Cost Analysis and Forecasting for Hospital Financial Performance

Group 4

Charlotte Xu, Ruiyang Dong, Xuyuan Zhang, Zihan Wang

December 11, 2024

Outline

1. Introduction
2. Data
3. Shiny
4. Model
5. ML
6. Conclusion

Introduction

Main Obejctive

- This project aims to analyze historical operating costs and revenue trends of hospitals using the *CMS Hospital Provider Cost Report dataset*.
- We use statistical models and machine learning tools to help hospital leaders make better decisions, plan budgets, and allocate resources more effectively.

All of the replication Code and Data can be found in Github https://github.com/sergiozxy/BIOSTAT625-Project

Research Background

Why this data

- This data includes a wide variety of covariates including but not limited to operating costs, total revenue, bad debt expense, and uncompensated care costs, which are essential for analyzing the financial health of hospitals.
- The data spans from 2011 to 2022, enabling the construction of panel data to capture temporal patterns and trends in hospital finances.
- In addition to financial data, it includes operational details like bed counts, urban vs. rural location, and other facility characteristics, which can serve as covariates to enhance model accuracy.
- The dataset consists of over 50,000 observations, with approximately 8,000 samples collected each year.

Variable Selection

Dependent Variables

We construct our key dependent variables as:

- Cost-to-Revenue Ratio = Operating Costs / Total Revenue.
- Revenue per Bed = Total Revenue / Number of Beds.

Duplicates

- Duplicates were identified based on the Provider CMS Certification Number (CCN) and year.
- For duplicates, the mean of numeric variables was taken, and the first occurrence was kept for non-numeric variables.
- Fully missing numeric columns within each group were filtered out.

Data Cleaning

Outlier Removal

- Cost-to-Revenue Ratio: Data points with a Cost-to-Revenue Ratio greater than 100 were removed as they were considered outliers.
- Revenue per Bed: Values of Revenue per Bed greater than 100 million (scaled by dividing by 1 million) were also removed.

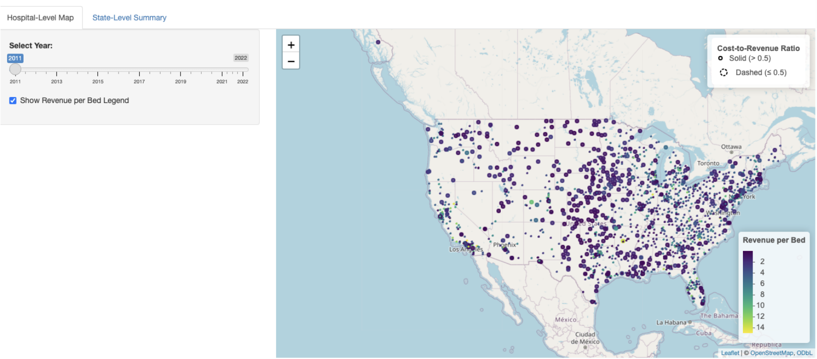
Missing Data Imputation

Numeric columns with missing values were interpolated using the *zoo::na.approx()* function based on the year if enough data points (more than two) were present. If data points were sparse, no interpolation was applied.

Interactive US Map - R Shiny Application

Deployed App: Access the interactive visualization at https://xxchar.shinyapps.io/hospital-financial-analysis-ui/.

Shiny Tab 1: Dot Distribution Map



Hospital-Level Map | State-Level Summary

Select Year: 2011 (range 2011–2022)

☑ Show Revenue per Bed Legend

Cost-to-Revenue Ratio
- ○ Solid (> 0.5)
- ◌ Dashed (≤ 0.5)

Revenue per Bed: 2, 4, 6, 8, 10, 12, 14

Shiny Tab 1: Dot Distribution Map

Objective: Visualize hospital-level financial metrics on an interactive map (*Sample of 4000 hospitals*). Slider from 2011 to 2022.

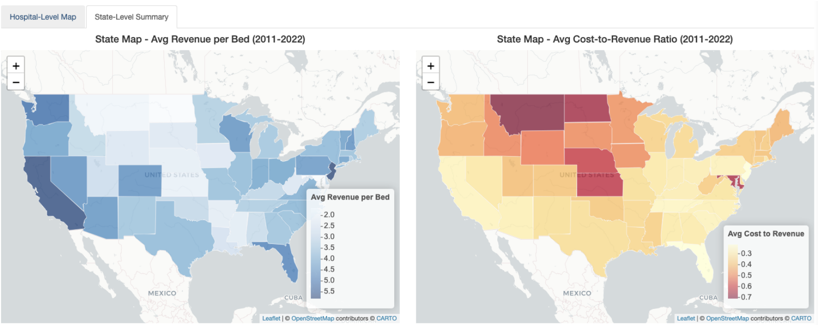
- **Dot Color:** Gradient reflects Revenue per Bed (darker = higher).
- **Dot Size:** Scaled by Cost-to-Revenue Ratio
- **Interactive Popups:** Show hospital name, location, Revenue per Bed, and Cost-to-Revenue Ratio.

Key Insights

- **Revenue per Bed:** Most hospitals earn less than 2 per bed, forming a dominant cluster of low-revenue facilities.
- **Sparse Distribution:** Rocky Mountain states (NV, UT, WY, SD, ND) have fewer hospitals.
- **Cost-to-Revenue:**

  Inland States (IL, MO, AL): Lower ratios (<0.5), indicating lower efficiency. **Coastal Areas:** Higher ratios (>0.5), suggesting better financial performance.

Shiny Tab 2: Aggregated State-Level Map


Hospital-Level Map
State-Level Summary
State Map - Avg Revenue per Bed (2011-2022)
+
−
UNITED STATES
MEXICO
CUBA
Avg Revenue per Bed
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5
Leaflet | © OpenStreetMap contributors © CARTO
State Map - Avg Cost-to-Revenue Ratio (2011-2022)
+
−
UNITED STATES
MEXICO
CUBA
Avg Cost to Revenue
0.3
0.4
0.5
0.6
0.7
Leaflet | © OpenStreetMap contributors © CARTO

Shiny Tab 2: Aggregated State-Level Map

Objective: Visualize state-level financial performance metrics.

- Interactive popups show aggregated metrics per state.
- Color gradients provide clear insights into state-level patterns.

Key Insights

- **West Coast:** Higher Revenue per Bed (e.g., California) driven by well-funded hospitals in urban regions.
- **East Coast:** Mixed trends; urbanized states exhibit higher financial efficiency.
- **Inland States:** Lower Revenue per Bed and higher Cost-to-Revenue Ratios indicate potential funding or operational challenges.

Summary Statistics

Covariates: key hospital-related variables such as inpatient and outpatient charges, total salaries, and hospital type.

Table: Summary Statistics for Revenue per Bed and Related Variables

Variable	**N (Hospitals)**= 7615
Revenue per Bed	105.4 (25.3)
Cost-to-Revenue Ratio	36.3 (25.2)
Total Discharges	5888.3 (8879.3)
Hospital Total Days	23787.2 (35755.5)
Total Salaries	65.9M (131.2M)
Inpatient Total Charges	338.7M (723.8M)
Outpatient Total Charges	307.4M (601.6M)
Total Income	14.9M (77.8M)
Total Other Income	17.9M (78.9M)
Total Liabilities and Fund Balances	246.2M (764.9M)
Accounts Payable	10.7M (62.2M)
Total Current Assets	83.2M (312.0M)
Total Fixed Assets	90.4M (217.8M)
General Fund Balance	141.3M (414.9M)
Inventory	3.6M (12.1M)
Total Patient Revenue	667.9M (1323.2M)
Number of Beds	168.8 (66.7)

Statistical Distribution

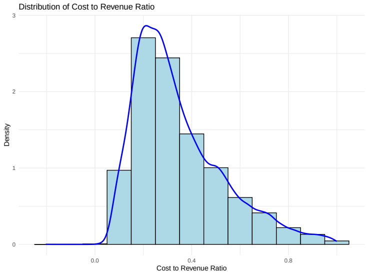


Figure: Distribution of Cost to Revenue Ratio

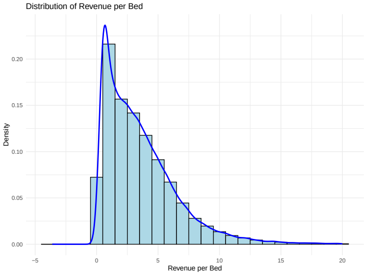


Figure: Distribution of Revenue per Bed

Linear Model

We consider a linear regression model to capture what is the correlation between the covariate variables with the dependent variables, specifically, we consider the model as:

$$Y_{it} = \beta_0 + \beta \mathbf{X_{it}} + \mu_{\mathbf{i}} + \tau_{\mathbf{t}} + \epsilon_{\mathbf{it}}$$

Description

Y_{it}: Target variable (e.g., total revenue or operating costs). β_i: Coefficients for predictors. $\mathbf{X_{it}}$ are the key independent variables that are reported in the statistics summary table, μ_i is the state level fixed effect and the τ_t is the time fixed effect (year). ϵ_{it}: Error term.

Linear Regression Model Result

Table: Linear Model: Dependent Variable - Cost to Revenue Ratio

Variable	Coefficient	SE
(Intercept)	6.064e+00 ***	(5.919e-01)
Total Discharges (V-XVIII, XIX, Unknown)	-1.349e-05 ***	(3.523e-07)
Hospital Total Days (V-XVIII, XIX, Unknown, Adults & Peds)	1.997e-06 ***	(9.172e-08)
Total Salaries (Worksheet A)	2.829e-10 ***	(2.038e-11)
Inpatient Total Charges	9.592e-11 ***	(8.149e-12)
Outpatient Total Charges	1.281e-11	(8.463e-12)
Total Income	-2.949e-12	(1.344e-11)
Total Other Income	8.889e-11 ***	(1.646e-11)
Total Liabilities and Fund Balances	6.370e-12 *	(2.511e-12)
Accounts Payable	-1.521e-11	(1.660e-11)
Total Current Assets	1.437e-11 ***	(4.202e-12)
Total Fixed Assets	5.201e-11 ***	(8.910e-12)
General Fund Balance	-8.667e-12 *	(3.803e-12)
Inventory	-1.587e-10	(9.254e-11)
Total Patient Revenue	-1.016e-10 ***	(7.822e-12)
Number of Beds	2.471e-08	(1.349e-07)

Note: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Even though the coefficient seems to be very small, and this is due to the fact that the data's value is pretty large.

Linear Regression Model Result II

Table: Linear Model: Dependent Variable - Revenue Per Bed

Variable	Coefficient	SE
(Intercept)	-3.296e+02 ***	(8.567e+00)
Total Discharges (V-XVIII, XIX, Unknown)	1.361e-04 ***	(5.099e-06)
Hospital Total Days (V-XVIII, XIX, Unknown, Adults & Peds)	-6.328e-05 ***	(1.328e-06)
Total Salaries (Worksheet A)	-2.273e-09 ***	(2.949e-10)
Inpatient Total Charges	-2.910e-09 ***	(1.179e-10)
Outpatient Total Charges	1.656e-10	(1.225e-10)
Total Income	7.628e-10 ***	(1.945e-10)
Total Other Income	-1.581e-09 ***	(2.382e-10)
Total Liabilities and Fund Balances	-2.736e-10 ***	(3.634e-11)
Accounts Payable	2.875e-10	(2.403e-10)
Total Current Assets	1.503e-10 *	(6.082e-11)
Total Fixed Assets	1.898e-10	(1.290e-10)
General Fund Balance	-3.383e-11	(5.504e-11)
Inventory	9.308e-10	(1.339e-09)
Total Patient Revenue	3.572e-09 ***	(1.132e-10)
Number of Beds	-3.784e-06	(1.952e-06)
Year	1.654e-01 ***	(4.246e-03)

Note: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Machine Learning Methods

To predict provider's Cost-To-Revenue using only variables concerning the scale of provider and account for the non-linear relationship, we choose support vector machine(SVM) and k-Nearest Neighbors(k-NN) with different kernels.

Variables included: Location, Type of providers, Total number of employees, Total number of inpatient days, The number of beds, The number of discharges, Assets, Cash on hand and in banks, Inventory, Buildings, Fund balances

Machine Learning Methods

Table: The RMSE on 5 folds of SVMs

	Linear	Polynomial	Radial
SVM	0.168	1.515	0.144

Table: The RMSE on 5 folds of k-NNs

	Rectangular	Triangular	Gaussian
k-NN	0.099	0.086	0.093

Computational Challenge: K-fold cross validation on large datasets are slow, and it is hard to choose the optimal parameters.

Limitation: The variables selection is subjective, later we may conduct PCA and Sensitivity analysis to choose variables.

Conclusion

- **Outcome:**
  - For the cost-to-revenue ratio, total income, salaries, and patient revenue strongly influence hospital efficiency, emphasizing the need for effective cost management and resource use.
  - For revenue per bed, patient discharges, income, and hospital size are key factors. Optimizing patient flow, resource use, and financial planning can enhance financial returns while maintaining quality care.
- **Future Work:**
  - R shiny could integrate more additional metrics, utilizing existing structure to enhance performance.
  - It is important to recognize that while predictive models can offer valuable insights, we still need to explore further to check which variables are of most significance.