A Regression Analysis on Physician Counts

Jason Tooch, Kyla Balquin, Philip Yoon

June 12, 2019

A Regression Analysis on Physician Counts

Jason Tooch, Kyla Balquin, Philip Yoon

June 12, 2019

Abstract

This project used regression analysis to investigate two linear models; both investigating the number of professionally active nonfederal physicians during 1990. For the first model, we looked for a linear relationship between the predictors and found total population was very significant to physician count whereas land area and income capita were not. Additionally, the variance with the log of total population was determined to be constant. For the second model, the geographic region was found to not be statistically significant; however, at least one of the following significantly affected physician count: age, poverty levels, education, crime rate, or personal income.

Problem and Motivation

The CDI data set contains information about 440 of the most populous counties in the United States. We are particularly interested in whether the number of professionally active nonfederal physicians in 1990 in a given county can be explained by factors such as population, land area, income, and others. These factors can be related to any county and we are looking to provide some explanation on how they affect the number of physicians through our analysis. During our research, we will explore the relationships between these factors and determine their significance in relation to the number of physicians in a given county. In this study we will find if the factors of interest have notable effects in the estimation of the number of professionally active nonfederal physicians. Our research uses data from counties within the United States. The associations and relationships concluded from our observations can be used as a basis for predictions or future analysis regarding the number of physicians in any county. Similarly, the factors that we determine to have a significant relationship with the number of physicians can provide surprising insight as to why more physicians may or may not reside in one area over another.

Data

This data set includes information about the most populous countries in the US. The variables we are interested are: Physicians - Number of professionally active nonfederal physicians during 1990 Land Area - Land area (square miles) IncPerCap - Per capita income of 1990 CDI population (dollars) Region - Geographic region classification is that used by the U.S. Bureau of the Census, where: 1 = NE; 2 = NC; 3 = S; 4 = W Pop65 - Percent of 1990 CDI population aged 65 years old and older Crimes - Total number of serious crimes in 1990, including murder, rape, robbery, aggravated asault, burglary, larceny-theft, and motor vehicle theft, as reported by law enforcement agencies Bachelor - Percent of adult population (percsons 25 years old or older) with bachelor's degree Poverty - Percent of 1990 CDI population with income below poverty level Personalinc - Total personal income of 1990 CDI population (in millions of dollars)

Questions of Interest

Part I: Our analysis aims to answer whether the number of physicians in a given county can be explained by the natural log of the total population, the land area, and the average income per person in a given county. How do the number of physicians change when any of these variables change? Do each of these variables have a significant effect on the number of physicians? Is the data sufficient as is or must we change it? What range of values can we be sure

Part II: In this part, we are looking to determine if geographic region is an effective factor in estimating the number of physicians given that the total population is also considered. We also are looking to see if the additional demographics- the percent of the population 65 years or older, the number of serious crimes, the percentage of adults with a bachelor's degree, percentage of population in poverty, and the total personal income- have a significant effect on estimation of the number of physicians.

Regression methods

First, we fit a linear model with physicians as response and land area, income per capita, and the log of total population as predictors. Looking at the Residuals vs Fitted, QQ, and Scale Location plot, we determined whether the dependent variables are sufficient or whether they must be transformed. Additionally, we performed a box-cox test to determine what transformation if any, needs to be performed on the Physicians variable. Then, using confidence intervals, we tested whether there was a linear relationship between the predictors and the response. Finally, we used a non-constant variance test to see whether we should refit the model using weighted-least squares.

For the second model, we looked at p-values for the F-statistic from numerical summaries to see if any total populations or regions have a significant effect on number of physicians. Again we used plots and the box-cox test to decide transformations. Then, we conducted a partial F-test using the ANOVA function to assess whether it was valuable to add relevant predictors to the model. Finally, we used the influenceIndexPlot function to identify data points with large influences.

7. Regression Analysis, Results and Interpretation:

Part I Analysis

In this part we are considering the Physicians (the number of professionally active nonfederal physicians) as the response variable and the three variables as predictors: log(totalPop), LandArea, IncPerCap (Income per capita). Before doing any statisitcal observations, the relationships we expect to see based on our intuition are as follows: We would expect to see a positive relationship between the total population and the number of physicians and we would expect it to be relatively strong since the bigger the total population increases both supply and demand for physicians. We would expect to see a weak or close to no relation between the land area and the physicians because the physical size of a county does not always correlate with the amount of people living there. We would expect to see a positive correlation between income per capita and the number of physicians. Since we believe that the more income the average person in the country generates allows for more disposable income to pay for medical expenses.

```
library(alr4)
```

```
## Loading required package: car
```

```
## Loading required package: carData
```

```
## Loading required package: effects
```

```
## lattice theme set by effectsTheme()
## See ?effectsTheme for details.
```

```r
CDI <- readRDS("C:/Users/Jason/Downloads/CDI.rds")
fit1 <- lm(CDI$Physicians ~ log(CDI$TotalPop))
fit2 <- lm(CDI$Physicians ~ CDI$LandArea)
fit3 <- lm(CDI$Physicians ~ CDI$IncPerCap)
summary(fit1)
```

```
##
## Call:
## lm(formula = CDI$Physicians ~ log(CDI$TotalPop))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1710.8  -485.0   -15.1   350.4  9925.9
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -17123.80     695.14  -24.63   <2e-16 ***
## log(CDI$TotalPop)   1447.06      55.73   25.97   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 864.1 on 423 degrees of freedom
## Multiple R-squared:  0.6145, Adjusted R-squared:  0.6136
## F-statistic: 674.3 on 1 and 423 DF,  p-value: < 2.2e-16
```

```r
summary(fit2)
```

```
##
## Call:
## lm(formula = CDI$Physicians ~ CDI$LandArea)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -854.7  -711.5  -514.1    92.9 14259.7
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.885e+02  8.164e+01  10.883   <2e-16 ***
## CDI$LandArea 5.071e-03  4.458e-02   0.114    0.909
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1392 on 423 degrees of freedom
## Multiple R-squared:  3.06e-05,   Adjusted R-squared:  -0.002333
## F-statistic: 0.01294 on 1 and 423 DF,  p-value: 0.9095
```

```r
summary(fit3)
```

```
##
## Call:
## lm(formula = CDI$Physicians ~ CDI$IncPerCap)
##
```

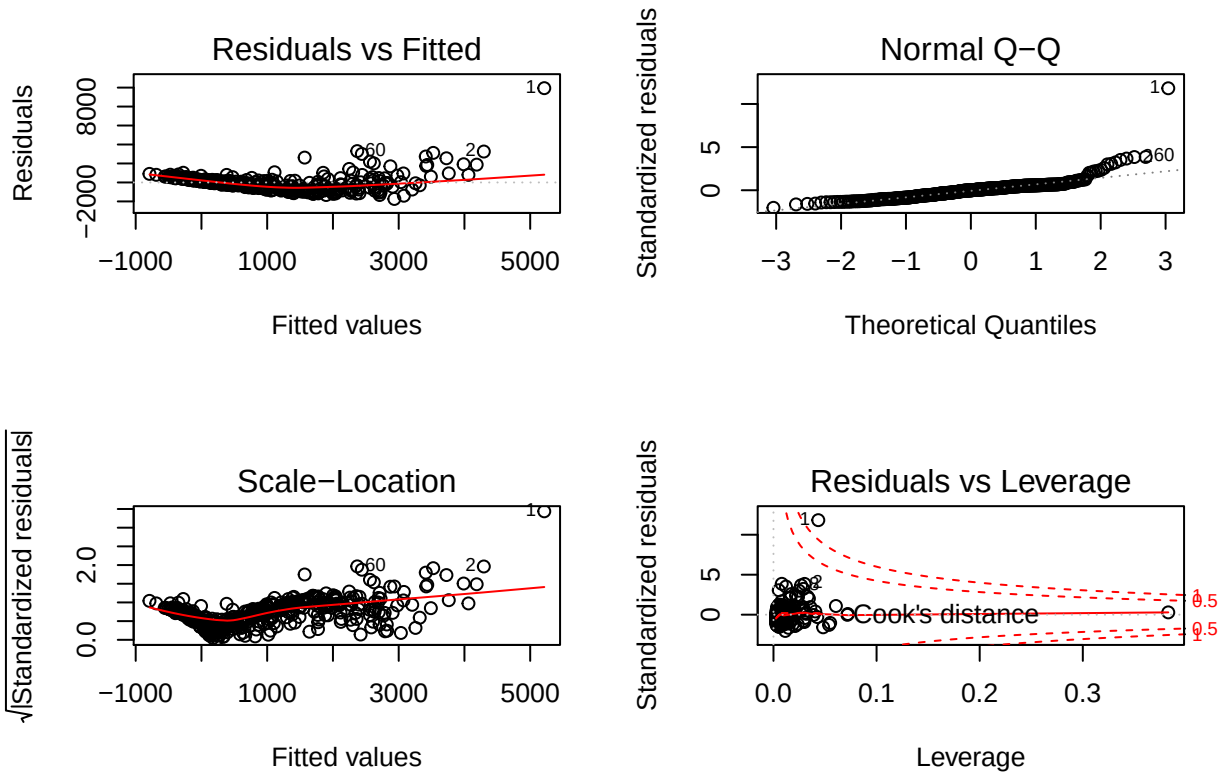
```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2425.8  -552.8  -302.8   103.0 13830.2
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -1.569e+03  2.923e+02  -5.366 1.33e-07 ***
## CDI$IncPerCap  1.331e-01  1.543e-02   8.622  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1283 on 423 degrees of freedom
## Multiple R-squared:  0.1495, Adjusted R-squared:  0.1475
## F-statistic: 74.34 on 1 and 423 DF,  p-value: < 2.2e-16
```

After drawing up the plot between the log(TotalPop) and the number of physicians we can see that there is a convex nonlinear association. Though a linear model is also sufficient as there is a strong positive correlation. 61.45% of the variability in the number of physicians per country is accounted for by a linear relationship with the total population. The scatterplot between the land area and the number of physicians shows that there is no discernible pattern between the area of land and the number of physicians. Because the correlation is so small,, we can conclude that our assumptions are correct. After plotting the distribution between the countries' income per capita and the number of physicians, we can see that there is a slight positive relationship between the two. They do not seem to have a strong correlation as 14.95% of the variability in the number of physicians is accounted for by a linear relationship with income per capita.

```r
library(alr4)
fit <- lm(CDI$Physicians ~ log(CDI$TotalPop) + CDI$LandArea + CDI$IncPerCap)
summary(fit)
```

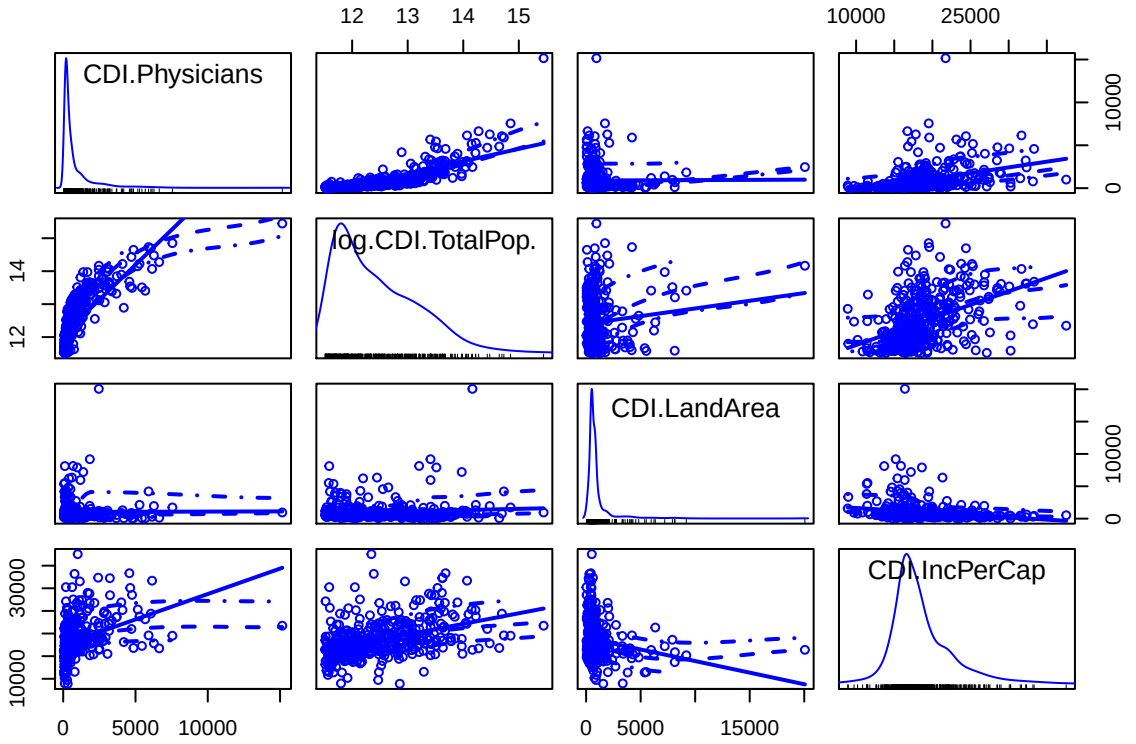
```
##
## Call:
## lm(formula = CDI$Physicians ~ log(CDI$TotalPop) + CDI$LandArea +
##     CDI$IncPerCap)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1739.9  -495.4    -5.4   375.4  9938.9
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)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

```
par(mfrow = c(2,2))
plot(fit)
```

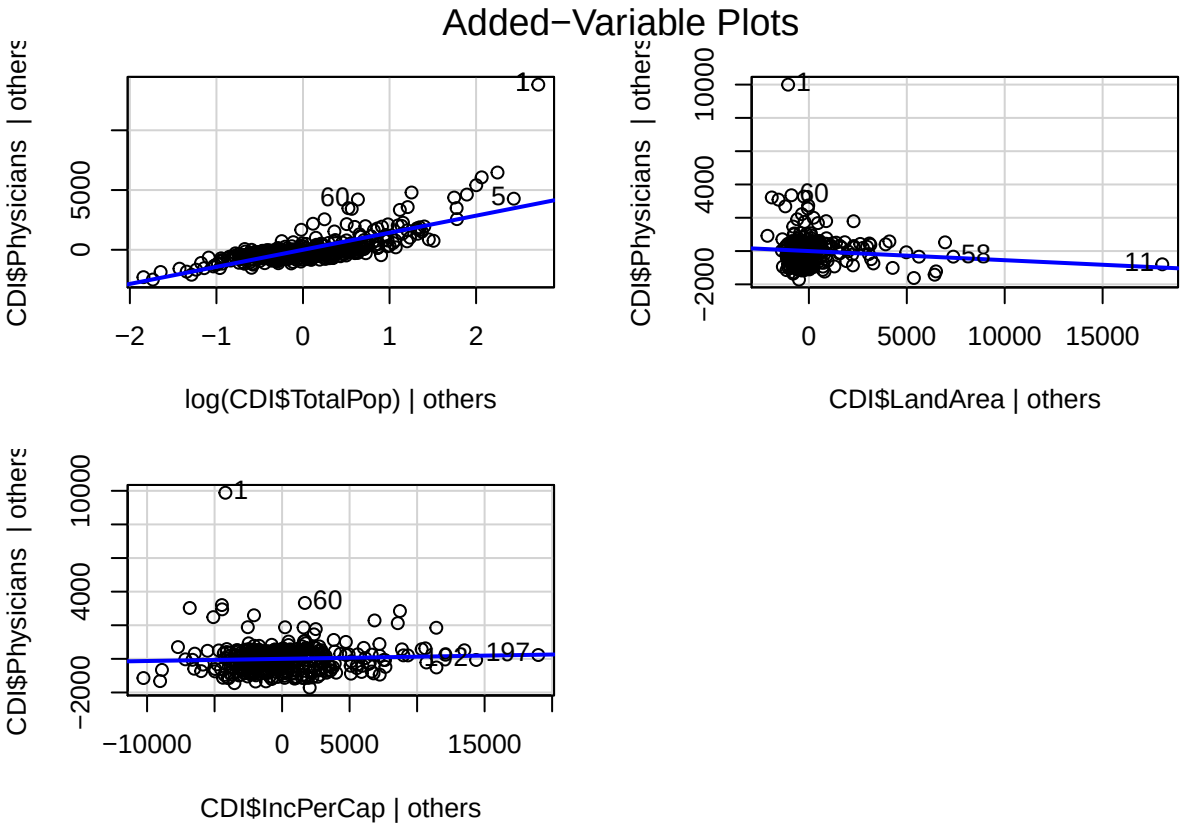


The interpretation of our original model is as follows: The number of physicians is expected to increase by 14.2 when the total population increases by 1%. The number of physicians is expected to change by -5.488e-02 when LandArea increases by one unit, this appears to be a very insignificant amount. Similarly, the number of physicians is expected to change by 1.285e-02 when IncPerCap increases by one unit. 62% of the variability in the number of physicians is explained by a linear relationship between the log of the total population, land area, and income per capita.

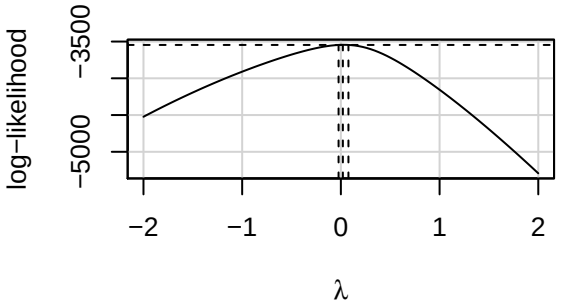
```
par(mfrow = c(2,2))
scatterplotMatrix(~CDI$Physicians+log(CDI$TotalPop) + CDI$LandArea + CDI$IncPerCap)
```


CDI.Physicians
log.CDI.TotalPop.
CDI.LandArea
CDI.IncPerCap
12
13
14
15
10000
25000
0
10000
12
14
0
10000
10000
30000
0
5000
10000
0
5000
15000

```
avPlots(fit)
```


Added−Variable Plots
CDI$Physicians | others
log(CDI$TotalPop) | others
CDI$LandArea | others
CDI$IncPerCap | others

```
boxCox(fit)
```



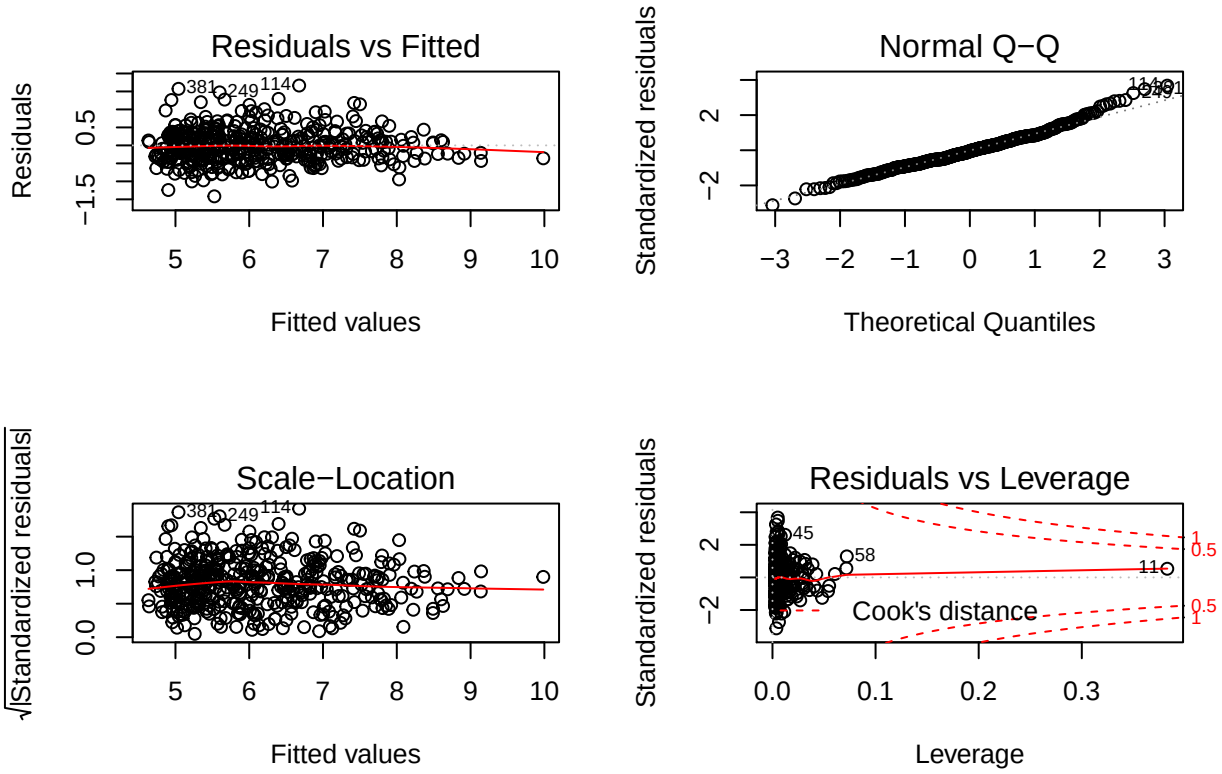
After reviewing diagnostics, the current model was not sufficient. We performed a BoxCox test and determined that our best option would be to log transform our response.This transformation allows our assumptions about linearity, normality, and equal variance to hold. Also, the scale location plot shows that the variance increases as we move from left to right.

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

```
##
## 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
```

```
par(mfrow = c(2,2))
plot(newfitt)
```



Our new model features a transformed response and as a result our coefficient interpretations have changed to the following: The expected value of physicians increases by 1.25% when the total population increases by 1%, all else constant. The expected value of physicians decreases by .003% when Land Area increases by one unit, all else constant. Similarly, the expected value of physicians increases by .003% when Income Per Capita increases by one unit. Similar to our original mode, Land Area and Income Per Capita appear to have a small affect on Physicians, while the total population has a significant affect. 83.4% of the variability in the log of the number of physicians is explained by the log of the total population, the land area, and the income per capita.

```
confint(newfitt)
```

```
##                             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
```

If you were to refit this model 100 times, we are 95% confident that the true values of these variables would be inside the following intervals:

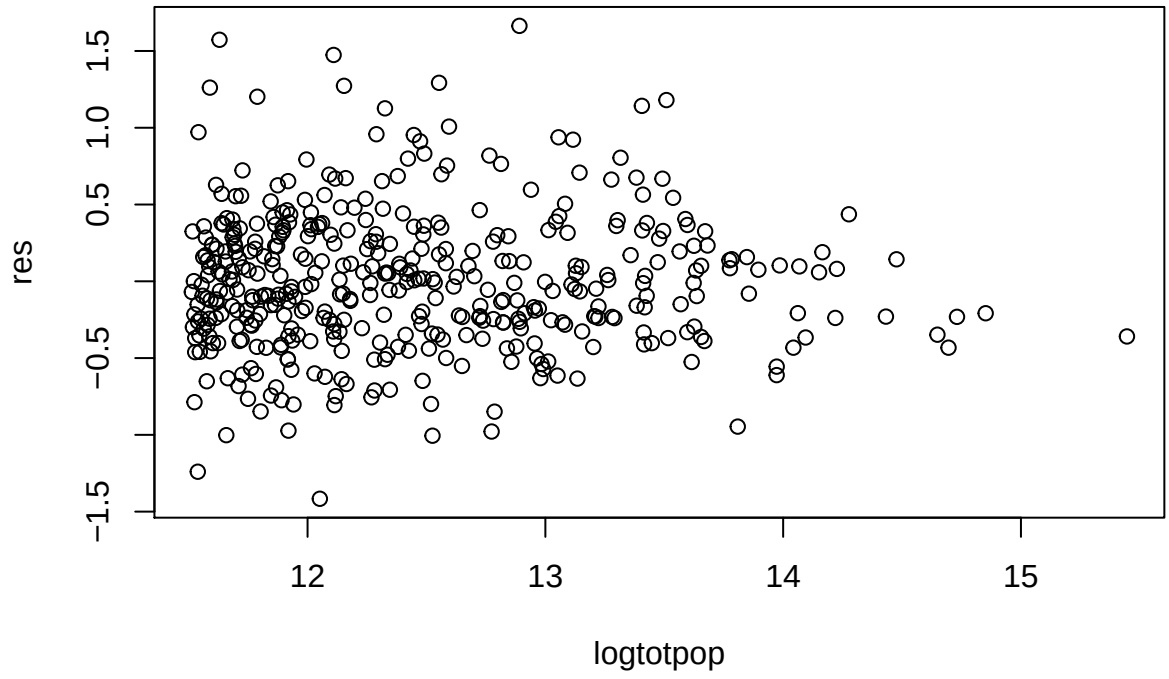
Intercept CI: (-1.087549e+01, -9.410109) log(TotalPop) CI: (1.189973, 1.320591) LandArea CI: (-5.952709e-05, -6.439696e-08) IncPerCap CI: (2.295433e-05, 4.766106e-05)

Ho: B0=B1=B2=B3= 0
Ha: Bj=/= 0 for some j=1,2,3

After drawing up the summary for our transformed model we see that our p-value of 2.2e-16, is less than the significance level = 0.01. Because the p-value is so small, we can reject the null hypothesis and conclude that at least one of the predictors has a significant effect in predicting the number of physicians.

Alternatively, no confidence intervals include zero so we can reject our null hypothesis that B0=B1=B2=B3=0

```
logtotpop <- log(CDI$TotalPop)
res <- newfitt$residuals
plot(logtotpop, res)
```




```
ncvTest(newfitt,~logtotpop)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ logtotpop
## Chisquare = 1.649145, Df = 1, p = 0.19908
```

Now we will take a look to see if our variance increases or decreases with Log(totalpop). From our residual plot above, our variance appears to be slightly decreasing. However, after performing a Non-constant variance test, our p-value from the test is 0.19 which is greater than any traditional significance level, so we fail to reject the null hypothesis and assume constant variance.

In conclusion, the Log of the total population has a very significant affect on the number of physicians in a given county. An interesting note is that the LandArea and Income per Capita variables seem to have inverse affects on the number of physicians in a given county. The number of physicians has a negative correlation with Land Area as the expected physicians decreases by .003% when Land Area increases by 1 unit. The income per capita variable has an inverse affect as it causes the expected physicians to increase by .003% when increased by 1 unit. Albeit miniscule, these inverse relationships might be worth investigating in a future study about the regression on physicians. All in all, the analysis was enlightening and provided some good intuition on how physicians is affected by these variables.

Part 2 Analysis

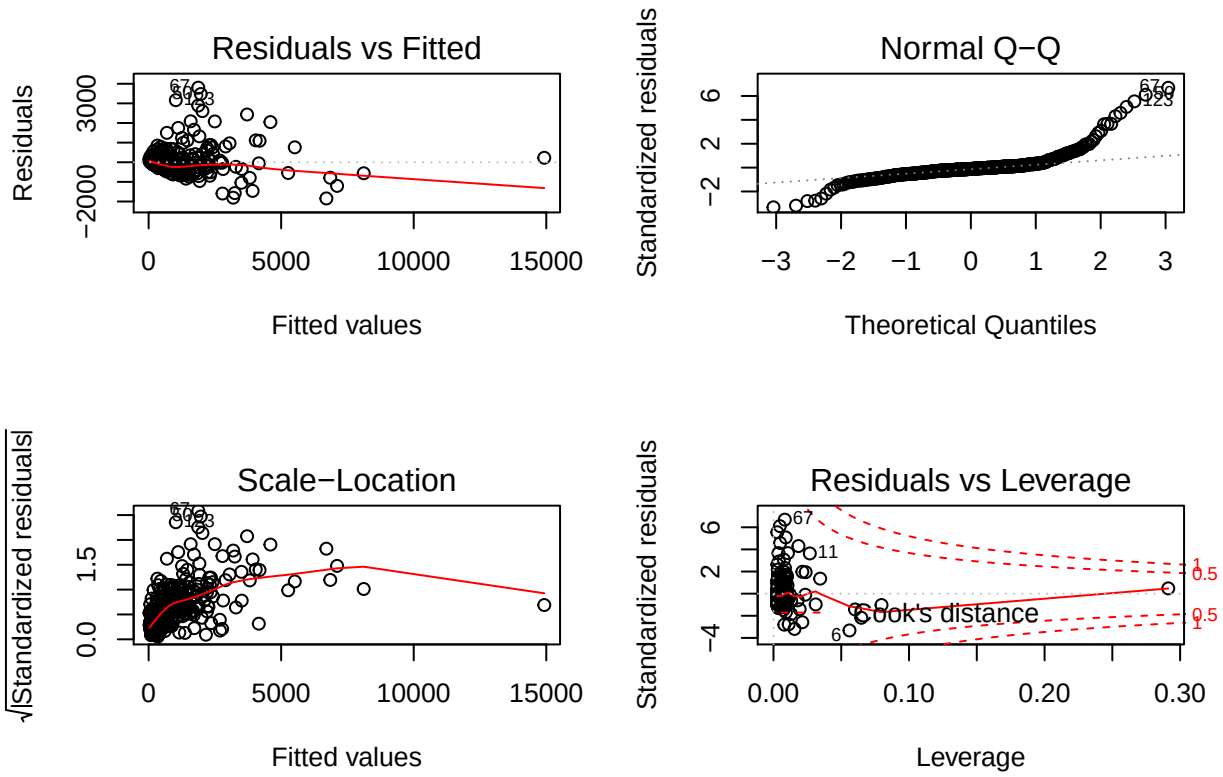
In the second part of our analysis, we are considering the model Physicians as response and TotalPop and Region as predictors

```
model0 <- lm(Physicians ~ TotalPop + Region, data = CDI)
summary(model0)
```

```
##
## 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
```

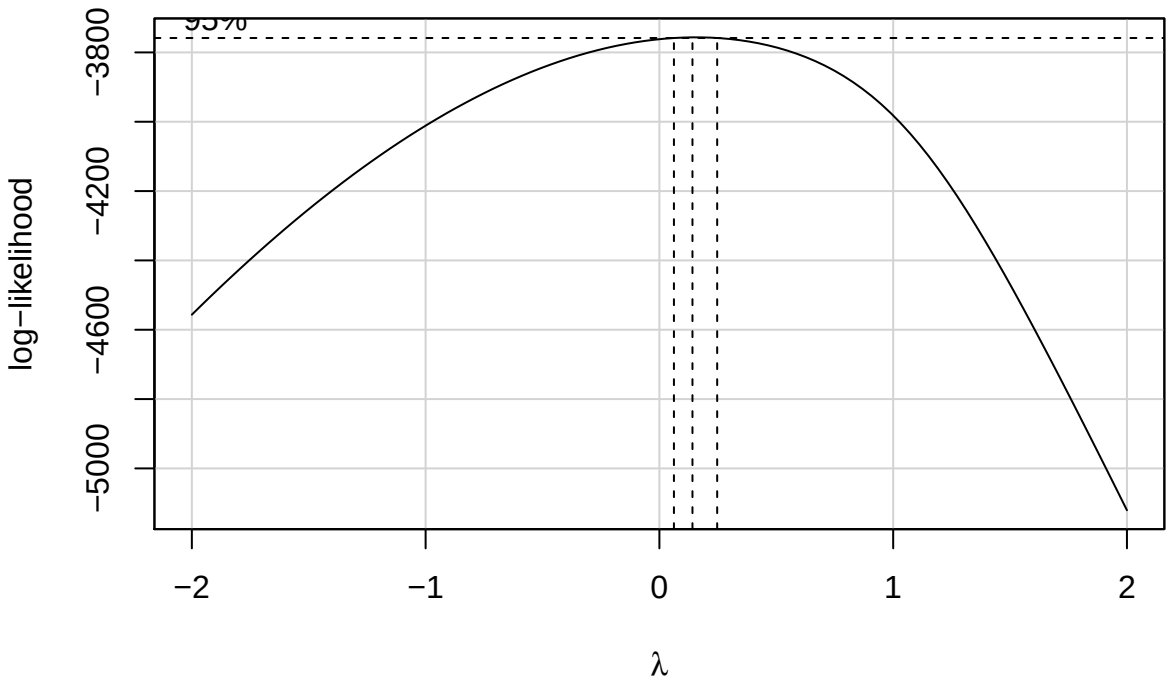
After fitting the model and viewing its numerical summary, we can see from the p-value of the F-statistic that it is lower than any signficance level; therefore, we can assume at least one of the predictors are signficant to Physician count.

```
par(mfrow = c(2,2))
plot(model0)
```



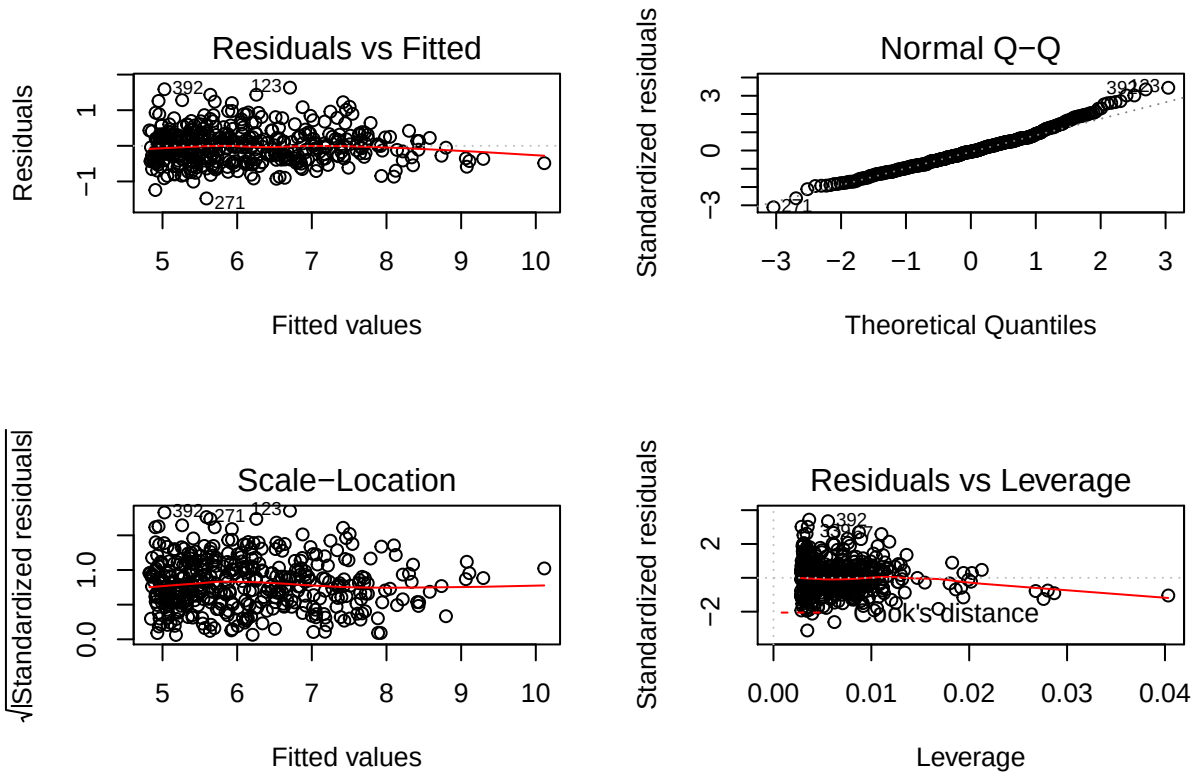
After viewing diagnostic plots, there seems to be a heavy-tail QQ Plot, which means normality is violated. The Scale-Location plot shows non-constant variance and the Residual vs Fitted plot shows violation of linearity.

```
boxCox(model0)
```



The Box Cox test shows lamda value of 0 should be sufficient, meaning we can log transform Physicians. Also, log transforming total population should help fix non-linearity.

```
model1 <- lm(log(Physicians) ~ log(TotalPop) + Region, data = CDI)
par(mfrow = c(2,2))
plot(model1)
```



After viewing the diagnostic plots after the transformation, we can see there are no more violations.

Now looking at each region separately:

```
model2 <- lm(Physicians ~ TotalPop + as.factor(Region), data = CDI)
summary(model2)
```

```
##
## Call:
## lm(formula = Physicians ~ TotalPop + as.factor(Region), data = CDI)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1845.2  -215.1   -67.5    96.0  3809.8
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -9.982e+01  6.240e+01  -1.600   0.1104
## TotalPop             2.958e-03  6.496e-05  45.542   <2e-16 ***
## as.factor(Region)2 -6.149e+01  8.011e+01  -0.768   0.4432
## as.factor(Region)3 -6.495e+01  7.422e+01  -0.875   0.3820
## as.factor(Region)4 -2.156e+02  8.733e+01  -2.469   0.0139 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 571.1 on 420 degrees of freedom
## Multiple R-squared:  0.8328, Adjusted R-squared:  0.8312
```

```
## F-statistic:   523 on 4 and 420 DF,  p-value: < 2.2e-16
```

If we write out the estimated mean of number of physicians as a function of total population and personal for each region we get:

Region 1: E(physicians) = 0.002958*TotalPop - 99.82 Region 2: E(physicians) = 0.002958*TotalPop - 61.49
Region 3: E(physicians) = 0.002958*TotalPop - 64.95 Region 4: E(physicians) = 0.002958*TotalPop - 215.6

From these equations we can see all betas for TotalPop, or slopes, are the same, hence the name parallel regression model.

If we want to determine whether Region is significant to our model, we will check the numerical summary for our transformed model.

```
summary(model1)
```

```
##
## Call:
## lm(formula = log(Physicians) ~ log(TotalPop) + Region, data = CDI)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.47765 -0.32390 -0.03421  0.25354  1.63280
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -10.42837    0.38558 -27.046   <2e-16 ***
## log(TotalPop)   1.33319    0.03070  43.427   <2e-16 ***
## Region         -0.02511    0.02230  -1.126    0.261
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4759 on 422 degrees of freedom
## Multiple R-squared:  0.8172, Adjusted R-squared:  0.8163
## F-statistic:   943 on 2 and 422 DF,  p-value: < 2.2e-16
```

Region has a high p-value greater than any signifiance level, so we can determine it is not needed in our model.

Now we will add the predictors Pop65, Crimes, Bachelor, Poverty, and PersonalInc to see if this full model is superior. We compare our full model with the original model to see if there is a statistically significant improvement in estimating the number of physicians

```
reduced_model <- lm(log(CDI$Physicians) ~ log(CDI$TotalPop))
full_model <- lm(log(CDI$Physicians) ~ log(TotalPop) + Pop65 + Crimes + Bachelor + Poverty + PersonalInc
anova(reduced_model, full_model)
```

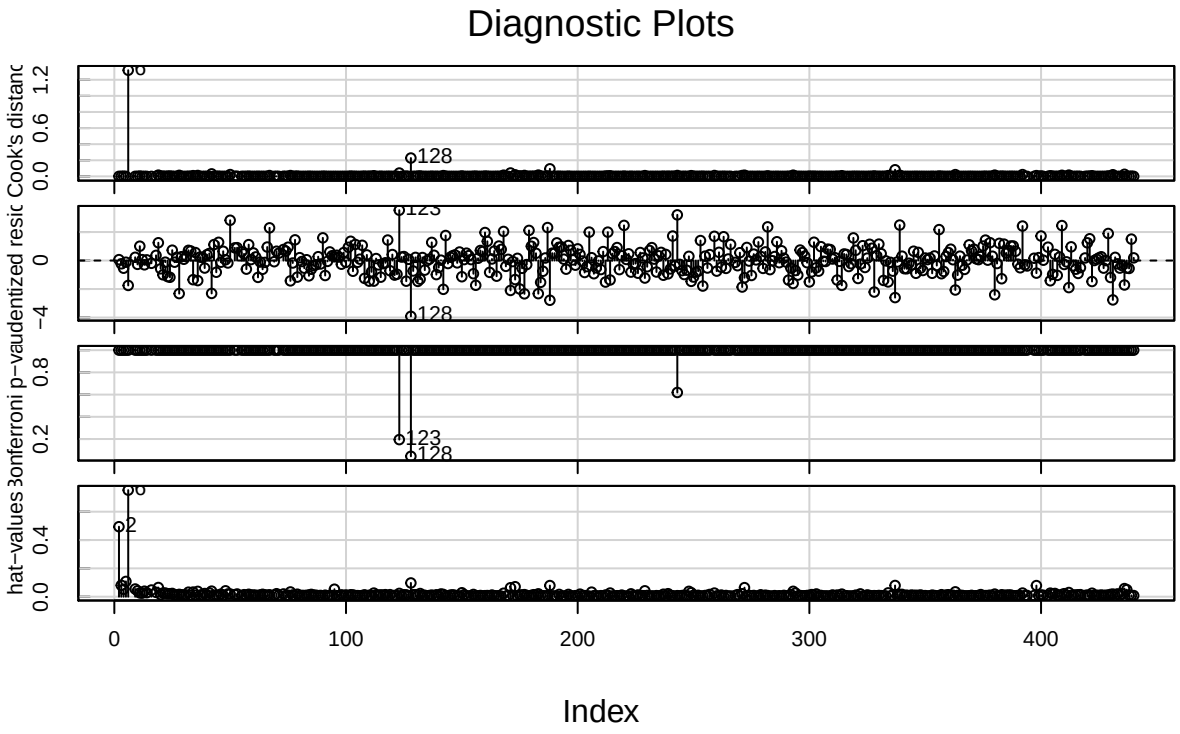
```
## Analysis of Variance Table
##
## Model 1: log(CDI$Physicians) ~ log(CDI$TotalPop)
## Model 2: log(CDI$Physicians) ~ log(TotalPop) + Pop65 + Crimes + Bachelor +
##     Poverty + PersonalInc
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    423 95.848
```

```
## 2    418 61.495  5    34.353 46.701 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

After conducting a partial F-test using the anova function we see that it yields a p value of less than 2.2e-16. Therefore at a significance level of 0.05, we have enough evidence to reject the null hypothesis that the coefficients for each of the added predictors are 0, meaning at least one of predictors is a useful predictor in estimating the number of physicians. According to the results of this test we can conclude that adding the new predictors to the model yields a statistically significant improvement.

Now we will identify any influential points in our new model.

```
influenceIndexPlot(full_model)
```



By using the influenceIndexPlot function we can see that points 2, 6, 123, and 128 are influential points. Points 2 and 6 have particularly high leverages meaning they are outliers in x (or log(totalPop) + region + pop65 + crimes + bachelor + poverty + personalInc). Points 123 and 128 have high studentized residuals meaning they have a high standardized difference between the expected value and the actual value. With the inclusion of these highly influential points, it may lessen our confidence to say all our predictors have a definite effect on physician count.

Conclusion

Upon checking the diagnostics for our original model, which holds the number of physicians as the response on the three predictors, log of the total population, the area of land, and the income per capita, we observed that the model violated the assumptions of linearity, normality, and constant variance, thus calling for a log

transformation on our response variable. After constructing confidence intervals for each of the predictors in our model, we found at least one of the predictors had a significant effect on the estimation of the number of physicians.

In our second model including the number of physicians as our response with the total population and geographic region classification as our predictors, our diagnostics showed we needed to log transform the total population. The categorical predictor Region was found to have a statistically significant effect on our response; in other words, the count of physicians does significantly differ depending on what region we are observing. We also found that adding Pop65, Crimes, Bachelor, Poverty, and PersonalInc as predictors showed to be a statistically significant improvement to the model; however, four data points were found to be highly influential which may hurt the ultimate reliability of this model.