Determinants of Hospital Charges in the United States

By: Rhythanyaa Nagaraja

Introduction:

Hospital charges vary widely throughout the United States. Hospitals submit charges for inpatient services that can vary significantly. The goal of this project was to understand which hospital factors are associated with higher hospital charges. Understanding this is important for understanding healthcare spending.

Data Collection:

The data used for this analysis are publicly available Medicare inpatient hospital data from the Centers for Medicare and Medicaid Services. The dataset contains hospital-level information on inpatient services, including total submitted charges, Medicare payments, patient volume, and patient risk characteristics. This data consists of 3,093 rows and 57 columns. The 3,093 rows correspond to the samples, and the 57 columns correspond to the different features for each patient. Features can be grouped into hospital financial variables, patient volume variables, and patient risk/demographic variables. For this project, the following features were used: ‘Tot_Submtd_Cvrd_Chrg', 'Tot_Pymt_Amt', 'Tot_Mdcr_Pymt_Amt', 'Tot_Dschrgs', and ‘Bene_Avg_Risk_Scre’. The predictor variables are 'Tot_Pymt_Amt', 'Tot_Mdcr_Pymt_Amt', 'Tot_Dschrgs', and ‘Bene_Avg_Risk_Scre’. The target variable is ‘Tot_Submtd_Cvrd_Chrg', as we are trying to predict the hospital charge. There are several limitations to this dataset. First, it only lists hospitals that participate in Medicare inpatient programs. Second, the data are recorded at the hospital level and lack many patient-related details, making it hard to identify the specific patient characteristics that influence high hospital charges.

Data Preparation:

To prepare the data, duplicates and rows with missing data were dropped. Additionally, a log transformation was applied to the variables “Tot_Submtd_Cvrd_Chrg” and “Tot_Mdcr_Pymt_Amt”. A new dataframe was created and included the variables ‘Log_Tot_Submtd_Cvrd_Chrg', 'Tot_Pymt_Amt', 'Log_Tot_Mdcr_Pymt_Amt', 'Tot_Dschrgs', and ‘Bene_Avg_Risk_Scre’.

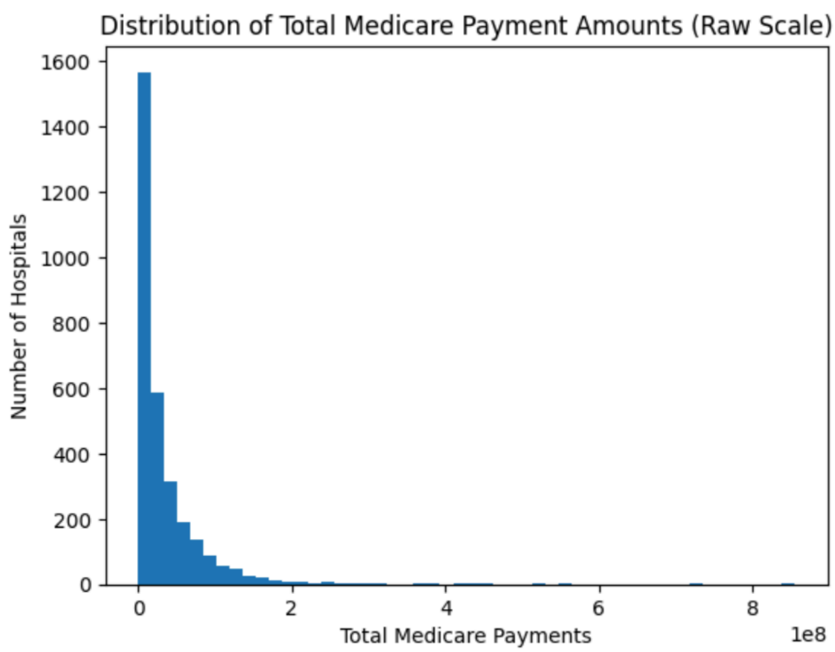

Figure 1. Distribution of Total Medicare Payment Amounts on Raw Scale

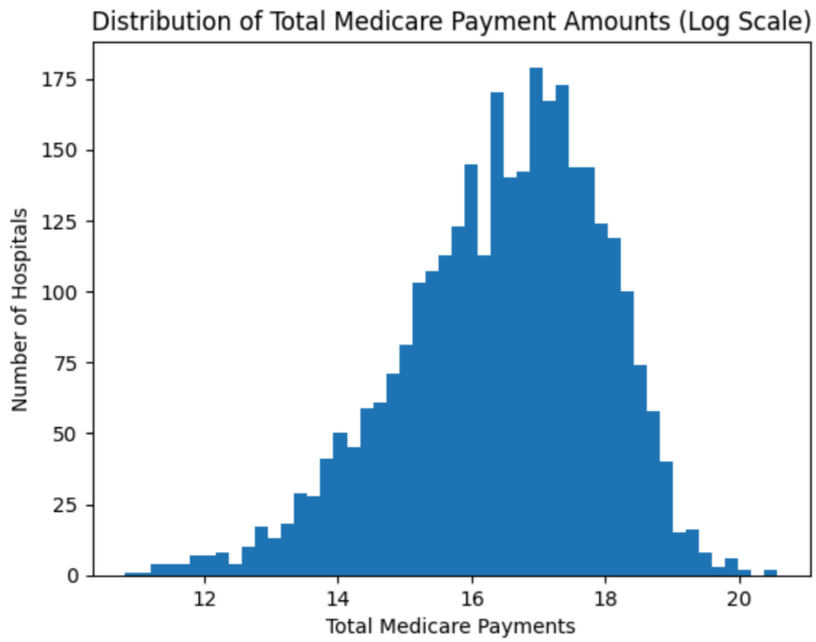

Figure 2. Distribution of Total Medicare Payment Amounts on Log Scale

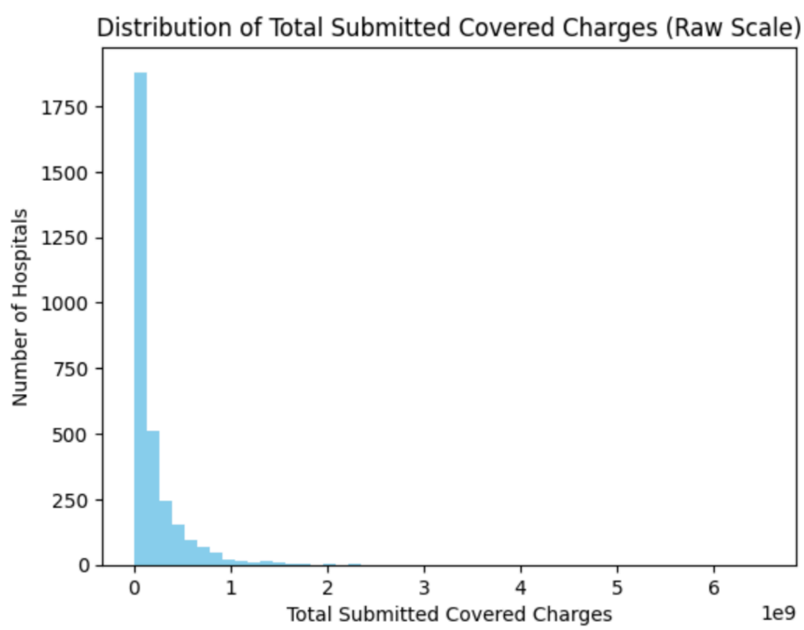

Figure 3. Distribution of Total Submitted Covered Charges on Raw Scale

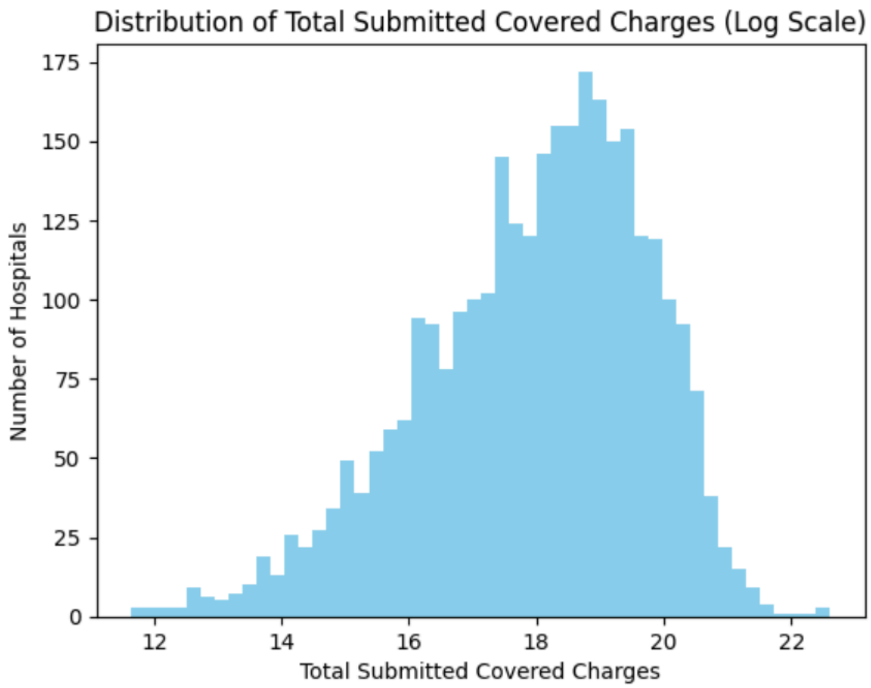

Figure 4. Distribution of Total Submitted Covered Charges on Log Scale

Applying a log transformation reduced skewness and made the distributions more balanced. This allowed patterns in the data to be more clearly visualized and analyzed. ‘Tot_Mdcr_Pymt_Amt’ and ‘Tot_Submtd_Cvrd_Chrg' were rightly skewed, meaning most of the hospitals have lower charges, while a small number of hospitals have higher charges. This skewness makes it hard to see patterns and trends. Logging the data reduces the impact of these extreme values, making the distribution more balanced and easier to visualize.

Linear Regression Results and Interpretation

Regression coefficients measure the strength and direction of the relationship between a variable and the hospital charges. A positive coefficient indicates that higher values of a variable are associated with higher hospital charges, while a negative coefficient indicates the opposite.

Variable	Coefficient
Total Discharges	0.0001
Total Covered Days	-0.00002
Avg Risk Score	0.13
Log Total Medicare Payments	1.11

Figure 5. Table of Predictor Variables and the Regression Coefficient

The coefficients in Figure 5 were estimated using a linear regression model, with the outcome variable being the log of total submitted covered charges. The coefficient on Medicare payment amounts is the largest and most positive, indicating that hospitals that receive higher payments tend to submit higher charges. The average patient risk score is positively associated with hospital charges, while total covered days and total discharges have relatively small effects. The train and test scores (R^2) were both 91 percent, meaning the model does not underfit nor overfit the data and is reliable. It explains about 91 percent of the variation in log hospital charges.

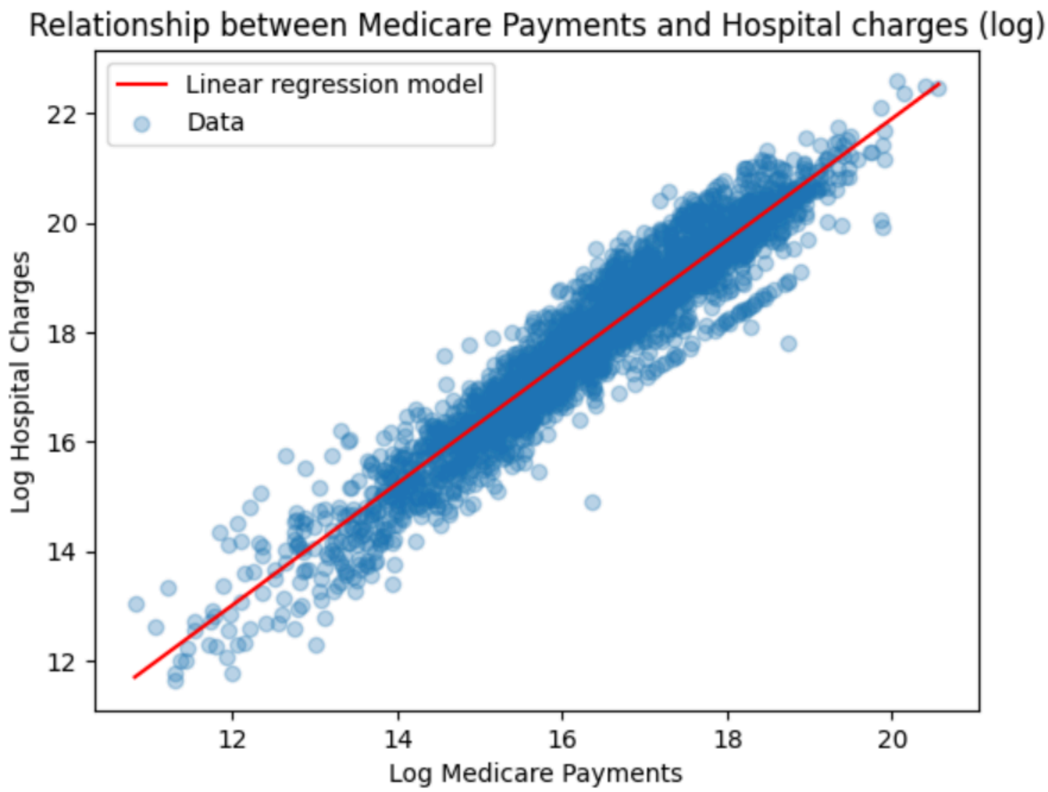

Figure 6. Relationship Between Log Medicare Payments and Log Hospital Charges With Fitted Linear Regression Line.

Figure 6 shows the relationship between log medicare payments and log hospital charges with a fitted regression line. There is a clear upward trend, indicating a strong positive association: hospitals with higher Medicare payments report higher hospital charges. The slope of the regression line reflects how percentage changes in Medicare payments relate to percentage changes in hospital charges.

Because hospital charges are log transformed, all regression coefficients represent proportional (percentage) changes in hospital charges rather than dollar amounts. For predictors that are also log transformed, such as Medicare payment amounts, a 1 percent increase in Medicare Payments is associated with a 1.11% increase in hospital charges based on the regression coefficient. However, for predictor variables that did not undergo log transformation, such as 'Bene_Avg_Risk_Scre', a one-unit increase in the average patient risk score is associated with a 12-13% increase in Medicare Payments.

Model Deployment:

This model demonstrates how hospital-level data can be used to predict hospital charges based on Medicare payments, risk characteristics, and patient volumes. Health care administrators and policymakers can use this model and regression coefficients to understand how hospital charges are influenced. Perhaps, if a hospital has higher-than-expected charges, this could indicate differences in pricing or operational inefficiencies that can be investigated further. It is observed that Medicare payments play a major role in the U.S. healthcare system, so the strong relationship between Medicare payments and hospital charges suggests that payment structures are important drivers of healthcare spending. Policymakers and researchers can use information from this analysis to better understand what contributes most to variation in healthcare spending. In real-world settings, understanding these relationships helps explain variation in hospital charging behavior.