p8130 Final Project

Imaani Easthausen, Shanshan Song, Huijuan Zhang, Xinyan Zheng

December 9, 2017

Abstract

We present a linear regression model for the prediction of hospital length of stay. The model has good internal validity and explains approximately 17.5% of the variation in hospital length of stay.

Introduction

In an era of rising healthcare expenses, one of the major drivers of the costs in an inpatient care setting is the length of hospital stay. A better understanding of the characteristics driving length of hospital stay will be needed in order identify patients at risk for long hospital stays and implement strategies designed to help these patients leave the hospital earlier.

In order to better understand characteristics associated with increased length of stay, we develop a linear regrestion model to predict length of stay. A dataset containing 3612 are used to fit the model. The following covariates are considered for inclusion: age, gender, race, religion, marital status, insurance type, vital status, severity of comobidity (measured using Charlson comorbidity index), number of visits to the ER in the prior six months, whether the patient as admitted to the ICU. Descriptive statistics for all continuous covariates considered for inclusion are presented in table 1.

	Mean	Median	Standard Error	Quantile_1st	Quantile_3nd	Minimum	Maximum
Length of stay (hours)	129.9	91.0	139.3	51.0	161.0	1.0	1861.0
Length of stay (days)	5.41	3.79	5.81	2.13	6.71	0.04	77.54
Evisit	1.73	1.00	1.58	0.00	3.00	0.00	4.00
Age	65	67	19	53	80	18	101
BMI	28.4	27.1	8.1	23.3	31.7	3.1	123.0
02sat	97.9	97.6	5.27	96.53	98.6	80	236.5
Temperature	36.75	36.73	0.79	36.61	36.87	30.00	52.28
Heart Rate	80.20	79.35	13.31	71.12	87.71	37.58	242.58
Respiration Rate	18.2	17.8	2.78	17.1	18.5	12.0	67.7
mean_arterial_pressure	140.2	139.0	16.4	128.9	149.7	90.9	252.3

Table 1: Summary statistics for continuous covariates considered for inclusion in predictive models.

Methods

Data Cleaning

Data cleaning and analyses were conducted in R and SAS. For subjects with more than one hospitalization, only the first hospitalization was included in analyses. Subjects that were missing any of the following data were excluded: patient ID, visit ID, length of stay, hospital admission witin the last 30 days, MEW score, C index, number of admissions in 6 months prior, age, gender, race, religion, marital status, insurance type.

Due to the high degree of anticipated collinearity between systolic and diastolic blood pressure, mean arterial pressure was calculated from these two variables and used in all subsequent model fitting analyses.

The below categorical and ordinal variables were reclassified in order to combine levels that provided no new information or that had too few observations for meaninful analysis:

- MEW score: Levels 0 and 1 were reclassified as “normal”, levels 2 and 3 were reclassified as “increase caution”, levels 4 and 5 were reclassified as “further deterioration”, levels 6 and above were re-classified as immediate action.
- C index: Level 0 reclassified as “normal”, levels 1 and 2 were reclassified as “mild”, levels 3 and 4 were reclassified as “moderate”, levels 5 and above were reclassified as “severe.”
- Religion: “Angelican” was reclassified as “Christian,” “Hebrew” was reclassified as “Jewish”, “Non-Denominational” was reclassified as “Other,” “Catholic” was reclassified as “Christian”, “Mormon” was reclassified as “Other.”
- Marital Status: “Married” and Civil Union" were reclassified as “partnered,” all other categories were reclassified as “not partnered.”

Preliminary Analyses

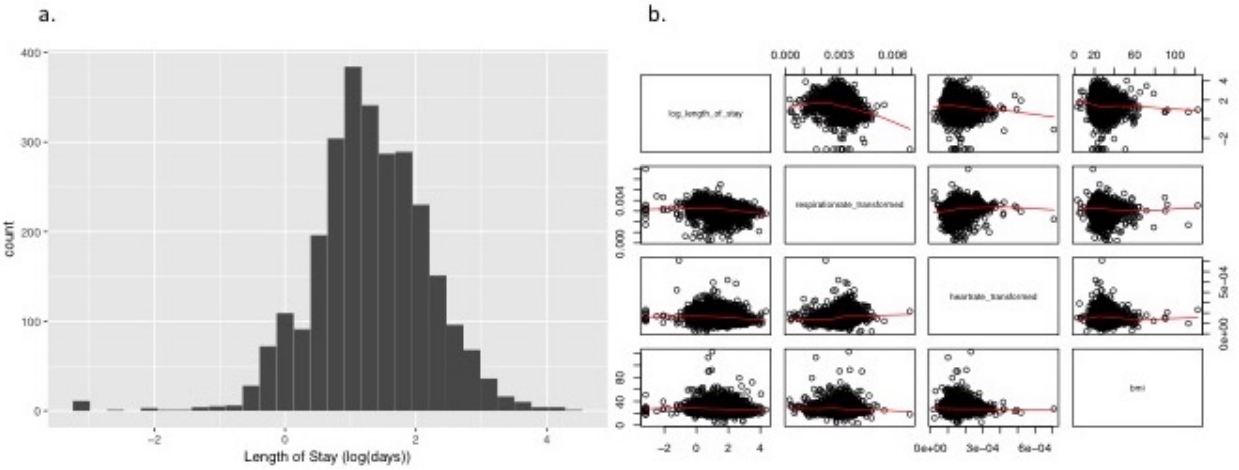

Figure 1: a) Histogram of log transformed outcome variable, and b) scatter plots showing relationships between continuous variables after transformations.

A histogram of LOS was visually inspected and substantial skewness was observed. Log transformation successfully rendered LOS approximatley normal (figure 1). Bivariate relationships between continuous predictors and length of stay were examined by visual inspection of scatter plots. Oxygen saturation, temperature, heart rate, and respiration rate appeared to have non-linear relationships with LOS. As such, respiration rate and heart rate were transformed using the below transformation equation:

$$T(X) = \frac{1}{X^2}$$

Oxygen saturation and temperature were re-classified into categorical variables as below:

- Oxygen saturation: saturations below 95% were defined as “low” and saturations from 95-100% were defined as “normal.”
- Temperature: Body temperatures bewteen 36.1 and 37.2 degrees were defined as “normal.” Temperatures below 36.1 degrees were defined as “low,” and temperatures above 37.2 degrees were defined as “high”.

After these adjustments were made, all continuous variables appeared to have approximately linear relationships with LOS by visual inspection of bivariate scatter plots (figure 1).

All pairs of continuous predictors were examined for collinearity by calculating all pairwise correlations. There did not appear to be substantial collinearity between any predictors (table 1).

	Age	Respiration Rate	Mean Arterial Pressure	Heart Rate	BMI
Age	1.0000000	-0.0789071	-0.0584134	0.1308305	-0.1327040
Respiration Rate	-0.0789071	1.0000000	0.0016222	0.1789010	-0.0783977
Mean Arterial Pressure	-0.0584134	0.0016222	1.0000000	-0.0653621	0.0900852
Heart Rate	0.1308305	0.1789010	-0.0653621	1.0000000	-0.0515141
BMI	-0.1327040	-0.0783977	0.0900852	-0.0515141	1.0000000

Table 2: Correlations between continuous covariates.

Variable Selection and Model Building

Variable selection was conducted using backward, forwards, and stepwise methods. Effect significance level of $alpha = 0.05$ was required for a variable to stay in or enter the model for all variable selection processes.

model	rss	adjr2	cp	aic	bic
With Outliers					
backward w/ outliers	1843.660	0.1352	11.0	6683.825	-317.9817
stepwise (p=0.05) w/outliers	1821.571	0.1446	14.0	6657.052	-327.0307
stepwise (p=0.10) w/outliers	1816.543	0.1463	16.0	6653.536	-318.7303
forward w/outliers	1818.582	0.1457	15.0	6654.587	-323.5877
Without Outliers					
backward w/o outliers	1425.000	0.1730	26.2	1013.000	-1693.0000
stepwise (p=0.05) w/o outliers	1425.000	0.1730	26.2	1013.000	-1693.0000
forward w/o outliers	1417.000	0.1750	26.3	1013.000	-1693.0000

Table 3: Diagnostic criteria for models with and without outlier data.

Variables selected by all methods include: 30 day readmit, number of ER visits in last six months, age, insurance type, heart rate, respiration rate, and mean arterial pressure. Additional variables selected using forward selection include: c-index, temperature, and partner status. Additional variables selected using backward selection include: gender and partner status. Additional varaibles choosen by stepwise selection include: c-index, temperature, and heart rate. The models selected by the forward and stepwise processes performed better that the model selected by the backwards variable selection process based on model diagnotic criteria including residual sum of squares, adjusted R-squared, BIC, and AIC (table 3). Diagnostic criteria for models selected in the forward and stepwise processes were similar to one another (table 3). Since the model selected in the stepwise process was a subset of the model selected in the forward process with only one less variable (marital status), we conducted an ANOVA test to compare nested models. Based on the results of the ANOVA test, we concluded that the addition of marital status significantly improved the predictive capability of the model ($F = 4.4441$; $p = 0.035$).

In order to improve residual normality, observations with outlier residual values (> 2.5 after standardization) were removed from the data, and the model was re-fit using the process described above. Models selected using the stepwise and backward processes were the same and included the following variables: 30 day readmit, c-index, number of ER visits in last six months, age, gender, marital status, insurance type, BMI, temperature, heart rate, respiration rate, and mean arterial pressure. Additional variables selected in the forward process include: ICU admission, religion, and oxygen saturation. All models performed similarly by diagnostic criteria (table 3). The additional variables selected in the forward process significantly improved the model by ANOVA testing ($F = 3.9807$; $p < 0.001$). Assumptions for this model (Model 2) were asssesed

by visual inspection of the appropriate plots; residuals associated with model 2 were approximately normal, uncorrelated with the fitted outcome values, and no influential points were identified by Cook's Distance.

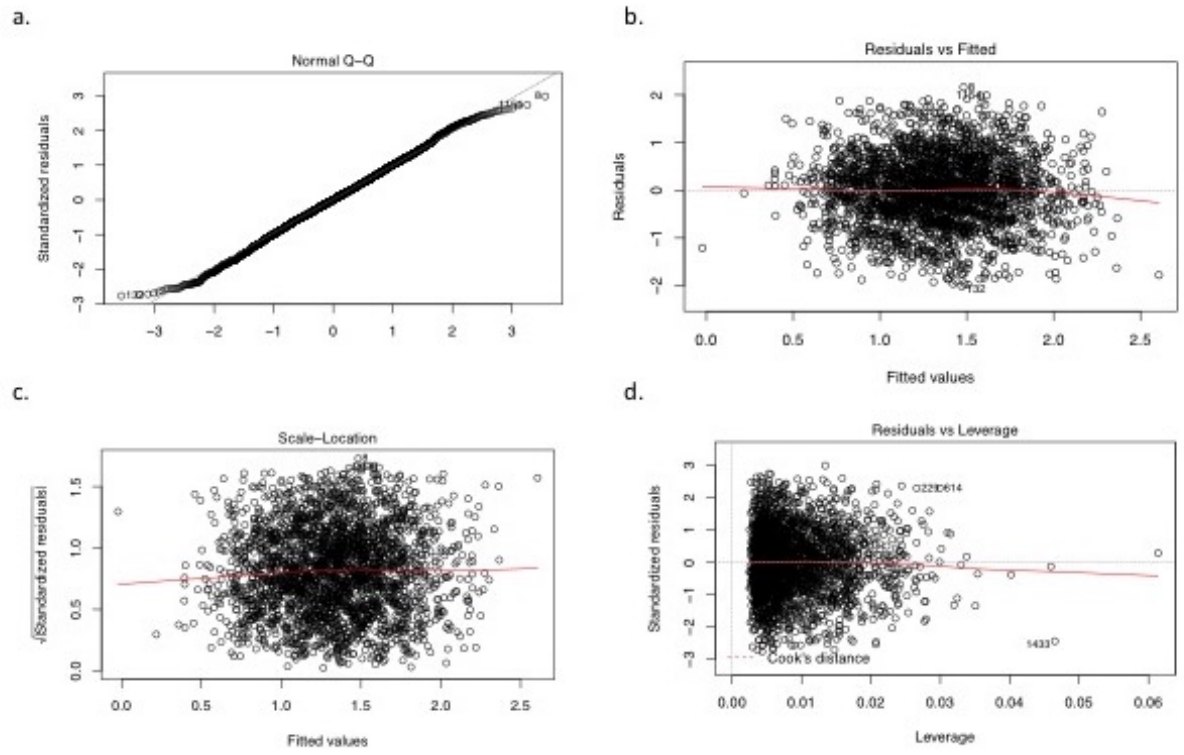

Figure 1: figure 2

Results

In table 4, we present a prediction model for length of stay. The model explains 17.5% of the variability in length of stay. Bootstrapping was used to assess interal validity, and good internal validity was observed.

Because model 1 violates the normal assumptions on the residuals, we decided to remove outliers and refit another model. Overall, the adjusted R-squared increases from ~14% in model 1 to ~17.5% in model 2. Model 2 selects additional covariates when compared with model 1. Further, coefficients of the predictors changed, especially for predictor "temperature_catlow", which changed from -0.09679 to -0.05544. As for direction, all the predictors are consistent before and after.

term	estimate	std.error	statistic	p.value
(Intercept)	2.7982159	0.1975741	14.1628691	0.0000000
is30dayreadmit	0.1821205	0.0425891	4.2762213	0.0000197
cindexmoderate	0.1415274	0.0480058	2.9481330	0.0032249
cindexnormal	0.0183115	0.0351937	0.5203068	0.6028932
cindexsevere	0.1776537	0.0413408	4.2972962	0.0000179
evisit	0.0643723	0.0095514	6.7395532	0.0000000
icu_flag	0.0825013	0.1041881	0.7918493	0.4285193
age	0.0089870	0.0009459	9.5006870	0.0000000
genderMale	0.0611664	0.0294735	2.0752998	0.0380549
religionHindu	-0.0527330	0.0814243	-0.6476320	0.5172790
religionIslam	-0.1790613	0.0844897	-2.1193259	0.0341555
religionJewish	-0.0271905	0.0423371	-0.6422398	0.5207730
religionNo Affiliation	-0.1134603	0.0660716	-1.7172325	0.0860534
religionOther	0.0951416	0.0704993	1.3495394	0.1772789
maritalstatusNot Married	0.0783656	0.0294386	2.6619991	0.0078147
insurancetypeMedicare	-0.1484534	0.0723563	-2.0517004	0.0402967
insurancetypePrivate	-0.1865536	0.0683891	-2.7278261	0.0064173
bmi	-0.0043811	0.0018093	-2.4214269	0.0155263
o2sat_catnormal	-0.1079565	0.0569153	-1.8967933	0.0579637
temperature_catlow	-0.0676050	0.0954025	-0.7086301	0.4786163
temperature_catnormal	-0.2495654	0.0575213	-4.3386630	0.0000149
heartrate_transformed	-2024.8069818	273.6346464	-7.3996733	0.0000000
respirationrate_transformed	-179.4469584	26.9501723	-6.6584717	0.0000000
mean_arterial_pressure	-0.0056365	0.0008823	-6.3881069	0.0000000

Table 4: Prediction model for length of stay.

Discussion

Because outlier data were removed in order to develop the final model, we urge extreme caution in using this model to make predictions. While this model is likely to have some predictive capability for specific types of patients, the model cannot be generalized to patients whose characteristics differ substantially from the patients used in the sample to develop the model.

Additional work will be needed in order to develop models with better predictive capability and increased generalizability. Different modeling methods such as generalized linear models may provide more robust prediction models.

Further, these variables may not be the best variables to predict hospital length of stay. It's possible that there are other variables that need to be explored in order to better understand characteristics that predict length of stay.