Data Analysis in Cancer Research: Mapping Socio-Economic Influences

Padmanabh Butala

May 20, 2024

1 Introduction

This analysis explores the intricate relationships between socioeconomic factors and cancer incidence and mortality rates across various counties in the United States. Utilizing a comprehensive dataset encompassing 3,072 counties, this study aims to discern how variables such as poverty levels, median income, and population estimates correlate with cancer statistics like incidence and death rates. The data, derived from reliable health and economic sources, includes diverse metrics such as the average annual count of cancer cases, median income levels, and population estimates as of 2015.

The significance of this analysis lies in its potential to inform public health policies and resource allocation. By understanding the socioeconomic underpinnings of cancer incidence and mortality, policymakers and health practitioners can better tailor interventions and support mechanisms to the needs of specific communities, ultimately aiming to reduce the burden of cancer across different socio-economic strata. This report seeks to provide a detailed statistical foundation to aid in such decision-making processes, emphasizing the importance of socio-economic context in health outcomes.

2 Methodology

This study employs a combination of statistical tools and software to analyze the relationships between socioeconomic factors and cancer-related metrics. The methodology is structured into two main segments: summary measures and regression analysis.

2.1 Summary Measures

For the summary measures, both Tableau and R were utilized to conduct descriptive statistical analysis. Tableau provided a robust platform for visual exploration of the data, allowing for the effective representation of distributions and trends across multiple variables such as poverty levels, median income, and cancer incidence rates. Concurrently, R was employed for its powerful statistical capabilities, facilitating the computation of central tendencies, variability, and other summary statistics which are critical in understanding the overall characteristics of the dataset.

2.2 Regression Analysis

The regression analysis was conducted exclusively using R due to its comprehensive suite of statistical modeling tools. The focus was on building linear regression models to identify and quantify the relationships between socioeconomic indicators (such as median income and poverty estimates) and cancer outcomes (incidence and death rates). The process involved data preparation steps including cleaning and normalization, followed by model selection, where various models were tested to find the best fit based on standard criteria like R-squared. Diagnostic plots and residual analysis were also performed to ensure the validity of the model assumptions.

The tools and analytical techniques chosen for this study were based on their proven efficacy in handling large datasets and their ability to provide deep insights through statistical analysis. This methodological approach aims to ensure the reliability and accuracy of the findings, providing a solid foundation for the subsequent results and discussion sections.

3 Results

3.1 Summary Measures

3.1.1 Geographic Distribution of Cancer Rates

This subsection presents a series of visualizations depicting the geographic distribution of cancer incidence and death rates across counties and states. These visualizations effectively highlight regional disparities and trends within the dataset.

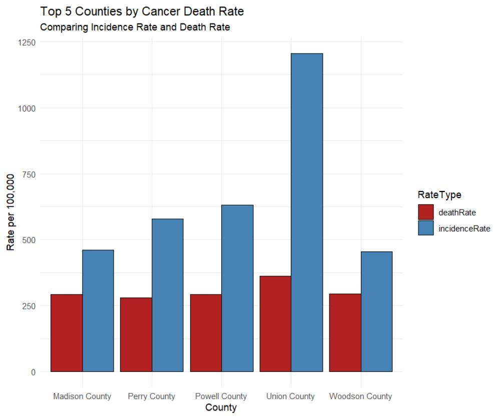

Figure 1: Top 5 Counties by Cancer Death Rate: This bar chart compares the incidence and death rates across the top five counties. It reveals significant variations in both metrics, suggesting disparities in cancer outcomes at the county level.

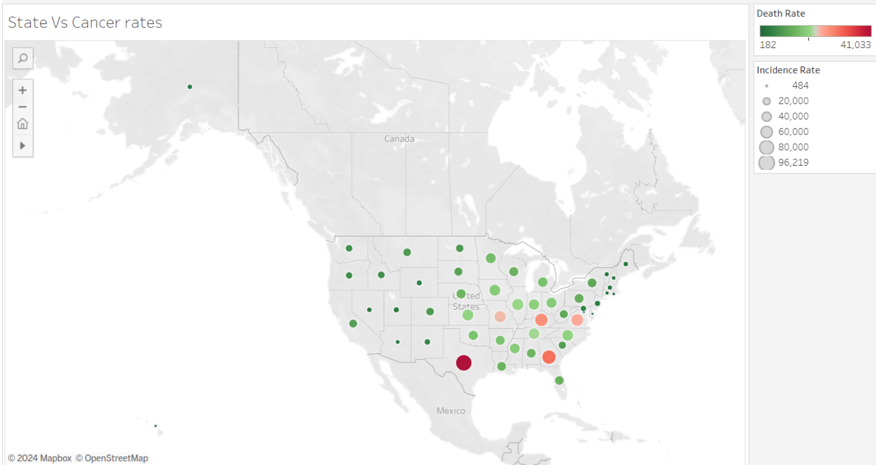

Figure 2: State vs Cancer Rates: This map provides a comprehensive overview of cancer incidence and death rates across the United States, using color coding to represent different intensities. The darkest red spot highlights the area with the highest death rate, serving as a crucial point for potential targeted health interventions.

3.1.2 Interactive Dashboard for Exploring Cancer Rates by Socioeconomic Factors

The interactive dashboard developed using Tableau enables users to explore variations in cancer rates across different socioeconomic contexts. Here are the key observations derived from the analysis:

1. **Cancer Rates vs. Median Income:** - The scatter plots display a negative correlation between median income and both cancer incidence and death rates. As median income increases, there is a general trend of decreasing cancer incidence and death rates. However, there is considerable variability, especially at lower income levels. - The regression lines, overlaid on the plots, show that while the trend is generally negative, the strength and consistency of this relationship vary across different income brackets.

2. **Cancer Rates vs. Poverty Levels:** - The relationship between poverty levels and cancer rates is illustrated through scatter plots for both death and incidence rates. Similar to income, there is a noticeable trend where higher poverty levels often correlate with higher cancer rates. - The plots show significant clustering at lower poverty levels, with a spread of values increasing as poverty levels rise. This suggests that poverty might be a more pronounced factor in cancer rate disparities at higher poverty thresholds.

3. **Geographic Variation in Cancer Rates:** - An interactive map highlights the geographic distribution of cancer rates across the United States, color-coded by recent trends in cancer rates (falling, rising, stable). This visualization points to regional disparities in cancer outcomes, with some states showing predominantly rising trends while others are stable or falling. - The map serves as a critical tool for visualizing how socio-economic factors and cancer rates are geographically intertwined, highlighting specific areas for targeted health interventions and policy focus.

These visualizations underscore the complex interplay between socioeconomic factors and cancer outcomes, providing a visual and interactive means to explore these relationships deeply.

3.1.3 Advanced Visualization of Multi-variable Relationships

This subsection features advanced visualizations that illustrate the complex interrelationships between cancer rates, population estimates, median income, and poverty levels, with state as a categorical variable.

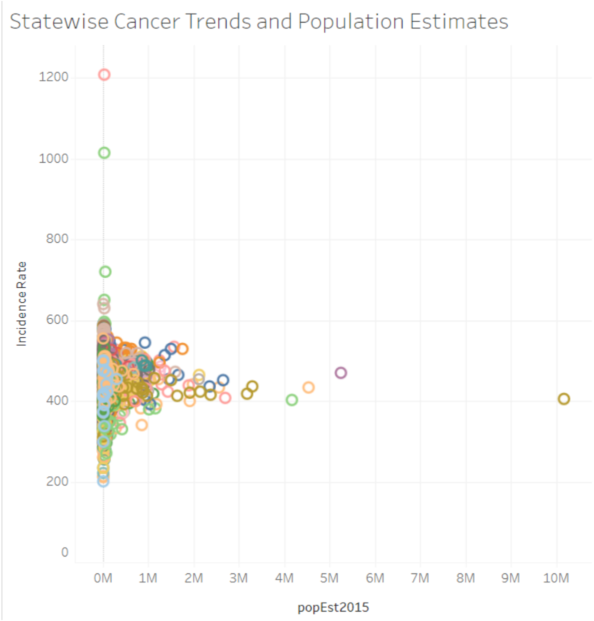

Figure 3: Statewise Cancer Trends and Population Estimates: This scatter plot shows the incidence rates of cancer against the population estimates for 2015. Each point is color-coded by state, highlighting the variations in cancer incidence relative to population sizes across different states. Larger populations do not necessarily correlate with higher incidence rates, suggesting other factors are at play.

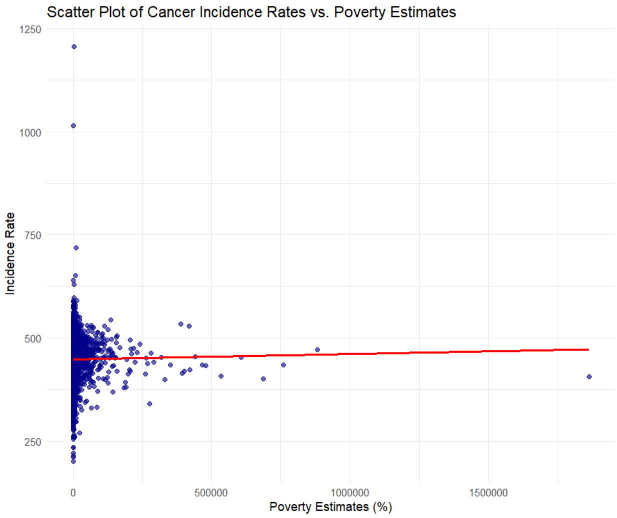

Figure 4: Scatter Plot of Cancer Incidence Rates vs. Poverty Estimates: This plot examines the relationship between cancer incidence rates and poverty levels, showing a wide distribution of incidence rates across varying poverty percentages. The trend line indicates a very slight upward trajectory, suggesting a weak correlation between higher poverty levels and increased cancer rates.

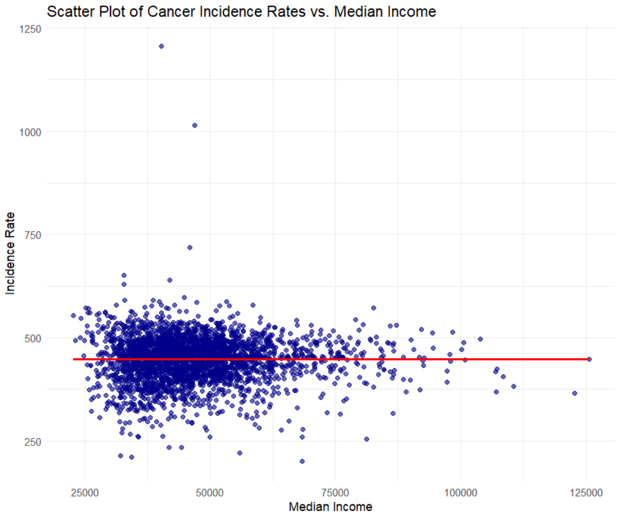

Figure 5: Scatter Plot of Cancer Incidence Rates vs. Median Income: This visualization explores the incidence rates of cancer in relation to median income. The relatively flat trend line across the income spectrum indicates that median income alone may not be a strong predictor of cancer incidence rates, highlighting the complexity of factors influencing cancer trends.

3.1.4 Identification of Outliers in Cancer Rates

This subsection features a visualization that identifies significant outliers in cancer rates across different counties, emphasizing those with exceptionally high or low cancer rates relative to the general population.

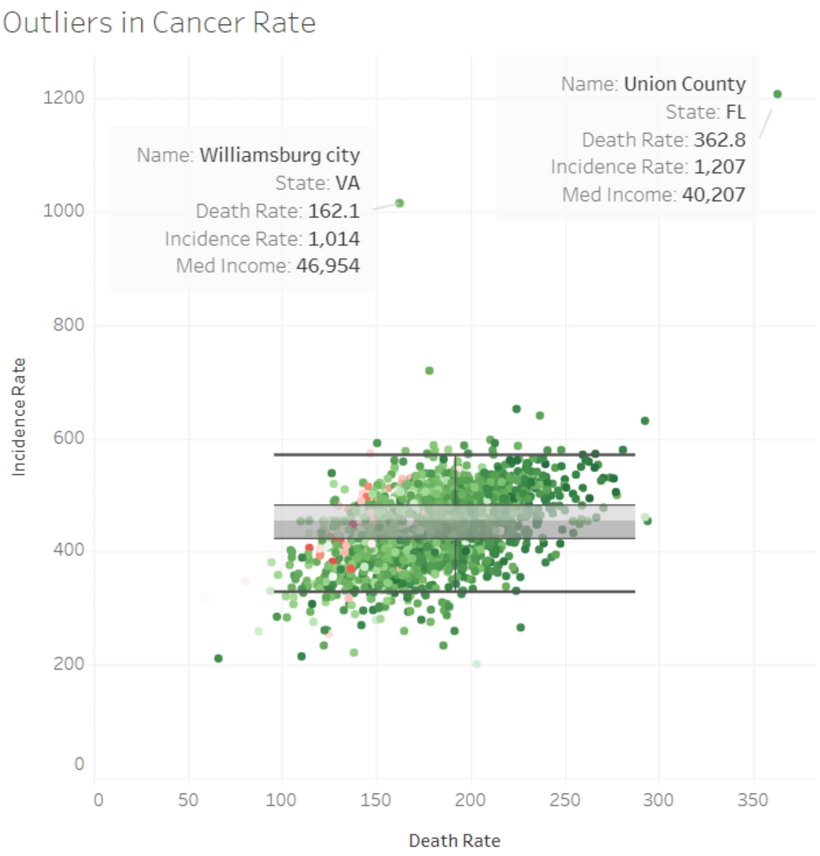

Figure 6: Outliers in Cancer Rate: This scatter plot identifies counties with notably high or low cancer incidence and death rates. Highlighted are Williamsburg City in Virginia and Union County in Florida, both of which show exceptional cancer incidence and death rates significantly diverging from the national average. These outliers are critical for targeted medical interventions and deeper epidemiological investigations to understand the factors contributing to these anomalies.

3.1.5 Descriptive Statistics

This subsection presents descriptive statistics including mean, median, mode, range, variance, and standard deviation for the key variables of median income, poverty estimates, and cancer incidence rates.

Table 1: Descriptive Statistics for Median Income, Poverty Estimates, and Cancer Incidence Rates

Variable	**Mean**	**Median**	**Mode**	**Range**	**Variance**	**Std. Deviation**
Median Income	$47,091	$45,201	$34,116	$22,640 - $125,635	145,925,542	$12,080
Poverty Estimates	15,680	4,436	1,544	91 - 1,863,025	3,152,457,771	56,147
Cancer Incidence Rates	448	454	454	201 - 1,207	2,968	54.5

3.1.6 Categorization of Counties by Income and Cancer Rate Quantiles

This subsubsection presents the categorization of counties based on quantiles of median income and cancer rates, facilitating the analysis of socio-economic patterns in relation to cancer prevalence.

Table 2: Categorization of Counties by Quantiles of Median Income and Cancer Rates

Income Quantile	Cancer Rate Quantile	Count
1	1	235
1	2	167
1	3	157
1	4	209
2	1	183
2	2	193
2	3	192
2	4	200
3	1	170
3	2	204
3	3	216
3	4	178
4	1	180
4	2	204
4	3	203
4	4	181

3.2 Regression Analysis

3.2.1 Interpretation of Regression Coefficients

The regression model summary provides insights into the relationship between socio-economic factors and cancer incidence rates, based on the following formula:

$$\text{incidenceRate} = \beta_0 + \beta_1 \times \text{medIncome} + \beta_2 \times \text{PovertyEst} + \epsilon$$

- **Intercept** (β_0): At a median income and poverty estimate of zero, the expected cancer incidence rate is approximately 449 per 100,000 people.
- **Median Income** (β_1): The coefficient of -0.00002062 indicates a small negative effect on cancer incidence rates, though not statistically significant (p-value = 0.801).
- **Poverty Estimate** (β_2): The coefficient of 0.00001353 suggests a slight positive association with cancer incidence rates, but this too is not statistically significant (p-value = 0.443).

Model Fit:

- **Multiple R-squared**: 0.0002005, indicating the model explains only 0.02% of the variance in cancer incidence rates.
- **Adjusted R-squared**: -0.000451, suggesting the model performs poorly even compared to a baseline model.
- **F-statistic**: The overall model significance is low (F-statistic = 0.3078, p-value = 0.7351), indicating it fails to capture the factors affecting cancer rates effectively.

Based on this regression analysis, neither median income nor poverty estimates have a statistically significant impact on cancer incidence rates according to the data used in this model. Moreover, the overall model does not effectively explain the variation in cancer rates, suggesting that other factors not included in the model might be more relevant or that the model needs refinement for more accurate predictions.

3.2.2 Assessment of Model Fit and Residual Analysis

The model's goodness-of-fit and the behavior of the residuals are crucial for validating the assumptions of linear regression. This analysis includes the examination of the R^2 and adjusted R^2 values, as well as the residuals to check for homoscedasticity and normality.

Goodness-of-Fit The R^2 value of the model is 0.0002005, and the adjusted R^2 is -0.000451, indicating that the model explains an extremely small portion of the variance in cancer incidence rates. This suggests that median income and poverty estimates may not be strong predictors of cancer incidence rates in this dataset.

Residual Analysis Residual plots are used to assess the assumptions of homoscedasticity and normality which are central to linear regression analysis.

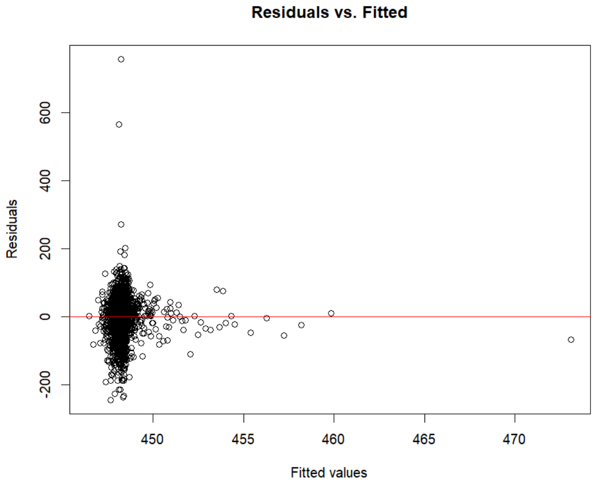

Figure 7: Residuals vs. Fitted Plot: This plot should ideally show no discernible pattern or systematic structure. The presence of a funnel shape or a pattern would suggest heteroscedasticity. Here, the residuals display a random pattern around the horizontal line at zero, which is a good indication of homoscedasticity. However, some outliers are evident, indicating potential anomalies in the data.

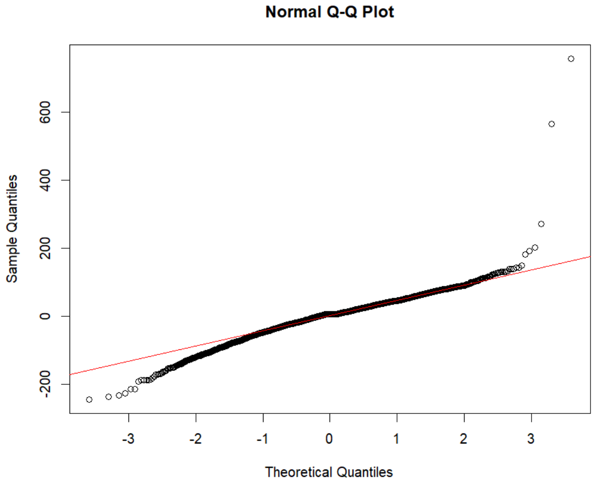

Figure 8: Normal Q-Q Plot: This plot checks the normality of residuals by comparing them with a theoretical normal distribution. The closer the points lie on the diagonal line, the more normal the distribution. Here, the points largely follow the line but deviate at the tails, suggesting some departure from normality, especially in the extremes.

These residual plots suggest that while there may be some issues with outliers and slight non-normality in the residuals, the basic assumptions of homoscedasticity are generally met. The model could benefit from further refinement or consideration of additional variables that might better explain the variability in cancer incidence rates.

4 Interpretations of Visual and Statistical Outputs

4.1 Figure 1: Top 5 Counties by Cancer Death Rate

Data-centric interpretation:
This bar chart highlights the cancer incidence and death rates in the top five counties with the highest death rates. There are notable disparities in these rates, indicating significant variation in cancer outcomes even within high-mortality regions.

Business/problem-centric interpretation:
Understanding these disparities is crucial for public health officials and policymakers. By identifying counties with disproportionately high cancer rates, targeted interventions can be planned, such as increased healthcare funding, cancer screening programs, and public health campaigns tailored to the specific needs of these communities.

4.2 Figure 2: State vs. Cancer Rates

Data-centric interpretation:
The map showcases the geographic distribution of cancer rates across states, with varying intensities indicated by color coding. It highlights regions with the highest burden of cancer, pinpointing where health disparities might be most severe.

Business/problem-centric interpretation:
This visualization assists in directing national health resources and interventions to states with the highest cancer rates. Policymakers can use this data to create localized health strategies that address the specific cancer-related challenges these states face.

4.3 Figure 3: Statewise Cancer Trends and Population Estimates

Data-centric interpretation:
The scatter plot correlates cancer incidence rates with population estimates by state, color-coded to show differences across states. The plot demonstrates that larger population sizes do not necessarily correlate with higher cancer rates, suggesting that other factors significantly influence cancer trends.

Business/problem-centric interpretation:
This information is pivotal for health departments to consider population density and distribution when planning cancer prevention and treatment services. States with high populations and lower cancer rates may have effective health policies that could serve as models for other states.

4.4 Figure 4: Scatter Plot of Cancer Incidence Rates vs. Poverty Estimates

Data-centric interpretation:
This plot illustrates a slight upward trend in cancer incidence as poverty levels increase, though the correlation is weak. The wide distribution of incidence rates at varying poverty levels indicates complex interactions between poverty and health outcomes.

Business/problem-centric interpretation:
The weak but present correlation suggests that poverty alleviation programs could indirectly reduce cancer incidence rates. Initiatives aimed at economic improvement in impoverished areas might also contribute to better health outcomes, reducing the cancer burden.

4.5 Figure 5: Scatter Plot of Cancer Incidence Rates vs. Median Income

Data-centric interpretation:
The relationship between median income and cancer incidence rates is shown to be relatively flat across the income spectrum, suggesting that income alone is not a strong predictor of cancer incidence rates.

Business/problem-centric interpretation:
This finding highlights the complexity of factors influencing cancer rates beyond simple economic metrics. Health policy development should consider a broader array of socioeconomic factors, possibly including environmental and lifestyle factors, in addition to income.

4.6 Table 1: Descriptive Statistics for Median Income, Poverty Estimates, and Cancer Incidence Rates

Data-centric interpretation:
The table provides a comprehensive overview of the central tendencies and variability of median income, poverty estimates, and cancer incidence rates, offering a statistical foundation for further analysis.

Business/problem-centric interpretation:
These statistics are crucial for identifying norms and deviations in socioeconomic and health data across counties. This insight helps in assessing the adequacy of current health policies and economic conditions, guiding future policy adjustments.

5 Discussion

5.1 Statistical Significance of Predictors

The regression analysis focused on evaluating the influence of median income and poverty estimates on cancer incidence rates. Hypothesis testing for the regression coefficients, which involves examining p-values and confidence intervals, provides insights into the statistical significance of these predictors.

- **Median Income:** The coefficient for median income is -2.062×10^{-5} with a standard error of 8.196×10^{-5} and a p-value of 0.801. This high p-value indicates that median income is not a statistically significant predictor of cancer incidence rates in this model.
- **Poverty Estimates:** Similarly, the coefficient for poverty estimates is 1.353×10^{-5} with a standard error of 1.763×10^{-5} and a p-value of 0.443, also indicating lack of statistical significance.

5.2 Assessment of Regression Assumptions

The validity of linear regression results is contingent upon several assumptions. Here, we discuss potential violations and their impacts:

5.2.1 Linearity

The assumption of linearity implies that the relationship between the independent variables and the dependent variable is linear. The lack of clear patterns in the residuals vs. fitted values plot generally supports the linearity assumption, although the presence of outliers suggests potential anomalies.

5.2.2 Independence and Homoscedasticity

The residuals should not only exhibit linearity but also independence and constant variance (homoscedasticity). The residuals plot displayed a random dispersion of residuals around the horizontal zero line, suggesting adequate homoscedasticity, but some outliers were noted, indicating possible exceptions.

5.2.3 Normality of Residuals

The normal Q-Q plot is utilized to assess the normality of residuals. If the residuals follow the theoretical line closely, it supports the assumption of normality. Deviations from this line, particularly in the tails, indicate potential issues with normality.

5.2.4 Absence of Multicollinearity

While not directly assessed here, multicollinearity involves high correlations among independent variables, which can inflate the variance of the estimated regression coefficients, leading to misleading results. Techniques such as variance inflation factor (VIF) can be employed to check for multicollinearity.

Impact on Model's Validity Violations of these assumptions, particularly if severe, could undermine the validity of the model's conclusions. The detected non-significance of predictors and the minimal R^2 value suggest that other unaccounted variables or non-linear relationships might be at play, or that the model may be oversimplified. Further investigations could include expanding the model to include more variables or using different modeling techniques such as logistic regression if the response variable is categorical.

6 Conclusion

This study sought to investigate the socio-economic factors affecting cancer incidence rates across various counties, utilizing extensive data analysis including summary measures and regression modeling. Our findings from the summary measures indicated a wide variation in median income and poverty levels across different states, which suggested potential disparities in health outcomes.

The regression analysis was particularly aimed at quantifying the impact of median income and poverty estimates on cancer incidence rates. However, the results indicated that these socio-economic factors did not significantly predict the cancer rates, as evidenced by the high p-values and negligible R-squared values. This lack of significant predictors suggests that other variables, possibly including lifestyle, environmental factors, or access to healthcare, might play more critical roles, which were not captured in the current model.

Further, the assessment of the regression model's assumptions revealed adequate adherence to linearity and homoscedasticity but pointed out potential issues with normality and outliers. These observations underline the need for a more robust model or the inclusion of additional variables that could better explain the variability in cancer incidence rates.

In conclusion, while the initial hypothesis regarding the impact of socio-economic factors on cancer incidence was not substantiated by significant statistical evidence, this study highlights the complexity of cancer epidemiology and the need for multifaceted research approaches. Future studies are recommended to incorporate broader data sets and potentially employ more complex statistical models that can account for a wider range of influential factors. Such efforts will enhance our understanding of cancer dynamics and aid in the development of targeted public health strategies and interventions.