Final Report 506

Jiayi Guo
Department of Applied Statistics
University of Michigan
Ann Arbor, United States
jiayiguo@umich.edu
Github link: https://github.com/waibibabodoge/STATS506

***Abstract*—Research Question**

This research project seeks to investigate the potential relationship between geographic and economic characteristics of U.S. states and the volume of Medicare services provided to individuals, as measured by the number of Medicare services (Tot_Srvcs) [1]. Specifically, the study aims to determine whether higher income levels, as indicated by reported salary and wage amounts (A00200) in state income tax statistics [2], are predictive of an increased number of Medicare services provided within that state.

I. Introduction

A. Datasets exploration

The CMS dataset utilized in this study provides detailed information on the amount of Medicare services provided, the types of services, payment amounts, etc., broken down by various factors including geographic region and service category. For this analysis, the data is examined at the state level to explore regional variations in service provision (Tot_Srvcs).

In addition to the CMS dataset, income data from the IRS Individual Income Tax Statistics (for 2020) provides information on salary and wage amounts (A00200) by state. This data enables an exploration of the relationship between income levels and the amount of Medicare services provided across different geographic regions. By analyzing Medicare utilization and payment data alongside state-level income tax statistics, this study will explore whether higher income populations within specific states correlate with a higher volume of Medicare services, and whether geographic and economic factors play a significant role in determining healthcare access and service provision.

B. Hypothesis

States with higher average wages and salaries (A00200) might have higher numbers of healthcare services provided (Tot_Srvcs) due to various factors, such as increased access to healthcare, better overall healthcare infrastructure, and higher state-level spending on healthcare services.

II. Approach

A. Setup

The data cleaning process involved merging healthcare services data (medical) and income tax data (tax) at the state level. The medical dataset was then cleaned to exclude national-level data and assign state abbreviations to a new column, State_Abbr. The tax dataset was processed to calculate the average income (A00200) per state. Finally, both datasets were merged on the state abbreviation, resulting in a cleaned merged_data dataset containing total healthcare services (Tot_Srvcs) and average income (A00200) for each state, ready for further analysis.

B. Regression Analysis

After preform the regression analysis, gets the following results:

The t-value for avg_income is 4.818, which is large enough to suggest that the relationship between average income and average services is statistically significant.

The p-value for avg_income is 1.5e-05, which is much smaller than the conventional significance level of 0.05, confirming that the effect of income on healthcare services is statistically significant.

The equation of the linear regression model is:

$$\text{avg\_services} = 1871 + 0.08694 \times \text{avg\_income}$$

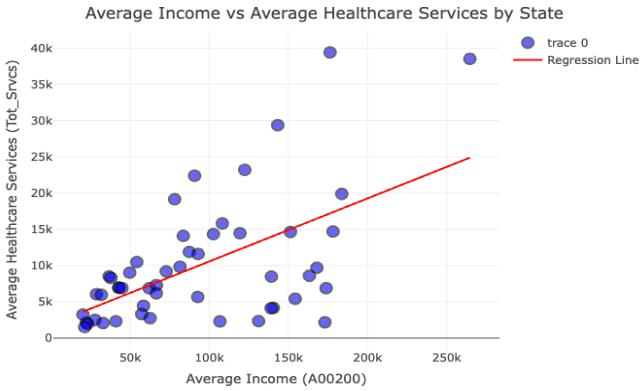

Fig. 1. Average Income vs Average Healthcare Services by State

There is a positive and statistically significant relationship between average income (avg_income) and the average number of healthcare services provided (avg_services) at the state level. However, the R-squared value around 32.6% (around 30.58% after performing log transformation) indicates that there are other factors affecting healthcare services that are not captured by average income alone. This means the model has explains parts of the model but is not a perfect fit.

The scatters show the a positive correlation between average income and average healthcare services, suggesting

that states with higher average incomes tend to have more healthcare services on average. States such as California (CA), Florida (FL), Texas (TX), New York (NY), Illinois (IL), and Pennsylvania (PA) are positioned above the regression line, indicating that these states have a higher number of healthcare services than what would be expected based on their average income. This could be due to various factors such as larger populations, more extensive healthcare infrastructure, or specific state policies that promote healthcare services. On the other hand, states like Hawaii (HI), Nevada (NV), Rhode Island (RI), Utah (UT), and Alaska (AK) are positioned below the regression line, suggesting that these states have fewer healthcare services than expected for their average income levels. This might be attributed to factors such as geographic challenges, lower population density, or different healthcare funding priorities. Which the result is reasonable considering the real-world situation of the United States.

C. Plot Analysis

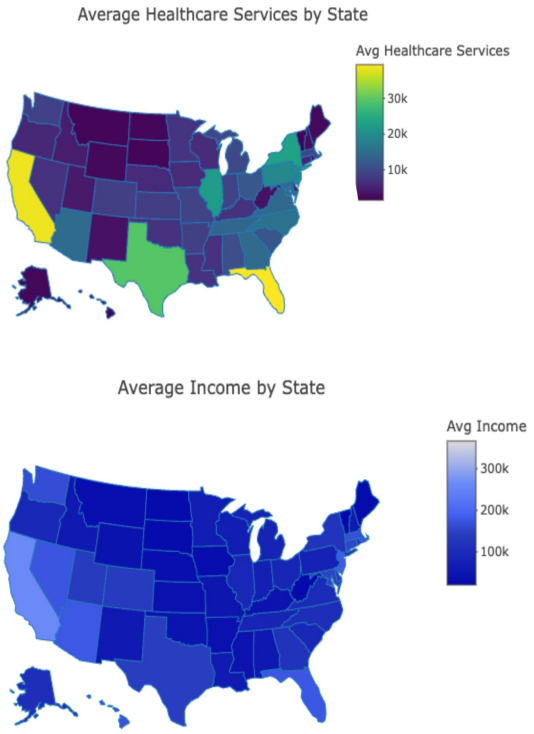

Fig. 2. Average Healthcare Services and Average Income by State

As in the scatter plot, this plot reveals more intuitive result of the relation between average healthcare services and income by state. The geographic maps further underscore the trend with states having high average incomes like those in the East, Southwest, and Southeast around the coastline, also showing higher average healthcare services. The Central of continent and Northwest areas showing lower average healthcare services. This plot confirms that while income is a significant predictor of healthcare service volume, other factors like geographic conditions play crucial roles, which is for the further study.

III. Conclusion

The analysis finds a statistically significant positive correlation between average income (A00200) and the average number of healthcare services provided (Tot_Srvcs). States with higher average incomes tend to have a greater number of healthcare services. However, the model's R-squared value of around 32.6% indicates that other factors also significantly influence the volume of Medicare services which requires more study.

Geographic factors, as seen in the plots, further emphasize the importance of regional disparities. States like California, Florida, and New York, despite their high income levels, show a higher-than-expected volume of services, suggesting that factors beyond income, such as healthcare infrastructure, population size, geographical position, and state policies, also play a key role in service provision. In contrast, states such as Hawaii and Utah, despite their relatively higher average incomes, fall below the expected service volume, hinting at the influence of geographic and demographic factors. It also highlights the complexity of healthcare access and suggests that a broader range of factors needs to be considered when assessing Medicare service utilization.

In this study, the method involved cleaning and merging two datasets: the Medicare Provider Utilization and Payment Data and IRS Income Tax Statistics. The data was processed using the data.table package in R, where state-level average income (A00200) and Medicare services (Tot_Srvcs) were computed since the dataset is too large to process. A linear regression model (also transformation) was used to examine the relationship between income and healthcare service volume, revealing a positive and statistically significant correlation. The plots, including scatter plots and geographic maps created with plotly, were used to illustrate regional variations and the relationship between income and services. The analysis highlighted the role of income in predicting Medicare service provision but also indicated that other geographic and policy factors contribute significantly to service availability for more research.

References

[1] Centers for Medicare & Medicaid Services, "Medicare Physician and Other Practitioners Provider Summary by Type of Service," 2024. [Online]. Available: https://data.cms.gov/provider-summary-by-type-of-service/medicare-physician-other-practitioners.

[2] Internal Revenue Service, "Individual Income Tax Statistics: 2020 Zip Code Data (SOI)," 2024. [Online]. Available: https://www.irs.gov/statistics/soi-tax-stats-individual-income-tax-statistics-2020-zip-code-data-soi.