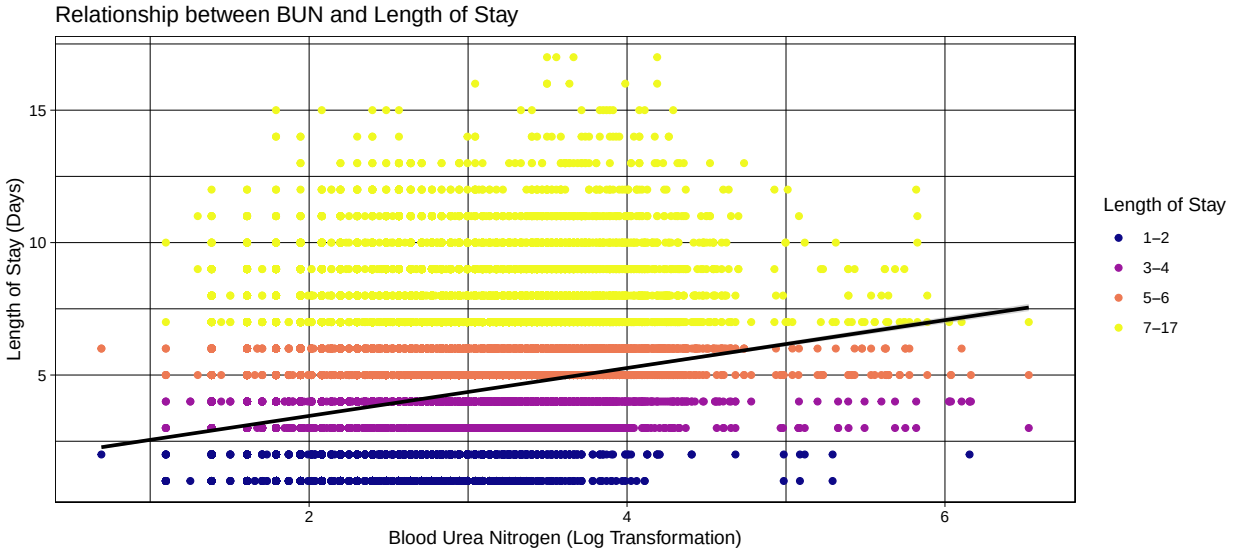
3.6 Scatterplots

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

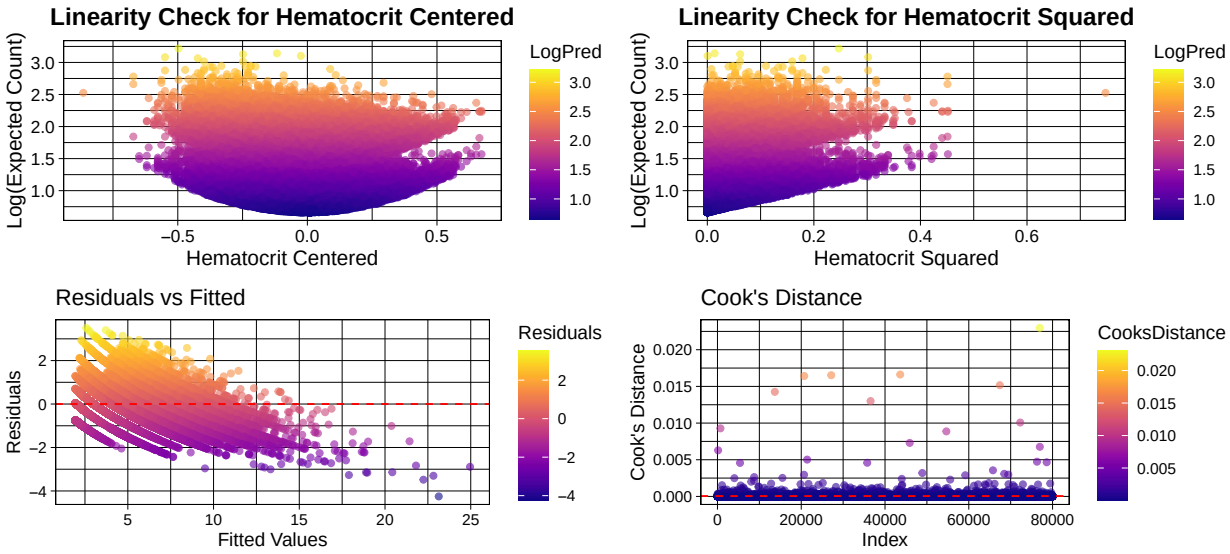
The final model was fit by removing insignificant terms and terms causing multicollinearity. The final model was once agained trained, yielding an RMSE of 1.63 and an R^2 of 0.5318, showing that the final model explains around 53.18% of the variance in the data.

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.

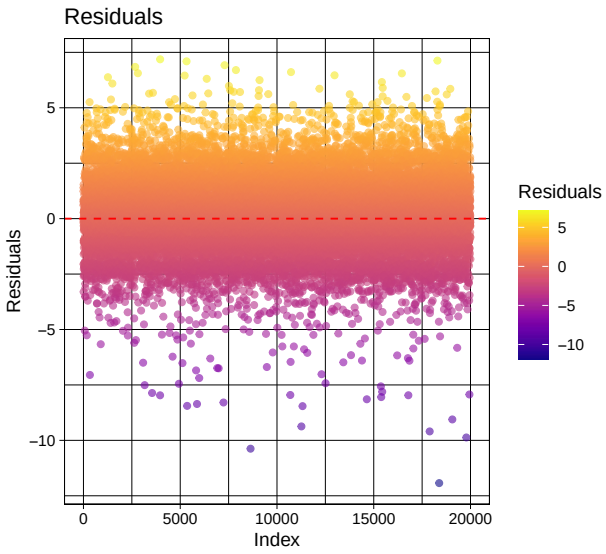


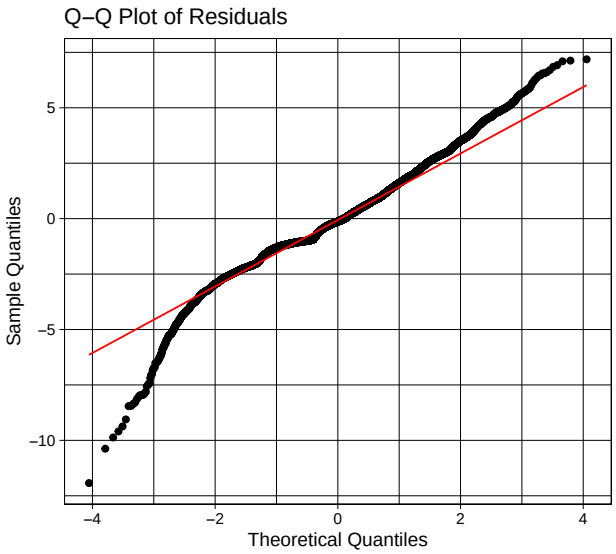
4.5 Prediction Analysis

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.

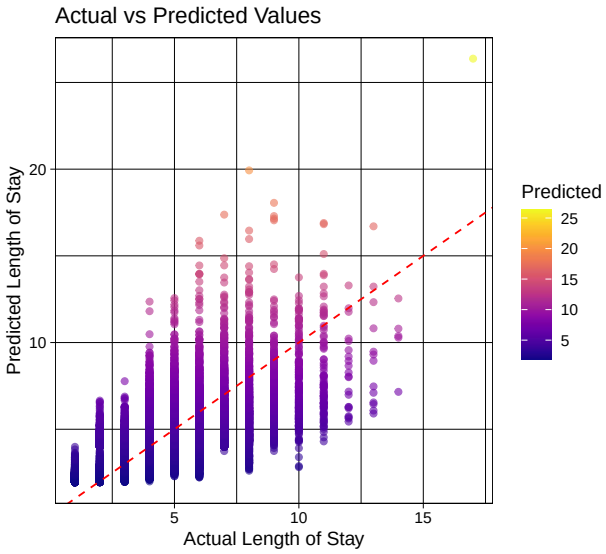
The residuals for the predictions appear to be clustered and evenly distributed around zero, but dispersion is seen as points move away from the line, indicating non-constant variance. This likely indicates that the model's accuracy varies across different ranges of predicted values. The QQ-plot shows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

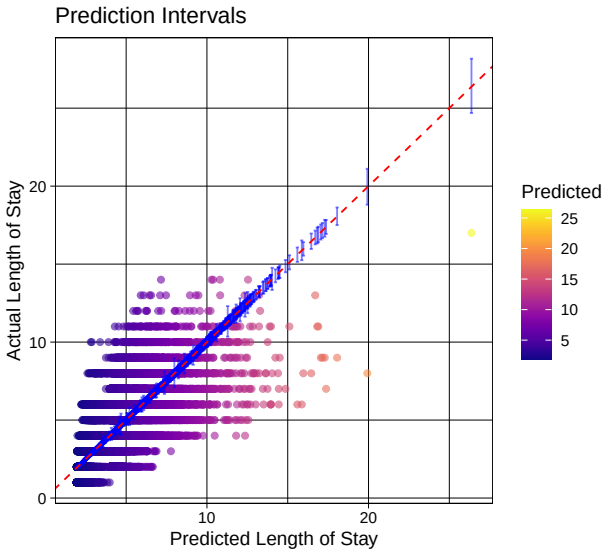




The standard errors of the predictions were calculated and used to create 95% prediction intervals. Plots were also created to visualize the action versus predicted values and prediction intervals. These plots reveal two outliers with predicted length of stays over 20 days, which are beyond the range 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.width
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
```