate**: The negative relationship between unemployment rate and survival rates aligns with the understanding that employment status is a critical determinant of health.

6. **Residual Analysis**: The residual plot indicated patterns suggesting potential non-linear relationships or omitted variables not captured by the linear model.

Next Steps

1. **Investigate SES**: The unexpected negative relationship between SES and survival rates should be explored further. This may involve re-evaluating the SES calculation or identifying confounders.
2. **Explore Non-linear Relationships**: Given the patterns observed in the residual plot, it may be beneficial to consider non-linear modeling approaches or interaction terms in the regression model.
3. **Incorporate Additional Variables**: With only 29% of the variance in survival rates explained by the current model, exploring other potential predictors could improve model performance.
4. **Detailed Urbanicity Analysis**: Urbanicity's impact on survival rates could be further dissected by examining more specific urban characteristics or categorizing different urban areas.
5. **Health Policy Implications**: The positive associations with Hispanic and Black populations should be explored further to understand how these findings can inform targeted health policies and interventions.
6. **Advanced Modeling Techniques**: Consider using more sophisticated models, such as random forests or gradient boosting machines, which can automatically account for complex interactions and non-linearities in the data.

Figure 4: *Residual Plot of Survival Rate Model*



Conclusion

This multiple regression analysis provides insight into the complex factors influencing cancer survival rates. While the model highlights the significant roles of education, race, and urbanicity, the unexpected findings related to SES call for further research. Future efforts should focus on refining the model, exploring non-linear relationships, and incorporating additional variables to enhance our understanding of survival rates and inform public health strategies.