Physicians

Fiara Causo, Taylor Villahermosa, Jingyue Zhou

6/9/2019

Abstract

We will analyze the diffeent factors that affect the total amounf of physicians in a county. Additionally, we will utilize a variety of regression methods to find the best model for predicting the amount of physicians in a county. From our analysis, we found that the variables that significantly affect the amound of physicians are TotalPop, Pop65, Bachelor, Poverty, and PersonalInc. We found that the best model to predict the number of physicians in a county is: $\text{Physicians}^{0.15} = \text{TotalPop}^{-0.5799} + \text{Pop65} + \text{Bachelor} + \text{Poverty} + \text{PersonalInc}$

Problem and Motivation

This data set provides selected county demographic information (CDI) for 440 of the most populous counties in the United States.The information generally pertains to the years 1990 and 1992. We want to know if total personal income, land area and population have effects on the number of professionally active non-federal physicians or if the relationship between them can be explained by the regression model. If not, our transformation of model can be useful. By the analysis we did, readers can get how do these 3 factors influence the number of professionally active non-federal physicians. For another part, what we focus on is the effect total population and region have on the number of professionally active non-federal physicians. Region here is a categorical variable represents Geographic region classification. Readers can learn whether region is useful in our model and see if the transformation is needed here. Also, issues of the original data set like influential data points will be presented. Readers can see why they are influential here.

Data

CDI contains county demographic information from the top 440 populous counties in the United States in 1990. In our analysis, we will focus on the following variables: Physicians, LandArea, TotalPop, Region, Pop65, Poverty, Crimes, and Bachelor. Physicians is the number of professionally active nonfederal physicians. LandArea is land area in square miles. TotalPop is the estimated population in 1990. Region is geographic regions NE, NC, S, and W. Pop65 is the percentage of the population that is over 65 years old. Poverty is the percentage of the population with an income below poverty level. Crimes is the total number of serious crimes in 1990. Bachelor is the percentage of adult population with a bachelor's degree.

Questions of Interest

In our analysis, we will investigate which variables affect the amount of physicians in a county. We will determine whether or not each variable has a relationship with the number of physicians and how significant that relationship is. Additionally, we will build a model in order to predict the amount of physicians in a county based on our analysis.

Regression Methods

First, we use F-test over a model to see if there is a linear relationship between the predictors and the responsor. We will also use a t-test to determine if each of the estimated coefficients is significant. Sometimes we need to do partial F-test to see if submodel holds or full model holds. ANOVA table helps here. Also, we can use backward or forward selection with AIC or BIC to choose predictors into our model. If the model holds after the two tests above, then R^2 can tell the goodness of the model by telling us the proportion in Y that can be explained by Xs. Then we will performd diagnostic checks to see if the linear regression assumptions seem to hold. Residuals Vs. Fitted and scale-location are about constant variance test, while the QQ-plot is about normality test. If not diagnostics assumptions are not met, we need to seek transformation for our model. Box-cox helps us find transformation for the response variable, and powerTransform function can find the best power lambda for the predictors. After the transformation, we will check diagnostics again

to see if all the violations are fixed. We can also use Non-constant Variance Score Test to see if the model has constant variance or not. If it's non-constant, then we can use weighted least squares mode to make the variance constant. Additionally, we need to see if there are outliers or high-influential points. If there are outliers present, we can carefully remove some of them or rebuild the model.

Regression Analysis, Results, and Interpretation

Part I

We expect that the relationship between the response physicians, the number of professionally active nonfederal physicians during 1990, would be positively correlated by the predictor's log(totalpop) and IncPerCap, but have no significant relationship with LandArea. The total population should have a positive correlation with the number of physicians due to the case of a higher likelihood for a person being a physician if the sample size of people increases. The land area in square miles should hold no significant relationship with the number of physicians, for the act and ability of a person becoming a physician has no dependency on the size of land. Lastly, the incPerCap should have a positive relationship/correlation with the response variable, physicians, for as the higher a total personal income of 1990 CDI population (in millions of dollars) is, the means of being able to pay for school and financially support yourself to become a physician increases.

We expect that there will be an association between log(TotalPop) and IncPerCap because the greater the population, there are a lot more jobs and people working, so IncPerCap is more than in less densely populated areas. We do not expect there to be an association between log(TotalPop) and LandArea because there are some very small yet densely populated cities, and also very large sparsely populated rural areas.

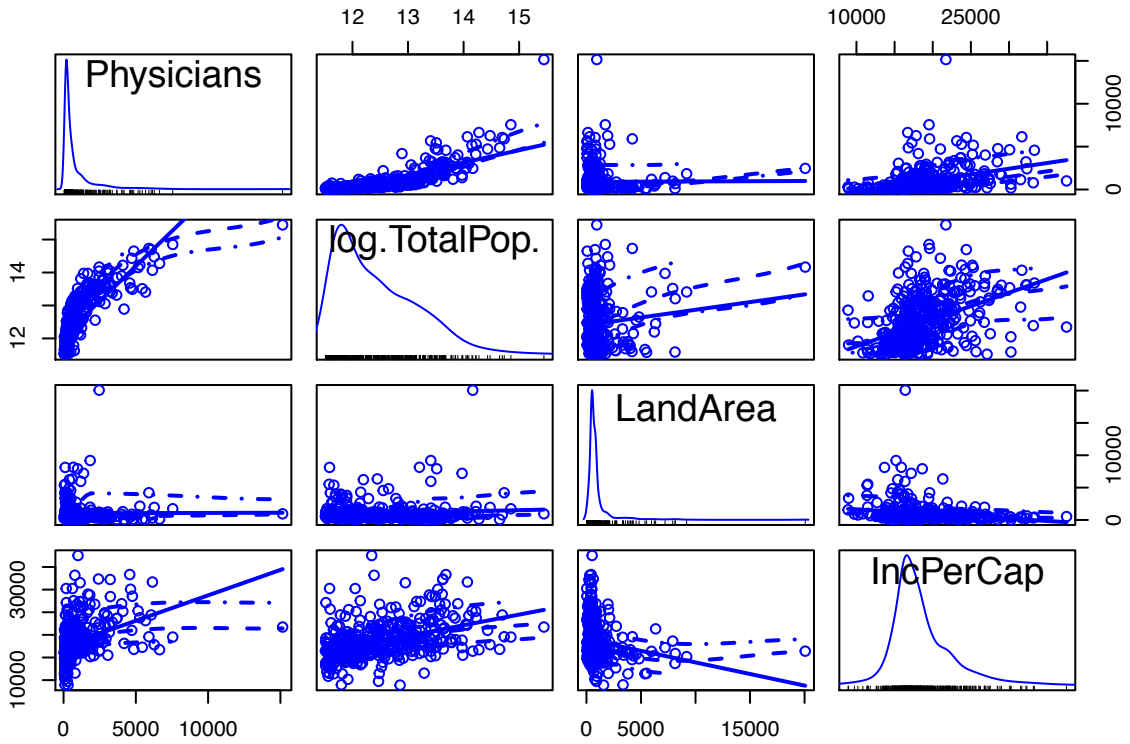
To test our intuition we decided to use the added variable plots and scatterplot matrix to get an accurate sense of the correlations between the response and predictors.

```
library(car)
```

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

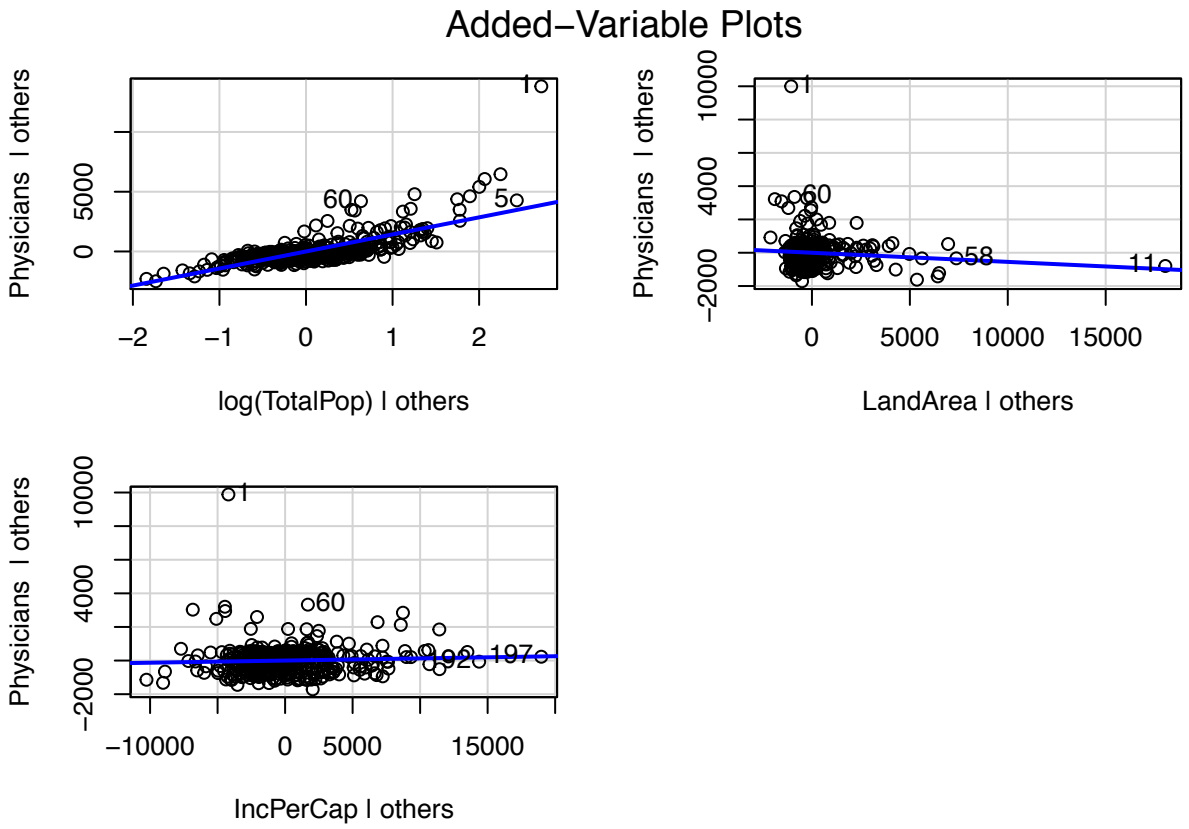
```
library(MASS)
CDI <- readRDS("CDI.rds", refhook = NULL)
attach(CDI)

fit1 = lm(Physicians~log(TotalPop)+LandArea+IncPerCap)
scatterplotMatrix(~Physicians+log(TotalPop)+LandArea + IncPerCap)
```



From the scatter plot matrix, Physicians is positively correlated with log(TotalPop) and IncPerCap. There is no apparent linear relationship can be observed between Physicians and LandArea. Additionally, the plots show that both LandArea and IncPerCap are negatively correlated with log.TotalPop. Consequently, the relationship between IncPerCap and log(TotalPop) matches our intuition but for LandArea and log(TotalPop) it does not. The relationship between log(TotalPop) and LandArea is nonlinear which is also the the same as our intuition.

```
avPlots(fit1)
```



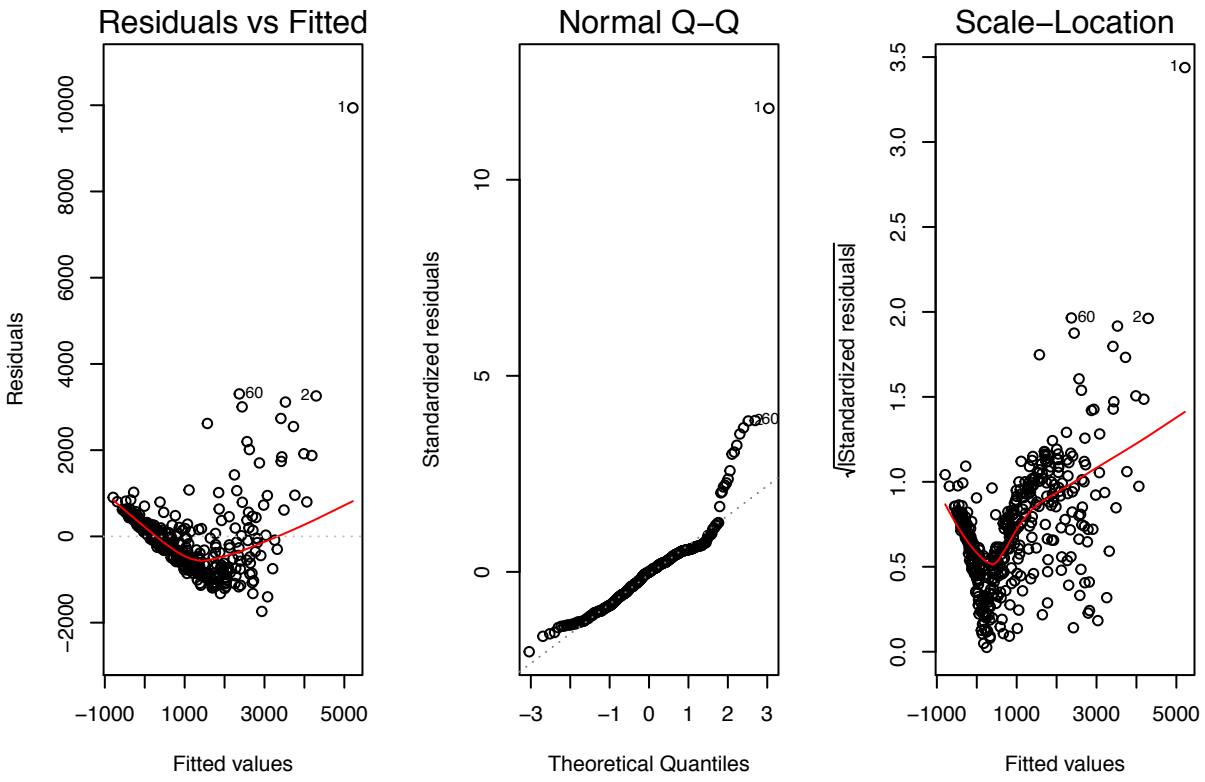
The Added-Variable plot for log(TotalPop) after LandArea and IncPerCap shows that log(TotalPop) is still positively correlated with Physicians even after accounting for the effects of LandArea and IncPerCap. Similarly to IncPerCap, log(TotalPop) is still positively correlated with Physicians after accounting for the effects of LandArea and log(TotalPop). However, that of LandArea after IncPerCap and IncPerCap shows that it is not useful when IncPerCap is already in the model.

```
summary(fit1)
```

```
##
## 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
```

β_0 which equals to -1.706e+04 represents the expected value of Physicians if log(TotalPop) , LandArea and IncPerCap all equals to zero. β_1 which equals to 1.427e+03 means the expected change in response which is Physicians if log(TotalPop) increase by one assuming all other predictors are held constant. β_2 which equals to -5.488e-02 represents the expected change in Physicians if LandArea increase by one assuming all other predictors are held constant. β_3 which equals to 1.285e-02 shows the expected change in Physicians if IncPerCap increase by one assuming all other predictors are held constant. $R^2 = 0.6175$. Thus, 61.75% of the variability in Physicians is accounted by log(TotalPop) , LandArea and IncPerCap.

```
# diagnostic checking original model
par(mfrow = c(1,3))
for (i in 1:3){
  plot(fit1, which = i)
}
```



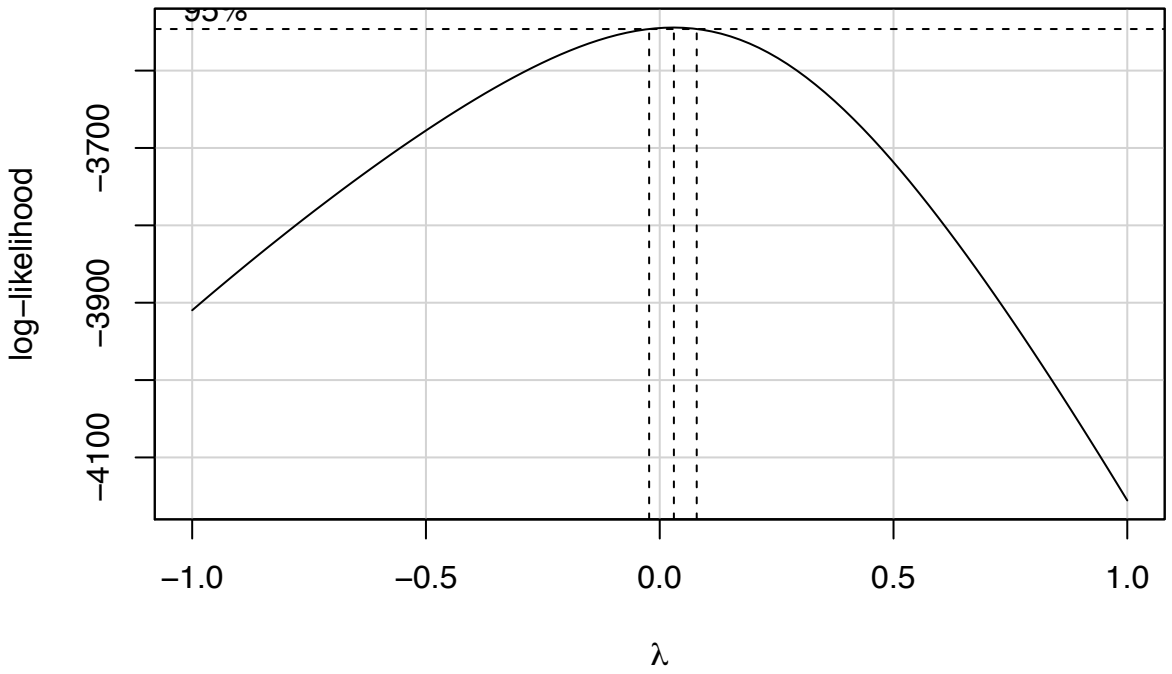
According to each diagnostics check, we see a violation for almost each of the linear regression assumptions. In the Residuals Vs. Fitted plot, residuals should form a horizontal bond along the 0 line (cloud of points). Residuals that seem to increase or decrease might indicate non-constant residual variance. By looking at our plot we can see that the residuals are doing just that, hence the constant variance test fails. A few large residuals may be indicative of outliers, here we can see more than a handful of large residuals. Curvature might indicate that the fitted mean function is inappropriate.

In the Normal Q-Q Plot, we are looking for strong violations of normality, points should form a roughly straight line showing the normality of errors. The plot doesn't look too bad except there appears some extreme right-skew to the residuals that indicate non-normality.

In the Scale - Location plot, the spread should show a similar pattern from left to right. Shows if residuals are spread equally along with a range of predictors. In the last plot, we can see that the constant variance assumption is for sure violate as the spread is uneven from side to side.

We will do a box-cox transformation for the response variable:

```r
# box cox
best.lambda <- boxCox(fit1, lambda = seq(-1, 1, by = 0.1))
```



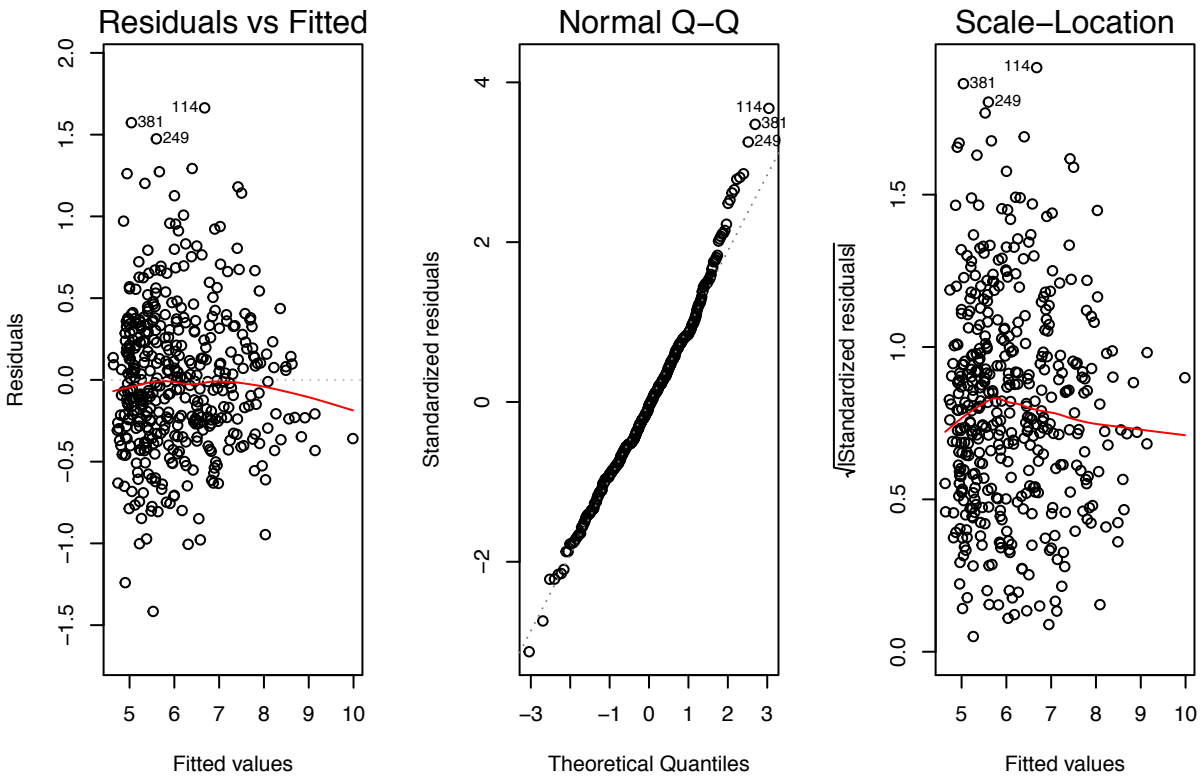

```r
best.lambda$x[which.max(best.lambda$y)]
```

```
## [1] 0.03030303
```

Since $\lambda \approx 0$, we will do a log transformation.

```r
# log transform because lambda = 0
fit1.tr = lm(log(Physicians)~log(TotalPop)+LandArea+IncPerCap)

# diagnostic checking transformed model
par(mfrow = c(1,3))
for (i in 1:3){
  plot(fit1.tr, which = i)
}
```



For our new log transformed model, the residual plots pass all of the diagnostics tests.
In the Residuals Vs. Fitted plot, the residuals have a constant scatter and no apparent outliers. It passes the constant variance test. In the Normal Q-Q Plot, we are looking for strong violations of normality, points should form a roughly straight line showing the normality of errors. The plot is a relatively straight line with only a few points on the tails falling off the line. In the Scale - Location plot, the spread should show a similar pattern from left to right. There is constant scatter, indicating normality.

We will use the boxTidwell() command to see if the predictors have a linear relationship with log(Physicians):

```
boxTidwell(log(Physicians) ~ log(TotalPop) + LandArea + IncPerCap)
```

```
##                MLE of lambda Score Statistic (z) Pr(>|z|)
## log(TotalPop)       -0.10589             -1.7537  0.07949 .
## LandArea            -0.71739              1.5808  0.11392
## IncPerCap           -0.62835             -0.3376  0.73567
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## iterations =  12
```

H_0: Each of the responses has a linear relationship with log(Physicians) vs H_1: Not H_0
The p-values for log(TotalPop), LandArea, and IncPerCap are 0.07949, 0.11392, and 0.73567, respectively. All of the p-values are greater than 0.05, so we fail to reject the null hypothesis and conclude that all of the predictors have a linear relationship with log(Physicians).

We will now test the significance of the predictors' coefficients.

```
summary(fit1.tr)
```

```
##
## Call:
## lm(formula = log(Physicians) ~ log(TotalPop) + LandArea + 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(TotalPop)  1.255e+00  3.323e-02  37.780  < 2e-16 ***
## LandArea      -2.980e-05  1.513e-05  -1.970   0.0495 *
## 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
```

```
confint(fit1.tr)
```

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

$H_0: \ \beta_j = 0 \ \forall j = 1, 2, 3$ vs $H_1: \ \beta_j \neq 0$ for some j
F = 705.2, p-value: $<$ 2.2e-16≈ 0
We reject the null hypothesis and conclude that $\beta_j \neq 0$ for some j, so at least one of the predictors has a coefficient that is not equal to 0.

T-test for each β_j:
$H_0: \ \beta_1 = 0$ (log(TotalPop)) vs $H_1: \ \beta_1 \neq 0$
t = 37.780, p-value $<$ 2e-16≈ 0
At $\alpha = 0.01$, we reject the null hypothesis and conclude that there is a linear relationship between log(TotalPop) and log(Physicians).
95% CI for β_1: (1.189973,1.320591)
We are 95% confident that true value of β_1 is in (1.189973,1.320591).

$H_0: \ \beta_2 = 0$ (LandArea) vs $H_1: \ \beta_2 \neq 0$
t = -1.970, p-value = 0.0495≈ 0
At $\alpha = 0.01$, we fail reject the null hypothesis and conclude that there is a not a significant linear relationship between log(TotalPop) and LandArea.
95% CI for β_2: (-5.952709e-05,-6.439696e-08)
We are 95% confident that true value of β_2 is in (-5.952709e-05,-6.439696e-08).

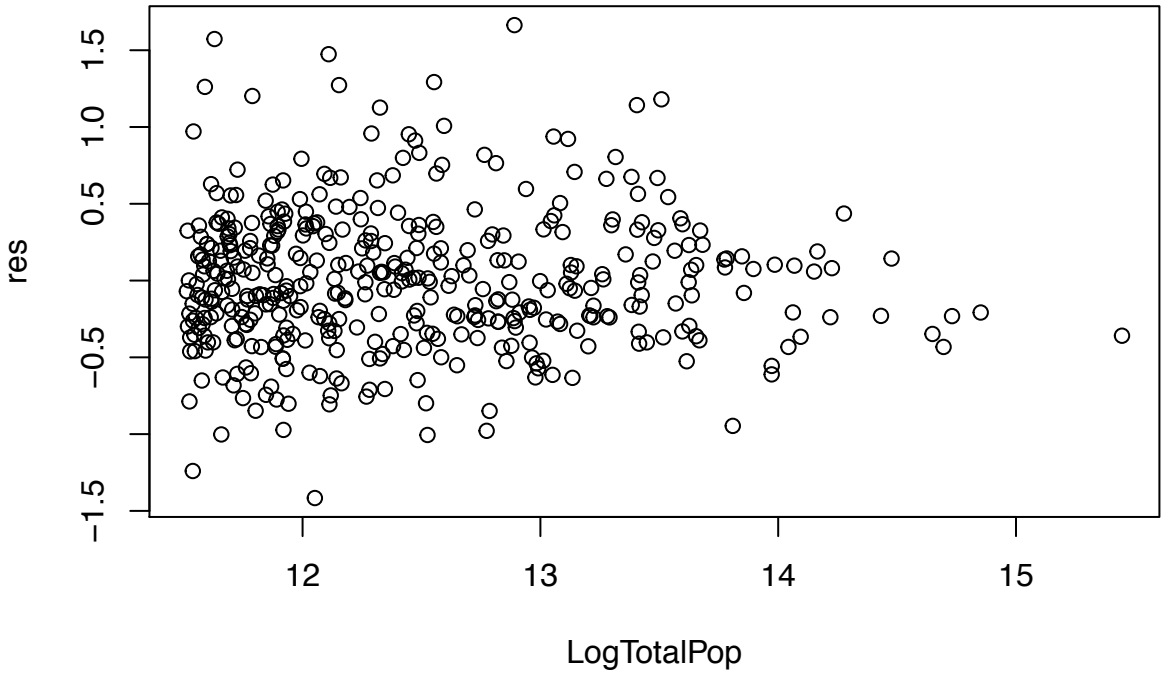
$H_0: \ \beta_3 = 0$ (IncPerCap) vs $H_1: \ \beta_3 \neq 0$
t = 5.618, p-value = 3.52e-08≈ 0
At $\alpha = 0.01$, we reject the null hypothesis and conclude that there is a linear relationship between log(TotalPop) and IncPerCap.
95% CI for β_3: (2.295433e-05,4.766106e-05)
We are 95% confident that true value of β_2 is in (2.295433e-05,4.766106e-05).

```
fit1 <- (log(Physicians) ~ log(TotalPop) + LandArea +IncPerCap)
lmvariance <- lm(log(Physicians) ~ log(TotalPop) + LandArea +IncPerCap)
#Variance as a function of the predictors
#Residuals versus predictors
```

```
res <- lmvariance$residuals
```

By looking at the residuals versus log(TotalPop) we can see that the variance decreases with Log(Totalpop).

```
#Plotting just log(TotalPop)
plot(log(TotalPop), res, xlab = 'LogTotalPop')
```



We performed a test to help make our conclusion:
H_0: Constant variance holds vs H_1:Non-constant variance holds

```
#testing whether the variance is a linear function of these predictors
ncvTest(lmvariance, ~log(TotalPop)) #Just Log(totalPop)
```

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

p = 0.19908 > 0.05 so we fail to reject the null hypothesis. Constant variance holds with just log(TotalPop)

```
#all the predictors test variance
ncvTest(lmvariance, ~ log(TotalPop) + LandArea +IncPerCap)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ log(TotalPop) + LandArea + IncPerCap
## Chisquare = 7.22885, Df = 3, p = 0.06495
```

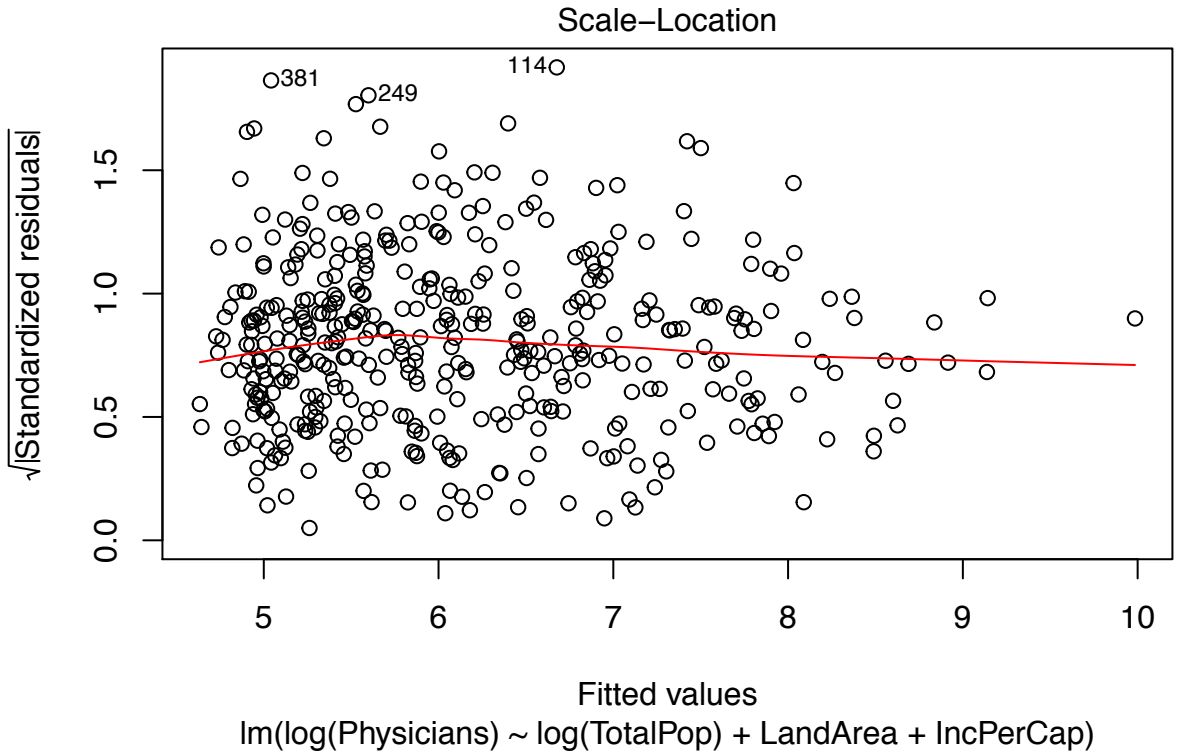
p = 0.06495 > 0.05 so we fail to reject the null hypothesis. Constant variance holds with the full model.

```
#ncv test with all both log(total) predictor
ncvTest(lmvariance, ~ LandArea + IncPerCap)
```

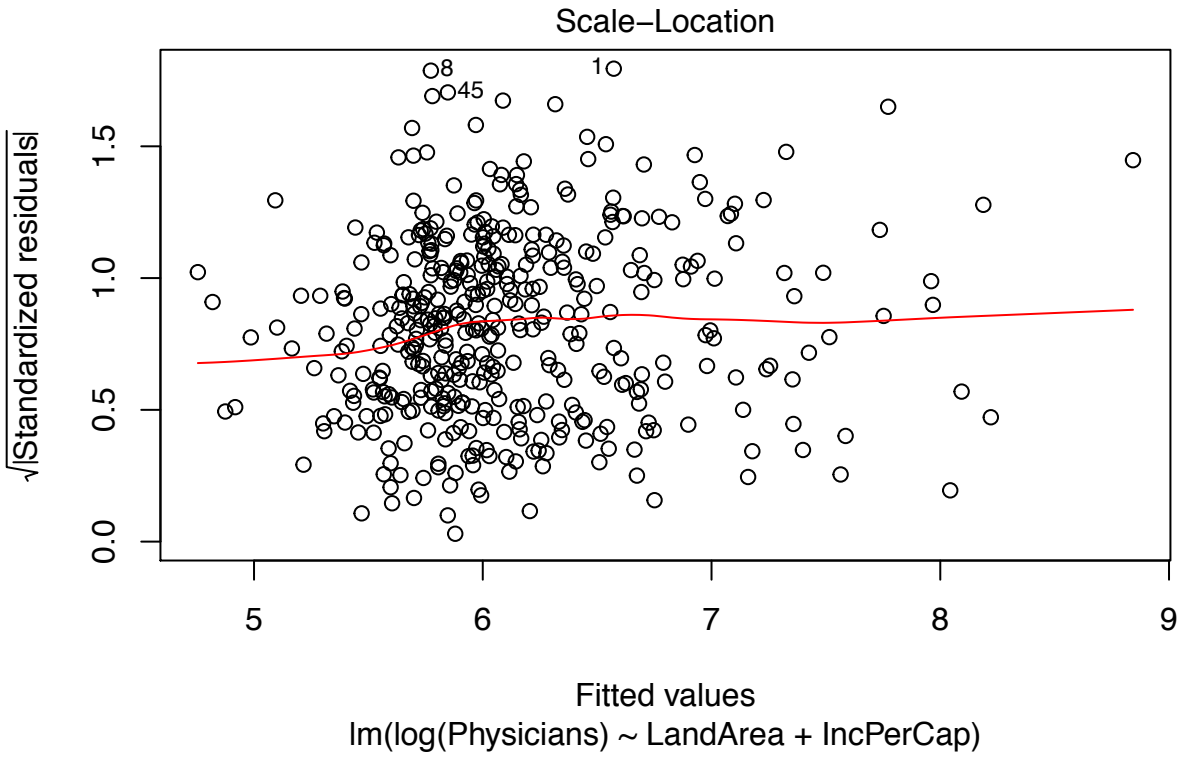
```
## Non-constant Variance Score Test
## Variance formula: ~ LandArea + IncPerCap
## Chisquare = 6.718986, Df = 2, p = 0.034753
```

p = 0.034753 < 0.05 so we reject the null hypothesis. Non-constant variance holds with LandArea and IncPerCap as the predictors.

```r
#Scale invariant with all of the predictors
plot(lmvariance, which = 3)
```




```r
#Scale Invariant without Log(TotalPop)
lm3withoutPoP <- lm(log(Physicians) ~ LandArea + IncPerCap)
plot(lm3withoutPoP, which = 3)
```



In the next test (1-pichisq) we need to test whether a smaller variance model just using log(totalPop) is preferable to the full variance model (alternative hypothesis). This is done by simply subtracting the test statistics, then comparing to the χ^2_{qf-qr} distribution. Here, qf and qr are the number of variables used in the full and reduced variance models, respectively, in this case qf = 3 and qr = 1.

```
#p-value for just using logTotalpop
1-pchisq(7.22885 - 1.649145, 2)
```

```
## [1] 0.06143027
```

```
#p-value for just using landArea and incpercap
1-pchisq(7.22885 - 6.718986, 2)
```

```
## [1] 0.7749692
```

So according to our p-values it looks like using logTotalPop works enough. This corresponds to the weights wi being the inverse of Log(TotalPop). So then we fitted the weighted least squares model and compared it to the ordinary least squares.

```
#fitting the weighted least squares model and comparing to the OLS
WLS <- lm(log(Physicians) ~ log(TotalPop) + LandArea + IncPerCap, weights = (1/log(TotalPop)), CDI)
summary(WLS)
```

```
##
## Call:
## lm(formula = log(Physicians) ~ log(TotalPop) + LandArea + IncPerCap,
##     data = CDI, weights = (1/log(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(TotalPop)  1.258e+00  3.408e-02  36.920  < 2e-16 ***
## LandArea      -2.876e-05  1.559e-05  -1.844   0.0659 .
## 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
```

$H_0: \ \beta_j = 0 \ \forall j = 1, 2, 3$ vs $H_1: \ \beta_j \neq 0$ for some j
F-statistic: 672.2 on 3 and 421 DF, p-value: $<$ 2.2e-16≈ 0
Since p $<$ 2.2e-16$\approx 0 < 0.05$, at level $\alpha = 0.05$ we reject the null hypothesis and state that at least one of the coefficients does not equal zero.

```
summary(lmvariance)
```

```
##
## Call:
## lm(formula = log(Physicians) ~ log(TotalPop) + LandArea + 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(TotalPop)  1.255e+00  3.323e-02  37.780  < 2e-16 ***
## LandArea      -2.980e-05  1.513e-05  -1.970   0.0495 *
## 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
```

$H_0: \ \beta_j = 0 \ \forall j = 1, 2, 3$ vs $H_1: \ \beta_j \neq 0$ for some j F-statistic: 705.2 on 3 and 421 DF, p-value: $<$ 2.2e-16
Since p $<$ 2.2e-16$\approx 0 < 0.05$, at level $\alpha = 0.05$ we reject the null hypothesis and state that at least one of the coefficients does not equal zero.

The fitted coefficients and the standard errors for the predictor log(totalPop) and LandArea, and the intercept increased with the weighted least squares model. The coefficient for IncPerCap increased with the weighted least squares model, but its standard error decreased.

We did a diagnostics test to see if all the linear regression assumption were help. Through the analyzation of the residuals Vs Fitted, Normal Q-Q, and Scale Location we concluded that the fit had a non-constant variance, a few large residuals, and some extreme right-skew to the residuals that indicate non-normality. Because of this, we decided to conduct a transformation. The boxCox showed that 0 was in the interval signifying that our response needed a log transformation. We used the boxtidwell command to check if any of the predictors needed a transformation and because they all had a > 0.05 we fail to reject H0 at alpha level .05 and conclude that the predictors have a linear relationship with log(Physicians) and we do not need to transform them. From the analysis, we reaffirm our initial intuition that LandArea does not have significant correlation with log(Physicians). Additionally, from the F test we concluded that log(TotalPop)

and IncPerCap have significant effects on log(Physicians). We also find that two of the predictors are not significant in the model. And the value of R2 is not big enough so the model is not good and we need to transform it. Additionally, we plot the residuals against the predictor LogTotalPop to see if the variance is increasing or decreasing. From the plot we clearly see the residuals decreasing as the values for logTotalPop increased. We then conducted a non constant variance test using the NCV test and saw that the null hypothesis of the constant variance holding is not rejected with just logTotalPop and the complete model but is rejected when the predictors are only LandArea and IncPerCap. So then we fitted the weighted least squares model and compared it to the ordinary least squares. From the summary of the weighted least regression test we see that the the p-value is less than 0.05 then at alpha level .05 we reject the null hypothesis and state that at least one of the coefffients does not equal to zero.

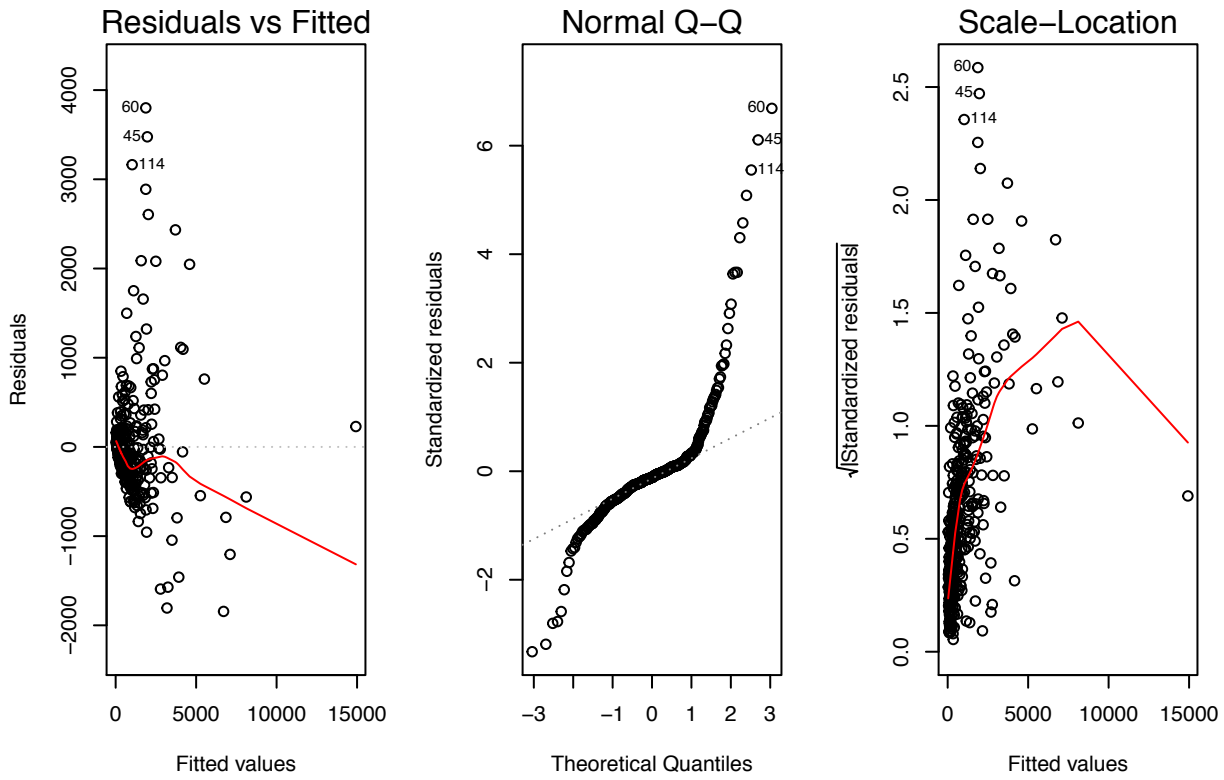
Part II

We will now investigate the model: Physicians∼TotalPop+Region

```
fit2 <- lm(Physicians~TotalPop + Region)
summary(fit2)
```

```
##
## Call:
## lm(formula = Physicians ~ TotalPop + Region)
##
## 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
```

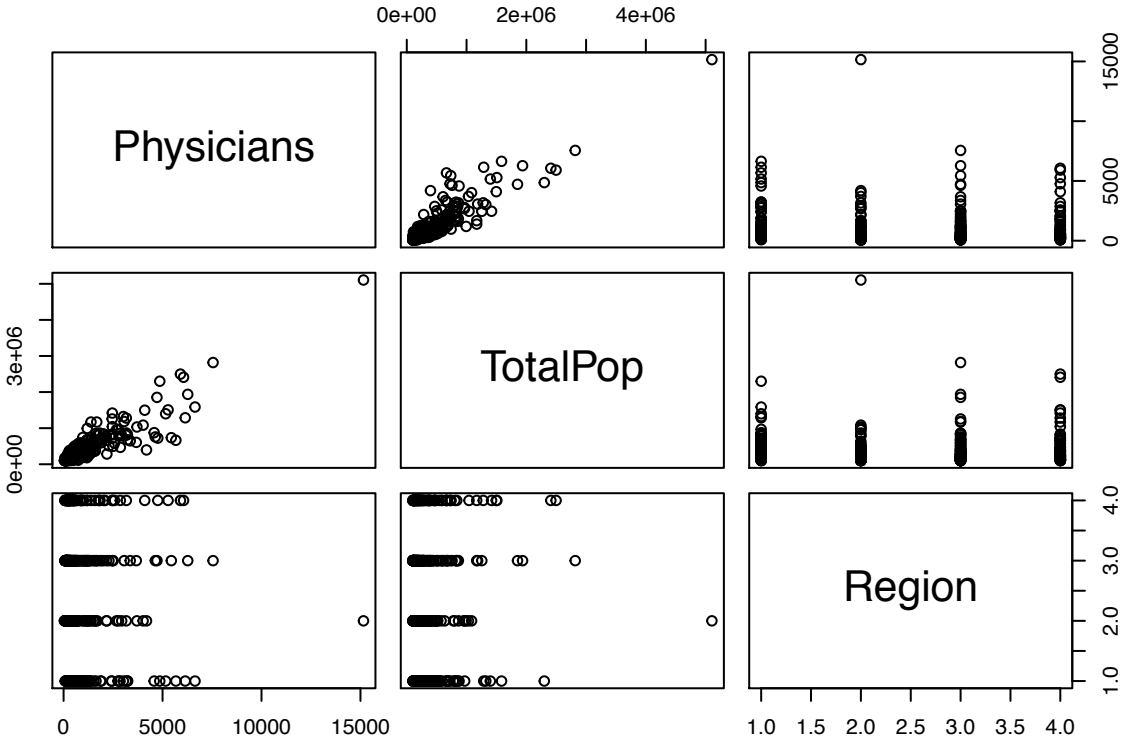
```
# diagnostic checking original model
par(mfrow = c(1,3))
for (i in 1:3){
  plot(fit2, which = i)
}
```



In the Residuals Vs. Fitted Plot, according to the assumption that variances of the error terms are equal and the relationship is linear is reasonable, the points should distribute around the 0 line randomly. But for plot here, the Residuals vs Fits plot shows the violations of Linearity and Constant (Error) Variance. In the normal Q-Q Plot, we expect to see the points forming a line that's roughly straight but the curve here is right-skewed so the model violates the assumption of normality. In the Scale-Location Plot, the residuals should be spread equally along with the ranges of predictors. Thus the model violates the assumption of constant variance.

We will transform the predictors to find a better fitted model:

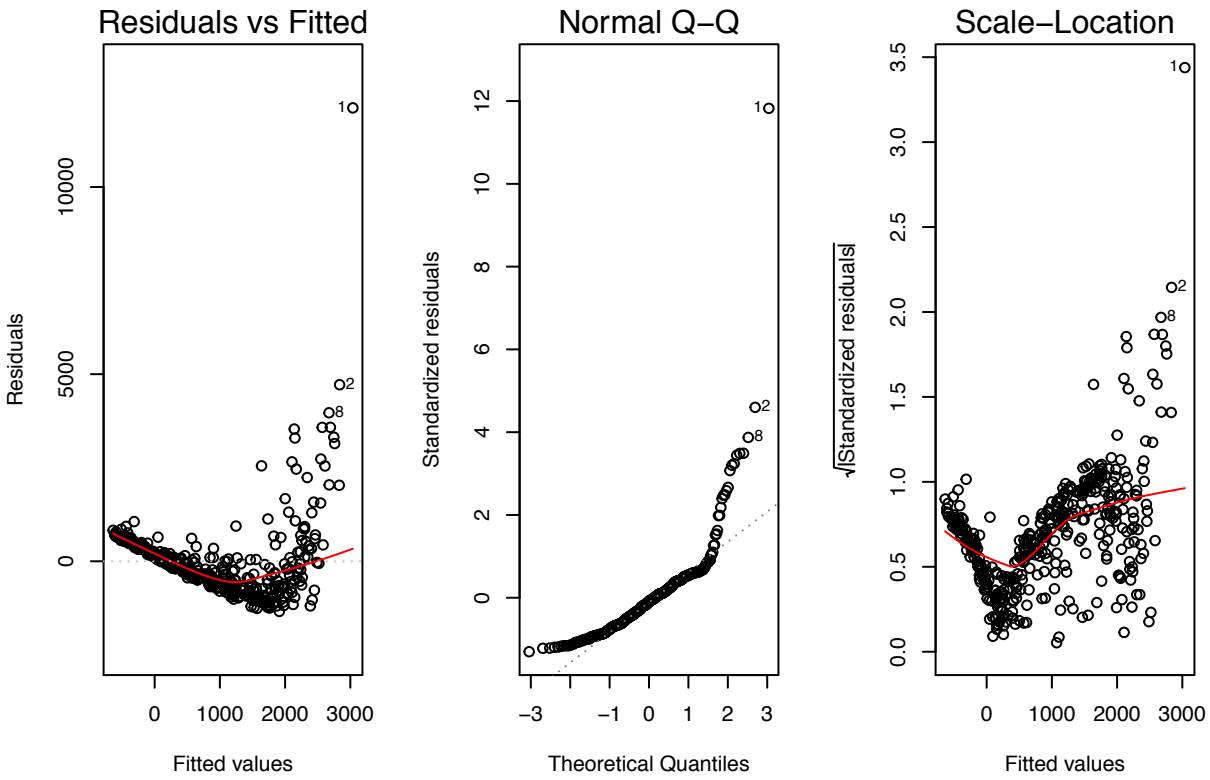
```
# transforming the predictors
pairs(CDI[c('Physicians','TotalPop','Region')])
```




```
hwy.pt = powerTransform(cbind(TotalPop) ~ 1, CDI)
summary(hwy.pt)
```

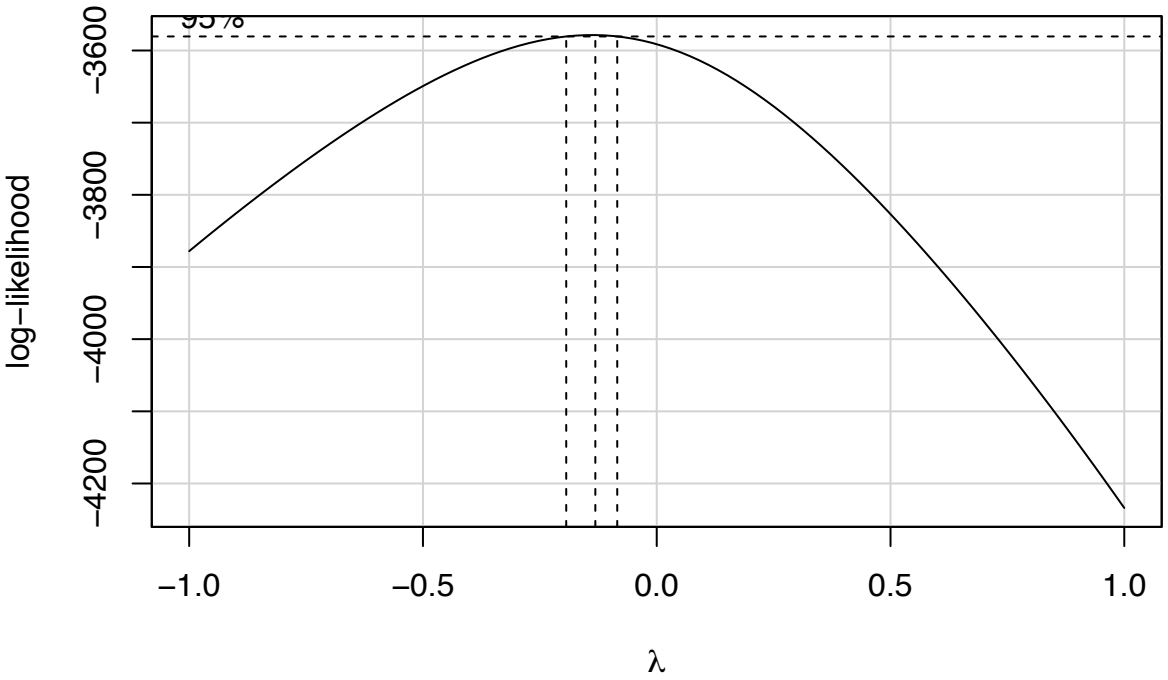
```
## bcPower Transformation to Normality
##    Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## Y1   -0.5799        -0.5      -0.7207       -0.439
##
## Likelihood ratio test that transformation parameter is equal to 0
##  (log transformation)
##                             LRT df       pval
## LR test, lambda = (0) 76.25795  1 < 2.22e-16
##
## Likelihood ratio test that no transformation is needed
##                             LRT df       pval
## LR test, lambda = (1) 759.2153  1 < 2.22e-16
```

```
# new model with transformed predictors
fit2.1 <- lm(Physicians~I(TotalPop^(-0.5799))+Region)
par(mfrow = c(1,3))
for (i in 1:3){
  plot(fit2.1, which = i)
}
```



The residual plots violate the diagnostics checks. We will now try to transform the response.

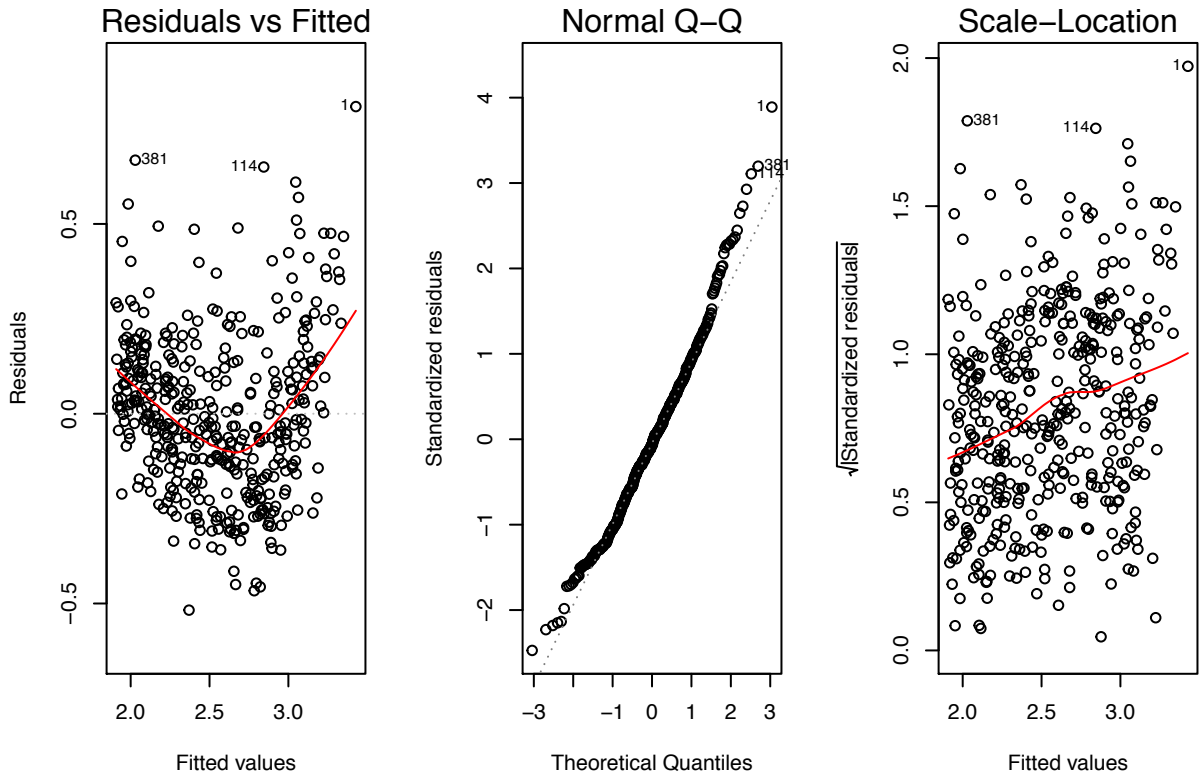
```
# transfroming the response
best.lambda<- boxCox(fit2.1,lambda=seq(-1,1,by=.1))
```




```
best.lambda$x[which.max(best.lambda$y)]
```

```
## [1] -0.1313131
```

```r
# new model with transofmred predictors and response
fit2.tr=lm(Physicians^0.15~I(TotalPop^(-0.5799)) + Region)
par(mfrow = c(1,3))
for (i in 1:3){
  plot(fit2.tr, which = i)
}
```



The new model does not violate any model diagnostics. The model after transformation is:
$\text{Physicians}^{0.15} \sim I(\text{TotalPop}^{-0.5799}) + \text{Region}$

```r
summary(fit2.tr)
```

```
##
## Call:
## lm(formula = Physicians^0.15 ~ I(TotalPop^(-0.5799)) + Region)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.51748 -0.14305 -0.01371  0.12341  0.80916
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              3.615e+00  3.843e-02  94.066   <2e-16 ***
## I(TotalPop^(-0.5799)) -1.333e+03  3.487e+01 -38.241   <2e-16 ***
## Region                -7.446e-03  9.822e-03  -0.758    0.449
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2096 on 422 degrees of freedom
## Multiple R-squared:  0.7761, Adjusted R-squared:  0.775
```

```
## F-statistic: 731.2 on 2 and 422 DF,  p-value: < 2.2e-16
```

```
fit2.tr.interaction <- lm(Physicians^0.15~I(TotalPop^(-0.5799)) * Region)
summary(fit2.tr.interaction)
```

```
##
## Call:
## lm(formula = Physicians^0.15 ~ I(TotalPop^(-0.5799)) * Region)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.51417 -0.14031 -0.01351  0.12715  0.79593
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   3.694e+00  7.351e-02   50.25   <2e-16 ***
## I(TotalPop^(-0.5799))        -1.438e+03  9.018e+01  -15.95   <2e-16 ***
## Region                       -3.898e-02  2.688e-02   -1.45    0.148
## I(TotalPop^(-0.5799)):Region  4.197e+01  3.331e+01    1.26    0.208
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2095 on 421 degrees of freedom
## Multiple R-squared:  0.7769, Adjusted R-squared:  0.7753
## F-statistic: 488.7 on 3 and 421 DF,  p-value: < 2.2e-16
```

Interaction term p-value is 0.208> 0.05, which implies that this is a parallel model as the interaction between predictors is not significant. Coefficient for region is not significant (p = 0.449 > 0.05) so each region has the same expected physicians based on total population.

The equations for estimated mean number of physicians as a function of the total population and geographic region are as follows: NE:
E(Yi) = 0 -1.333e+03xi4
NC
E(Yi) = 0 -1.333e+03xi4
S
E(Yi) = 0 -1.333e+03xi4
W
E(Yi) = 0 -1.333e+03xi4
Our model is a arallel model because they all have the same slope.

```
fit2.sub <- lm(Physicians^0.15~I(TotalPop^(-0.5799)))
fit2.full <- lm(Physicians^0.15~I(TotalPop^(-0.5799)) + Region)
anova(fit2.sub, fit2.full)
```

```
## Analysis of Variance Table
##
## Model 1: Physicians^0.15 ~ I(TotalPop^(-0.5799))
## Model 2: Physicians^0.15 ~ I(TotalPop^(-0.5799)) + Region
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    423 18.569
## 2    422 18.544  1  0.025254 0.5747 0.4488
```

The p-value .4488, is more than .05. Hence we fail to reject H_0: Region = 0 at an α level of .05, and can state that region is an insignificant predictor when other predictors are included in the model.

We will use forward and backward seleciton with AIC to determine if Pop65, Crimes, Bachelor, Poverty, and

PersonalInc are relevant predictors.

```
mod.0 <- lm(Physicians^0.15 ~ 1)
mod.full <- lm(Physicians^0.15 ~ I(TotalPop^-0.5799) + Pop65 + Crimes + Bachelor + Poverty + PersonalInc

# forward selection with AIC
step(mod.0, scope = list(lower = mod.0, upper = mod.full), direction = "forward")
```

```
## Start:  AIC=-693.12
## Physicians^0.15 ~ 1
##
##                        Df Sum of Sq    RSS      AIC
## + I(TotalPop^-0.5799)  1    64.238 18.569 -1326.50
## + PersonalInc          1    51.160 31.647 -1099.92
## + Crimes               1    31.134 51.673  -891.54
## + Bachelor             1    17.443 65.363  -791.65
## <none>                              82.807  -693.12
## + Pop65                1     0.026 82.781  -691.25
## + Poverty              1     0.013 82.794  -691.19
##
## Step:  AIC=-1326.5
## Physicians^0.15 ~ I(TotalPop^-0.5799)
##
##                Df Sum of Sq    RSS     AIC
## + PersonalInc  1    4.2973 14.272 -1436.4
## + Crimes       1    2.1663 16.403 -1377.2
## + Bachelor     1    1.9697 16.599 -1372.2
## + Poverty      1    0.2432 18.326 -1330.1
## + Pop65        1    0.1368 18.432 -1327.6
## <none>                     18.569 -1326.5
##
## Step:  AIC=-1436.36
## Physicians^0.15 ~ I(TotalPop^-0.5799) + PersonalInc
##
##            Df Sum of Sq    RSS     AIC
## + Bachelor  1   1.46099 12.811 -1480.3
## + Poverty   1   0.45129 13.821 -1448.0
## + Pop65     1   0.14199 14.130 -1438.6
## <none>                  14.272 -1436.4
## + Crimes    1   0.04220 14.230 -1435.6
##
## Step:  AIC=-1480.26
## Physicians^0.15 ~ I(TotalPop^-0.5799) + PersonalInc + Bachelor
##
##           Df Sum of Sq    RSS     AIC
## + Poverty  1   1.63670 11.174 -1536.4
## + Pop65    1   0.75692 12.054 -1504.2
## + Crimes   1   0.23973 12.571 -1486.3
## <none>                 12.811 -1480.3
##
## Step:  AIC=-1536.36
## Physicians^0.15 ~ I(TotalPop^-0.5799) + PersonalInc + Bachelor +
##     Poverty
##
##          Df Sum of Sq     RSS     AIC
```

```
## + Pop65   1   1.22541  9.9488 -1583.7
## <none>                11.1742 -1536.4
## + Crimes  1   0.00963 11.1645 -1534.7
##
## Step:  AIC=-1583.72
## Physicians^0.15 ~ I(TotalPop^-0.5799) + PersonalInc + Bachelor +
##     Poverty + Pop65
##
##          Df Sum of Sq    RSS     AIC
## <none>                9.9488 -1583.7
## + Crimes  1  0.028418 9.9203 -1582.9
```

```
##
## Call:
## lm(formula = Physicians^0.15 ~ I(TotalPop^-0.5799) + PersonalInc +
##     Bachelor + Poverty + Pop65)
##
## Coefficients:
##         (Intercept)  I(TotalPop^-0.5799)          PersonalInc
##           2.466e+00           -8.693e+02            1.440e-05
##            Bachelor              Poverty                Pop65
##           1.589e-02            1.686e-02            1.454e-02
```

```
# backward selsction with AIC
step(mod.full, scope = list(lower = mod.0, upper = mod.full), direction = 'backward')
```

```
## Start:  AIC=-1582.94
## Physicians^0.15 ~ I(TotalPop^-0.5799) + Pop65 + Crimes + Bachelor +
##     Poverty + PersonalInc
##
##                       Df Sum of Sq     RSS     AIC
## - Crimes               1    0.0284  9.9488 -1583.7
## <none>                              9.9203 -1582.9
## - Pop65                1    1.2442 11.1645 -1534.7
## - Poverty              1    1.7838 11.7042 -1514.7
## - PersonalInc          1    1.7994 11.7197 -1514.1
## - Bachelor             1    3.7574 13.6777 -1448.4
## - I(TotalPop^-0.5799)  1   12.7141 22.6344 -1234.4
##
## Step:  AIC=-1583.72
## Physicians^0.15 ~ I(TotalPop^-0.5799) + Pop65 + Bachelor + Poverty +
##     PersonalInc
##
##                       Df Sum of Sq     RSS     AIC
## <none>                              9.9488 -1583.7
## - Pop65                1    1.2254 11.1742 -1536.4
## - Poverty              1    2.1052 12.0540 -1504.2
## - Bachelor             1    3.7302 13.6790 -1450.4
## - PersonalInc          1    3.9152 13.8640 -1444.7
## - I(TotalPop^-0.5799)  1   12.7427 22.6915 -1235.3
```

```
##
## Call:
## lm(formula = Physicians^0.15 ~ I(TotalPop^-0.5799) + Pop65 +
##     Bachelor + Poverty + PersonalInc)
##
```

```
## Coefficients:
##         (Intercept)  I(TotalPop^-0.5799)                Pop65
##           2.466e+00           -8.693e+02            1.454e-02
##            Bachelor              Poverty          PersonalInc
##           1.589e-02            1.686e-02            1.440e-05
```

Both forward and backward selection with AIC determined that our new model should be $\text{Physicians}^{0.15} \sim I(\text{TotalPop}^{-0.5799})$+Pop65 + Bachelor + Poverty + PersonalInc

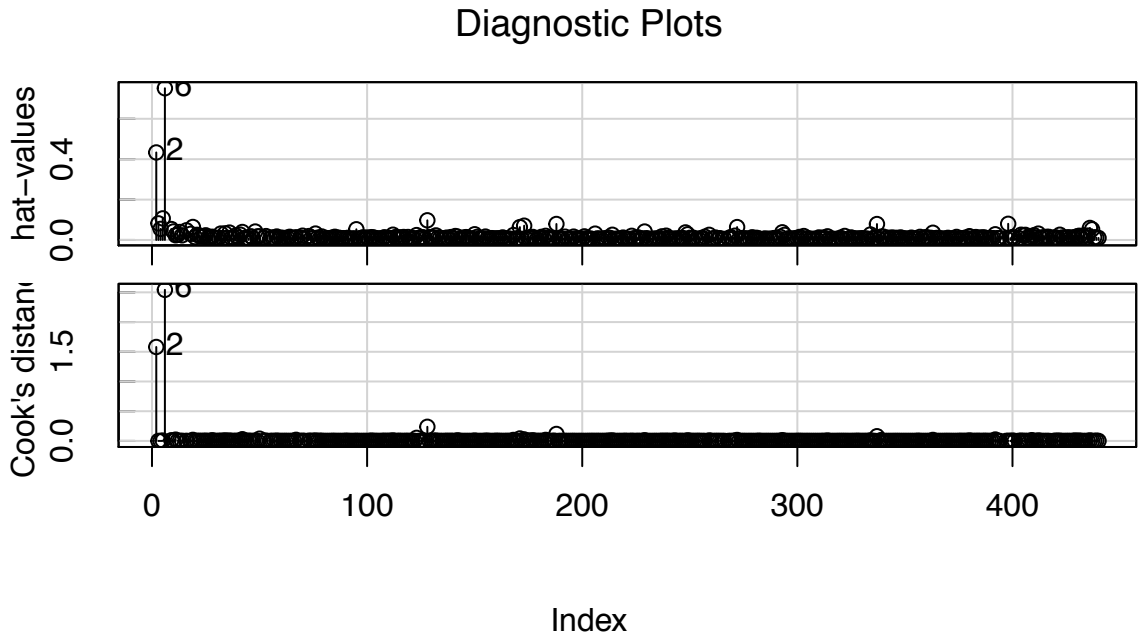
We will perform a partial F-test to assess whether the improvement from adding these predictors compared to the first model:

```
# comparing our model from AIC selection with submodel
mod.0 <- lm(Physicians^0.5 ~ I(TotalPop^-0.5799))
mod.full <- lm(Physicians^0.5 ~ I(TotalPop^-0.5799) + Pop65 + Bachelor + Poverty + PersonalInc)
anova(mod.0,mod.full)
```

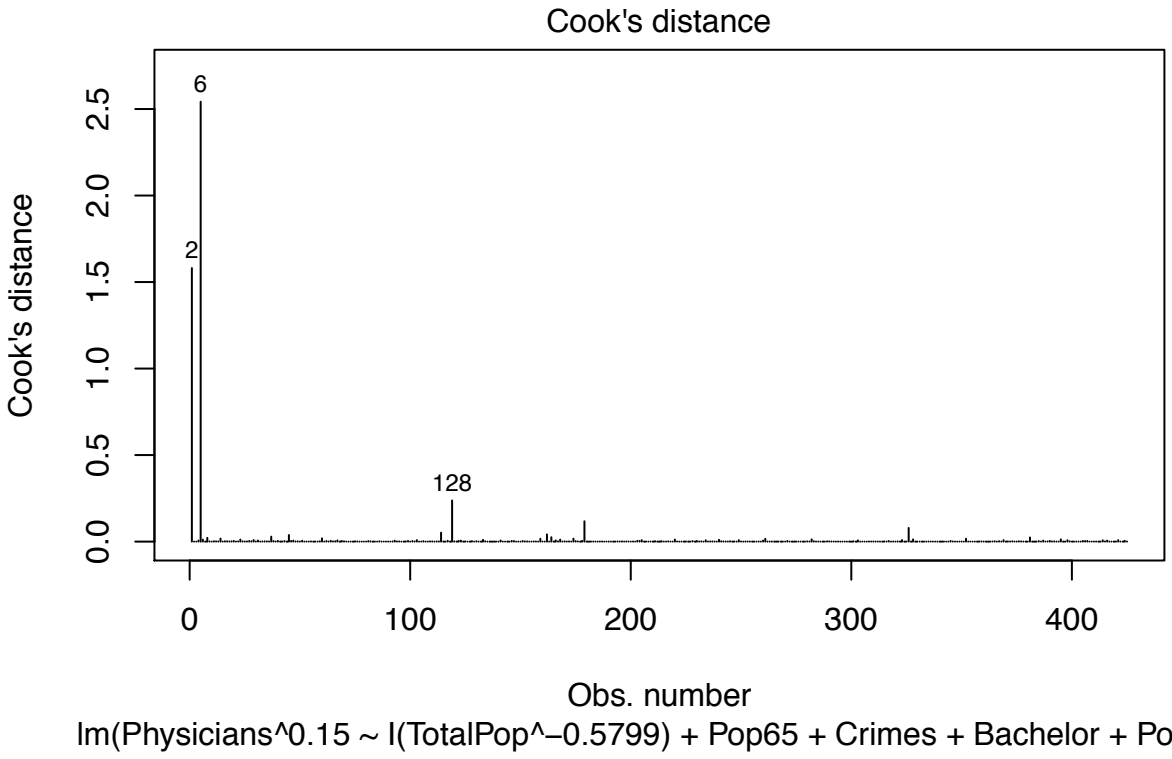
```
## Analysis of Variance Table
##
## Model 1: Physicians^0.5 ~ I(TotalPop^-0.5799)
## Model 2: Physicians^0.5 ~ I(TotalPop^-0.5799) + Pop65 + Bachelor + Poverty +
##     PersonalInc
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1    423 36819
## 2    419 14048  4     22771 169.79 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

H_0: Submodel holds vs H_1: Full model holds
F = 169.79, p=value < 2.2e-16$\approx 0 < 0.05$, we reject the null hypothesis at alpha level .05. Therefore, we can conclude that the submodel does not hold, hence we prefer the full model.

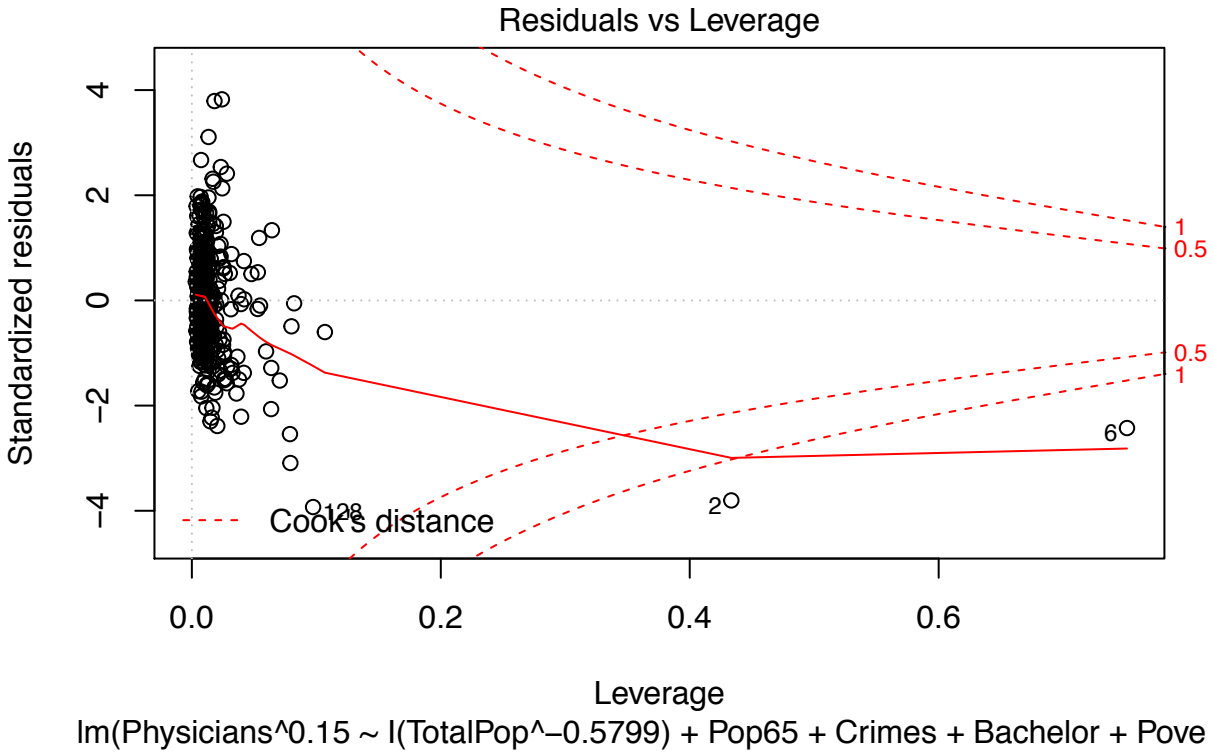
```
mod.full <- lm(Physicians^0.15 ~ I(TotalPop^-0.5799) + Pop65 + Crimes + Bachelor + Poverty + PersonalInc
influenceIndexPlot(mod.full, vars = c('hat', 'Cook'))
```




```
plot(mod.full,which=4)
```




```
plot(mod.full,which=5)
```



So data 2, 6 are influential. For data 2 and data 6, they both have high leverage, residuals, and high cook's distances. But for data 128, even though it's pointed out in two of these plots as a deviation from the scutter of data points, it doesn't have high leverage or high leverage. So data 128 is not influential here.

With our new fitted model, we concluded that region does not have an affect on the amount of Physicians.

We conducted a partial F test using the anova command to see if the geographic region has a significant effect on the number of physicians in a county. After making two separate models, one including the region predictor and another not we conducted the test. The The p-value ended up being .4488, which is more than .05. Hence we fail to reject Ho: Region = 0 at an alpha level of .05, and can state that region is an insignificant predictor when other predictors are included in the model. Thus, we can disregard region and use a model that focuses on total population. From forward and backward model selection, we concluded that we should add Pop65, Bachelor, Poverty, and PersonalInc to the model. In part d we performed a partial F-test to assess whether the improvement from adding these predictors compared to the first model is statistically significant at $\alpha = 0.05$. By utilizing the anova command again, our test game us a P-value of 2.2eˆ-16 signifying that we prefer the full model that included all the new predictors.

Conclusion

In the first part of our project, we conducted a test on the fitted model Physicians $\sim$ *log*(TotalPop) + LandArea + IncPerCap. From initial observations of the scatterplot matrix, we concluded that Physicians is positively correlated with log(TotalPop) and IncPerCap, and that there's no apparent linear relationship between Physicians and LandArea. Additionally, LandArea and IncPerCap are negatively correlated with logTotalPop. Through the use of diagnostic plots, we used a Box-Cox transformation on the response, which led us to our final model log(Physicians)$\sim$log(TotalPop) + LandArea + IncPerCap. In the second part, we tested the model Physicians$\sim$TotalPop+Region and found that both our predictor TotalPop and response had to be transformed. We tested if adding Region would be useful for our model, and found it held no significance on the response. Because IncPerCap showed a straight regression line in the AV plot from our first model, we were surprised to see that when the test of transforming the predictors was made, IncPerCap seemed to hold a strong linear relationship with the response. Lastly, with AIC and BIC, we found that our best model for the second part to be Physiciansˆ0.15$\sim$I(TotalPop(-0.5799))+Pop65 + Bachelor + Poverty + PersonalInc.