Project 2: STA 108
Fall 2024

Submitted To: Amy Kim, Dongchan Kim

Darren Lam

Contents

I. Introduction

The objective of this project is to create an effective multiple linear regression model that **correctly** maps the average Length of Stay of patients in hospitals (in days) using various explanatory variables from the SENIC2 dataset (8 numeric and 2 categorical). The SENIC2 dataset contains data from 113 hospitals surveyed during the 1975-76 study period. We will use multiple linear regression techniques to understand the relationship between the Length of Stay and selected predictor variables which could have real world implications for healthcare providers or patients.

II. Exploratory Data Analysis

According to the summary statistics for each numeric variable, (Table 1) patients tended to stay in the hospital for approximately 10 days and were generally on the older side, having an average age of 53 years. The hospitals also had fewer average nurses than beds, suggesting that a single nurse likely had to treat multiple patients. Looking at the histogram for Length of Stay, (Figure 1) the distribution of Length of Stay is skewed towards the right, indicating most hospitals have a shorter average stay. There also seems to be an uneven distribution of hospital Medical School Affiliation and Geographic Region. (Figure 2, 3) The grouped scatter plots indicate a generally weak positive relationship between the Length of Stay and each individual numeric variable. (Figure 4) Patients in hospitals with Medical School Affiliation tend to stay longer than patients in those without and hospitals in the NE Geographic Region tend to yield the longest patient stay. (Figure 5) Using these distributions, we can now combine multiple variables for the best mapping of patient Length of Stay.

III. Model Selection

To develop the model, we first assessed possible multicollinearity between explanatory variables using both Variance Inflation Factors (VIF) (Table 2) and a correlation matrix. (Figure 6) We found that the variables: Number of Beds (X5), Daily Census (X8), and Number of Nurses (X9) had high probabilities of being correlated to other explanatory variables so we decided to further investigate. According to our correlation matrix, X5 and X8 had very high correlation so one of them needed to be removed from our model. We trained two models on the full dataset with each variable (X5 and X8) removed respectively and found that removing X5 decreased BIC (Table 3) and increased Adjusted R^2 more than removing X8, therefore we removed X5. X9 also had high correlation with X8 so we removed that variable as well. With our final set of possible predictors that did not include X5 and X9, we used forward-backward selection (to minimize overfitting) using BIC to identify the optimal predictors for our correct model. We chose to include: Patient Age (X1), Infection Risk (X2), Geographic Region (X7), and Daily Census (X8) in our final model (Y ~ X1 + X2 + X7 + X8). The final model before removing outliers/influential points was:
$Y = 3.512 + 0.080X_1 + 0.509X_2 - 0.967X_{7NC} - 1.296X_{7S} - 2.270X_{7W} + 0.004X_8$ (Table 4), yielding a BIC of 408.8766 (Table 3).

IV. Model Diagnostics

When creating our model, we must fulfill a set of assumptions that are standard for conducting regressions and allow us to ensure that our model is reliable. We tested the same set of assumptions for the model we trained before and after removing outliers, except for the assumption that the model predictors are independent, which we only need to test once because we use the same predictors for both models.

Assumption: Predictors are Independent (Satisfied)

In the previous model selection section, we checked for collinearity between each explanatory variable and removed the two highly correlated variables from our subset before performing model selection. (Table 2, Figure 6) Therefore, we can conclude that the assumption for independent predictors is satisfied.

Pre-Outlier Removal

Assumption 1: Linear Relationship Between Variables (Satisfied)

According to Figure 7, the fitted values show a moderately strong linear relationship with the Length of Stay, with a few outliers. This satisfies the assumption for a linear relationship between the response variable and explanatory variables for the model before removing outliers.

Assumption 2: Errors are Normally Distributed and Mean = 0 (Violated)

To check the assumption for normality in errors, we created a histogram and QQ plot of the model residuals. (Figure 8, 9) The errors do not seem especially normally distributed, especially given the multiple outliers displayed in both plots. According to the histogram, the residuals did not seem to be centered around a mean of 0. We also checked this assumption using the Shapiro-Wilks Test. (Table 5) This test checks the H_o: errors are normally distributed and H_a: errors are not normally distributed. We obtained a test statistic of 0.8791 and a p-value of almost 0, suggesting that we would reject our null hypothesis and conclude that the errors are not normally distributed.

Assumption 3: Errors have Constant Variance and are Uncorrelated (Satisfied)

To check the assumption for constant variance in errors, we assessed a scatter plot of our fitted values plotted against the residuals. (Figure 10) Based on this graph, the errors seemed to have constant variance and seemed to be uncorrelated (no pattern) with a few outliers. We double-checked this assumption using the Fligner-Killeen Test, which checks the H_o: errors have constant variance and H_a: errors do not have constant variance. (Table 5) We obtained a test statistic of 0.66194 and a p-value of 0.4159, suggesting that we would fail to reject our null hypothesis and conclude that the errors have constant variance.

Because the diagnostic graphs for the model before removing outliers showed that there were multiple outliers that could affect the effectiveness of the model, and multiple assumptions were violated, we next looked at possible points to remove to address these concerns.

Outliers

When determining possible outliers, we examined the scatter plot of our fitted values plotted against residuals. (Figure 10) We decided that most of the data was between an absolute residual value of 2.5 and used that as our threshold to determine possible outliers. We found 5 possible outliers that exceeded our outlier threshold and listed them in a table. (Table 6)

High Leverage Points

To determine possible high-leverage points, we calculated the hat values for each data point and created a hat matrix. We then compared these values to a high-leverage threshold of 0.125 (2p / n, p = 7, n = 112) and plotted them on a high-leverage plot. (Figure 11) We found 6 possible high-leverage points that exceeded our high-leverage threshold and listed them in a table. (Table 7)

Influential Points

To determine the influential points in our data, we used the Cooks Distance for each point and created a vector of these values. We then compared these values to a threshold of 0.0625 (p / n) and plotted them on an influential point plot. (Figure 12) We found 4 possible influential points that exceeded our threshold and listed them in a table. (Table 8) We removed these values and refitted the model on the cleaned data set for our final model (Table 9).

Post-Outlier Removal (Final Model)

Assumption 1: Linear Relationship Between Variables (Satisfied)

According to Figure 13, the fitted values show a strong linear relationship with the Length of Stay. This satisfies the assumption for a linear relationship between the response variable and explanatory variables for the model before removing outliers.

Assumption 2: Errors are Normally Distributed and Mean = 0 (Satisfied)

To check the assumption for normality in errors after removing influential points, we created a histogram and QQ plot of the model residuals. (Figure 14, 15) The errors seem more normally distributed now. According to the histogram, the residuals now seem to be centered around a mean of 0. We double this assumption using the Shapiro-Wilks Test. (Table 10) Again we checked the H_o: errors are normally distributed and H_a: errors are not normally distributed. We obtained a test statistic of 0.99476 and a p-value of almost 0.9576, suggesting that we would fail to reject our null hypothesis and conclude that the errors are normally distributed.

Assumption 3: Errors have Constant Variance and are Uncorrelated (Satisfied)

To check the assumption for constant variance in errors, we assessed a scatter plot of our fitted values plotted against the residuals. (Figure 16) Based on this graph, the errors seemed to have more constant variance and were uncorrelated (no pattern). We double-checked this assumption using the Fligner-Killeen Test, again checking the H_o: errors have constant variance and H_a: errors do not have constant variance. (Table 10) We obtained a test statistic of 0.20333 and a p-value of

0.652, suggesting that we would fail to reject our null hypothesis and strengthening our conclusion that the errors have constant variance.

Having removed our outliers, we observed a large decrease in BIC (Table 3) and satisfied each of our regression model assumptions allowing us to ensure the reliability of our linear regression model and move on to analysis and interpretations.

V. Analysis and Interpretation

Now that we have cleaned our data we have accomplished our **main goal** of training a "correct" model that effectively maps the patient's Length of Stay at a hospital. The **final model** we obtained was: (Y ~ X1 + X2 + X7 + X8) $Y = 4.016 + 0.067X_1 + 0.512X_2 - 0.542X_{7NC} - 0.960X_{7S} - 2.123X_{7W} + 0.002X_8$ (Table 9), yielding a **BIC** of 319.9867 (Table 3). To accomplish this, we focused on decreasing BIC. We observed this decrease in BIC because we eliminated multicollinear variables as well as outliers to satisfy model assumptions and train the most optimal model. The **B_0** predictor value indicates that we estimate that a patient that is 0 years old, has an infection risk of 0, at a hospital in the NE region with a daily census of 0 stays at the hospital for 4.016 days, however this is obviously uninterpretable. **B_1** indicates that we estimate that when patient Age increases by 1, we expect the average patient Length of Stay to increase by 0.067 days, holding all other variables constant. **B_2** indicates that we estimate that when Infection Risk increases by 1%, we expect the average patient Length of Stay to increase by 0.512 days, holding all other variables constant. **B_3 , B_4** , and **B_5** indicate that a hospital in the NC, S, and W region respectively, has an average patient Length of Stay 0.542, 0.96, and 2.123 days lower than a patient at a hospital in the NE region, holding all other variables constant. Finally, **B_6** indicates that we estimate that when the Daily Census increases by 1, we expect the average patient Length of Stay to increase by 0.002, holding all other variables constant. We found that the final model had an **adjusted R_2** of 0.6027 which suggests that the model explains 60.27% of the variation in the Length of Patient Stay. Having obtained this model, we were able to calculate simultaneous **confidence intervals** for each predictor using the Bonferroni multiplier and conclude with 95% confidence that the true effect of each predictor on a patient's Length of Stay within the hospital is between each upper and lower bound. (Table 11) **General Linear Test (F):** We also tested whether or not the reduced model was a better fit for the data than the full model using H_o: B3 = B4 = B5 = B6(Yes) = B9 = B10 = 0 (Reduced Model Fits Better) and H_a: At least 1 $B_i \neq 0$, i = (3, 4, 5, 6, 9, 10) (Full Model Fits Better). We obtained an F-statistic of 1.0187 and a p-value of 0.418, (Table 10) leading to the conclusion that we fail to reject the null hypothesis and conclude that the reduced model fits the data better than the full model. **Slope Test (T):** According to the summary table (Table 9), under the H_o: $B_i = 0$ and H_a: $B_i \neq 0$, the given T-statistics for each predictor yield p-values that allow us to conclude that each value is an important predictor for patient Length of Stay. This also shows us that Infection Risk (X2) is the most significant predictor of Length of Stay because it has the lowest p-value.

VI. Conclusion

The final model provides a **correct** mapping of the average patient Length of Stay in hospital. With our assumptions for regression satisfied, our analysis shows that average patient Age, Infection Risk, hospital Geographic Region, and Daily Census are significant predictors of a patient's Length of Stay. This notion is supported by our test-statistics and p-values, which indicate a significant linear relationship. Despite this, our somewhat low adjusted R^2 value suggests that there may be other factors outside of the data set that need to be taken into account when mapping a patient's Length of Stay in Hospital. This leads us to draw an important conclusion about the flawed nature of our model: that given the limited number of patients in our sample, it is difficult to draw conclusions about a patient's Length of Stay that would be applicable to real world scenarios. The simple nature of our measured response variable being patient "Length of Stay" does not account for factors such as the type of infection or severity of infection thus leading to difficulty in understanding factors that contribute to a patient's length of stay in hospitals. A possible solution of this would be to collect more data to further fine tune the model and increase mapping accuracy. Since the dataset we trained the model on was a subset of the entire SENIC dataset, using the entire SENIC dataset would be appropriate in increasing our model reliability and usability.

Appendix: Plots

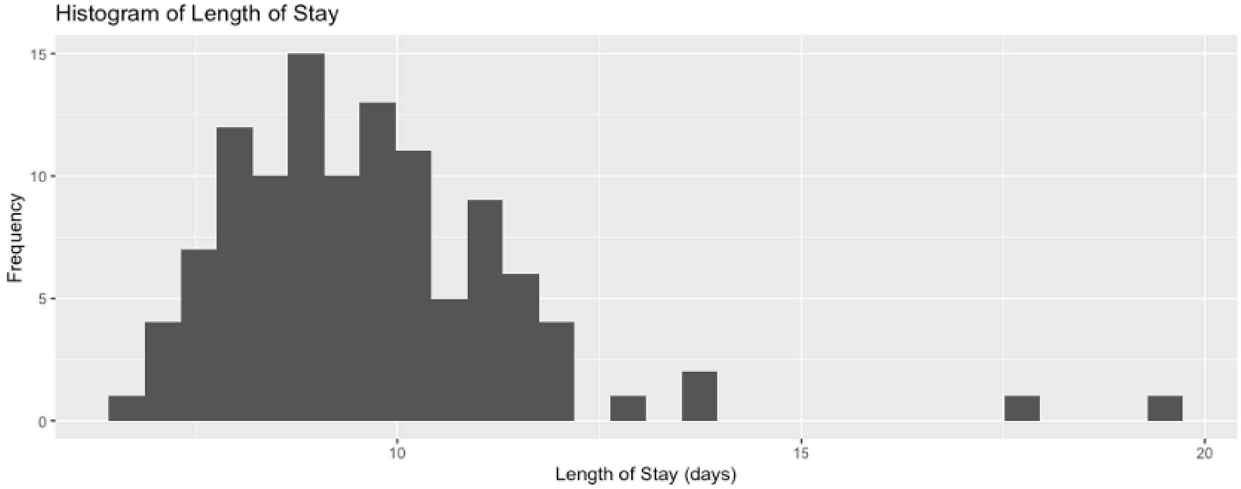


Figure 1: Histogram of Average Patient Length of Stay in Hospital (days)

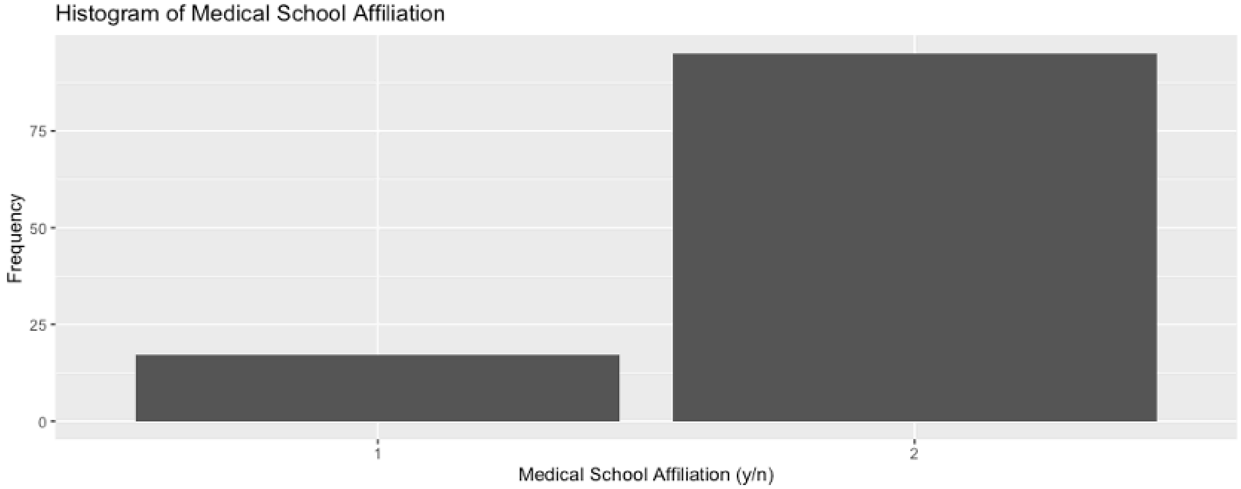


Figure 2: Histogram of Hospital Medical School Affiliation (1 = Yes, 2 = No)

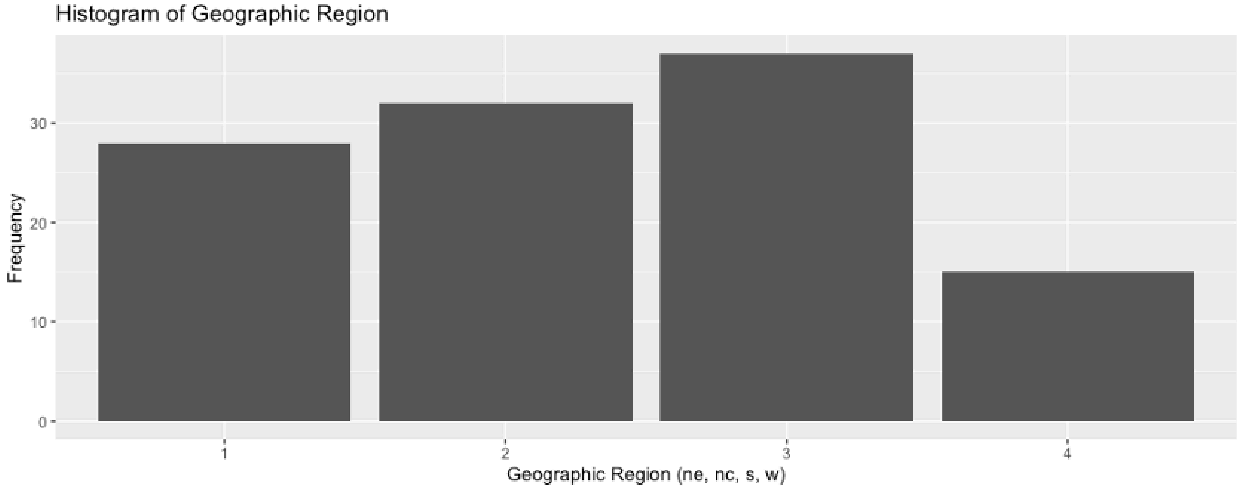


Figure 3: Histogram of Hospital Geographic Region (1 = NE, 2 = NC, 3 = S, 4 = W)

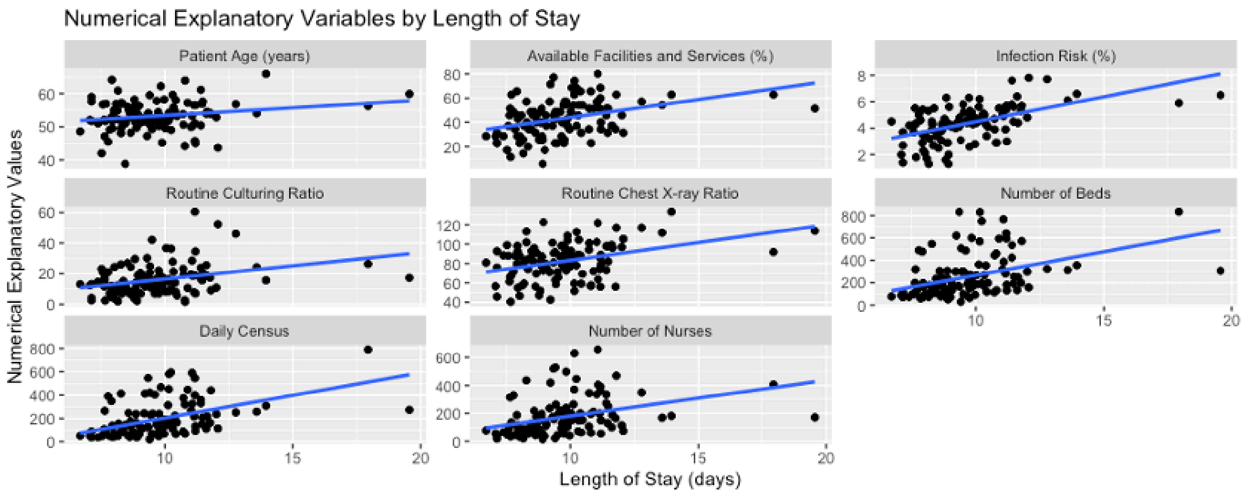


Figure 4: Faceted Scatter Plots of Senic Numerical Variables Plotted Against Average Patient Length of Stay (days)

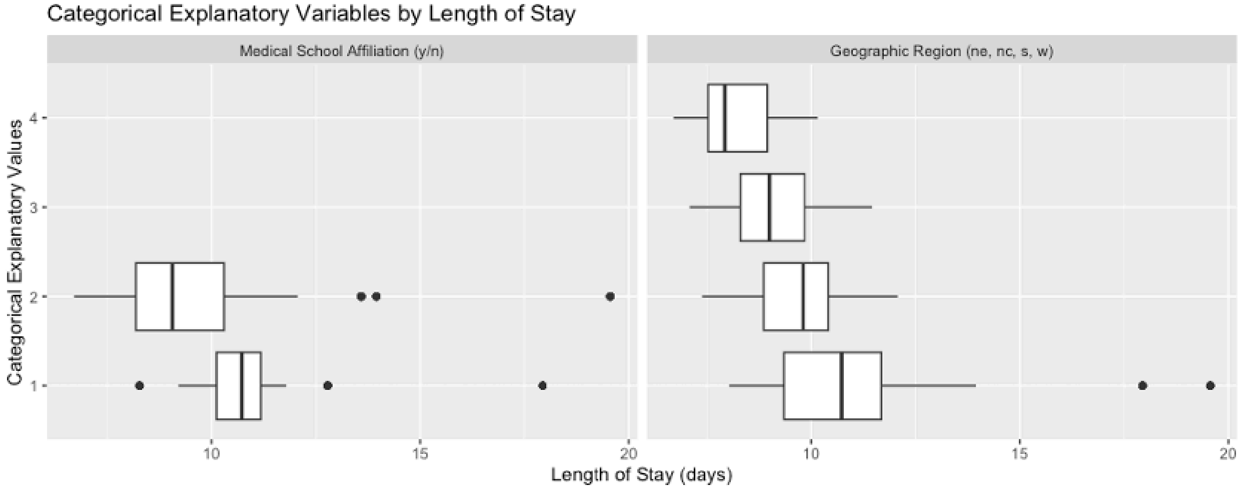


Figure 5: Faceted Boxplots of Senic Categorical Variables Plotted Against Average Patient Length of Stay (days)

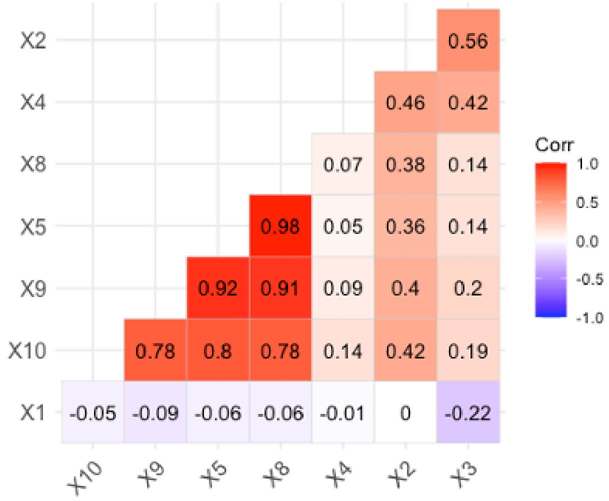


Figure 6: Correlation Matrix between Senic Numerical Variables

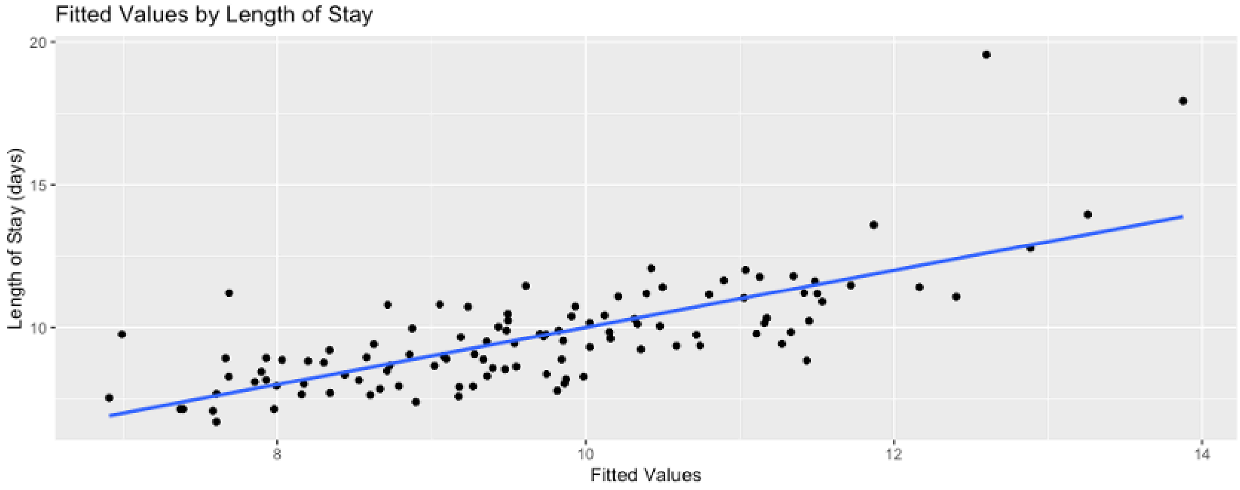


Figure 7: Scatter Plot of Fitted Values Plotted Against Average Patient Length of Stay (days) for the Final Model Before Outlier Removal

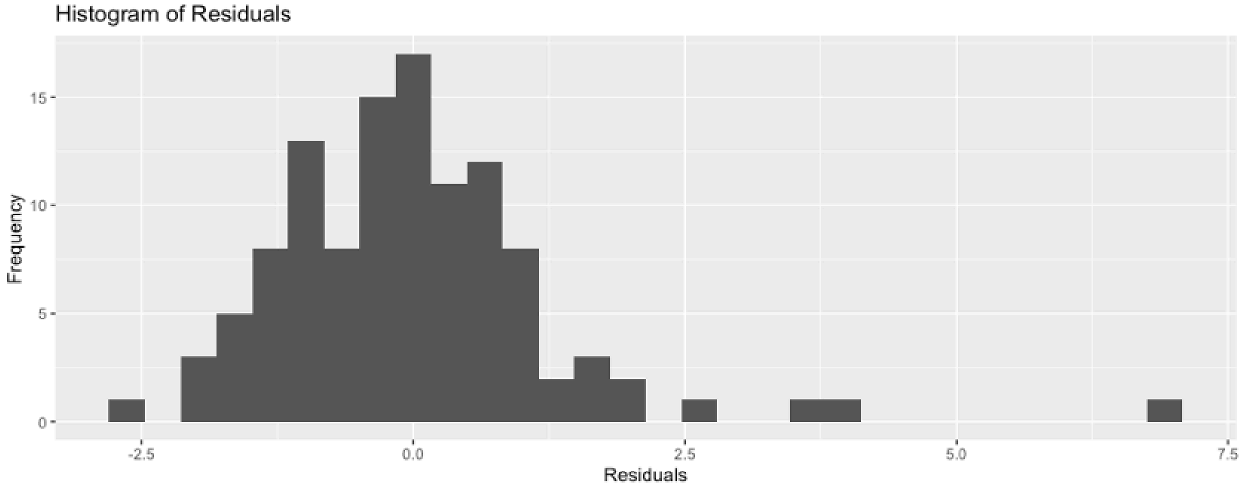


Figure 8: Histogram of Residuals for the Final Model Before Outlier Removal

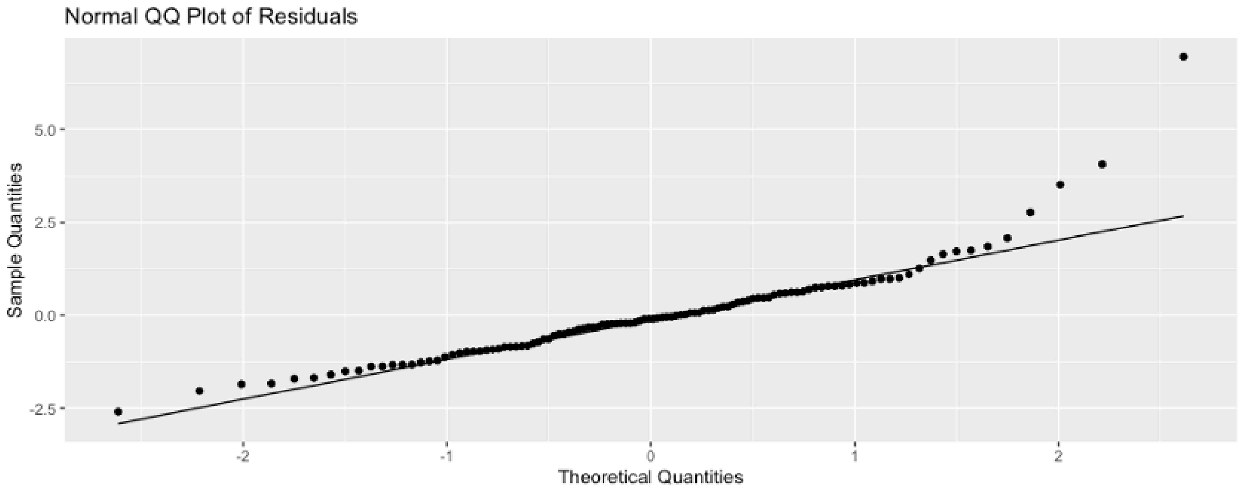


Figure 9: Normal QQ Plot of Residuals for the Final Model Before Outlier Removal

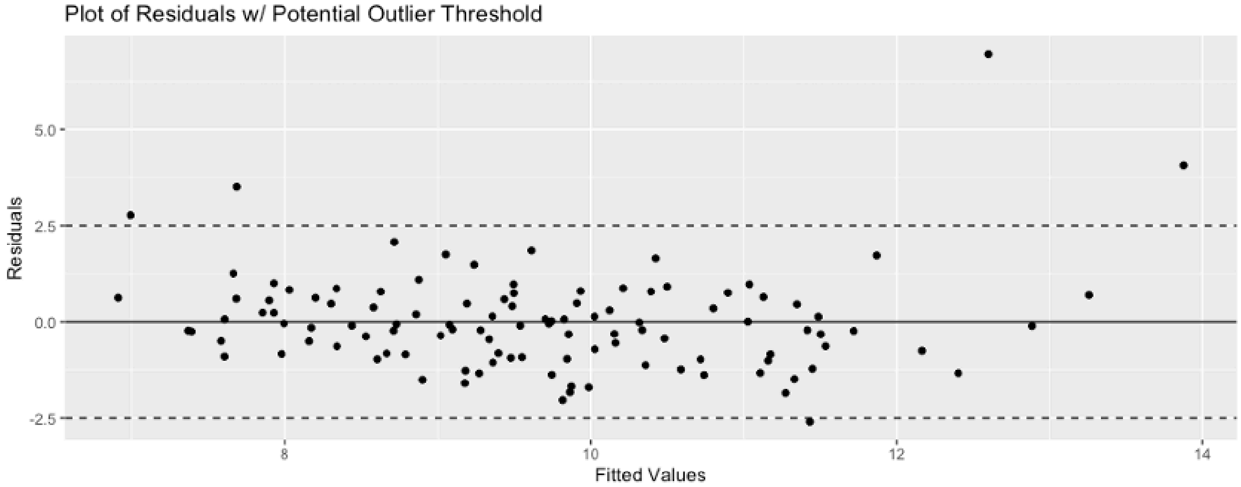


Figure 10: Scatter Plot of Fitted Values Plotted Against Residuals for the Final Model Before Outlier Removal with the Outlier Threshold Marked

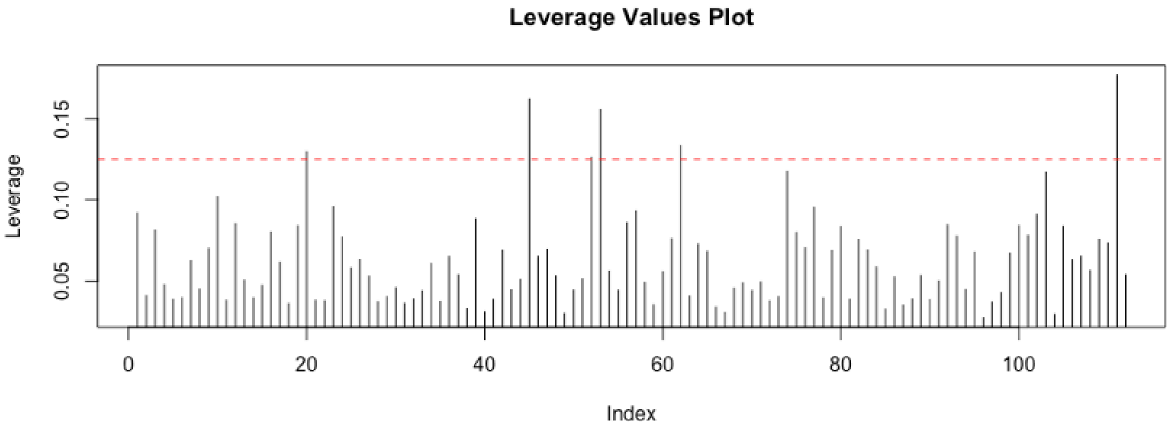


Figure 11: Plot of High Leverage Points for the Final Model Before Outlier Removal

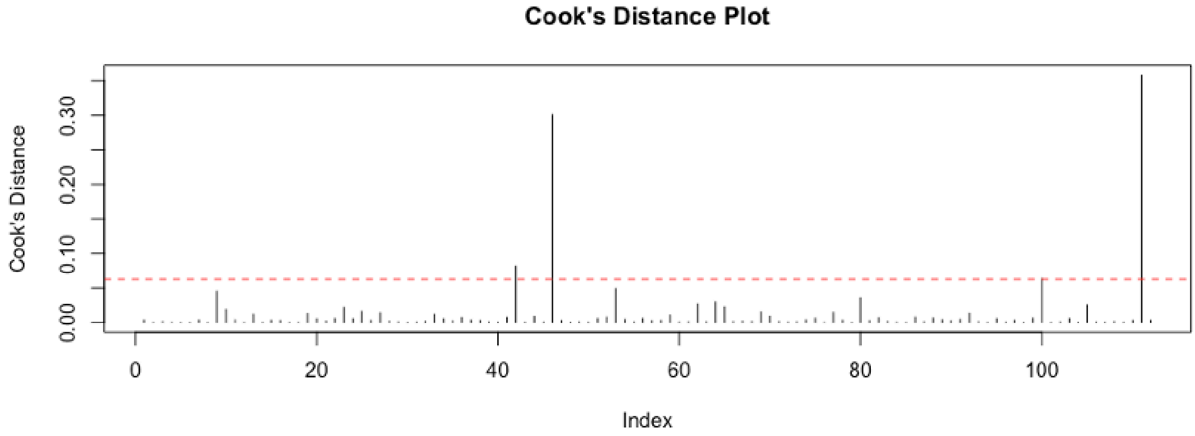


Figure 12: Plot of Influential Points for the Final Model Before Outlier Removal

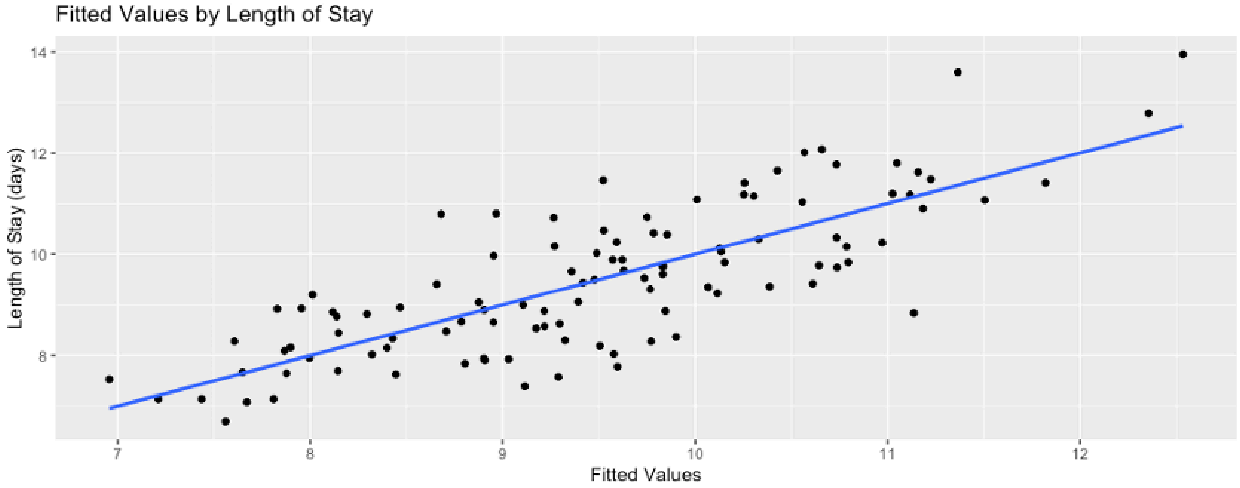


Figure 13: Scatter Plot of Fitted Values Plotted Against Average Patient Length of Stay (days) After Outlier Removal

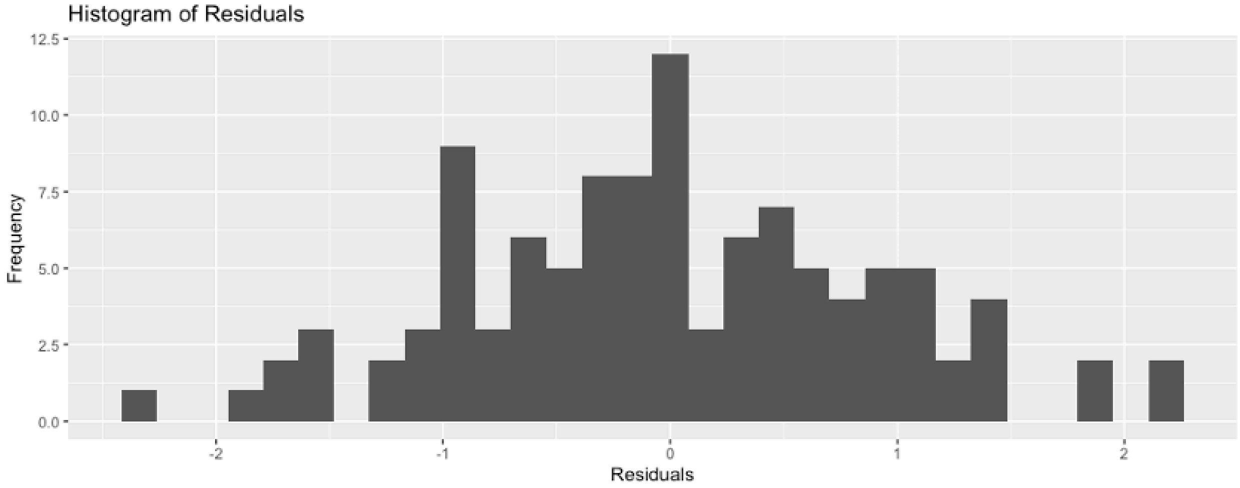


Figure 14: Histogram of Residuals After Outlier Removal

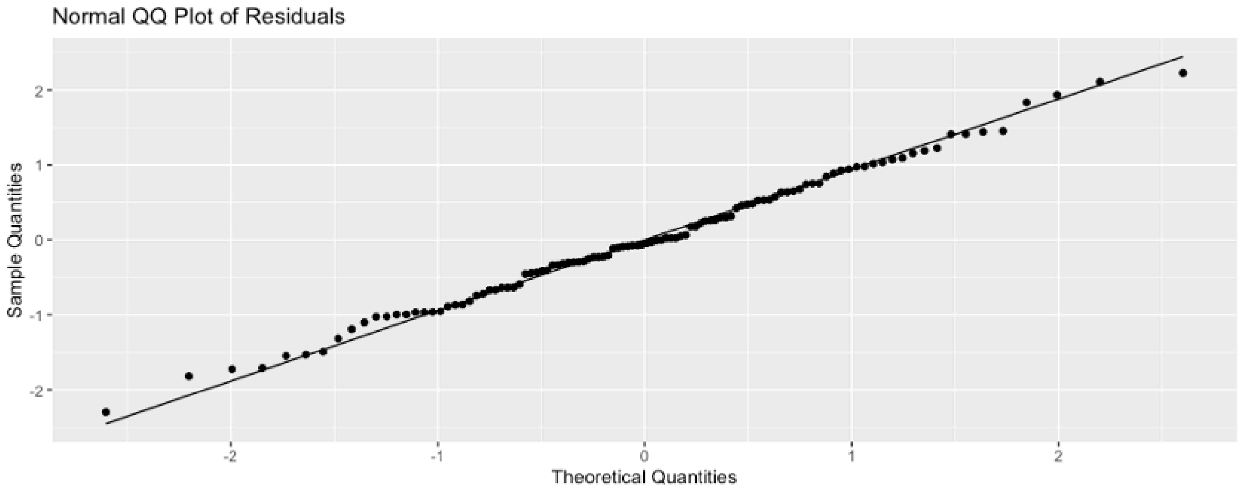


Figure 15: Normal QQ Plot of Residuals After Outlier Removal

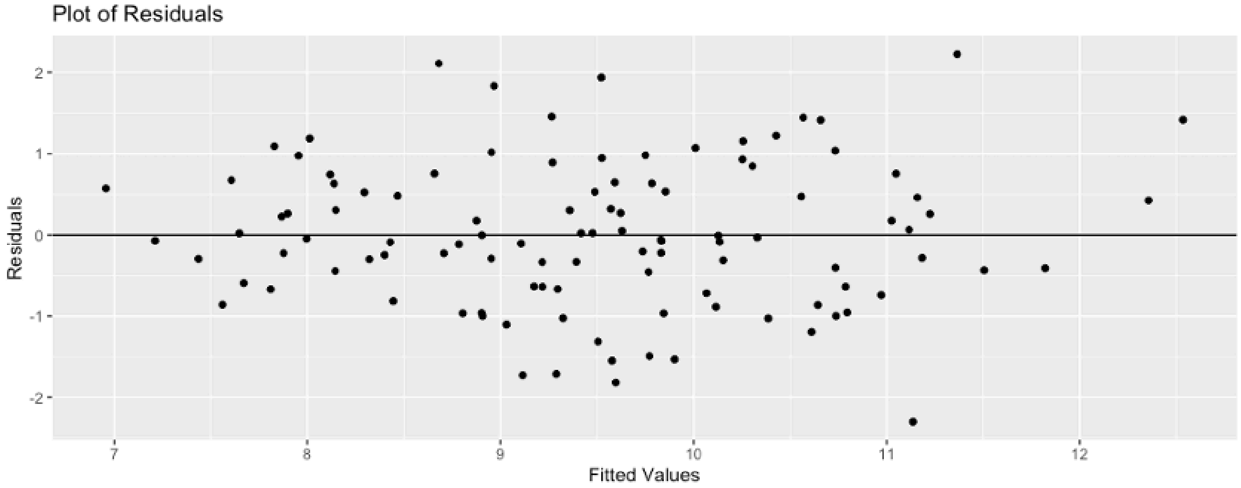


Figure 16: Scatter Plot of Fitted Values Plotted Against Residuals After Outlier Removal

Appendix: Tables

Table 1: Summary Statistics for Numeric Senic Variables

Variable	Mean	SD
Length of Stay (Y)	9.670804	1.904976
Patient Age (X1)	53.209821	4.475478
Infection Risk (X2)	4.357143	1.346715
Culturing Ratio (X3)	15.853571	10.260288
X-ray Ratio (X4)	82.003571	19.033664
Beds (X5)	251.928571	193.692508
Daily Census (X6)	191.232143	154.443434
Nurses (X9)	172.642857	139.742096
Facilities Offered (X10)	43.008929	15.184532

Table 2: VIF Values for Senic Variables

Variable	VIF
Patient Age (X1)	1.176911
Infection Risk (X2)	2.155635
Culturing Ratio (X3)	1.978736
X-ray Ratio (X4)	1.413107
Beds (X5)	36.114296
Medical School (X6)	1.874651
Region (X7)	1.718537
Daily Census (X8)	34.421151
Nurses (X9)	7.050241
Facilities Offered (X10)	3.311386

Table 3: Model BICs

Model	BIC
Full	417.4897
Full - X5	414.9036
Full - X8	425.4263
X1 + X2 + X7 + X8 (Pre-removal)	408.8766
X1 + X2 + X7 + X8 (Post-removal)	319.9867

Table 4: Final Model (Pre-Outlier-Removal)

Coefficient	Estimate
Intercept	3.5122223
X1	0.0800081
X2	0.5086450
X7 (NC)	-0.9669117
X7 (S)	-1.2957230
X7 (W)	-2.2695829
X8	0.0036260

Table 5: Diagnostic Tests (Pre-Outlier-Removal)

Test	Statistic	P-value
Shapiro-Wilk	0.8791	0.00000004457
Fligner-Kileen	0.66194	0.4159

Table 6: Outliers

Row#	Y	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
9	8.84	56.3	6.3	29.6	82.6	85	2	1	59	66	40.0
42	11.20	45.0	3.0	7.0	78.9	130	2	3	95	56	34.3
46	19.56	59.9	6.5	17.2	113.7	306	2	1	273	172	51.4
100	9.76	53.2	2.6	6.9	80.1	64	2	4	47	55	22.9
111	17.94	56.2	5.9	26.4	91.8	835	1	1	791	407	62.9

Table 7: High Leverage Points

Row#	Y	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
20	7.53	42.0	4.2	23.1	98.9	95	2	4	47	49	17.1
45	10.16	54.2	4.6	8.4	51.5	831	1	4	581	629	74.3
52	11.41	61.1	7.6	16.6	97.9	535	2	3	330	273	51.4
53	12.07	43.7	7.8	52.4	105.3	157	2	2	115	76	31.4
62	7.93	64.1	5.4	7.5	98.1	68	2	4	42	49	28.6
111	17.94	56.2	5.9	26.4	91.8	835	1	1	791	407	62.9

Table 8: Influential Points

Row#	Y	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
42	11.20	45.0	3.0	7.0	78.9	130	2	3	95	56	34.3
46	19.56	59.9	6.5	17.2	113.7	306	2	1	273	172	51.4
100	9.76	53.2	2.6	6.9	80.1	64	2	4	47	55	22.9
111	17.94	56.2	5.9	26.4	91.8	835	1	1	791	407	62.9

Table 9: Final Model Summary (Post-Outlier-Removal)

Coefficient	Estimate	Std. Error	T-value	P-Value
Intercept	4.0156674	1.1936912	3.364	0.001086 **
X1	0.0666492	0.0207443	3.213	0.001764 **
X2	0.5120519	0.0748493	6.841	0.000000000616 ***
X7 (NC)	-0.5426321	0.2508467	-2.163	0.032885 *
X7 (S)	-0.9602391	0.2461705	-3.901	0.000173 ***
X7 (W)	-2.1230293	0.3096017	-6.857	0.000000000570 ***
X8	0.0024271	0.0006803	3.568	0.000553

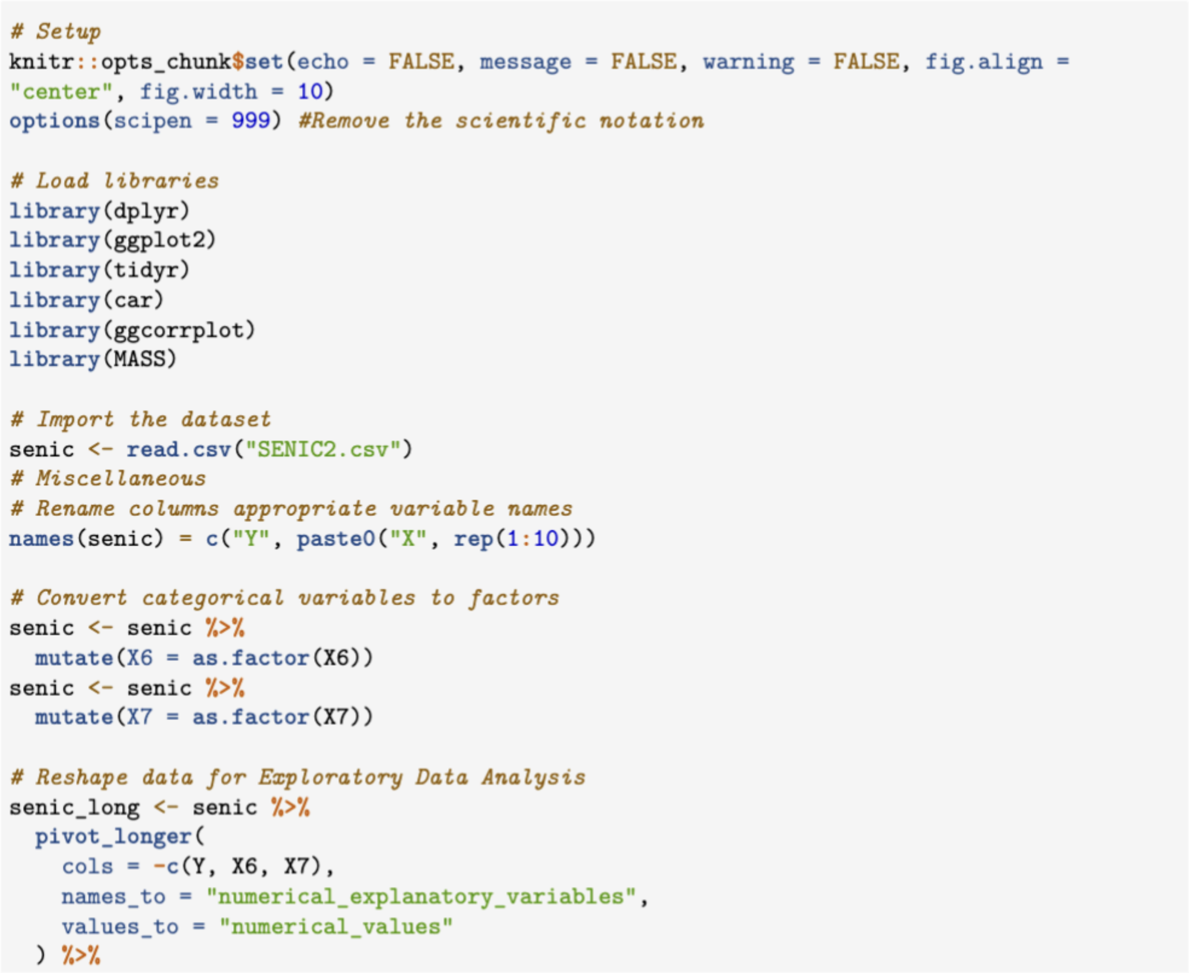
Table 10: Diagnostic Test (Post-Outlier-Removal)

Test	Statistic	P-value
Shapiro-Wilk	0.99476	0.9576
Fligner-Kileen	0.20333	0.652
General Linear	1.0187	0.418

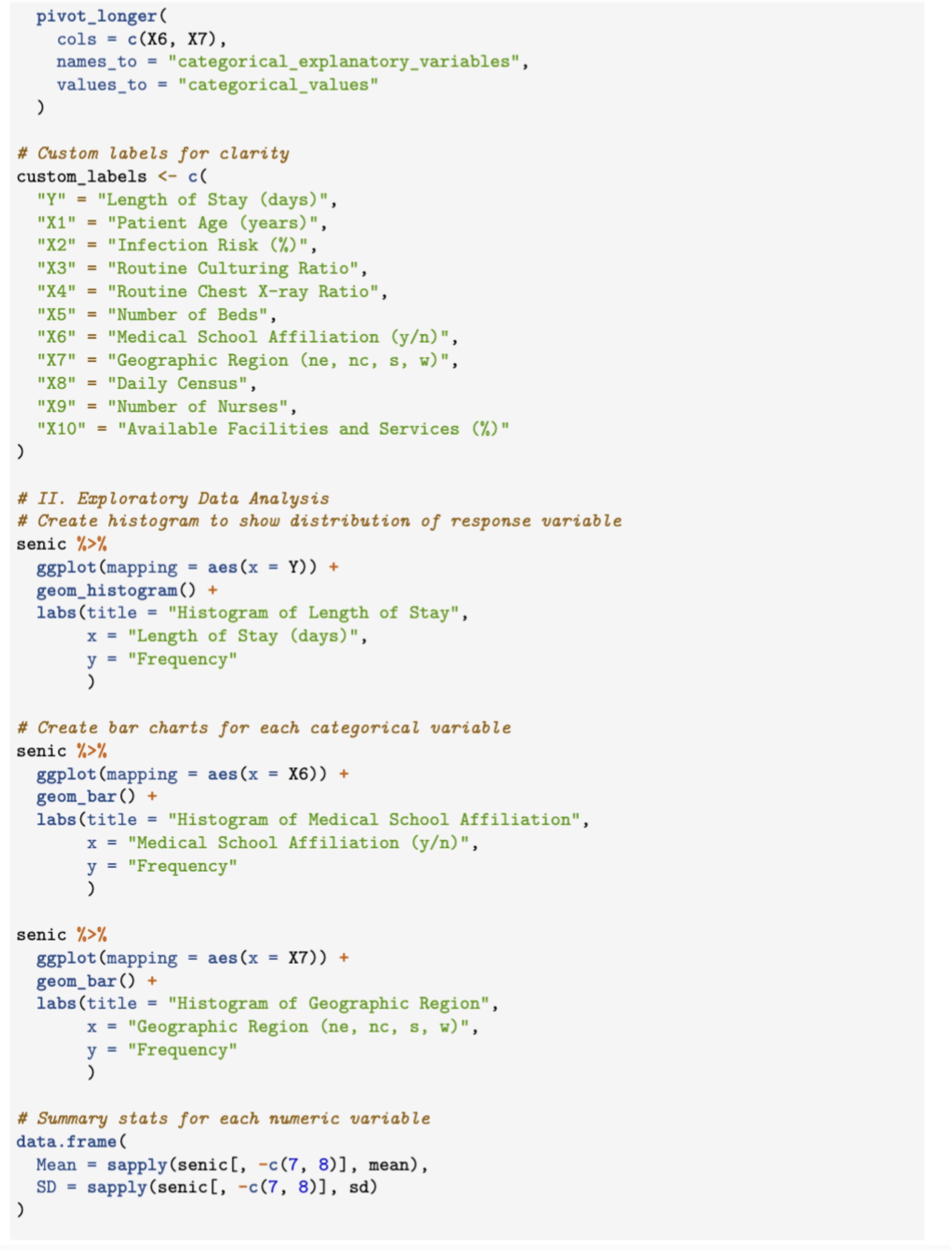
Table 11: Simultaneous Confidence Intervals for Model Coefficients (Bonferroni)

Coefficient	Estimate	Lower	Upper
Intercept	4.015667418	0.7377912995	7.293543537
X1	0.066649223	0.0096852708	0.123613174
X2	0.512051894	0.3065157280	0.717588060
X7 (NC)	-0.542632130	-1.2314571913	0.146192930
X7 (S)	-0.960239115	-1.6362232308	-0.284254999
X7 (W)	-2.123029350	-2.9731957124	-1.272862987
X8	0.002427056	0.0005588894	0.004295223

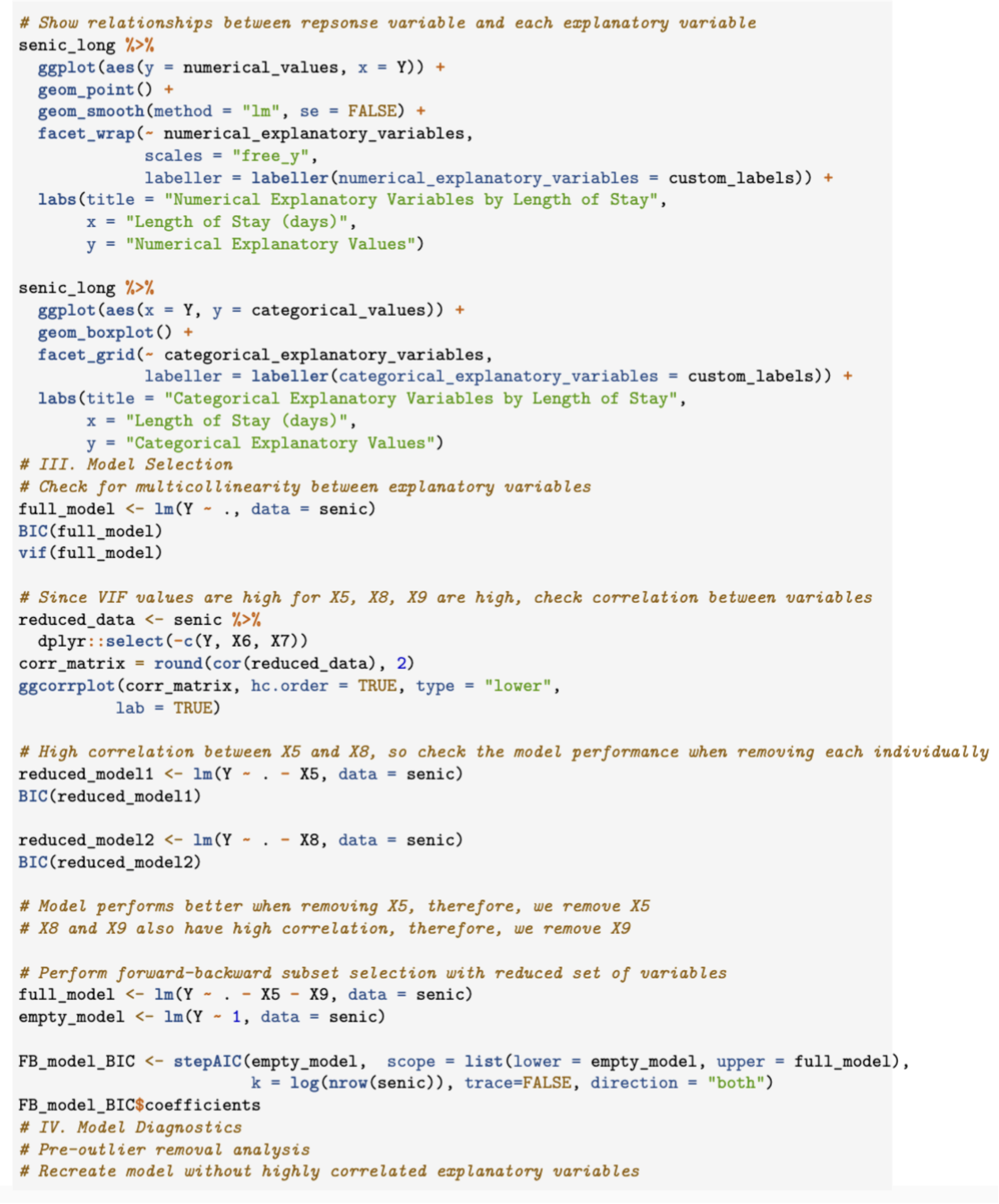
Appendix: R Code



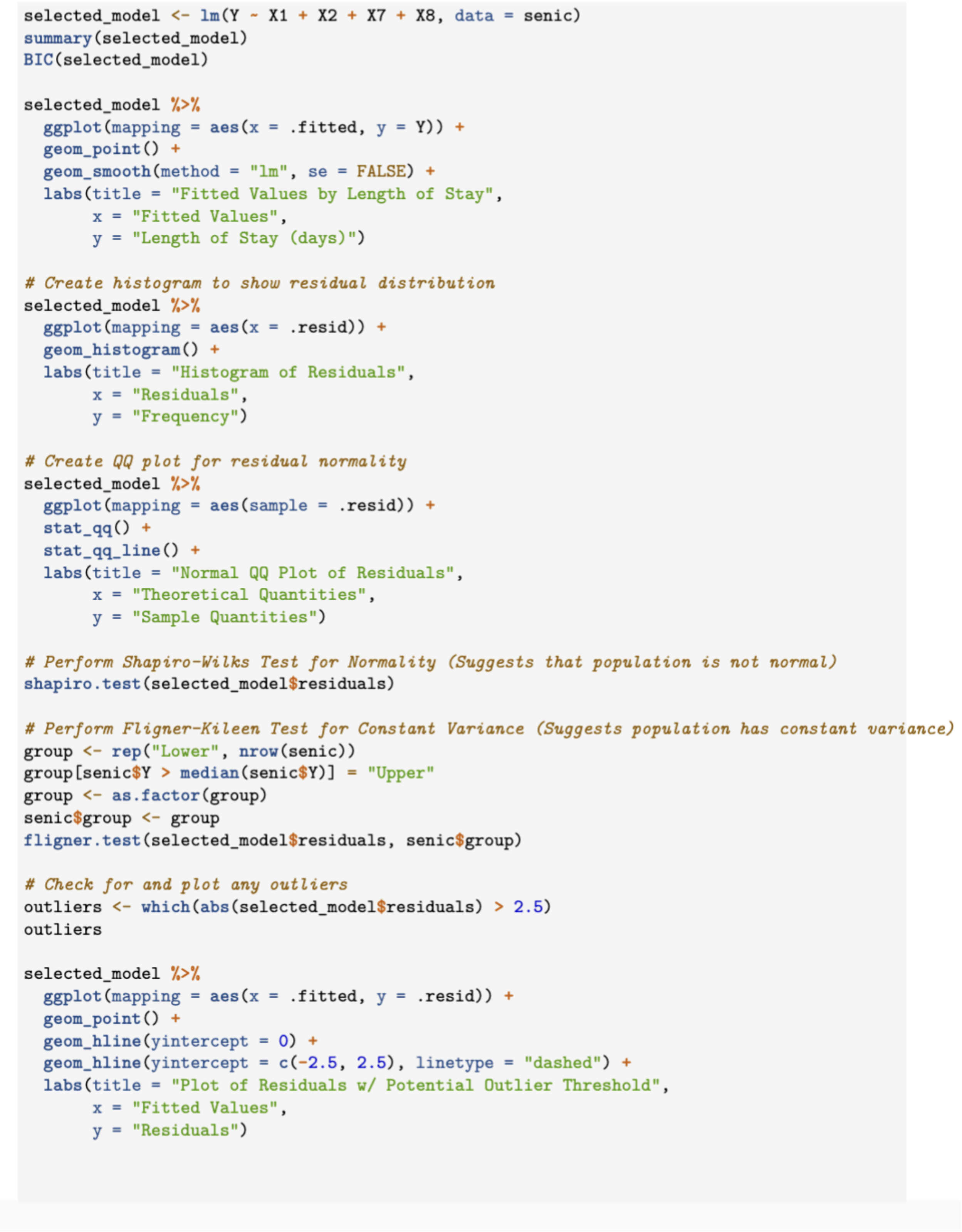
```r
# Setup
knitr::opts_chunk$set(echo = FALSE, message = FALSE, warning = FALSE, fig.align =
"center", fig.width = 10)
options(scipen = 999) #Remove the scientific notation

# Load libraries
library(dplyr)
library(ggplot2)
library(tidyr)
library(car)
library(ggcorrplot)
library(MASS)

# Import the dataset
senic <- read.csv("SENIC2.csv")
# Miscellaneous
# Rename columns appropriate variable names
names(senic) = c("Y", paste0("X", rep(1:10)))

# Convert categorical variables to factors
senic <- senic %>%
  mutate(X6 = as.factor(X6))
senic <- senic %>%
  mutate(X7 = as.factor(X7))

# Reshape data for Exploratory Data Analysis
senic_long <- senic %>%
  pivot_longer(
    cols = -c(Y, X6, X7),
    names_to = "numerical_explanatory_variables",
    values_to = "numerical_values"
  ) %>%
```



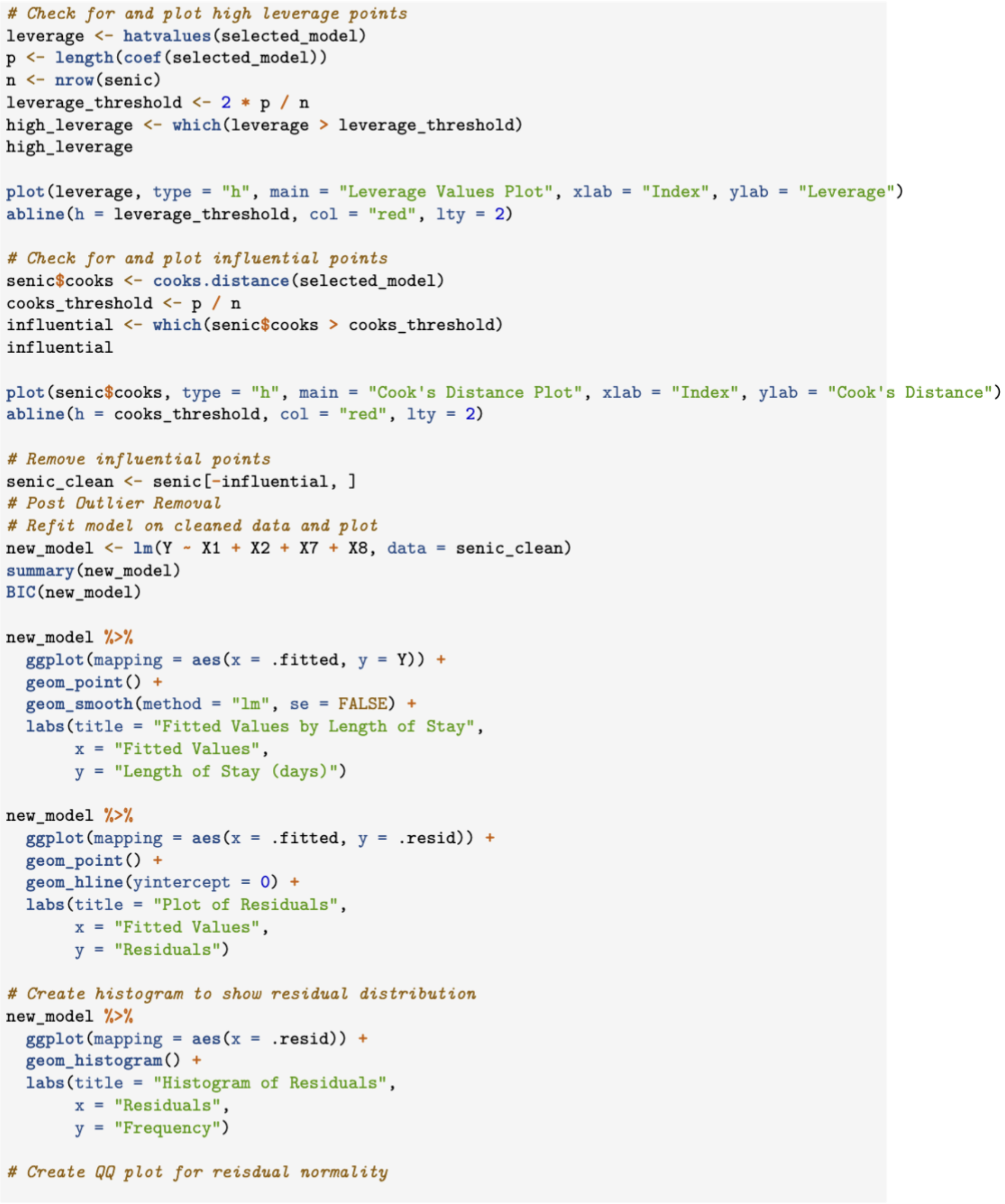
```r
  pivot_longer(
    cols = c(X6, X7),
    names_to = "categorical_explanatory_variables",
    values_to = "categorical_values"
  )

# Custom labels for clarity
custom_labels <- c(
  "Y" = "Length of Stay (days)",
  "X1" = "Patient Age (years)",
  "X2" = "Infection Risk (%)",
  "X3" = "Routine Culturing Ratio",
  "X4" = "Routine Chest X-ray Ratio",
  "X5" = "Number of Beds",
  "X6" = "Medical School Affiliation (y/n)",
  "X7" = "Geographic Region (ne, nc, s, w)",
  "X8" = "Daily Census",
  "X9" = "Number of Nurses",
  "X10" = "Available Facilities and Services (%)"
)

# II. Exploratory Data Analysis
# Create histogram to show distribution of response variable
senic %>%
  ggplot(mapping = aes(x = Y)) +
  geom_histogram() +
  labs(title = "Histogram of Length of Stay",
       x = "Length of Stay (days)",
       y = "Frequency"
       )

# Create bar charts for each categorical variable
senic %>%
  ggplot(mapping = aes(x = X6)) +
  geom_bar() +
  labs(title = "Histogram of Medical School Affiliation",
       x = "Medical School Affiliation (y/n)",
       y = "Frequency"
       )

senic %>%
  ggplot(mapping = aes(x = X7)) +
  geom_bar() +
  labs(title = "Histogram of Geographic Region",
       x = "Geographic Region (ne, nc, s, w)",
       y = "Frequency"
       )

# Summary stats for each numeric variable
data.frame(
  Mean = sapply(senic[, -c(7, 8)], mean),
  SD = sapply(senic[, -c(7, 8)], sd)
)
```



```r
# Show relationships between repsonse variable and each explanatory variable
senic_long %>%
  ggplot(aes(y = numerical_values, x = Y)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  facet_wrap(~ numerical_explanatory_variables,
             scales = "free_y",
             labeller = labeller(numerical_explanatory_variables = custom_labels)) +
  labs(title = "Numerical Explanatory Variables by Length of Stay",
       x = "Length of Stay (days)",
       y = "Numerical Explanatory Values")

senic_long %>%
  ggplot(aes(x = Y, y = categorical_values)) +
  geom_boxplot() +
  facet_grid(~ categorical_explanatory_variables,
             labeller = labeller(categorical_explanatory_variables = custom_labels)) +
  labs(title = "Categorical Explanatory Variables by Length of Stay",
       x = "Length of Stay (days)",
       y = "Categorical Explanatory Values")
# III. Model Selection
# Check for multicollinearity between explanatory variables
full_model <- lm(Y ~ ., data = senic)
BIC(full_model)
vif(full_model)

# Since VIF values are high for X5, X8, X9 are high, check correlation between variables
reduced_data <- senic %>%
  dplyr::select(-c(Y, X6, X7))
corr_matrix = round(cor(reduced_data), 2)
ggcorrplot(corr_matrix, hc.order = TRUE, type = "lower",
          lab = TRUE)

# High correlation between X5 and X8, so check the model performance when removing each individually
reduced_model1 <- lm(Y ~ . - X5, data = senic)
BIC(reduced_model1)

reduced_model2 <- lm(Y ~ . - X8, data = senic)
BIC(reduced_model2)

# Model performs better when removing X5, therefore, we remove X5
# X8 and X9 also have high correlation, therefore, we remove X9

# Perform forward-backward subset selection with reduced set of variables
full_model <- lm(Y ~ . - X5 - X9, data = senic)
empty_model <- lm(Y ~ 1, data = senic)

FB_model_BIC <- stepAIC(empty_model,  scope = list(lower = empty_model, upper = full_model),
                        k = log(nrow(senic)), trace=FALSE, direction = "both")
FB_model_BIC$coefficients
# IV. Model Diagnostics
# Pre-outlier removal analysis
# Recreate model without highly correlated explanatory variables
```



```r
selected_model <- lm(Y ~ X1 + X2 + X7 + X8, data = senic)
summary(selected_model)
BIC(selected_model)

selected_model %>%
  ggplot(mapping = aes(x = .fitted, y = Y)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Fitted Values by Length of Stay",
       x = "Fitted Values",
       y = "Length of Stay (days)")

# Create histogram to show residual distribution
selected_model %>%
  ggplot(mapping = aes(x = .resid)) +
  geom_histogram() +
  labs(title = "Histogram of Residuals",
       x = "Residuals",
       y = "Frequency")

# Create QQ plot for residual normality
selected_model %>%
  ggplot(mapping = aes(sample = .resid)) +
  stat_qq() +
  stat_qq_line() +
  labs(title = "Normal QQ Plot of Residuals",
       x = "Theoretical Quantities",
       y = "Sample Quantities")

# Perform Shapiro-Wilks Test for Normality (Suggests that population is not normal)
shapiro.test(selected_model$residuals)

# Perform Fligner-Kileen Test for Constant Variance (Suggests population has constant variance)
group <- rep("Lower", nrow(senic))
group[senic$Y > median(senic$Y)] = "Upper"
group <- as.factor(group)
senic$group <- group
fligner.test(selected_model$residuals, senic$group)

# Check for and plot any outliers
outliers <- which(abs(selected_model$residuals) > 2.5)
outliers

selected_model %>%
  ggplot(mapping = aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0) +
  geom_hline(yintercept = c(-2.5, 2.5), linetype = "dashed") +
  labs(title = "Plot of Residuals w/ Potential Outlier Threshold",
       x = "Fitted Values",
       y = "Residuals")
```



```
# Check for and plot high leverage points
leverage <- hatvalues(selected_model)
p <- length(coef(selected_model))
n <- nrow(senic)
leverage_threshold <- 2 * p / n
high_leverage <- which(leverage > leverage_threshold)
high_leverage

plot(leverage, type = "h", main = "Leverage Values Plot", xlab = "Index", ylab = "Leverage")
abline(h = leverage_threshold, col = "red", lty = 2)

# Check for and plot influential points
senic$cooks <- cooks.distance(selected_model)
cooks_threshold <- p / n
influential <- which(senic$cooks > cooks_threshold)
influential

plot(senic$cooks, type = "h", main = "Cook's Distance Plot", xlab = "Index", ylab = "Cook's Distance")
abline(h = cooks_threshold, col = "red", lty = 2)

# Remove influential points
senic_clean <- senic[-influential, ]
# Post Outlier Removal
# Refit model on cleaned data and plot
new_model <- lm(Y ~ X1 + X2 + X7 + X8, data = senic_clean)
summary(new_model)
BIC(new_model)

new_model %>%
  ggplot(mapping = aes(x = .fitted, y = Y)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Fitted Values by Length of Stay",
       x = "Fitted Values",
       y = "Length of Stay (days)")

new_model %>%
  ggplot(mapping = aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0) +
  labs(title = "Plot of Residuals",
       x = "Fitted Values",
       y = "Residuals")

# Create histogram to show residual distribution
new_model %>%
  ggplot(mapping = aes(x = .resid)) +
  geom_histogram() +
  labs(title = "Histogram of Residuals",
       x = "Residuals",
       y = "Frequency")

# Create QQ plot for reisdual normality
```

```
new_model %>%
  ggplot(mapping = aes(sample = .resid)) +
  stat_qq() +
  stat_qq_line() +
  labs(title = "Normal QQ Plot of Residuals",
       x = "Theoretical Quantities",
       y = "Sample Quantities")

# Perform Shapiro-Wilks Test for Normality (Suggests that population is normal)
shapiro.test(new_model$residuals)

# Perform Fligner-Kileen Test for Constant Variance (Suggests population has constant variance)
group <- rep("Lower", nrow(senic_clean))
group[senic_clean$Y > median(senic_clean$Y)] <- "Upper"
group <- as.factor(group)
senic_clean$group <- group
fligner.test(new_model$residuals, senic_clean$group)
# Confidence Intervals and Hypothesis Tests
# Find best multiplies for multiple CI
mult.fun <- function(n, p, g, alpha){
  bon = qt(1 - alpha / (2 * g), n - p)
  WH = sqrt(p * qf(1 - alpha, p, n - p))
  all.mul = c(bon, WH)
  all.mul = round(all.mul, 3)
  names(all.mul) = c("Bon", "WH")
  return(all.mul)
}

multipliers <- mult.fun(nrow(senic_clean), length(new_model$coefficients), 7, 0.05)
multipliers

coefs <- coef(summary(new_model))

# Find multiple CI's using bonferroni
data.frame(
  Estimate = coefs[, "Estimate"],
  Lower = coefs[, "Estimate"] - multipliers[1] * coefs[, "Std. Error"],
  Upper = coefs[, "Estimate"] + multipliers[1] * coefs[, "Std. Error"]
)

full_model <- lm(Y ~ . - group - cooks, data = senic_clean)
anova(new_model, full_model)
# General Linear Test
# Null Hypothesis: B3 = B4 = B5 = B6(Yes) = B9 = B10 = 0 (Reduced Model Fits Better)
# Alternative Hypothesis: At least 1 Bi =/ 0, i = 3, 4, 5, 6, 9, 10 (Full Model Fits Better)
# F stat: 1.0187
# P-value: 0.418
# Since the p-value is 0.418 > alpha = 0.05, we fail to reject the null hypothesis and
# conclude that the reduced model fits the data better than the full model.
tinytex::install_tinytex()
```