Bayesian Analysis on Standardized Medicare Payments and State Tax Payments Revenues

Xiayang Sun (Uniqname: xysun)

December 14, 2024

1 Introduction

Abnormal heart rhythms, such as atrial fibrillation, is a significant public health concern in the United States because of the increasing amount of aging population. Some procedures like catheter ablation are implemented to ensure the effectiveness of treatments for these situations.[6] However, the associated costs, including relative anesthesia services, has been served as a challenge not only for patients themselves but also for the broader healthcare system. For instance, Medicare, the primary source of health insurance for older adults in the United States, contribute to a large portion of high-cost medical procedures.[1] Among these, procedures to correct abnormal heart rhythms, are particularly vital considering their high prevalence and risk. Moreover, factors such as Anesthesia services which is critical to the success of such procedures, contribute a significant portion of to the costs, making Medicare's reimbursement policies a key determinant of access and affordability.

While Medicare payments are standardized to remove geographic differences caused by local wages or input costs, state-level factors such as tax revenues remain an important consideration. States with higher tax revenues contribute more to the federal budget, potentially influencing Medicare's capacity to fund high-cost services. It is worthwhile to explore the influence of state tax revenues on Medicare's standardized reimbursements for specific services, such as anesthesia in heart rhythm correction procedures as a result.

A general idea is to assume a linear connection between the response and predictor. However, if we apply traditional frequentist linear regression models where parameters are fixed, large residuals and weak correlation could be witnessed, leading to feeble interpretability of the relationship.

This study would apply Bayesian analysis to investigate this relationship. The linear assumption would be made between the response and predictor. Posterior distributions of linear coefficients would be derived to observe the plausibility of the linear model with help of prior information and regional data. The results will provide valuable insights into the financial dynamics of Medicare funding and contribute to policy discussions on equitable healthcare resource allocation.

2 Data

Two datasets are obtained from websites. The first dataset includes Medicare provider utilization and payment data, which tracks information on services that physicians perform which are sent to Medicare for insurance coverage in 2020. All kinds of Medicare payments are listed by state in *Rndrng Prvdr Geo Desc* and category in *HCPCS Desc*. More precisely, we would take deep insight into *Average Medicare Standardized Payment Amount* (where geographic differences in payment rates for individual services have been removed) of "Anesthesia for procedure to correct abnormal heart rhythm" in *HCPCS Desc* of all states [3].

Another dataset is The Statistics of Income (SOI) Division's ZIP code data. This dataset includes various factors relative to income and taxes integrated by state. We would lay major emphasis on *Total taxes paid amount*[5].

As each state's Medicare payment amount is calculated summarily rather than door-to-door, we would apply the average tax paid amount of each state rather than data of each zip-code in the state. In the next section, a linear assumption would be established to unfold the relationship between these two variables.

3 Model

Assumptions

Suppose $x_1, x_2, ..., x_n$ be all states' inner average tax revenues data, $Y_1, Y_2, ..., Y_n$ denote the r.v. representing the Medicare payment amount. And $\boldsymbol{\beta} = (b, a)'$ is the linear regression coefficents between each x_i and Y_i. Let $\boldsymbol{X} = (1, x_1; 1, x_2; ...; 1, x_n)$, $\boldsymbol{Y} = (y_1, y_2, ..., y_n)'$, and $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, ..., \epsilon_n)' = \boldsymbol{Y} - \boldsymbol{X\beta}$ denotes the residuals.

- $(x_i, Y_i)_{i=1}^n$ should follow basic assumptions for *Linear Regression*[2],

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}, \tag{1}$$

$$r\left(\boldsymbol{X}\right) = 2, \tag{2}$$

$$\epsilon_1, \epsilon_2, ..., \epsilon_n \sim N\left(0, \sigma^2\right), \text{ i.i.d.} \tag{3}$$

- Suppose prior [4]

$$\boldsymbol{\beta} \sim N_2\left(\boldsymbol{\beta_0}, \boldsymbol{\Sigma_0}\right), \tag{4}$$

$$\gamma = \frac{1}{\sigma^2} \sim \text{Gamma}\left(\frac{\nu_0}{2}, \frac{\nu_0\sigma_0^2}{2}\right). \tag{5}$$

  where $\boldsymbol{\beta}$ and γ are independent.

Bayesian Model

Based on all above assumptions,

$$\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\beta}, \sigma \sim N_n\left(\boldsymbol{X\beta}, \sigma^2\boldsymbol{I}\right), \tag{6}$$

Therefore posterior for $\boldsymbol{\beta}|\boldsymbol{X}, \boldsymbol{\beta}, \sigma$ is

$$N_2\left(\mathbb{E}[\boldsymbol{\beta}, \sigma^2], \mathbb{D}[\boldsymbol{\beta}, \sigma^2]\right), \tag{7}$$

where

$$\mathbb{E}[\boldsymbol{\beta}, \sigma^2] = \left(\boldsymbol{\Sigma_0}^{-1} + \boldsymbol{X}'\boldsymbol{X}/\sigma^2\right)^{-1} \left(\boldsymbol{\Sigma_0^{-1}\beta_0} + \boldsymbol{X}'\boldsymbol{Y}/\sigma^2\right), \tag{8}$$

$$\mathbb{D}[\boldsymbol{\beta}, \sigma^2] = \left(\boldsymbol{\Sigma_0}^{-1} + \boldsymbol{X}'\boldsymbol{X}/\sigma^2\right)^{-1}. \tag{9}$$

While

$$p(\gamma|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\beta}) \propto p(\gamma)p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\beta}, \gamma), \tag{10}$$

Therefore posterior for $\gamma|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\beta}$ is

$$\text{gamma}\left(\frac{\nu_0 + n}{2}, \frac{\nu_0\sigma_0^2 + \text{SSR}(\boldsymbol{\beta})}{2}\right), \tag{11}$$

where $\text{SSR}(\boldsymbol{\beta}) = \|\boldsymbol{Y} - \boldsymbol{X\beta}\|_2^2$[4].

Sampling Algorithm

It would be hard to obtain joint distribution of $\boldsymbol{\beta}, \sigma^2|\boldsymbol{X}, \boldsymbol{Y}$, so we apply *Gibbs sampler* to sample $\boldsymbol{\beta}, \sigma^2$. Input $\boldsymbol{Y}, \boldsymbol{X}$, hyperparameters ν_0, σ_0^2, initial values $\boldsymbol{\beta}^{(0)}$, $\sigma^{2(0)}$

Repeat

1. Update $\boldsymbol{\beta}^{(s+1)} \sim N_2\left(\mathbb{E}[\boldsymbol{\beta}, \sigma^{2(s)}], \mathbb{D}[\boldsymbol{\beta}, \sigma^{2(s)}]\right)$;
2. Update $\sigma^{2(s+1)} \sim \text{inverse-gamma}\left(\frac{\nu_0 + n}{2}, \frac{\nu_0\sigma_0^2 + \text{SSR}(\boldsymbol{\beta^{(s+1)}})}{2}\right)$.

4 Numerical Analysis

In this experiment, parameters in priors are given to be $\boldsymbol{\beta_0} = \boldsymbol{0}$, $\boldsymbol{\Sigma_0} = \text{diag}\left(2, 2\right)$, $\nu_0 = 2$, $\sigma_0^2 = 1$. We employed the Gibbs MCMC method to sample the model parameters, running a total of 10,000 iterations. To enhance the reliability of the results and avoid potential bias from the initial unstable phase (burn-in period), we conducted convergence checks and excluded the first 500 iterations from the analysis. This ensures that the parameter estimates for $\boldsymbol{\beta}, \sigma^2$ are based on the final 9,500 samples, which represent a more stable and reliable portion of the posterior distribution. By focusing on these iterations, we improve the robustness and scientific validity of the findings.

The histograms of $\boldsymbol{\beta} = (b, a)'$ generated from the posterior samples illustrate smooth, bell-shaped distributions, indicating proper convergence and reliable posterior estimates. For b, the distribution with a normal shape is centered around a mean of 0.63, suggesting a positive baseline level of Medicare payments even when per capita taxes are near zero. However, the relatively long tail in the positive direction reflects variability across states, likely influenced by outliers with unusually high Medicare payments. Similarly, the distribution of a has a sharp peak near its mean of 0.161, showing a strong and consistent positive relationship between per capita taxes and Medicare payments. This implies that higher state taxes are associated with higher Medicare spending.

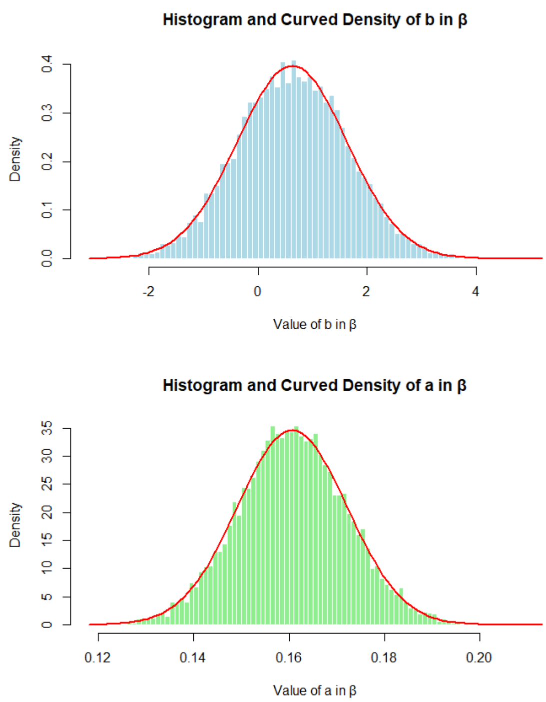

Figure 1: Histogram and curved density function of sampled b(up plot), a(down plot) in $\boldsymbol{\beta}$ by MCMC.

The residual variance σ^2 has a median of 903.0, with a wide range from 123.4 to 2390.3, and a distribution of Inverse-Gamma$(23.32, 20731.02)$, indicating moderate variability in Medicare payments that is not explained by per capita taxes alone. This unexplained variance suggests the influence of additional state-level factors such as income payment amount, return deductions under various loan or insurance, and so on.

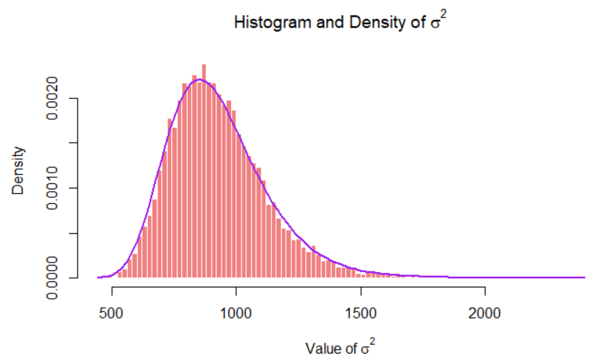

Figure 2: Histogram and curved density function of sampled σ^2 by MCMC.

The smooth histograms of Figure 1 and 2 indicate that the Gibbs sampling process converged effectively, providing reliable posterior estimates of the parameters.

5 Conclusion

The analysis in section 4 reveals a statistically significant positive relationship between per capita taxes and Medicare payments, implying that states with higher tax revenues are likely to allocate more resources to Medicare spending. However, the variability in the residual σ^2 highlights the importance of unmeasured factors that influence healthcare costs at the state level. Policymakers should consider these findings when designing tax and healthcare funding policies, as higher state taxes may provide better support for Medicare-funded procedures. Future research could enhance this model by incorporating additional predictors, such as state population health metrics, urbanization levels, and healthcare infrastructure quality, to better understand the drivers of Medicare payment variability. Identifying and analyzing outlier states with extreme values of b and σ^2 could provide further insights into unique challenges or best practices in managing.

References

[1] Envisioning a better u.s. health care system for all: Coverage and cost of care. *Annals of Internal Medicine*, 172(2_Supplement):S7–S32, 2020. PMID: 31958805.

[2] Julian J. Faraway. *Linear Models with R*. Chapman and Hall/CRC, 2nd edition, 2014.

[3] Centers for Medicare Medicaid Services (CMS). Medicare physician other practitioners - by geography and service, 2022. Accessed: 2024-12-07.

[4] Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 3rd edition, 2013.

[5] Internal Revenue Service (IRS). Soi tax stats - individual income tax statistics — 2020 zip code data (soi), 2022. Accessed: 2024-12-07.

[6] George Katritsis and Hugh Calkins. Catheter ablation of atrial fibrillation - techniques and technology. *Arrhythmia Electrophysiology Review*, 1(1):29–33, Sep 2012.

Appendix 1: OLS model VS Bayesian model

With given data, we could estimate $\boldsymbol{\beta}$ by Least Squares Method. The result and relative plot is given as following:

Table 1: OLS estimation for $\boldsymbol{\beta}$

$\hat{\beta}_{OLS}$	**Estimate**	**Std. Error**	**t value**	**Pr(>|t|)**
(Intercept) b	66.550470	4.393140	15.149	$< 2 \times 10^{-16}$
(Slope) a	0.005125	0.011406	0.449	0.655

- Residual Standard Error: 12.82 on 47 degrees of freedom
- Multiple R^2: 0.004278
- Adjusted R^2: -0.01691

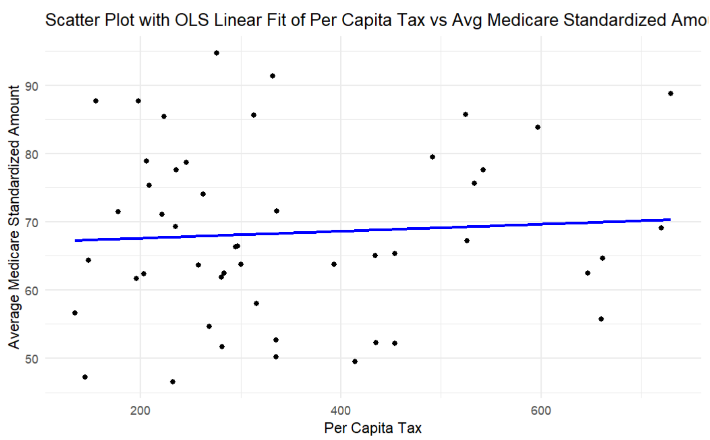

Figure 3: OLS result for the linear regression model

On the one hand, OLS does not serve as a good method to estimate $\boldsymbol{\beta}$ given the large p-value for the slope, the large MSE and the small R^2

On the other hand, there exist significant differences between the OLS and the Bayesian model. While considering the actual situation, there should be a proportional relationship between the response and the predictor, which means the Bayesian Model is more reasonable.

To investigate the reason for the difference, we should check the plausibility of the assumptions for the model. Figure 4: QQ-plot below shows that the residuals fit Assumption 3. Besides, the Breusch-Pagan test's p-value equals to 0.5767, making the heteroscedasticity hypothesis hard to be rejected.

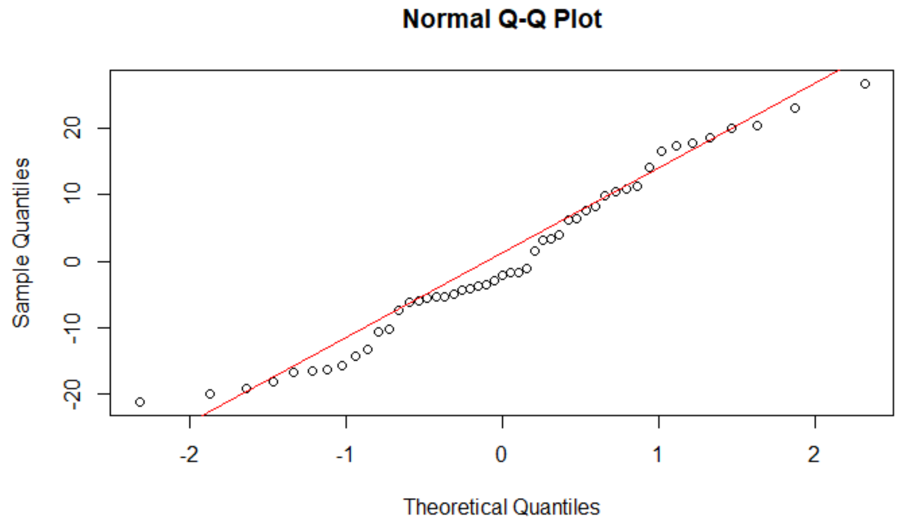

Figure 4: QQ-plot

Appendix 2: Statistics for Posterior Parameters Samples

In the numerical experiment, we performed 10,000 iterations and excluded the first 500 samples. The table below lists the statistical result for the 9,500 samples. The result is cited in the main part of the essay.

Table 2: Statistics for Posterior Parameters

Statistic	b	a	σ^2
Min.	-2.86831	0.1155	452.3
1st Qu.	-0.04061	0.1530	790.6
Median	0.62116	0.1607	902.4
Mean	0.62750	0.1608	928.9
3rd Qu.	1.29525	0.1686	1038.9
Max.	4.43436	0.2075	2335.2

Appendix 3:

Link for the project:

https://github.com/xysun-pixel/STATS-551-final