A Regression Analysis of Physician Distribution

Gunica Sharma

Introduction

The project analyzes the relationship between demographic, economic, healthcare, and geographic factors and the number of active physicians in counties across the United States. We will be using the CDI2 dataset, which provides data on the number of active physicians (response variable) and other county-level variables such as land area, population demographics, economic indicators, and healthcare resources. We added a header to the dataset and renamed the columns accordingly to our response variable (Y) and explanatory variables (X1-X10). Our main goal is to develop the "best" multiple linear regression model to predict the response variable number of active physicians in counties (Y) using our multitude of demographic, economic, healthcare, and geographic explanatory variables (X1-X10). This analysis aims to identify the key predictors of physician distribution and evaluate the strength and significance of these relationships. By doing so, it can be better understood how demographic, economic, healthcare, and geographic characteristics predict the availability of physicians in counties, which is critical for resource planning and addressing disparities in healthcare access.

Exploratory Data Analysis

We ran a few preliminary analyses on the dataset. First we ran numerical summaries on each of the variables. We created a Figure 1.1 summarizing all of the variables with their respective mean, median, quartiles, standard deviations, and minimum and max values. There are a few notable features that stand out. For the X2 values it shows that there is a wide range of population from around 100,000 to over 8 million. There is also skewness in the total personal income (X9) since the range is very large for this category. These variables show skewed distributions to the right suggesting the presence of outliers. This seems consistent with other variables such as Hospital Resources (X6), Elderly Population (X4) along with the Crime and the Poverty Rates (X7 & X8). Additionally we created scatter plots of our response variable vs our different explanatory variables to see if we could identify outliers and see a linear relationship. From the graphs as seen in Figures 1.2 - 1.10 variables such as X2 (total population), X5 (number of hospital beds), X6 (total serious crimes), and X9 (total personal income) show a linear relationship while the other variables do not show a statistically significant relationship from the regression line. We can also see that other variables have a lot of outliers such as X1 (land area), X3 & X4 (percentage of population ages), and X7 & X8 (poverty and crime rates). These are key points that we noted to better understand our data as we continue with our analysis.

Model Selection

Since our goal is prediction, developing the "best" multiple linear regression model to predict the number of active physicians in a county (Y), we will run different regressions to determine which is the best model. We first ran an all-subsets regression, which gives a very large output as seen in Figure 2.0. Our output showed that explanatory variables X5 (number of hospital beds) and X9 (total personal income) are the top models of each size–having a large influence/statistical significance on the final model. This is backed up by the visualization scatter plots we got from our exploratory data analysis, where X5 (number of hospital beds) and X9 (total personal income) had a positive linear relationship with Y. However, since we have a large number of predictors given in our dataset, we will further our model selection by using forward, backward, forward-backward, and backward-forward step-wise regression. The AIC model (Figure 2.1) returns with X2, X3, X5, X6, X7, X8, X9, and X10. Since X10 aka the geographic regions is a categorical variable that is split into $k = 4$ levels, the final model will have $k - 1 = 3$ dummy variables. The BIC model (Figure 2.2) returns with X2, X3, X5, X9, and X10. In other words, both models exclude X1 and X4. The BIC model additionally excludes X6, X7, and X8. Now that we have narrowed the model down to 2 models using AIC and BIC, we will determine the best model by further examining the model criteria using the PRESS and adjusted R^2. By running through all criteria, our final model selection is based on the AIC model because it demonstrates higher prediction accuracy with lower AIC, lower PRESS, and higher adjusted R^2 compared to the BIC model (Figure 2.3). Thus, with the coefficients calculated by R, our final best linear model is $\hat{Y} = -661.988 - 0.00191X_2 + 20.531X_3 + 0.499X_5 - 0.00114X_6 + 11.903X_7 - 15.819X_8 + 0.143X_9 - 84.753X_{10,2} - 35.699X_{10,3} + 152.019X_{10,4}$.

Model Diagnostics

We will run diagnostics using the best final linear model we selected before and after removing outliers, high lineage points, and influential points from the dataset. The three assumptions for this model are normality, constant variance, and linearity. The null and alternative hypotheses for normality tests is H_0: the residuals follow a normal distribution vs. H_a: the residuals do not follow a normal distribution. The null and alternative hypotheses for constant variance is H_0: the variance of the residuals is constant across all levels of fitted values (homoscedasticity) vs. H_a: the variance of the residuals changes across fitted values (heteroscedasticity).

Original Dataset To assess normality, we used the Normal Q-Q plot and Shapiro-Wilks Normality test. The points in the Q-Q plot (Figure 3.1) followed a

straight line approximately between the theoretical quantiles -2 to 2. Outside of the range, the points deviated significantly from the straight line, most likely from outliers. This suggests that the residuals is approximately normal within the central portion of the distribution but has issues in the tails. The S-W test produced a p-value<0.00000000000000022. This is less than the significance level $\alpha=0.05$, so we reject the null hypothesis and conclude the residuals do not follow a normal distribution. For the linearity and constant variance (homoscedasticity), we refer to the Errors vs. Fitted Values plot (Figure 3.2) and Fligner-Killeen test of homogeneity of variances. There is a cluster of points in the left that gradually increases in variance across the fitted values axis. There are very distinguishable outliers past ~5000+ in the fitted value axis. The residuals' spread is in the shape of an increasing funnel shape, violating the assumption of constant variance. The points are not random, so this violates the assumption of linearity too. The F-K test produced a p-value< 0.00000000000000022, which is less than $\alpha=0.05$, rejecting the null hypothesis again.

Cleaned Dataset To compare our diagnostics with the original data, we cleaned our dataset to remove rows containing observations with outliers, high leverage points, and influential points (which may significantly affect and change the regression model). By using different detection methods and R-functions, we were able to remove a total of 53 unique observations/rows from the dataset found in Figure 3.3. We ran the same diagnostics on our cleaned dataset. The Q-Q plot now (Figure 3.4) shows a y-axis of Sample Quantiles with a smaller range of -500 through 1000, compared to the previous Q-Q plot from the original dataset of -1000 through 2000. Normality significantly improved, with the points in the central portion following the straight line the closest, but the points at the tails are not as far from the straight line, except the top right. Visually, the points look approximately normal like last time. The S-W test produced a p-value = 0.0000000000003933, which slightly worsened, but is still very small–rejecting the null hypothesis and concluding that the residuals are not normally distributed. The new Errors vs. Fitted Values (Figure 3.5) shows a clearer picture of the spread of the residuals. The fitted values now range from 0 through 5000. The observation still remains the same as last time: there is an increasing funnel shape with the points increasing in variance across the fitted values. Therefore, although the variance has improved and become more constant, it is still not constant variance. The points appear more spread out and random, but still suggest non-linearity. The F-K test produces a p-value<0.00000000000000022, rejecting the null and concluding with non-constant variance.

Analysis and Interpretation

Following the diagnostics, we started our analysis and interpretation of the dataset using our data that had been cleaned up by removing outliers and such. Figure 4.1 shows the different analysis we ran. First was our final model we obtained, $\hat{Y} = -400.375 - 0.00151X_2 + 13.643X_3 + 0.460X_5 - 0.00121X_6 + 5.057X_7 - 16.571X_8 + 0.113X_9 - 32.766X_{10,2} - 2.398X_{10,3} + 124.635X_{10,4}$. From this we determined that the intercept represents the predicted number of active physicians when all predictor variables are 0 but, since having a negative number of physicians is not possible in practice, the intercept does not carry practical meaning in this context. The coefficient −0.00151 for X2 suggests that for every additional person in the population, the number of active physicians decreases by 0.00151, holding all other variables constant. This small negative effect might indicate inefficiencies or other factors affecting healthcare availability as populations increase. The coefficient 13.643 implies that for every 1% increase in the percentage of individuals aged 18–34, the number of active physicians increases by approximately 13.64, suggesting that younger individuals positively correlate to better healthcare. The coefficient 0.460 indicates that for every additional hospital bed, the number of active physicians increases by 0.46, suggesting that greater hospital infrastructure means more physicians. For X6 (number of serious crimes), the coefficient −0.00121 indicates a decrease of 0.00121 active physicians per new crime, this shows that higher crime means less doctors. The coefficient for X7 (percentage below the poverty level) is 5.057, suggesting that a 1% increase in poverty correlates with an increase of 5.06 active physicians, suggesting that lower income people may have more healthcare in their area. For X8 (percentage unemployed), the coefficient −16.571 suggests that a 1% increase in unemployment correlates with a decrease of 16.57 active physicians, suggesting that less doctors are in areas with more unemployment. The coefficient for X9 (total personal income) is 0.113, meaning that for every additional million dollars in total personal income, the number of active physicians increases by 0.113, showing the positive relationship between wealth and healthcare. Lastly, the geographic region variables (X10) indicate regional differences relative to the reference category. Counties in the North Central region (X10,2) have approximately 32.77 fewer physicians compared to the reference region, while those in the South (X10,3) have about 2.40 fewer physicians. In contrast, counties in the West (X10,4) have approximately 124.64 more active physicians than the reference region, showing significant regional variation in physician distribution.

Next we ran our confidence intervals at 95% for each of the variables for the intercept we obtained that the true intercept value is between (-600.945, -199.805)

and for each of the variables in the model we obtained values seen in Figure 4.2. From these values we can determine which variables are significant based on if the interval contains 0, as seen by our data X_6 ,X_7 ,$X_{10,2}$,$X_{10,3}$ contains 0 which suggest that these variables are likely not significant. For our null hypothesis we stated that H_0 is equal to 0 for all of the coefficients while our alternative H_A was that at least one of the coefficients is not equal to 0. Next we calculated the test-statistic for which we got 497.3 with 376 degrees of freedom. Since we obtained a relatively large test statistic we can say that the predictors are significantly associated with the response variable. For the p-value of the model we obtained that it is less than 2.2e-16 and with an α value of less than 0.05 we can easily say that our model is statistically significant and we can reject our null hypothesis. We also calculated the t and p-values for each of the variables and the ones that were significant were the intercept, X2, X3, X5, X8, X9 and X104 these suggest that we can reject the null hypothesis since the corresponding values as seen from Figure 4.1. These are less than our significance threshold which means those variables contribute significantly to our final linear model and the t-values are large suggesting a significant predictor. The residual standard error is 189.9 on 376 degrees of freedom which suggests that it's a relatively good model fit. The multiple and adjusted R^2 values are 0.9297, 0.9278 respectively and since they are above 0.90 or 90% that means our model is strong. So our final model with the CI considerations is $\hat{Y}$ = -400.375 - 0.00151X_2 + 13.643X_3 + 0.460X_5 - 16.571X_8 + 0.113X_9 + 124.635$X_{10,4}$. This analysis matches up well with our initial considerations in the exploratory data analysis as variables X2, X5, X6, and X9 could be visually seen to have a linear relationship.

Conclusion

In this analysis of the CDI dataset we developed a multiple linear regression model in order to predict the number of physicians in different counties using multiple different explanatory variables relating to demographics, economics and regions. We found that several predictors, such as total population, percentage of young adults, number of hospital beds, unemployment rate, and regional differences, significantly impact physician distribution. Our model was able to show that there is truth in different socioeconomic and geographic factors impacting healthcare. One limitation we thought of was that since we are under a linear assumption, we may not be getting a full picture of the model by assuming a simple linear model between the variables, additionally we have no way of knowing the quality of the data and if our data has all of the variables needed to understand all of the relationships. To improve our model, we could check the data quality, explore more variables, and even look at non linear relationships and see if they are a better model. We also did not consider how the variables interact with each other apart from our response variable, perhaps relationships there may be explored for a more comprehensive model. For example, the impact of poverty on the number of physicians might differ based on geographic region and that would need to be explored. Future analysis could target these areas for an even more robust model.

Figures and Graphs

Figure 1.1

Variable	Mean	Median	Q1	Q3	Std_Dev	Min	Max
X1	1.041411e+03	656.50	451.250	946.750	1.549922e+03	15.0	20062.0
X2	3.930109e+05	217280.50	139027.250	436064.500	6.019870e+05	100043.0	8863164.0
X3	2.856841e+01	28.10	26.200	30.025	4.191083e+00	16.4	49.7
X4	1.216977e+01	11.75	9.875	13.625	3.992666e+00	3.0	33.8
X5	1.458627e+03	755.00	390.750	1575.750	2.289134e+03	92.0	27700.0
X6	2.711162e+04	11820.50	6219.500	26279.500	5.823751e+04	563.0	688936.0
X7	8.720682e+00	7.90	5.300	10.900	4.656737e+00	1.4	36.3
X8	6.596591e+00	6.20	5.100	7.500	2.337924e+00	2.2	21.3
X9	7.869273e+03	3857.00	2311.000	8654.250	1.288432e+04	1141.0	184230.0

Figure 1.2

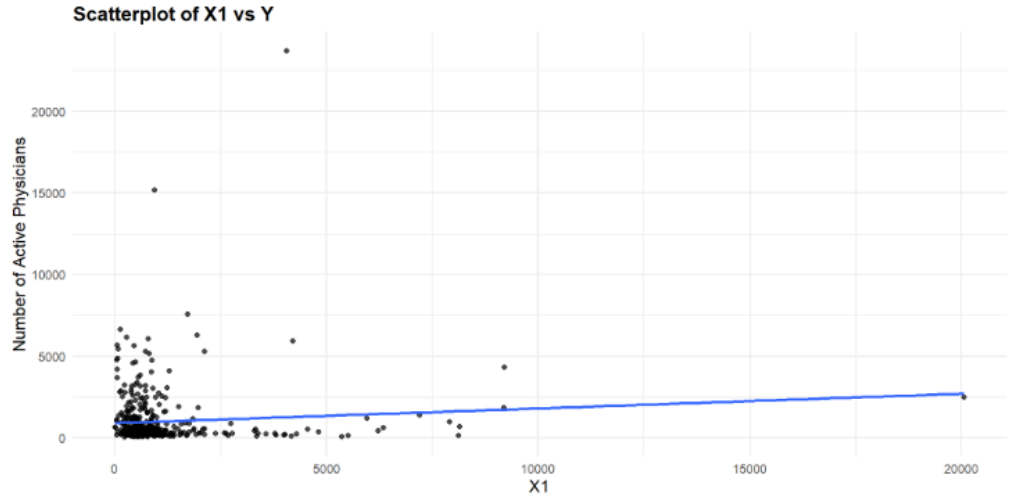


Figure 1.3

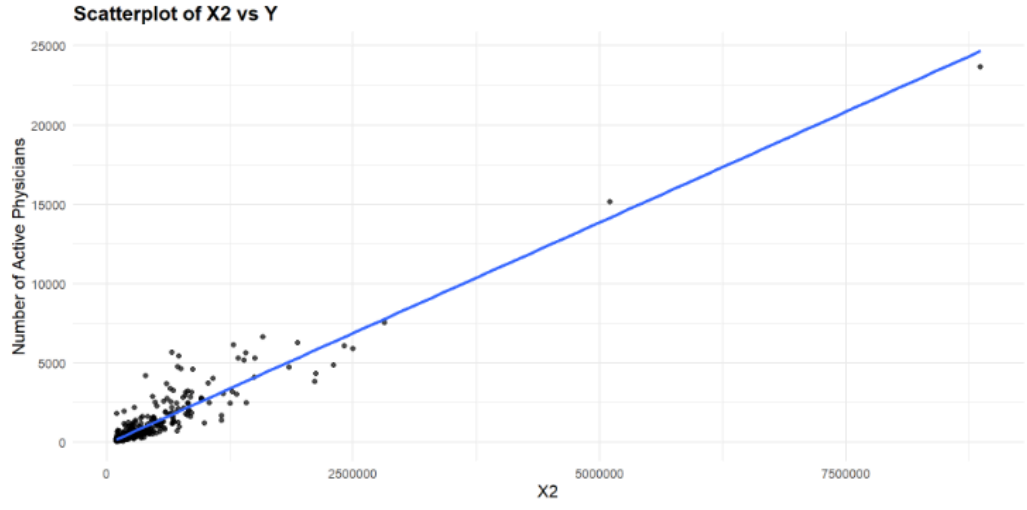


Figure 1.4

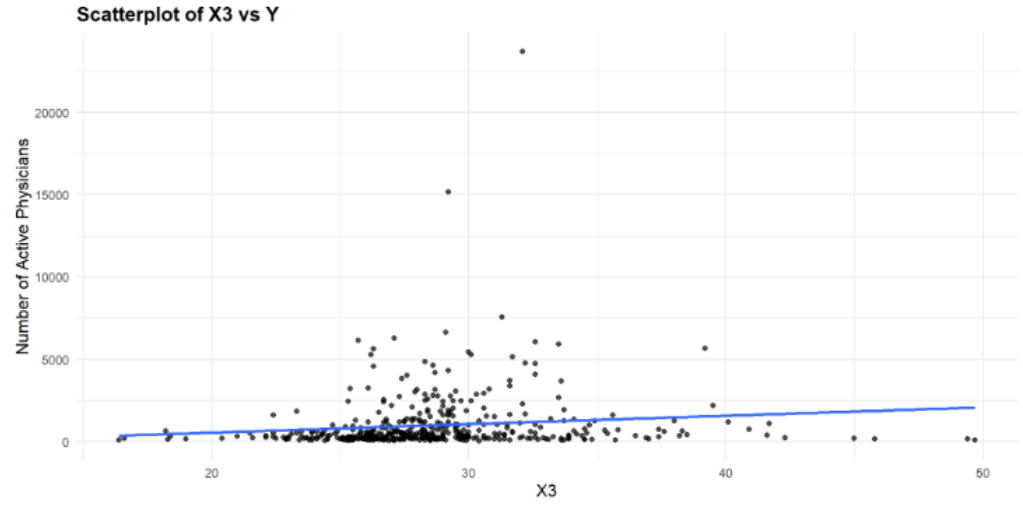


Figure 1.5

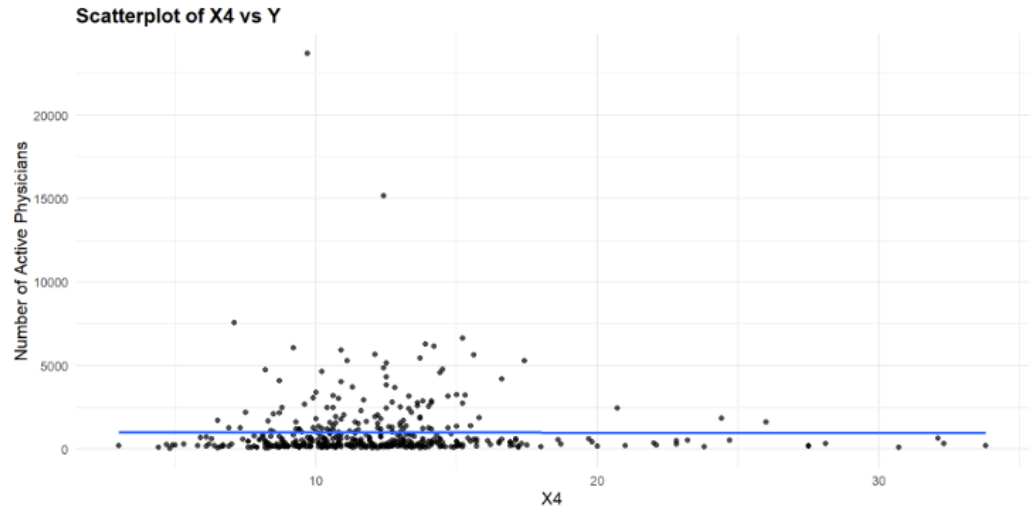


Figure 1.6

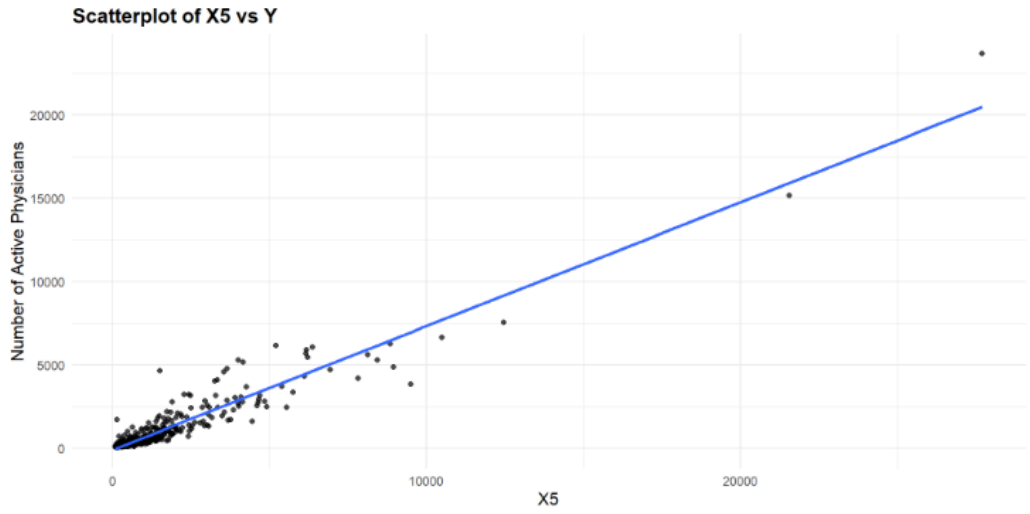


Figure 1.7

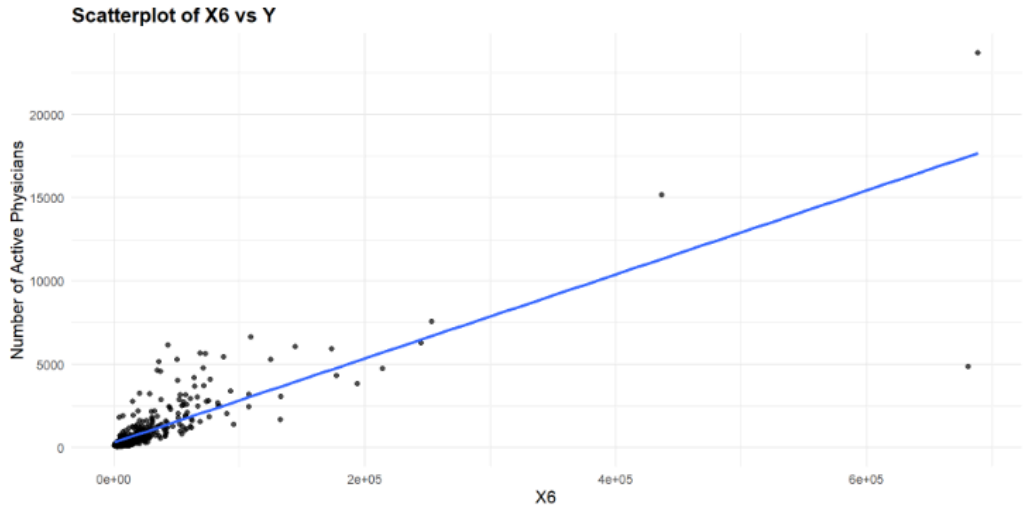


Figure 1.8

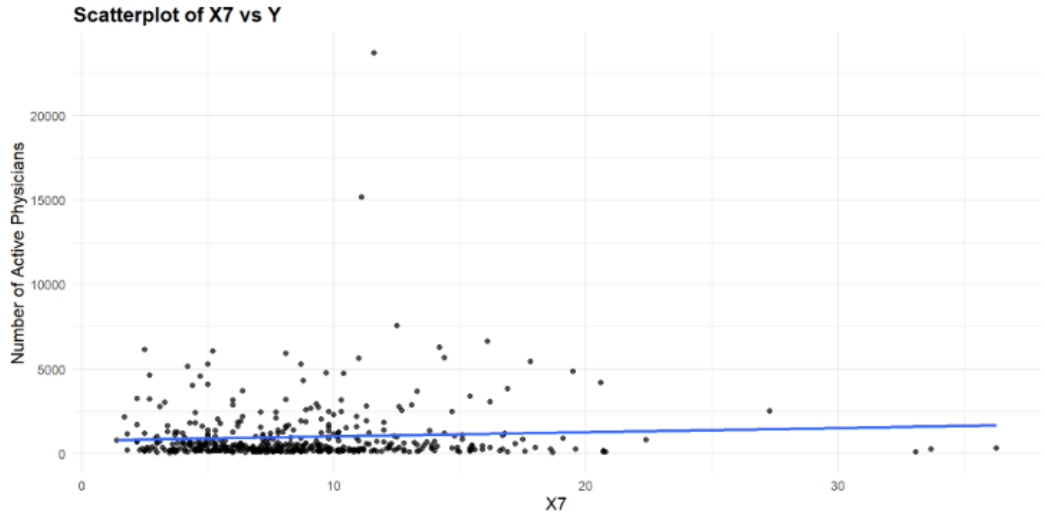


Figure 1.9

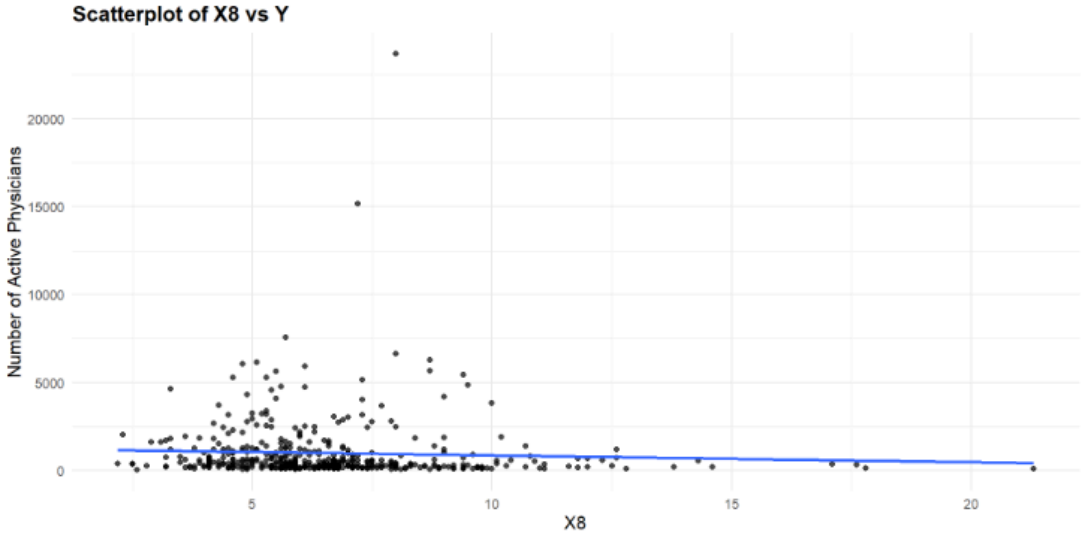


Figure 1.10

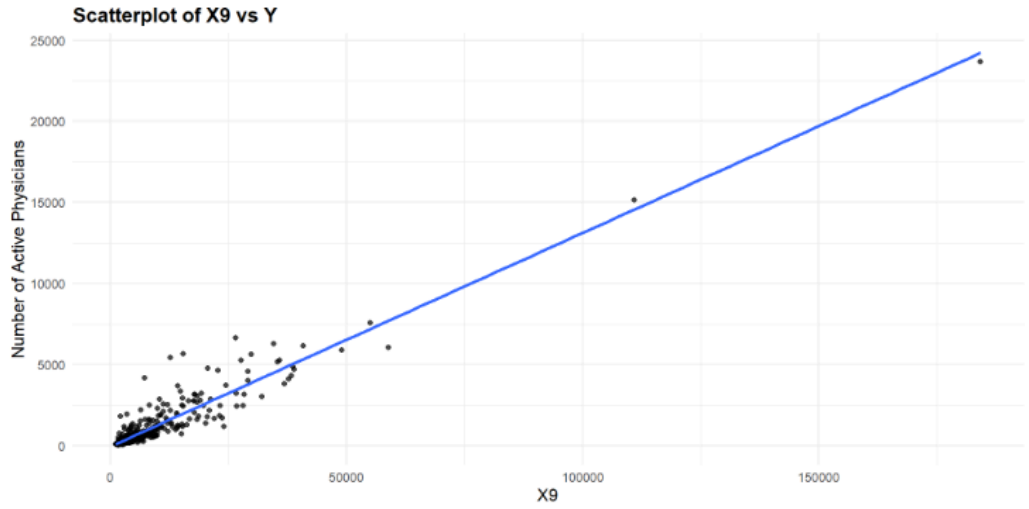

Scatterplot of X9 vs Y
Number of Active Physicians
25000
20000
15000
10000
5000
0
0
50000
100000
150000
X9

Figure 2.0

```
##                           model p     adjR2         CP       BIC
## 1                           Y,X5 2 0.9031620 609.237813 -1016.105
## 2                           Y,X5 2 0.9473357 609.237813 -1016.105
## 3                           Y,X5 2 0.9549641 609.237813 -1016.105
## 4                           Y,X5 2 0.9574870 609.237813 -1016.105
## 5                           Y,X5 2 0.9590979 609.237813 -1016.105
## 6                           Y,X5 2 0.9592780 609.237813 -1016.105
## 7                           Y,X5 2 0.9593880 609.237813 -1016.105
## 8                           Y,X5 2 0.9595137 609.237813 -1016.105
## 9                         Y,X5,X9 3 0.9031620 133.142853 -1279.029
## 10                        Y,X5,X9 3 0.9473357 133.142853 -1279.029
## 11                        Y,X5,X9 3 0.9549641 133.142853 -1279.029
## 12                        Y,X5,X9 3 0.9574870 133.142853 -1279.029
## 13                        Y,X5,X9 3 0.9590979 133.142853 -1279.029
## 14                        Y,X5,X9 3 0.9592780 133.142853 -1279.029
## 15                        Y,X5,X9 3 0.9593880 133.142853 -1279.029
## 16                        Y,X5,X9 3 0.9595137 133.142853 -1279.029
## 17                      Y,X2,X5,X9 4 0.9031620  51.882766 -1342.800
## 18                      Y,X2,X5,X9 4 0.9473357  51.882766 -1342.800
## 19                      Y,X2,X5,X9 4 0.9549641  51.882766 -1342.800
## 20                      Y,X2,X5,X9 4 0.9574870  51.882766 -1342.800
## 21                      Y,X2,X5,X9 4 0.9590979  51.882766 -1342.800
## 22                      Y,X2,X5,X9 4 0.9592780  51.882766 -1342.800
## 23                      Y,X2,X5,X9 4 0.9593880  51.882766 -1342.800
## 24                      Y,X2,X5,X9 4 0.9595137  51.882766 -1342.800
## 25                    Y,X2,X3,X5,X9 5 0.9031620  25.728185 -1363.090
## 26                    Y,X2,X3,X5,X9 5 0.9473357  25.728185 -1363.090
## 27                    Y,X2,X3,X5,X9 5 0.9549641  25.728185 -1363.090
## 28                    Y,X2,X3,X5,X9 5 0.9574870  25.728185 -1363.090
## 29                    Y,X2,X3,X5,X9 5 0.9590979  25.728185 -1363.090
## 30                    Y,X2,X3,X5,X9 5 0.9592780  25.728185 -1363.090
## 31                    Y,X2,X3,X5,X9 5 0.9593880  25.728185 -1363.090
## 32                    Y,X2,X3,X5,X9 5 0.9595137  25.728185 -1363.090
## 33               Y,X2,X3,X5,X9,X104 6 0.9031620   9.452473 -1375.012
## 34               Y,X2,X3,X5,X9,X104 6 0.9473357   9.452473 -1375.012
## 35               Y,X2,X3,X5,X9,X104 6 0.9549641   9.452473 -1375.012
## 36               Y,X2,X3,X5,X9,X104 6 0.9574870   9.452473 -1375.012
## 37               Y,X2,X3,X5,X9,X104 6 0.9590979   9.452473 -1375.012
## 38               Y,X2,X3,X5,X9,X104 6 0.9592780   9.452473 -1375.012
## 39               Y,X2,X3,X5,X9,X104 6 0.9593880   9.452473 -1375.012
## 40               Y,X2,X3,X5,X9,X104 6 0.9595137   9.452473 -1375.012
## 41          Y,X2,X3,X5,X7,X9,X104 7 0.9031620   8.522417 -1371.882
## 42          Y,X2,X3,X5,X7,X9,X104 7 0.9473357   8.522417 -1371.882
## 43          Y,X2,X3,X5,X7,X9,X104 7 0.9549641   8.522417 -1371.882
## 44          Y,X2,X3,X5,X7,X9,X104 7 0.9574870   8.522417 -1371.882
## 45          Y,X2,X3,X5,X7,X9,X104 7 0.9590979   8.522417 -1371.882
## 46          Y,X2,X3,X5,X7,X9,X104 7 0.9592780   8.522417 -1371.882
## 47          Y,X2,X3,X5,X7,X9,X104 7 0.9593880   8.522417 -1371.882
## 48          Y,X2,X3,X5,X7,X9,X104 7 0.9595137   8.522417 -1371.882
## 49       Y,X2,X3,X5,X7,X8,X9,X104 8 0.9031620   8.347403 -1368.003
## 50       Y,X2,X3,X5,X7,X8,X9,X104 8 0.9473357   8.347403 -1368.003
## 51       Y,X2,X3,X5,X7,X8,X9,X104 8 0.9549641   8.347403 -1368.003
## 52       Y,X2,X3,X5,X7,X8,X9,X104 8 0.9574870   8.347403 -1368.003
## 53       Y,X2,X3,X5,X7,X8,X9,X104 8 0.9590979   8.347403 -1368.003
## 54       Y,X2,X3,X5,X7,X8,X9,X104 8 0.9592780   8.347403 -1368.003
## 55       Y,X2,X3,X5,X7,X8,X9,X104 8 0.9593880   8.347403 -1368.003
## 56       Y,X2,X3,X5,X7,X8,X9,X104 8 0.9595137   8.347403 -1368.003
## 57 Y,X2,X3,X5,X6,X7,X9,X102,X104 9 0.9031620   8.011975 -1364.299
## 58 Y,X2,X3,X5,X6,X7,X9,X102,X104 9 0.9473357   8.011975 -1364.299
## 59 Y,X2,X3,X5,X6,X7,X9,X102,X104 9 0.9549641   8.011975 -1364.299
## 60 Y,X2,X3,X5,X6,X7,X9,X102,X104 9 0.9574870   8.011975 -1364.299
## 61 Y,X2,X3,X5,X6,X7,X9,X102,X104 9 0.9590979   8.011975 -1364.299
## 62 Y,X2,X3,X5,X6,X7,X9,X102,X104 9 0.9592780   8.011975 -1364.299
## 63 Y,X2,X3,X5,X6,X7,X9,X102,X104 9 0.9593880   8.011975 -1364.299
## 64 Y,X2,X3,X5,X6,X7,X9,X102,X104 9 0.9595137   8.011975 -1364.299
```

Figure 2.1

AIC models

```
##    (Intercept)             X5             X9             X2             X3
## -661.987638193    0.498547114    0.142792162   -0.001907178   20.531222951
##           X102           X103           X104             X6             X7
##  -84.752933811  -35.699015771  152.019285844   -0.001135843   11.902511843
##             X8
##  -15.819294399
```

```
##    (Intercept)             X2             X3             X5             X6
## -661.987638193   -0.001907178   20.531222951    0.498547114   -0.001135843
##             X7             X8             X9           X102           X103
##   11.902511843  -15.819294399    0.142792162  -84.752933811  -35.699015771
##           X104
##  152.019285844
```

```
##    (Intercept)             X5             X9             X2             X3
## -661.987638193    0.498547114    0.142792162   -0.001907178   20.531222951
##           X102           X103           X104             X6             X7
##  -84.752933811  -35.699015771  152.019285844   -0.001135843   11.902511843
##             X8
##  -15.819294399
```

```
##    (Intercept)             X2             X3             X5             X6
## -661.987638193   -0.001907178   20.531222951    0.498547114   -0.001135843
##             X7             X8             X9           X102           X103
##   11.902511843  -15.819294399    0.142792162  -84.752933811  -35.699015771
##           X104
##  152.019285844
```

Figure 2.2

BIC models

```
##   (Intercept)            X5            X9            X2            X3
## -747.62435462    0.51505293    0.14397971   -0.00208566   22.64568403
##          X102          X103          X104
##  -60.97359368   14.47483469  194.50302545
```

```
##   (Intercept)            X2            X3            X5            X9
## -747.62435462   -0.00208566   22.64568403    0.51505293    0.14397971
##          X102          X103          X104
##  -60.97359368   14.47483469  194.50302545
```

```
##   (Intercept)            X5            X9            X2            X3
## -747.62435462    0.51505293    0.14397971   -0.00208566   22.64568403
##          X102          X103          X104
##  -60.97359368   14.47483469  194.50302545
```

```
##   (Intercept)            X2            X3            X5            X9
## -747.62435462   -0.00208566   22.64568403    0.51505293    0.14397971
##          X102          X103          X104
##  -60.97359368   14.47483469  194.50302545
```

Figure 2.3

Model Criteria

```
##           LL            p            n          AIC          BIC
##    -3208.3928      11.0000     440.0000    6440.7857    6489.8270
##        PRESS           R2adj
## 64823451.0180       0.9596
```

```
##           LL            p            n          AIC          BIC
##    -3212.1203       8.0000     440.0000    6442.2407    6479.0217
##        PRESS           R2adj
## 65522448.8057       0.9592
```

Figure 3.1

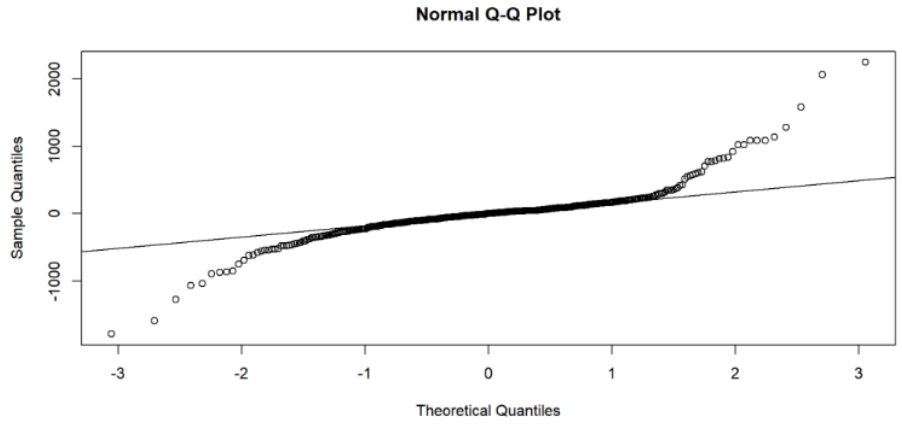


Figure 3.2

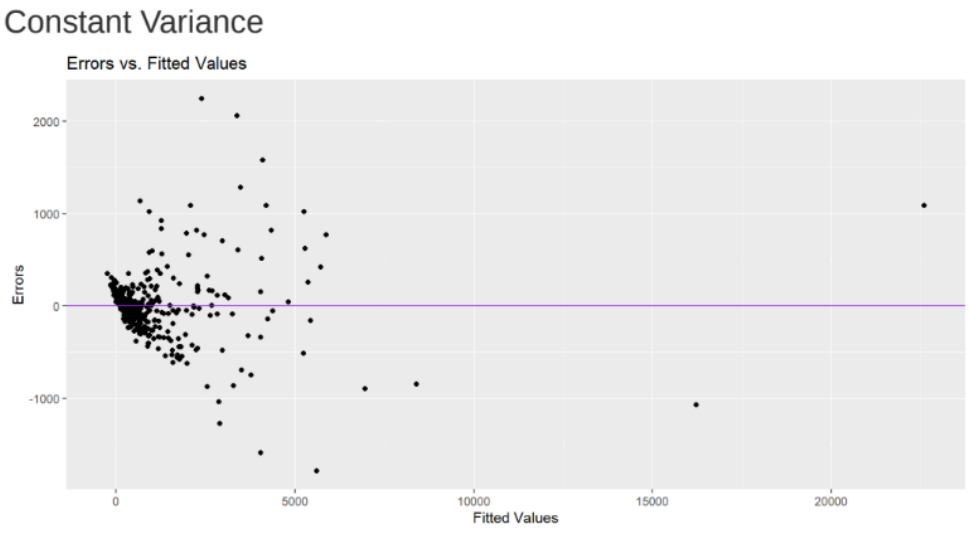


Figure 3.3

Outliers

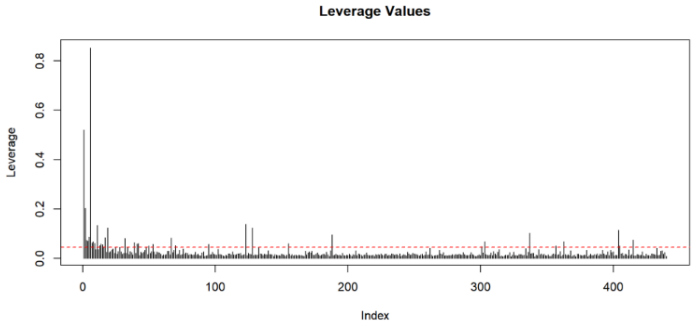
```
##   1   2   3   5   8   9  11  12  16  18  21  22  24  34  36  39  43  48  50  52
##   1   2   3   5   8   9  11  12  16  18  21  22  24  34  36  39  43  48  50  52
##  53  58  67  72 168 258 418
##  53  58  67  72 168 258 418
```

```
## [1] 27
```

High Leverage Points

```
##   1   2   3   4   5   6   7   8   9  11  13  14  15  16  17  19  32  39  41  42
##   1   2   3   4   5   6   7   8   9  11  13  14  15  16  17  19  32  39  41  42
##  50  53  67  70  95 123 128 133 155 188 301 303 337 357 363 404 405 415
##  50  53  67  70  95 123 128 133 155 188 301 303 337 357 363 404 405 415
```

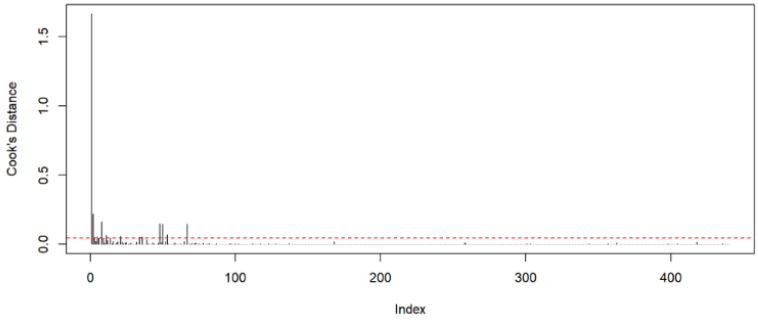
```
## [1] 38
```



Influential Points

```
##  1  2  5  6  8 11 14 21 34 36 48 50 53 67
##  1  2  5  6  8 11 14 21 34 36 48 50 53 67
```

```
## [1] 14
```




```
##  [1]   1   2   3   5   8   9  11  12  16  18  21  22  24  34  36  39  43  48  50
## [20]  52  53  58  67  72 168 258 418   4   6   7  13  14  15  17  19  32  41  42
## [39]  70  95 123 128 133 155 188 301 303 337 357 363 404 405 415
```

```
## [1] 53
```

```
## [1] 387  14
```

Figure 3.4

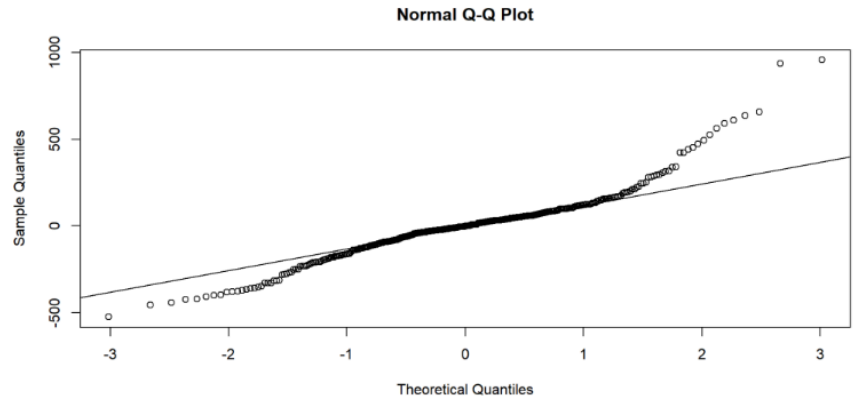


Figure 3.5

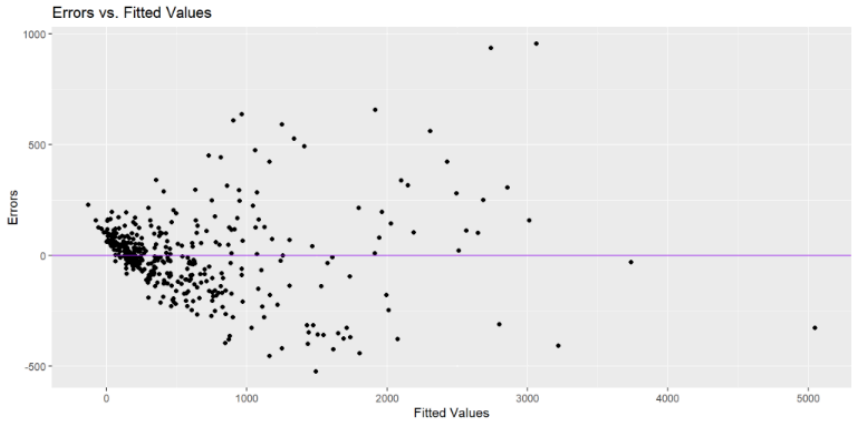


Figure 4.1

```
Call:
lm(formula = Y ~ X2 + X3 + X5 + X6 + X7 + X8 + X9 + X10, data = cdi2_clean)

Residuals:
    Min      1Q  Median      3Q     Max
-523.77  -90.78   -1.01   77.89  956.88

Coefficients:
               Estimate   Std. Error t value             Pr(>|t|)
(Intercept) -400.3753111  102.0043722  -3.925             0.000103 ***
X2            -0.0015146    0.0002641  -5.736           0.00000002 ***
X3            13.6428298    2.7642214   4.936           0.00000120 ***
X5             0.4599894    0.0219183  20.987 < 0.0000000000000002 ***
X6             0.0012165    0.0014658   0.830             0.407133
X7             5.0574293    3.7863803   1.336             0.182459
X8           -16.5715901    6.2725610  -2.642             0.008588 **
X9             0.1133118    0.0098231  11.535 < 0.0000000000000002 ***
X102         -32.7658012   29.1654238  -1.123             0.261965
X103          -2.3983125   32.8925262  -0.073             0.941914
X104         124.6347801   36.6550411   3.400             0.000746 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 189.9 on 376 degrees of freedom
Multiple R-squared:  0.9297, Adjusted R-squared:  0.9278
F-statistic: 497.3 on 10 and 376 DF,  p-value: < 0.00000000000000022
```

Figure 4.2

```
                    2.5 %          97.5 %
(Intercept) -600.945817466 -199.8048046443
X2            -0.002033890   -0.0009953749
X3             8.207560013   19.0780996849
X5             0.416891593    0.5030871424
X6            -0.001665776    0.0040986849
X7            -2.387704627   12.5025632440
X8           -28.905284429   -4.2378958664
X9             0.093996705    0.1326268782
X102         -90.113576530   24.5819740438
X103         -67.074664031   62.2780390160
X104          52.560221020  196.7093391445
```