p8130 Final Project

Imaani Easthausen, Shanshan Song, Huijuan Zhang, Xinyan Zheng

December 9, 2017

Methods

Data Cleaning

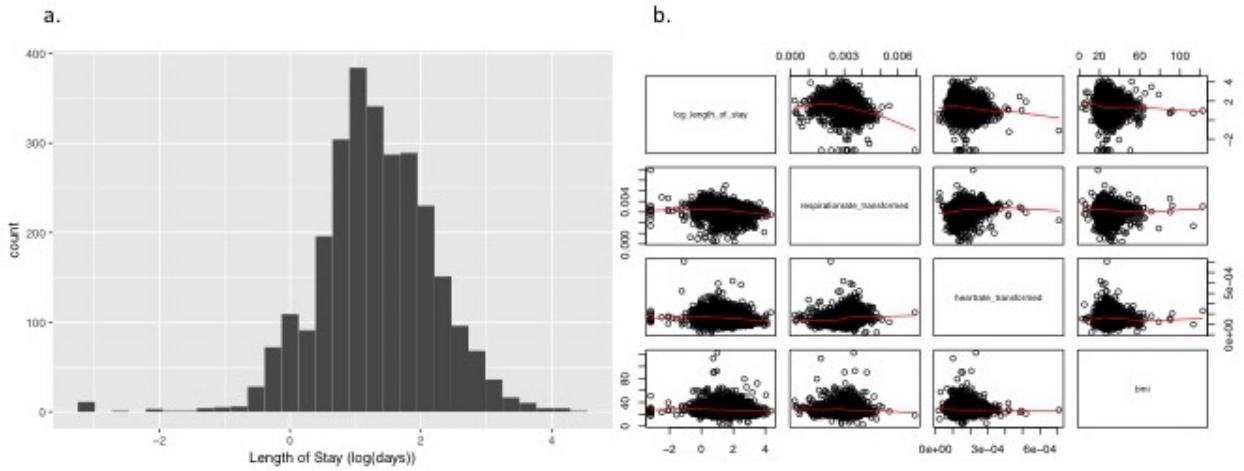
Data cleaning was conducted in R. For subjecta with more than one hospitalization, only the first hospitalization was included in analyses. Subjects that were missing any of the following data were excluded: patient ID, visit ID, length of stay, hospital admission witin the last 30 days, MEW score, C 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



A histogram of LOS was visually inspected and substantial skewness was observed. Log transformation successfully rendered LOS approximatley normal (figure XXX). 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:

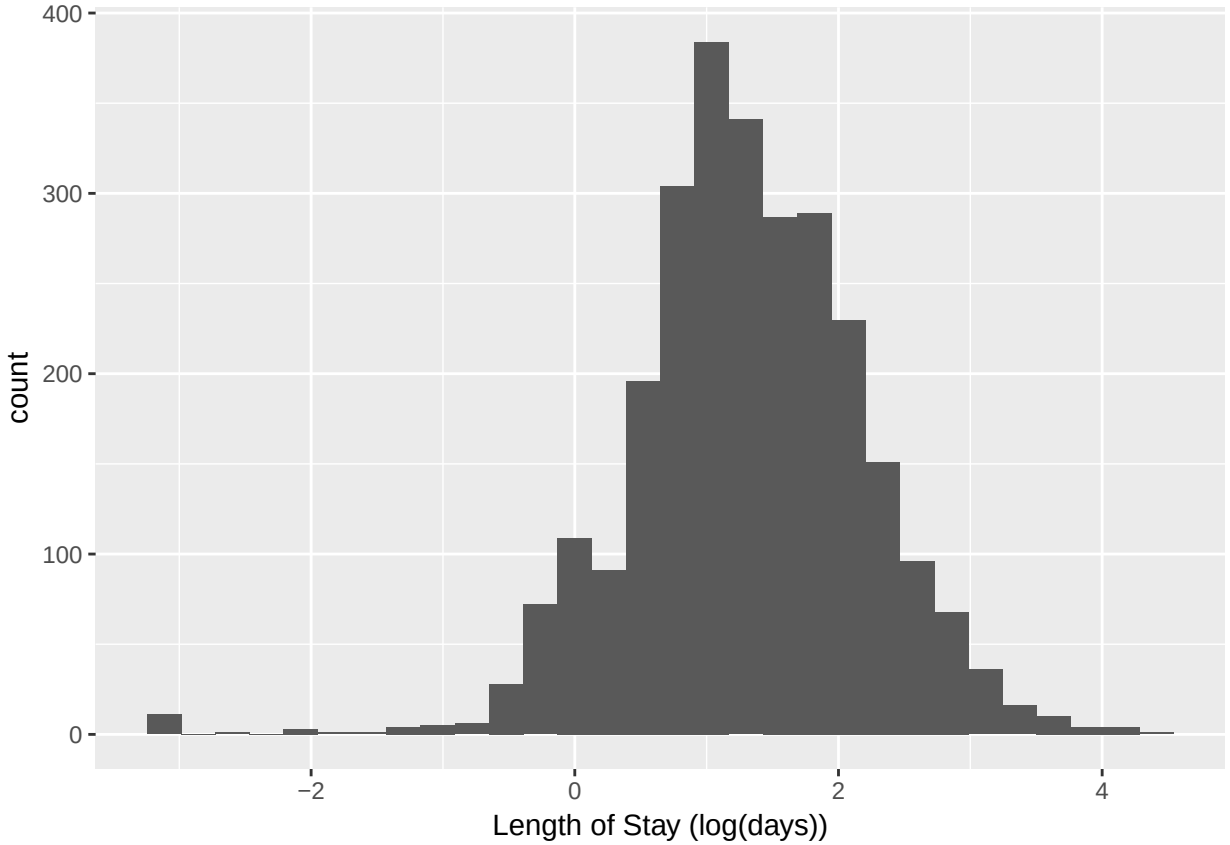
$$T(X) = \frac{1}{X^2}$$

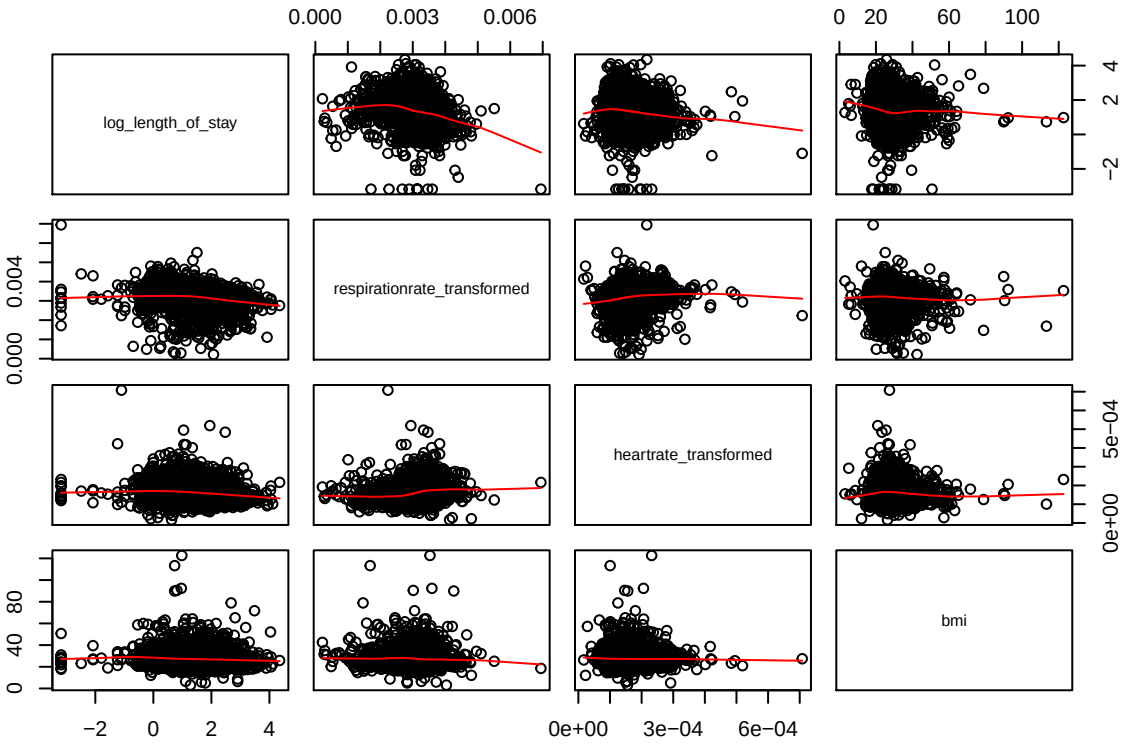
Oxygen saturation and temperature were re-classified into categorical variables as below:

- 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 XXX).

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 XXX).





	age	respirationrate_transformed	mean_arterial_pressure	heartrate_transforme
age	1.0000000	-0.0789071	-0.0584134	0.130830
respirationrate_transformed	-0.0789071	1.0000000	0.0016222	0.178901
mean_arterial_pressure	-0.0584134	0.0016222	1.0000000	-0.065362
heartrate_transformed	0.1308305	0.1789010	-0.0653621	1.000000
bmi	-0.1327040	-0.0783977	0.0900852	-0.051514

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.

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 XXX). Diagnostic criteria for models selected in the forward and stepwise processes were similar to one another (table XXX). 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 XXX). 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.

Results

We present the following model to predict length of hospital stay: $$Y_i = 2.80 + 0.18X_{i1} + 0.18X_{i2} + 0.14X_{i1}