Exploring the Relationship Between AGI Levels and Healthcare Service Settings in Michigan ZIP Codes

Yufan Li
LSA Department of Statistics
University of Michigan
Ann Arbor, United States
Email: liyufan@umich.edu

***Abstract*—This study explores the relationship between Adjusted Gross Income (AGI) levels and the likelihood of healthcare services being provided in facility versus non-facility settings across Michigan ZIP codes. Using data from the IRS and Medicare datasets, multiple predictive models, including logistic regression, random forests, and exploratory data analysis (EDA), are applied to assess the role of AGI and other factors in determining service settings. The findings show that total services and total beneficiaries are stronger predictors of facility-based services, with AGI playing a minor role.**

I. Introduction

The aim of this project is to explore the relationship between Adjusted Gross Income (AGI) levels and the likelihood of healthcare services being provided in facility-based versus non-facility settings in Michigan. Understanding these relationships is crucial for healthcare policy and resource allocation. The primary research question guiding this study is:

> *"Are ZIP codes with higher AGI levels more likely to have services provided in facility settings compared to non-facility settings in the state of Michigan?"*

This question seeks to explore the impact of income levels on the distribution of healthcare services, particularly in terms of facility vs. non-facility settings across Michigan ZIP codes.

II. Data Sources and Methodology

Two primary datasets were used:

- **IRS Data (2020)**: Contains AGI data for ZIP codes across Michigan [2].
- **Medicare Data (2020)**: Provides information on healthcare services by service setting (facility or non-facility) for Michigan ZIP codes [1].

The data was cleaned by filtering for Michigan ZIP codes and merging based on ZIP codes. The response variable was binary, with 1 indicating facility-based services and 0 indicating non-facility services. Two major modeling techniques were used to explore this relationship, including **Logistic Regression** and **Random Forest**[3].

III. Exploratory Data Analysis (EDA)

Exploratory data analysis was performed to understand the distribution of AGI and healthcare service settings. The **box plots** and **bar charts** (placed in the Appendix) revealed that higher AGI values tended to be associated with non-facility settings, though there was significant overlap between categories. The following **Leaflet maps** provide a geographic visualization of healthcare service distribution in Michigan. The first map displays the weighted AGI stub median across the state, highlighting areas with varying levels of AGI. The second map shows the geographic distribution of facility-based (blue markers) and non-facility (red markers) healthcare services, with urban centers like **Detroit** and **Ann Arbor** showing a higher prevalence of facility-based services.

A. AGI Distribution Map

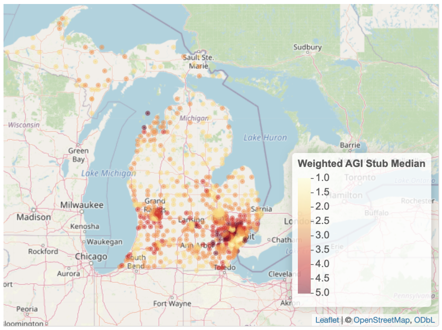

Fig. 1. Geographic Distribution of Weighted AGI Stub Median across Michigan

B. Service Distribution Map

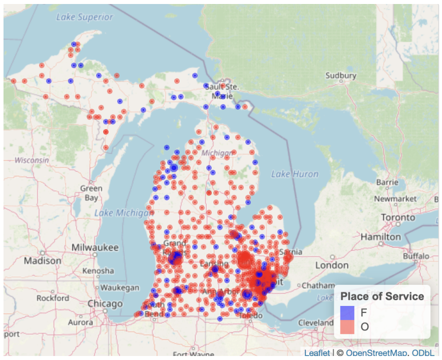

Fig. 2. Geographic Distribution of Facility vs Non-Facility Services in Michigan

IV. Models and Approaches

A. Logistic Regression (Original Model)

A logistic regression model was fit with **AGI** as the only predictor of facility-based services. The model estimated the coefficient for AGI to assess its impact on the likelihood of facility-based services. The analysis was conducted using a binomial family with a logit link, and the model was evaluated using deviance and Akaike Information Criterion (AIC) for model fit.

B. Logistic Regression (Controlled Model)

In the controlled model, additional variables, namely **Total services** and **Total beneficiaries**, were included to assess their impact alongside AGI on the likelihood of facility-based services. The model was fit using a binomial family with a logit link. The logistic regression equation for this model is:

$$\begin{aligned}\text{logit}(P(\text{Facility})) = &\beta_0 + \beta_1 \times \text{AGI}\\ &+ \beta_2 \times \text{Total Services}\\ &+ \beta_3 \times \text{Total Beneficiaries}\end{aligned}$$

The model was also evaluated using deviance and Akaike Information Criterion (AIC) for model fit.

C. Random Forest Model

A **Random Forest** model was applied to capture non-linear relationships between the predictors and the response variable. The model was trained using 500 trees, with the number of variables tried at each split set to 1. The model estimated the importance of each predictor using metrics such as IncMSE and IncNodePurity. The model performance was assessed using the mean squared residuals and the percentage of variance explained.

$$\text{Random Forest Output: Importance of Variables} \tag{1}$$

V. Results

A. Logistic Regression Results

The first logistic regression model, with only agi_stub_weighted_median as the predictor, showed a marginally significant negative relationship between AGI and the likelihood of facility-based services. The coefficient for AGI was -0.2212 with a p-value of 0.06599, suggesting that as AGI increases, the likelihood of facility-based services decreases. However, since the p-value exceeds 0.05, we fail to reject the null hypothesis and conclude that AGI likely does not have a significant effect on the place of services in Michigan. The odds ratio for AGI was 0.801, indicating that for each unit increase in AGI, the odds of facility-based services decrease by approximately 19.9%.

In the second model, which included Total_services and Total_beneficiaries, the coefficient for AGI remained negative (-0.3200), and was statistically significant with a p-value of 0.011494. Total_beneficiaries had a significant positive relationship with facility-based services (1.378×10^{-5}, p-value 0.000888), while Total_services had a marginally negative effect. The model's AIC improved to 571.64, suggesting a better fit than the first model.

B. Random Forest Results

The random forest model, which included agi_stub_weighted_median, Total_services, and Total_beneficiaries, explained 23.01% of the variance in service settings. The variable importance scores revealed that Total_beneficiaries and Total_services were the most important predictors of facility-based services, with scores of 42.27 and 32.58 respectively, while agi_stub_weighted_median had a much smaller importance score of 9.48. The model provided more accurate predictions than the logistic regression models, further confirming that healthcare infrastructure, such as the number of beneficiaries and service volume, plays a stronger role in determining service type than AGI levels.

VI. Conclusion

The analysis suggests that Adjusted Gross Income (AGI) levels have a **minor role** in determining whether services are provided in facility settings. The results indicate that Total services and Total beneficiaries are much stronger predictors of facility-based services. This finding directly answers the research question: "ZIP codes with higher AGI levels are **NOT** more likely to have services provided in facility settings compared to non-facility settings in Michigan". The analysis challenges the initial hypothesis that wealthier areas are more likely to have facility-based services. Instead, it suggests that healthcare infrastructure, including the number of beneficiaries and service availability, plays a more significant role in determining whether services are provided in facility-based settings.

A. Policy Implications

The findings highlight the need for policymakers to focus on improving access to services in both **facility** and **non-facility settings**, especially in regions with high numbers of beneficiaries but fewer resources. The model also provides insights into areas where additional facilities or services may be required.

VII. Future Work

Future research could explore:

- Interactions between AGI and other factors, such as urban vs. rural locations.
- Advanced machine learning models like **XGBoost** or **Neural Networks** to capture more complex patterns.
- Additional variables like insurance coverage or proximity to healthcare providers to better understand the factors influencing service type.

VIII. References

References

[1] Centers for Medicare and Medicaid Services, ”Physician and Other Practitioner Services Data,” 2020. Available at: https://data.cms.gov/provider-summary-by-type-of-service/medicare-physician-other-practitioners/medicare-physician-other-practitioners-by-provider-and-service/data/2020.

[2] Internal Revenue Service (IRS), ”Individual Income Tax Statistics: 2020 ZIP Code Data,” Available at: https://www.irs.gov/statistics/soi-tax-stats-individual-income-tax-statistics-2020-zip-code-data-soi.

[3] L. Breiman, ”Random Forests,” *Machine Learning*, vol. 45, no. 1, pp. 5-32, 2001.

Appendix

This appendix contains the boxplots, bar charts, and histograms used for exploratory data analysis. The following figures provide visual representations of the relationships between Adjusted Gross Income (AGI) and healthcare service settings.

A. Boxplots

The boxplot below shows the distribution of **AGI** across different service settings (facility vs. non-facility). It highlights how AGI varies between the two categories of service types.

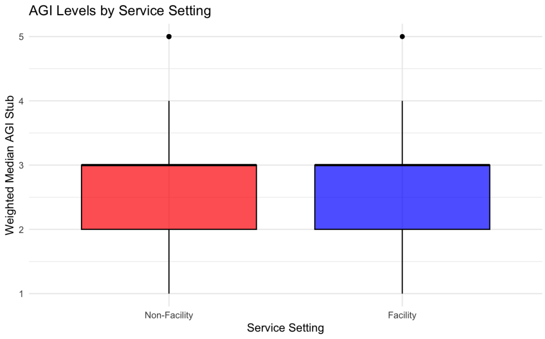

Fig. 3. Boxplot of AGI by Service Type (Facility vs. Non-Facility)

B. Bar Charts

The bar chart below illustrates the count of healthcare services provided in facility vs. non-facility settings across different ranges of AGI. It provides a comparative view of how the service distribution varies by AGI levels.

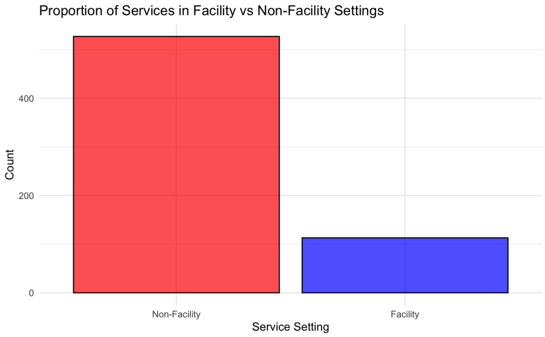

Fig. 4. Bar Chart of Service Type by AGI Range

C. Histogram of Weighted Median AGI Levels

The histogram below shows the distribution of the weighted median AGI levels across the ZIP codes. It provides insight into the spread and concentration of AGI values across Michigan.

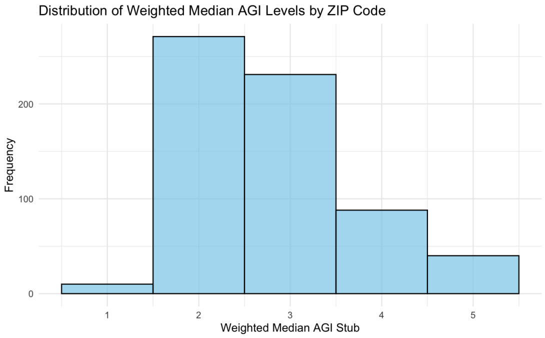

Fig. 5. Histogram of Weighted Median AGI Levels

D. GitHub Repository

The complete code used for the analysis, including the generation of the boxplots, bar charts, and histograms, as well as the HTML output, is available in the GitHub repository. Access the repository at the following link:

https://github.com/YufanLi2002/STATS506/tree/main/Final%20Project

E. Attribution of Source

This section provides attribution for any external sources, datasets, or tools used in this report. All data used in the analysis was sourced from publicly available datasets, including the IRS data and Medicare data. The link can be found in the Reference section. Additionally, the statistical methods and machine learning algorithms used in this report were implemented with the help of various open-source libraries, including:

- The `glm` function from the *stats* package in R for logistic regression.
- The `randomForest` package in R for the random forest model.
- Plotting libraries such as `ggplot2` for visualizations.

Minor tasks such as debugging code, reviewing formatting of the report were assisted by ChatGPT and Claude, which were used to enhance the workflow and ensure the quality of the document.

Full attribution for these tools and datasets can be found in their respective documentation or repositories.