Creating Better Healthcare Through the Use of Data

Matthew Coleman & Daniel Lai

5/27/2019

1 Abstract:

For our project, we investigated different aspects of a county that would affect the number of physicians in a county. Variables we explored included: total population, land area, income per capita, personal income, region, crimes, population over 65, and poverty.
In the first section of our analysis, we explored the effects that logarithm of the total population, land area, and income per capita had on the logarithm of physicians. From our analysis we discovered $\frac{1}{\sqrt{TotalPopulation}}$ and $log(Physicians)$ are the best transformations to have the variables adhere to normality. We also found a positive relationship between log(physicians) and the logarithm of income per capita, as well as a negative relationship with log(physicians) and log(land area); note, all of these relationships occur when all other predictors are held constant. Overall, many of our initial assumptions about the relationships were confirmed as being correct.
In the analysis of our second model we questioned the initial model adequacy, and explored larger models to learn whether we could improve upon the initial model. We found the addition of poverty, population over 65, bachelor's degree and personal income as appropriate in the creation of a better model. All of these predictors increased our coefficient of determination, meaning we could explain more of the variation in the response with the linear relationship with the predictors. After fitting the new model, we discovered the initial model was not the best. A larger model improved our analysis, and we found more predictors that had associations with the logarithm of the physicians in counties.

2 Problem and Motivation:

i) Background:
Our data set covers the 1990 county demographic information of the 440 most populous counties in the United States. There are 16 variables total and 8 of which we will be exploring as possible predictor to the number of physicians in a county. We are going to be exploring two different models of physicians being regressed onto different predictors. The first model is going to contain total population, land area, and income per capita as predictors. The second model is going to begin with total population and region, with possible exploration into population above 65 years of age, bachelor, poverty rates, and personal income.

ii) Motivation:
This report is important because the results of our models will allow us to create a predictive model which can take in a number of predictors and give a number of physicians in a county. The predictive model can also allow us to see how many physicians a county needs to properly serve the population based on different factors. This is an important problem because there are many areas that are underserved, and looking at factors such as age above 65 allows us to see there needs to be an increase in health coverage.

3 Questions of Interest

- Which factors have a positive linear relationship with the number of physicians? Which ones have a negative relationship?
- Do all the assumptions of linear regression hold? If they do not, which transformations will correct these failed assumptions?
- Are there better models that will allow us to remove variables which are not influential to our analyses?
- Are there better models with more variables which can explain a greater amount of variability in physicians?
- How much of the variability in physicians can be explained by a linear relationship with our variables?

Regression Methods

Regression Methods used for Part I outlined below:

- Diagnostic Plots:
  - Scatterplot Matrices
  - Added-Variable Plots
  - Residuals vs. Fitted
  - Scale-Location
  - Q-Q Plots
- Ordinary Least Squares Regression Models
- Power Transformations and Box-Cox Method
- Partial F-Tests (including Global F-Tests)
- Non-Constant Variance Tests

Regression Methods used for Part II outlined below:

- Diagnostic Plots:
  - Added-Variable Plots
  - Residuals vs. Fitted
  - Scale-Location
  - Q-Q Plots
- Power Transformations and Box-Cox Methods
- t-Tests
- Partial F-Tests
- Outlier Tests
- Influence Index Plots
- Residuals vs. Leverage Plots

4 Analysis

4.1 Model 1

In this report, we are going to explore the relationship between the number of physicians and total population, land area and income per capita with the multiple linear regression model:
We will begin with the model below, and explore into possibly better models.

$$Physicians \sim log(TotalPop) + LandArea + IncPerCap$$

The data for this report is the county demographic information (CDI), and contains demographic information on the 440 most populous counties in the United States. The variables we will be working with and their descriptions are listed below:

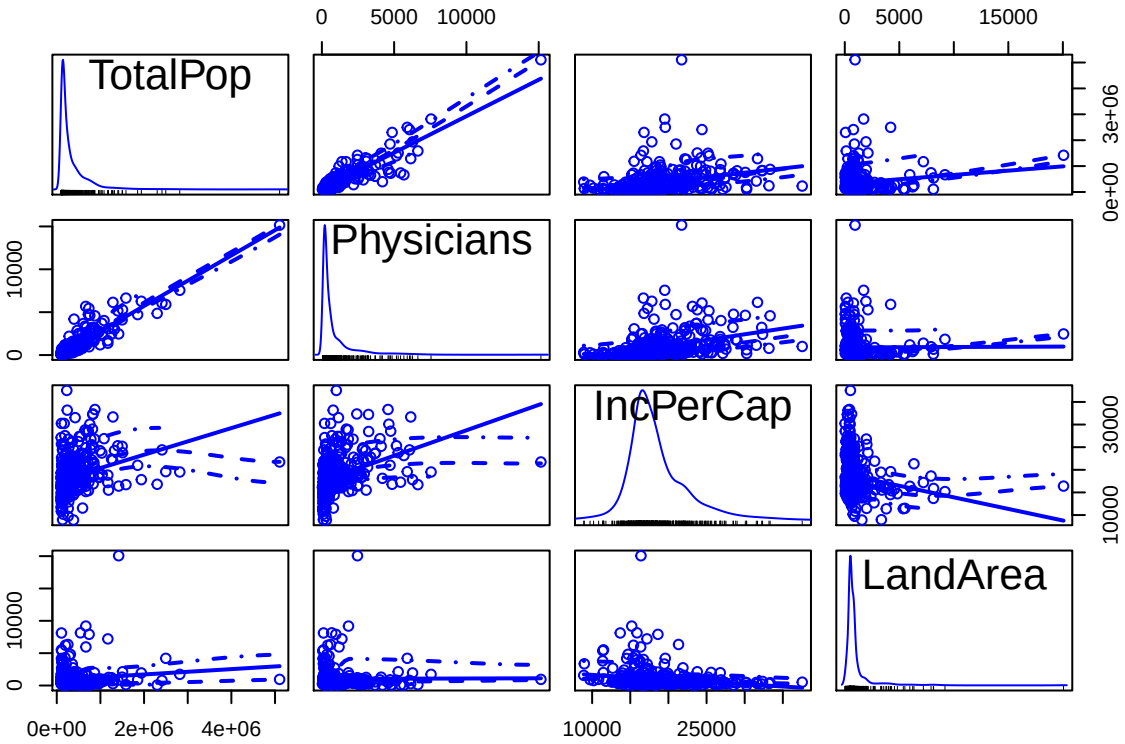
Variable	Description
Physicians	Number of professionally active nonfederal physicians during 1990.
TotalPop	Estimated 1990 population.
LandArea	Land area (square miles).
IncPerCap	Per capita income of 1990 CDI population (dollars).

4.1.1 Exploratory Data Analysis

One relationship I will want to explore is the relationship between population and physicians. I infer that there is a positive relationship between population and physicians because as the number of people increases, there is likely to also be more physicians. I would also expect income per capita to have a positive association with physicians because areas with higher income would have better access to healthcare as a result of better

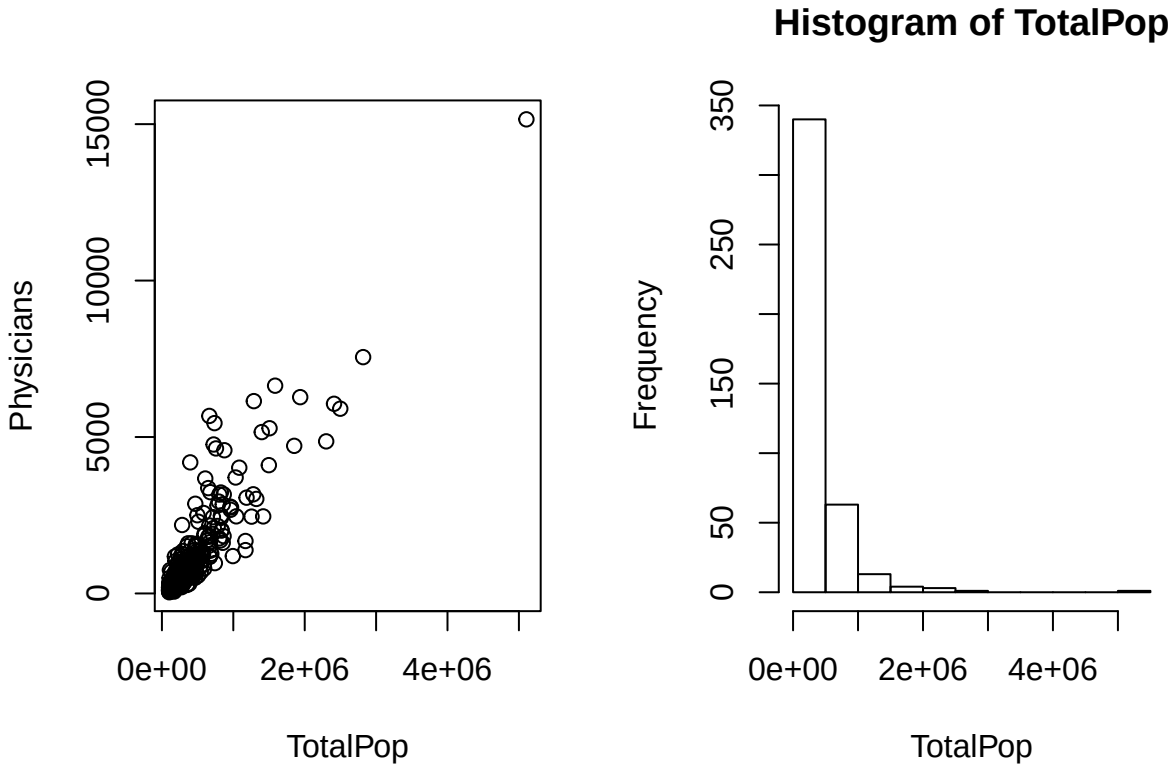
infrastructure and a more lucrative customer base. I am not expecting there to be a relationship between land area and the number of physicians. I believe population is going to be most important, and regardless of whether a county is large or small, if the population is small there will not be many physicians and if it is large there will be many. Between predictors, I would not expect association between income per capita and any of the other predictors. There could be some association between land area and total population, but population density becomes an issue. There can be large counties with low population density such as rural counties, as well as large areas with large populations, such as Los Angeles county.

```
scatterplotMatrix(cbind(TotalPop,Physicians,IncPerCap,LandArea))
```



The scatterplot matrix reveals that the total population has a positive linear relationship with physicians. Another noteworthy relationship is that between income per capita and physicians. The distribution of each variable demonstrates that total population, physicians, and land area are right skewed. Income per capita is slightly right skewed, but is more close to normal than the other variables.
Having right skewed total population data means there are more lower population counties than there are higher population counties Plotting a histogram of the total population will give a better view of this idea.

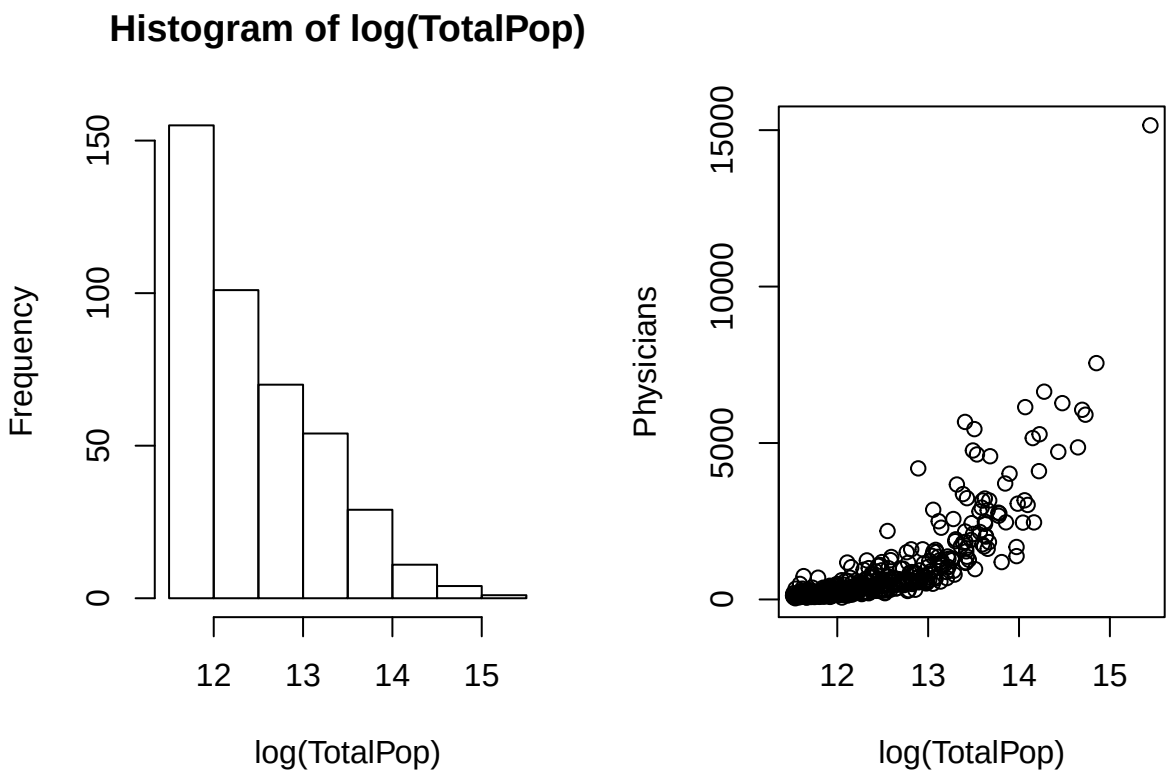
```
par(mfrow = c(1,2))

plot(x = TotalPop, y = Physicians)
hist(TotalPop)
```




```
#ggplot(data = CDI, aes(x = TotalPop, y = Physicians)) +
#  geom_point()
```

These plots confirm that there is a heavy right skew to the total population. A Log transformation on the total population would help make the data more distributionally similar to normal.

```
par(mfrow = c(1,2))
hist(log(TotalPop))
plot(log(TotalPop), Physicians)
```



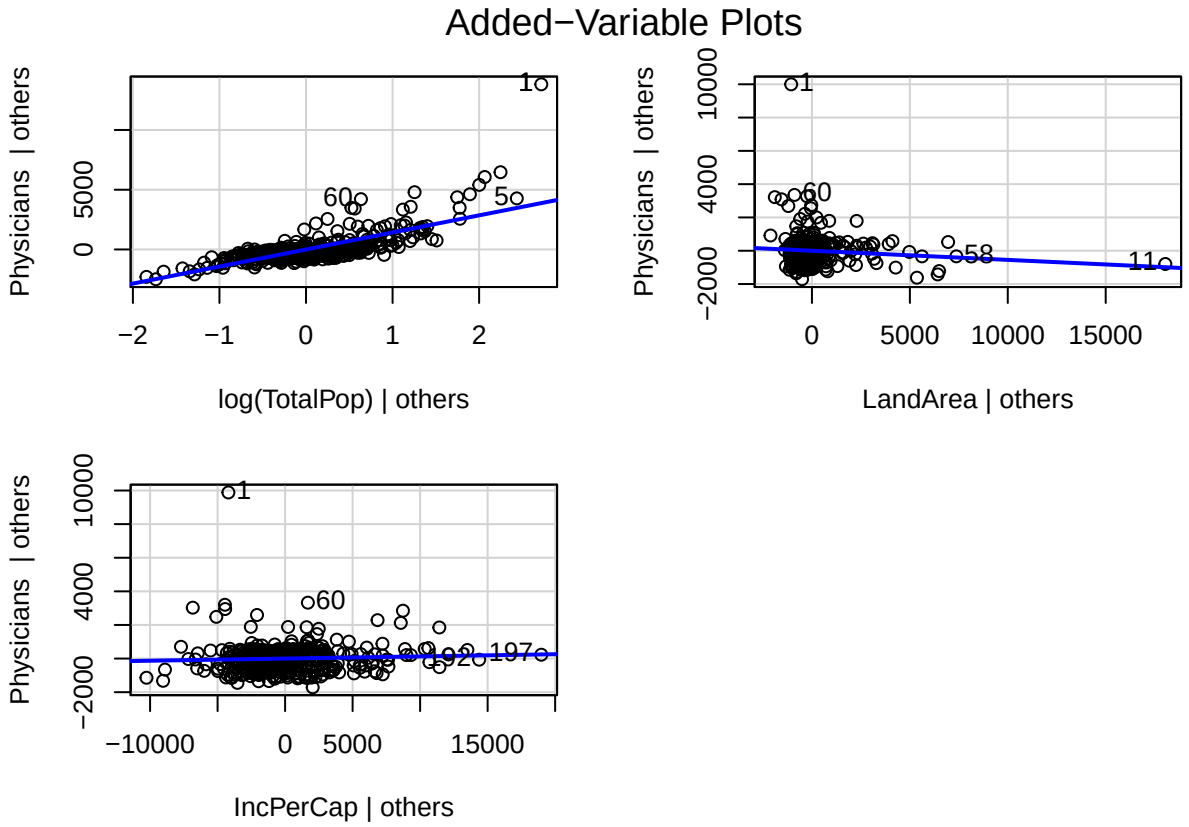
While this is not perfect, it is an improvement from the untransformed variable.

4.1.2 Model Fitting

```
full.lm <- lm(Physicians ~ log(TotalPop) +  LandArea + IncPerCap)
```

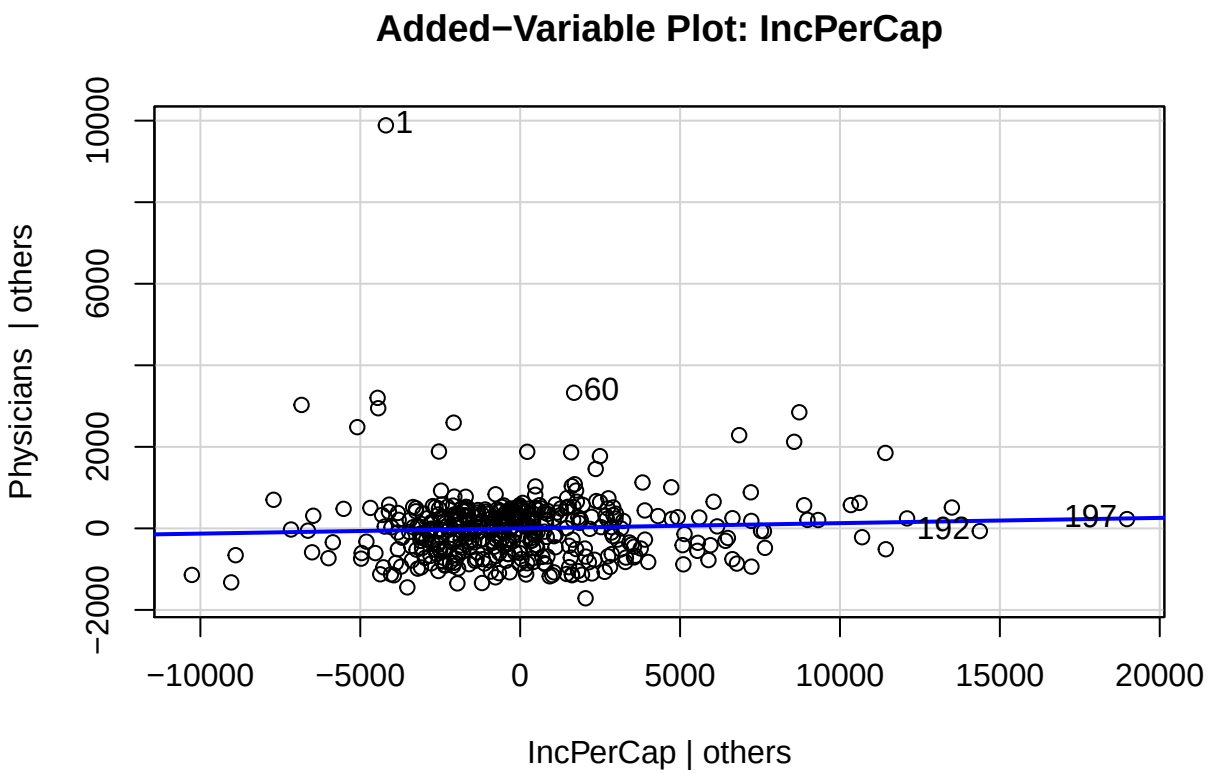
We can look at the marginal relationships after removing the effects of the other predictors with an added variable plot

```
avPlots(full.lm)
```



When adjusting for all the other predictors, physicians has a positive relationship with the logarithm of the total population. Additionally the added variable plot for land area shows a small negative relationship between physicians and land area when adjusting for all other predictors. It appears that physicians and income per capita have no relationship, so I will observe it more closely.

```
avPlot(full.lm, variable = IncPerCap)
```



A closer look at the added variable plot of income per capita shows that when adjusting for all other predictors, there is almost no relationship between the number of physicians and the income per capita. Running the summary on the linear model will show us relevant constants in the regression:

```
summary(full.lm)
```

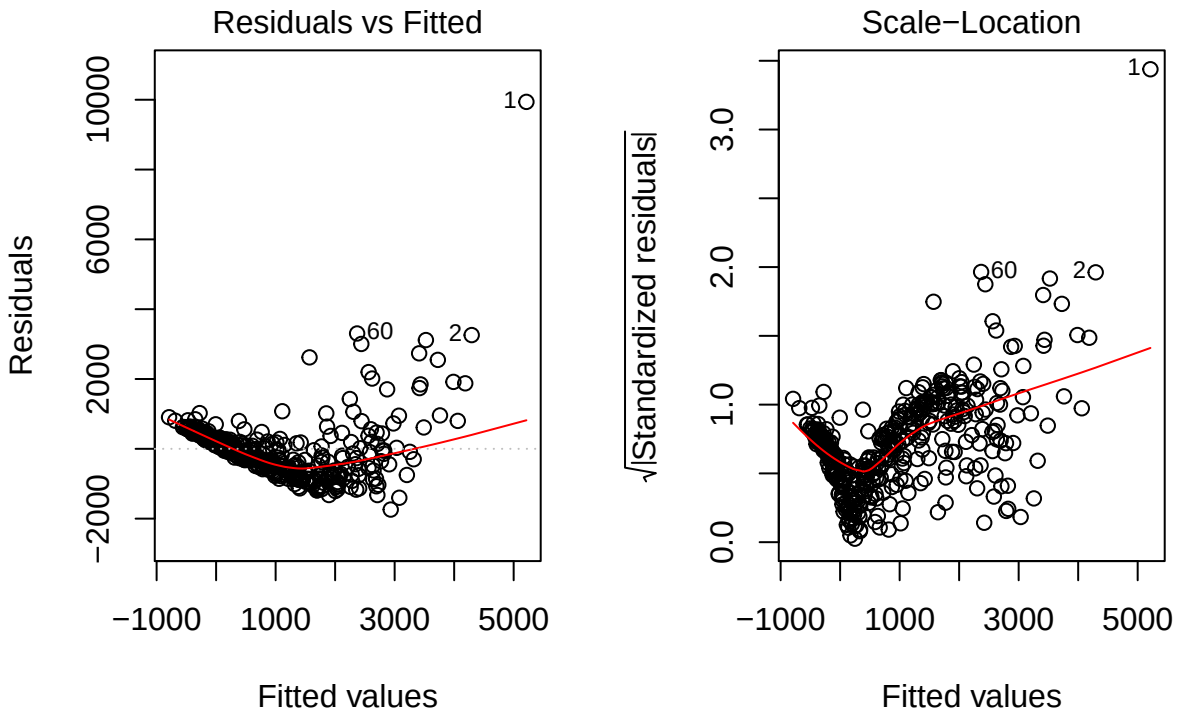
```
##
## Call:
## lm(formula = Physicians ~ log(TotalPop) + LandArea + IncPerCap)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1739.9  -495.4    -5.4   375.4  9938.9
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -1.706e+04  7.060e+02 -24.165   <2e-16 ***
## log(TotalPop)  1.427e+03  6.293e+01  22.683   <2e-16 ***
## LandArea      -5.488e-02  2.865e-02  -1.916   0.0561 .
## 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
```

The summary output shows that $R^2 = .6202$. This means that 62.02% of the variability in physicians can be be explained by land area, income per capita and the logarithm of the total population.
Interpretation of the regression coefficients:

- The expected number of physicians is -17060 when all predictors are zero. This does not have much meaning, because if all the other predictors are zero, then the city does not exist.
- When all other predictors are held constant, 30.23 physicians is the expected change when there is a 5% increase in total population.
- When all other predictors are held constant, the number of physicians is expected to go down by 1 when there is a 20 square mile increase in land area.
- The number of physicians is expected to go up by 1 with every 77.8 increase in income per capita, all other predictors held constant.

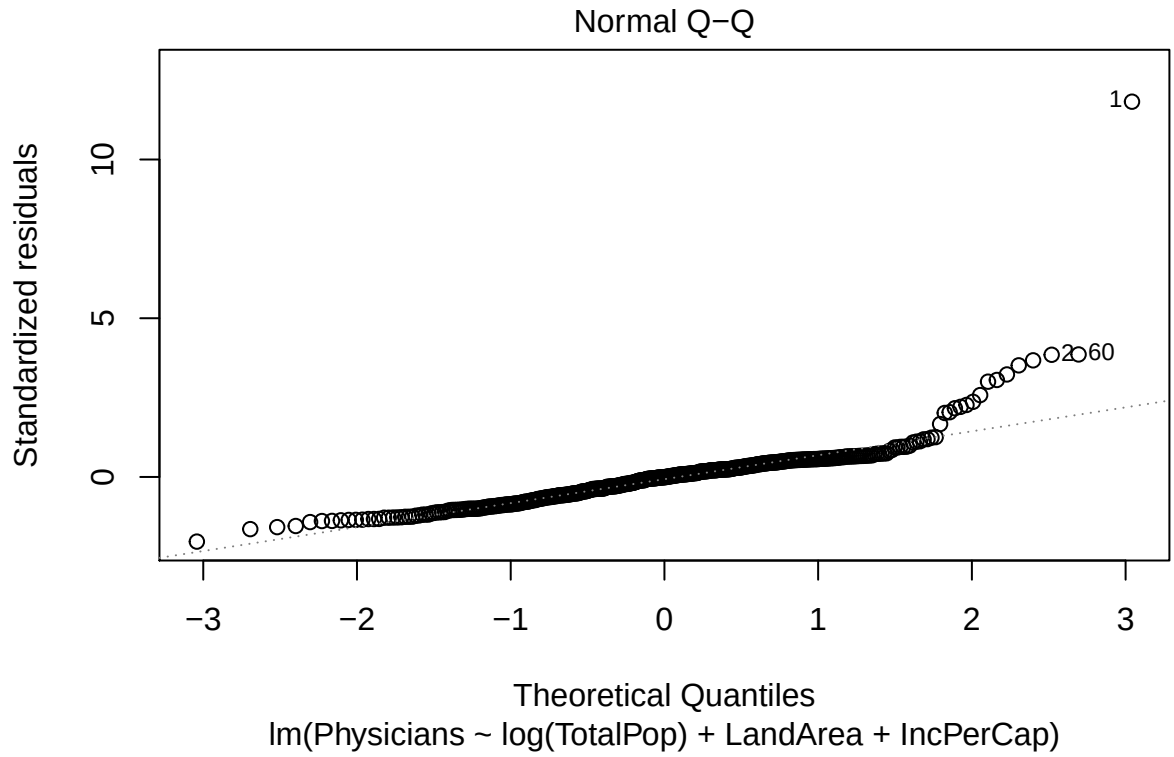
It seems that in this fit, log(TotalPop) and IncPerCap have a positive relationship with the number of physicians, meanwhile LandArea has a negative relationship with the number of physicians.
After checking the fit, we will want to see whether the fit meets the linear assumptions. There are 4 key assumptions for linear regression: Linearity, Independence, Normality, and Equal Variance. We asses Linearity, Normality and Equal Variance below with the use of multiple residual diagnostic plots:
Linearity:

```
par(mfrow = c(1,2))
plot(full.lm, which = 1)
plot(full.lm, which = 3)
```



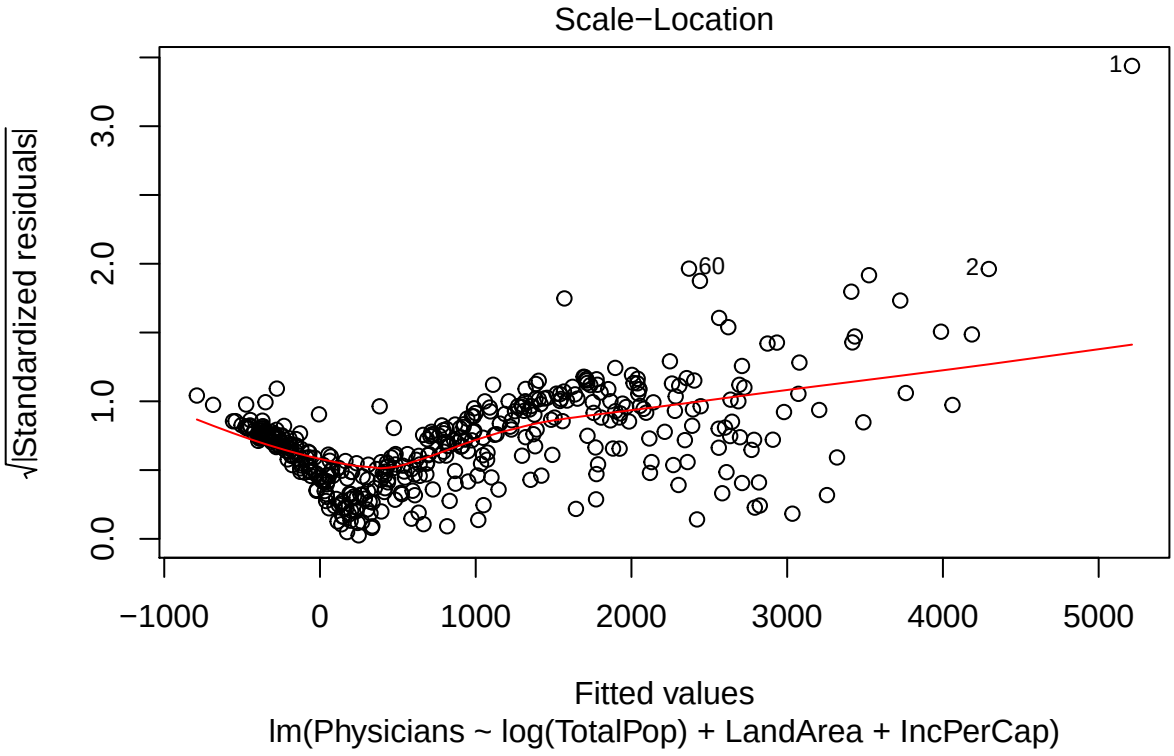
Looking at the Residuals vs. Fitted plot and the Scale-Location plot reveals that linearity is violated.
Normality:

```
plot(full.lm, which = 2)
```



Observing the Normal Q-Q plot reveals that the data has a right skew because of the upward-opening shape of the points.
Equal Variance for ϵ_i, σ^2:

```
plot(full.lm, which = 3)
```



Again, observing the scale-location plot reveals that equal variance for ϵ_i, σ^2 is violated because the points

are not equally scattered.
Because all of the linear regression assumptions were violated, It would be proper procedure to attempt transformations to correct these violations.

```
pwrtransform <-  powerTransform(cbind(TotalPop, LandArea ,IncPerCap) ~ 1, CDI )
summary(pwrtransform)
```

```
## bcPower Transformations to Multinormality
##           Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## TotalPop    -0.5589        -0.5      -0.6941      -0.4236
## LandArea    -0.0080         0.0      -0.0727       0.0567
## IncPerCap   -0.3156        -0.5      -0.5978      -0.0334
##
## Likelihood ratio test that transformation parameters are equal to 0
##  (all log transformations)
##                                 LRT df       pval
## LR test, lambda = (0 0 0) 81.81138  3 < 2.22e-16
##
## Likelihood ratio test that no transformations are needed
##                                 LRT df       pval
## LR test, lambda = (1 1 1) 1714.683  3 < 2.22e-16
```

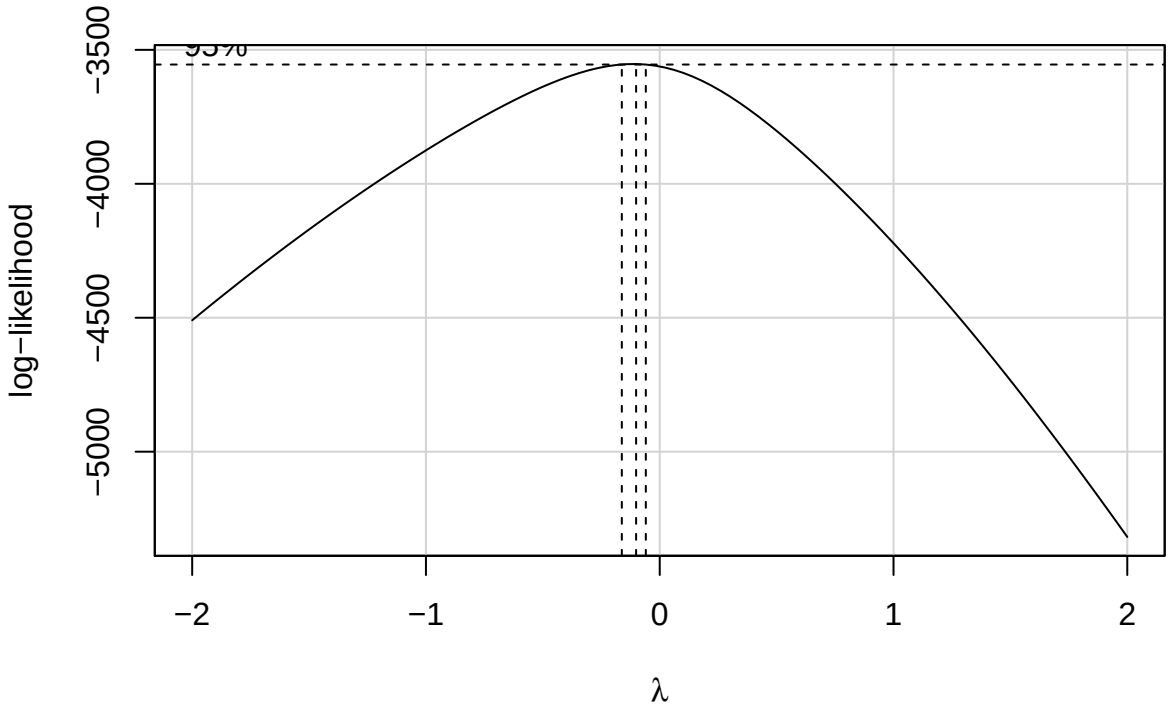
Running a power transform on the data reveals that recommended values of lambda for total population, land area, and income per capita are -0.5, 0, and -0.5, respectively. The lower bound for the transformation on income per capita is very close to 0, so I am going to choose a log transformation for a more easily interpretable model.

```
testTransform(pwrtransform, lambda = c(-.5,0,0))
```

```
##                                  LRT df    pval
## LR test, lambda = (-0.5 0 0) 5.50348  3 0.13843
```

Testing the values of lambda corresponding to each predictor shows that it will be a proper transformation. of all variables.
Next, I will use the box-cox method to determine the best transformation for physicians.

```
phystrnsfrm.lm <- lm(Physicians~ I(TotalPop^(-1/2))  +log(LandArea) + log(IncPerCap))
boxCox(phystrnsfrm.lm)
```

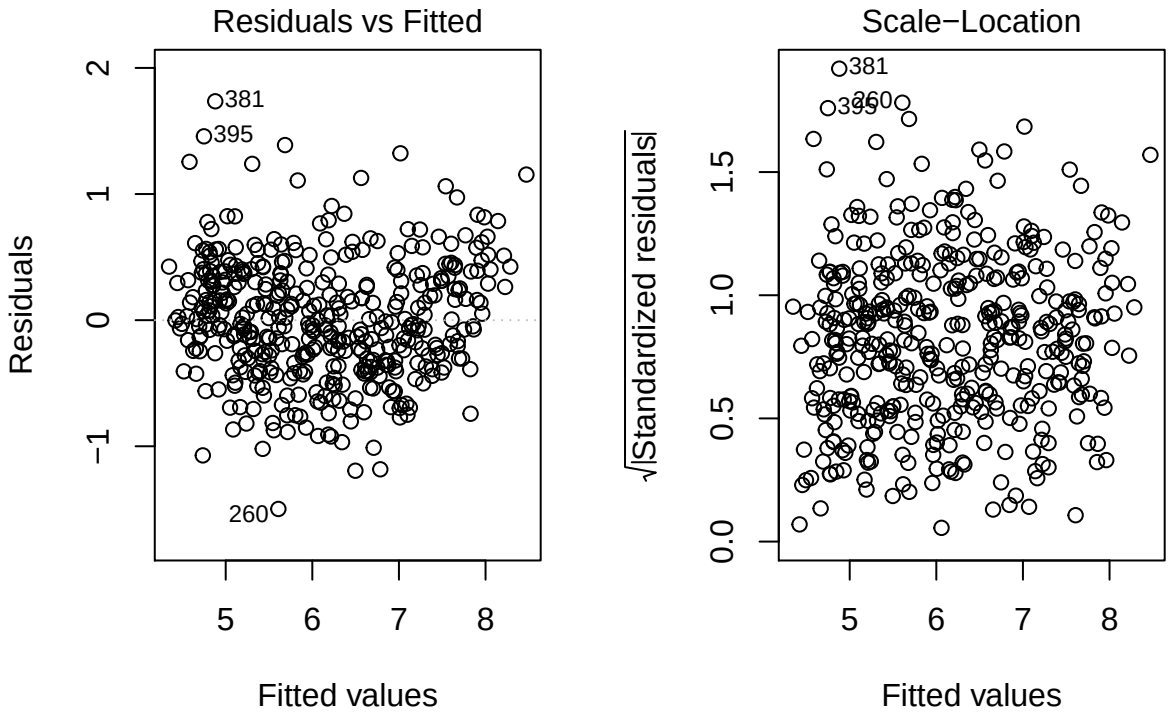


Again, the confidence interval for lambda is very close to 0, so I will choose a log transformation for physicians as well.
Now to create a linear model of the final transformation and recheck the diagnostics.

```
final.tran <- lm(log(Physicians)~ I(TotalPop^(-1/2)) + log(LandArea) + log(IncPerCap) )
```

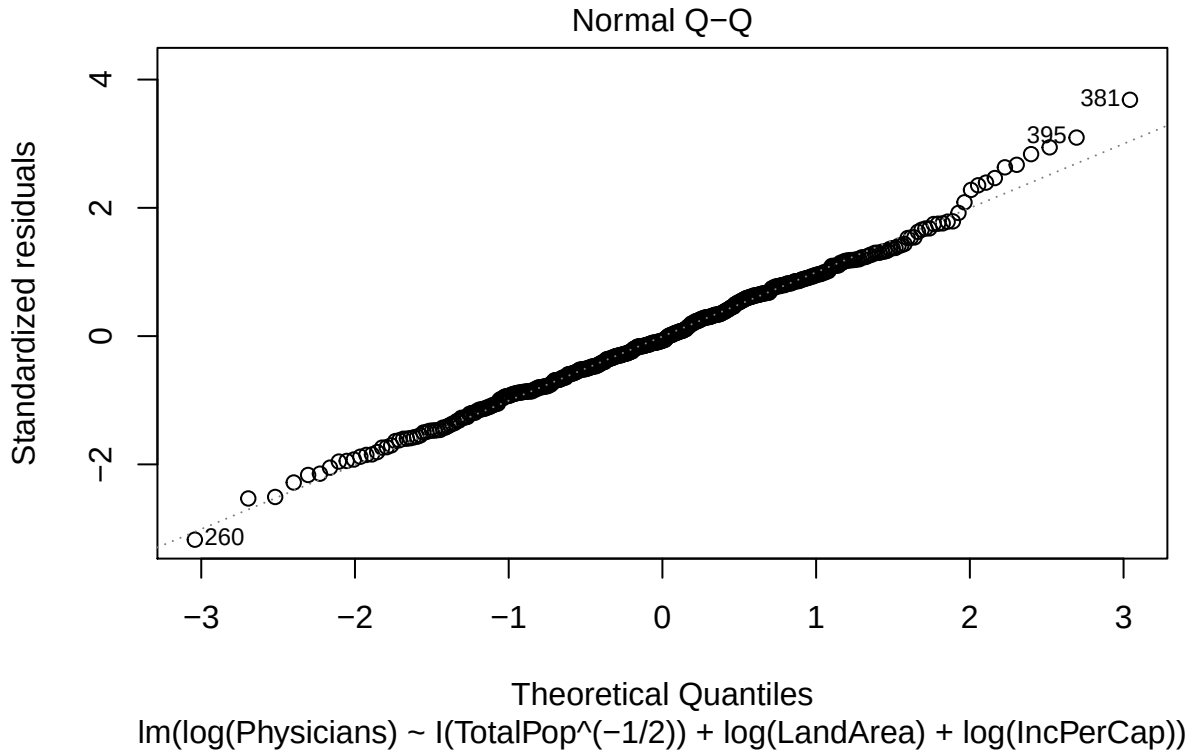
Linearity

```
par(mfrow = c(1,2))
plot(final.tran, add.smooth = FALSE, which = 1)
plot(final.tran, add.smooth = FALSE, which = 3)
```



The residuals vs fitted plot has improved immensely from the untransfromed linear model. There is not a random scattering of residuals, without a clear trend in residuals. The same applies to the scale-location plot, which has a random scattering of residuals.
Normality:

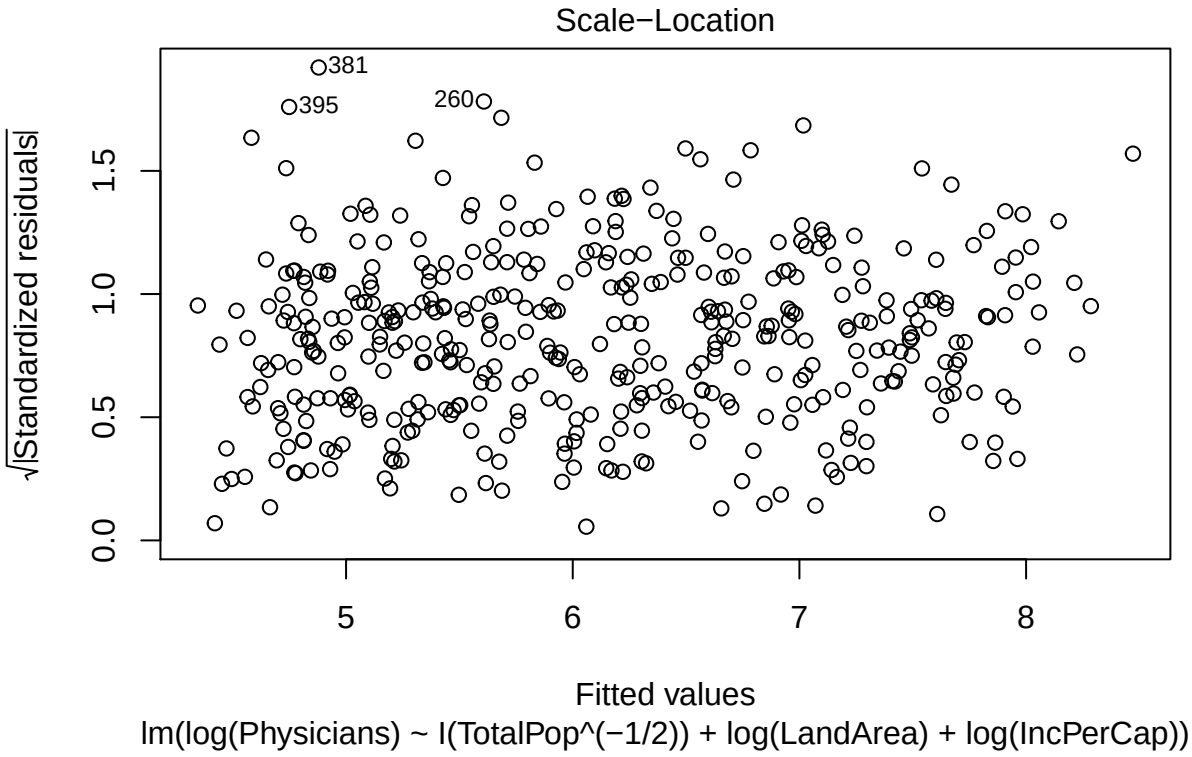
```
plot(final.tran, which = 2)
```



The normal Q-Q plot demonstrates normality.

Equal variance for ϵ_i, σ^2:

```
plot(final.tran, add.smooth = FALSE, which = 3)
```



The scale-location plot demonstrates even scattering of the $\sqrt{StandardizedResiduals}$, which suggests equal variance.

```
summary(final.tran)
```

```
##
## Call:
## lm(formula = log(Physicians) ~ I(TotalPop^(-1/2)) + log(LandArea) +
##     log(IncPerCap))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.49706 -0.31982 -0.03117  0.31495  1.73544
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         3.250e+00  1.393e+00   2.333  0.02013 *
## I(TotalPop^(-1/2)) -1.365e+03  3.792e+01 -36.004  < 2e-16 ***
## log(LandArea)      -8.281e-02  2.842e-02  -2.914  0.00376 **
## log(IncPerCap)      6.402e-01  1.312e-01   4.880 1.51e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4725 on 421 degrees of freedom
## Multiple R-squared:  0.8202, Adjusted R-squared:  0.8189
## F-statistic:   640 on 3 and 421 DF,  p-value: < 2.2e-16
```

The summary output of our transformed model shows that $R^2 = .8199$. This means that 81.99% of the variability in the log(Physicians) can be be explained by Log(LandArea), Log(IncPerCap) and log(TotalPop). Interpretation of the coefficients in this model are as follows.

- When all predictors are 0, the value of log(Physicians) is 3.25, This coefficient does not mean much because $\frac{1}{\sqrt{TotalPop}}$ is never equal to 0.
- When all predictors are held constant, as the total population increases by 1 unit, the number of $-1365\frac{1}{\sqrt{TotalPop}}$, increases by $-1365\frac{1}{\sqrt{x_i}} + 1365\frac{1}{\sqrt{x_{i-1}}}$, where x_{i-1} is the previous value of x_i. This means that as the value of population gets larger, the value of -1365 approaches 0. As a result, the value of log(Physicians) increases as the term $-1365 \times \frac{1}{\sqrt{x_i}}$ becomes less negative.
- When all other predictors are held constant, there is a -8.24% change in the expected value of physicians when LandArea is increased by 1%.
- The expected number of physicians increases by 6.37% when income per capita increases by 1%.

```
confint(final.tran)
```

```
##                              2.5 %        97.5 %
## (Intercept)             0.5114455  5.988861e+00
## I(TotalPop^(-1/2)) -1439.8107475 -1.290737e+03
## log(LandArea)         -0.1386730 -2.695424e-02
## log(IncPerCap)         0.3823317  8.980242e-01
```

Interpretations:

- I am 95% confident that the parameter β_0 is within the interval (0.511,5.988). In our case, this is not an especially important interval.
- I am 95% confident that when the total population increases by one unit, all other predictors constant, the expected change of Y_i is within the interval $(-1439.8 \times \frac{1}{\sqrt{x_i}} + 1439.8 \times \frac{1}{\sqrt{x_i+1}}, -1290.7 \times \frac{1}{\sqrt{x_i}} + 1290.7 \times \frac{1}{\sqrt{x_i+1}})x_i \neq 0$ Tthe interpretation of this coefficient is that each side of the interval is the change in the slope between the previous point and the next point.
- I am 95% confident that when land area increases by 7.5%, the expected number of physicians will change by a percentage within the interval (-.1947,-.9974), all other predictors held constant.
- I am 95% confident that when all other predictors are held constant, increasing income per capita is increased by 2.5% will increase the physicians by a percentage the interval (.9485,2.242)

To check whether there is a linear relationship, we can run a global F-test on all the predictors at an $\alpha = .01$ level. Our hypothesis will be: $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ vs. $H_0 : \beta_i \neq 0$ for some $i = 1, 2, 3$. In this case β_1 is the coefficient for $TotalPop^{-\frac{1}{2}}$, β_2 is the coefficient for $log(LandArea)$ and β_3 is the coefficient for $log(IncPerCap)$.

```
null.lm <- lm(log(Physicians)~1)
anova(null.lm, final.tran)
```

```
## Analysis of Variance Table
##
## Model 1: log(Physicians) ~ 1
## Model 2: log(Physicians) ~ I(TotalPop^(-1/2)) + log(LandArea) + log(IncPerCap)
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    424 522.62
## 2    421  93.98  3    428.64 640.04 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value for this test is 2.2×10^{-16} , therefore, I reject the null hypothesis and suggest there is a linear relationship between the predictors and log(physicians).
Finally, to check whether the variance increases or decreases with $TotalPop^{-\frac{1}{2}}$ included.
H_0 : The model has constant variance vs H_a : The variance decreases with $TotalPop^{-\frac{1}{2}}$.

```
res <- final.tran$residuals
#par(mfrow = c(1,3))
#plot(I(TotalPop^(-1/2)), res)
#plot(log(LandArea), res)
#plot(log(IncPerCap), res)

ncvTest(final.tran, ~ I(log(LandArea) + log(IncPerCap)))
```

```
## Non-constant Variance Score Test
## Variance formula: ~ I(log(LandArea) + log(IncPerCap))
## Chisquare = 5.875905, Df = 1, p = 0.015349
```

```
ncvTest(final.tran, ~ -I(TotalPop^(-1/2)) + log(LandArea) + log(IncPerCap))
```

```
## Non-constant Variance Score Test
## Variance formula: ~ -I(TotalPop^(-1/2)) + log(LandArea) + log(IncPerCap)
## Chisquare = 6.00296, Df = 2, p = 0.049713
```

Because the p value for the non-constant variance test is greater than $\alpha = .05$, I failed to reject the null and concluded there is constant variance with $TotalPop^{-\frac{1}{2}}$

4.1.3 Model 1 Final Analysis:

After working with our data, we came to some conclusions on how different predictors effected the logarithm of physicians within counties. After doing some exploratory analysis into proper transformations, we determined $log(LandArea)$, $Log(IncomePerCap)$, and the reciprocal square root of the total population ($\frac{1}{\sqrt{TotalPopulation}}$) would be the most appropriate transformations for analyzing the logarithm of physicians. From our analysis, we determined that $\frac{1}{\sqrt{TotalPopulation}}$ and $log(Physicians)$ have the highest correlation. This is logical because as the population increases there are more patients to treat, as well as a grater pool of trained physicians in the workforce. Similarly, when income per capita increases the number of physicians increases. An interesting insight from our analysis is that physicians had a negative relationship with land area. This is intuitively correct because land area does not interact with total population; a county with large land area could be dense similar to Los Angeles, or it could be desolate similar to many cities in the midwest. Through linear models and testing, we were able to confirm our initial hypothesis about the data.

4.2 Model 2

The model $Physicians \sim TotalPop + Region$ is also of interest, and will allow us to look at another factor, Region, and its effect on the number of physicians.

```
fit1 <- lm(Physicians ~ TotalPop + Region, data = CDI)
summary(fit1)
```

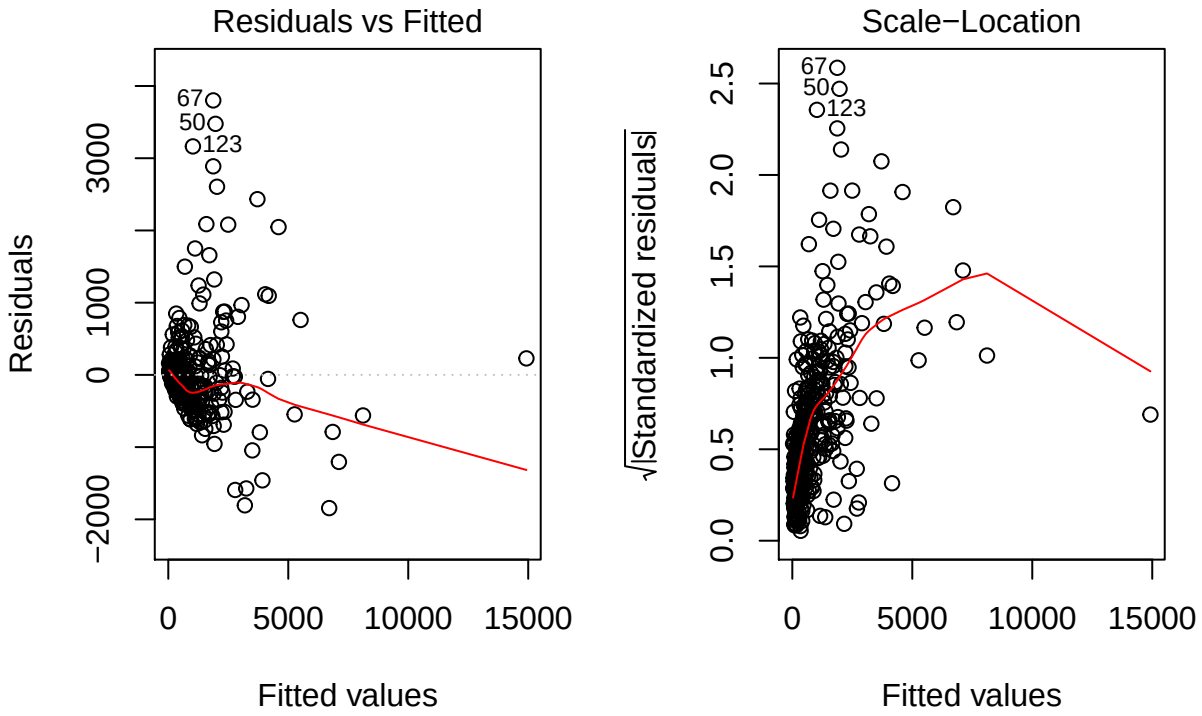
```
##
## Call:
## lm(formula = Physicians ~ TotalPop + Region, data = CDI)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -1844.2  -218.7   -62.9    66.6  3800.5
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.706e+01  7.447e+01  -0.363   0.7165
## TotalPop     2.952e-03  6.453e-05  45.748   <2e-16 ***
## Region      -5.927e+01  2.675e+01  -2.216   0.0272 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 570.7 on 422 degrees of freedom
## Multiple R-squared:  0.8322, Adjusted R-squared:  0.8314
## F-statistic:  1047 on 2 and 422 DF,  p-value: < 2.2e-16
```

4.2.1 Diagnostic check:

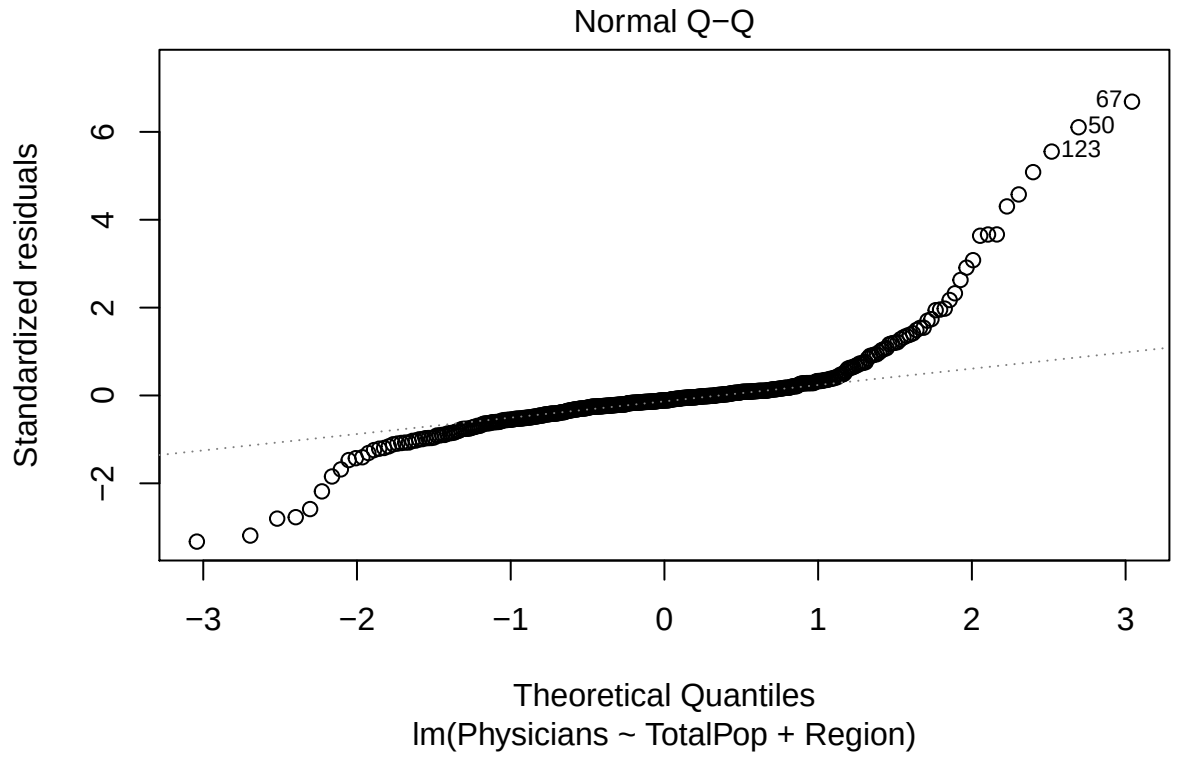
Linearity:

```
par(mfrow = c(1,2))
plot(fit1, which = 1)
plot(fit1, which = 3)
```



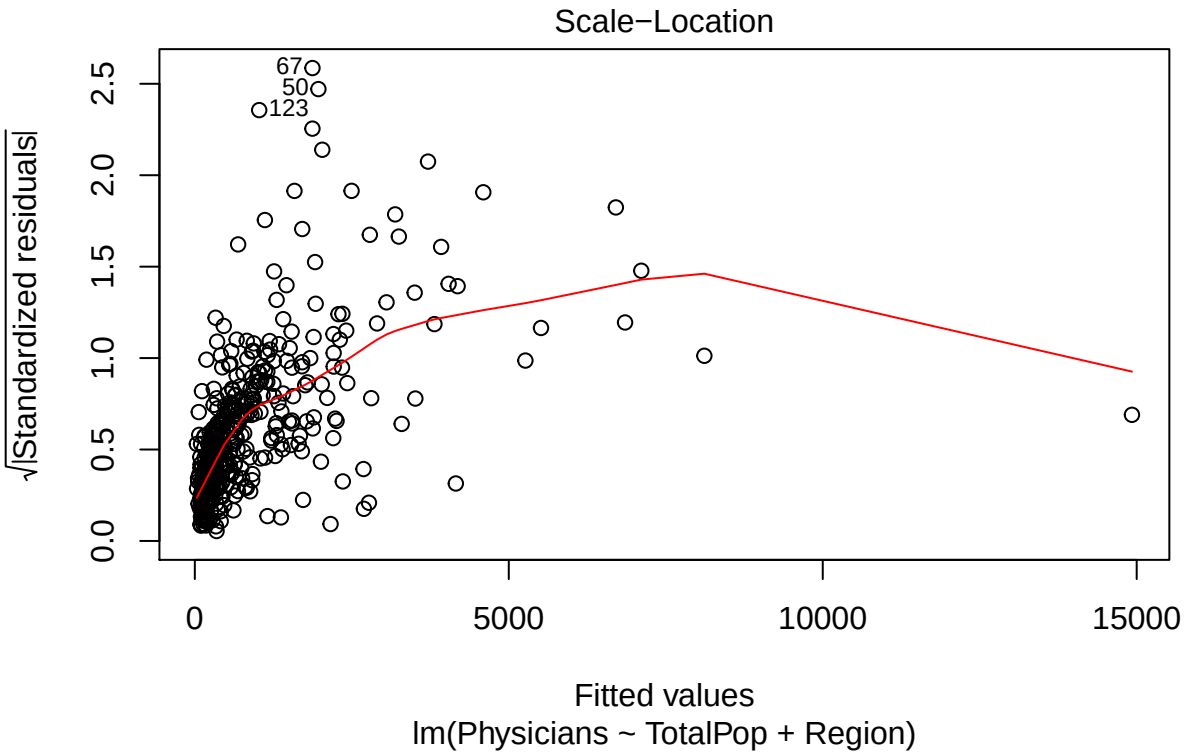
After observing the residuals vs. fitted and scale-location plots, we can see that linearity is violated.
Normality assumption:

```
plot(fit1, which = 2)
```



The Q-Q plot has heavy tails, so we know that the normality assumption is violated.
Equal Variance Assumption:

```
plot(fit1, which = 3)
```



The scale-location plot shows that equal variance is violated.

4.2.2 Variable transformation

```
pwrtrans.fit1 <- powerTransform(cbind(TotalPop, Region)~1, CDI)
summary(pwrtrans.fit1)
```

```
## bcPower Transformations to Multinormality
##          Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## TotalPop   -0.5798        -0.5      -0.7207      -0.4390
## Region      0.7945         1.0       0.5677       1.0213
##
## Likelihood ratio test that transformation parameters are equal to 0
##  (all log transformations)
##                               LRT df       pval
## LR test, lambda = (0 0) 124.3438  2 < 2.22e-16
##
## Likelihood ratio test that no transformations are needed
##                               LRT df       pval
## LR test, lambda = (1 1) 762.0031  2 < 2.22e-16
```

```
pwrtrans.dat <- with(CDI, data.frame(Physicians, I(TotalPop^(-1/2)), Region))

region.tran <- lm(Physicians ~ I(TotalPop^(-1/2)) + Region, data = pwrtrans.dat)

summary(powerTransform(region.tran))
```

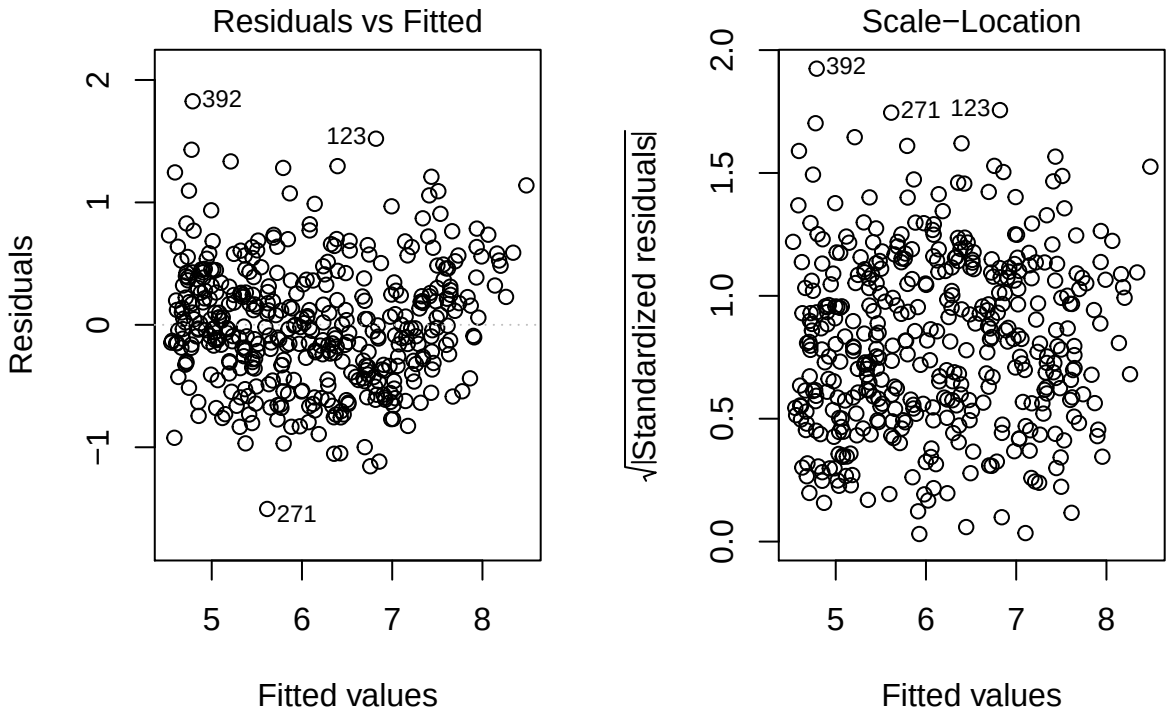
```
## Warning in model.matrix.default(mt, mf, contrasts): non-list contrasts
## argument ignored
```

```
## bcPower Transformation to Normality
##    Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## Y1   -0.1191       -0.12      -0.1733       -0.065
##
## Likelihood ratio test that transformation parameter is equal to 0
##  (log transformation)
##                             LRT df       pval
## LR test, lambda = (0) 19.19385  1 1.1809e-05
##
## Likelihood ratio test that no transformation is needed
##                             LRT df       pval
## LR test, lambda = (1) 1304.455  1 < 2.22e-16
```

The interval for the box-Cox method has $\lambda = -0.065$, which is very close to 0. To make interpretability of the model easier, I will choose $\lambda = 0$, which will be a logarithmic transform of the response.

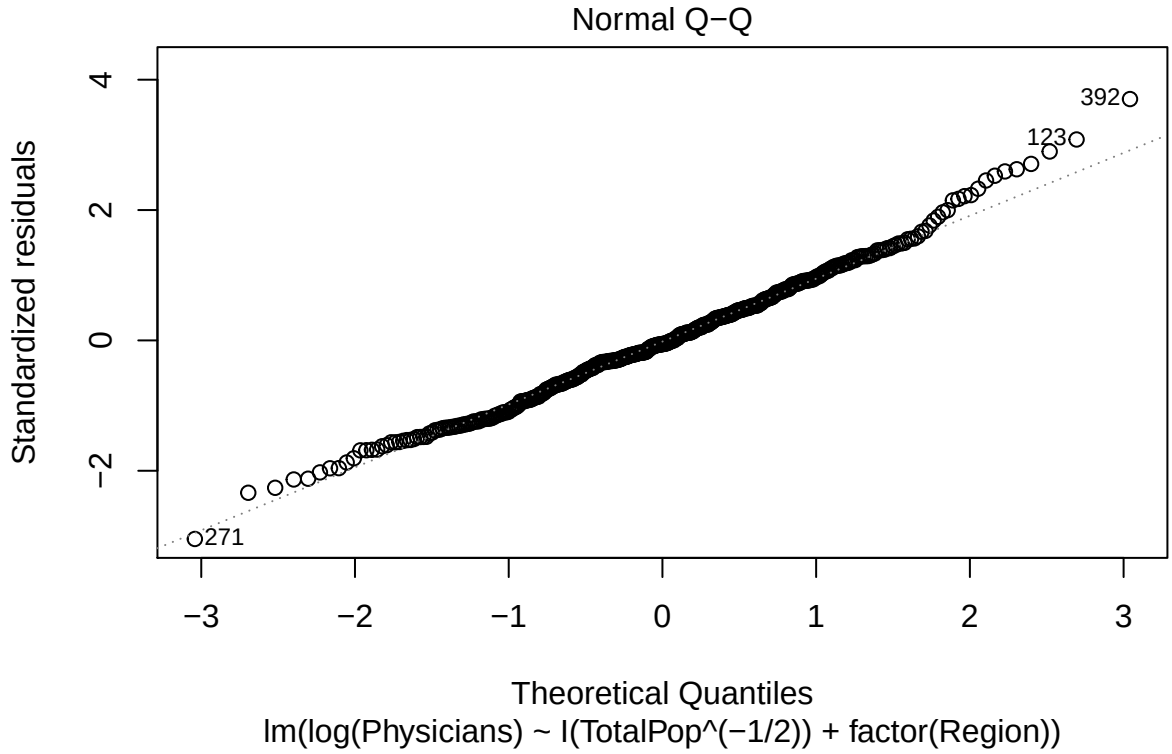
4.2.3 Transformation Checking and Model Exploration

```
region.final <- lm(log(Physicians) ~ I(TotalPop^(-1/2)) + factor(Region), data = CDI)
par(mfrow = c(1,2))
plot(region.final, add.smooth = FALSE, which = 1)
plot(region.final, add.smooth = FALSE, which = 3)
```



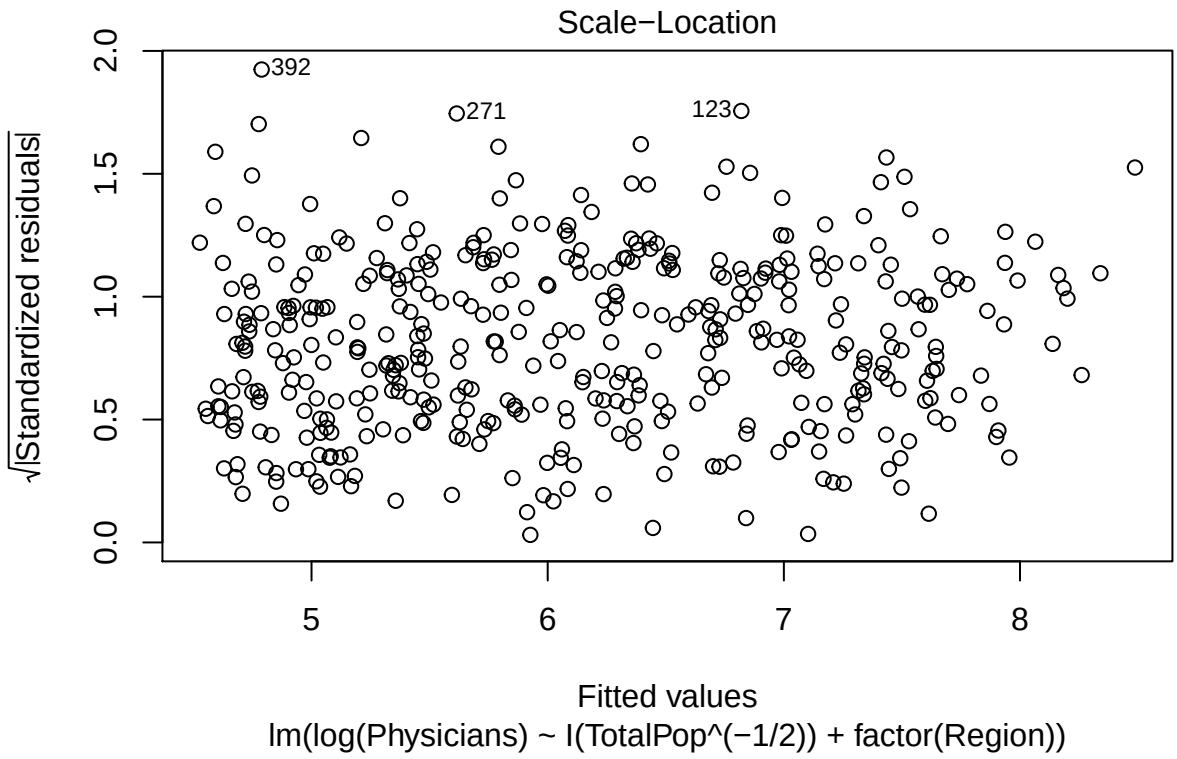
The residuals vs fitted plot has improved immensely from the transformed linear model. There is a random scattering of residuals, without a clear trend in residuals. The same applies to the scale-location plot, which has a random scattering of residuals.

```
plot(region.final, which = 2)
```



The new Q-Q plot demonstrates normality.

```
plot(region.final, add.smooth = FALSE, which = 3)
```



The $\sqrt{StarndardizedResiduals}$ are randomly scattered, demonstrating equal variance.

```
nonpar.model <- lm(log(Physicians) ~ factor(Region)*I(TotalPop^(-1/2)), data = CDI)
summary(nonpar.model)
```

```
##
## Call:
## lm(formula = log(Physicians) ~ factor(Region) * I(TotalPop^(-1/2)),
##     data = CDI)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.49382 -0.34016 -0.02396  0.30388  1.87114
##
## Coefficients:
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)                          9.282e+00  1.560e-01  59.516   <2e-16
## factor(Region)2                     -1.132e-02  2.352e-01  -0.048    0.962
## factor(Region)3                     -1.296e-01  2.096e-01  -0.618    0.537
## factor(Region)4                     -2.833e-01  2.284e-01  -1.241    0.215
## I(TotalPop^(-1/2))                  -1.486e+03  7.407e+01 -20.065   <2e-16
## factor(Region)2:I(TotalPop^(-1/2)) -3.131e+01  1.055e+02  -0.297    0.767
## factor(Region)3:I(TotalPop^(-1/2))  5.615e+01  9.650e+01   0.582    0.561
## factor(Region)4:I(TotalPop^(-1/2))  9.401e+01  1.094e+02   0.859    0.391
##
## (Intercept)                         ***
## factor(Region)2
## factor(Region)3
```

```
## factor(Region)4
## I(TotalPop^(-1/2))                  ***
## factor(Region)2:I(TotalPop^(-1/2))
## factor(Region)3:I(TotalPop^(-1/2))
## factor(Region)4:I(TotalPop^(-1/2))
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4969 on 417 degrees of freedom
## Multiple R-squared:  0.803,  Adjusted R-squared:  0.7996
## F-statistic: 242.8 on 7 and 417 DF,  p-value: < 2.2e-16
```

There is no interaction between total population and region, so we will use a parallel model. The model is parallel because the total pop does not have a different affect on the log(physicians) when in one region as opposed to another. The mean functions will look like the following. Group 0 is region 1 (NE), group 1 is region 2 (NC), group 2 is region 3 (S), and group 3 is region 4 (W).

```
par.model <- lm(log(Physicians) ~ factor(Region) + I(TotalPop^(-1/2)), data = CDI)
summary(par.model)
```

```
##
## Call:
## lm(formula = log(Physicians) ~ factor(Region) + I(TotalPop^(-1/2)),
##     data = CDI)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.50401 -0.32945 -0.02823  0.31192  1.82588
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          9.221e+00  8.710e-02 105.862   <2e-16 ***
## factor(Region)2     -8.957e-02  7.009e-02  -1.278    0.202
## factor(Region)3     -1.321e-02  6.469e-02  -0.204    0.838
## factor(Region)4     -9.894e-02  7.581e-02  -1.305    0.193
## I(TotalPop^(-1/2)) -1.456e+03  3.587e+01 -40.581   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4961 on 420 degrees of freedom
## Multiple R-squared:  0.8022, Adjusted R-squared:  0.8003
## F-statistic: 425.8 on 4 and 420 DF,  p-value: < 2.2e-16
```

Group	Mean Function
0	$9.221 + -1456 \times \frac{1}{\sqrt{TotalPop}}$
1	$8.541 - .08957 - 1456 \times \frac{1}{\sqrt{TotalPop}}$
2	$8.541 - .01321 - 1456 \times \frac{1}{\sqrt{TotalPop}}$
3	$8.541 - .09894 - 1456 \times \frac{1}{\sqrt{TotalPop}}$

Fitting the model with region included allows us to verify whether it is significant in our regression.

```
summary(lm(log(Physicians) ~ Region + I(TotalPop^(-1/2)), data = CDI))
```

```
##
## Call:
## lm(formula = log(Physicians) ~ Region + I(TotalPop^(-1/2)), data = CDI)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.55742 -0.32961 -0.03473  0.30852  1.77342
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          9.225e+00  9.771e-02  94.403   <2e-16 ***
## Region              -1.793e-02  2.326e-02  -0.771    0.441
## I(TotalPop^(-1/2)) -1.457e+03  3.536e+01 -41.211   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4965 on 422 degrees of freedom
## Multiple R-squared:  0.801,  Adjusted R-squared:    0.8
## F-statistic: 849.2 on 2 and 422 DF,  p-value: < 2.2e-16
```

Carrying out a t-test on the hypothesis $\beta_{region} = 0$ vs. $\beta_{region} \neq 0$ at an $\alpha = .05$ significance level shows that at a p-value .441>.05, I fail to reject the null, and conclude that $\beta_{region} = 0$. I will not remove it from the model.

```
non.region <- lm(log(Physicians) ~ I(TotalPop^(-1/2)), data = CDI)
summary(non.region)
```

```
##
## Call:
## lm(formula = log(Physicians) ~ I(TotalPop^(-1/2)), data = CDI)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.5493 -0.3276 -0.0192  0.2987  1.7813
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          9.179e+00  7.826e-02  117.29   <2e-16 ***
## I(TotalPop^(-1/2)) -1.457e+03  3.534e+01  -41.22   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4962 on 423 degrees of freedom
## Multiple R-squared:  0.8007, Adjusted R-squared:  0.8002
## F-statistic:  1699 on 1 and 423 DF,  p-value: < 2.2e-16
```

There are many other variables such as crimes, bachelor's degrees, and poverty that can build a better explained model. We will begin to explain different models through AIC and BIC model selection. We will be choosing from the following variables:

Variable	Description
Pop65	Percent of 1990 CDI population aged 65 years and older
Crimes	Total number of serious crimes in 1990, as reported by law enforcement agencies
Bachelor	Percent of adult population (25 years or older) with a bachelor's degree
Poverty	Percent of 1990 CDI population with income below poverty level
PersonalInc	Total income of 1990 CDI population (in millions of dollars)

The first exploration into model selection is going to be with forwards AIC:

```
#AIC
mod.full <- lm(log(Physicians) ~ I(TotalPop^(-1/2)) + Pop65 + Crimes +
                Bachelor + Poverty + PersonalInc, data = CDI)
mod.0 <- lm(log(Physicians) ~I(TotalPop^(-1/2)) , data = CDI)
step(mod.0, scope = list(lower = mod.0, upper = mod.full), direction = "forward")
```

```
## Start:  AIC=-593.62
## log(Physicians) ~ I(TotalPop^(-1/2))
##
##               Df Sum of Sq     RSS     AIC
## + Bachelor     1   14.0197  90.140 -653.06
## + PersonalInc  1    9.8940  94.265 -634.04
## + Crimes       1    4.7517  99.408 -611.47
## + Pop65        1    1.1671 102.992 -596.41
## + Poverty      1    1.1292 103.030 -596.25
## <none>                     104.159 -593.62
##
## Step:  AIC=-653.06
## log(Physicians) ~ I(TotalPop^(-1/2)) + Bachelor
##
##               Df Sum of Sq    RSS     AIC
## + Poverty      1    8.2651 81.875 -691.93
## + PersonalInc  1    7.9233 82.216 -690.16
## + Crimes       1    6.9874 83.152 -685.35
## + Pop65        1    6.7320 83.408 -684.05
## <none>                     90.140 -653.06
##
## Step:  AIC=-691.93
## log(Physicians) ~ I(TotalPop^(-1/2)) + Bachelor + Poverty
##
##               Df Sum of Sq    RSS     AIC
## + Pop65        1    9.8708 72.004 -744.53
## + PersonalInc  1    8.8518 73.023 -738.56
## + Crimes       1    4.4318 77.443 -713.58
## <none>                     81.875 -691.93
##
## Step:  AIC=-744.53
## log(Physicians) ~ I(TotalPop^(-1/2)) + Bachelor + Poverty + Pop65
##
```

```
##               Df Sum of Sq    RSS     AIC
## + PersonalInc  1    8.4323 63.572 -795.47
## + Crimes       1    4.8719 67.132 -772.31
## <none>                     72.004 -744.53
##
## Step:  AIC=-795.47
## log(Physicians) ~ I(TotalPop^(-1/2)) + Bachelor + Poverty + Pop65 +
##     PersonalInc
##
##          Df Sum of Sq    RSS     AIC
## <none>                 63.572 -795.47
## + Crimes  1   0.12983 63.442 -794.34

##
## Call:
## lm(formula = log(Physicians) ~ I(TotalPop^(-1/2)) + Bachelor +
##     Poverty + Pop65 + PersonalInc, data = CDI)
##
## Coefficients:
##        (Intercept)  I(TotalPop^(-1/2))            Bachelor
##          6.484e+00          -1.085e+03           4.327e-02
##            Poverty               Pop65         PersonalInc
##          4.087e-02           4.038e-02           2.162e-05
```

Forwards AIC seems to choose all predictors other than Crimes.
Next, we will explore backwards AIC to see whether it yield a distinct model.

```
step(mod.full, scope = list(lower = mod.0, upper = mod.full), direction = "backward")
```

```
## Start:  AIC=-794.34
## log(Physicians) ~ I(TotalPop^(-1/2)) + Pop65 + Crimes + Bachelor +
##     Poverty + PersonalInc
##
##               Df Sum of Sq    RSS     AIC
## - Crimes       1    0.1298 63.572 -795.47
## <none>                     63.442 -794.34
## - PersonalInc  1    3.6903 67.132 -772.31
## - Pop65        1    9.5543 72.996 -736.72
## - Poverty      1   10.5759 74.018 -730.81
## - Bachelor     1   27.7979 91.240 -641.91
##
## Step:  AIC=-795.47
## log(Physicians) ~ I(TotalPop^(-1/2)) + Pop65 + Bachelor + Poverty +
##     PersonalInc
##
##               Df Sum of Sq    RSS     AIC
## <none>                     63.572 -795.47
## - PersonalInc  1    8.4323 72.004 -744.53
## - Pop65        1    9.4513 73.023 -738.56
## - Poverty      1   12.3657 75.937 -721.93
## - Bachelor     1   27.7083 91.280 -643.72

##
## Call:
## lm(formula = log(Physicians) ~ I(TotalPop^(-1/2)) + Pop65 + Bachelor +
```

```
##     Poverty + PersonalInc, data = CDI)
##
## Coefficients:
##        (Intercept)  I(TotalPop^(-1/2))              Pop65
##          6.484e+00          -1.085e+03          4.038e-02
##           Bachelor             Poverty        PersonalInc
##          4.327e-02           4.087e-02           2.162e-05
```

Finally, stepwise AIC:

```
step(mod.0, scope = list(lower = mod.0, upper = mod.full))
```

```
## Start:  AIC=-593.62
## log(Physicians) ~ I(TotalPop^(-1/2))
##
##               Df Sum of Sq     RSS     AIC
## + Bachelor     1   14.0197  90.140 -653.06
## + PersonalInc  1    9.8940  94.265 -634.04
## + Crimes       1    4.7517  99.408 -611.47
## + Pop65        1    1.1671 102.992 -596.41
## + Poverty      1    1.1292 103.030 -596.25
## <none>                     104.159 -593.62
##
## Step:  AIC=-653.06
## log(Physicians) ~ I(TotalPop^(-1/2)) + Bachelor
##
##               Df Sum of Sq     RSS     AIC
## + Poverty      1    8.2651  81.875 -691.93
## + PersonalInc  1    7.9233  82.216 -690.16
## + Crimes       1    6.9874  83.152 -685.35
## + Pop65        1    6.7320  83.408 -684.05
## <none>                      90.140 -653.06
## - Bachelor     1   14.0197 104.159 -593.62
##
## Step:  AIC=-691.93
## log(Physicians) ~ I(TotalPop^(-1/2)) + Bachelor + Poverty
##
##               Df Sum of Sq     RSS     AIC
## + Pop65        1    9.8708  72.004 -744.53
## + PersonalInc  1    8.8518  73.023 -738.56
## + Crimes       1    4.4318  77.443 -713.58
## <none>                      81.875 -691.93
## - Poverty      1    8.2651  90.140 -653.06
## - Bachelor     1   21.1555 103.030 -596.25
##
## Step:  AIC=-744.53
## log(Physicians) ~ I(TotalPop^(-1/2)) + Bachelor + Poverty + Pop65
##
##               Df Sum of Sq     RSS     AIC
## + PersonalInc  1    8.4323  63.572 -795.47
## + Crimes       1    4.8719  67.132 -772.31
## <none>                      72.004 -744.53
## - Pop65        1    9.8708  81.875 -691.93
## - Poverty      1   11.4039  83.408 -684.05
```

```
## - Bachelor     1   29.8617 101.866 -599.09
##
## Step:  AIC=-795.47
## log(Physicians) ~ I(TotalPop^(-1/2)) + Bachelor + Poverty + Pop65 +
##     PersonalInc
##
##               Df Sum of Sq    RSS     AIC
## <none>                     63.572 -795.47
## + Crimes       1    0.1298 63.442 -794.34
## - PersonalInc  1    8.4323 72.004 -744.53
## - Pop65        1    9.4513 73.023 -738.56
## - Poverty      1   12.3657 75.937 -721.93
## - Bachelor     1   27.7083 91.280 -643.72

##
## Call:
## lm(formula = log(Physicians) ~ I(TotalPop^(-1/2)) + Bachelor +
##     Poverty + Pop65 + PersonalInc, data = CDI)
##
## Coefficients:
##        (Intercept)  I(TotalPop^(-1/2))            Bachelor
##          6.484e+00          -1.085e+03           4.327e-02
##            Poverty               Pop65         PersonalInc
##          4.087e-02           4.038e-02           2.162e-05
```

Both forwards, backwards, and stepwise AIC yielded the same models.
Trying BIC with k=log(n) ,where n is the number of observations, may yield a smaller model as the sample size is taken into account.
Forwards BIC:

```
step(mod.0, scope = list(lower = mod.0, upper = mod.full),
     direction = "forward", k=log(length(Physicians)))
```

```
## Start:  AIC=-585.52
## log(Physicians) ~ I(TotalPop^(-1/2))
##
##               Df Sum of Sq     RSS     AIC
## + Bachelor     1   14.0197  90.140 -640.90
## + PersonalInc  1    9.8940  94.265 -621.88
## + Crimes       1    4.7517  99.408 -599.31
## <none>                     104.159 -585.52
## + Pop65        1    1.1671 102.992 -584.25
## + Poverty      1    1.1292 103.030 -584.10
##
## Step:  AIC=-640.9
## log(Physicians) ~ I(TotalPop^(-1/2)) + Bachelor
##
##               Df Sum of Sq    RSS     AIC
## + Poverty      1    8.2651 81.875 -675.72
## + PersonalInc  1    7.9233 82.216 -673.95
## + Crimes       1    6.9874 83.152 -669.14
## + Pop65        1    6.7320 83.408 -667.84
## <none>                     90.140 -640.90
##
```

```
## Step:  AIC=-675.72
## log(Physicians) ~ I(TotalPop^(-1/2)) + Bachelor + Poverty
##
##               Df Sum of Sq    RSS     AIC
## + Pop65        1    9.8708 72.004 -724.27
## + PersonalInc  1    8.8518 73.023 -718.30
## + Crimes       1    4.4318 77.443 -693.32
## <none>                     81.875 -675.72
##
## Step:  AIC=-724.27
## log(Physicians) ~ I(TotalPop^(-1/2)) + Bachelor + Poverty + Pop65
##
##               Df Sum of Sq    RSS     AIC
## + PersonalInc  1    8.4323 63.572 -771.16
## + Crimes       1    4.8719 67.132 -747.99
## <none>                     72.004 -724.27
##
## Step:  AIC=-771.16
## log(Physicians) ~ I(TotalPop^(-1/2)) + Bachelor + Poverty + Pop65 +
##     PersonalInc
##
##          Df Sum of Sq    RSS     AIC
## <none>                 63.572 -771.16
## + Crimes  1   0.12983 63.442 -765.97

##
## Call:
## lm(formula = log(Physicians) ~ I(TotalPop^(-1/2)) + Bachelor +
##     Poverty + Pop65 + PersonalInc, data = CDI)
##
## Coefficients:
##        (Intercept)  I(TotalPop^(-1/2))            Bachelor
##          6.484e+00          -1.085e+03           4.327e-02
##            Poverty               Pop65         PersonalInc
##          4.087e-02           4.038e-02           2.162e-05
```

Backwards BIC:

```
step(mod.full, scope = list(lower = mod.0, upper = mod.full),
     direction = "backward", k = log(length(Physicians)))
```

```
## Start:  AIC=-765.97
## log(Physicians) ~ I(TotalPop^(-1/2)) + Pop65 + Crimes + Bachelor +
##     Poverty + PersonalInc
##
##               Df Sum of Sq    RSS     AIC
## - Crimes       1    0.1298 63.572 -771.16
## <none>                     63.442 -765.97
## - PersonalInc  1    3.6903 67.132 -747.99
## - Pop65        1    9.5543 72.996 -712.40
## - Poverty      1   10.5759 74.018 -706.50
## - Bachelor     1   27.7979 91.240 -617.59
##
## Step:  AIC=-771.16
## log(Physicians) ~ I(TotalPop^(-1/2)) + Pop65 + Bachelor + Poverty +
```

```
##     PersonalInc
##
##               Df Sum of Sq    RSS     AIC
## <none>                     63.572 -771.16
## - PersonalInc  1    8.4323 72.004 -724.27
## - Pop65        1    9.4513 73.023 -718.30
## - Poverty      1   12.3657 75.937 -701.67
## - Bachelor     1   27.7083 91.280 -623.46

##
## Call:
## lm(formula = log(Physicians) ~ I(TotalPop^(-1/2)) + Pop65 + Bachelor +
##     Poverty + PersonalInc, data = CDI)
##
## Coefficients:
##        (Intercept)  I(TotalPop^(-1/2))              Pop65
##          6.484e+00          -1.085e+03          4.038e-02
##           Bachelor             Poverty         PersonalInc
##          4.327e-02           4.087e-02           2.162e-05
```

Stepwise BIC:

```
step(mod.0, scope = list(lower = mod.0, upper = mod.full),
     k = log(length(Physicians)))
```

```
## Start:  AIC=-585.52
## log(Physicians) ~ I(TotalPop^(-1/2))
##
##               Df Sum of Sq     RSS     AIC
## + Bachelor     1   14.0197  90.140 -640.90
## + PersonalInc  1    9.8940  94.265 -621.88
## + Crimes       1    4.7517  99.408 -599.31
## <none>                     104.159 -585.52
## + Pop65        1    1.1671 102.992 -584.25
## + Poverty      1    1.1292 103.030 -584.10
##
## Step:  AIC=-640.9
## log(Physicians) ~ I(TotalPop^(-1/2)) + Bachelor
##
##               Df Sum of Sq     RSS     AIC
## + Poverty      1    8.2651  81.875 -675.72
## + PersonalInc  1    7.9233  82.216 -673.95
## + Crimes       1    6.9874  83.152 -669.14
## + Pop65        1    6.7320  83.408 -667.84
## <none>                      90.140 -640.90
## - Bachelor     1   14.0197 104.159 -585.52
##
## Step:  AIC=-675.72
## log(Physicians) ~ I(TotalPop^(-1/2)) + Bachelor + Poverty
##
##               Df Sum of Sq     RSS     AIC
## + Pop65        1    9.8708  72.004 -724.27
## + PersonalInc  1    8.8518  73.023 -718.30
## + Crimes       1    4.4318  77.443 -693.32
## <none>                      81.875 -675.72
```

```
## - Poverty      1    8.2651  90.140 -640.90
## - Bachelor     1   21.1555 103.030 -584.10
##
## Step:  AIC=-724.27
## log(Physicians) ~ I(TotalPop^(-1/2)) + Bachelor + Poverty + Pop65
##
##               Df Sum of Sq     RSS     AIC
## + PersonalInc  1    8.4323  63.572 -771.16
## + Crimes       1    4.8719  67.132 -747.99
## <none>                      72.004 -724.27
## - Pop65        1    9.8708  81.875 -675.72
## - Poverty      1   11.4039  83.408 -667.84
## - Bachelor     1   29.8617 101.866 -582.88
##
## Step:  AIC=-771.16
## log(Physicians) ~ I(TotalPop^(-1/2)) + Bachelor + Poverty + Pop65 +
##     PersonalInc
##
##               Df Sum of Sq    RSS     AIC
## <none>                     63.572 -771.16
## + Crimes       1    0.1298 63.442 -765.97
## - PersonalInc  1    8.4323 72.004 -724.27
## - Pop65        1    9.4513 73.023 -718.30
## - Poverty      1   12.3657 75.937 -701.67
## - Bachelor     1   27.7083 91.280 -623.46

##
## Call:
## lm(formula = log(Physicians) ~ I(TotalPop^(-1/2)) + Bachelor +
##     Poverty + Pop65 + PersonalInc, data = CDI)
##
## Coefficients:
##        (Intercept)  I(TotalPop^(-1/2))            Bachelor
##          6.484e+00          -1.085e+03           4.327e-02
##            Poverty               Pop65         PersonalInc
##          4.087e-02           4.038e-02           2.162e-05
```

Again, the BIC criterion gave the same models with Crimes not included. We will then fit the model with all the variables except Crimes.
After fitting the new model, we can run a partial-F test to check whether the new model or the submodel is better:
Hypothesis: H_0 : The submodel is better vs H_a : The full model is better.

```
old.lm <- lm(log(Physicians) ~ I(TotalPop^(-1/2)) + Region, data = CDI)
new.lm <- lm(log(Physicians) ~ I(TotalPop^(-1/2)) + Region + Pop65 +
               Bachelor + Poverty + PersonalInc, data = CDI)
anova(old.lm,new.lm, data = CDI)
```

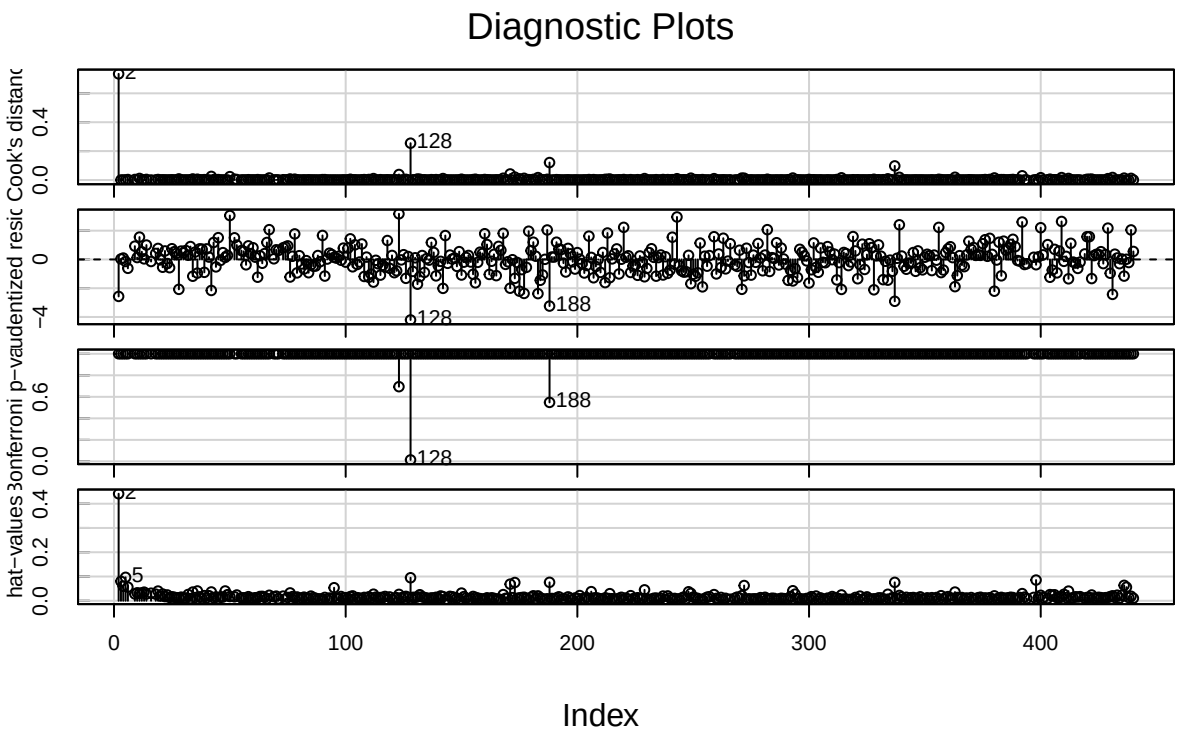
```
## Warning in anova.lmlist(object, ...): models with response '"NULL"' removed
## because response differs from model 1
```

```
## Analysis of Variance Table
##
## Model 1: log(Physicians) ~ I(TotalPop^(-1/2)) + Region
## Model 2: log(Physicians) ~ I(TotalPop^(-1/2)) + Region + Pop65 + Bachelor +
```

```
##     Poverty + PersonalInc
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1    422 104.013
## 2    418  62.619  4    41.394 69.079 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$2.2 \times 10^{-16} < 0.05$, so at an $\alpha = 0.05$ level, I reject the null hypothesis and suggest that the new model is better than the full model.
Finding any influential points in our new model will be important to choosing the best model:

```
influenceIndexPlot(new.lm)
```

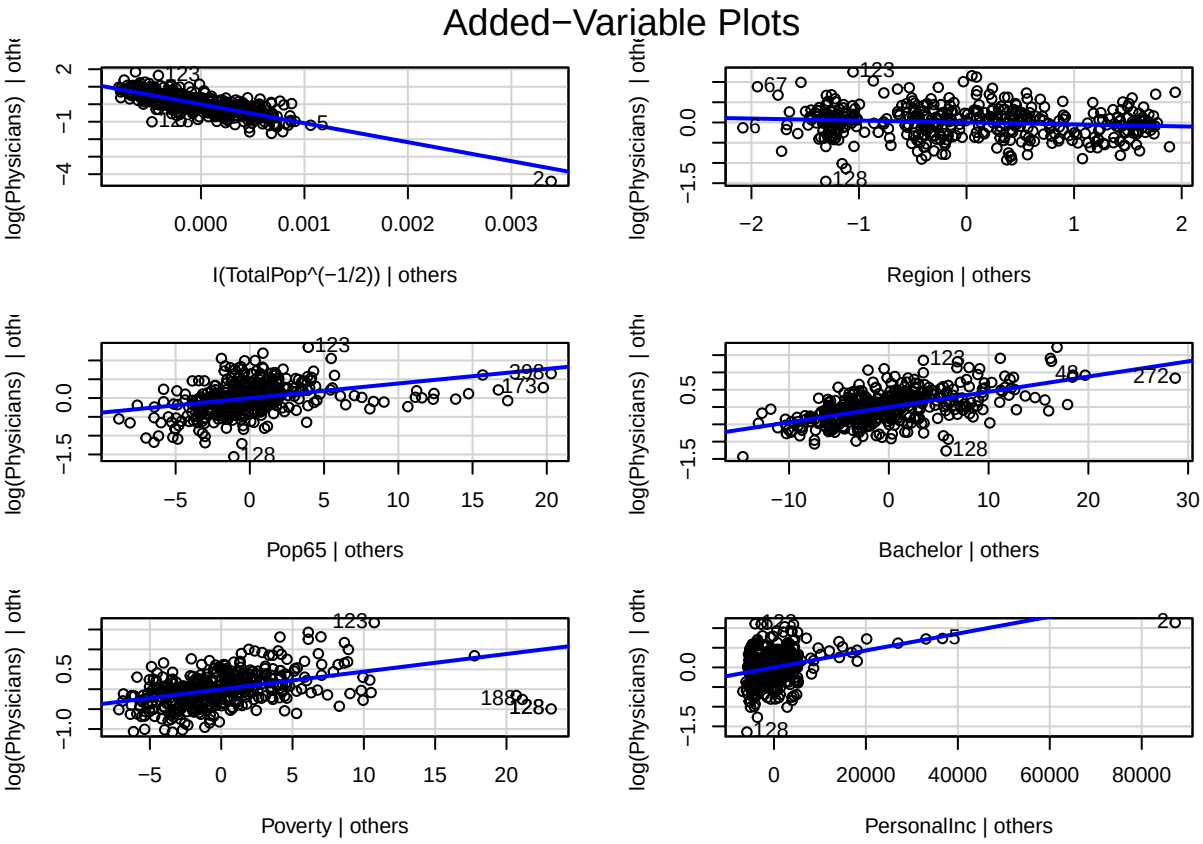



```
outlierTest(new.lm)
```

```
##      rstudent unadjusted p-value Bonferroni p
## 128 -4.205958         3.1844e-05     0.013534
```

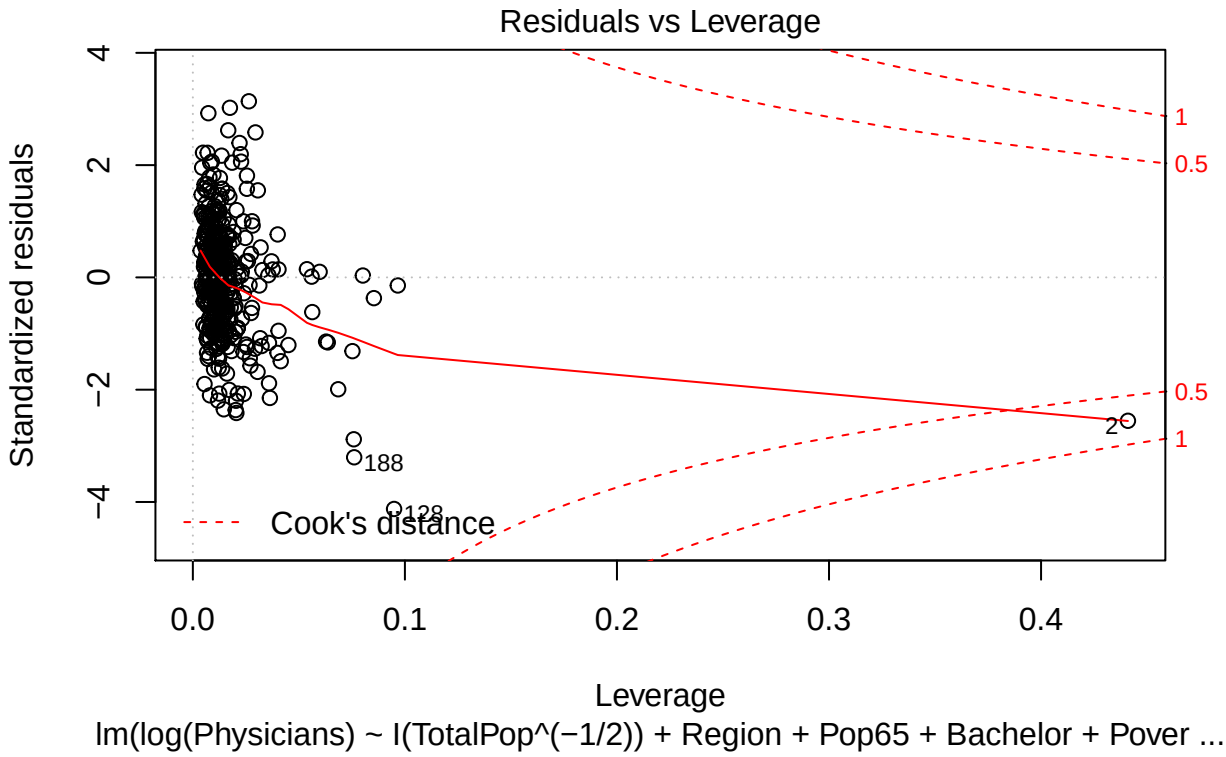
The influence index plot demonstrates that observations 2 and 128 have the largest Cook's distance. Running an outlier test shows that the Bonferoni-corrected p-value of 0.013891 for observation 128 makes it an outlier. Based on the outlier test and the Cook's distance, it would be safe to assume this is an influential point in the dataset.

```
avPlots(new.lm)
```



The added variable plots demonstrate how observations 2 and 128 are points of interest. Observation 2 has high leverage because it is an "outlier in x" as its distance from all the other points is very large. This means that h_{ii} is very large, and would explain why the Cook's distance is so large.

```
plot(new.lm, which = 5)
```




```
summary(new.lm)
```

```
##
## Call:
## lm(formula = log(Physicians) ~ I(TotalPop^(-1/2)) + Region +
##     Pop65 + Bachelor + Poverty + PersonalInc, data = CDI)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.51872 -0.23985  0.01679  0.22363  1.19797
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          6.579e+00  1.811e-01  36.322  < 2e-16 ***
## I(TotalPop^(-1/2)) -1.085e+03  4.127e+01 -26.290  < 2e-16 ***
## Region             -4.841e-02  1.920e-02  -2.521   0.0121 *
## Pop65               3.879e-02  5.123e-03   7.571  2.4e-13 ***
## Bachelor            4.400e-02  3.195e-03  13.772  < 2e-16 ***
## Poverty             4.423e-02  4.692e-03   9.427  < 2e-16 ***
## PersonalInc         2.153e-05  2.881e-06   7.471  4.7e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.387 on 418 degrees of freedom
## Multiple R-squared:  0.8802, Adjusted R-squared:  0.8785
## F-statistic: 511.8 on 6 and 418 DF,  p-value: < 2.2e-16
```

Looking at the plot of the leverage against standardized residuals demonstrates approximate values of interest with observation 2 and 128. observation 2 has a leverage of ~ 0.433 and a standardized residual of ~ 2.9. The Cook's distance of this point is ~ 0.75, which, while not above 1, would mean it is influential and

removing it would have an effect on the regression. The standardized residual of observation 128 is higher at ~ 4.15, making it an "outlier in Y." Despite this, the Cook's distance is less than .5 and would not have as large of an influence on the regression if it was removed. Another value of interest is observation 188, which has a large residual of ~ 3.5, making it an "extreme value in Y," but not characterized as a Bonferoni-outlier.

4.2.4 Model 2 Analysis

We began our second analysis with a model of physicians regressed on total population and county region. After testing different transformations, we found $\frac{1}{\sqrt{TotalPop}}$ was the best transformation, meanwhile region remained untransformed. Carrying out a t-Test on the coefficient for region revealed it was insignificant, so we removed it from the model. After doing appropriate AIC and hypothesis testing, we determined a larger model with population over 65, bachelors degree's, poverty, and personal income added was a better choice. Adding these predictors is logical intuitively. When a county has a larger population of people over 65, it is going to be necessary to have more physicians as older people will have more health issues. Personal income and bachelors degree were other factors that had positive relationships with the number of log physicians. One can assume when income increases, the number of physicians will increase because there is better access to healthcare in higher income areas. In areas with higher education rates, it is likely that there are educated individuals with medical school degrees. It is logical region does not have any effect on number of physicians given all the counties were sorted on population. If all of the counties had similar populations, education levels, and incomes, then region would not have a large effect on number of physicians. Linear models and testing allowed us to conclude that our initial hypothesis was incorrect.