Multiple Linear Regression of US Demographic Information on Number of Physicians

By Hans Christensen, Noemi Rieber, and Sam Sheridan

PSTAT 126, Spring 2019

Abstract

The number of physicians in an area is heavily influenced by the total population. The impact of other predictors is almost negligible. Based on demographic information for 440 of the most populous counties in the United States a regression model was constructed to predict the average number of physicians in a given country. We were especially curious about the influence of income on the number of physicians, as it could provide insights into socio-economic healthcare disparities and distributions. Our expectation of correlations between income and physicians proved true, but ultimately was insignificant in the construction of our model, where the total population proved to be the defining factor.

Problem and Motivation

The data set used in our project is the County Demographic Information (CDI) of the United States between the years 1990 and 1992. For each county, we examined how factors such as land area, total population, geographical region, and income per capita affected the number of physicians.

Our project aims to explore the relationship between these variables and the number of physicians. Examining the factors underlying the distribution of physicians in the United States of America provides numerous insights to both the local and federal government, doctors, insurers, and other medical professionals; discovering trends in the demographic information could let hospitals and governments allocate healthcare resources efficiently and drive down costs in an ever inflating marketplace . We intend to discover non-intuitive relationships which may explain and possibly affect the number of physicians for each county, state, and region of the United States.

Readers can expect a thorough examination of this demographic data. We hope to help them gain insight into the demographic distributions and correlations of not only their own counties, but of counties throughout the United States. After finishing this report, the reader should come away with a significant understanding of the primary factors driving physician population in a county.

Data

The data set we examined includes 16 variables and 440 observations from the most populated counties in the United States. Each observation is categorized by the county name, the state the county is located in, and the region of the U.S. the county pertains to. We studied the linear relationship between various predictors and the number of physicians in a county through two models. In our first model, relevant variables we considered as predictors were the total population, land area in square miles, and income per capita in dollars. For our second model, we considered the geographic region classification, as well as total population once again, as relevant predictors. We also briefly considered the percent of the population aged 65 or older, number of serious crimes, percentage of adult population with a bachelor's degree, percent of population below the poverty level, and total personal income as predictors when pursuing model selection.

The smallest county has an area of 26 square miles while the largest has an area of 20062 square miles; the average county area is 1030 square miles. The smallest population of a county is 100043, while the largest is 5105067; the average population is 361341. The smallest number of physicians in a county is 39 while the largest is 15153; the average number of physicians in a county is 894. There is a significant correlation between total population and number of physicians. There are slight correlations between total population and land area, income per capita and population, and income per capita and land area.

Questions of Interest

We are quite interested in the relationship between income and the number of physicians in a county due to the socioeconomic implications this nature of a relationship may have. We are also interested in the relationship between population and physicians, as we believe there will be a high correlation. The relationship between region and number of physicians could explain the distribution of physicians between the counties in the center of the country and their coastal counterparts. Our general questions are as follows. Can the number of physicians in a county be described by the total population, land area, and per capita income? Additionally, can the number of physicians in a county be described by the total population and the region?

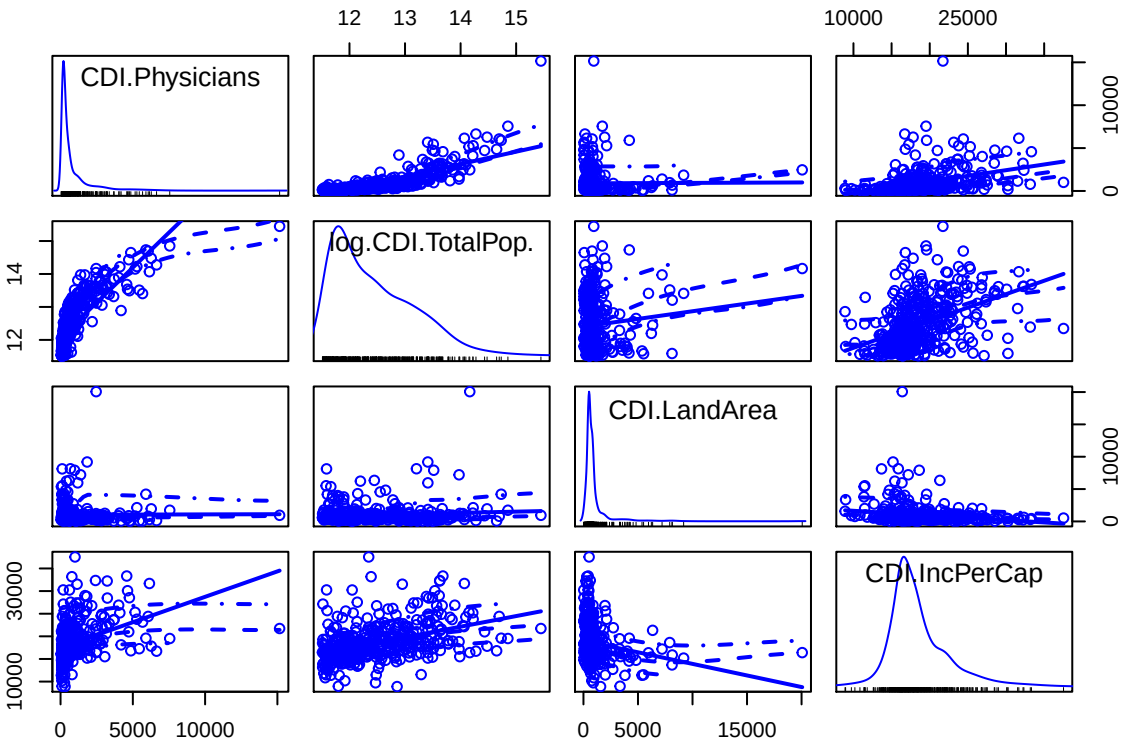
Regression Methods

The first model we investigated was the linear relationship between the total number of physicians and predictors such as the log of the total population, land area, and income per capita. Our first step was to take a look at a brief overview of the relationships between our response variable and each of the predictors, and between the predictors themselves. After gaining an initial understanding of these linear relationships, we fit our model to see the effect a change in any of the variables has on the number of physicians, given each of the other variables remain constant. We fit our regression model and estimated the effect each variable has on the response by calculating the coefficient estimates. We then took a look at diagnostic plots in order to see if the linear regression model assumptions held, transforming variables as necessary. For each transformation we made, we refit our model and checked the diagnostic plots once again to ensure our fit was improving. After refitting our model, we calculated confidence intervals for each of the coefficient estimates from our new model. We tested the hypothesis that none of the predictors have a linear relationship with the number of physicians against the hypothesis that at least one of the predictors have a linear relationship with the number of physicians, and concluded in favor of the alternate. We then tested our model for constant variance by calculating the variance of the model with and without the predictor Total Population. We found that our variance decreased without this predictor, and concluded that we do not have constant variance. In order to account for this, we refit our model using weighted least squares and found the changes to be very minimal. The second model we looked at was the linear relationship between the total number of physicians and total population and region. We looked at the initial fit, checked the diagnostic plots to verify the linear regression model assumptions, and transformed as necessary. Region is a categorical variable, so we considered the total population and number of physicians in each of the four regions. We found that the coefficient estimates for each of the predictors was similar in each region, with varying intercepts, defining a parallel regression model. This led us to hypothesize whether region had a significant effect on the number of physicians in a county, so we performed an analysis of variance to check our theory. In the first, we compared a model containing the interaction between Region and Total Population, and a model without the interaction. We found that the interaction did not have a significant effect on the model, so a second analysis of variance between a model including region as a predictor, and a model without it. From this test, we concluded Region could be removed as a predictor from our model altogether. Next, we considered a few extra predictors in order to select the optimal linear regression model. We defined the least amount of predictors we would consider in a model to be the log of the Total Population, and the maximum amount of predictors we would consider to be the log(Total Population), Pop65, Crimes, Bachelor, Poverty, and Personal Income. When going about model selection, we chose to prioritize finding the best fit for our model, therefore performing backwards selection. We were left with a model that included log(Total Population), Pop65, Bachelor, and Poverty as predictors. This led us to question whether the added complexity of including additional predictors in our model significantly improved our fit, so we tested our full model (which now included three new predictors), against our minimum model (which only included log(Total Population) as a predictor). We concluded in favor of the full model, and decided the added complexity statistically significantly improved our fit. With this best fit, we identified any influential points in our model. The first kind of influential points we looked for were outliers, of which we found a few. We then looked for influential points which are outliers in both x and y. We found points 119 and 179 to be influential points in our model.

Regression Analysis, Results, and Interpretation

To gain an idea of what relationships between preditors and our response variable will look like, we take a look at a scatterplot matrix.

```
scatterplotMatrix(~CDI$Physicians + log(CDI$TotalPop) + CDI$LandArea + CDI$IncPerCap)
```



We expect to see a positive correlation between physicians and total populations, generally there are more doctors for more people. We expect to see a positive correlation between physicians and land area. Larger swafts of land could have more doctors then smaller ones, though this could be negated by the fact that smaller municipalities may be denser and thus have more doctors. We expect there to be a positive correlation between income and number of physicians; higher incomes are generally in more urban areas where there are more doctors. We expect a positive correlation between income and population, denser areas generally have higher income than rural areas. We expect a negative correlation between income and land area for the same reason.

Now that we have an understanding of the relationships between our variables, we fit our model and look at some summary statistics.

```
summary(lm(CDI$Physicians ~ log(CDI$TotalPop) + CDI$LandArea + CDI$IncPerCap))
```

```
##
## Call:
## lm(formula = CDI$Physicians ~ log(CDI$TotalPop) + CDI$LandArea +
##     CDI$IncPerCap)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1739.9  -495.4    -5.4   375.4  9938.9
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -1.706e+04  7.060e+02 -24.165   <2e-16 ***
```

```
## log(CDI$TotalPop)  1.427e+03  6.293e+01  22.683   <2e-16 ***
## CDI$LandArea      -5.488e-02  2.865e-02  -1.916   0.0561 .
## CDI$IncPerCap      1.285e-02  1.190e-02   1.079   0.2811
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 859.7 on 421 degrees of freedom
## Multiple R-squared:  0.6202, Adjusted R-squared:  0.6175
## F-statistic: 229.2 on 3 and 421 DF,  p-value: < 2.2e-16
```

Our coefficient of determination, R^2, is 0.6202 which means that 62.02% of the variance in the number of Physicians in explained by the variance in the predictors.

The intercept is $-1.706 * 10^4$, meaning that when all predictors are zero the expected number of physicians is -17060. This is unimportant, as it does not make sense to determine the number of physicians in a place that has population 0 and land area of 0 square miles.

The remaining coefficients tell us the expected change in the number of physicians when one of our predictors increase by one unit, with all other variables in our model remaining constant.

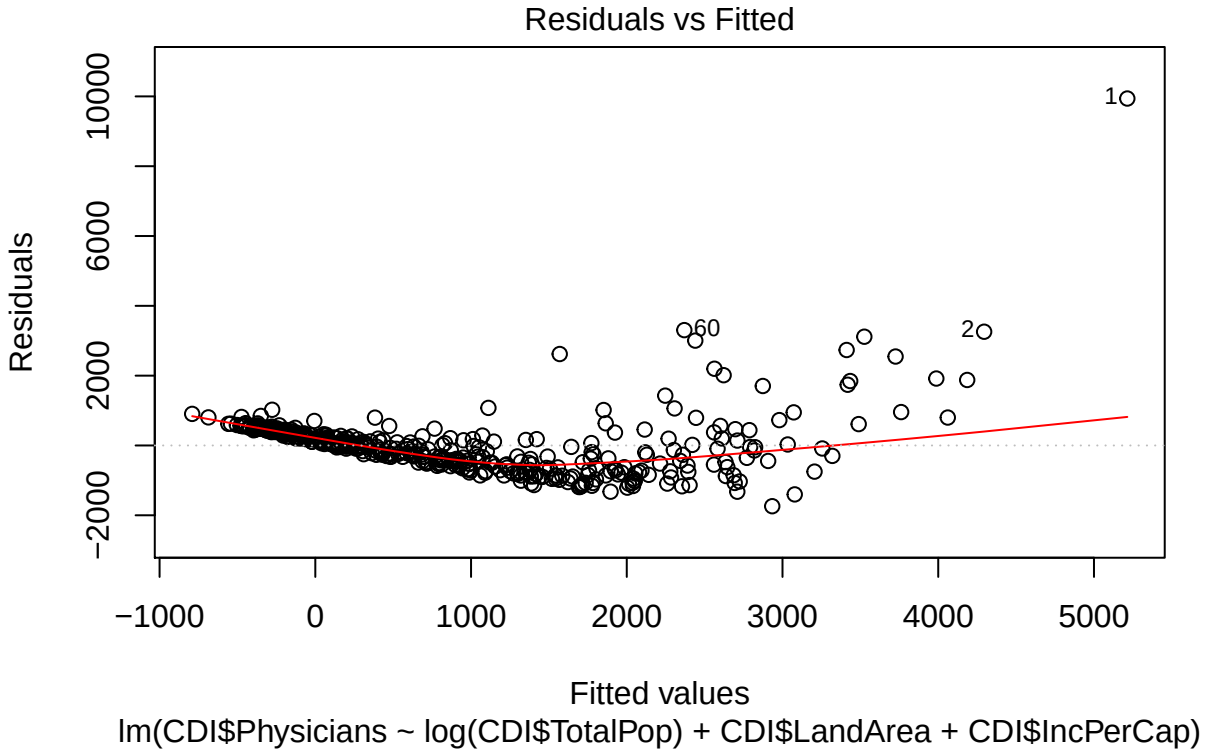
As per β_1, the expected change in the number of physicians is $(1.427 * 10^3)log(1+p)$ when the total population changes by 100p%, all else held constant. For example, when p = 0.1, the number of physicians increase by about 59 physicians when the total population increases by 10%.

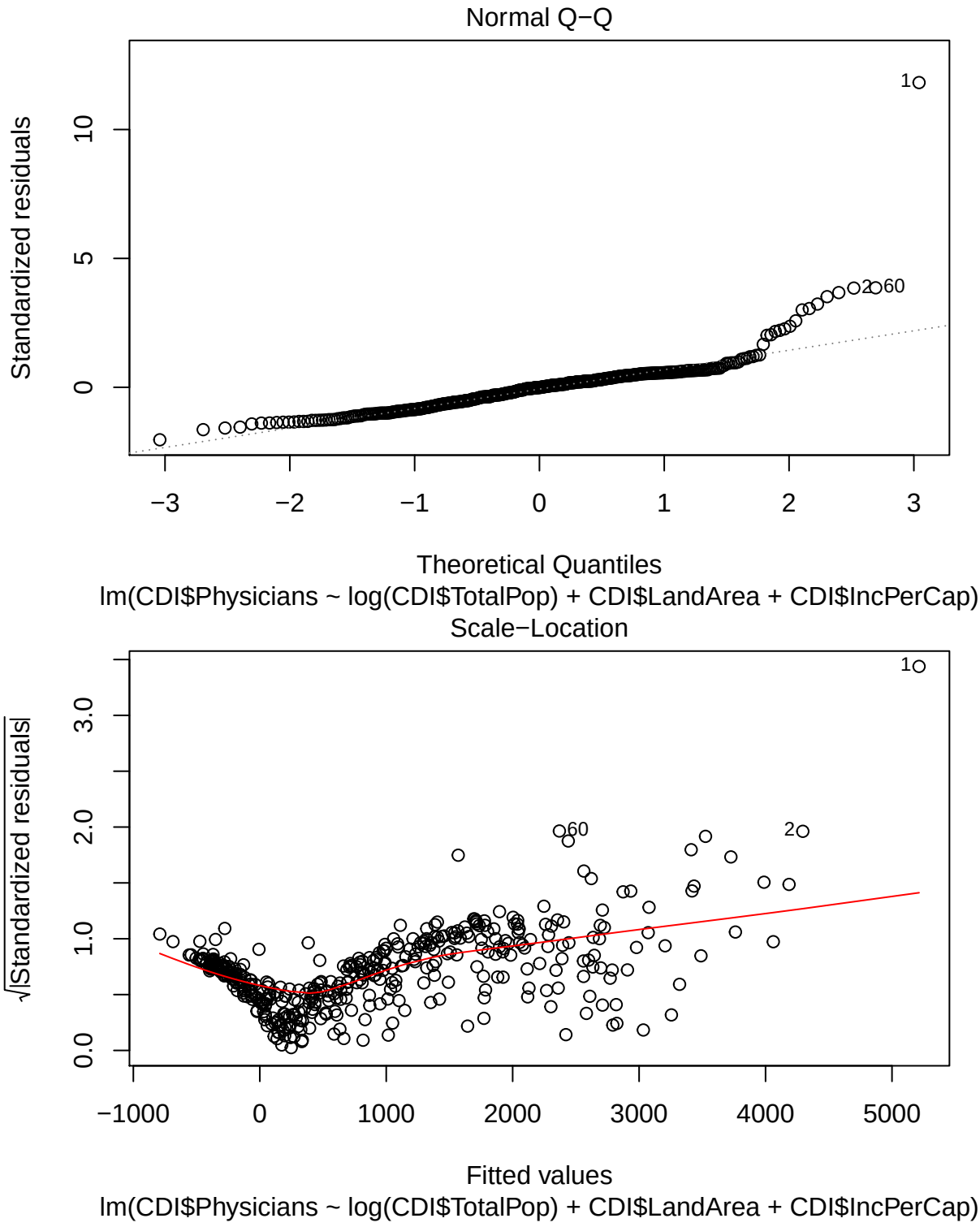
As per β_2, the number of physicians decreases by 0.06 when the land area increases by one square mile all else held constant.

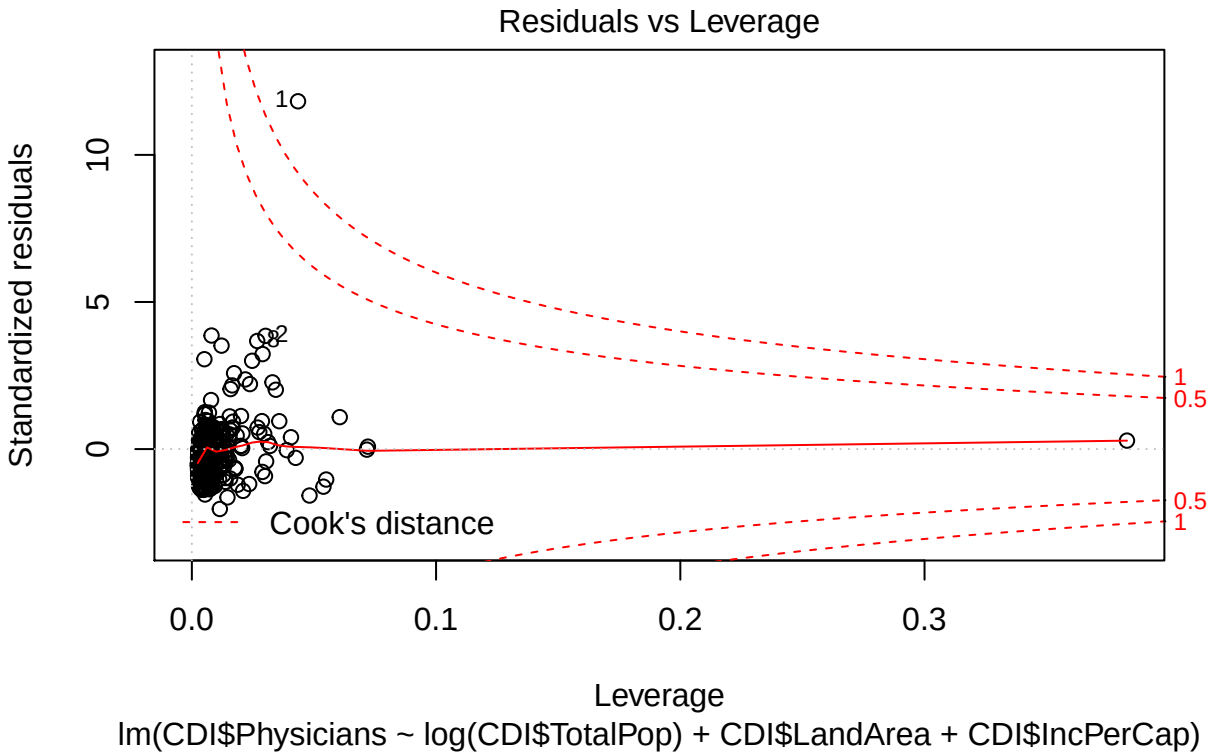
As per β_3, the number of physicians increases by 0.01 when income per capita increases by one dollar, all else held constant.

We will now check the linear regression model assumptions by looking at our diagnostic plots.

```
plot(lm(CDI$Physicians ~ log(CDI$TotalPop) + CDI$LandArea + CDI$IncPerCap))
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(CDI$Physicians ~ log(CDI$TotalPop) + CDI$LandArea + CDI$IncPerCap)
Scale−Location
√|Standardized residuals|
Fitted values
lm(CDI$Physicians ~ log(CDI$TotalPop) + CDI$LandArea + CDI$IncPerCap)



Initial diagnostic plots reveal the linear regression assumptions do not hold.

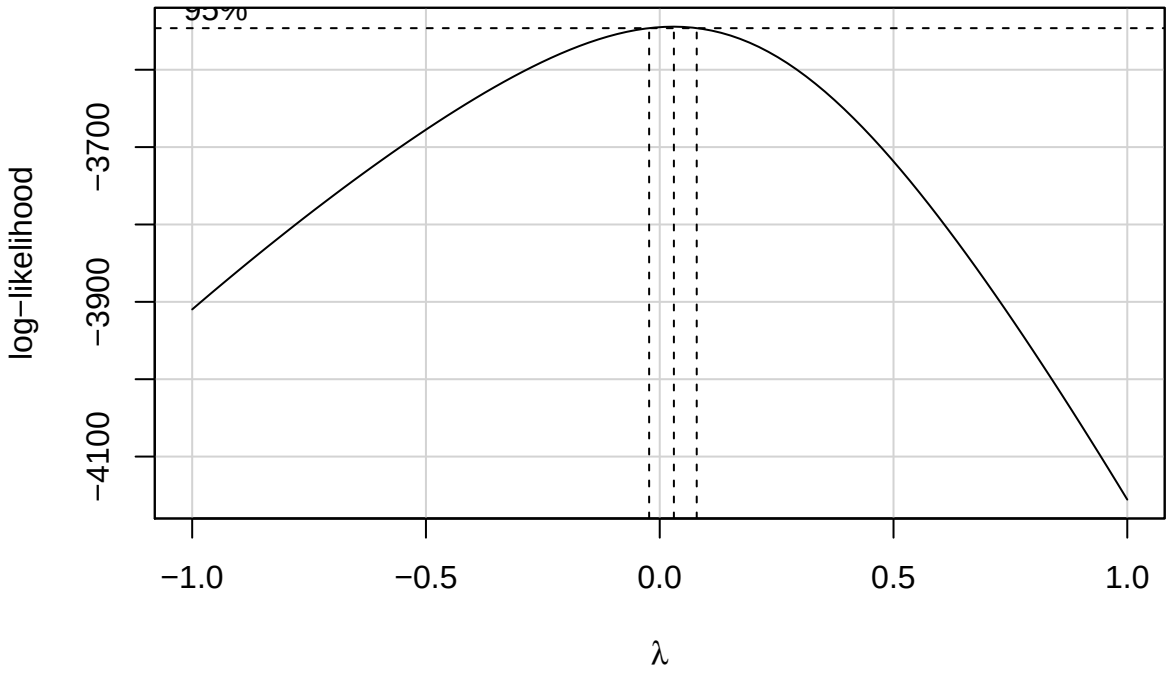
The Residuals vs. Fitted Values plot violates linearity, as the points are clustered together, then move further apart as the fitted values grow larger.

The QQ-Plot violates the normality assumption, as the data indicates a right skew with heavy tails, as the points begin to inch up in the top left corner of the plot.

The Scale-Location plot shows a violation to the constant variance assumption, as the highest points are on an upward slant, instead of distributed along an even line.

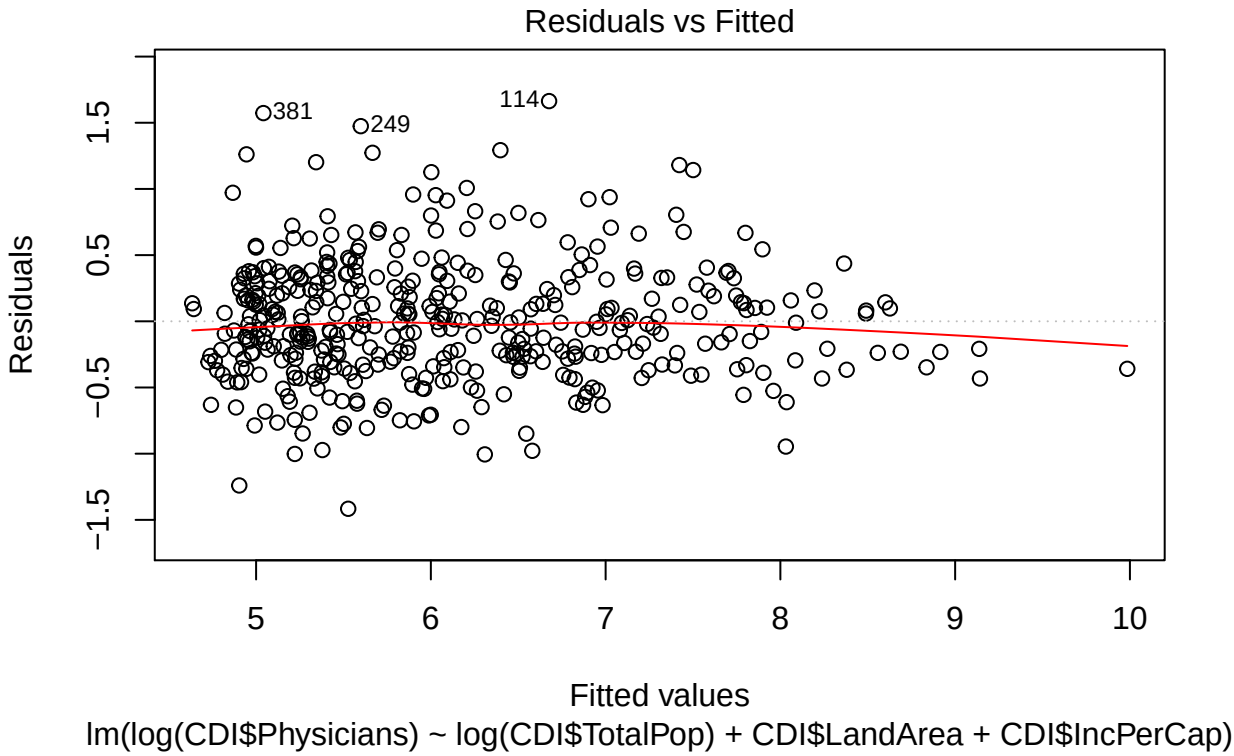
It is clear that a transformation of some kind is necessary. We use the BoxCox function to find the confidence interval for the best transformation for our response.

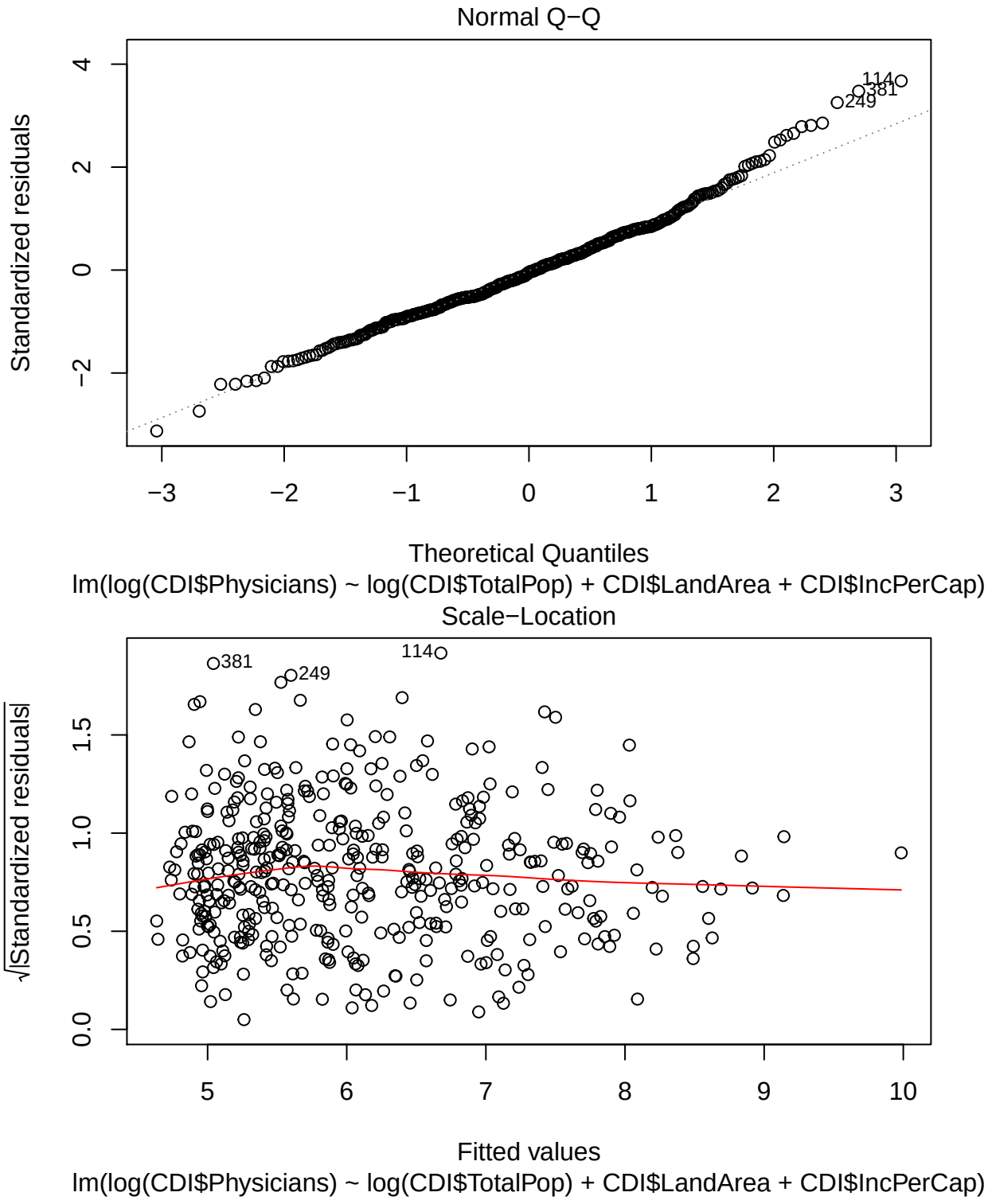
```
boxCox(lm(CDI$Physicians ~ log(CDI$TotalPop) + CDI$LandArea + CDI$IncPerCap),lambda=seq(-1,1,by=.1))
```

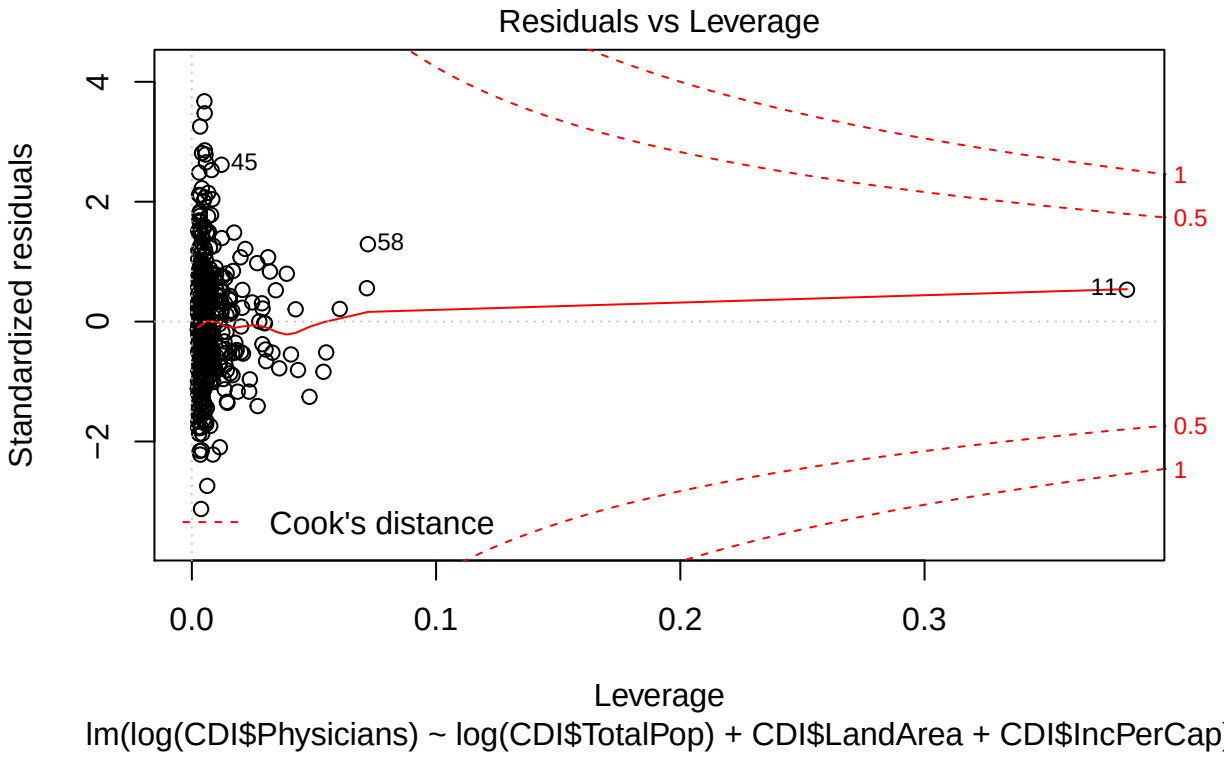


The transformation $\lambda = 0$ is in our interval, which indicates a log transformation for our response variable, Physicians. We can check the diagnostic plots for this new model to determine if our linearity assumptions now hold.

```
plot(lm(log(CDI$Physicians) ~ log(CDI$TotalPop) + CDI$LandArea + CDI$IncPerCap))
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(log(CDI$Physicians) ~ log(CDI$TotalPop) + CDI$LandArea + CDI$IncPerCap)
Scale−Location
√|Standardized residuals|
Fitted values
lm(log(CDI$Physicians) ~ log(CDI$TotalPop) + CDI$LandArea + CDI$IncPerCap)
114
381
249



This looks much better. The points in our Residuals vs Fitted Values plot are evenly distributed, resembling a null plot, which indicates linearity. Our points follow the QQ-Plot line nicely, which indicates normality. Lastly, the top-most points in the Scale-Location plot are roughly lined up, which indicates constant variance.

Now that we have found a suitable transformation, we can take a look at the summary statistics of our new model.

```
summary(lm(log(CDI$Physicians) ~ log(CDI$TotalPop) + CDI$LandArea + CDI$IncPerCap))
```

```
##
## Call:
## lm(formula = log(CDI$Physicians) ~ log(CDI$TotalPop) + CDI$LandArea +
##     CDI$IncPerCap)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.41621 -0.29509 -0.02084  0.28492  1.66362
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -1.014e+01  3.728e-01 -27.210  < 2e-16 ***
## log(CDI$TotalPop)  1.255e+00  3.323e-02  37.780  < 2e-16 ***
## CDI$LandArea      -2.980e-05  1.513e-05  -1.970   0.0495 *
## CDI$IncPerCap      3.531e-05  6.285e-06   5.618 3.52e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4539 on 421 degrees of freedom
## Multiple R-squared:  0.834,  Adjusted R-squared:  0.8328
## F-statistic: 705.2 on 3 and 421 DF,  p-value: < 2.2e-16
```

Our new R^2 is 0.834 which means that 83.4% of the variance in the number of physicians is explained by the

variance in the predictors. Note that our previous R^2 was 62.02%; the variance in the number of physicians explained by the variance in the predictors increased by about 20% in our new model, which indicates a better fit.

The intercept is -10.14, meaning that when all predictors are zero the expected number of physicians is approximately -10. Again, this is unimportant.

As per β_1, the expected percent change in the number of physicians is $100[(1+p)^(1.26) - 1]$ percent when the total population changes by 100p%, all else held constant. For example, when p = 0.1, the expected number of physicians increases by about 12.76% when the total population increases by 10%.

As per β_2, the number of physicians decreases by 0.003% when the land area increases by one square mile all else held constant.

As per β_3, the number of physicians increases by 0.004% when income per capita increases by one dollar, all else held constant.

We can now look at confidence intervals for each coefficient in our new linear model.

```
confint(lm(log(CDI$Physicians) ~ log(CDI$TotalPop) + CDI$LandArea + CDI$IncPerCap), level = .95)
```

```
##                            2.5 %        97.5 %
## (Intercept)       -1.087549e+01 -9.410109e+00
## log(CDI$TotalPop)  1.189973e+00  1.320591e+00
## CDI$LandArea      -5.952709e-05 -6.439696e-08
## CDI$IncPerCap      2.295433e-05  4.766106e-05
```

We are 95% confident that the true value of the expected number of physicians is between -0.1088 and -9.41 when all predictors are zero.

We are 95% confident that the true value of the percent change in the expected number of physicians is between $100[(1+p)^(1.19) - 1]$ percent and $100[(1+p)^(1.32) - 1]$ percent when the total population changes by 100p% and all else remains constant.

We are 95% confident that the true value of the percent change in the expected number of physicians is between -0.006% and -0.00006% when the land area increases by one square mile and all else is held constant.

We are 95% confident that the true value of the percent change in the expected number of physicians is between 0.003% and 0.005% when the income per capita increases by one dollar and all else is held constant.

```
summary(lm(log(CDI$Physicians) ~ log(CDI$TotalPop) + CDI$LandArea + CDI$IncPerCap))
```

```
##
## Call:
## lm(formula = log(CDI$Physicians) ~ log(CDI$TotalPop) + CDI$LandArea +
##     CDI$IncPerCap)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.41621 -0.29509 -0.02084  0.28492  1.66362
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -1.014e+01  3.728e-01 -27.210  < 2e-16 ***
## log(CDI$TotalPop)  1.255e+00  3.323e-02  37.780  < 2e-16 ***
## CDI$LandArea      -2.980e-05  1.513e-05  -1.970   0.0495 *
## CDI$IncPerCap      3.531e-05  6.285e-06   5.618 3.52e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.4539 on 421 degrees of freedom
## Multiple R-squared:  0.834,  Adjusted R-squared:  0.8328
## F-statistic: 705.2 on 3 and 421 DF,  p-value: < 2.2e-16
```

An $\alpha = 0.01$ level test of the linear relationship between the predictors and response tests the null hypothesis–no linear relationship between any of the predictors and the response–versus the alternative hypothesis–there is a linear relationship between any one of the predictors and the response. Our test statistic is 705.2 and our p-value is 2.2e-16. Our p-value is significantly small, therefore, we reject the null and conclude there is a linear relationship between the response and at least one predictor.

We can test for non-constant variance by seeing if the variance is a function of log(TotalPop).

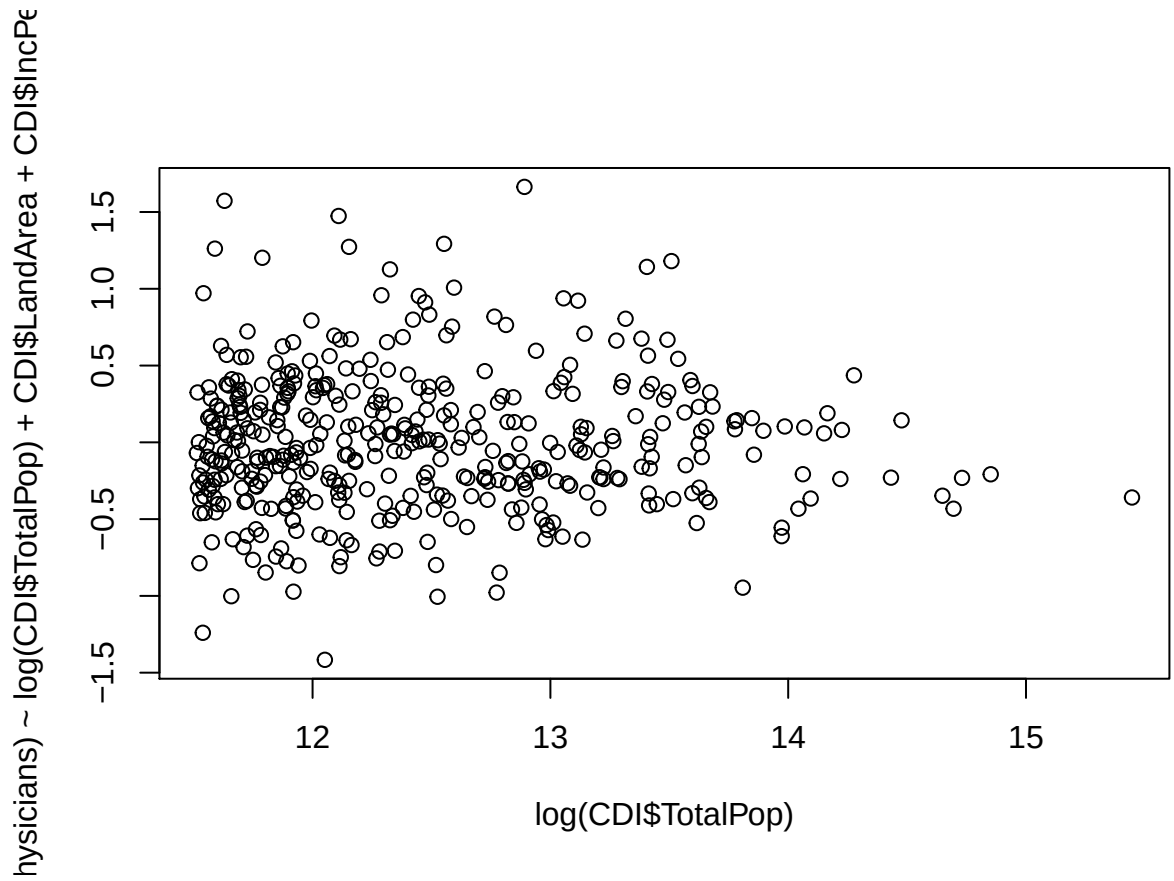
```
ncvTest(lm(log(CDI$Physicians)~log(CDI$TotalPop)+CDI$LandArea+CDI$IncPerCap), ~ log(CDI$TotalPop))
```

```
## Non-constant Variance Score Test
## Variance formula: ~ log(CDI$TotalPop)
## Chisquare = 1.649145, Df = 1, p = 0.19908
```

Our p-value, 0.19908, is statistically significant, so we can reject the null hypothesis–the variance is not a function of log(TotalPop)–for the alternate hypothesis–the variance is a function of log(TotalPop).

Is our variance increasing or descreasing with log(TotalPop)? We can plot the residuals against log(TotalPop) to see how changes with predictor log(TotalPop).

```
plot(log(CDI$TotalPop), lm(log(CDI$Physicians)~log(CDI$TotalPop)+CDI$LandArea+CDI$IncPerCap)$residuals)
```



We see can see a downward trend in the variance as a function of log(TotalPop). To reaffirm our results, we can calculate the variance before and after including log(TotalPop) in our model.

```
summary(lm(log(CDI$Physicians)~+CDI$LandArea+CDI$IncPerCap))$sigma
```

```
## [1] 0.9499664
```

```
summary(lm(log(CDI$Physicians)~log(CDI$TotalPop)+CDI$LandArea+CDI$IncPerCap))$sigma
```

```
## [1] 0.4539108
```

The variance decreases from 0.9499664 to 0.4539108 when log(TotalPop) is included in the model, therefore we can conclude that variance decreases as a function of log(TotalPop).

We will now refit our model using weighted least squares and compare to the ordinary least squares estimates of our initial model.

```
summary(lm((log(CDI$Physicians)~log(CDI$TotalPop)+CDI$LandArea+CDI$IncPerCap), weights = 1/log(CDI$Tota
```

```
##
## Call:
## lm(formula = (log(CDI$Physicians) ~ log(CDI$TotalPop) + CDI$LandArea +
##     CDI$IncPerCap), weights = 1/log(CDI$TotalPop))
##
## Weighted Residuals:
##      Min       1Q   Median       3Q      Max
## -0.40735 -0.08359 -0.00555  0.08175  0.46331
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -1.018e+01  3.819e-01 -26.667  < 2e-16 ***
## log(CDI$TotalPop)  1.258e+00  3.408e-02  36.920  < 2e-16 ***
## CDI$LandArea      -2.876e-05  1.559e-05  -1.844   0.0659 .
## CDI$IncPerCap      3.552e-05  6.353e-06   5.591 4.07e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1292 on 421 degrees of freedom
## Multiple R-squared:  0.8273, Adjusted R-squared:  0.8261
## F-statistic: 672.2 on 3 and 421 DF,  p-value: < 2.2e-16
```

```
summary(lm(log(CDI$Physicians)~log(CDI$TotalPop)+CDI$LandArea+CDI$IncPerCap))
```

```
##
## Call:
## lm(formula = log(CDI$Physicians) ~ log(CDI$TotalPop) + CDI$LandArea +
##     CDI$IncPerCap)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.41621 -0.29509 -0.02084  0.28492  1.66362
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -1.014e+01  3.728e-01 -27.210  < 2e-16 ***
## log(CDI$TotalPop)  1.255e+00  3.323e-02  37.780  < 2e-16 ***
## CDI$LandArea      -2.980e-05  1.513e-05  -1.970   0.0495 *
## CDI$IncPerCap      3.531e-05  6.285e-06   5.618 3.52e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.4539 on 421 degrees of freedom
## Multiple R-squared:  0.834,  Adjusted R-squared:  0.8328
## F-statistic: 705.2 on 3 and 421 DF,  p-value: < 2.2e-16
```

From comparing the summary tables for the weighted least squares model (top), and our initial model (bottom), we can see that our residual standard error decreases from 0.4539 in our initial model, to 0.1292 in our weighted least squares model. The coefficient estimate for log(TotalPop) decreases from 1.255 in our initial model to 1.258 in our weighted least squares model. The standard error for log(TotalPop) increases from 0.03323 in our initial model to 0.3408 in our weighted least squares model. Our coefficient of determination decreases by 1% from our initial model to our weighted least squares model. These changes are very minimal.

We can test for non-constant variance once again in our weighted least squares model.

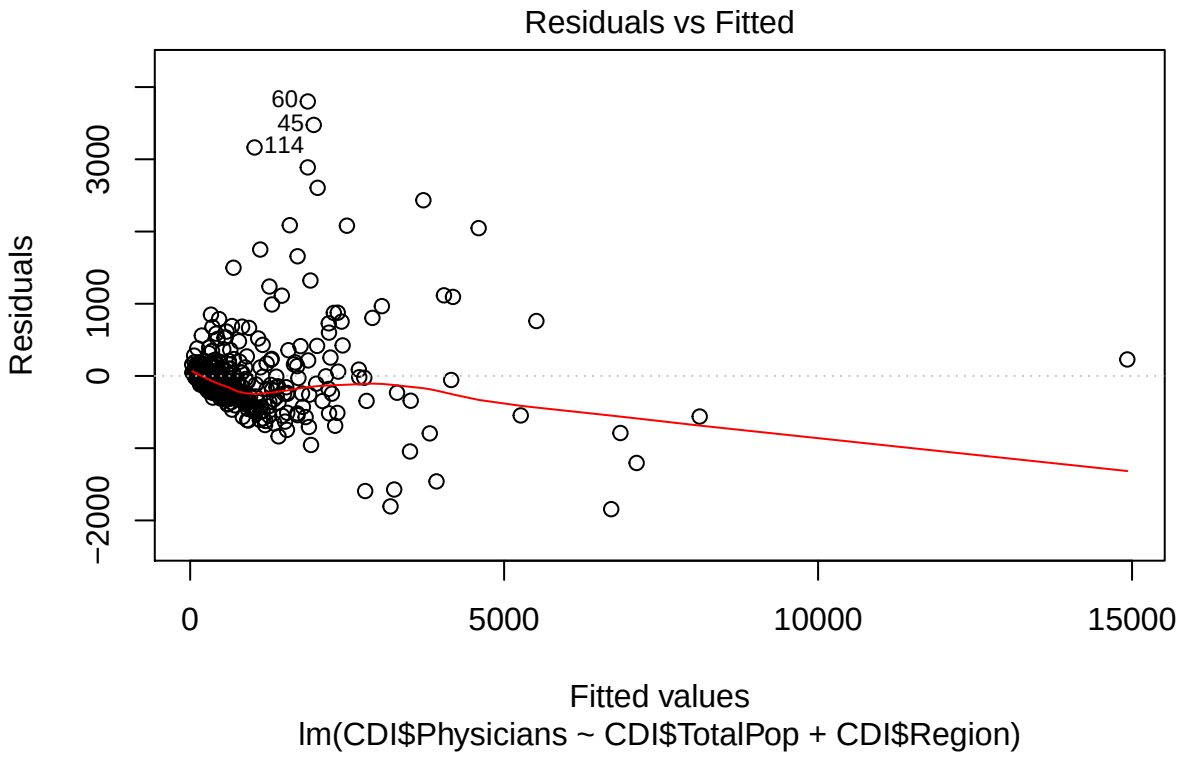
```
ncvTest(lm((log(CDI$Physicians)~log(CDI$TotalPop)+CDI$LandArea+CDI$IncPerCap), weights = 1/log(CDI$Tota
```

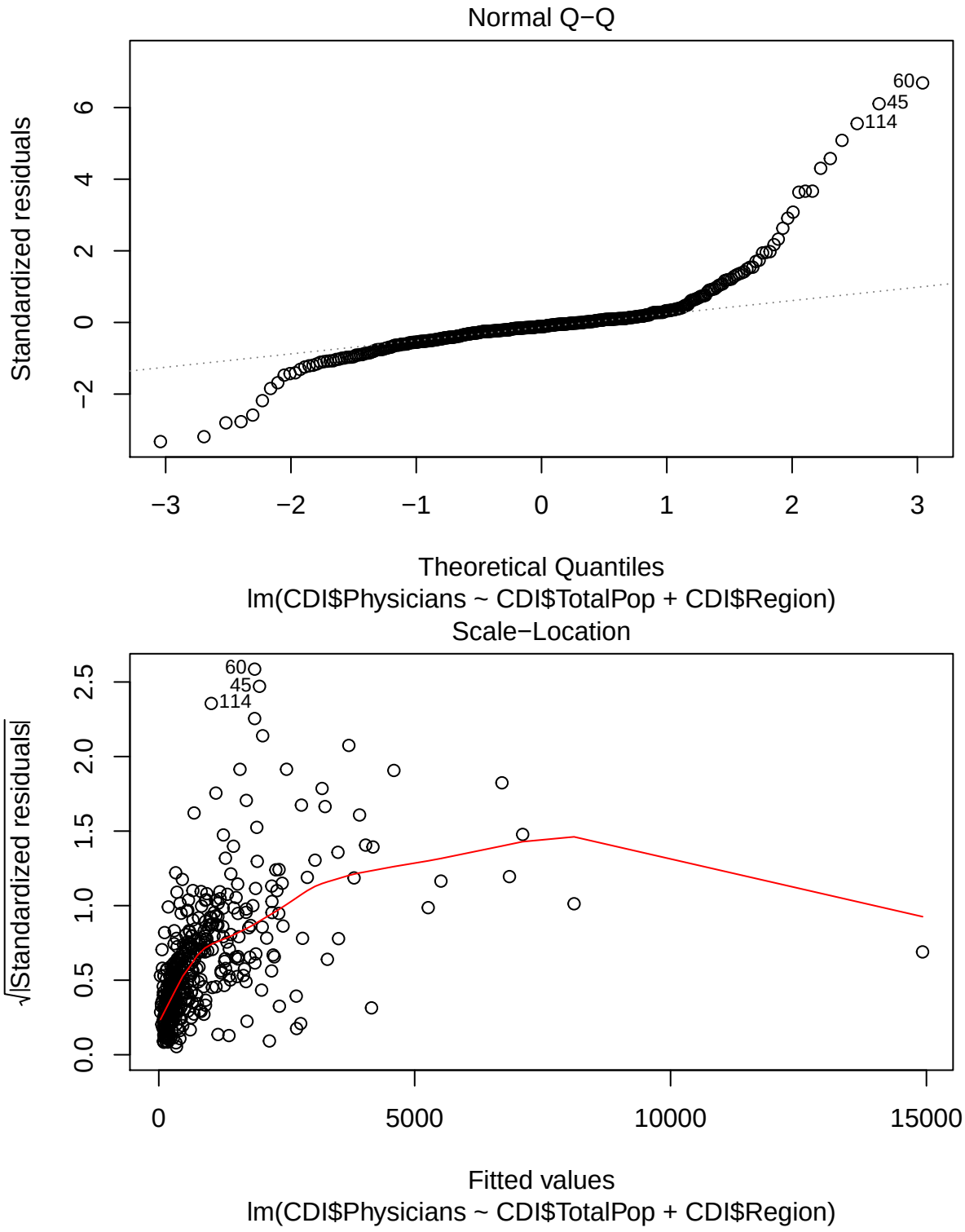
```
## Non-constant Variance Score Test
## Variance formula: ~ log(CDI$TotalPop)
## Chisquare = 3.698627, Df = 1, p = 0.054457
```

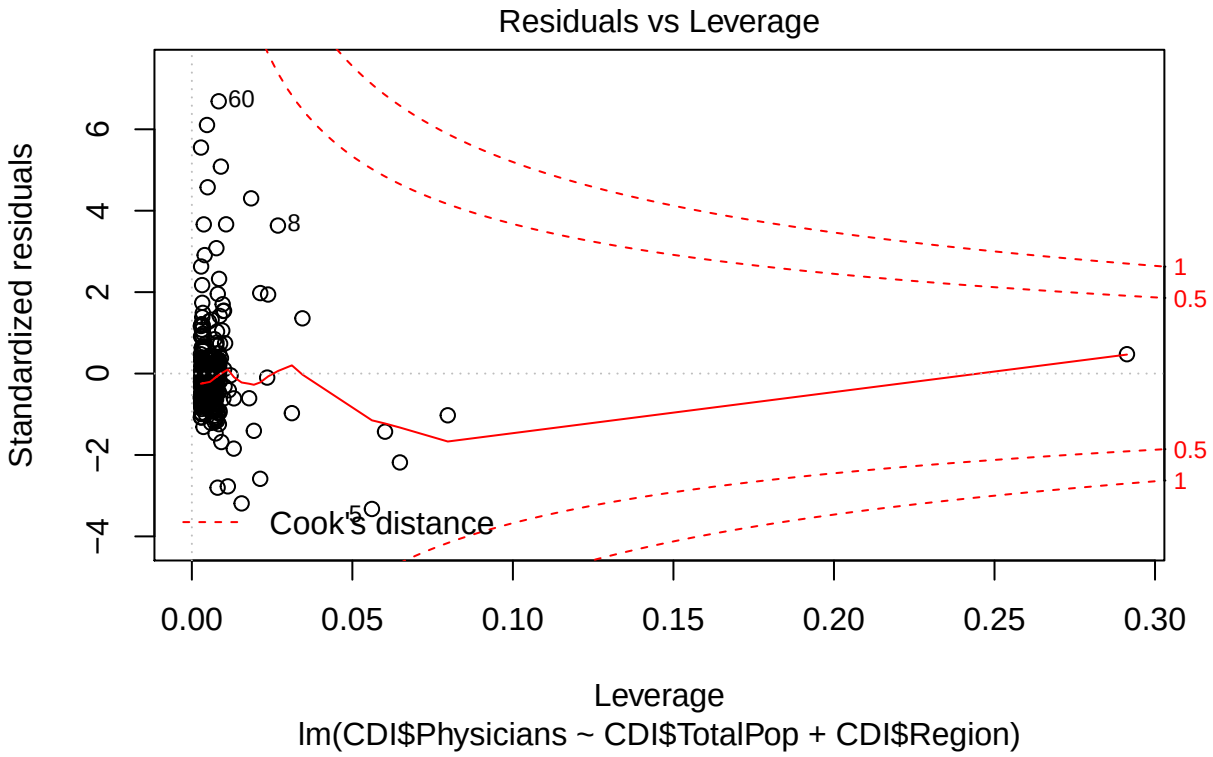
Our p-value, 0.054457, is smaller than it was when we tested for non-constant variance in our initial model; however, it is still large when testing at an $/alpha = 0.05$ level, so we reject our null of constant variance and accept that there is still non-constant variance in the weighted least squares model. Therefore, we suspect our variance is a function of one of the other predictors–either LandArea, IncPerCap, or both–as well.

We will now consider a new model.

```
plot(lm(CDI$Physicians ~ CDI$TotalPop + CDI$Region))
```




Normal Q−Q
Standardized residuals
60
45
114
−3
−2
−1
0
1
2
3
−2
0
2
4
6
Theoretical Quantiles
lm(CDI$Physicians ~ CDI$TotalPop + CDI$Region)
Scale−Location
√|Standardized residuals|
60
45
114
0
5000
10000
15000
0.0
0.5
1.0
1.5
2.0
2.5
Fitted values
lm(CDI$Physicians ~ CDI$TotalPop + CDI$Region)



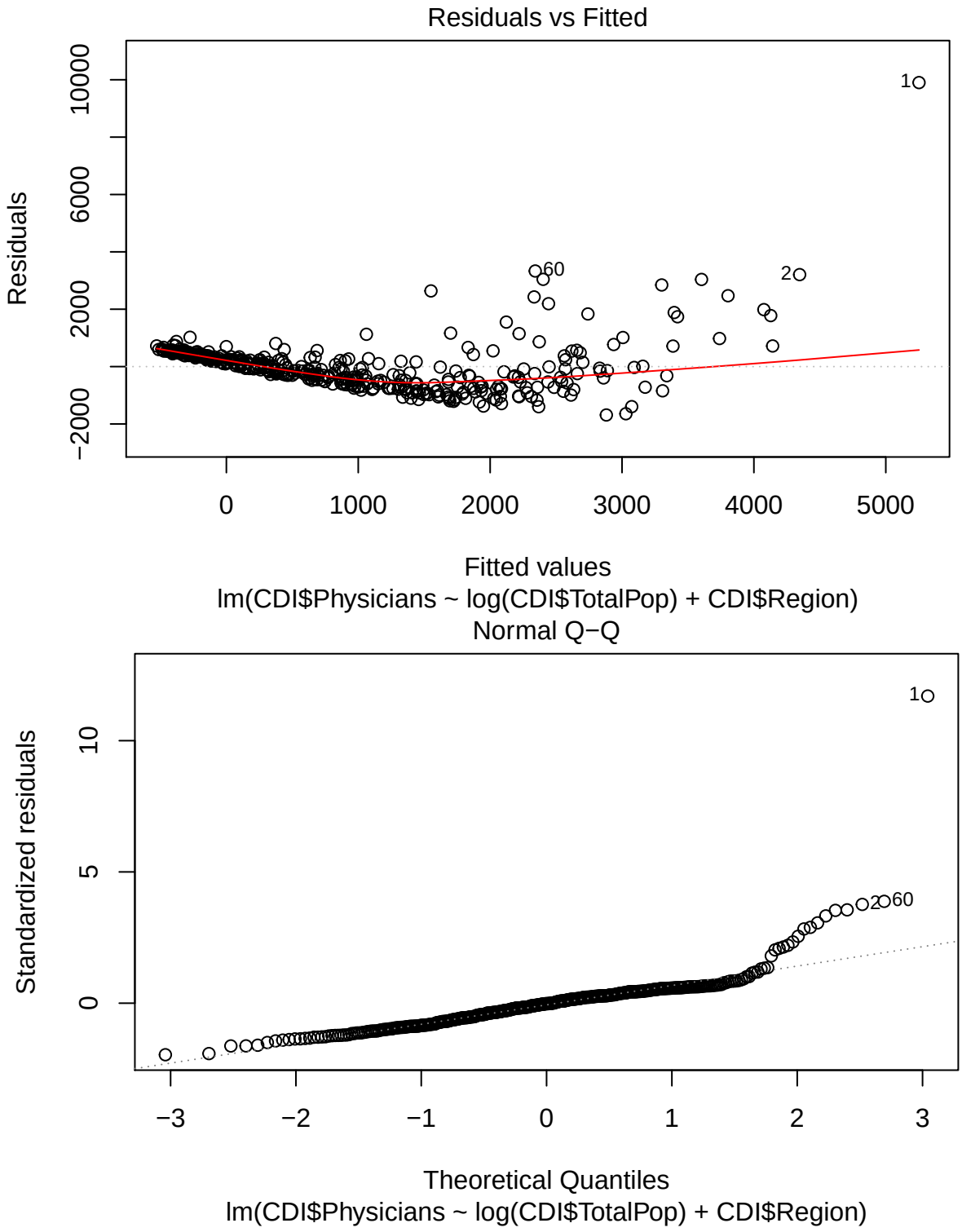
Looking at the diagnostic plots for our initial model, it is clear that our linearity assumptions are violated.

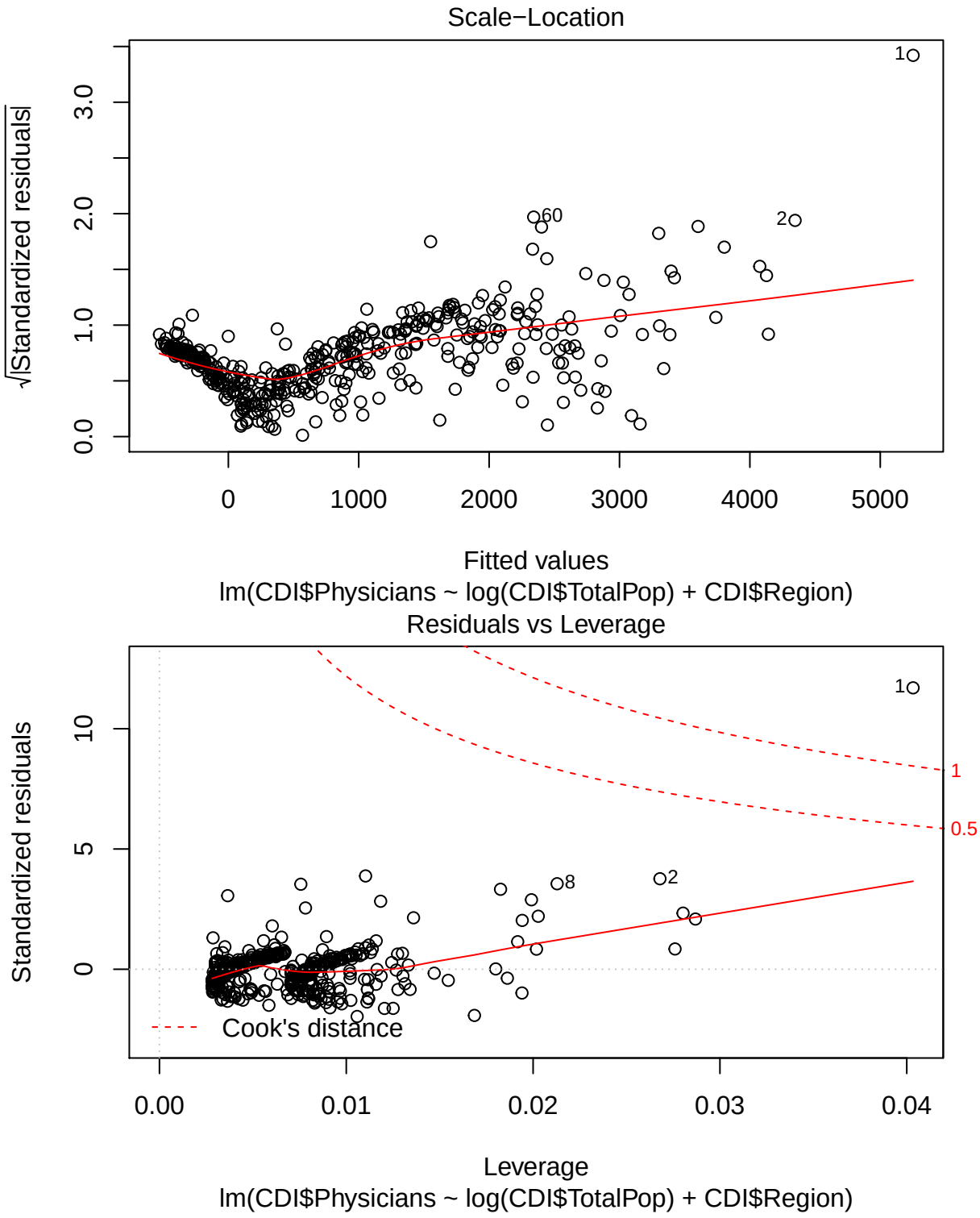
In the Residuals vs Fitted Values plot, we can see the linearity assumption is violated as there is a clear downward curvature in the points, rather than an evenly spread distribution of points throughout.

The QQ plot displays heavy tails at both ends, violating normality assumption.

The Scale vs. Location plot shows curvature, violating constant variance assumption. We perform a log transformation on the response variable Total Population, then check our diagnostic plots again to see if our predictor variable must be transformed.

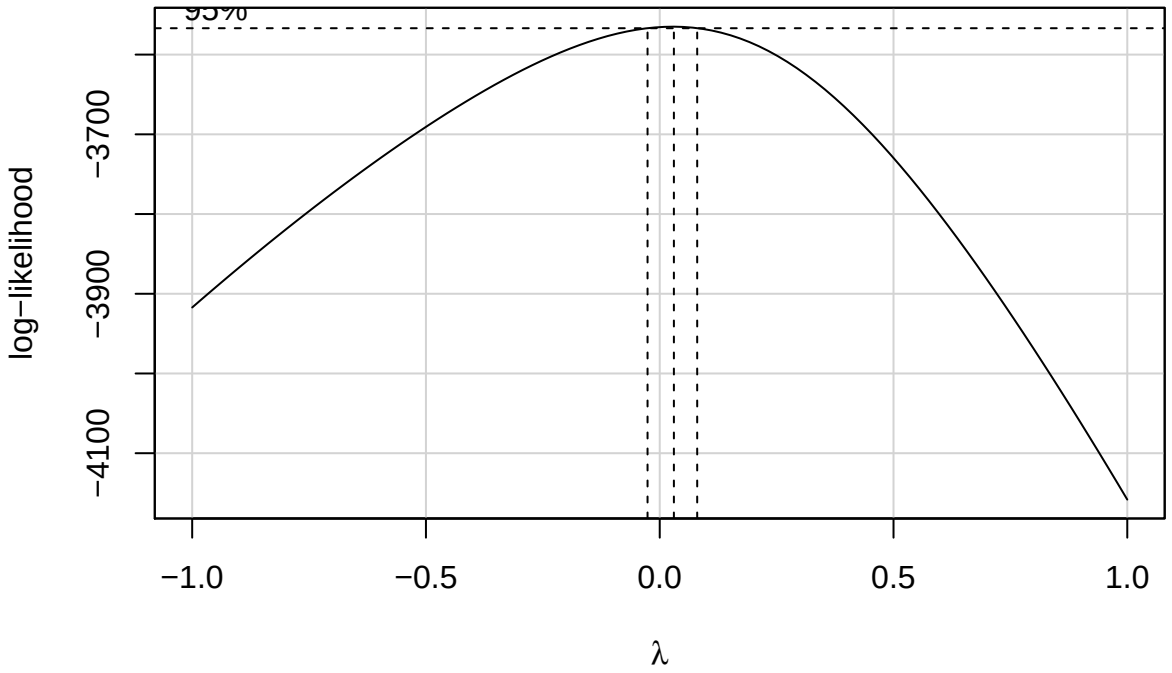
```
plot(lm(CDI$Physicians ~ log(CDI$TotalPop) + CDI$Region))
```


Residuals vs Fitted
Residuals
10000
6000
2000
−2000
1
60
2
0
1000
2000
3000
4000
5000
Fitted values
lm(CDI$Physicians ~ log(CDI$TotalPop) + CDI$Region)
Normal Q−Q
Standardized residuals
10
5
0
1
2
60
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(CDI$Physicians ~ log(CDI$TotalPop) + CDI$Region)



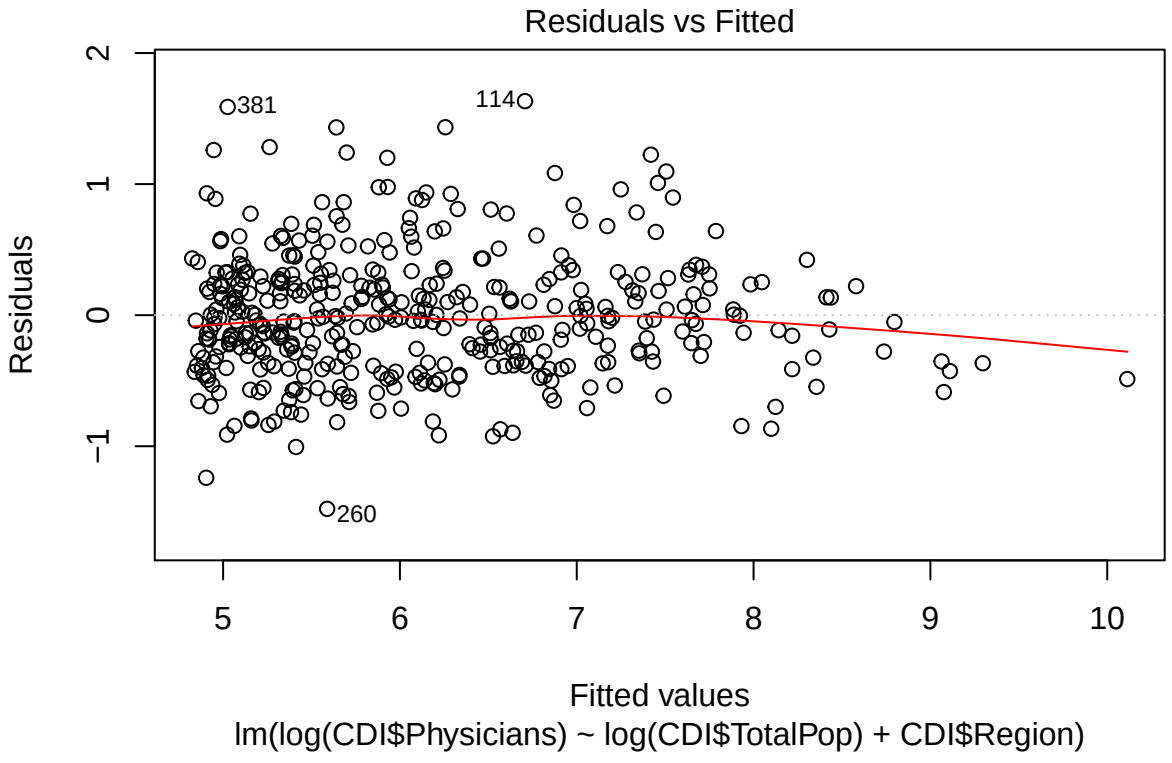
Once again, we see our linear regression assumptions are violated. Therefore, we must transform our response variable. We use the BoxCox function to find the ideal transformation.

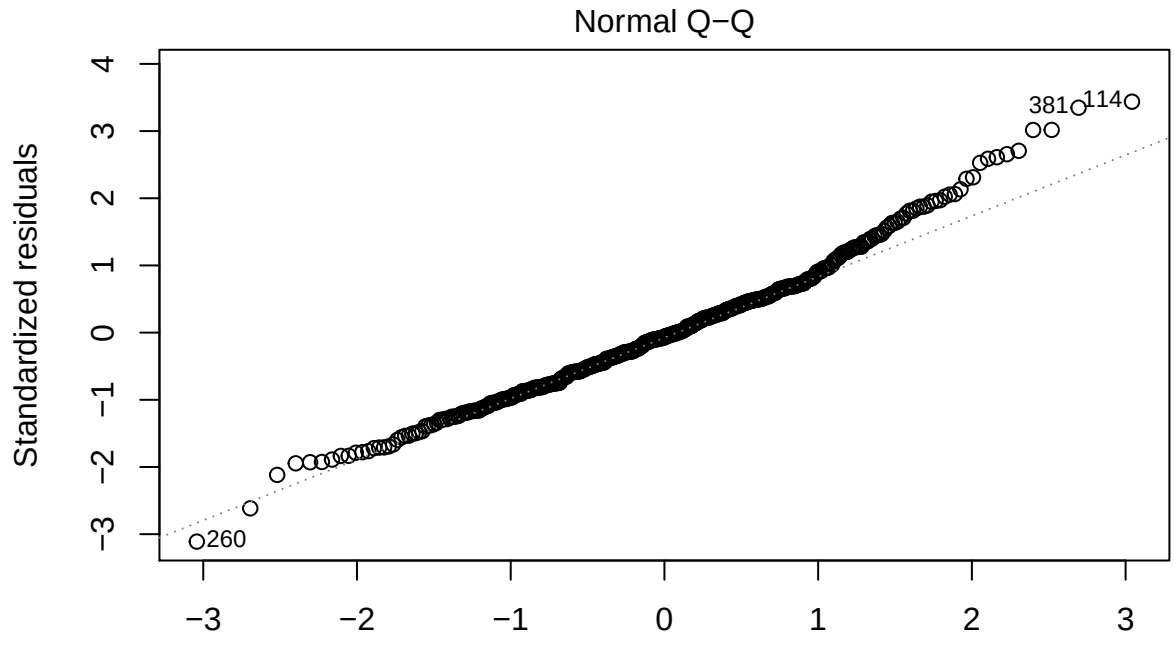
```
boxCox(lm(CDI$Physicians ~ log(CDI$TotalPop) + CDI$Region),lambda=seq(-1,1,by=.1))
```

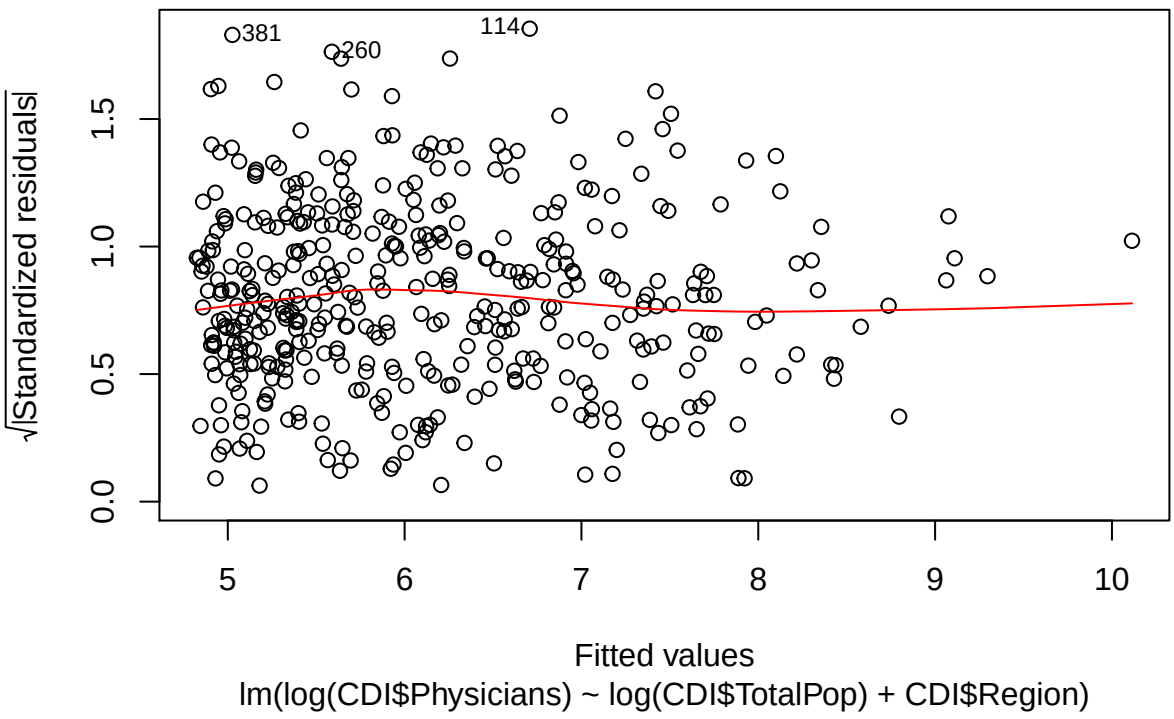


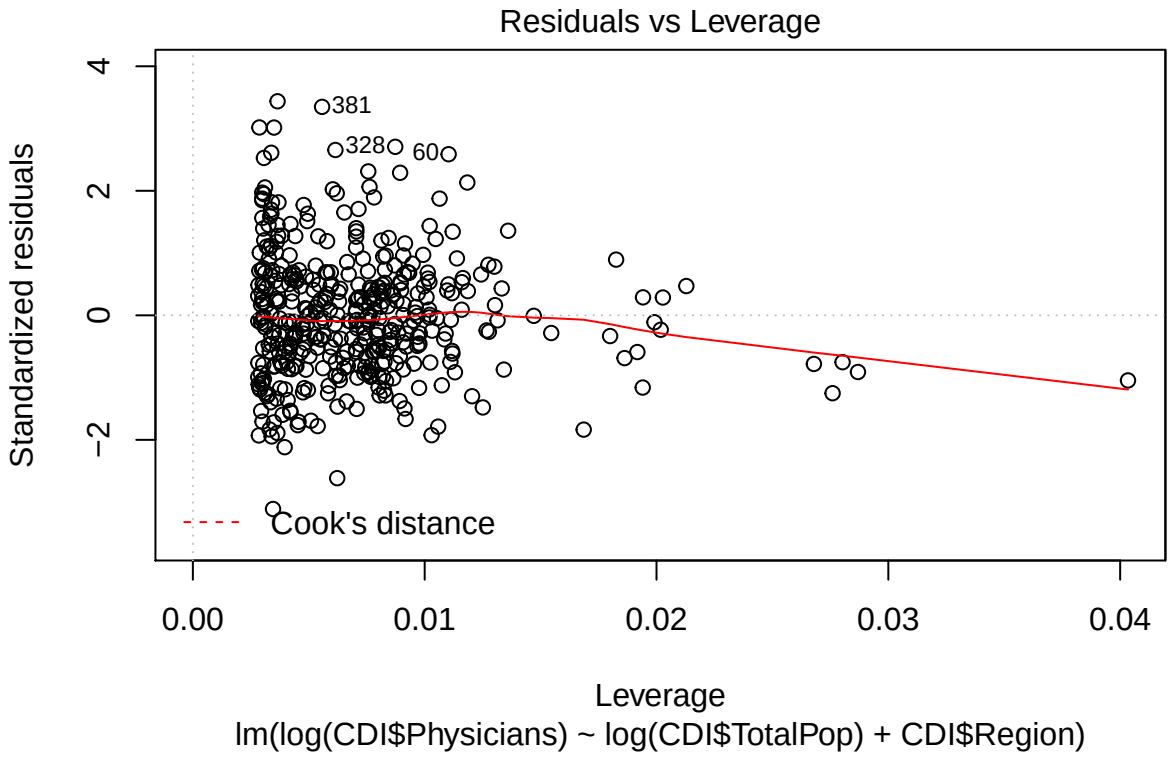
The BoxCox function provides a λ value close to zero, indicating to log transform the response variable. We can now check our diagnostic plots one more time.

```
plot(lm(log(CDI$Physicians) ~ log(CDI$TotalPop) + CDI$Region))
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(log(CDI$Physicians) ~ log(CDI$TotalPop) + CDI$Region)
381
114
260


Scale−Location
√|Standardized residuals|
Fitted values
lm(log(CDI$Physicians) ~ log(CDI$TotalPop) + CDI$Region)
381
260
114



Now our Residuals vs Fitted Values plot shows a much nicer, evenly spaced distribution of points in accordance with the linearity assumption, the QQ Plot shows normality, as our points closely follow the ideal line, and the Scale-Location plot shows constant variance as the top-most points are more of less in line. We are happy with this model, and will continue to use it throughout the rest of our analysis.

For the second linear model, we are considering Region as a predictor. Region is a categorical variable, so we subset it and included it as an interaction with the other predictor we are investigating, Total Population.

```
reg1 <- subset(CDI, Region == 1)
reg2 <- subset(CDI, Region == 2)
reg3 <- subset(CDI, Region == 3)
reg4 <- subset(CDI, Region == 4)

summary(lm(log(reg1$Physicians) ~ log(reg1$TotalPop) + (reg1$PersonalInc)))$coefficients
```

```
##                         Estimate   Std. Error   t value     Pr(>|t|)
## (Intercept)        -8.443367e+00 1.443182e+00 -5.850523 6.631879e-08
## log(reg1$TotalPop)  1.161565e+00 1.212958e-01  9.576301 1.102750e-15
## reg1$PersonalInc    1.964724e-05 1.058677e-05  1.855830 6.651391e-02
```

The estimated mean of the number of physicians in region 1 is (-8.443 + 1.162*log(TotalPop) + 1.965e-05*PersonalInc).

```
summary(lm(log(reg2$Physicians) ~ log(reg2$TotalPop) + (reg2$PersonalInc)))$coefficients
```

```
##                         Estimate   Std. Error   t value     Pr(>|t|)
## (Intercept)        -1.287215e+01 1.382070e+00 -9.313676 2.915171e-15
## log(reg2$TotalPop)  1.527403e+00 1.149931e-01 13.282562 6.821442e-24
## reg2$PersonalInc   -9.578606e-06 7.020942e-06 -1.364291 1.755084e-01
```

The estimated mean of the number of physicians in region 2 is (-12.87 + 1.527*log(TotalPop) - 9.579e-06*PersonalInc).

```
summary(lm(log(reg3$Physicians) ~ log(reg3$TotalPop) + (reg3$PersonalInc)))$coefficients
```

```
##                          Estimate   Std. Error   t value     Pr(>|t|)
## (Intercept)        -1.160619e+01 1.388619e+00 -8.358078 5.101416e-14
## log(reg3$TotalPop)  1.431890e+00 1.168503e-01 12.254051 4.724235e-24
## reg3$PersonalInc   -1.180837e-05 1.123909e-05 -1.050652 2.951908e-01
```

The estimated mean of the number of physicians in region 3 is (-11.61 + 1.432*log(TotalPop) - 1.181e-05*PersonalInc).

```
summary(lm(log(reg4$Physicians) ~ log(reg4$TotalPop) + (reg4$PersonalInc)))$coefficients
```

```
##                          Estimate   Std. Error    t value     Pr(>|t|)
## (Intercept)        -8.162386e+00 1.375631e+00 -5.9335581 9.510339e-08
## log(reg4$TotalPop)  1.137870e+00 1.140876e-01  9.9736521 3.301527e-15
## reg4$PersonalInc    7.901167e-06 8.784415e-06  0.8994528 3.714099e-01
```

The estimated mean of the number of physicians in region 4 is (-8.162 + 1.138*log(TotalPop) + 7.901e-06*PersonalInc).

This is referred to as the parallel regression model because for each separate region, the functions for estimated number of physicians all share similar slopes, but have different intercepts. This happens because there is not a large difference between the four regions.

In order to verify whether or not region has a significant effect on number of predictors, we perform two partial-F tests. In the first, we compare a model with the interaction between Region and Total Population, and a model without the interaction.

```
anova(lm(log(CDI$Physicians)~log(CDI$TotalPop) + CDI$Region), lm(log(CDI$Physicians)~log(CDI$TotalPop) +
```

```
## Analysis of Variance Table
##
## Model 1: log(CDI$Physicians) ~ log(CDI$TotalPop) + CDI$Region
## Model 2: log(CDI$Physicians) ~ log(CDI$TotalPop) + CDI$Region + log(CDI$TotalPop):CDI$Region
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1    422 95.561
## 2    421 94.766  1   0.79409 3.5278 0.06104 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can see the p-value = 0.06 is large, so we decide that including the interaction between Region and Total Population does not have a significant effect on our model. We can now check if Region should be included as a predictor at all through a second partial F-test between a model with Region and without.

```
anova(lm(log(CDI$Physicians)~log(CDI$TotalPop)), lm(log(CDI$Physicians)~log(CDI$TotalPop) + CDI$Region)
```

```
## Analysis of Variance Table
##
## Model 1: log(CDI$Physicians) ~ log(CDI$TotalPop)
## Model 2: log(CDI$Physicians) ~ log(CDI$TotalPop) + CDI$Region
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    423 95.848
## 2    422 95.561  1   0.28708 1.2677 0.2608
```

Again, we can see our p-value = 0.26 is large, so we decided to remove Region altogether, as it does not have a significant effect when added to the model.

We can now choose which predictors are relevant to our model and which are not. We perform chose to perform backwards model selection and prioritize finding the best fit. Since are looking at a small number of

predictor variables, we are not concerned with interpretability. Had there been a larger number of variables, we would have chose forward selection instead to prioritize the ease of interpretability for our model.

The initial model includes the minimum number of predictors we are willing to consider, which in this case is log(Total Population). The full model includes the maximum number of predictors we are willing to consider; log(Total Population), Pop65, Crimes, Bachelor, Poverty, and Personal Income.

```
step(lm(log(CDI$Physicians) ~ log(CDI$TotalPop) + CDI$Pop65 + CDI$Crimes + CDI$Bachelor + CDI$PersonalI
```

```
##
## Call:
## lm(formula = log(CDI$Physicians) ~ log(CDI$TotalPop) + CDI$Pop65 +
##     CDI$Bachelor + CDI$Poverty)
##
## Coefficients:
##       (Intercept)  log(CDI$TotalPop)          CDI$Pop65
##         -10.50625            1.18840            0.04226
##      CDI$Bachelor        CDI$Poverty
##           0.04632            0.03820
```

We can see that the model with the best fit includes log(Total Population), Pop65, Bachelor, and Poverty as predictors.

We can assess whether adding these predictors significantly improves our fit by testing the complex model with the newly added predictors against the model with only log(Total Population) in a partial-F test.

```
anova(lm(log(CDI$Physicians)~log(CDI$TotalPop)), lm(log(CDI$Physicians) ~ log(CDI$TotalPop) + CDI$Pop65
```

```
## Analysis of Variance Table
##
## Model 1: log(CDI$Physicians) ~ log(CDI$TotalPop)
## Model 2: log(CDI$Physicians) ~ log(CDI$TotalPop) + CDI$Pop65 + CDI$Bachelor +
##     CDI$Poverty
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    423 95.848
## 2    420 61.638  3     34.21 77.702 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can see that our p-value is significantly small, so we conclude the added predictors of percentage of population over 65, the percentage of people over age 25 with a bachelor's degree, and the percentage of the population below the poverty line have a significant effect on the number of physicians and should therefore be included in the model.

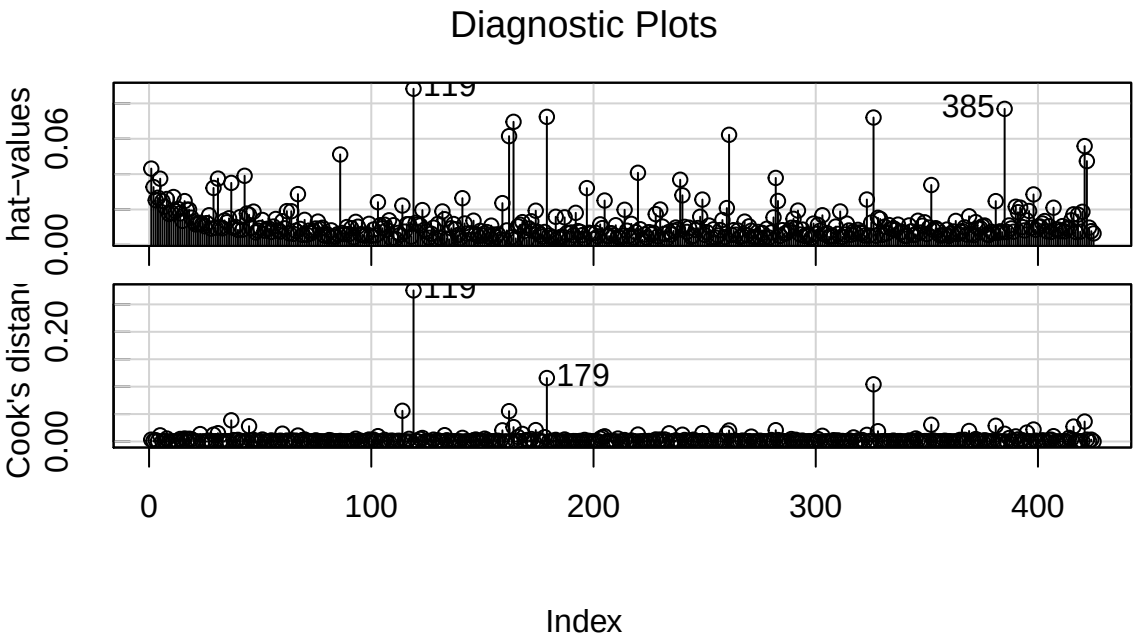
We will now identify any influential points in our new model. The first kind of influential point we will look for are outliers, or points for which our regression model is not an adequate mean function.

```
outlierTest(lm(log(CDI$Physicians) ~ log(CDI$TotalPop) + CDI$Pop65 + CDI$Bachelor + CDI$Poverty))
```

```
## No Studentized residuals with Bonferonni p < 0.05
## Largest |rstudent|:
##     rstudent unadjusted p-value Bonferonni p
## 119  -3.8332         0.00014589     0.062004
```

We see that there are no outliers in our model. We will now look for high influence points. These are points that are outliers in x and/or y.

```
influenceIndexPlot(lm(log(CDI$Physicians) ~ log(CDI$TotalPop) + CDI$Pop65 + CDI$Bachelor + CDI$Poverty)
```



The top diagnostic plot reveals points that are outliers in x, the bottom plot reveals plots that are outliers in both x and y.

Point 119 stands out the most, as it is clearly an outlier in x, as shown by the top plot. More importantly, however, it is an outlier in y as well, which increases its leverage on the total fit. Consequently, point 119 shows up in the bottom plot. We see that 179 also appears to be a high leverage point, as it is also in the bottom plot.

Conclusions

After performing analysis on the CDI dataset using a variety of regression and model building methods, we found that the best model for predicting the number of physicians in a county included total population, the percentage of the population over 65 years old, the percentage of the population over 25 with a bachelor's degree, and the percentage of the population below the poverty line. We decided to transform the response, physicians, and the predictor total population with a logarithmic scale because they both showed exponential trends. We were surprised to find that personal income did not have a significant effect on the model, as most people would assume counties with wealthier citizens would receive better healthcare, and thus have a greater number of physicians. As this data set is nearly 30 years old, these findings may not be relevant to counties that have experienced a lot of change in the past 30 years, but the findings could still hold for counties whose demographics have not altered much since 1990. These findings should not be generalized to today's prediction of physicians, as any underlying or confounding variables have likely changed during the past 30 years, and we can not rely on our model to confidently make conclusions about today's counties.