David DeStephano

Amanda Howarth

Data Science II

Midterm Report

**Background / Motivation**

Total hospital charges vary greatly for patients across the United States. In 2012, it was estimated that the mean hospital cost per stay was $10,400 (Moore et al., 2014). However, the cost per stay varies by different factors, including a patient's demographic information, the severity of their health condition, and type of admission. The costs of having a baby and the costs of infant hospitalizations also differ based on many different factors. In 2004, March of Dimes reported that the average cost of having a baby was $8,802. This cost varied by factors such as type of birth, whether it be a Cesarean-section of vaginal birth ($10,958 vs $7,737) (March of Dimes, 2007). This cost was also affected by the birthweight of the newborn, with the average cost per stay increasing to $15,100 for pre-term/low birthweight babies (Russell et al, 2007). For babies born at a weight less than 1000 grams, the average cost of hospital stay was reported to be $65,600 (Russell et al, 2007).

The ability to predict the total cost of a patient’s stay at a hospital based on known patient characteristics, hospital information, and diagnosis would provide significant benefits for health insurance companies, patients, and hospital systems. If a soon-to-be mother was knowledgeable of the predicted charges of her upcoming birth at different hospitals, she would be better informed in choosing which hospital she would like to be admitted to. Additionally, a newborn/infant's parent can choose a hospital that is best for their family economically if their child is having health complications.

The goal of this analysis is to determine the best model to predict the total healthcare charges of newborn and infant patients who were admitted to hospitals in New York City. We used the Hospital Inpatient Discharges (SPARCS De-Identified) 2013 data set. This data was provided by the New York State Department of Health’s Office of Quality and Patient Safety and includes 2.43 million observations and 34 variables, such as features for demographic information, hospital stay, payment typology, cost information.

In order to load the dataset to my computer, this dataset was limited (through the NY State's website filter) to the hospital service area of New York City and includes information from hospitals in all five boroughs. The dataset was limited to include only patients who had birthweight information available, recorded during their hospital stay. Over 150+ CCS diagnosis codes were included in the original dataset. The CCS diagnosis codes were categorized into 17 diagnosis types, such as "infectious diseases" and" pregnancy and childbirth complications." Individuals with missing values (n = 119) were dropped from the dataset, totaling 121,380 patients in the final dataset.

Nineteen variables were dropped from the dataset. Variables were dropped if they were redundant of other variables or if they were not deemed to be clinically meaningful in predicting healthcare hospitalization charges. All categorical variables were turned into dummy variables with referent groups. The final dataset used for models in this analysis (with dummy variables) included 92 predictors and 1 response variable (total charges).

**Our research questions were the following:**

* Which approach (linear regression, lasso regression, ridge regression, principle component regression, or generalized additive model) best predicts a patient’s total healthcare charges per hospital stay in New York?
* What is the root mean square error (RMSE) of each model?
* What percent of the variance in total healthcare charges is explained by the predictors available in SPARCS?

**Exploratory Analysis Findings / Visualization**

This dataset includes 121,380 patients who had their birthweight recorded at the time of their hospital visit. The mean total charges for a patient at discharge across the 49 hospital facilities was $19,787.95. The median total charges for a patient was $6,590.02. For exploratory analysis, we compared the median total charges because there were extreme outliers with very high charges. The median is a better measure of central tendency than the mean. The median total charges varied widely across the 49 facilities included in this analysis. The median total charges for a patient at discharge were highest at Montefiore Medical Center - Henry & Lucy Moses Div ($25,160.36) and lowest at North Central Bronx Hospital ($2,407.01) (Table 1).

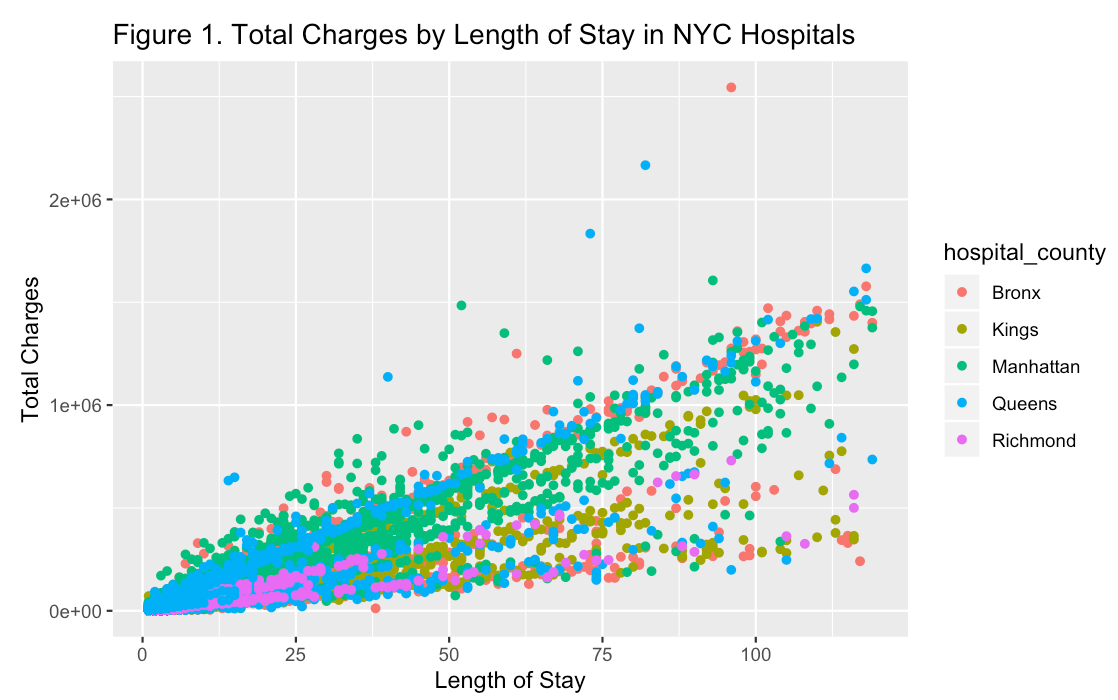
In Figure 1, there is a positive correlation between length of stay and the total charges at discharge. The relationship between length of stay and total charges demonstrates linearity. Thus, as the length of stay increases, the total charges increased. The charges appear to be relatively lower at hospitals in Richmond County (Staten Island), even as the number of days spent in the hospital increases. Additionally, it appears that the charges at hospitals in Manhattan are typically higher compared to other counties across the lengths of stay.

Figure 2 demonstrates that the total charges generally decrease across increasing birthweight. There is a moderately negative correlation between baby birthweight and total charges. Thus, patients with lower birthweights pay the highest total charges at discharge.

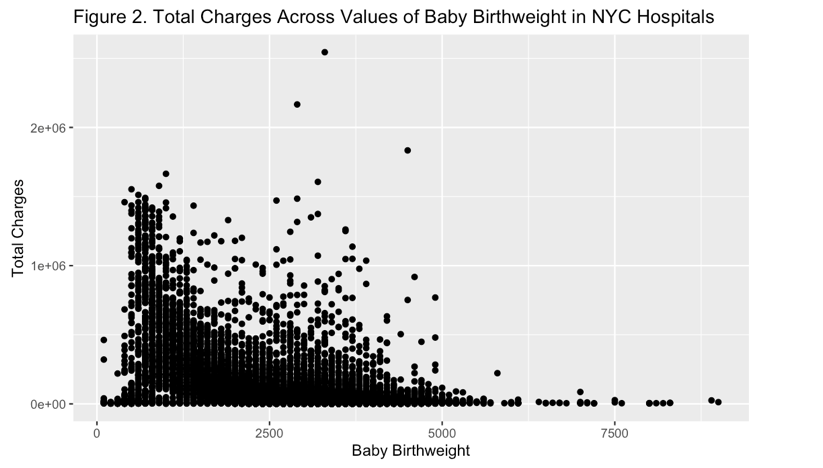
**Table 1. Median Total Charges by Hospital Facility**

|  |  |  |
| --- | --- | --- |
| Facility Name | Hospital County | Median Total Charge |
| Montefiore Medical Center - Henry & Lucy Moses Div | Bronx | 25160.36 |
| Brookdale Hospital Medical Center | Kings | 11561.00 |
| Long Island Jewish Schneiders Children’s Hospital Division | Queens | 11498.00 |
| NYU Hospitals Center | Manhattan | 10573.70 |
| New York Presbyterian Hospital - Columbia Presbyterian Center | Manhattan | 9383.75 |
| NewYork-Presbyterian/Queens | Queens | 9361.96 |
| New York Presbyterian Hospital - Allen Hospital | Manhattan | 9335.04 |
| New York Presbyterian Hospital - New York Weill Cornell Center | Manhattan | 8879.03 |
| Lenox Hill Hospital | Manhattan | 8868.30 |
| Montefiore Medical Center-Wakefield Hospital | Bronx | 8520.85 |
| Montefiore Medical Center - North Division | Bronx | 8083.35 |
| Forest Hills Hospital | Queens | 7946.70 |
| Mount Sinai West | Manhattan | 7897.00 |
| New York Hospital Medical Center of Queens | Queens | 7592.46 |
| New York Hospital Medical Center of Queens | Queens | 7522.46 |
| Montefiore Med Center - Jack D Weiler Hosp of A Einstein College Div | Bronx | 7136.95 |
| University Hospital of Brooklyn | Kings | 6771.00 |
| New York-Presbyterian/Queens | Queens | 6744.78 |
| Coney Island Hospital | Kings | 6331.11 |
| Mount Sinai Roosevelt | Manhattan | 6314.50 |
| Lutheran Medical Center | Kings | 6286.70 |
| St Lukes Roosevelt Hospital Center - Roosevelt Hospital Division | Manhattan | 5980.50 |
| New York Methodist Hospital | Kings | 5719.94 |
| Beth Israel Medical Center/Petrie Campus | Manhattan | 5676.00 |
| Mount Sinai Beth Israel | Manhattan | 5673.40 |
| SUNY Downstate Medical Center at LICH | Kings | 5367.00 |
| Maimonides Medical Center | Kings | 4993.00 |
| St Johns Episcopal Hospital So Shore | Queens | 4910.00 |
| Staten Island University Hosp-North | Richmond | 4820.00 |
| Mount Sinai Hospital | Manhattan | 4787.20 |
| Woodhull Medical & Mental Health Center | Kings | 4547.34 |
| Kings County Hospital Center | Kings | 4540.02 |
| Queens Hospital Center | Queens | 4527.31 |
| Bellevue Hospital Center | Manhattan | 4527.26 |
| Brooklyn Hospital Center - Downtown Campus | Kings | 4470.44 |
| Jacobi Medical Center | Bronx | 4461.91 |
| Richmond University Medical Center | Richmond | 4438.52 |
| St Barnabas Hospital | Bronx | 4394.39 |
| SBH Health System | Bronx | 4391.21 |
| Lincoln Medical & Mental Health Center | Bronx | 4376.42 |
| Harlem Hospital Center | Manhattan | 4361.67 |
| Wyckoff Heights Medical Center | Kings | 4335.00 |
| Metropolitan Hospital Center | Manhattan | 4299.00 |
| Elmhurst Hospital Center | Queens | 4285.44 |
| Jamaica Hospital Medical Center | Queens | 3692.00 |
| New York-Presbyterian/Lower Manhattan Hospital | Manhattan | 3498.26 |
| Flushing Hospital Medical Center | Queens | 3397.00 |
| Bronx-Lebanon Hospital Center - Concourse Division | Bronx | 3370.57 |
| North Central Bronx Hospital | Bronx | 2407.01 |

**Figure 1: Total Charges by Length of Stay in NYC Hospitals**

****

**Figure 2: Total Charges Across Values of Baby Birthweight in NYC Hospitals**

****

**Models**

We fit 5 different models using cross validation to determine which model would fit the data best:

* One linear model,
* Three regularized linear models (Lasso regression, Ridge regression, PCR), and
* One non-linear model (Generalized Additive Model).

Predictor variables included: 1) hospital county (Manhattan, Bronx, King's County, Richmond County, and Queens), 2) hospital facility, 3)gender, 4) race, 5) ethnicity, 6) length of stay, 7) type of admission, 8) APR severity of illness code, 9) APR medical-surgical description (medical or surgical), 10) type of payment, 11) birth weight, 12) emergency department indicator, 13) APR risk of mortality, 14) total costs, and 15) type of CCS diagnosis. All categorical variables were made into dummy variables. Three predictors were continuous (length of stay, birthweight, and total costs) as well as the outcome variable (total charges).

Our dataset (92 predictors total, with dummy variables) was partitioned into a training and test data set. Five models were fitted using the training data and the mean squared error (MSE) was calculated for each model using the test data. We measured the mean squared error to quantify the extent to which the predicted response value for a given observation is close to the true response value for that observation.

First, we fit a linear model using least squares on all the predictors in the training data. We found that the MSE calculated on the test data was 937604771. By looking at the coefficient values for the linear model, some of the dummy variables were not significant at an alpha value of 0.05 in predicting total charges, including ethnicity, emergency department indicator, and certain CCS diagnosis types.

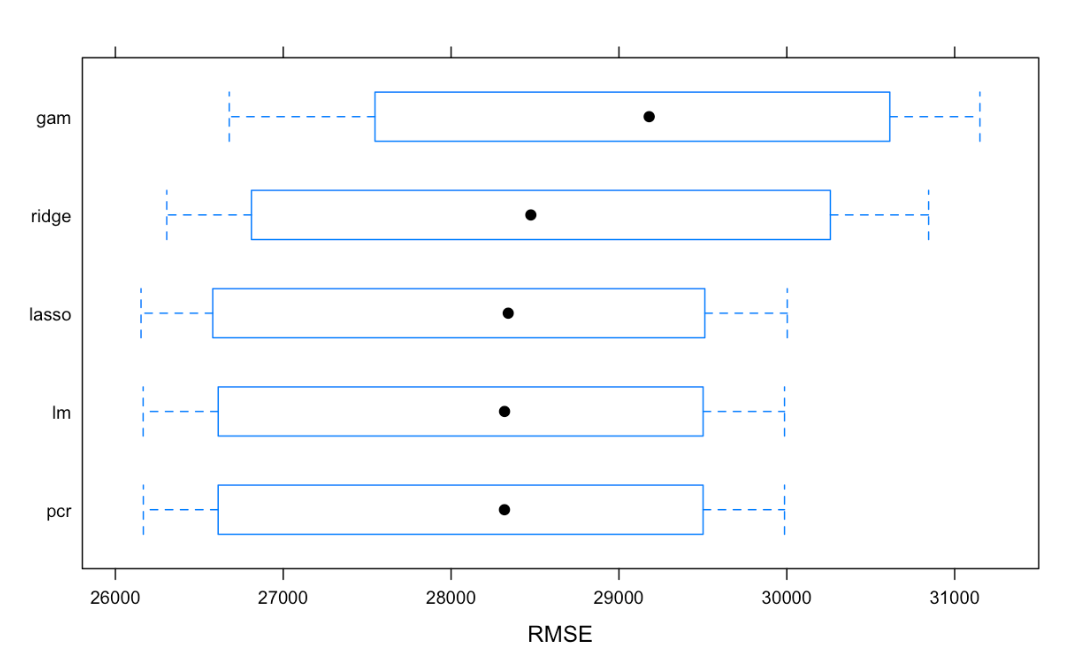
Next, we fit two models on all the predictor variables using two different techniques that "shrink" the coefficient estimates towards zero, which reduces variance. We fit a ridge regression model on the training data. Alpha was held at a value of 0 and our final lambda value chosen by cross-validation was 5355.389. Using our test data, we found that test error (MSE) was 9790669585. Next, we fit a lasso model on the training data. Alpha was held at a value of 1 and our final lambda value chosen by cross-validation was 28.03162. Our test error was 937116653.

We also fit a principle component regression (PCR) model on the training data with M chosen by cross validation. The PCR method constructs the first M principal components. Our M-value was 87 and our test error was 937635353. A benefit of using PCR is that this method avoids multicollinearity between variables in the dataset. A limitation of PCR is that there is no guarantee that Z­m­ are the best linear combinations of the variables in predicting the response.

Lastly, we fit a GAM model to our training dataset. The benefit of using a GAM model is that it will automatically model non-linear relationships that standard linear regression would miss. S functions were applied to the three continuous variables in the model (total costs, length of stay, and birthweight).

The best model for predicting total charges was PCR. This model was chosen because it has the smallest cross validation RMSE, a value of 28173.68 (Figure 3). For the PCR model, 85.69% of the variance in total charges is explained by the predictors in the model. The mean RMSE value of the lasso regression model (28174.24) and linear regression model (28180.84) were very close to the PCR model (depicted in Figure 4). Overall, GAM had the lowest RMSE (29076.52), which was unexpected because it was the only model to account for non-linear relationships that a standard linear regression model would miss.

**Figure 3: RMSE of GAM, Ridge, Lasso, Linear, and PCR Models**

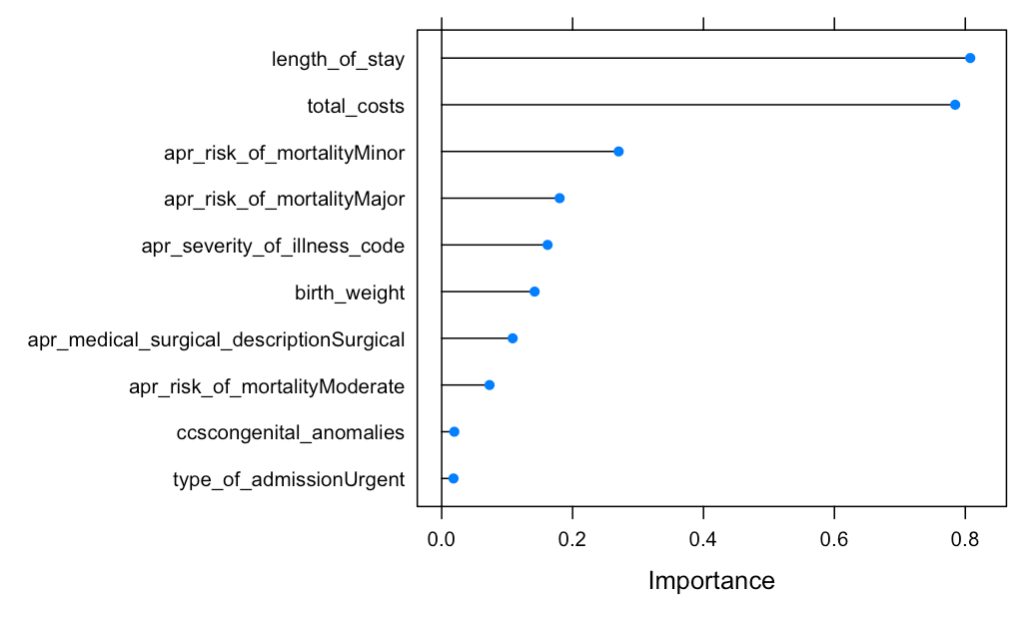
****

**Table 2. R-Squared Values Per Model**

|  |  |
| --- | --- |
| Model | Mean R-squared |
| Linear | 0.85688 |
| Lasso | 0.85683 |
| PCR | 0.85689 |
| GAM | 0.84753 |
| Ridge | 0.85482 |

For the PCR model, the top 10 variables that played important roles in predicting a patient’s total charges are displayed in Figure 4. These variables were chosen using the varImp() function, which automatically chooses a measure of variable importance that is appropriate for given algorithms. The top five variables included length of stay, total costs, APR risk of mortality (major), APR risk of mortality (minor), and the APR severity of illness code. The greatest importance in predicting total healthcare charges was length of stay in the hospital, which aligns with the linear relationship depicted in Figure 1.

**Figure 4: Top 10 Important Variables in Predicting Patient Total Charges at Discharge**



Limitations of this analysis included that non-significant variables were included in models and not dropped to improve the model. This was done in an effort to compare models, by first including all predictors in the model building phase. Thus, if we were to drop variables that are neither clinically meaningful nor significant, this may have created more parsimonious and efficient models. Additionally, the highest R^2 value was 85.69%. The remaining variance of total charges are thus unexplained and may be due to variables that were dropped (such as patient’s zip code) or by features not collected in this dataset.

**Conclusion**

Of all models, the PCR model best predicted total healthcare charges of patients at time of discharge. The PCR model had very similar RMSE values to the linear model and lasso regression model, and thus, they have similar prediction performance. By listing the most important variables in predicting charges, we gained insight into the most important predictors for healthcare charges, particularly gaining knowledge that length of stay and total costs have much greater importance in predicting the charges compared to other variables. Birthweight is the top 6th highest importance, which is expected, as it often recorded in literature. It was of valuable insight to learn that the APR risk of mortality (minor, major, and moderate) has strong importance in determining total charges as well as congenital anomalies and Urgent admission types.

**References**

Moore, B., Levit, K., Elixhauser, A. (2012). Costs for Hospital Stays in the United States, 2012. Retrieved from <https://www.hcup-us.ahrq.gov/reports/statbriefs/sb181-Hospital-Costs-United-States-2012.pdf>

The March of Dimes. (2007). The Healthcare Costs of Having a Baby. Retrieved from <https://www.marchofdimes.org/glue/materials/partner-the-healthcare-costs-of-having-a-baby.pdf>

Russell, R., Green, N., Steiner, C., Meikle, S., Howse, J., Poschman, K., Dias, T., Potetz, L., Davidkoff, M., Damus, K., Petrini, J. (2007). Cost of hospitalization for preterm and low birthweight infants in the United States. *Pediatrics*, 120(1): E1-E9. Retrieved from

<https://pediatrics.aappublications.org/content/120/1/e1>

**Dataset**

**Dataset can be found here:** <https://health.data.ny.gov/Health/Hospital-Inpatient-Discharges-SPARCS-De-Identified/npsr-cm47>

**RMD file:** data\_science2\_midterm.RMD