Analysis of Physicians in the CDI Data

Kelly Xiao
Cassie Guo
Xin Chen

ABSTRACT:

In this statistical report, we are interested in how the expected mean number of professionally active nonfederal physicians during 1990 (denoted as "Physicians") is affected by different variables. After realizing transformations are needed, our objective changes to analyze the estimated median of the number of physicians. We want to know if there exists a linear relationship between the median of Physicians with various variables that are also in CDI.

In part I, our investigation involves three continuous variables, while in part II, we analyzed one continuous variable and one categorical variable. By utilizing Multiple Linear Regression techniques, we came to the conclusion that after transformations and selections of different models, there exists a linear relationship between the median of the numbers of physicians and some predictors of interest.

PROBLEM AND MOTIVATION:

Physicians play a significant role in the society as people's health are largely depended on them. Many factors can be taken into consideration while talking about the number of physicians in a specific county. How do we know which factors are the most important? This statistical report will investigate this question by the CDI data.

This data set provides selected demographic information for 440 of the most populous counties in the United States during the years of 1990 and 1992. In our report, in particular, we are interested in the number of professionally active nonfederal physicians in 1990. The motivation of our report is to show how certain factors such as the environment and a population's living standards can have impacts on physicians in a county. There are 16 variables given in the CDI data, which some are more closely related with physicians than others. From this report, we pick out a few and investigate their significance. Readers will be able to see some initial expectations that followed by statistical analysis, and some unexpected findings between the number of physicians and multiple predictors.

DATA:

The data set used in this report is county demographic information (CDI). The relevant variables of interest in this report are the number of professionally active nonfederal physicians during 1990 (denoted as "Physicians"), the estimated 1990 population (TotalPop), land area (LandArea), per capita income of 1990 CDI population (IncPerCap), and geographic regions (Region) which is classified by the U.S. Bureau of the Census. Later in the report we also consider variables such as the percent of aged 65 and older population, crimes, percent of adult with bachelor's degree, poverty level, and total personal income.

QUESTIONS OF INTEREST:

In this report, our analysis are interested in whether there exists some linear relationship between the average number of physicians in 1990 and some relevant variables. We'll answer some questions for readers who might be interested in knowing what kinds of factors impact the number of active physicians during 1990. And also among those factors, what kinds of associations exist. On the other hand, we'll also show how some variables do not have an effect on physicians in 1990.

REGRESSION METHODS:

In this report, we use Multiple Linear Regression tools to analyze our questions of interest. For a given model, we first do diagnostic checks using plots and numerical summaries to see which assumptions of MLR are violated. Then, according to the specific violations, we conduct transformations to either the response or the predictors or both. Next, we do diagnostic checks again and test the new model after transforming with the old model to see whether the transformation made improvements. Transformations involve with log transform and power transform. With the fitted model, we compute confidence intervals and conduct hypothesis tests to see whether each predictor is indeed significant to our model. Moreover, we might need to use some other methods to modify our model, such as weighted least square method (WLS). However, when we only have some potential predictors but do not have a specific model yet, we need to use model selection techniques to choose the best model. In addition, we also need to identify any influential points that stand out in the data set.

REGRESSION ANALYSIS, RESULTS and INTERPRETATION:

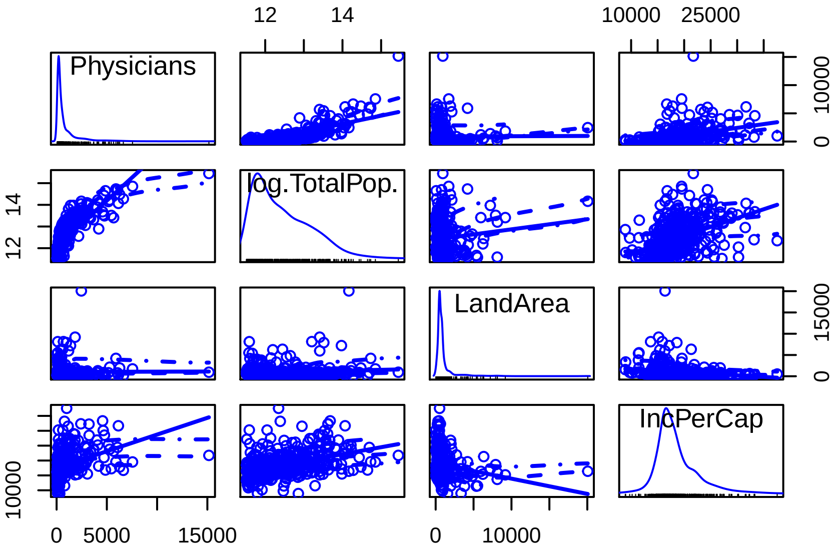
NOTE: For the sake of a more concise report, we hid some codes (but kept the results) by using "echo = FALSE", "Include = False" and so on, for example, we hid the codes of "avPlots()" but kept the plots. To see the full codes of our report, please go to the Rmarkdown file that we uploaded as well.

Part I

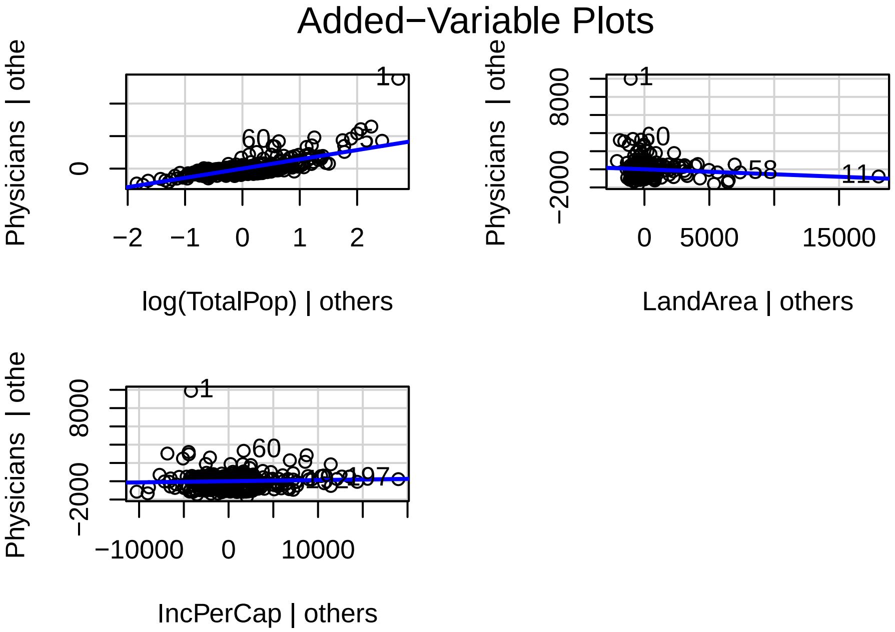
We will investigate the model Physicians ~ log(TotalPop) + LandArea + IncPerCap.

At a first glance, before we do anything to the model, we expected to see a positive relationship between the number of professionally active nonfederal physicians and total population since we think a more populated area would lead to higher demands for physicians intuitively. A higher per capita income might also attract more physicians to a county since there are more resources as well as a higher consumption level. However, we did not have any idea about how Land area by itself is related to the mean number of physicians in general. So, now we do some explanatory analysis:

First, we examine the relationship between Physicians, the response, with each predictor.



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



From the matrix scatterplots, we see positive correlations between Physicians and both log(TotalPop) and Land Area; positive relationship between log(TotalPop) and both Land Area and IncPerCap, but Land Area and IncPerCap seem to have no relationship. Then from the added-variable plots, log(TotoalPop) has a strong positive linear relationship with Physicians, the response, after controlling the effects of the other two predictors, but LandArea and IncPerCap seem to share almost no relationship with Physicians respectively (indicating transformations are needed later).

We fit the model: Physicians ~ log(TotoalPop) + LandArea + IncPerCap

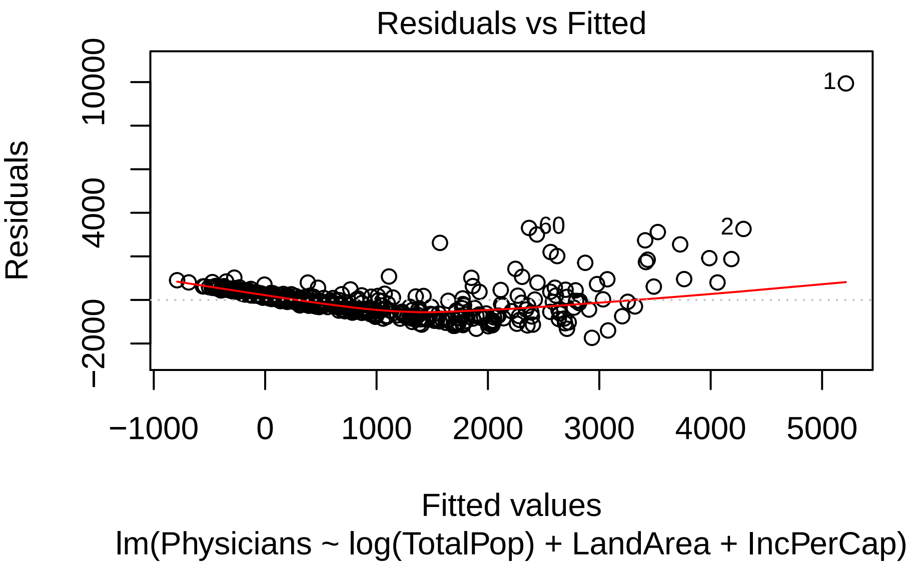
```
##                    Estimate   Std. Error    t value     Pr(>|t|)
## (Intercept)   -1.705931e+04 705.95857946 -24.164745 1.472441e-81
## log(TotalPop)  1.427327e+03  62.92620162  22.682557 5.313205e-75
## LandArea      -5.487678e-02   0.02864665  -1.915644 5.608750e-02
## IncPerCap      1.284710e-02   0.01190268   1.079346 2.810517e-01
```

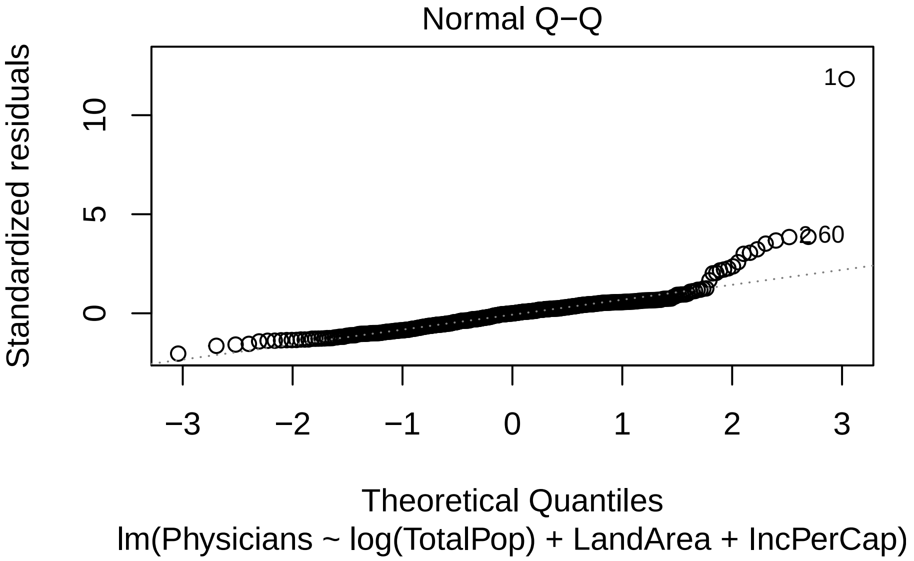
```
## [1] 0.6175083
```

Using the summary table to extract the coefficients and adjusted-R^2. We see that as log(TotalPop) increased by one unit, the mean change in Physicians is 1427 after controlling LandArea and IncPerCap. Similarly, for one unit increase in IncPerCap, mean number of Physicians is changed by 0.0128. For one unit increase in LandArea, the mean number of Physicians is decreased by 0.0549 after controlling Land Area and log(TotalPop). Before transformation, the latter two seem to contribute little to our model.

We also notice a large standard error and only 61.75% of variability in the response is explained by the multiple linear regression model with the three predictors collectively.

Diagnostic checks and transformations:





After running diagnostics using Residuals vs. Fitted and Normal Q-Q plot, we see severe violations of non-linearity, non-constant variance, and non-normality. Therefore, we need to do transformations on both predictors and the response.

```
Trans.cdi <- powerTransform(cbind(LandArea, IncPerCap) ~ 1, CDI)
summary(Trans.cdi)
```

```
## bcPower Transformations to Multinormality
##           Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## LandArea    -0.0154         0.0      -0.0799       0.0490
## IncPerCap   -0.3741        -0.5      -0.6779      -0.0704
##
## Likelihood ratio test that transformation parameters are equal to 0
##  (all log transformations)
##                               LRT df     pval
## LR test, lambda = (0 0) 5.993834  2 0.049941
```

```
##
## Likelihood ratio test that no transformations are needed
##                               LRT df       pval
## LR test, lambda = (1 1) 915.2652  2 < 2.22e-16
```

We first transform predictors, from the powerTranform function above, taking log transformation on LandArea, and inverse of the square root of IncPerCap seem the most optimal. We leave log(TotalPop) as it is since it's already a useful predictor in the model. Now move on to the transformation of response:

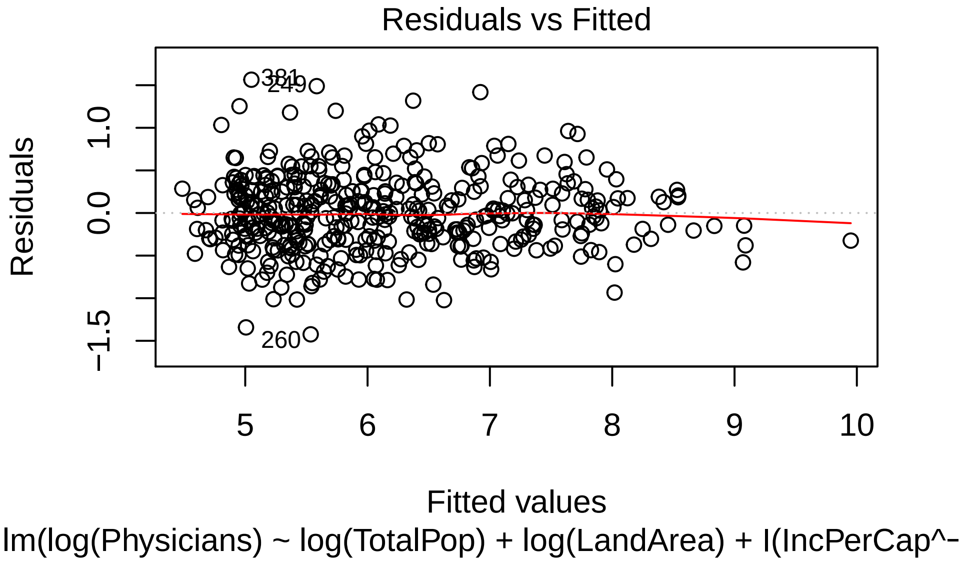
```
CDItrans <- with(CDI, data.frame(Physicians, log(TotalPop), log(LandArea), I(IncPerCap)^-0.5))
cdi.power <- powerTransform(Physicians ~.,CDItrans )
summary(cdi.power)
```

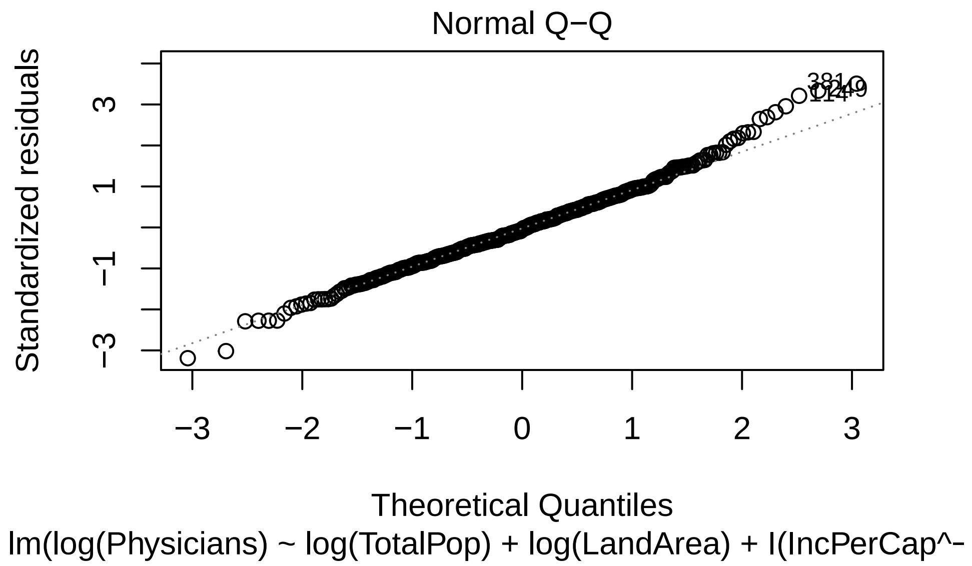
```
## bcPower Transformation to Normality
##    Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## Y1    0.0329           0      -0.0184       0.0842
##
## Likelihood ratio test that transformation parameter is equal to 0
##  (log transformation)
##                             LRT df    pval
## LR test, lambda = (0) 1.544583  1 0.21394
##
## Likelihood ratio test that no transformation is needed
##                             LRT df       pval
## LR test, lambda = (1) 1229.647  1 < 2.22e-16
```

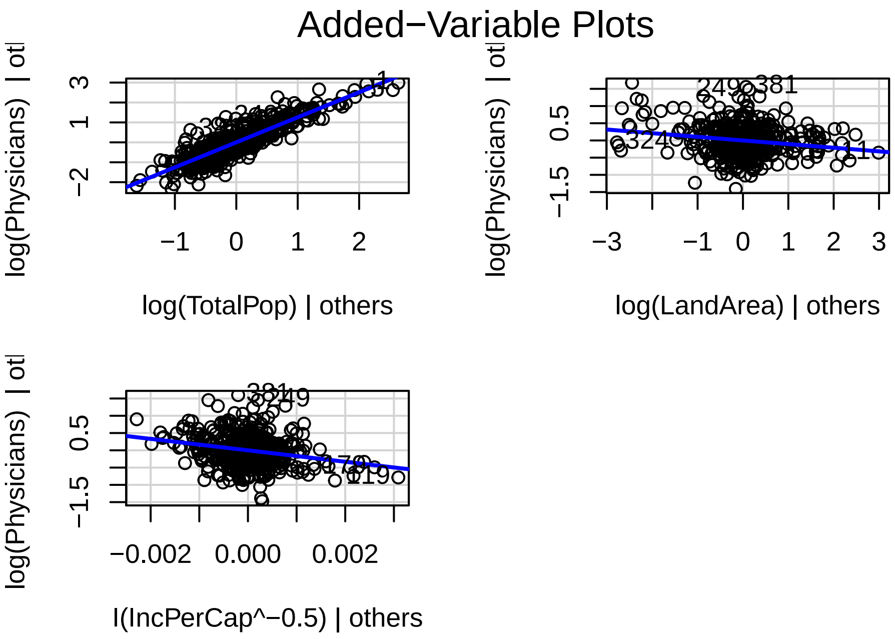
Using the same method, log transforming the response is the best option. The transformed model is: log(Physicians) ~ log(TotalPop) + log(LandArea) + IncPerCap$^{-0.5}$.

Now we do diagnostic checks for the new model, Residual v.s Fitted plot, Normal Q-Q plot, and Added-Variable plot, as follows:

```
newCDI <- lm(log(Physicians) ~ log(TotalPop) + log(LandArea) + I(IncPerCap^-0.5))
```







By comparing the new plots with the plots before transforming, significant improvements on linearity and equal variance are shown. Although there is still a little deviation from normality at the tail end by looking at the Q-Q plot, we decide to keep the current model. The final model is log(Physicians) ~ log(TotalPop) + log(LandArea) + IncPerCap$^{-0.5}$.

95% Confidence intervals and hypothesis tests:

```
##                          2.5 %       97.5 %
## (Intercept)          -8.685567  -6.52759789
## log(TotalPop)         1.192687   1.31987357
## log(LandArea)        -0.158806  -0.05277783
## I(IncPerCap^-0.5) -231.257174 -99.63312984
```

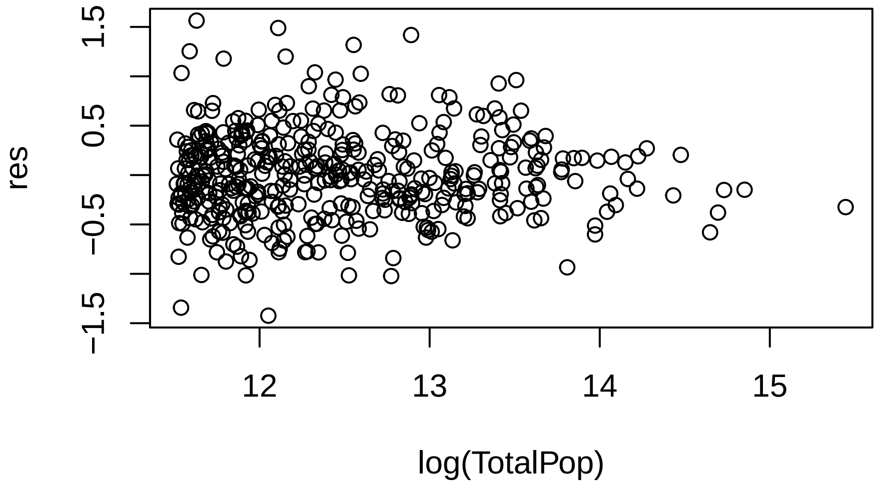
```
##
## Call:
```

```
## lm(formula = log(Physicians) ~ log(TotalPop) + log(LandArea) +
##     I(IncPerCap^-0.5))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.42421 -0.29010 -0.01979  0.27075  1.56441
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -7.60658    0.54893 -13.857  < 2e-16 ***
## log(TotalPop)         1.25628    0.03235  38.831  < 2e-16 ***
## log(LandArea)        -0.10579    0.02697  -3.922 0.000102 ***
## I(IncPerCap^-0.5) -165.44515   33.48165  -4.941 1.12e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4472 on 421 degrees of freedom
## Multiple R-squared:  0.8389, Adjusted R-squared:  0.8377
## F-statistic: 730.7 on 3 and 421 DF,  p-value: < 2.2e-16
```

Conducting the 95% confidence interval, we are 95% confidence that log(TotalPop) is between 1.1927 and 1.3199; log(LandArea) is between -0.1588 and -0.0528; IncPerCap$^{-0.5}$ is between -231.2572 to -99.6331. Then we can use the summary function to conduct a test for linear relationships between predictors and response, with H_0: $\beta_i = 0$ vs H_1: $\beta_i \neq 0$, with $\alpha = 0.01$. The value of the T-statistic and its p-value is shown in the summary table above. Comparing each p-value with 0.01, we can reject all three null hypothesis and conclude that all predictors are useful and should be added in the model.

Now we do non-constant variance tests:

```r
res <- newCDI$residuals
plot(log(TotalPop), res, xlab = "log(TotalPop)")
```



We notice the variance decreases as log(TotalPop) increases from the plot above. We then perform ncv tests on the fitted values as well as the other two predictors, and refit using weighted least squares:

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 2.450299, Df = 1, p = 0.1175
```

P-value of ncvTest on fitted values is 0.1175 > 0.05, we fail to reject the null hypothesis and conclude that the assumption of constant variance on fitted values is NOT violated, but it is still useful to check ncv tests on predictors, as follows:

```
## Non-constant Variance Score Test
## Variance formula: ~ log(TotalPop)
## Chisquare = 4.30121, Df = 1, p = 0.038085
```

```
## Non-constant Variance Score Test
## Variance formula: ~ log(LandArea)
## Chisquare = 13.09886, Df = 1, p = 0.00029548
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.8079593, Df = 1, p = 0.36872
```

```
## Non-constant Variance Score Test
## Variance formula: ~ I(IncPerCap^-0.5)
## Chisquare = 0.1246725, Df = 1, p = 0.72402
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 1.009114, Df = 1, p = 0.31512
```

We refit three new models, which resulted that using log(LandArea) and IncPerCap$^{-0.5}$ are both adequate to be z_i. This corresponds to the weights being the inverse of log(LandArea) and IncPerCap$^{-0.5}$ respectively. Now we compare their fitted coefficients, standard error, and adjusted R^2.

```
summary(wlm2)$coefficients
```

```
##                       Estimate  Std. Error    t value      Pr(>|t|)
## (Intercept)         -7.6439259  0.56219658 -13.596536  3.682449e-35
## log(TotalPop)        1.2695606  0.03289260  38.597144 2.300205e-140
## log(LandArea)       -0.1297871  0.02653323  -4.891492  1.426952e-06
## I(IncPerCap^-0.5) -161.6325912 33.82346022  -4.778712  2.441440e-06
```

```
summary(wlm2)$sigma
```

```
## [1] 0.1797049
```

```
summary(wlm2)$adj.r.squared
```

```
## [1] 0.8366731
```

```
summary(wlm3)$coefficients
```

```
##                       Estimate  Std. Error    t value      Pr(>|t|)
## (Intercept)         -7.5337637  0.54540920 -13.813048  4.681239e-36
## log(TotalPop)        1.2542997  0.03201804  39.174776 1.721418e-142
## log(LandArea)       -0.1118265  0.02705700  -4.132997  4.322116e-05
## I(IncPerCap^-0.5) -166.6231335 33.50487988  -4.973100  9.612798e-07
```

```
summary(wlm3)$sigma
```

```
## [1] 5.198199
```

```
summary(wlm3)$adj.r.squared
```

```
## [1] 0.8416299
```

We conclude that wlm3 <- lm(log(Physicians) ~ log(TotalPop) + log(LandArea) + IncPerCap$^{-0.5}$, CDI, weights = 1/log(LandArea)) is a better refitted model since the standard error is lower than using

IncPerCap$^{-0.5}$ as z_i.

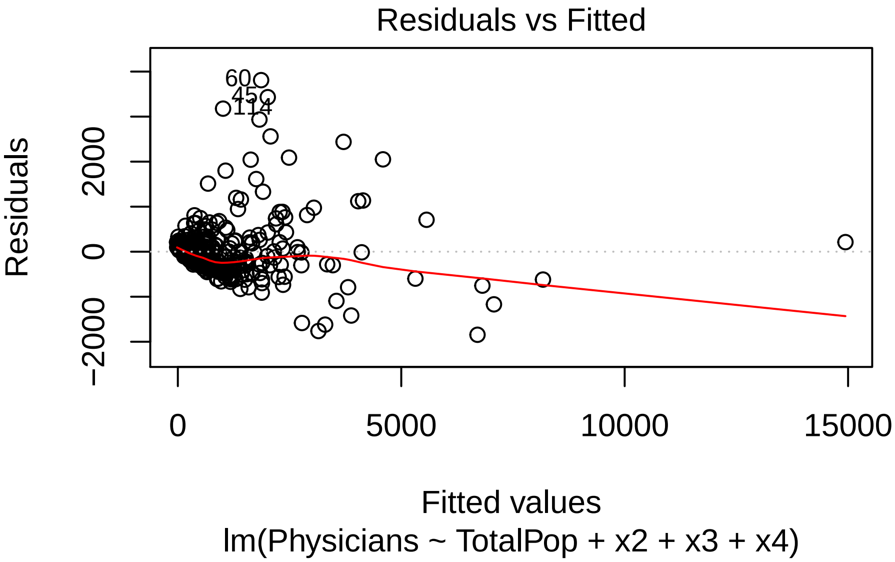
Comparing to the model: log(Physicians) ~ log(TotalPop) + log(LandArea) + I(IncPerCap^-0.5), adding a weight of the inverse of the log(LandArea) and refitting the model, results in a constant variance. There are minor changes to the fitted coefficients, but the standard error has lowered from 0.4472 to 0.1797 while keeping the adjusted R^2 constant.

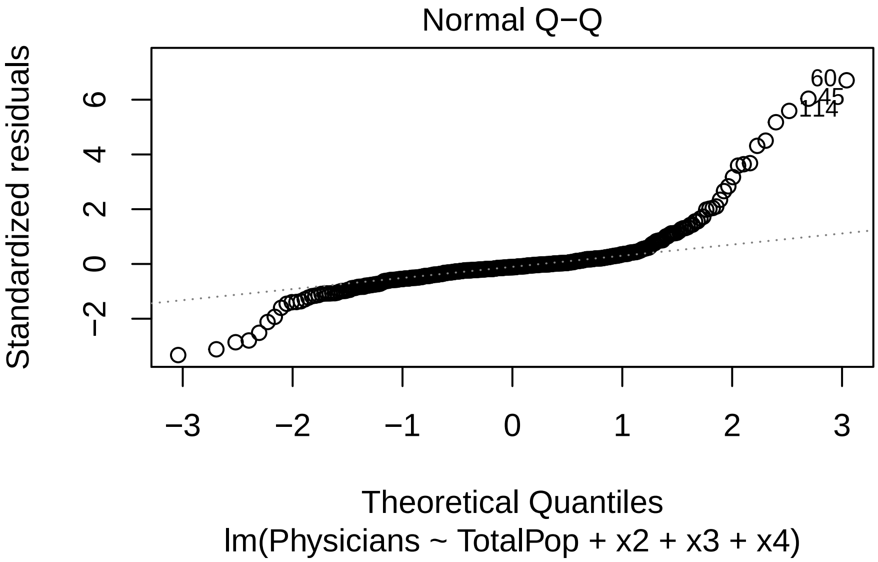
Part II

In the second part of this project, we will investigate the model: Physicians ~ TotalPop + Region

Since there are four regions, we first create three dummy variables for this categorical variable. Let $x1$ represent TotalPop, and since there are four regions, three dummy variables are needed for this categorical variable. Let $x2$ represent Region 1, Northeast; $x3$ represent Region 2, Northcentral; and $x4$ represent Region 3, South.

```
CDI2 <- lm( Physicians ~ TotalPop + x2 + x3 + x4)
```





After fitting the model and running diagnostics, we see severe violations of linearity, normality and equal variance. Thus, we need to do transformations:

```
Transx <- powerTransform(TotalPop ~ 1, CDI)
summary(Transx)
```

```
## 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
```

```
CDI2.lm <- lm(Physicians ~ I(TotalPop^-0.5) + Region)
CDItrans <- with(CDI, data.frame(Physicians, I(TotalPop^-0.5), Region))
cdipower <- powerTransform(Physicians ~.,CDItrans )
summary(cdipower)
```

```
## 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
```

Similar to Part I, we perform powerTransform test to both predictor TotalPop, leaving out Region, and to the response Physicians. We decides to log transform Physicians as the transformation provides significant improvement to linearity and equal variance, and take TotalPop to the negative 0.5 power as the powerTransform test suggests.

Our transformed model is:

```
TransCDI <- lm(log(Physicians) ~ I(TotalPop^-0.5) + x2 + x3 + x4)
```

```
##                        Estimate  Std. Error     t value      Pr(>|t|)
## (Intercept)       9.121828e+00  0.09026571 101.0552917 7.771396e-297
## I(TotalPop^-0.5) -1.455684e+03 35.87128212 -40.5807757 2.036235e-147
## x2                9.893740e-02  0.07580894   1.3050887  1.925772e-01
## x3                9.369966e-03  0.07595727   0.1233584  9.018823e-01
## x4                8.573170e-02  0.07092505   1.2087649  2.274332e-01
```

Let β_0 be the intercept, β_1, β_2, β_3, and β_4 be the parameters of $x1$, $x2$, $x3$, and $x4$ respectively.

Region 1 (Northeast): $E(\log(\text{Physicians})) = \beta_0 + \beta_2 + \beta_1 x_{i1} =$ 9.1218 + 0.0989 - 1455.684x_{i1}

Region 2 (Northcentral): E(log(Physicians)) = β_0 + β_3 + $\beta_1 x_{i1}$ = 9.1218 + 0.0094 - 1455.684x_{i1}

Region 3 (South): E(log(Physicians)) = β_0 + β_4 + $\beta_1 x_{i1}$ = 9.1218 + 0.0857 - 1455.684x_{i1}

Region 4 (West): E(log(Physicians)) = β_0 + $\beta_1 x_{i1}$ = 9.1218 - 1455.684x_{i1}

This model is called a parallel regression model becasue they all have β_1 = -1455.684 as their slopes.

Our initial thought is that geographic regions do not have a significant effect on the number of physicians in a county since each region covers a wide range of different states. Without considering other demographic information such as income and age groups, region itself cannot say much about the estimated mean number of physicians.

We conduct test to see if Region can be removed from the model:

```r
red <- lm(log(Physicians) ~ I(TotalPop^-0.5))
full <- lm(log(Physicians) ~ I(TotalPop^-0.5) + Region)
anova(red,full)
```

```
## Analysis of Variance Table
##
## Model 1: log(Physicians) ~ I(TotalPop^-0.5)
## Model 2: log(Physicians) ~ I(TotalPop^-0.5) + Region
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    423 104.16
## 2    422 104.01  1   0.14645 0.5942 0.4412
```

Using an anova test with null hypothesis H_0: log(Physicians) ~ TotalPop$^{-0.5}$ (reduced model), and alternative hypothesis H_1: log(Physicians) ~ TotalPop$^{-0.5}$ + Region (full model). The p-value is 0.4412 > 0.04, we fail to reject the null hypothesis, so the reduced model (without Region) is preferred, i.e. the geographic region does NOT have a significant effect on the number of physicians, so we can remove it from now on.

We will build on the current model log(Physicians) ~ TotalPop$^{-0.5}$ by selecting relevant predictors from Pop65, Crimes, Bachelor, Poverty, and PersonalInc.

```r
library(leaps)
CDI3.lm <- lm(log(Physicians) ~ I(TotalPop^-0.5))
x2 <- Pop65
x3 <- Crimes
x4 <- Bachelor
x5 <- Poverty
x6 <- PersonalInc
```

```r
mod.0 <- lm(log(Physicians) ~ I(TotalPop^-0.5), data = CDI)
mod.full <- lm(log(Physicians) ~ I(TotalPop^-0.5) + x2 + x3 + x4 + x5 + x6, data = CDI)
forward <- step(mod.0, scope = list(lower = mod.0, upper = mod.full),
direction = "forward")
backward <- step(mod.full, scope = list(lower = mod.0, upper = mod.full),
direction = "backward")
```

Using forward and backward selection, we found the optimal model with the smallest AIC to be log(Physicians) ~ TotalPop$^{-0.5}$ + Pop65 + Bachelor + Poverty + PersonalInc.

And now let's try the best subsets regression method:

```r
mod.reg <- regsubsets(cbind(I(TotalPop^-0.5),x2,x3,x4,x5, x6), log(Physicians), data = CDI)
summary.reg <- summary(mod.reg)
names(summary.reg)
```

```
## [1] "which"  "rsq"    "rss"    "adjr2"  "cp"     "bic"    "outmat" "obj"
```

```
summary.reg$which
```

```
##   (Intercept)         x2    x3    x4    x5    x6
## 1        TRUE TRUE FALSE FALSE FALSE FALSE FALSE
## 2        TRUE TRUE FALSE FALSE  TRUE FALSE FALSE
## 3        TRUE TRUE FALSE FALSE  TRUE  TRUE FALSE
## 4        TRUE TRUE  TRUE FALSE  TRUE  TRUE FALSE
## 5        TRUE TRUE  TRUE FALSE  TRUE  TRUE  TRUE
## 6        TRUE TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
```

```
summary.reg$rsq
```

```
## [1] 0.8006995 0.8275250 0.8433396 0.8622266 0.8783611 0.8786095
```

```
summary.reg$adjr2
```

```
## [1] 0.8002284 0.8267076 0.8422233 0.8609144 0.8769096 0.8768671
```

```
summary.reg$cp
```

```
## [1] 265.277844 174.906165 122.449591  59.413621   5.855423   7.000000
```

```
summary.reg$bic
```

```
## [1] -673.3960 -728.7825 -763.6035 -812.1510 -859.0343 -853.8511
```

Again, using adjusted R^2, Mallow's Cp, and BIC as model selection criteria, give us the best model is log(Physicians) ~ TotalPop$^{-0.5}$ + Pop65 + Bachelor + Poverty + PersonalInc.

```
CDI3.lm <- lm(log(Physicians) ~ I(TotalPop^-0.5))
newmod <- lm(log(Physicians) ~ I(TotalPop^-0.5) + x2 + x4 + x5 + x6)
anova(CDI3.lm, newmod)
```

```
## Analysis of Variance Table
##
## Model 1: log(Physicians) ~ I(TotalPop^-0.5)
## Model 2: log(Physicians) ~ I(TotalPop^-0.5) + x2 + x4 + x5 + x6
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1    423 104.159
## 2    419  63.572  4    40.588 66.879 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

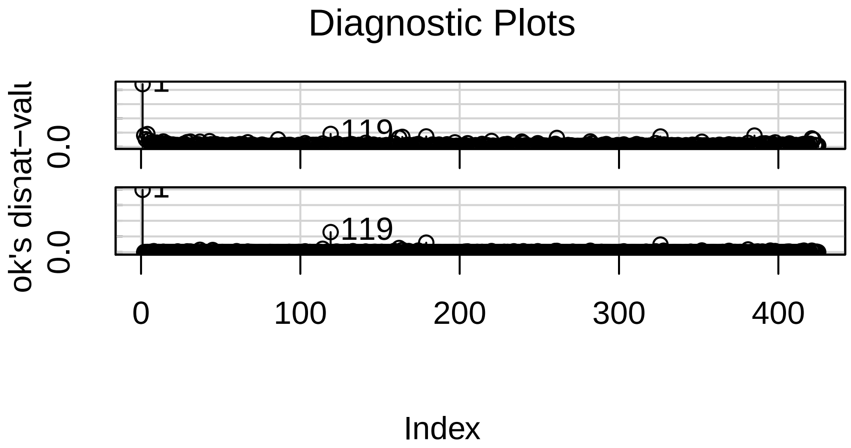
Using a partial F-test with H_0: log(Physicians) ~ TotalPop$^{-0.5}$, and H_1: log(Physicians) ~ TotalPop$^{-0.5}$ + Pop65 + Bachelor + Poverty + PersonalInc. From the anova table, we reject the null hypothesis, and conclude that the new model with relevant predictors added has a significant improvement comparing with the first model we started with.

Now, let's see if there are influential points:

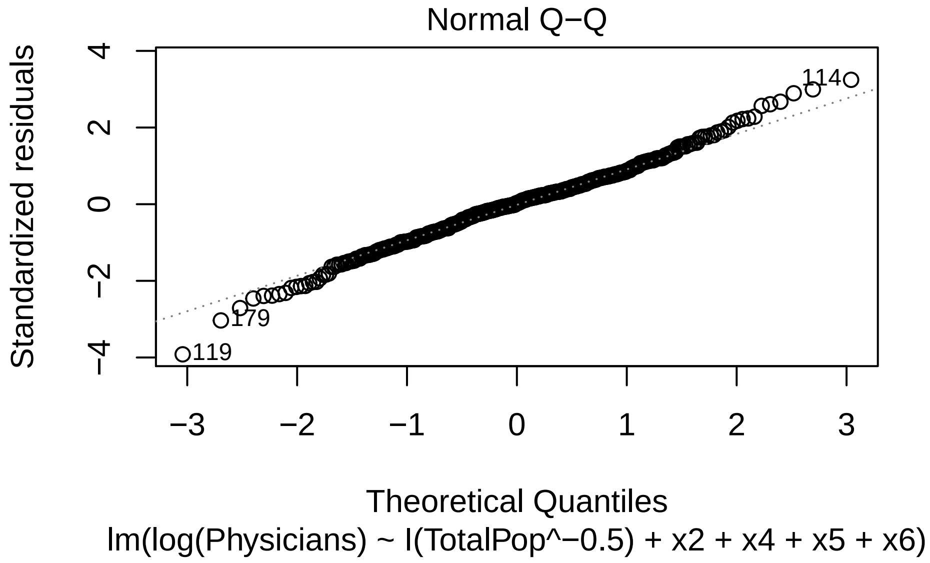
```
s.resid <- rstudent(newmod)
as = abs(s.resid)
which(as == max(as))
```

```
## 119
## 119
```

```
influenceIndexPlot(newmod, var = c('hat','Cook'),id=TRUE)
```



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



By finding the maximum studentized residual and looking at the cook's distance diagonistic plot above, 119 seems to be the only influential point while all other x_i's follow the linear fit quite well. This can also be ensured by the Q-Q plot of our model log(Physicians) ~ log(TotalPop) + Pop65 + Bachelor + Poverty + PersonalInc. As shown in plot, 119 is at the most lower left which is far away from the other x_i's.

CONCLUSION: In part I, we find that physicians lhas positive linear relationship with log of total population, and almost none with land area. There is almost no association between land area and per capita income which also parallels with our initial expectation. But, unlike what we expected initially, per capita income actually has almost no linear relationship with the number of physicians, before making any transformations.

However, after log and power transformations to both the response and the predictors, we have a much better linear model: log(Physicians) with log(TotalPop), log(LandArea), and $1/\sqrt{IncPerCap}$. We see a linear relationship between each predictor to the response physicians, and the estimated median of the number of physicians changes as each predictor increases by one unit.

In part II, through the investigation of total population and region on physician, we came to remove region from our model eventually as it made no significant impact to the model even after transformation. In general, our analysis should be reliable not only because we used scientific tools to determine and modify our statistical models, but the models we concluded also make sense intuitively: A more populated area would lead to higher demands for physicians; we can also see that a county with larger percentage of elderly people, and higher education level/ personal income can be contributing factors to the number of physicians. While total population always have impact to the response, poverty level can also be taken into consideration. However, the geographic location of a county — whether the county is located in the west or south or northeast or northcentral — has no significant effect to the number of physicians since it is the demographic information of a county that matters, rather than the geographic information.