Analysis of Factors Influencing Length of Hospital Stays

Angelina Cottone

2024-12-11

Introduction

Understanding the factors that lead to longer hospital stays is crucial for improving patient outcomes and optimizing resource allocation in healthcare settings. Longer hospital stays can increase the risk of complications and hospital-acquired infections, and place financial strain on facilities and patients alike. Identifying factors that influence LOS can help clinicians enhance quality of care and improve hospital operations.

Several factors, such as patient demographics, comorbidities, and laboratory results can potentially influence the length of hospital stays (LOS). Identifying and understanding the key contributors can provide valuable insights for clinicians and healthcare facilities to enhance efficiency and quality of care for patients. The primary question this research aims to answer is: *What are the key factors that influence the length of hospital stays for patients?* Additional sub-questions have been identified to refine the focus of the research question. These sub-questions include:

- How do comorbidities, such as asthma, iron deficiency, and renal disease, impact the length of stay?
- What is the role of mental health in determining length of stay?
- How do laboratory values, such as hematocrit, neutrophil levels, and blood urea nitrogen, influence length of hospital stays?

By addressing these questions, this study aims to identify what the most influential factors are in predicting LOS to improve patient outcomes and optimize healthcare practices through data driven approaches.

2 Data Acquisition & Processing

2.1 Data Overview

The dataset (`LengthOfStay.csv`) contains 100,000 rows and 28 columns, with information on comorbidities, laboratory results, and vital signs. This datset contains information on a variety of patient characteristics, including demographics, health conditions, laboratory results, and vital signs. Variables irrelevant to this analysis, including date columns, patient IDs, and facility IDs were excluded, leaving 24 variables for analysis.

2.2 Data Preprocessing

A check for missing data confirmed that no there were no missing values in the dataset, so no data-cleaning techniques or imputations were required. An initial summary of the dataset shows a mix of binary, continuous, and categorical variables:

- **Binary**: Indicators for health conditions including `asthma`, `pneum`, and `malnutrition`, as well as mental health indications including `depress` and `psychologicaldisordermajor`.

- **Continuous**: Laboratory results such as `hematocrit`, `neutrophils`, and vital signs such as `pulse`.
- **Categorical**: Demographic information including `gender` and `rcount`.
- **Response variable**: `lengthofstay` is a **discrete count** variable.

Since `lengthofstay` is a discrete count variable, representing the number of days a patient stays in the hospital as an integer, its mean and variance were checked for over dispersion. The variance (5.571) was found to be greater than the mean (4.001), suggesting that a Quasi-Poisson model is more appropriate for this dataset than a standard Poisson model, since it accounts for over dispersion.

2.3 Feature Engineering

Feature engineering was performed for select variables to improve model performance and interpretability. The following changes were made:

- `rcount` (readmission count) was converted into a binary variable: 0 for no readmissions in the past 180 days, and 1 for one or more readmissions.
- `gender` was also converted to binary, with `1` for male and `0` for female.
- Rare categories of `secondarydiagnosisnonicd9` (all categories beside `1`) were combined into "Other" categories to simplify analysis.

The prevalence of outliers in continuous variables was assessed using the interquartile range (IQR) method, where points beyond 1.5 times the IQR from the first and third quartile were identified as potential outliers. Results from the IQR method also found that 50.79% of the the dataset contains values that would be considered outliers. These outliers, given the context of healthcare, likely represent real-world variation (extreme cases and conditions). Therefore, outliers were retained to preserve the natural variation of the dataset and to prevent overfitting.

3 Exploratory Data Analysis (EDA)

3.1 Summary Statistics

The exploratory data analysis (EDA) began by analyzing the key characteristics of the dataset using descriptive statistical summaries, which includes mean, standard deviation, skewness, and more.

Key insights from the summary statistics:

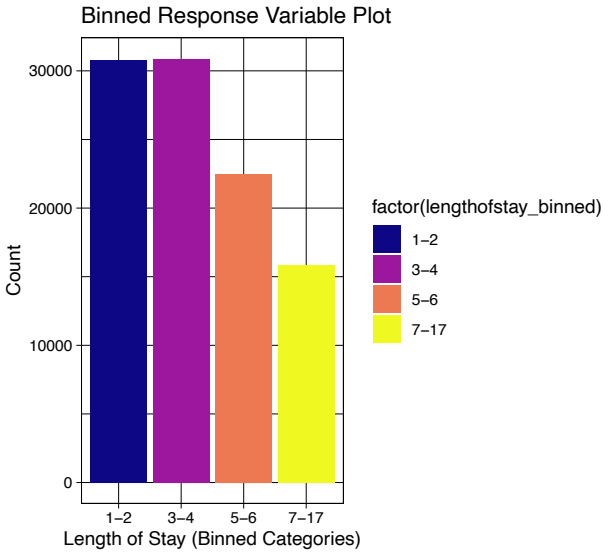
- `neutrophils` and `bloodureanitro` exhibit strong positive skewness, with values greater than 1, suggesting most observations are clustered towards the lower values.
- `hematocrit` and `lengthofstay` exhibit moderate positive skew with values greater than 0.5, suggesting most values are still on the lower side but not as extreme.
- `respiration` has a moderate negative skew with value less than -0.5, suggesting most values are on the higher side.
- `glucose` has the highest standard deviation of all the variables, meaning that data points for this variable display greater variability than those of other variables
- `bloodureanitro` has a biggest range of the variables, at 681.50.

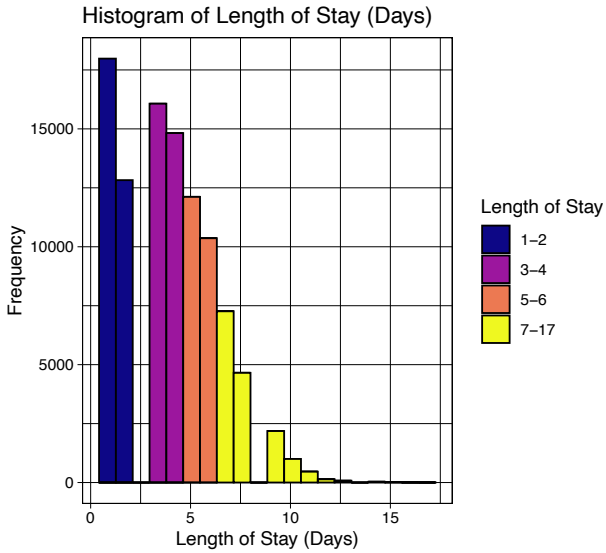
3.2 Response Variable Binning

To facilitate visualization of the response and predictor variables, `lengthofstay` was binned into four intervals.

Since LOS is skewed toward shorter stays, the intervals ("1-2", "3-4", "5-6", and "7-17") were chosen to ensure relative balance of frequency across levels. This allows for clearer visual comparisons of LOS with other variables. By visualizing the distribution of the `lengthofstay` categories, it is evident that the variable is highly imbalanced, with a majority of patients have shorter stay ("1-2" and "3-4" days).

This binning strategy ensures that longer stays, which are less common, are still visible in the plots.

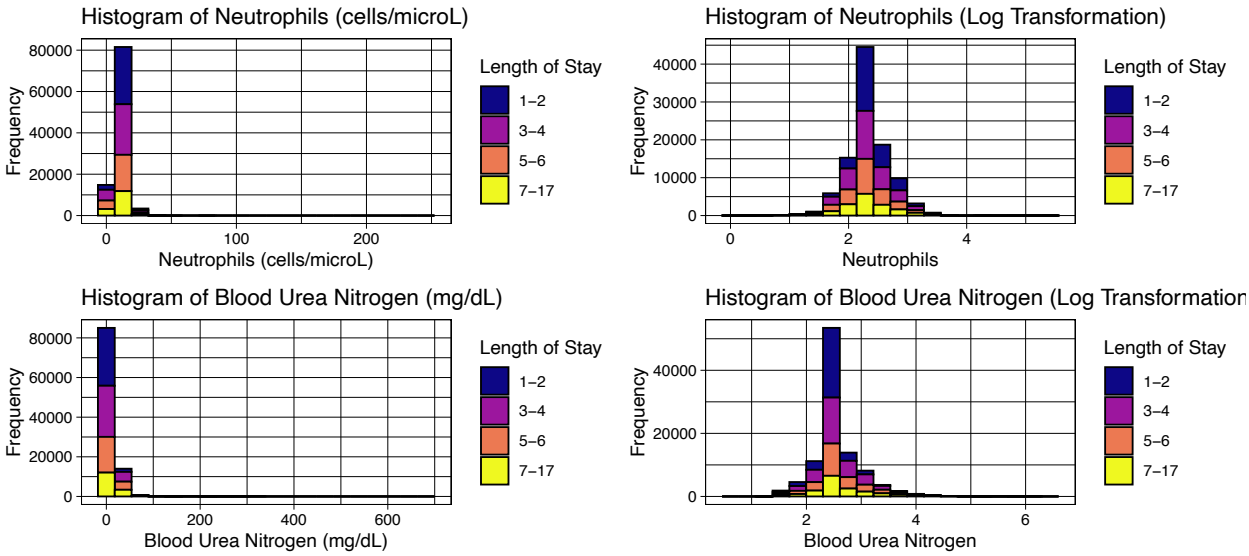




3.3 Visualizing Distributions of Variables

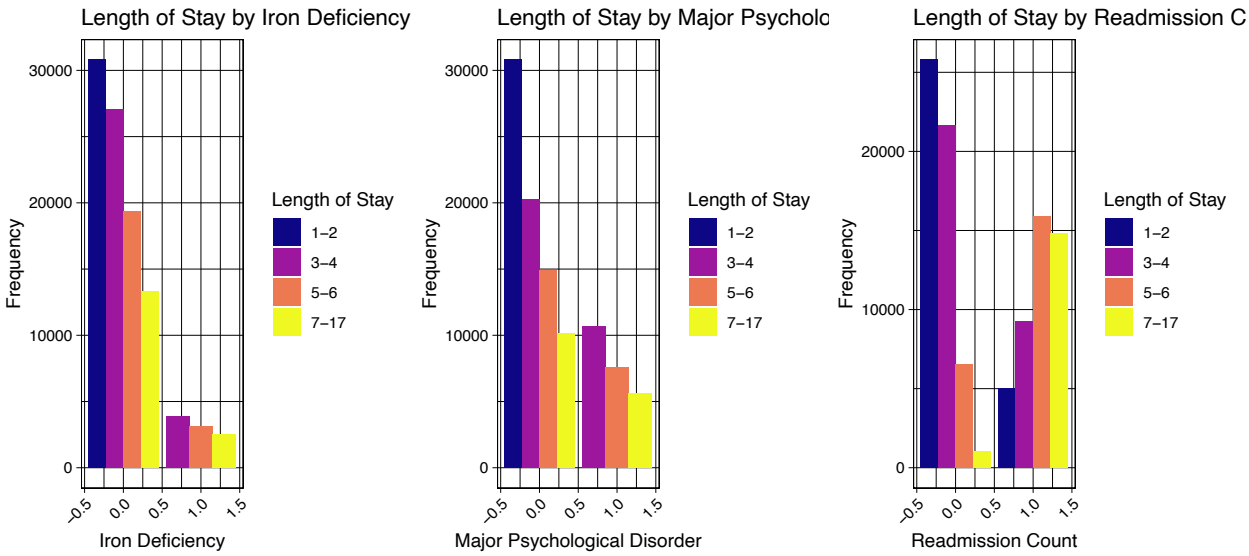
To further assess the distribution of predictor variables in relation to LOS, various visualization methods were used. For continuous variables, histograms were generated, filled by the binned `lengthofstay` variable. These plots confirmed the skewness of certain variables found previously:

- `hematocrit` shows a slight right skew, indicating a higher concentration of lower values.
- `neutrophils` and `bloodureanitro` exhibit more extreme positive skews, with most values being clusters to left of the plot.
- `respiration` exhibits a left skew, with most values lying on the higher end of the plot. To normalize skewed distributions, log transformations were applied to the positively skewed variables (`hematocrit`, `neutrophils`, and `bloodureanitro`), while `respiration` was squared to address its negative skew. All other continuous variables showed a normal distribution of values.



Bar plots were used to visualize the distributions of binary and categorical variables, also filled by the binned `lengthofstay` variable. These plots revealed several trends:

- For all health and mental health indicator (binary) variables, length of stays from 1-2 days were present if the patient did not have that condition, but all stays were three days or longer if the patient did have that condition.
- The bar plot for readmission count also show that length of stays for five days or more are most common in patients that have at least one readmission in the past 180 days compared to those that don't. Short stay lengths (1-4 days) are also less common in those with recent readmissions.



3.4 Principal Component Analysis (PCA)

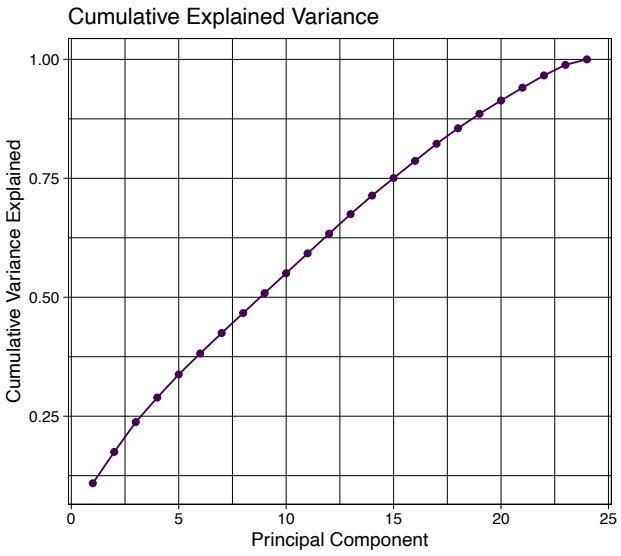
Principal Component Analysis (PCA) was conducted to explore the underlying structure of the numerical variables in the data. Scaling was applied to ensure that the differing units and scales of variables did

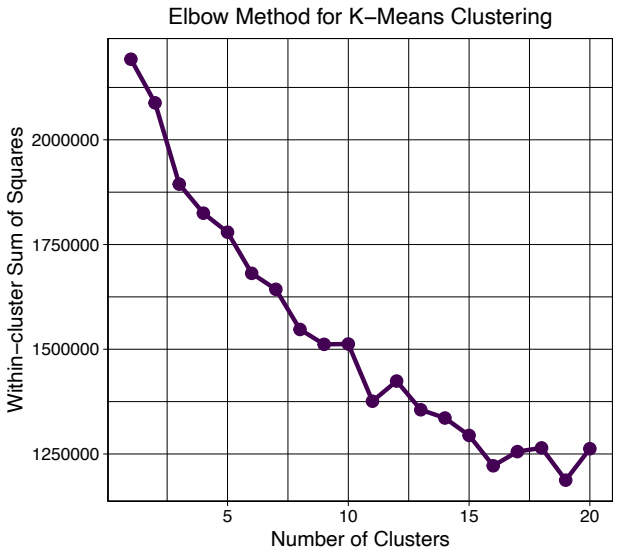
not influence the analysis disproportionately. The goal was to reduce the dimensionality of the data while retaining as much variance as possible. A cumulative variance plot was generated to see how much variance is explained by the first few principal components. The plot did not show a distinct increase or elbow point, indicating that many components would be needed to capture the variance. This suggests that the dataset is quite complex and several components would be needed.

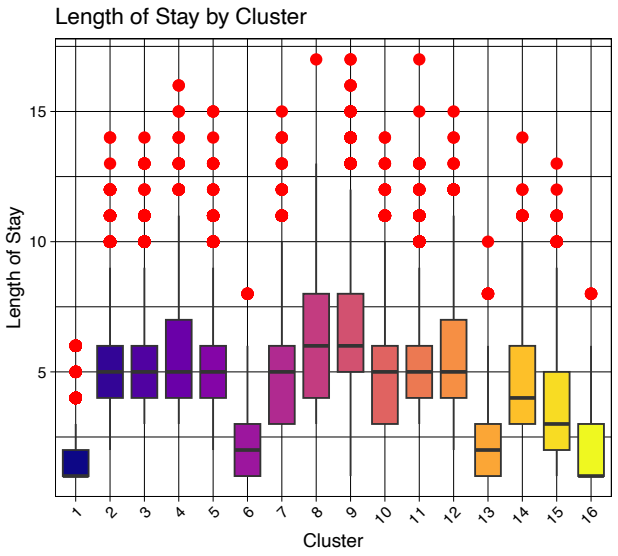
To identify potential clusters in the data, K-means clustering was applied using the first 20 principal components. The plot for the elbow method displayed more than one distinct 'elbow', it was decided that 16 clusters would be an adequate number to capture the most variance possible while still reducing dimensionality. After applying K-means clustering for 16 clusters, the results were visualized using a scatterplot of the first two principal components. The data was colored according to its cluster.

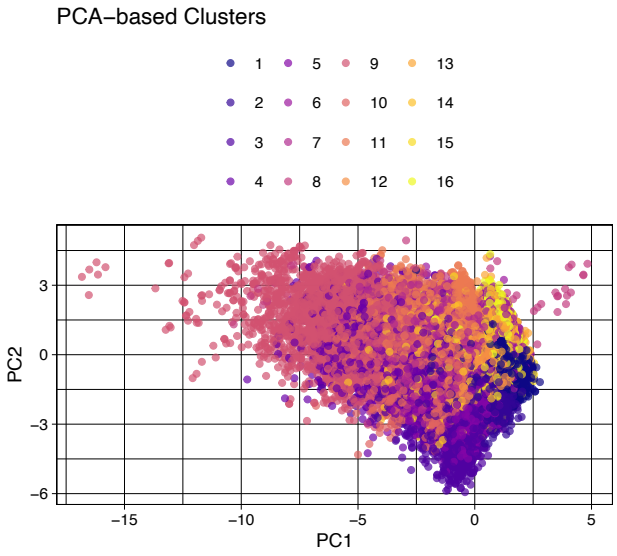
The relationship between the clusters and length of stay was assessed using boxplots, which helped identify which clusters corresponded to shorter or longer stays. Clusters 1 and 16 had the shortest average length of stay (around 1 day), and cluster 8 and 9 had the longest average stay (around 6 days). Clusters 6 and 13 had stays averaging around 2 days, while cluster 15 averaged 3 days and cluster 14 averaged 4 days. All other clusters had averages around 5 days.

To better understand the variation in `lengthofstay` for each cluster, summary statistics were generated. These summary statistics supported what was seen in the boxplot, but also shows that cluster 1 had the lowest average length of stay overall at 1.70 days, while cluster 9 had the highest average at 6.46 days.







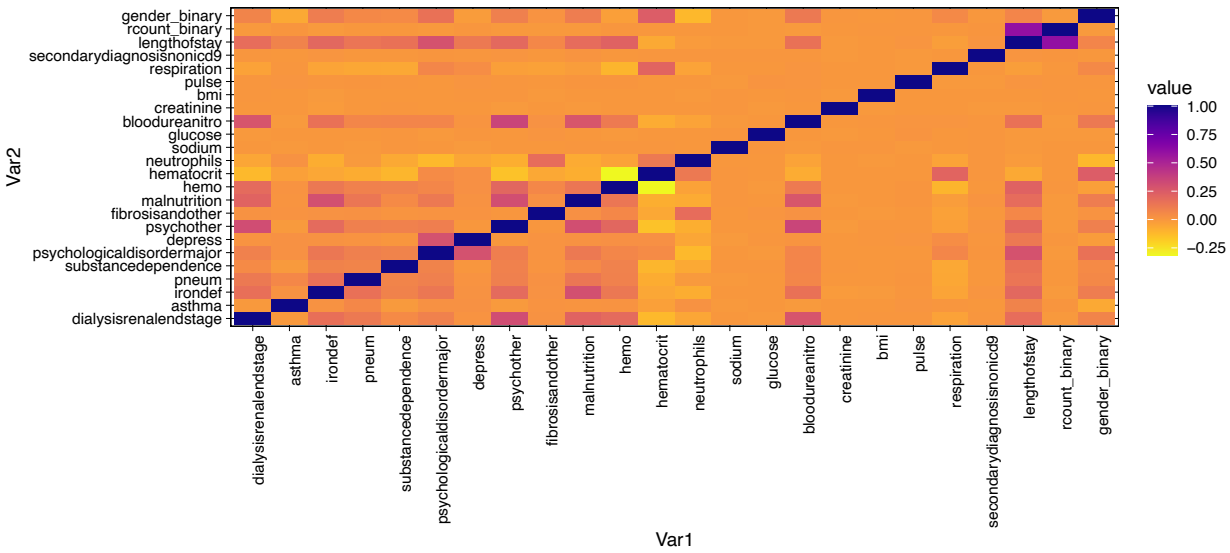


3.5 Correlation Matrix

A correlation matrix was generated for all numeric variables to identify any strong relationships between them. Several positive correlations were identified, including:

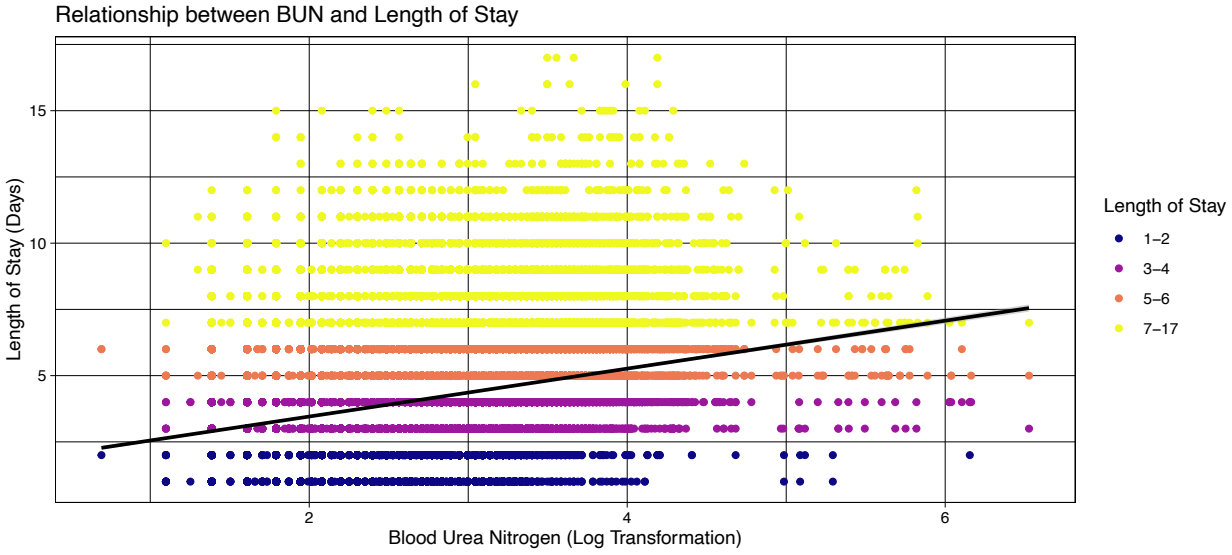
- `psychother`, `dialysisrenalendstage`, `bloodureanitro`, and `malnutrition`
- `malnutrition` and `irondef`, `depress` and `psychologicaldisordermajor`
- `lengthofstay` and `psychologicaldisordermajor`, `rcount_binary` and `lengthofstay`

A negative correlation was also observed between `hematocrit` and `hemo`.



3.6 Scatterplots

Scatterplots were used to assess the linear relationship between continuous predictors and LOS. However, the discrete nature of LOS makes a clear linear relationship difficult to see.



4 Model Selection

4.1 Data Split

To ensure strong model evaluation, the dataset was split into training and testing sets (80-20% split). Stratified sampling was used to ensure the training and testing sets had a proportional distribution of all LOS values.

4.2 Full Model

Model selection began with fitting a full model using a generalized linear model (GLM) with quasi-Poisson distribution, due to the count nature of `lengthofstay` and the overdispersion present. Variables that were changed during feature engineering, such as `rcount` and `secondarydiagnosisnonicd9` were excluded. Transformed variables were included in place of previously identified skewed variables (e.g., `hematocrit`).

The summary for this model indicates that certain variables, such as `sodium`, `bmi`, and `pulse` do not have strong associations with the outcome, with higher p-values suggesting limited predictive power. The dispersion parameter was shown to be 0.697, confirming the presence of overdispersion.

4.2 Lasso Regression

Lasso regression was performed to select a subset of predictors with the most importance to the model. Coefficients that were shrunk to zero, indicating less importance, were removed for future models. The Lasso model removed many of the predictors that were found to not have significance in the full model, with some exceptions (`sodium`, `pulse`). Based on these results, another GLM was fit using the predictors selected by Lasso. This model was expected to have better performance by focusing on the most relevant variables. The summary for the Lasso selected model contained less insignificant variables, although some, such as `sodium` and `pulse` still do not have much significance in the model.

To assess the difference of the two models, a Chi-squared test was performed to compare the deviance of the two models. This indicated whether the reduced Lasso model outperforms the full model in explaining variability in the data. The Chi-squared test results show that the full model does not significantly outperform the Lasso model, with a p-value of 0.8055. Therefore, we fail to reject H_0 (the difference in deviance between the two models is not statistically significant) and conclude that the difference in deviance between the two models is not statistically significant.

4.3 Polynomial Terms

In order to account for the non-linear relationships between the predictors and `lengthofstay`, polynomial terms were created for continuous variables in the model. These terms allow the model to capture nonlinear relationships which are likely present in the data. These variables were also centered to improve interpretation and mitigate multicollinearity.

A new generalized linear model (GLM) was fit to predict the length of hospital stay with the polynomial terms. The dispersion parameter, which is now .615, was reduced when compared to the Lasso-selected model. A Chi-squared test was also performed to compare the polynomial and Lasso-selected model. The p-value in this case is very small, so we can reject H_0 (the difference in deviance between the two models is not statistically significant) and conclude that the model with polynomial terms results in a better fit.

The two models were also compared based on their predictive power using cross-validation training. The model without polynomial terms had an root mean squared error (RMSE) of 1.70 units, and an R^2 of 0.4873. The RMSE for the model with polynomial terms was 1.63 units, with an R^2 of 0.5293, showing an improvement in both for the polynomial model.

4.4 Interaction Terms

To investigate the influence of interaction terms, another model incorporating interaction terms was fit. This model included both main effects and interaction terms between predictors that were previously identified to be associated. Another Chi-squared test was performed between the model with interaction terms added and the model without interaction terms. This test yielded similar results, with a very small p-value, allowing us to reject H_0 (the difference in deviance between the two models is not statistically significant) and conclude that the model with interaction terms results in a better fit.

The predictive accuracy of the model with interaction terms was also assessed with cross-validation training, which resulted in an RMSE of 1.63 and an R^2 of 0.5316. Compared to the model without interaction terms, there was no change in RMSE but a slight increase in R^2, suggesting slightly more variation in the data can be explained with the addition of interactions.

4.5 Final Model

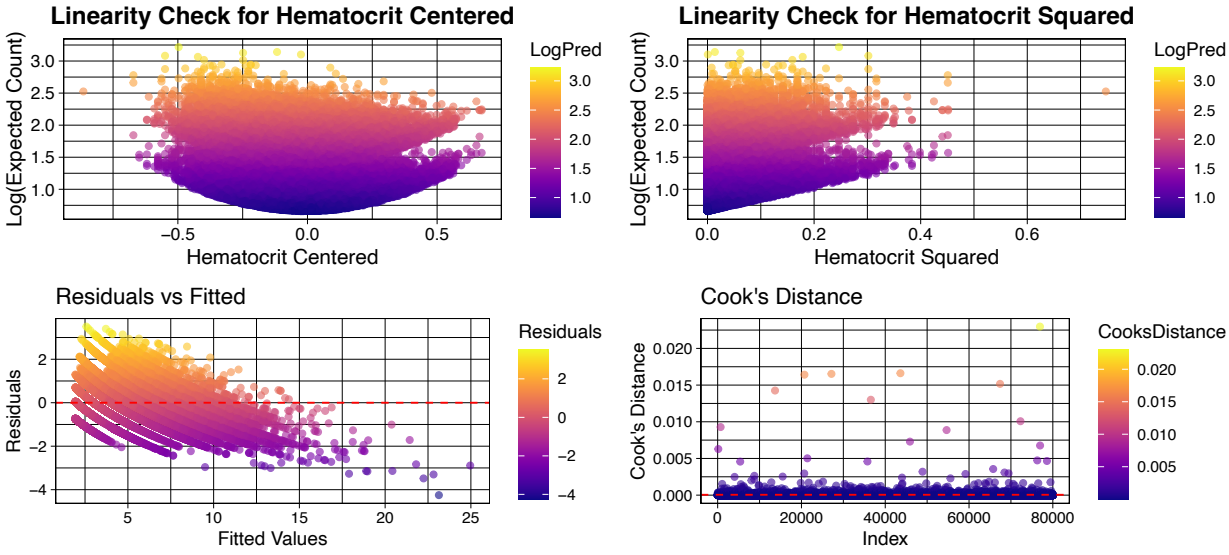
The final model was fit by removing insignificant terms and terms causing multicollinearity. The final model was once agained trained, yielding an RMSE of 1.63 and an R^2 of 0.5318, showing that the final model explains around 53.18% of the variance in the data.

In the total, the model contains 29 predictor coefficients. A subset of the equation for this model can be written as: $lengthofstay = 0.6635 + 0.1406 * \text{dialysisrenalendstage} + 0.1893 * \text{asthma} + 0.1769 * \text{irondef} + 0.1123 * \text{pneum} + 0.1713 * \text{substancedependence} + 0.2882 * \text{psychologicaldisordermajor} + 0.2615 * \text{depress} + 0.1604 * \text{psychother} + 0.1454 * \text{malnutrition} + 0.1963 * \text{hemo} + ...$

Diagnostics were performed on the final model to assess any violations for quasi-Poisson:

- **Durbin-Watson Test for independence**: Confirmed no significant autocorrelation in the residuals, with test statistics close to 2 and p-value of 0.589.

- **Linearity**: Scatterplots of predictors versus log-transformed fitted values show curved relationships with centered variables and linear relationships with polynomial terms.
- **Multicollinearity**: Variance inflation factors (VIFs) showed no significant multicollinearity, but the condition number (5465.351) indicated some multicollinearity is still present.
- **Constant variance**: Residuals vs. fitted values plot showed residuals had a pattern and were not randomly scattered around 0, suggesting a violation of this assumption.
- **Deviance and deviance ratio**: To assess the overall fit of the model, deviance and deviance ratio were calculated. The deviance of the model came out to 48823.16, with a deviance ratio of 0.6105, suggesting a relatively good fit to the data.
- **Influential points**: Cook's Distance was used to assess the presence of influential points.

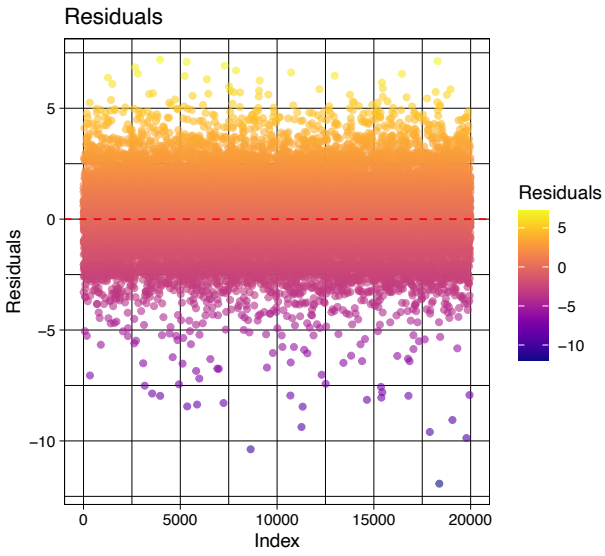


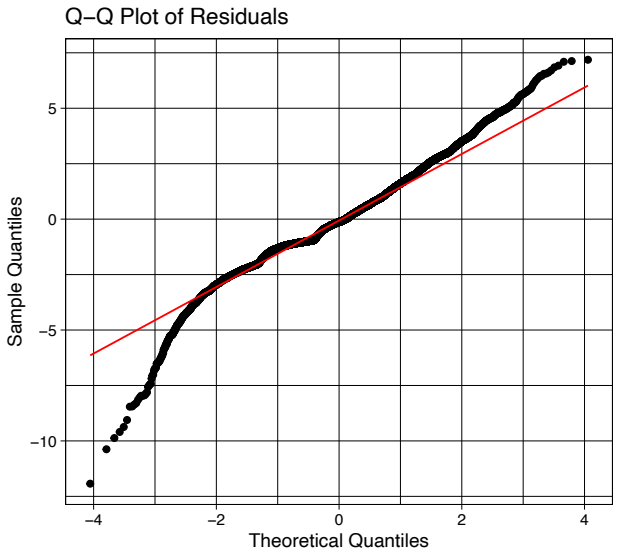
4.5 Prediction Analysis

Confidence intervals were created for the coefficients of the final model. The intervals for `hematocrit_centered` and `pulse_centered` contain 0, suggesting that these variables may have no significant effect on the model.

The average prediction error for this model is -0.004838201, indicating that predictions, on average, are close to the actual values. The negative value suggests that the model tends to slightly overestimate the length of stay for patients. In a healthcare setting, it is often preferable to over-predict rather than under-predict, as it allows facilities to better better prepared for resource allocation and patient care.

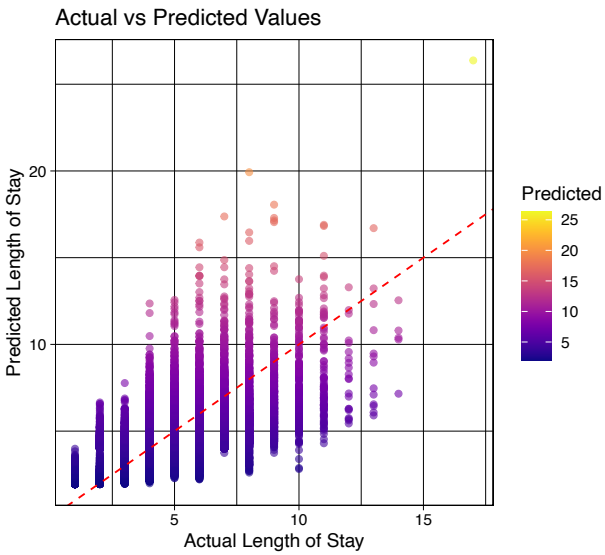
The residuals for the predictions appear to be clustered and evenly distributed around zero, but dispersion is seen as points move away from the line, indicating non-constant variance. This likely indicates that the model's accuracy varies across different ranges of predicted values. The QQ-plot shows heavy tails, suggesting a deviation from normality.

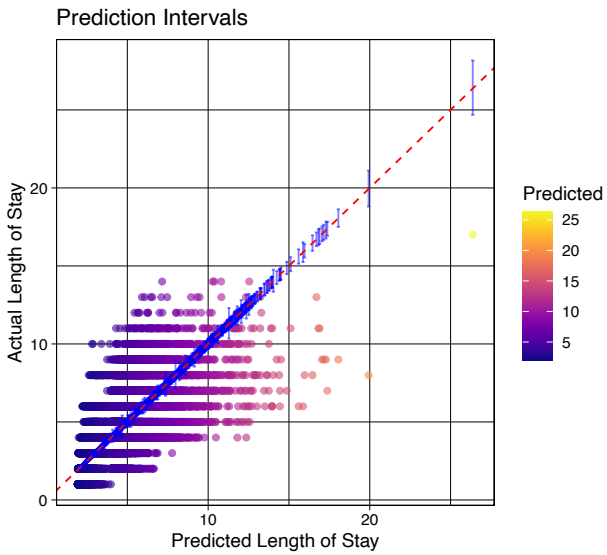




The standard errors of the predictions were calculated and used to create 95% prediction intervals. Plots were also created to visualize the action versus predicted values and prediction intervals. These plots reveal two outliers with predicted length of stays over 20 days, which are beyond the range of the data. This indicates some inaccuracy as the model struggles to predict values at extreme ends of the distribution.

```
##    Predicted Lower_Pred Upper_Pred
## 1   2.391003   2.368766   2.413449
## 5   3.234944   3.185682   3.284968
## 6   4.763576   4.724182   4.803299
## 14  3.605989   3.547513   3.665428
## 15  4.051727   4.028992   4.074589
## 20  2.105840   2.088332   2.123495
```





Conclusion

From our final model it is evident that all of the comorbidities and mental health indicators included in the dataset, with the exception of `fibrosisandother`, have significant effects on length of stay, with them all being positive. This indicates that the presence of health and mental health conditions increase the lengths of stay for patients. The laboratory values of `hematocrit`, `neutrophils`, `bloodureanitro`, `sodium`, and the vital sign `pulse` also have positive effects on length of stay, with higher values increasing the length of stay.

Overall, the complexity of the dataset posed a challenge for finding an adequate model for lengths of stays, but this was eventually achieved through generalized linear models with quasi-Poisson distributions. The insights gained from this analysis for a strong foundation for future refinements and models.

These results contribute to the broader goal of improving patient outcomes and optimizing healthcare resources by better understanding the factors influencing LOS.

R Appendix

```
knitr::opts_chunk$set(echo = FALSE, message = FALSE, warning = FALSE, fig.align = "center", fig.widt
options(scipen = 999)
# 2.1 Data Overview
# Libraries
library(psych)
library(gridExtra)
library(ggplot2)
library(MASS)
library(car)
library(caret)
library(lmtest)
library(glmnet)
library(corrplot)
library(reshape2)
library(GGally)
library(viridis)
library(dplyr)

setwd("~/Desktop/sta 141a")
hospital_data <- read.csv("LengthOfStay.csv")
hospital <- hospital_data[, c("rcount", "gender", "dialysisrenalendstage",
                              "asthma", "irondef", "pneum",
                              "substancedependence",
                              "psychologicaldisordermajor",
                              "depress", "psychother", "fibrosisandother",
                              "malnutrition", "hemo", "hematocrit",
                              "neutrophils", "sodium", "glucose",
                              "bloodureanitro", "creatinine", "bmi", "pulse",
                              "respiration", "secondarydiagnosisnonicd9",
                              "lengthofstay")]
# 2.2 Data Preprocessing
# Summarize data structure
summary(hospital)
str(hospital)

# Check for missing values
```

See PredictingLengthofStay.Rmd file for full code