Cornell University
Operations Research and
Information Engineering

ORIE 4741

Learning with Big Messy Data

Midterm Report

Matthew Dalton
mgd67

Katelyn Glassman
kmg237

October 29, 2016

1 Problem statement

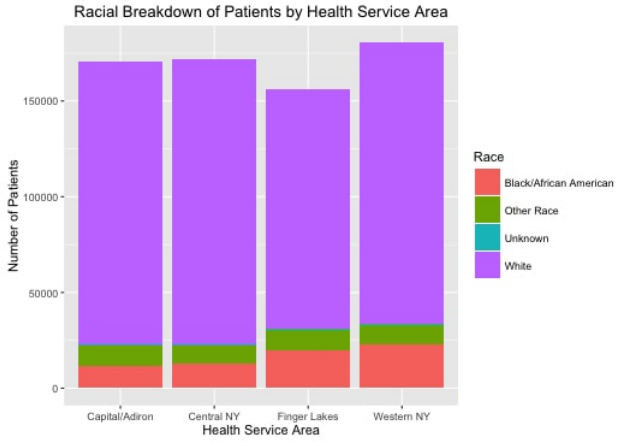
Our project's goal is to provide for a patient valuable insights regarding medical care based on their condition and some basic information. Given their personal information and medical status we hope to develop a system that can provide the patient with an estimate of their length of stay in a hospital, expected outcome, and costs incurred. In addition, we may try and make a recommendation on which hospital would best fit their needs.

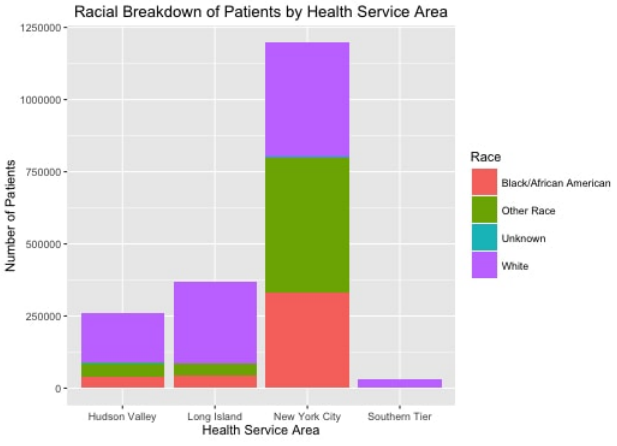
2 About the dataset

Our project examines the Statewide Planning and Research Cooperative System's (SPARCS) Hospital Inpatient Discharges for New York State. This dataset provides basic information about all patients that have been discharged from a New York hospital in the year 2012, including age group, gender, and race. It also provides information about each patient's length of stay, reason for admission, severity of illness, method of payment and costs incurred. Specifically, the dataset quantifies severity of illness on a scale from 1 to 4, with 1 being minor and 4 being extreme, as well as an ordinal field for risk of mortality.

The dataset is comprehensive and complete, with intuitive, colloquial values. Data validation was certainly employed in the collection of the data, as nominal fields like Diagnosis Description and Ethnicity are sensible and consistent. Furthermore, age values were pregrouped into ranges (0-17, 18-29, etc.), and lengths of stay beyond 120 days were truncated to be displayed as `120+`.

The histograms below show the racial breakdown by health service area of the regions provided in our dataset. Race was a a highly statistically significant feature in a preliminary regression across all features, motivating its inclusion in our models.





3 Data cleaning

To make our dataset more approachable we converted some nominal values into integer values using one-hot encoding to allow for regression. For example, we converted the single `Race` column which contained text values of a patient's race into three columns, `White, Black` and `Other Race`, and for each patient we placed a 1 in the applicable column and a 0 in the others. We used the same logic for the `Gender` column, splitting it into two binary columns for `Male` and `Female`, and for the `Method of Payment` column into `Private Insurance, Medicare`, and `Medicaid`.

Also, manipulation was required for some numerical values. `Length of Stay` values given as `120+` were converted to `120` in order to allow for analysis and eliminate outliers.

4 Preliminary models

To specify our study, we sought to examine cases of a sudden onset condition that requires immediate care. Our preliminary models were fit to data on heart attack diagnoses in New York hospitals and attempt to predict patient outcome. Our preliminary models formulated parameters to predict survival rate based on varying features and, through splitting the data into train and test sets, tested the accuracy of these fits on untrained data.

As shown in the file `Heart Attack Linear Models.ipynb` in our GitHub repository, we developed 3 models to be fit to the dataset that predict survival rate. Each model is fit to and cross-validated against a training set to develop a predictive fit parameter w, then this w is tested on a previously unvisited test set.

In order to determine which features are most influential and should be isolated and examined, we ran a preliminary regression across all of the features and picked those that were the most statistically significant.

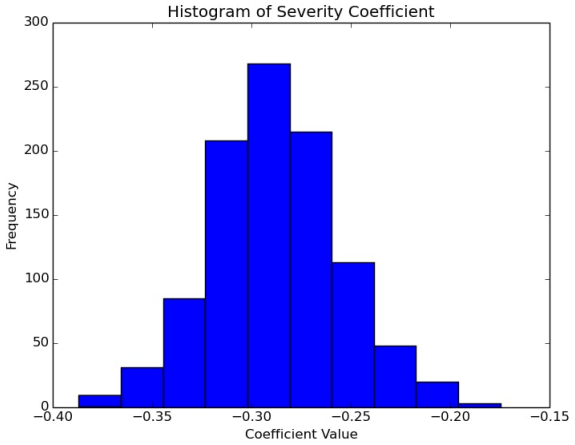
- Our first model fits a parameter against severity of illness and length of stay to predict survival rate.
- Our second model fits a parameter against severity of illness, length of stay, and gender to predict survival rate.
- Our third model fits a parameter against severity of illness, length of stay, gender, and age to predict survival rate.

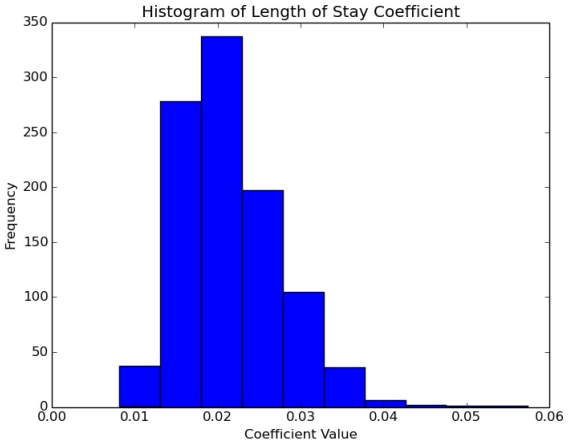
5 Validation

We've developed several methods to test our models' ability to generalize and avoid overfitting. In our training stage, our models are cross-validated against different partitions of the training set. We developed a bootstrap estimator that randomly resamples from the training set and generates an average w parameter that avoids overfitting. We then move to a test set and find an out-of-sample error calculation of our models' parameters w and the bootstraps estimation parameters across all three of our models.

At this stage, our out-of-sample error measure has high variance. We are working to develop less variant measures for our test set error evaluation, and would like to include measures of bias and variance into our error calculation so as to offer a more stable error measure.

The histograms below show the frequency of coefficient values from our parameter fits for Severity and Length of Stay as they relate to survival rate. The Severity coefficients appear to be near normally distributed, whereas the Length of Stay coefficient appears to be more skewed.





6 Moving forward

Going forward with our analysis, we would like to work to automate as much of this process as possible. We would like to have a program in place that can, given input, pull the existing data, run all models automatically, and determine through error evaluation which model would be most accurate. Once a system is in place that can accomplish this, we can manually go back through and refine models as needed.

As mentioned prior, our current out-of-sample error calculation produces high variance, and so we would like to refine this to offer a less variant measure of error. By including measures of bias and variance in this error calculation, we can stabilize our out-of-sample error measure. Also, we hope to add regularizers to models to further avoid overfitting and reduce variance of models.