Hospital Multiple Linear Regression Project

By: Ethan Hershman

I. Introduction

Our goal is to understand the relationship between the number of active physicians in a county (response variable) and various explanatory variables, ultimately identifying the most accurate multiple linear model. To achieve this, we utilized the County Demographic Information (CDI) dataset, which provides detailed demographic and socioeconomic data for 440 of the most populous counties in the United States. This dataset includes variables such as land area, population demographics, economic indicators, and healthcare resources, with each row representing a specific county identified by its name and state abbreviation. From this dataset, we focused on the response variable, Number of Active Physicians, and analyzed its relationship with ten explanatory variables through statistical methods. Based on our analysis, we identified the model Y ~ X2 + X3 + X5 + X10 as the best multiple linear model for correctly capturing the relationship between Y and explanatory variables.

II. Exploratory Data Analysis

Our summary plots are scatterplots of each of the explanatory variables (X's) compared to the response variable, which is the number of active physicians in a county (Y). Each scatterplot (Figure 2.1) reveals the presence of outliers and noticeable variance between the variables, which will need to be addressed before proceeding with further model selection.

For numerical summaries (Figure 2.2), we extracted the mean and standard deviation for each explanatory variable. Among these, X2 (Total Population) had the highest mean and standard deviation, with a mean of 3.74 * 10^5 and a standard deviation of 4.461 * 10^5. In contrast, X10 (Geographic Region) had the lowest mean and standard deviation, with a mean of 2.458 and a standard deviation of 1.033. This is expected, as X10 is a categorical variable with levels ranging from 1 to 4, resulting in smaller values compared to variables like X2, which represents the total population of a county.

Based on the summary models and the numerical summaries, we are able to see that there is a possibly linear trend between some of the explanatory variables and the response variable, and we will further analyze to see which combination of explanatory variables have a relationship with the number of active physicians in a county.

III. Model Selection

Since our goal is focused on finding a correct model, we can further test whether a multiple linear regression model is the best fit for each of the scatterplots by using model selection. We decided to find the best model for correction and found our chosen model as Y ~ X2 + X3 + X5 + X9 + X10. To conclude this model, we applied different model selection techniques: Forward Step-wise Regression, Backward Step-wise Selection, Forward Backward Step-wise Selection, and Backward Forward Step-wise Selection to find the AIC and BIC values for each (Figure 3.1). As a result from all the model selection techniques, the lowest AIC value (6414.469) was for the model: Y ~ X1 + X2 + X3 + X5 + X9 + X10, and the lowest BIC value (6415.489) was from the model: Y ~ X2 + X3 + X5 + X9 + X10.

Additionally, to provide context, the number of active physicians in a single county (Y) can be explained through the relationship between Total Population (X2), Percent of Population aged 18-34 (X3), Number of Hospital beds (X5), Total Personal Income (X9), and Geographic Region (X10).

IV. Model Diagnostics

The diagnostic plots are a residual plot, Q-Q plot, Scale-Location plot, Residuals vs. Leverage plot, and a histogram of residuals (Figure 4.1). In regards to linearity, the residual plot shows deviation from the horizontal trend, indicating potential non-linearity in the relationship between the predictors and the response variable. As for predictor independence (Figure 4.7.1), we performed a Variance Inflation Factor (VIF) analysis to check for multicollinearity, which occurs when explanatory variables are strongly correlated. A VIF value over 5 suggests potential multicollinearity issues. In our analysis, X2 (Total Population), X5 (Number of Hospital Beds), and X9 (Total Personal Income) crossed this threshold. To better understand the relationships between variables, we used a correlation matrix to measure the strength and direction of their linear connections (Figure 4.7.1). The pairs with the strongest correlations were X2~X5 (0.896), X2~XX9 (0.977), and X5~X9 (0.859). This indicates multicollinearity, and potentially violates the assumption that the predictors are independent. In the residual plot, the values appear to be centered around zero, suggesting that the errors have a mean of approximately zero. To check constant variance (homoscedasticity), the Scale-Location plot shows a slight cone shape, where the variances increase with higher fitted values. This goes against homoscedasticity, meaning the errors do not have constant variance. Based on the Q-Q plot, there are some deviations as shown at the ends of the tails, suggesting that the errors are not perfectly normally distributed. Additionally, the histogram of errors has a slight right skew, and appears to have outliers, indicating that outliers may need to be identified to improve the correct model.

For the Fligner-Killeen Test, the null hypothesis is that the variances of the two groups are equal, and the alternative hypothesis is that the variances of the two groups are not equal. For the Shapiro Wilks Test, the null hypothesis was the population is normally distributed, and alternative hypothesis is population is not normally distributed.

Overall, our assumptions of linearity, normality, and constant variance are violated, so this requires the removal of outliers.

To improve our current correct model, we performed tests to remove potential outliers, high leverage points, and influential points from the current best model. We used a residual plot and standardized residuals to identify outliers (Figure 4.10). High leverage points were removed by identifying leverage points above a certain threshold (Figure 4.3). Finally, influential points were detected by calculating Cook's Distance (Figure 4.4). After all these tests were performed, there were 77 outliers identified (Figure 4.7.1). However, from the influence plot, we saw that 5 data points had the highest impact in influencing our chosen model (Figure 4.5). We know that we have more power with more data so it is not necessary to remove all 77 outliers. So, we decided to remove these 5 data points as they had the highest influence in affecting the final model results (Figure 4.6.2). Now, we can proceed with our model of $Y \sim X2 + X3 + X5 + X9 + X10$ with the removal of outliers.

To eliminate redundant variables without compromising correctness, we considered the context of the dataset by re-analyzing our VIF and correlation matrix. X2 (Total Population) is a key variable that's often correlated with others, so we decided to keep it in the model. On the other hand, X9 (Total Personal Income) is closely tied to X2, as a county's population grows, its total income also increases, since income can't be negative. Due to this overlap, we removed X9 from the model. After removing X9, we recalculated the VIF values and confirmed that none of the variables exceeded the threshold of 5. This shows our revised model no longer shows signs of multicollinearity (Figure 4.6.2).

We also performed simultaneous confidence intervals to analyze whether each predictor is within their respective true population parameters, which is further analyzed in the analysis section.

V. Analysis and Interpretation

After removing X9 from the model as determined in the diagnostics section, the new model has a decreased adjusted R^2 of 0.8832 compared to the previous model with X9 of 0.9371. Since we are looking for the best correct model, we will still continue to remove X9 as we prioritize accurately understanding the relationship between response and explanatory variables and not overfitting, therefore, favoring a simpler model with only the most important predictors.

For the tests (Figure 5.1), we assume an alpha level of 0.05. Before removing outliers, the p-value for the Shapiro Wilks Test was 2.2 * 10^-16, and t-statistic of 0.81253. After removing outliers, the p-value was 2.2 * 10^-16, and a t-statistic of 0.765. In both cases, we reject the null hypothesis that the population is normally distributed, therefore, there is sufficient evidence to conclude that the population is not normally distributed. Before removing outliers, the p-value for the Fligner-Killeen Test was 2.2 * 10^-16, and t-statistic of 105.87. After removing outliers, the p-value was 2.2 * 10^-16, and t-statistic of 104.83. In both cases, we reject the null hypothesis that the variances of the two groups are equal. Therefore, there is sufficient evidence to conclude that the population does not have constant variance.

Since we rejected the null hypotheses for the Shapiro-Wilk and Fligner-Killeen tests, this indicates that the distribution of the average number of active physicians in the CDI data does not follow a normal distribution, nor are the variances equal. As these do not satisfy the assumptions for linear regression, it is important to exercise caution when moving forward with the model, as the violations of these assumptions could impact the validity of the results.

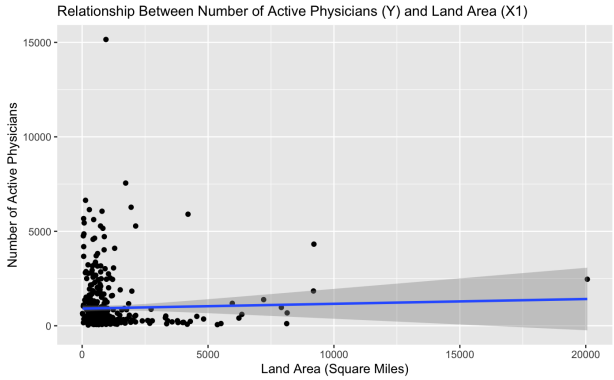
We are simultaneously 90% confident that the intercept and each predictor's estimated slope lies within the bounds of its confidence interval (Figure 5.2). This indicates that the true value of the coefficient is likely to fall within that interval, so the variable's effect is not trivially small or inconsistent. However, the Geographic Region (X10) does include 0 in the confidence interval. This is understandable because X10 is a categorical variable, and its numeric coding does not imply a natural order. As such, it may not accurately capture the true relationship between geographic region and the number of active physicians. Therefore, the effect of Geographic Region may be minimal or only slightly influential in predicting the average number of active physicians in our final model.

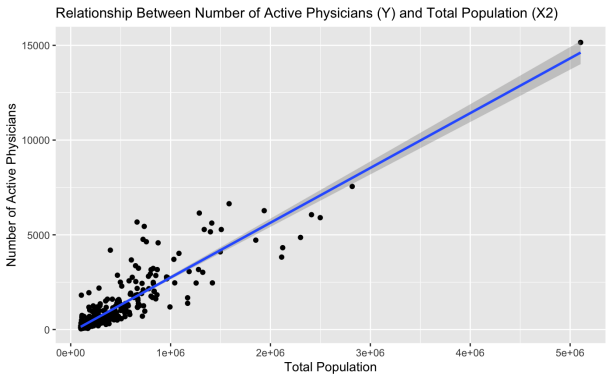
VI. Conclusion

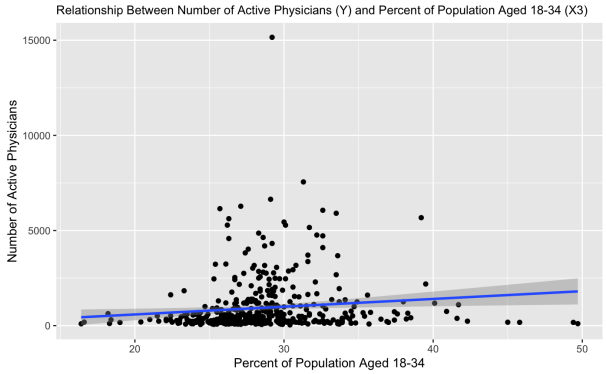
From our findings, the best model to understand the correct relationship between the number of active physicians in a county (Y) and the explanatory variables (X's) is Y ~ X2 + X3 + X5 + X10. Through statistical measures (AIC and BIC), the chosen model had the lowest resulting BIC value, which is fitting for finding model correctness. After performing the model diagnostics and checking assumptions, we concluded that all the assumptions were fairly violated except that the errors were independent, as shown by the non-normality, multicollinearity, and non-constant variance in the residual plot, histogram of errors, VIF table and Q-Q plot. However, even after a further step of removing influential points and dropping a redundant variable, the only assumption to change was multicollinearity. This introduces caution in the correctness of our model to accurately represent the relationship between active physicians in a single county and the explanatory variables. However, we should note a limitation that our final model may fail to capture potential non-linear relationships or might not represent the most accurate model due to omitted explanatory variables. A suggestion would be to test and try alternative models, like polynomial regression rather than linear regression, that could improve the model's accuracy in understanding the relationship between the response and explanatory variables.

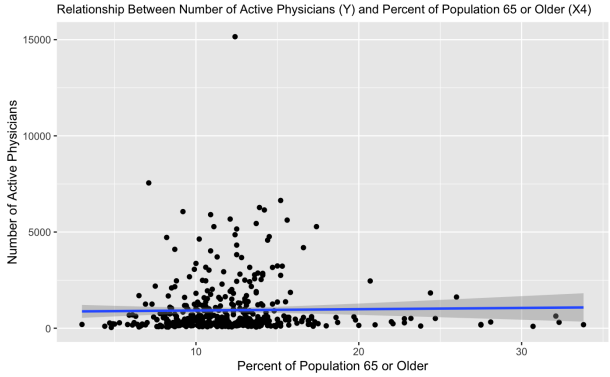
II. Exploratory Data Analysis

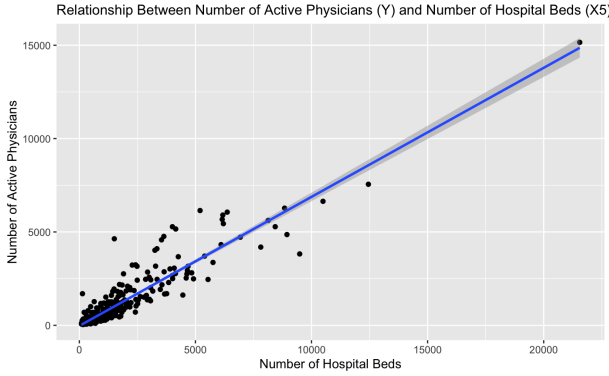
Figure 2.1: Scatterplots for each Relationship Between Number of Active Physicians (Y) and each Predictor Variable (X)

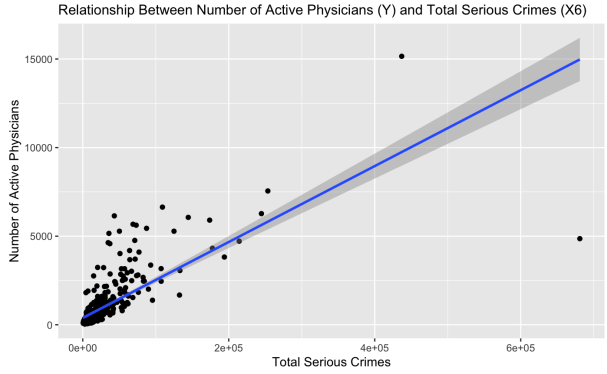












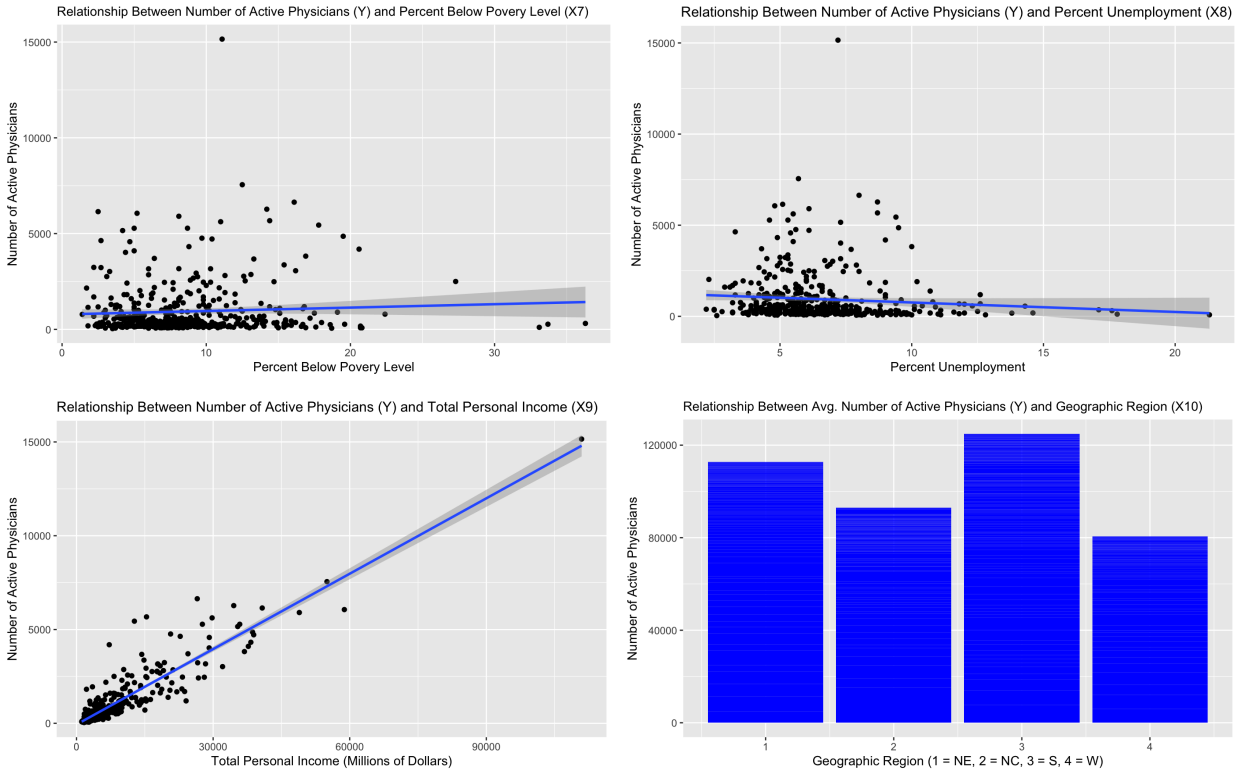


Figure 2.2: Sample Means and Standard Deviations of the Number of Active Physicians (Y) for each Predictor Variable (X)

Variable	Sample Means:	Standard Deviations:
X1	1.034535 * 10^3	1.544957 * 10^3
X2	3.737167 * 10^5	4.461443 * 10^5
X3	2.856036 * 10^1	4.192462
X4	1.21754 * 10^1	3.995475
X5	1.398852 * 10^3	1.917381 * 10^3
X6	2.560405 * 10^4	4.895964 * 10^4
X7	8.714123	4.660015
X8	6.593394	2.339628
X9	7.46754 * 10^3	9.757629 * 10^3
X10	2.457859	1.032542

Figure 2.2: The table above organizes and displays the sample means and standard deviations calculated for each predictor variable. The data helps us identify the average and variation of the average number of physicians based on each predictor for comparison.

Figure 2.3: Summary Table for Full Model

```
Call:
lm(formula = Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9 +
    X10, data = cdi2)

Residuals:
     Min       1Q   Median       3Q      Max
-1613.07  -131.20    -9.76    95.63  2202.37

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -7.933e+02  2.246e+02  -3.532 0.000457 ***
X1           2.088e-02  1.330e-02   1.570 0.117075
X2          -2.434e-03  2.871e-04  -8.479 3.71e-16 ***
X3           2.165e+01  5.387e+00   4.020 6.88e-05 ***
X4           8.315e-01  5.761e+00   0.144 0.885304
X5           5.230e-01  2.536e-02  20.628  < 2e-16 ***
X6          -4.417e-04  7.337e-04  -0.602 0.547491
X7           6.374e+00  5.361e+00   1.189 0.235161
X8          -7.806e+00  9.265e+00  -0.843 0.399948
X9           1.554e-01  1.045e-02  14.876  < 2e-16 ***
X10          4.241e+01  1.975e+01   2.148 0.032311 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 357.5 on 428 degrees of freedom
Multiple R-squared:  0.9385,    Adjusted R-squared:  0.9371
F-statistic: 653.6 on 10 and 428 DF,  p-value: < 2.2e-16
```

III. Model Selection

Figure 3.1: Table for Model Selection

Model Selection Type	Selected Model	Value
Forward Model - AIC	Y ~ X1 + X2 + X3 + X5 + X9 + X10	-818.75397378
Backward Model - AIC	Y ~ X1 + X2 + X3 + X5 + X9 + X10	-818.75397378
Forward Backward Model - AIC	Y ~ X1 + X2 + X3 + X5 + X9 + X10	-818.75397378
Backward Forward Model - AIC	Y ~ X1 + X2 + X3 + X5 + X9 + X10	-818.75397378
Forward Model - BIC	Y ~ X2 + X3 + X5 + X9 + X10	-8.027981e+02
Backward Model - BIC	Y ~ X2 + X3 + X5 + X9 + X10	-8.027981e+02
Forward Backward Model - BIC	Y ~ X2 + X3 + X5 + X9 + X10	-8.027981e+02
Backward Forward Model - BIC	Y ~ X2 + X3 + X5 + X9 + X10	-8.027981e+02

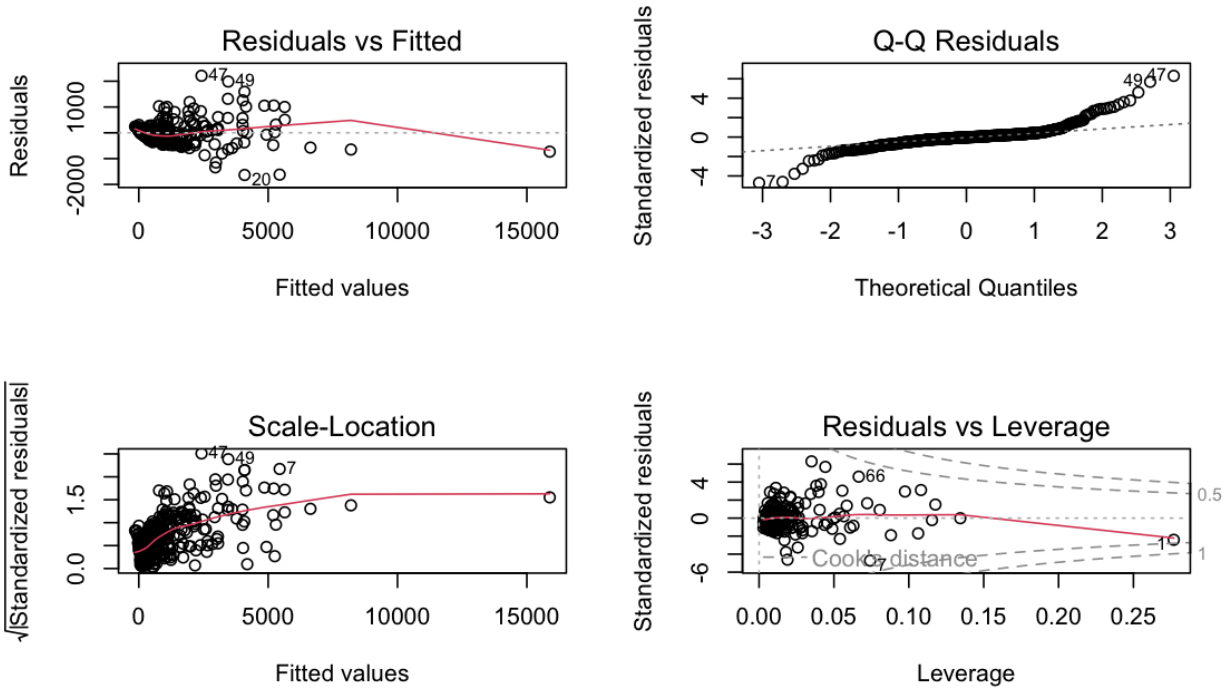
Figure 3.1: The table above shows the different model selections performed and the selected model based on the lowest AIC/BIC value.

Figure 3.2: Summary Table for Chosen Model: Y ~ X2 + X3 + X5 + X9 + X10

```
##
## Call:
## lm(formula = Y ~ X2 + X3 + X5 + X9 + X10, data = cdi2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1633.81  -127.22    -9.79    87.96  2210.85
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -8.028e+02  1.232e+02  -6.518 1.98e-10 ***
## X2          -2.381e-03  2.189e-04 -10.876  < 2e-16 ***
## X3           2.166e+01  4.090e+00   5.295 1.89e-07 ***
## X5           5.265e-01  2.096e-02  25.121  < 2e-16 ***
## X9           1.510e-01  8.544e-03  17.668  < 2e-16 ***
## X10          5.965e+01  1.726e+01   3.455 0.000604 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 357.4 on 433 degrees of freedom
## Multiple R-squared:  0.9379, Adjusted R-squared:  0.9371
## F-statistic:  1307 on 5 and 433 DF,  p-value: < 2.2e-16
```

IV. Model Diagnostics

Figure 4.1: Diagnostic Plots for Chosen Model: Y ~ X2 + X3 + X5 + X9 + X10



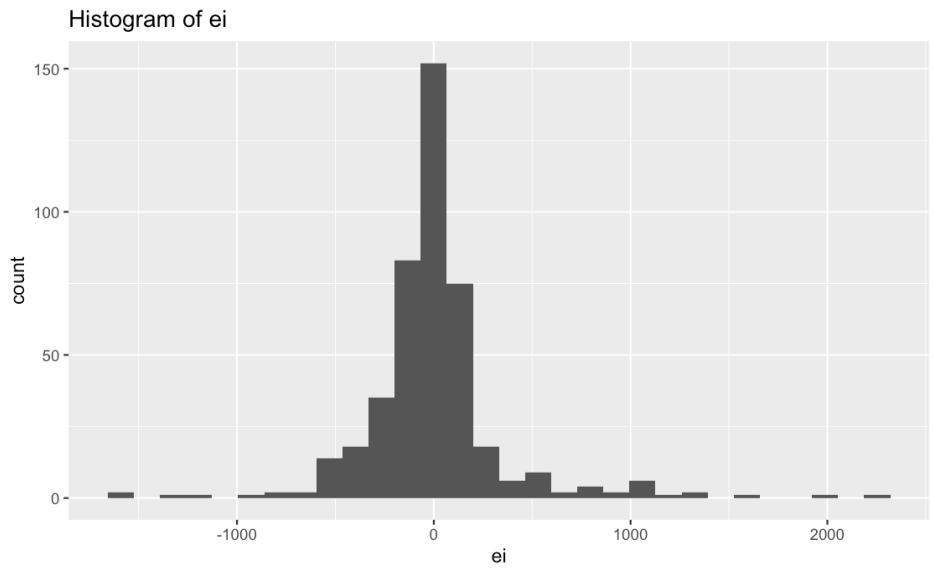


Figure 4.1: The diagnostics plots above are for Y ~ X2 + X3 + X5 + X9 + X10, which is the chosen model.

Figure 4.2: Hypothesis Tests

Hypothesis Tests:	
Shapiro-Wilks Test	H0: The population is normally distributed. HA: The population is not normally distributed.
Fligner-Killeen Test	H0: The variances of the two groups are equal. HA: The alternative hypothesis is that the variances of the two groups are not equal.

Figure 4.3: Leverage Values

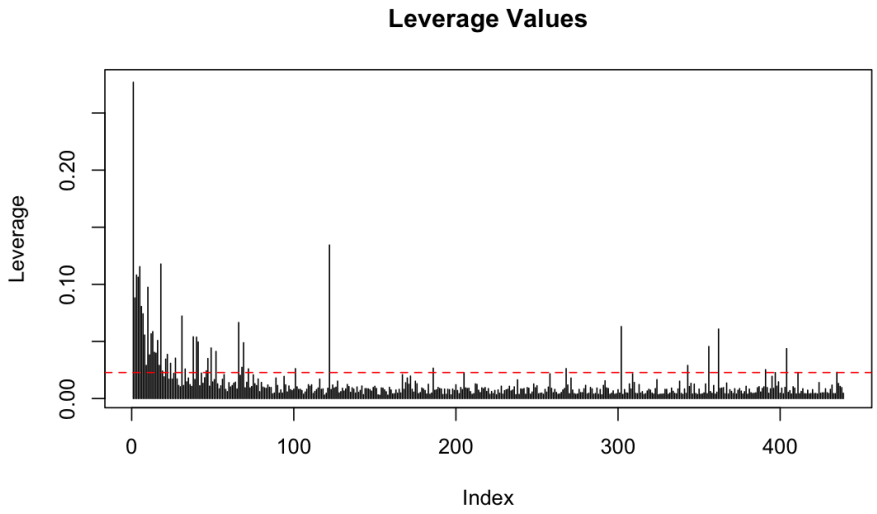


Figure 4.3: The figure above displays the leverage values for individual data points in the model Y ~ X2 + X3 + X5 + X9 + X10. The leverage measures how far an observation's predictor values are from the mean of the predictor values. The horizontal red line represents a threshold (2p/n), where points above are considered high-leverage points.

Figure 4.4: Cook's Distance

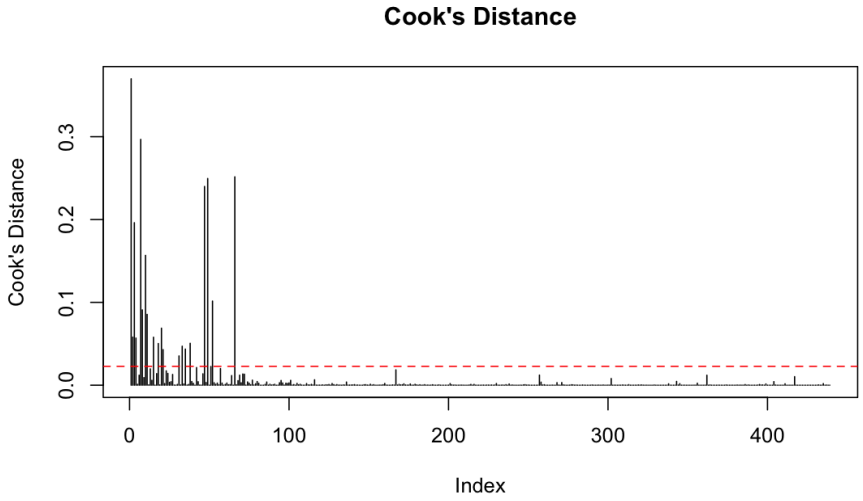


Figure 4.4: The graph above measures the influence of individual points for the chosen model: Y ~ X2 + X3 + X5 + X9 + X10. The red dashed line represents the threshold value (2p/n), where points above may be considered influential.

Figure 4.5: Influence Plot

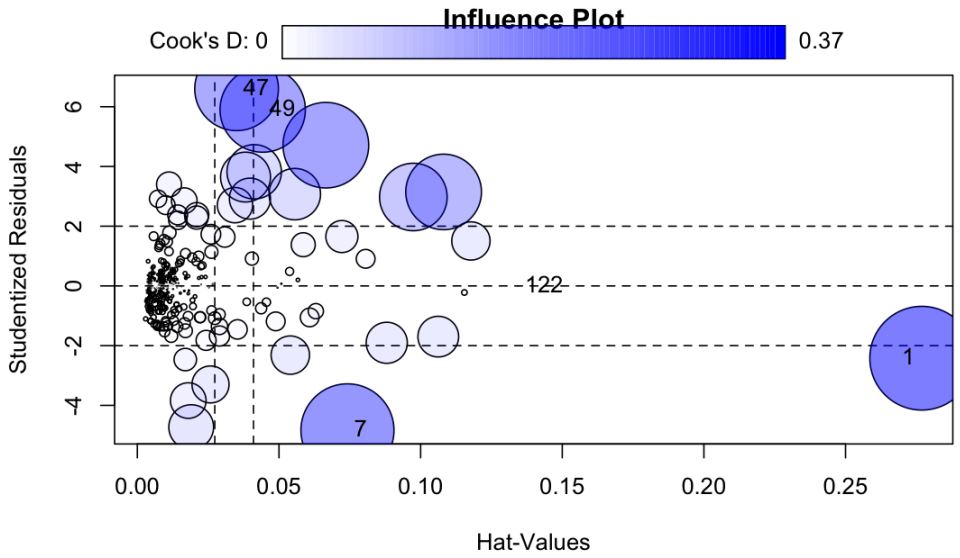


Figure 4.5: The plot above combines Cook's Distance, Leverage, and standardized residuals, where larger bubbles indicate more influence. Specific observations, such as "47" and "49" are labeled, identifying the impact and potential removal from the model.

Figure 4.6: Outlier Table

Figure 4.6.1: Raw Data:

Description: df [77 × 14]

	X1 <int>	X2 <int>	X3 <dbl>	X4 <dbl>	Y <int>	X5 <int>	X6 <int>	X7 <dbl>	X8 <dbl>
1	946	5105067	29.2	12.4	15153	21550	436936	11.1	7.2
3	4205	2498016	33.5	10.9	5905	6179	173821	8.1	6.1
7	614	2111687	27.4	12.5	3823	9490	193978	16.9	10.0
8	1945	1937094	27.1	13.9	6274	8840	244725	14.2	8.7
10	135	1585577	29.1	15.2	6641	10494	109148	16.1	8.0
11	2126	1507319	30.1	11.1	5280	4009	124959	5.0	4.6
15	824	1398468	31.7	12.5	5158	4152	35825	4.2	7.3
20	1209	1255488	25.3	20.7	2456	5543	107386	7.1	7.4
21	1247	1185394	29.5	9.9	3062	4086	133098	16.2	6.7
23	864	1170103	32.2	8.3	1677	3672	132495	8.2	6.6

1–10 of 77 rows | 1–10 of 14 columns

Previous 1 2 3 4 5 6 ... 8 Next

Figure 4.6.1: From the raw data, the tests identified 77 outliers.

Figure 4.6.2: Outliers Removed from Final Model:

	Y <int>	X2 <int>	X3 <dbl>	X5 <int>	X9 <int>	X10 <int>
2	7553	2818199	31.3	12449	55003	3
9	4718	1852810	32.6	6934	38911	3
51	2094	725956	27.8	2076	11179	4
53	705	717400	28.2	1202	15011	2
127	311	383545	26.4	860	3413	3

Figure 4.6.2: We decided to remove these 5 outliers from the final model and its dataset.

Figure 4.7: Variance Inflation Factor (VIF) and Correlation Matrix

```
##        X2        X3        X5        X9       X10
## 24.949577  1.009849  4.567092 18.204732  1.099171
```

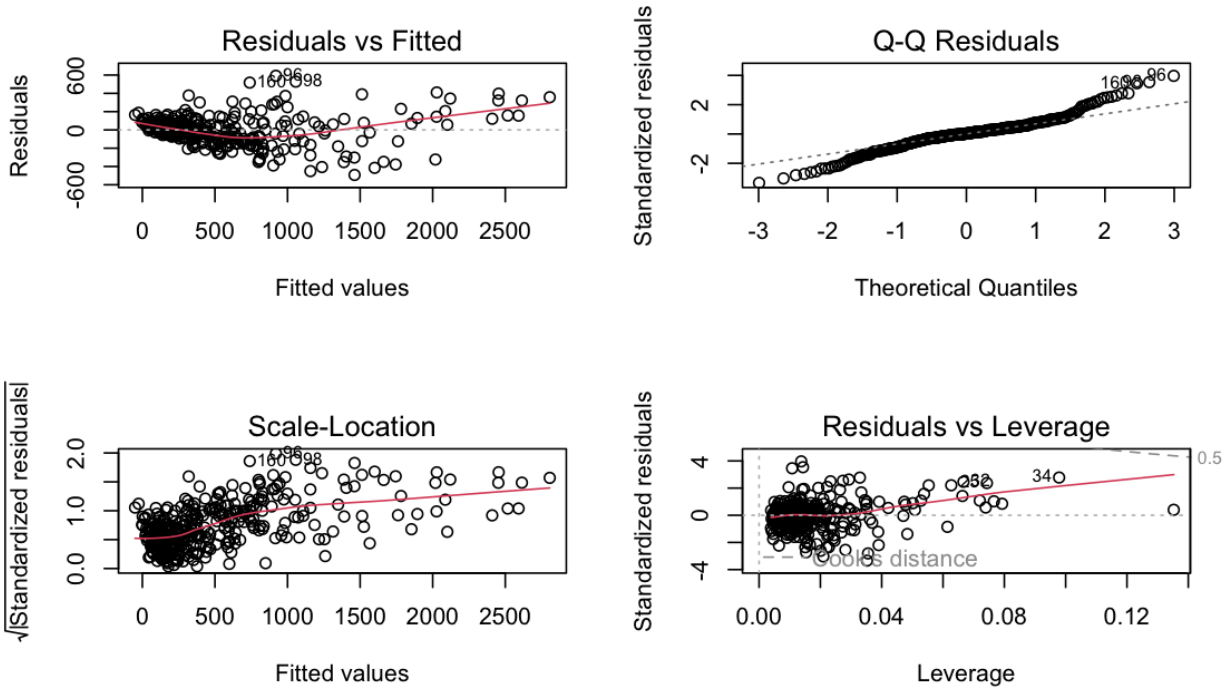
```
##              X2         X3          X5           X9           X10
## X2   1.00000000 0.08040312  0.87394520  0.9696189710  0.0501842308
## X3   0.08040312 1.00000000  0.07302039  0.0674787195  0.0492792712
## X5   0.87394520 0.07302039  1.00000000  0.8281272573 -0.0431375636
## X9   0.96961897 0.06747872  0.82812726  1.0000000000 -0.0005108083
## X10 0.05018423 0.04927927 -0.04313756 -0.0005108083  1.0000000000
```

Figure 4.7.1: The image above shows the VIF table for the chosen model: Y ~ X2 + X3 + X5 + X9 + X10, as well as the correlation matrix.

```
##       X2       X3       X5      X10
## 4.379486 1.008790 4.374699 1.038110
```

Figure 4.7.2: The image above shows the VIF table for the final model: Y ~ X2 + X3 + X5 + X10.

Figure 4.8: Diagnostic Plots (after removing outliers) for Final Model: Y ~ X2 + X3 + X5 + X10



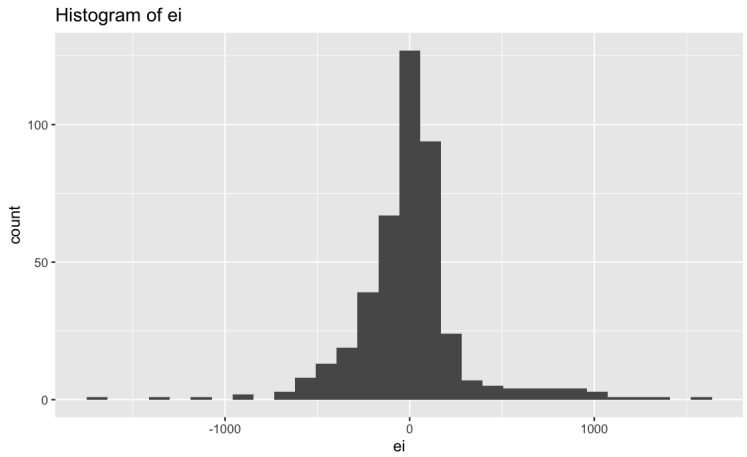


Figure 4.8: The diagnostic plots for the chosen model: Y ~ X2 + X3 + X5 + X10 after outliers were removed from the dataset and model.

Figure 4.9: Summary Table for Final Model: Y ~ X2 + X3 + X5 + X10

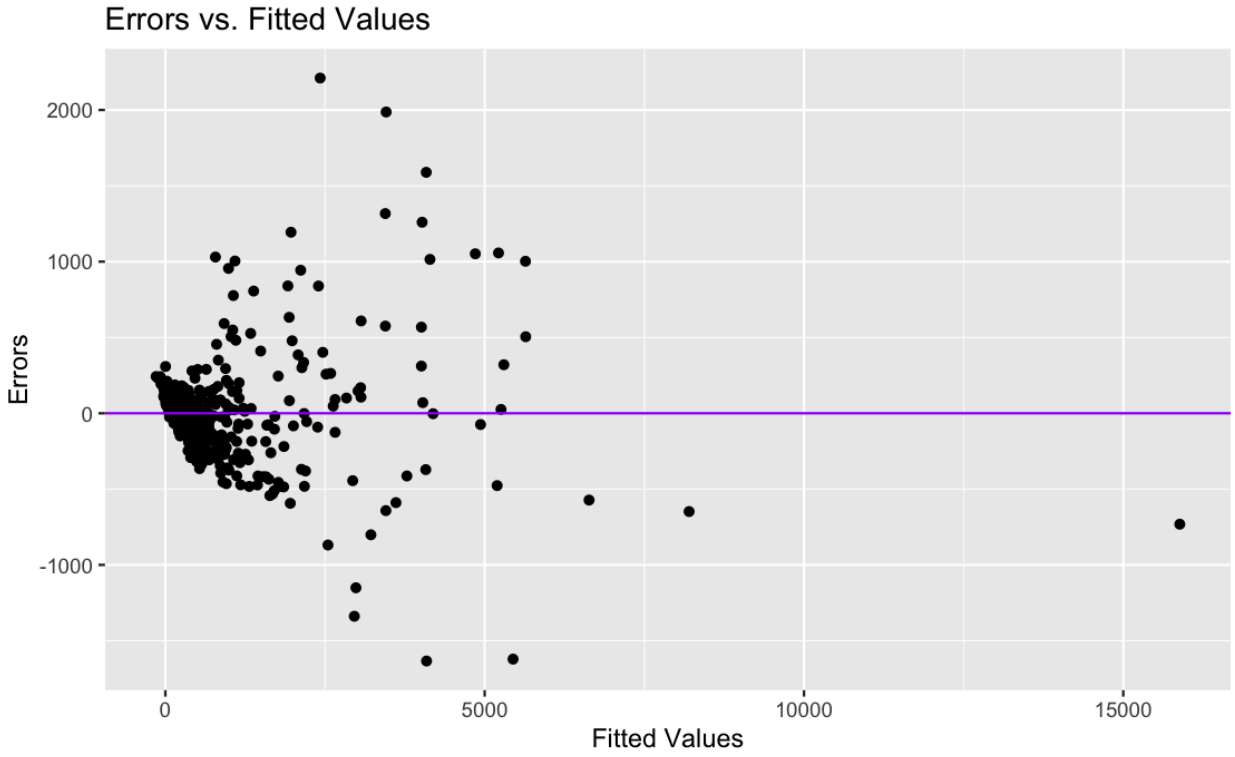
```
Call:
lm(formula = Y ~ X2 + X3 + X5 + X10, data = cleaned_data)

Residuals:
     Min       1Q   Median       3Q      Max
-1825.54  -163.10   -24.73    82.51  2422.57

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.303e+02  1.422e+02  -4.433 1.18e-05 ***
X2           1.229e-03  1.104e-04  11.137  < 2e-16 ***
X3           1.836e+01  4.732e+00   3.880 0.000121 ***
X5           4.459e-01  2.641e-02  16.882  < 2e-16 ***
X10         -1.854e+01  1.951e+01  -0.950 0.342418
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 413.2 on 429 degrees of freedom
Multiple R-squared:  0.8843,    Adjusted R-squared:  0.8832
F-statistic: 819.9 on 4 and 429 DF,  p-value: < 2.2e-16
```

Figure 4.10: Residual Plot before Removing Outliers for Chosen Model: Y ~ X2 + X3 + X5 + X9 + X10



V. Analysis

Figure 5.1: Hypothesis Test Results

Hypothesis Tests:	
Shapiro-Wilks Test	Before Outlier Removal: p-value = 2.2e-16 t-statistic: 0.81253 Conclusion: Reject null hypothesis After Outlier Removal: p-value = 2.2e-16 t-statistic: 0.765 Conclusion: Reject null hypothesis
Fligner-Killeen Test	Before Outlier Removal: p-value = 2.2e-16 t-statistic: 105.87 Conclusion: Reject null hypothesis After Outlier Removal:

	p-value = 2.2e-16 t-statistic: 104.83 Conclusion: Reject null hypothesis
Adjusted R^2	0.8832

Figure 5.2: Simultaneous Confidence Intervals for Final Model (90% confidence interval, alpha = 0.10): Y ~ X2 + X3 + X5 + X10

Variable	**Estimate**	**Lower Bound: 0.5%**	**Upper Bound: 99.5%**
(Intercept)	-6.303e+02	-9.981650e+02	-2.624180e+02
X2	1.229e-03	9.437908e-04	1.515011e-03
X3	1.836e+01	6.117961e+00	3.060262e+01
X5	4.459e-01	3.775715e-01	5.142502e-01
X10	-1.854e+01	-6.902142e+01	3.193524e+01

Figure 5.2: The table above displays the simultaneous confidence interval performed at alpha = 0.10.

R Appendix:

```
knitr::opts_chunk$set(echo = FALSE)
#II. EXPLORATORY DATA ANALYSIS (EDA)
library(tidyverse)
library(ggplot2)
library(leaps)
library(MASS)
library(car)

cdi2 <- read.csv("~/Downloads/CDI2.csv")
#Response = number of active physicians in a county (variable #5)
names(cdi2) = c("X1", "X2", "X3", "X4", "Y", "X5", "X6", "X7", "X8", "X9", "X10")

full.model = lm(Y~X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9 + X10, data = cdi2)
empty.model = lm(Y ~ 1, data = cdi2)
n = nrow(cdi2)
#Scatterplots
qplot(X1, Y, data = cdi2, xlab = "Land Area (Square Miles)",
      ylab = "Number of Active Physicians",
      main = "Relationship Between Number of Active
      Physicians (Y) and Land Area (X1)") + geom_smooth(method = 'lm') +
      theme(plot.title = element_text(size=12))
qplot(X2, Y, data = cdi2, xlab = "Total Population",
      ylab = "Number of Active Physicians",
      main = "Relationship Between Number of Active Physicians (Y) and
      Total Population (X2)") + geom_smooth(method = 'lm') +
      theme(plot.title = element_text(size=12))
qplot(X3, Y, data = cdi2,
      xlab = "Percent of Population Aged 18-34",
      ylab = "Number of Active Physicians",
      main = "Relationship Between Number of Active Physicians (Y) and
      Percent of Population Aged 18-34 (X3)") + geom_smooth(method = 'lm') +
      theme(plot.title = element_text(size=10.5))
qplot(X4, Y, data = cdi2, xlab = "Percent of Population 65 or Older",
      ylab = "Number of Active Physicians",
      main = "Relationship Between Number of Active Physicians (Y) and
      Percent of Population 65 or Older (X4)") + geom_smooth(method = 'lm') +
      theme(plot.title = element_text(size=10.5))
qplot(X5, Y, data = cdi2, xlab = "Number of Hospital Beds",
      ylab = "Number of Active Physicians",
      main = "Relationship Between Number of Active Physicians (Y) and
      Number of Hospital Beds (X5)") + geom_smooth(method = 'lm') +
      theme(plot.title = element_text(size=12))
qplot(X6, Y, data = cdi2, xlab = "Total Serious Crimes",
      ylab = "Number of Active Physicians",
      main = "Relationship Between Number of Active Physicians (Y) and
      Total Serious Crimes (X6)") + geom_smooth(method = 'lm') +
      theme(plot.title = element_text(size=12))
qplot(X7, Y, data = cdi2, xlab = "Percent Below Povery Level",
      ylab = "Number of Active Physicians",
      main = "Relationship Between Number of Active Physicians (Y) and
      Percent Below Povery Level (X7)") + geom_smooth(method = 'lm') +
      theme(plot.title = element_text(size=11))
qplot(X8, Y, data = cdi2, xlab = "Percent Unemployment",
```

```
      ylab = "Number of Active Physicians", main = "Relationship Between
      Number of Active Physicians (Y) and Percent Unemployment (X8)") +
      geom_smooth(method = 'lm') + theme(plot.title = element_text(size=12))
qplot(X9, Y, data = cdi2, xlab = "Total Personal Income (Millions of Dollars)",
      ylab = "Number of Active Physicians",
      main = "Relationship Between Number of Active Physicians (Y) and
      Total Personal Income (X9)") + geom_smooth(method = 'lm') +
      theme(plot.title = element_text(size=12))
ggplot(data = cdi2, aes(x = X10, y = Y)) +
      geom_bar(stat = "identity", fill = "blue") +
      labs(x = "Geographic Region (1 = NE, 2 = NC, 3 = S, 4 = W)",
      y = "Number of Active Physicians",
      title = "Relationship Between Avg. Number of Active Physicians (Y)
      and Geographic Region (X10)") + theme(plot.title = element_text(size=11))

#Mean and Standard Deviation of each Variable

# Calculating means
means <- apply(cdi2[, c("X1", "X2", "X3", "X4", "X5", "X6", "X7", "X8", "X9",
                        "X10")], 2, mean, na.rm = TRUE)

# Calculating standard deviations
sds <- apply(cdi2[, c("X1", "X2", "X3", "X4", "X5", "X6", "X7", "X8", "X9",
                      "X10")], 2, sd, na.rm = TRUE)

# Combine results into a data frame
summary_table <- data.frame(
  Variable = colnames(cdi2[, c("X1", "X2", "X3", "X4", "X5", "X6", "X7", "X8",
                               "X9", "X10")]),
  Mean = means,
  Standard_Deviation = sds
)
summary_table
#III. MODEL SELECTION

#Forward Step-wise Regression
forward.model.AIC = stepAIC(empty.model, scope = list(lower = empty.model,
              upper= full.model), k = 2,direction = "forward",trace = FALSE)
forward.model.BIC = stepAIC(empty.model,  scope = list(lower = empty.model,
              upper= full.model), k = log(n),trace=FALSE,direction = "forward")

#Backward Stepwise selection
backward.model.AIC = stepAIC(full.model, scope = list(lower = empty.model,
              upper= full.model), k = 2,direction = "backward",trace = FALSE)
backward.model.BIC = stepAIC(full.model,  scope = list(lower = empty.model,
              upper= full.model), k = log(n),trace=FALSE,direction = "backward")

#Forward Backward Stepwise Selection
FB.model.AIC = stepAIC(empty.model, scope = list(lower = empty.model,
              upper= full.model), k = 2,direction = "both",trace = FALSE)
FB.model.BIC = stepAIC(empty.model,  scope = list(lower = empty.model,
              upper= full.model), k = log(n),trace=FALSE,direction = "both")
```

```r
#Backward Forward Stepwise Selection
BF.model.AIC = stepAIC(full.model,
                       scope = list(lower = empty.model, upper= full.model),
                       k = 2,direction = "both",trace = FALSE)
BF.model.BIC = stepAIC(full.model,
                       scope = list(lower = empty.model, upper= full.model),
                       k = log(n),trace=FALSE,direction = "both")
forward.model.AIC$coefficients
backward.model.AIC$coefficients
FB.model.AIC$coefficients
BF.model.AIC$coefficients
forward.model.BIC$coefficients
backward.model.BIC$coefficients
FB.model.BIC$coefficients
BF.model.BIC$coefficients
#Summary Plots and Numerical Summaries
best.model <- lm(Y ~ X2 + X3 + X5 + X9 + X10, data = cdi2)
summary(best.model)
#IV. MODEL DIAGNOSTICS
plot(best.model)
best.data <- model.frame(best.model)
best.data$ei <- resid(best.model)

#Plot ei vs. Xi
plot(best.model$fitted.values, resid(best.model),
     xlab = "Fitted Values", ylab = "Residuals",
     main = "Residuals vs. Fitted Values")
abline(h = 0, col = "red")

#Histogram of ei
ggplot(best.data, mapping = aes(x = ei)) + geom_histogram() +
  ggtitle("Histogram of ei")

#Assessing Normality
cdi2$ei <- best.model$residuals
cdi2$yhat <- best.model$fitted.values

#Assessing constant variance (homoscedasticity)
library(ggplot2)
qplot(yhat, ei, data = cdi2) +  ggtitle("Errors vs. Fitted Values") +
  xlab("Fitted Values") +
  ylab("Errors") + geom_hline(yintercept = 0,col = "purple")
#Detect Outliers
# Calculate residuals
residuals <- rstandard(best.model)  # Standardized residuals
outliers <- which(abs(residuals) > 2)
outliers

# Boxplot to detect outliers
boxplot(cdi2$Y,
        main = "Boxplot of Average Number of Active Physicians",
        xlab = "Number of Active Physicians",
        horizontal = TRUE,
```

```
        col = "lightblue",
        border = "darkblue",
        notch = TRUE)

#Detect High Leverage Points
leverage <- hatvalues(best.model)
p <- length(coef(best.model)) - 1  # Number of predictors
n <- nrow(cdi2)  # Number of observations
leverage_threshold <- 2 * (p) / n # some other may use 2(p+1)/n
high_leverage_points <- which(leverage > leverage_threshold)
high_leverage_points
plot(leverage, type = "h", main = "Leverage Values", xlab = "Index",
     ylab = "Leverage")
abline(h = leverage_threshold, col = "red", lty = 2)

#Detect Influential Points

# Calculate Cook's Distance
cooks_d <- cooks.distance(best.model)
cooks_threshold <- (2 * p) / n # 2p/n
influential_points <- which(cooks_d > cooks_threshold)
influential_points
plot(cooks_d, type = "h", main = "Cook's Distance", xlab = "Index",
     ylab = "Cook's Distance")
abline(h = cooks_threshold, col = "red", lty = 2)

# Calculate DFFITS
dffits_values <- dffits(best.model)
dffits_threshold <- 2 * sqrt((p + 1) / n)

influential_dffits <- which(abs(dffits_values) > dffits_threshold)
influential_dffits

dfbetas_values <- dfbetas(best.model)
dfbetas_threshold <- 2 / sqrt(n)

# Identify influential points for each coefficient
influential_dfbetas <- apply(dfbetas_values, 2,
                             function(x) which(abs(x) > dfbetas_threshold))
influential_dfbetas
covratio_values <- covratio(best.model)
covratio_threshold_upper <- 1 + 3 * p / n
covratio_threshold_lower <- 1 - 3 * p / n

# Identify influential points using COVRATIO
influential_covratio <- which(covratio_values <
          covratio_threshold_lower | covratio_values > covratio_threshold_upper)
influential_covratio

all_outliers <- unique(c(
  outliers,
  high_leverage_points,
  influential_points,
```

```r
  influential_dffits,
  unlist(influential_dfbetas),
  influential_covratio
))
cdi2[all_outliers, ]

# Influence plot (combines Cook's distance, leverage, and residuals)
library(car)
influencePlot(best.model, id.method = "identify", main = "Influence Plot")
remove <- as.numeric(rownames(influencePlot(best.model, id.method = "identify",
                                            main = "Influence Plot")))
remove

cleaned_data <- cdi2[-remove, ]
cleaned_data

# Fit the model again with the cleaned dataset
cleaned_model <- lm(Y ~ X2 + X3 + X5 + X9 + X10, data = cleaned_data)

# Summary of the cleaned model
summary(cleaned_model)

#Create outlier table
cleaned.data <- model.frame(cleaned_model)
cleaned.data

outlier_table <- cleaned.data[remove, ]
outlier_table

# Diagnostic plots for the final model
par(mfrow = c(2, 2))
plot(cleaned_model)
#Histogram of ei (after removing outliers)
ggplot(cleaned_data, mapping = aes(x = ei)) + geom_histogram() +
  ggtitle("Histogram of ei")

# Add DFBETAS if needed
dfbetas_table <- apply(dfbetas_values, 2, function(x) x[outliers])
colnames(dfbetas_table) <- paste0("DFBETAS_", colnames(dfbetas_values))

#Assessing Normality Post-Outlier Removal
cleaned_data$ei <- cleaned_model$residuals
cleaned_data$yhat <- cleaned_model$fitted.values

#Variance Inflation Factor (VIF) and Correlation Matrix
vif(cleaned_model)
matrix_correlation <- cor(cleaned_data[,c("X2","X3","X5","X9","X10")])
matrix_correlation
final_model <- lm(Y ~ X2 + X3 + X5 + X10, data = cleaned_data)
vif(final_model)

#Value greater than 5: potentially severe correlation -> found for X2, X5, X9
final_data <- model.frame(final_model)
```

```r
summary(final_model)

#V. ANALYSIS

#Simultaneous CIs
alpha =0.10
SCI =confint(final_model,level = 1-alpha/ (2 * p))
SCI

#Testing Normality (Shapiro-Wilks) - BEFORE REMOVING OUTLIERS
ei = best.model$residuals
the.SWtest = shapiro.test(ei)
the.SWtest


#Fligner Killeen Test (FK Test) - BEFORE REMOVING OUTLIERS
Group = rep("Lower",nrow(cdi2)) #Creates a vector that repeats "Lower" n times
Group[cdi2$Y < median(cdi2$Y)] = "Upper" #Changing the appropriate values to "Upper"
Group = as.factor(Group) #Changes it to a factor, which R recognizes as a grouping variable.
cdi2$Group = Group
the.FKtest= fligner.test(cdi2$ei, cdi2$Group)
the.FKtest
#Testing Normality (Shapiro-Wilks) - ON FINAL MODEL
ei = final_model$residuals
final_data$ei <- ei
the.SWtest = shapiro.test(ei)
the.SWtest

#Fligner Killeen Test (FK Test) - ON FINAL MODEL
Group = rep("Lower",nrow(final_data)) #Creates a vector that repeats "Lower" n times
Group[cleaned_data$Y < median(final_data$Y)] = "Upper" #Changing the appropriate values to "Upper"
Group = as.factor(Group) #Changes it to a factor, which R recognizes as a grouping variable.
final_data$Group = Group
the.FKtest= fligner.test(final_data$ei, final_data$Group)
the.FKtest
```