Modeling Hospital Length of Stay

Morgan de Ferrante, Shuang Wu, Qing Xu, Weijie Liu

Abstract

The purpose of this analysis was to investigate variables that could be used to predict a patient's length of stay in the hospital. After adjusting and transforming variables, we used automatic and criteria based procedures to create a model to best fit the data. Our resulting model was a multiple linear regression including the variables *heartrate, is30dayreadmit, cindex*, *evisits, ageyear, temperature, respirationrate, insurancetype, bpsystolic, and bpdiastolic*,** and the resulting variables had a highly significant linear relationship with log length of stay. Although the relationship was significant, goodness of fit measures suggest a multiple linear regression may not be the best fit.

Introduction

Based on data from hospitals in 2012, the average cost of stay for an individual (at the average length of stay of 4.5 days) was $10,500. [1] Minimizing length of stay in a hospital is of major concern to a hospital's administration, as longer lengths of stay can lead to an increased financial burden for both the hospital and the patient. As a result of a program enacted by Medicare called the prospective payment system, hospitals are directly incentivized financially to minimize length of stay for patients. Of major concern with minimizing length of stay is ensuring the shorter stay does not come at the cost of quality - studies have shown that reducing length of stay can come at the risk of increased readmission. [2] By looking at the data set which contains a total of 3682 records from 3612 patients, we would like to find out which variables are associated with length of stay in hospital (in days) and build a model that could be used to predict a patient's length of stay.

Methods

Data Cleaning and Manipulation

Before building our multiple linear regression model, we cleaned the data and made various adjustments. Since it was possible a patient visited the hospital more than once, we chose each patient's first visit. Additionally, we

[1] Weiss AJ (Truven Health Analytics), Elixhauser A (AHRQ). Overview of Hospital Stays in the United States, 2012. HCUP Statistical Brief #180. October 2014. Agency for Healthcare Research and Quality, Rockville, MD. http://www.hcup-us.ahrq.gov/reports/statbriefs/sb180-Hospitalizations-United-States- 2012.pdf.

[2] Heggestad, Torhild. "Do Hospital Length of Stay and Staffing Ratio Affect Elderly Patients' Risk of Readmission? A Nation-Wide Study of Norwegian Hospitals."*Health Services Research*, Blackwell Science Inc, June 2002, www.ncbi.nlm.nih.gov/pmc/articles/PMC1434663/.#

** more details about variables can be found in the appendix

excluded observations from patients that visited the Intensive Care Unit (ICU) since it is likely that ICU patients will have more hospital-related conditions and thus a longer length of stay. By looking at the range of continuous variables, we found some "funny" values that were not possible. We converted them into missing values if they met the following criteria: *bmi* (< 7 and > 50), *heartrate*: (> 220), *o2sat* (>200), *respirationrate* (< 50) and *temperature*(< 24 and > 47). ** After inspecting tables of individual categorical variables, we concluded there were too many levels in some factors, and found some levels only contained a few observations. Thus we combined different levels of *MEWS, cindex, race, religion* and *maritalstatus* into fewer levels based on frequency and practical importance. Levels of the original *MEWS* variable were combined into 2 levels (0-3 = normal/increase caution, >3 = further deterioration/immediate action required). Levels of the original *cindex* variable were combined into 4 levels (0 = normal (32.6%), 1-2 = mild (36.9%), 3-4 = moderate (11.8%) and >4 = severe (18.7%). For *race*, we re-coded "Natv Hawaii/Pacf Isl " into "Other/Multiracial" since the level only had a few observations and modeling predictions based on a level with only a few observations would not be practical . Additionally for r*eligion*, we combined "Angelican", "Hebrew", "Mormon" and "Non Denominational" into the level "Other". For *Maritalstatus*, since there was only one case in "*Civil Union*", we combined "*Civil Union*" into "*Married*".

Missing Values

We also made adjustments to variables with missing values. The continuous *bpsystolic, o2sat, temperature, heartrate, respirationrate,* and *bpdiastolic* variables each had less than five missing values, so we imputed these missing values by the mean of each variable. *BMI* had a very large number of missing values and was excluded from all of our models. The categorical variables *MEWS*, *maritalstatus* and *insurancetype* had too many missing observations to be imputed by mode, and too few missing observations to be entirely excluded, thus observations with missing data for these variables were excluded from our models. The resulting dataset contained 3282 observations with 17 variables.

Linear Regression

We built several simple linear models between the outcome (log of length of stay) and each variable to investigate which variables were significantly associated with the outcome. We excluded *gender, race,* and *maritalstatus,* due to the fact that they did not have significant linear relationships with outcome. Religion had only a few significant

[3] **descriptive statistics on odd variables found in appendix

levels, and we ultimately decided to remove the *religion* variable from our model since the R-squared was pretty low, showing religion was not capturing much of the variability in log of length of stay. Not all of the levels of *insurancetype* were significant, but we felt a patient's health insurance would have a major impact on how long they could stay in the hospital and we kept it in our models. We then used both stepwise regression based on AIC and Criterion-Based Procedures in R to find the best model. The stepwise regression returned a model with ten variables. This model agreed with the model generated from criterion-based procedures, using Adjusted R-squared, Cp and BIC. The final model showed that log of length of time is associated with *is30dayreadmit*, *cindex, evisit, ageyear , temperature , heartrate , respirationrate, bpdiastolic, bpsystolic* and *insurancetype.*

In checking model assumptions, we found that the assumption of normality was met towards the center of our QQ-plot, but there were wide tails at the end, implying outliers.** We identified many outliers in our outcome variable (54 observations) using studentized residuals, and ultimately decided to remove these outliers. We investigated how this subpopulation of outliers in our data differed from the rest of our study population, and using t-tests and Chi-squared tests, we found that the group of outliers was significantly different in age and insurance type. In the group of outliers, there were significantly more individuals who had private insurance versus Medicaid and individuals were younger in general. Length of stay in the outliers had a much greater range, but the mean length of stay in the outliers data and the mean length stay in the data without outliers are relatively similar (Fig. 1)

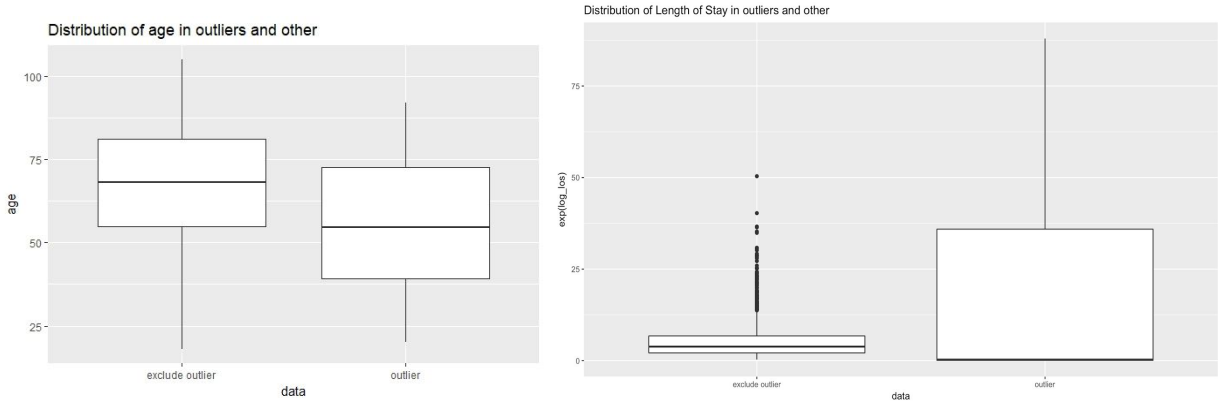

Fig.1 Comparison of Outlier Data and Full Data without Outliers

[4] ** plots with full data before removing outliers are in appendix

The following plots (Fig. 2) assisted us in checking the assumptions of our final model after the removal of outliers. From the Normal Q-Q plot, we can see that the residuals are normally distributed, and the standardized residuals versus fitted plot shows that the assumption of homoskedasticity is met (variance is constant across predictors), since the residuals are randomly scattered around zero.

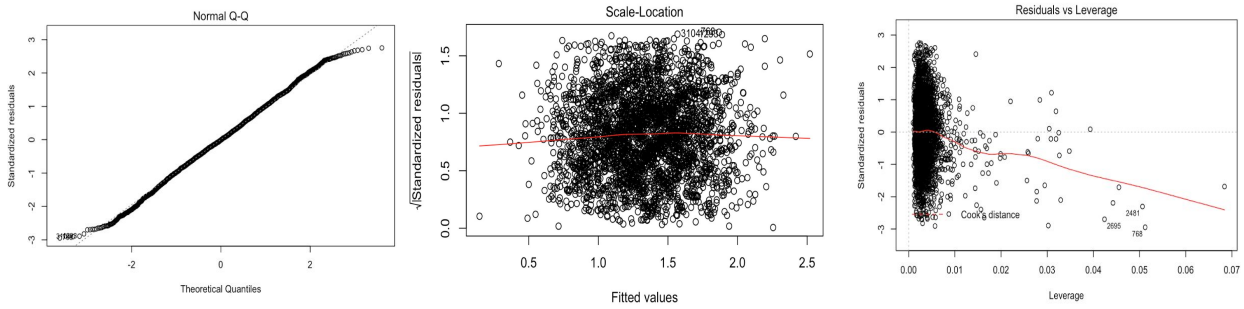

Fig.2 Final Model Diagnostics

Results

The model was found to be highly significant, with a p-value of < 2.2e-16, showing that there is a significant linear relationship between log of length of stay and our chosen variables (Table 1). The Adjusted R-Squared of 0.1508 shows us that 15.08% of the variability in log length of stay is explained by the combination of these variables.

Variable	Coefficients	Standard Error	95% Conf Intervals
Charlson comorbidity index (moderate)	0.1695	0.0439	0.0834, 0.2555
Charlson comorbidity index (normal)	-0.0102	0.0326	-0.0742, 0.0537
Charlson comorbidity index (severe)	0.1538	0.0375	0.0803, 0.2274
Emergency Dept Visits	0.0641	0.0088	0.0469, 0.0814
Age	0.0096	9e-04	0.0078, 0.0113
Temperature	0.0637	0.0197	0.025, 0.1024
Heart Rate	0.0074	0.0011	0.0053, 0.0095
Respiration Rate	0.0314	0.0059	0.0198, 0.0431
Diastolic BP	-0.0049	0.0016	-0.008, -0.0017
Systolic BP	-0.005	0.001	-0.0069, -0.0031
Insurance Type (Medicare)	-0.1418	0.0681	-0.2753, -0.0082
Insurance Type (Private)	-0.2101	0.0646	-0.3368, -0.0835

Table. 1

term	MSE
best model we choose	0.5498823
N-fold	0.5528415
10-fold	0.5853504

Table. 2

We investigated the predictive capability of our model in two parts: goodness of fit and predictive generalization. Since the adjusted R-squared was less than 20%, we conclude that our model with this dataset might not provide accurate predictions. For this reason, we believe other (non-linear) methods might be a better fit for this data, and possibly in general for predicting length of stay in hospitals. For predictive generalization, we used N-folds cross-validation and 10-folds cross-validation. The results are shown in Table 2. We can see that the mean squared error (MSE) of our final model is approximately 0.5499, while the MSE of N-fold cross validation and 10-fold cross

validation are approximately 0.5528 and 0.5854, respectively. There is little difference between the mean squared error in these cross validation methods and that of our final model, indicating our model has good predictive generalization.

Conclusion

For those patients who did not visit the ICU, we were able to make necessary adjustments to our data and build a multiple regression linear model for predicting the length of stay in the hospital. The model results indicated that there was a significant linear relationship between log length of stay in the hospital and the following variables: admission into the hospital within past 30 days, Charlson comorbidity index (CCI), number of times the patient visited an emergency department in the six months prior to admission, age , temperature, heart rate, respiration rate, diastolic blood pressure, systolic blood pressure and insurance type. Our final model met the necessary assumptions of normality and homoskedasticity, and the results of cross validation show that our model has relatively good predictive capability. Adjusted R-squared for our model was relatively small, and even though the F-test showed this was a statistically significant linear relationship, we suggest non-linear models be investigated.

Discussion

Further analysis should incorporate results from other statistical methods. One alternative to multiple linear regression for predicting length of stay is random forest regression. Random forests handle categorical variables well, and they require no assumptions about the underlying distribution. Random forests can also be easy to interpret and visualize, and be computationally quick for large dataset. In addition to random forests, we also suggest a hierarchical linear model (mixed model) be investigated. This model is particularly appropriate for research designs where data for participants are organized at more than one level. In this dataset, we found that the observations are collected from different hospitals and that observations within each hospital might be correlated, resulting in a violation of the assumption of independence. Thus, a mixed model would be a viable solution.

Appendix

Variables

LOSDays2: length of stay in the hospital (days)

Ls30DayReadmit: 1=admission into the hospital within past 30 days; 0=otherwise.

MEWS: The Modified Early Warning Score (MEWS) determines the degree of illness of a patient based on respiratory rate, oxygen saturation, temperature, blood pressure, heart rate, AVPU response; 0-1=normal, 2-3=increase caution, 4-5=further deterioration, >5 immediate action required

Cindex: Charlson comorbidity index (CCI) ranks patients based on severity of comorbidity: 0=normal, 1-2=mild, 3-4=moderate and >5=severeEvisit: number of times the patient visited an emergency department in the six months prior to admission (not including the emergency department visit immediately preceding the current admission)

ICU_Flag: 1=if during hospitalization, the patient had a visit in the intensive care unit (ICU); 0=otherwise. Note that ICU patients tend to have more hospital related conditions and thus a longer length of stay.

AgeYear: patient's age in yearsGender: patient's gender

Race: patient's race

Religion: patient's religion

MaritalStatus: patient's marital status

InsuranceType: patient's insurance

Vital Signs: respiration rate, blood pressure diastolic (BPD), oxygen saturation (O2), blood pressure diastolic (BPS), temperature, heart rate, and body mass index (BMI).

Descriptive Statistics of Numeric Variables (odd values in min/max)

continuous_variable <chr>	min <dbl>	first_quantile <dbl>	median <dbl>	third_quantile <dbl>	max <dbl>
ageyear	18.00000	54.00000	68.00000	81.00000	105.00000
bmi	3.10000	23.30000	27.08333	31.70000	122.65000
bpdiastolic	29.56349	66.05583	71.86660	77.96097	154.40000
bpsystolic	88.78261	118.00000	129.30385	141.47806	193.96296
heartrate	37.58333	71.27929	79.17029	87.54519	242.58333
o2sat	80.00000	96.50000	97.57143	98.61538	236.52632
respirationrate	12.00000	17.12500	17.76471	18.46154	67.71795
temperature	11.85000	36.61333	36.72727	36.86667	52.27500

Fig 3. Descriptive statistics of numerical variables

Skewness of Our Outcome

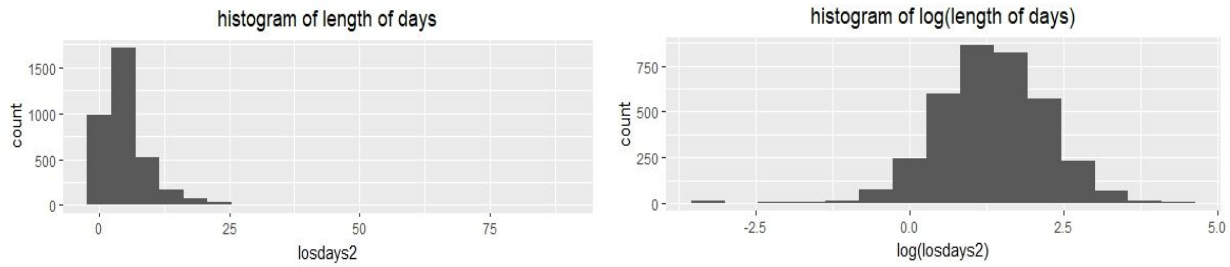

Fig.4 Histogram of *losdays* and log(*losdays*)

From the above histograms of length of stay, we can see that length of stay is skewed right, and that the natural logarithm of length of stay in days has a much more normal distribution. For this reason, all of our models use the outcome of log length of stay.

Final Model Plots before Outliers were removed

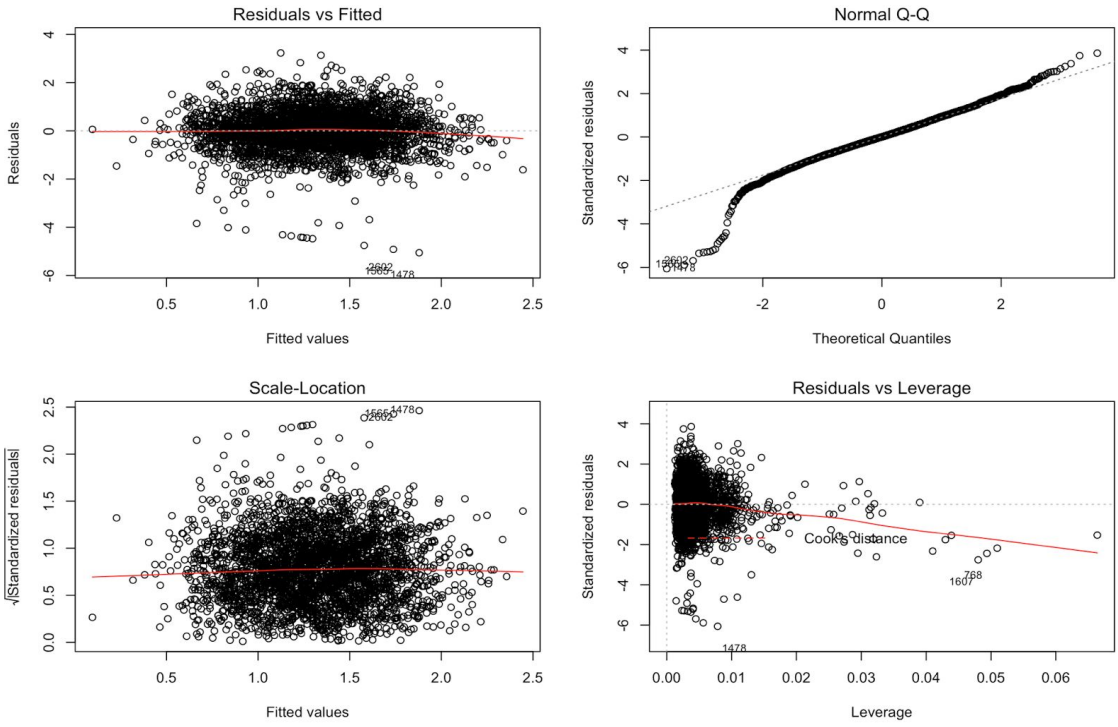

Fig. 5 Model Diagnostics