PSTAT 126 Project

Physicians and US Counties

Kelly Wang, Qirong He, Angel Chen

A project about the regression of physicians on multiple variables

Department of Statistics & Applied Probability
University of California, Santa Barbara
United States
June 9, 2019

PSTAT 126 Project

Physicians and US Counties

Kelly Wang, Qirong He, Angel Chen

Abstract

Abstract: We are interested in investigating how to best estimate the number of physicians in a US county using various predictors. In the first part of our investigation, we wanted to know how we can improve our model if we only had **log(TotalPop)**, **LandArea**, and **IncPerCap** as predictors. By finding the suitable transformations, we included **log(TotalPop)**, **log(LandArea)**, and **IncPerCap** in our model with **log(Physicians)** as the response.

In the second part, we examined the model with only **TotalPop** and **Region**. To build on our model, we selected other relevant predictors. Finally, we got **log(TotalPop)**, **Region**, **Bachelor**, **Poverty**, and **Pop65** as predictors with **log(Physicians)** as the response.

Problem and Motivation

The United States is made up of over 3000 counties, which are administrative subdivisions of states. The government collects information on counties through the Census Bureau. The dataset we are working with, for example, contains data on the number of physicians in a specific county. With these data, we can then come up with the best model to estimate the number of physicians in a county based on other variables in the dataset. This is useful to us in real life because we may be interested in knowing if a bigger-sized county implies more physicians, or if more residents implies more physicians. Exactly what variable(s) affects how many physicians are in a county? Maybe physicians need to know these questions so they can find counties where physicians are in high demand. We will investigate these questions in our project.

Data

Our data set contains county demographic information (CDI) for 440 of the most populous counties in United States around 1990 to 1992. Since the missing data has been removed in our data set, we actually have 425 observations instead of 440. Inside the data set, we have country name, state abbreviation and 14 other variables for each observation. And we will use only 5 out of the 14 variables in our project. The 5 variables are **Physicians**, **TotalPop**, **LandArea**, **IncPerCap** and **Region**. Here are some short descriptions about our variables.

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

Variable Name	Description
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

We are interested which factors have associations with our target **Physicians**, which is the number of professionally active nonfederal physicians during 1990. Specifically, we will explore the relationship of **physicians** and some features, especially **TotalPop**, **LandArea**, **IncPerCap**, **Region**, **Pop65**, **Crimes**, **Bachelor**, **Poverty**, **PersonalInc**. If there exists relationship between **Physicians** and other features, is the relationship in positive proportionor negative proportion? And how strong is the relationship? How could we interpret their associations? If the association is not easy to explain, how could we apply some transformtions to make the association more linear and thus easier to interpret? We will answer all these questions during our analysis.

Regresson Methods

For the first part of our investigation, we did some exploratory analysis on our initial model then checked the diagnostic plots. We wanted to fix the diagnostic plots with transformation. To figure out if we needed to transform any variables, we used powerTransform(), the multivariate version of the Box Cox transformation. This is to find the optimal lambdas for the predictors. Next, we moved onto transforming the response by using Box Cox to get our final model, which had diagnostic plots that look way better than the ones we started with in our initial model. To make sure that our new model works, we conduct t-tests on each of its coefficients and their p-values tell us that they are all needed in the model. However, we were concerned if the variance was non-constant for log(TotalPop), so we used ncvTest() to find out that it was constant after all.

For the second part of our investigation, again, we did some exploratory analysis on our initial model then checked the diagnostic plots. They violated assumptions of a linear model, so we knew that we had to transform variables. powerTransform() and boxCox() showed us that transformations were needed for one of the predictors and the response. After checking that the diagnostic plots looked okay, we considered whether we needed one of the predictors or not, so we conducted an anova partial F-test. We still had other predictors to choose from, though, so the AIC forward, backward, and stepwise selection helped us choose the best ones to include. Finally, we looked for outliers and influential points using outlierTest() and influenceIndexPlot().

Regression Analysis, Results and Interpretation

Exploratory Analysis

To start off, in the first part of our investigation, we are interested in predicting the number of physicians with predictors log(total population), land area, and per capita income. We expect to see a positive linear relationship between the number of physicians and log(total population) because it would make sense for the number of physicians to increase as the total population increases.

We are not sure if a strong linear relationship exists between the number of physicians and land area because sometimes a county can be sparsely population (so there are fewer physicians on average), but have a large area. Also, sometimes a county can be heavily populated (so there are more physicians on average), but have a small area.

As for the last predictor, we expect to see a positive linear relationship between the response and per capita income because it would make sense for the number of physicians to increase as the county becomes wealthier.

Since we are examining a model with multiple predictors, we must also look at the effects the predictors will have on each other. For example, we expect to see a positive linear relationship between log(total population) and land area because it is often, but not always, the case that the larger areas imply larger populations. Hence, the relationship will most likely be a weak one.

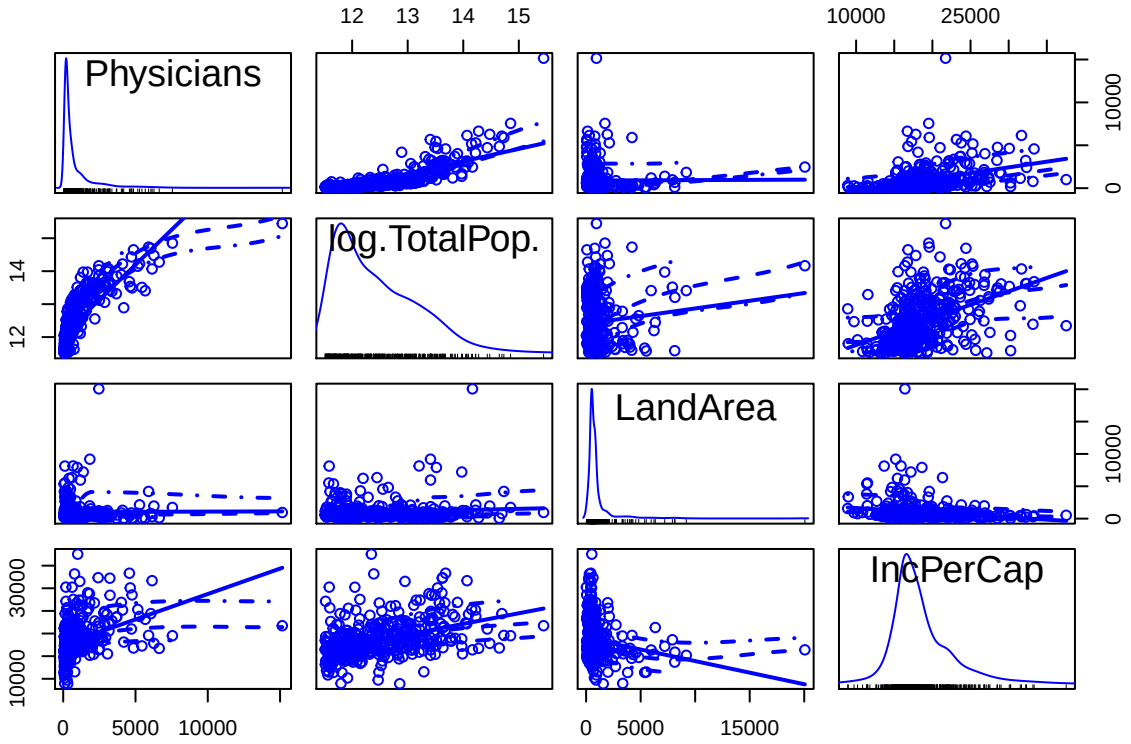
Our hunch tells us to expect to see another positive linear relationship between log(total population) and per capita income because more people implies that per capita income, a mean value, will increase as well.

Finally, we expect to see a negative linear relationship between land area and per capita income because there are counties with a large area but few people. That means that its per capita income will not be as high as other counties with a smaller area but more people. So, as land area increases, per capita income may decrease. This, however, may not always be the case, so the association between these two predictors will most likely be a weak one.

To check our assumptions about marginal relationships, we plot the scatterplot matrix for our model. Let's load all the needed libraries.

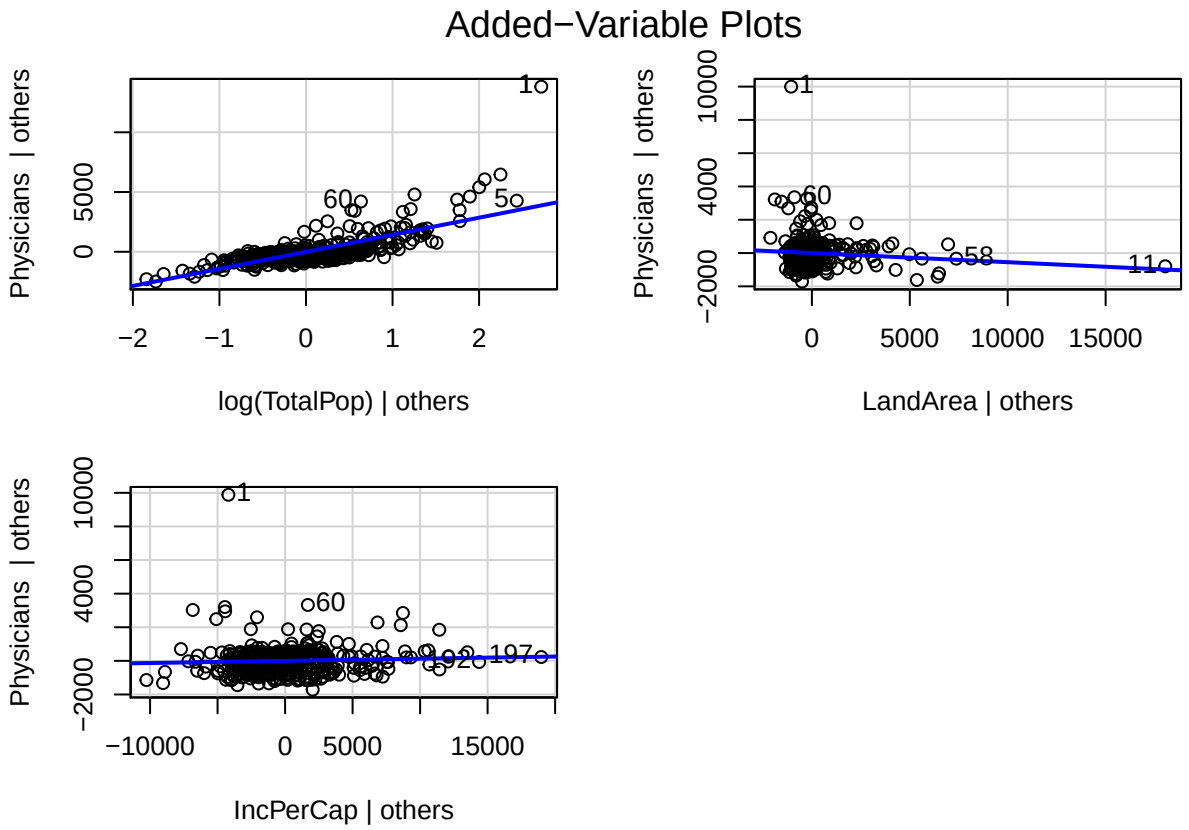
```
library(alr4) #to get the scatterplotMatrix function
library(outliers)
library(dplyr)
library(ggplot2)
```

```
CDI <- readRDS("~/rmd files/CDI.rds")
attach(CDI)
CDI_I.lm <- lm(Physicians~log(TotalPop)+LandArea+IncPerCap)
scatterplotMatrix(~Physicians+log(TotalPop)+LandArea+IncPerCap)
```



Our assumptions about marginal relationships seem to hold well. However, the scatterplot matrix does not tell us the full story because some predictors may be correlated. We need to consider how useful a predictor will be if the other predictors are already included in the model. To check this, we look at the added-variable plots.

```
avPlots(CDI_I.lm)
```



In the marginal plot of physicians versus per captia income, we see a positive linear relationship. Nonetheless, in their added-variable plot, we only see a horizontal line. This means that per capita income is not useful to include in our model if it already has log(total population) and land area. We will keep this in mind as we move on to analyze our model in more detail.

To examine our model further, we look at its summary output.

```
summary(CDI_I.lm)
```

```
##
## Call:
## lm(formula = Physicians ~ log(TotalPop) + LandArea + IncPerCap)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1739.9  -495.4    -5.4   375.4  9938.9
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -1.706e+04  7.060e+02 -24.165   <2e-16 ***
## log(TotalPop)  1.427e+03  6.293e+01  22.683   <2e-16 ***
## LandArea      -5.488e-02  2.865e-02  -1.916   0.0561 .
## IncPerCap      1.285e-02  1.190e-02   1.079   0.2811
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 859.7 on 421 degrees of freedom
```

```
## Multiple R-squared:  0.6202, Adjusted R-squared:  0.6175
## F-statistic: 229.2 on 3 and 421 DF,  p-value: < 2.2e-16
```

From this output, we can see that our fitted model is:

$\widehat{Physicians} = (-17060) + (1427)log(TotalPop) + (-0.05488)LandArea + (0.01285)IncPerCap$

The intercept is −17060, which is the expected number of physicians when all the predictors equal zero.

The coefficient for log(TotalPop) is 1427. Then 1427log(1+p) is the expected change in the number of physicians when TotalPop changes by 100p%, assuming that all the other predictors are held constant.

The coefficient for LandArea is −0.05488. This means that if all the other predictors are held constant, a 1 square mile increase in land area results in an estimated average decrease of −0.05488 physicians.

The coefficient for IncPerCap is 0.01285. This means that if all the other predictors are held constant, a $1 increase in IncPerCap results in an estimated average increase of 0.01285 physicians.

The summary output gives us the goodness-of-fit measure R^2, too. In this case, $R^2 = 0.6202$, which means that 62.02% of the variability in the number of physicians can be explained by all the predictors. Although R^2 is extremely useful in simple linear regression, we must be careful in using R^2 in multiple linear regression; when an additional predictor is added to the model, R^2 almost always goes up.

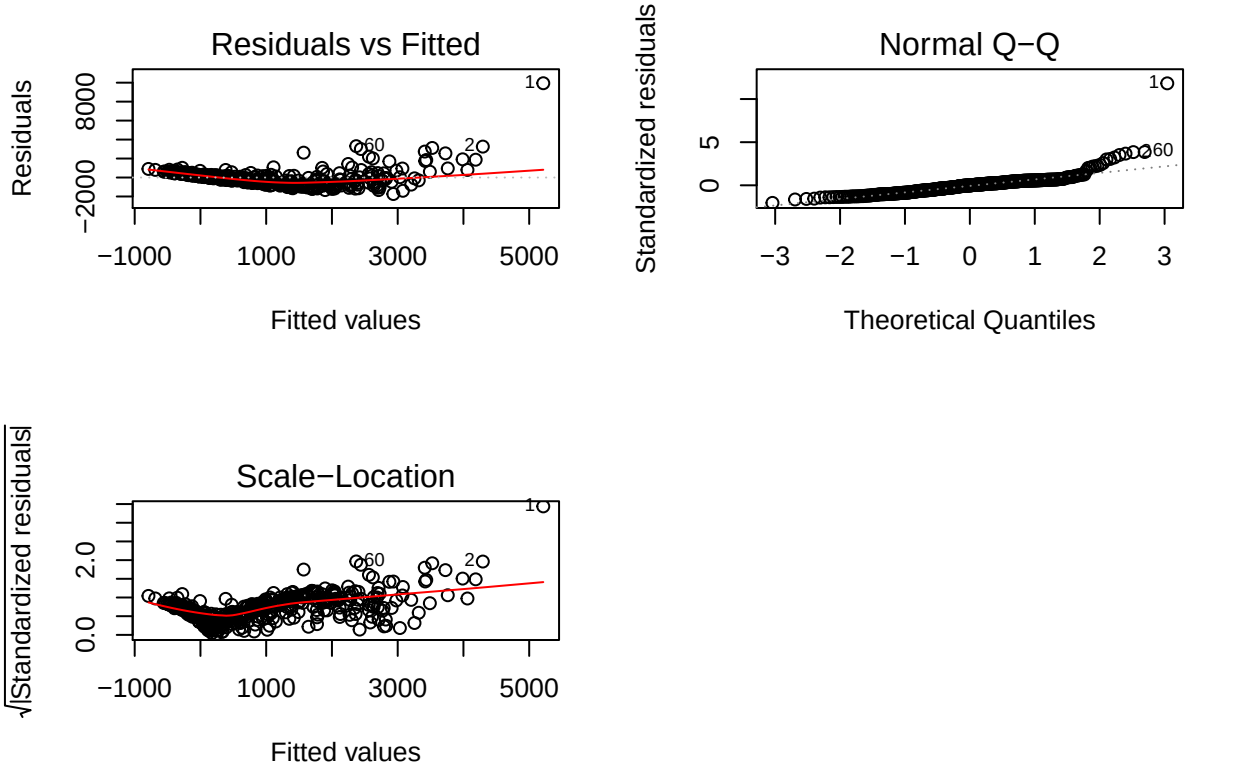
```
head(CDI)
```

```
##      County State LandArea TotalPop Pop18 Pop65 Physicians  Beds Crimes
## 2      Cook    IL      946  5105067  29.2  12.4      15153 21550 436936
## 3    Harris    TX     1729  2818199  31.3   7.1       7553 12449 253526
## 4 San_Diego    CA     4205  2498016  33.5  10.9       5905  6179 173821
## 5    Orange    CA      790  2410556  32.6   9.2       6062  6369 144524
## 6     Kings    NY       71  2300664  28.3  12.4       4861  8942 680966
## 9      Dade    FL     1945  1937094  27.1  13.9       6274  8840 244725
##   HSGrad Bachelor Poverty Unemp IncPerCap PersonalInc Region
## 2   73.4     22.8    11.1   7.2     21729      110928      2
## 3   74.9     25.4    12.5   5.7     19517       55003      3
## 4   81.9     25.3     8.1   6.1     19588       48931      4
## 5   81.2     27.8     5.2   4.8     24400       58818      4
## 6   63.7     16.6    19.5   9.5     16803       38658      1
## 9   65.0     18.8    14.2   8.7     17823       34525      3
```

Diagnostic Plots

From our previous scatterplot matrix, we can see log(TotalPop) and IncPerCap) have slightly linear relationship. In the plot between LandArea and Physicians, apart from one separated point in the right, the clumping of points in the lower left of the plot hides any visual information, the values of LandArea range over more than one order of magnitude, this suggests that we might want find some proper transform for LandArea.

```
CDI_I.lm <- lm(Physicians~log(TotalPop)+LandArea+IncPerCap)
# do the diagnostic checks
par(mfrow = c(2,2))
plot(CDI_I.lm,which = c(1,2,3))
```



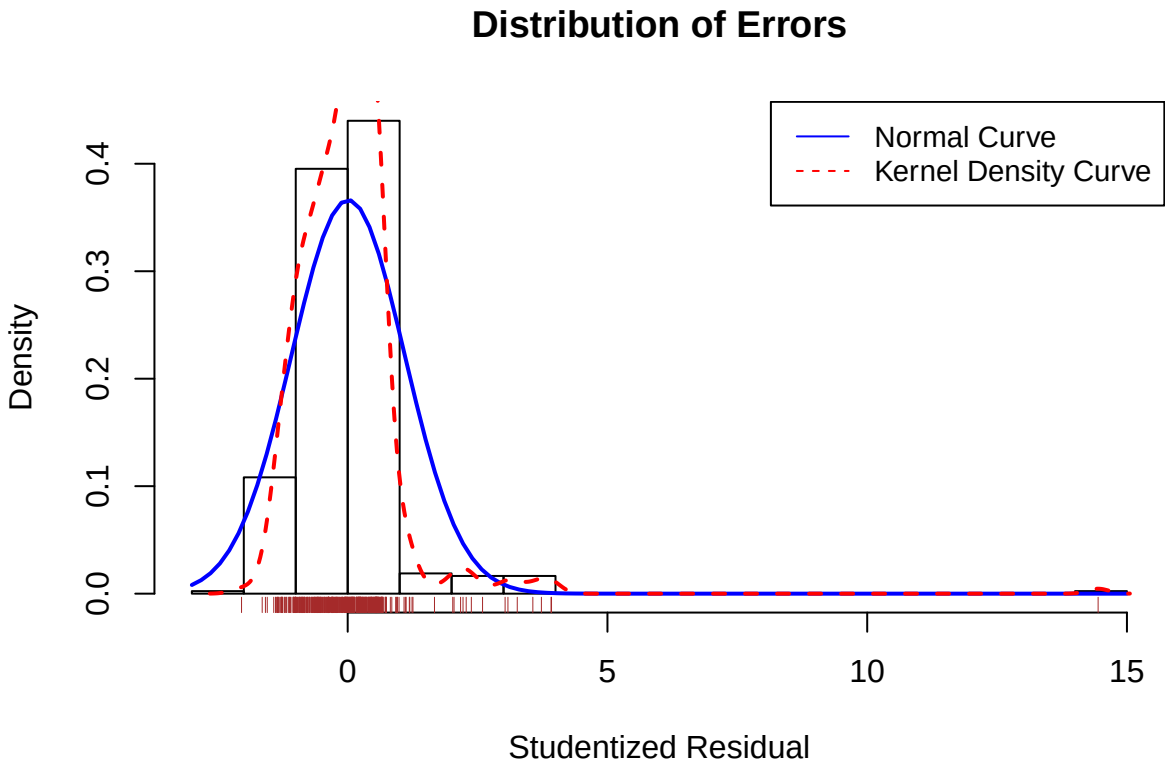
From the plot above we can do the diagnostic tests.

- Linearity—From the Residuals vs. Fitted graph (upper left), the residuals seems to have a linear pattern and don't bounce randomly around the 0-line. We can see that there is evidence of a little curved relationship, which suggests that we may want to add a nonlinear term to the regression. This suggest that the assumption that the relationship is linear is not reasonable.
- Normality—From the Normal Q-Q plot (upper right), it is severely right-skewed, so it doesn't meet the normality assumption.

We can make a histgram to verify that.

```
residplot <- function(CDI_I.lm, nbreaks=20) {
  z <- rstudent(CDI_I.lm)
  hist(z, breaks=nbreaks, freq=FALSE,
       xlab="Studentized Residual",
       main="Distribution of Errors")
```

```
      rug(jitter(z), col="brown")
      curve(dnorm(x, mean=mean(z), sd=sd(z)),
            add=TRUE, col="blue", lwd=2)
      lines(density(z)$x, density(z)$y,
            col="red", lwd=2, lty=2)
      legend("topright",
             legend = c( "Normal Curve", "Kernel Density Curve"),
  }
residplot(CDI_I.lm)
```



As we can see, the errors don't follow a normal distribution quite well, with the exception of a large outlier. I found there is evdience of right skew of a distribution from a histogram and density plot, because compare to the middle the right part of the hisgram is so small. Which meets the analysis we made from the Q-Q Plot.

- Constant Variance—If we've met the constant variance assumption, the points in the Scale-Location graph (bottom left) should be a random band around a horizontal line. However, the points seems to have a curved pattern, so we seem to violate from this assumption.
- Independence—We can't tell if the dependent variable values are independent from these plots. We have to use our understanding of how the data was collected.

In conclusion, all of the diagnostic assumptions do not hold for this model.

Transformations

Before considering transformations for the response Physicians, we will choose transformations for the predictors. We can use a multivariate version of the Box-Cox method which will try to choose power transformations so that the predictors have approximately a multivariate normal distribution.

```
pt = powerTransform(cbind(LandArea,IncPerCap)~1,CDI)
summary(pt)
```

```
## 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
```

The columns labeled "Wald Lower Bound" and "Wald Upper Bound" are the boundaries of 95% confidence intervals for the maximum likelihood power estimates. Pwr of LandArea is 0 and we can see that intervals of LandArea is [-0.0799,0.0490] contains 0, so we do the log transformation for LandArea.

What about IncPerCap?

```
summary(powerTransform(IncPerCap))
```

```
## bcPower Transformation to Normality
##           Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## IncPerCap    -0.369        -0.5      -0.6776      -0.0605
##
## Likelihood ratio test that transformation parameter is equal to 0
##  (log transformation)
##                              LRT df     pval
## LR test, lambda = (0) 5.42254  1 0.019878
##
## Likelihood ratio test that no transformation is needed
##                              LRT df       pval
## LR test, lambda = (1) 72.10332  1 < 2.22e-16
```

The likelihood ratio tests inidicate that using log transformations for IncPerCap is not appropriate, neither should we use no transformations.

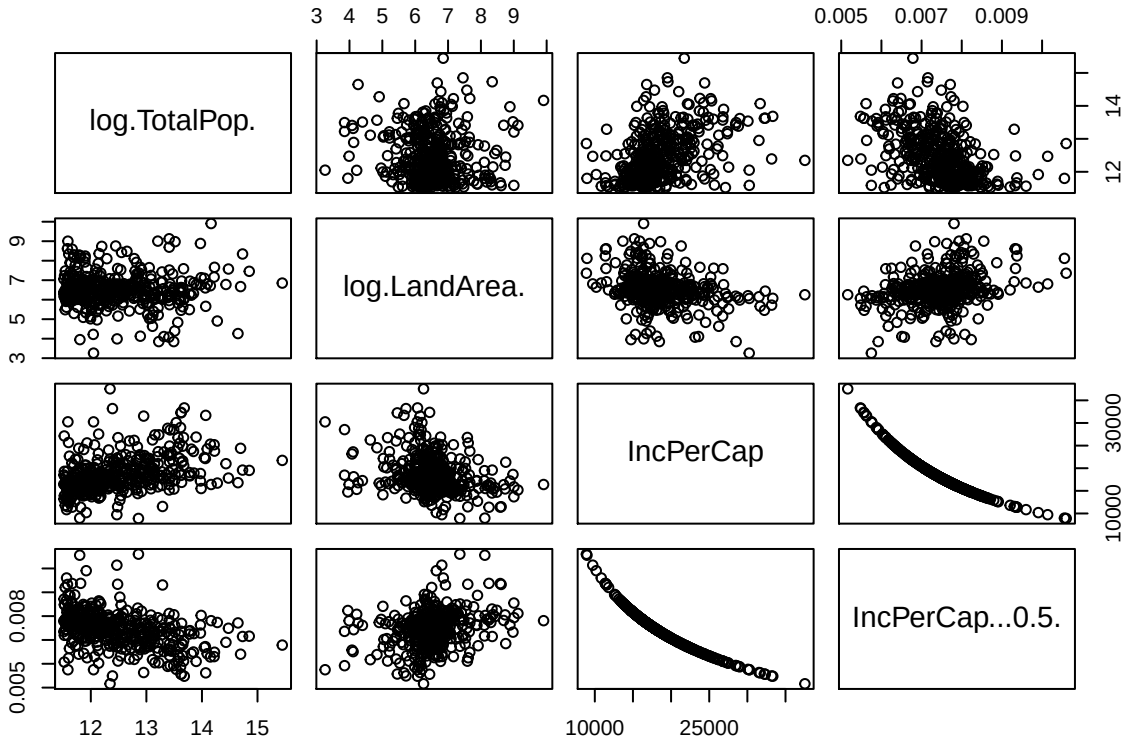
As the result from the powerTransform, it indicates Pwr = -0.5, I just add the power of -0.5 for IncPerCap in the model.

```
#New fitted model:
CDI_I.lm2<-lm(Physicians~log(TotalPop)+log(LandArea)+ IncPerCap+I(IncPerCap^(-0.5)))
testTransform(pt, lambda = c(0, -0.5))
```

```
##                                  LRT df    pval
## LR test, lambda = (0 -0.5) 0.8748981  2 0.64568
```

The p-value for this new model is 0.64568 which is very large, the hypothesis that $\lambda = -0.5$ can't be rejected, so there's no strong evidence that a transformation is needed in this case.

```
CDI_trsf = with(CDI, data.frame(log(TotalPop),log(LandArea),IncPerCap,I(IncPerCap^(-0.5))))
pairs(CDI_trsf)
```



From the pairplot, we can see that there is no obvious relationship between any two predictors, so our transformation is reasonable.

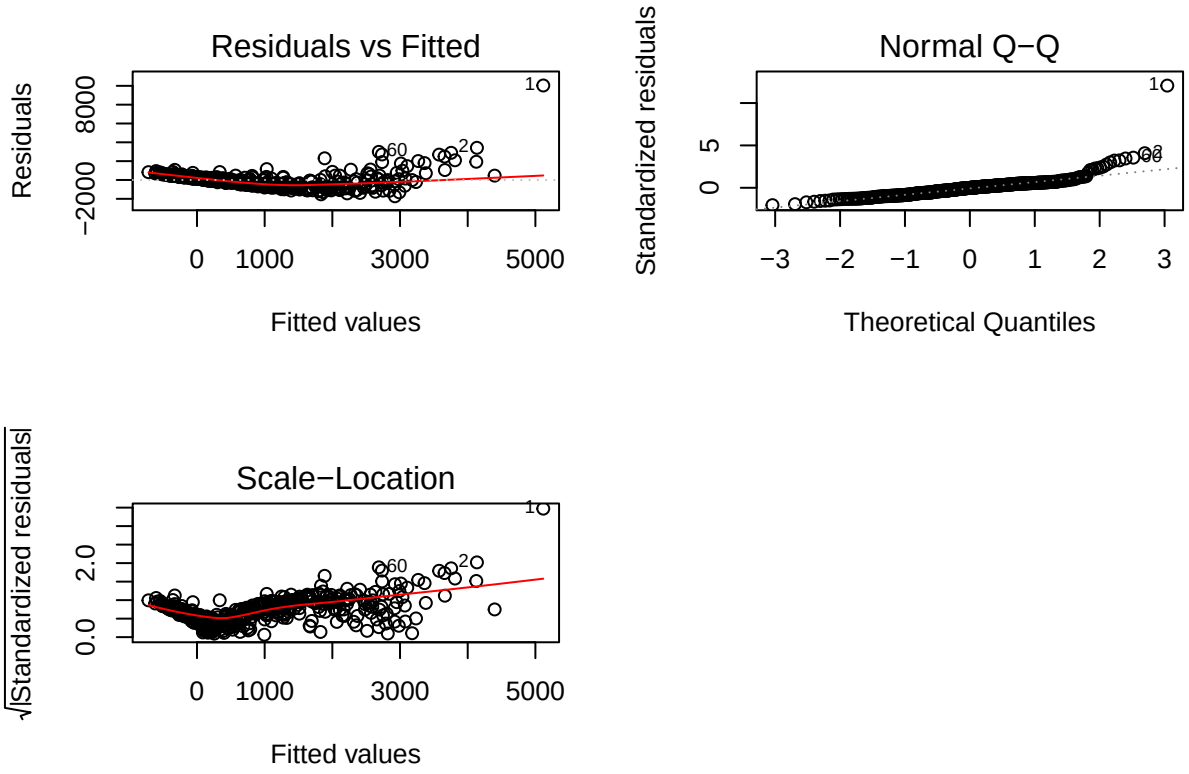
```
summary(CDI_I.lm2)
```

```
##
## Call:
## lm(formula = Physicians ~ log(TotalPop) + log(LandArea) + IncPerCap +
##     I(IncPerCap^(-0.5)))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1730.6  -495.1    -9.5   363.6 10036.3
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -2.033e+04  2.398e+03  -8.479 3.91e-16 ***
## log(TotalPop)        1.436e+03  6.165e+01  23.297  < 2e-16 ***
## log(LandArea)       -1.703e+02  5.130e+01  -3.320 0.000978 ***
## IncPerCap            7.883e-02  3.904e-02   2.019 0.044115 *
## I(IncPerCap^(-0.5))  4.013e+05  2.079e+05   1.930 0.054292 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 850.7 on 420 degrees of freedom
## Multiple R-squared:  0.629,  Adjusted R-squared:  0.6255
## F-statistic:   178 on 4 and 420 DF,  p-value: < 2.2e-16
```

From the summary we can see that the p-value of $IncPerCap^{-0.5}$ is $0.054292 > 0.05$, so the model still seems to not work well.

Do diagnostic test for our current model:

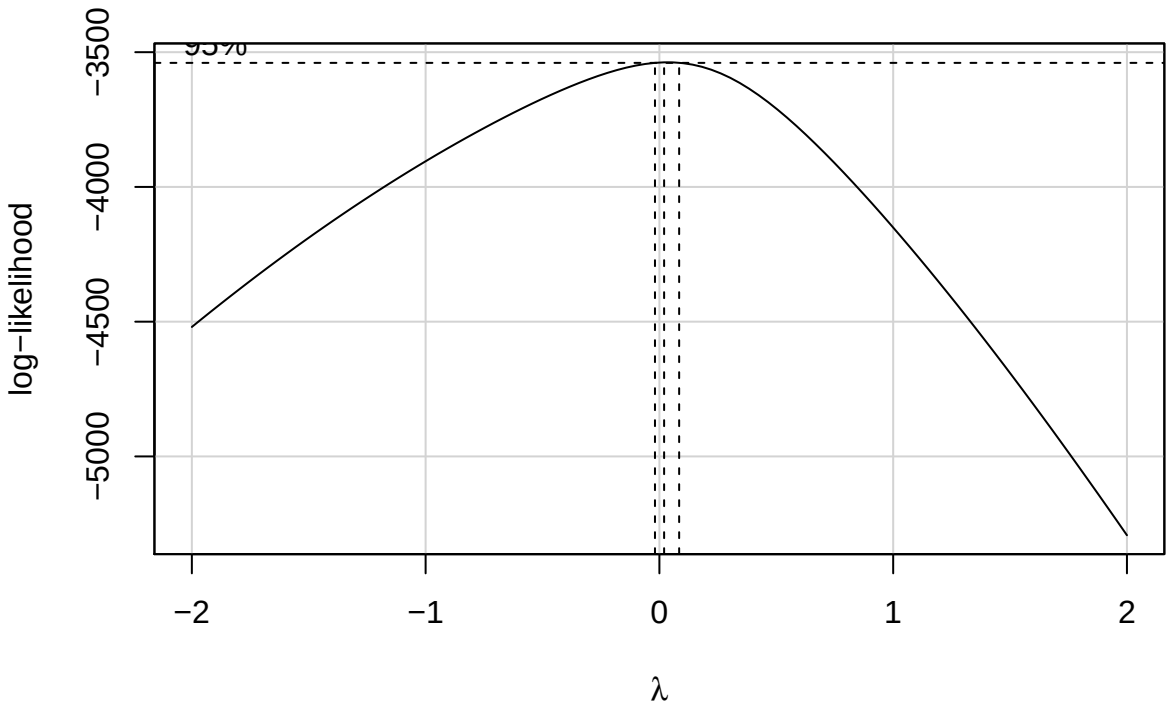
```
par(mfrow = c(2,2))
plot(CDI_I.lm2,which = c(1,2,3))
```



We do the diagnostic checks, there is no plot seems to meet the assumptions.

We don't gain a perfect result after transforming the predictors. So, we decide to do the transform for the response.

```
bc <- boxCox(CDI_I.lm2, data = CDI)
```




```
bc$x[which.max(bc$y)]
```

```
## [1] 0.02020202
```

The result is 0.02 which is very close to 0, so it suggests that we should take $\lambda = 0$ and transform the predictors with log-transformation.

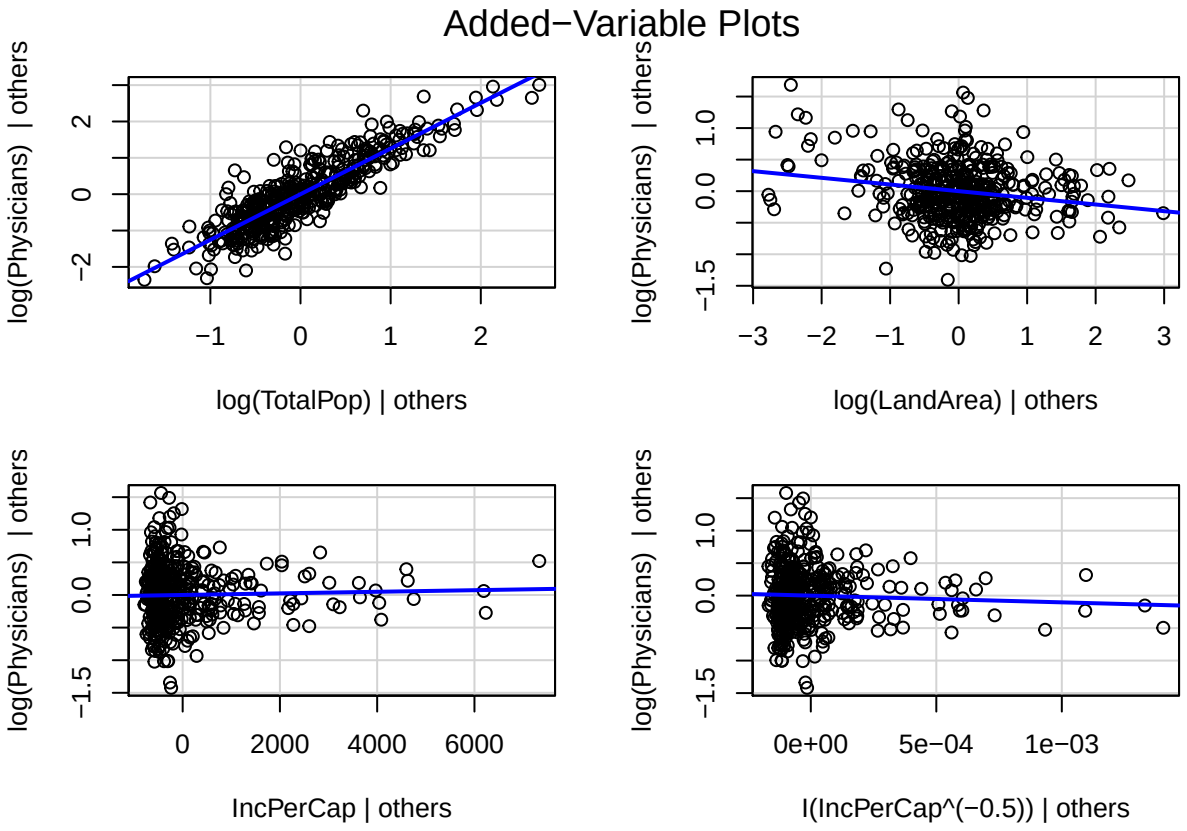
And now our new model is goning to be *log(Physicians) = log(TotalPop) + log(LandArea) + IncPerCap*

```
CDI_new.redu<-lm(log(Physicians)~log(TotalPop)+log(LandArea)+ IncPerCap)
CDI_new.full<-lm(log(Physicians)~log(TotalPop)+log(LandArea)+IncPerCap+I(IncPerCap^(-0.5)))
summary(CDI_new.full)
```

```
##
## Call:
## lm(formula = log(Physicians) ~ log(TotalPop) + log(LandArea) +
##     IncPerCap + I(IncPerCap^(-0.5)))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.42132 -0.29612 -0.01559  0.26925  1.56984
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -8.282e+00  1.262e+00  -6.564 1.55e-10 ***
## log(TotalPop)        1.255e+00  3.243e-02  38.699  < 2e-16 ***
## log(LandArea)       -1.057e-01  2.699e-02  -3.914 0.000106 ***
## IncPerCap            1.222e-05  2.054e-05   0.595 0.552257
```

```
## I(IncPerCap^(-0.5)) -1.035e+02  1.094e+02  -0.946 0.344646
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4476 on 420 degrees of freedom
## Multiple R-squared:  0.839,  Adjusted R-squared:  0.8375
## F-statistic: 547.3 on 4 and 420 DF,  p-value: < 2.2e-16
```

```
avPlots(CDI_new.full, id=FALSE)
```



After we transform the response, now it seems like IncPerCap and its -0.5 power is not significant. To keep the principle of simplification, we try to compare the full model with the $InCerCap^{-0.5}$ and the reduced model without $InCerCap^{-0.5}$.

conduct F-test to compare these two models

```
anova(CDI_new.redu,CDI_new.full)
```

```
## Analysis of Variance Table
##
## Model 1: log(Physicians) ~ log(TotalPop) + log(LandArea) + IncPerCap
## Model 2: log(Physicians) ~ log(TotalPop) + log(LandArea) + IncPerCap +
##     I(IncPerCap^(-0.5))
##   Res.Df   RSS Df Sum of Sq      F Pr(>F)
## 1    421 84.31
## 2    420 84.13  1    0.1793 0.8951 0.3446
```

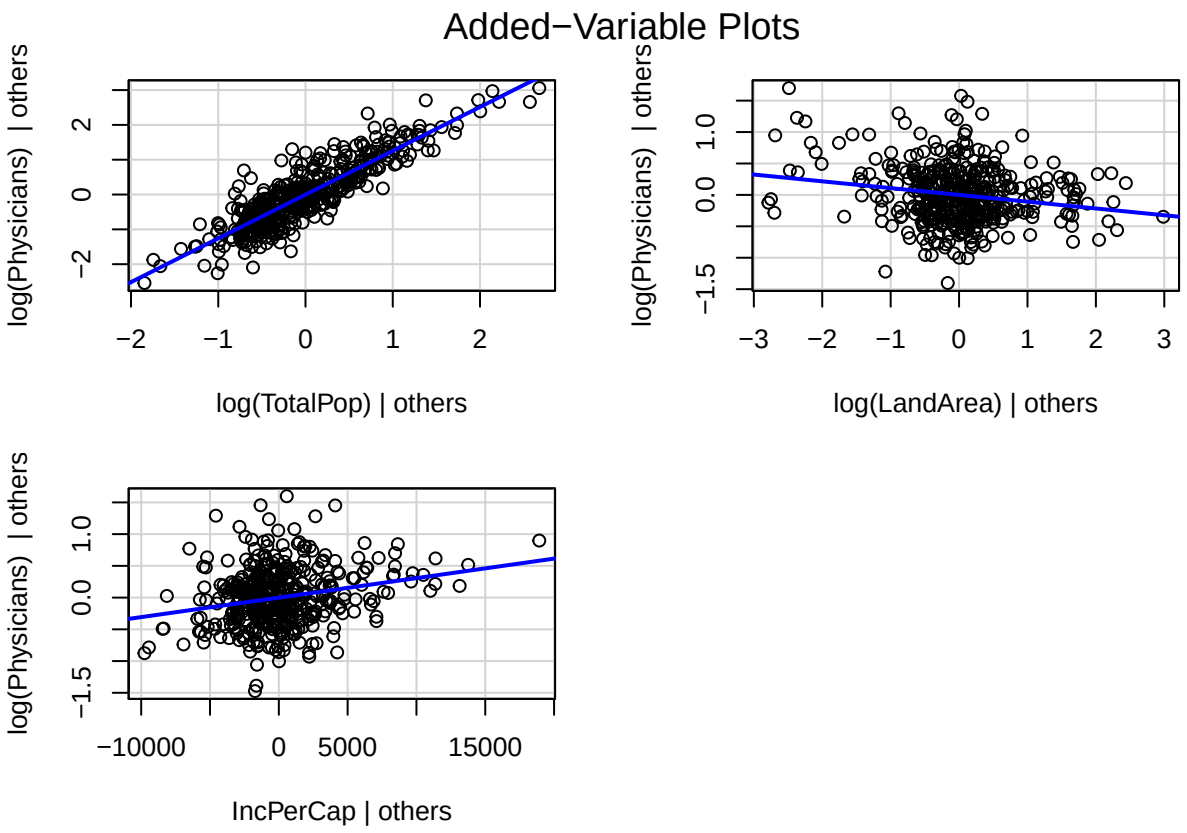
The p-value is $0.3446 > 0.05$, we fail to reject the null hypothesis, so we assume the reduced model(i.e.the

model without $IncPerCap^{-0.5}$) is better.

```
summary(CDI_new.redu)
```

```
##
## Call:
## lm(formula = log(Physicians) ~ log(TotalPop) + log(LandArea) +
##     IncPerCap)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.41980 -0.29642 -0.02003  0.27359  1.58001
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -9.414e+00  4.003e-01 -23.517  < 2e-16 ***
## log(TotalPop)  1.258e+00  3.232e-02  38.914  < 2e-16 ***
## log(LandArea) -1.080e-01  2.688e-02  -4.016 7.00e-05 ***
## IncPerCap      3.072e-05  6.291e-06   4.883 1.49e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4475 on 421 degrees of freedom
## Multiple R-squared:  0.8387, Adjusted R-squared:  0.8375
## F-statistic: 729.6 on 3 and 421 DF,  p-value: < 2.2e-16
```

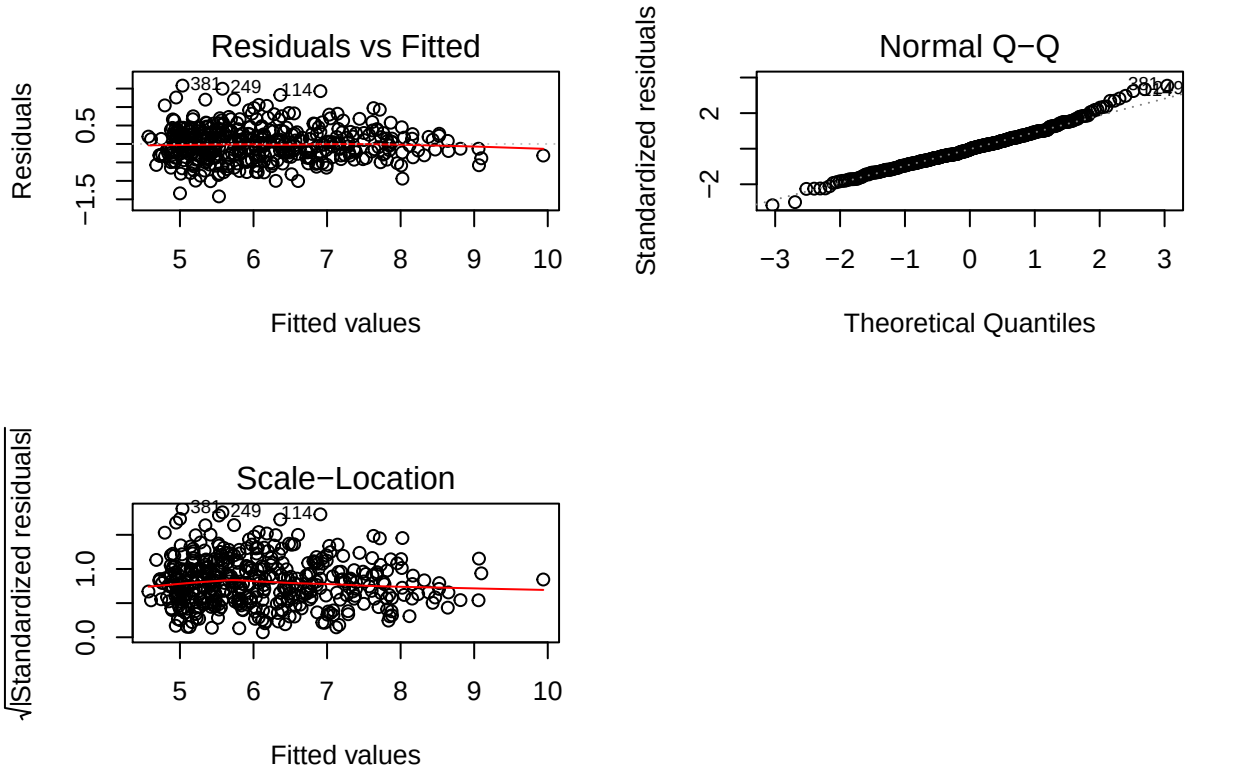
```
avPlots(CDI_new.redu, id=FALSE)
```



By looking at the added variable plot and summary tables, all of the predictors seem to be important.

Now, we do the diagnostic checks

```
par(mfrow = c(2,2))
plot(CDI_new.redu,which= c(1,2,3))
```



For linearity,from the Residuals vs. Fitted graph, the residuals "bounce randomly" around the 0 line. This suggests that the assumption that the relationship is linear is reasonable.

For normality, from the Q-Q Plot we can see that the relationship between the theoretical percentiles and the sample percentiles is approximately linear. Therefore, the normal probability plot of the residuals suggests that the error terms are indeed normally distributed.

For constant variance, the residuals roughly form a "horizontal band" around the 0 line in the scale-location plot. However, there are more points on the right side than the left. We will check for constant variance later.

This new model perform very well and we'll use this model for the remainder of Part I.

The confidence interval:

```
confint(CDI_new.redu,level = 0.95)
```

```
##                       2.5 %        97.5 %
## (Intercept)   -1.020117e+01 -8.627415e+00
## log(TotalPop)  1.194152e+00  1.321205e+00
## log(LandArea) -1.607933e-01 -5.512559e-02
## IncPerCap      1.835277e-05  4.308287e-05
```

Interpretation: The results suggest that :

- We can be 95% confident that the interval $[1.12, 1.32]$ contains the logarithm of true value of estimated 1990 population.
- We can be 95% confident that the interval $[-0.16, -0.055]$ contains the logarithm of true value of Land Area(square mile).
- We can be 95% confident that the interval $[1.84 \times 10^{-5}, 4.3 \times 10^{-5}]$ contains the true value of Per capita income of 1990 CDI population (dollars).

Additionally, no confidence interval contains 0, we can conclude that a change in every variable have influence to response, holding the other variables constant. But our faith in these results is only as strong as the evidence we have that our data satisfies the statistical assumptions underlying the model.

Perform test for the existence of a linear relationship between the predictors and response at $\alpha = 0.01$

```
summary(CDI_new.redu)
```

```
##
## Call:
## lm(formula = log(Physicians) ~ log(TotalPop) + log(LandArea) +
##     IncPerCap)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.41980 -0.29642 -0.02003  0.27359  1.58001
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -9.414e+00  4.003e-01 -23.517  < 2e-16 ***
## log(TotalPop)  1.258e+00  3.232e-02  38.914  < 2e-16 ***
## log(LandArea) -1.080e-01  2.688e-02  -4.016 7.00e-05 ***
## IncPerCap      3.072e-05  6.291e-06   4.883 1.49e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4475 on 421 degrees of freedom
## Multiple R-squared:  0.8387, Adjusted R-squared:  0.8375
## F-statistic: 729.6 on 3 and 421 DF,  p-value: < 2.2e-16
```

The null hypothesis and alternative for each test is:

$H_0 : \beta_1 = 0$ vs $\beta_1 \neq 0$

$H_0 : \beta_2 = 0$ vs $\beta_2 \neq 0$

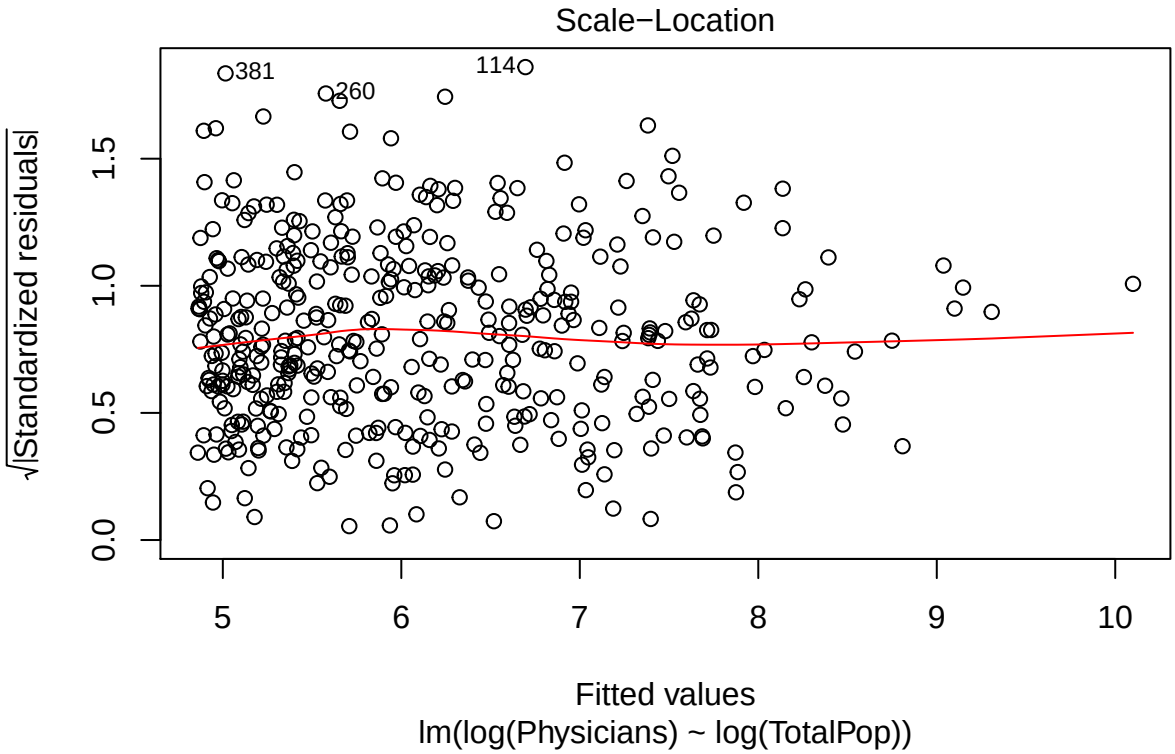
$H_0 : \beta_3 = 0$ vs $\beta_3 \neq 0$

In fact, the R summary output above does tests all of the null hypotheses. From the summary we can see that the p-values for each of the predictors log(TotalPop), log(LandArea) and IncPerCap are less than 2×10^{-16}, 7×10^{-5} and 1.49×10^{-6} respectively. Coefficients for all of log(TotalPop), log(LandArea) and IncPerCap are significant at $\alpha = 0.01$.

From the scale-location plot, we notice that the variance seems to decrease. We will explore this further by extracting only **log(TotalPop)** since it has largest estimated coefficient.

Scale-location Plot

```
CDI_I.lm5 <- lm(log(Physicians)~log(TotalPop))
plot(CDI_I.lm5, which=3)
```



From the scale-location residual plot, we notice that there are more points on the left area than on the right area. Besides, it seems that variance decreases with **log(TotalPop)** but we are not sure about that. To conclude our guess is correct or not, we will perform a test for constant variance.

Test For Constant Variance

H_0: Constant variance holds
H_1: Non-constant variance holds

```
ncvTest(CDI_I.lm5)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.2953407, Df = 1, p = 0.58682
```

Since p-value = 0.58682 which is greater than usual test level $\alpha = 0.05$, we fail to reject the hull hypothesis. Therefore, we conclude that the variance is actually constant.
Since our model is fitted well now, we will summerize our analysis and some findings here.

Summary

In our original model, we notice that **LandArea** is not significant since its p-value is much greater than usual test level $\alpha = 0.05$. **IncPerCap** has value around 0.05, which means that it is not very significant but useful than **LandArea**. **log(TotalPop)** has smallest p-value and thus we could conclude that the estimated coefficient for **log(TotalPop)** is definitely not 0. We can get same conclusion from the added variable plots. After controlling the effects of other variables, **log(TotalPop)** has a strong positive linear relationship with **Physicians**. Besides, **LandArea** has a weak negative linear relationship with **Physicians** and **IncPerCap** has very weak relationship with **Physicians** since the linear line is almost horizontal. And is this model, R^2 = 61.75% of variability in response **Physicians** is explained by a linear relationship with all predictors.

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

However, after checking the diagnostics, we find that our original model does not satisfy any of the linear regression assumptions. Therefore, we need some transformations for our original model. After investigate possible transformations, we apply logarithms on our reponse **Physicians** and predictor **LandArea** and refit our model as following. And this time, $R^2 = 83.75\%$ of variability in **log(Physicians)** is explained by by a linear relationship with all predictors.. And higher R^2 here means that our new model is better than our original model. And our 95% confident intervals for each coefficients in this new model also confirm the existence of a linear relationship between the predictors and response.

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

Though our new model is much better than before, we find that the variance looks like non-constant with **log(TotalPop)** from the scale-location residual plot. However, after performing a test, we conclude that the variance is actually constant.

After analyzing our model, we find that **TotalPop** has a positive relationship with **Physicians** and **LandArea** has a negative relationship with **Physicians**. Though **IncPerCap** also has a positive relationship with **Physicians**, the estimated coefficient is very close to 0 and we conclude that **IncPerCap** almost has no effect on **Physicians**. Specifically, if **TotalPop** changes by 100p% with all other predictors constant, $100[(1+p)^{1.258} - 1]$ percentage change will apply on expected **Physicians**; If **log(LandArea)** changes by 100p% with all other predictors constant, $100[(1+p)^{-0.1080} - 1]$ percentage change will apply on expected **Physicians**

It is very interesting that the **Physicians** almost has no relationship with **IncPerCap**. Before analyzing the data, we thought **IncPerCap** will have a positive relationship with **Physicians**, since more money usually stands for better education and better education is likely to produce more physicians. However, the data told us that was not what really happened. Our speculation that the more money producing better education does not hold. In order to be a physicians, you have to have good education. But not everyone invest their money to their or their children's education. More money could provide better privacy school and expensive books, but it could also provide more video games, more spoiling and more temptations. With these two aspects balancing with each other, **IncPerCap** could probably turn to be useless in our model in the end.

From our model above, we already know that **TotalPop** has a strong association with our target **Physicians**. In the following analysis, we will study the linear relationship of **Physicians** with **TotalPop** and **Region**.. We will start with fitting our new model.

$$Physicians \sim TotalPop + Region$$

```
CDI_II.lm <- lm(Physicians~factor(Region)+TotalPop+factor(Region):TotalPop)
anova(CDI_II.lm)
```

```
## Analysis of Variance Table
##
## Response: Physicians
##                            Df    Sum Sq   Mean Sq   F value    Pr(>F)
## factor(Region)              3   5865657   1955219    6.2005 0.0003962 ***
## TotalPop                    1 676386442 676386442 2144.9892 < 2.2e-16 ***
## factor(Region):TotalPop     3   5473352   1824451    5.7858 0.0006980 ***
## Residuals                 417 131493969    315333
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Fit Model

```
CDI_II.lm <- lm(Physicians~TotalPop+factor(Region))
summary(CDI_II.lm)
```

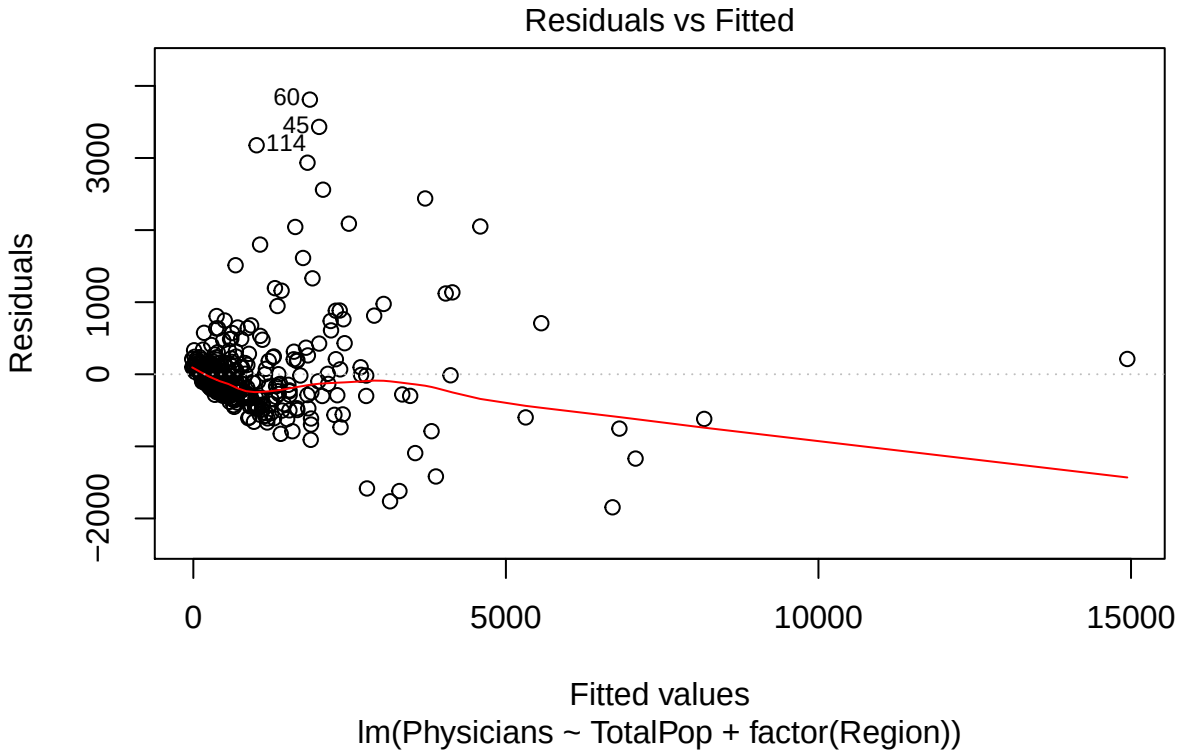
```
##
## Call:
## lm(formula = Physicians ~ TotalPop + factor(Region))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1845.2  -215.1   -67.5    96.0  3809.8
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -9.982e+01  6.240e+01  -1.600   0.1104
## TotalPop         2.958e-03  6.496e-05  45.542   <2e-16 ***
## factor(Region)2 -6.149e+01  8.011e+01  -0.768   0.4432
## factor(Region)3 -6.495e+01  7.422e+01  -0.875   0.3820
## 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
```

Now we will check diagnostics to see whether whether our model satisfies three linear assumptions or not.

Check Diagnostics

Assumption1 - Linearity: Residual vs Fitted plot

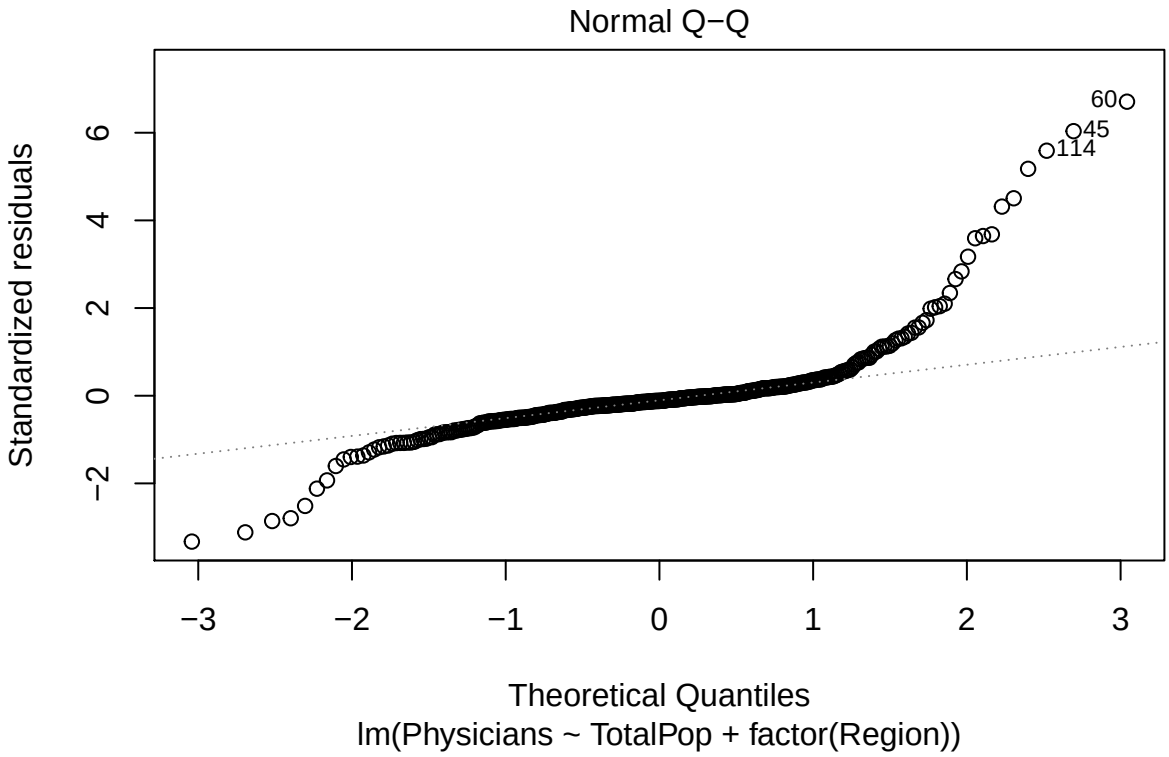
```
plot(CDI_II.lm, which=1)
```



The linearity is violated because the residuals does not bounce randomly around the horizontal line residuals = 0.

Assumption2 - Normality: QQ-plot

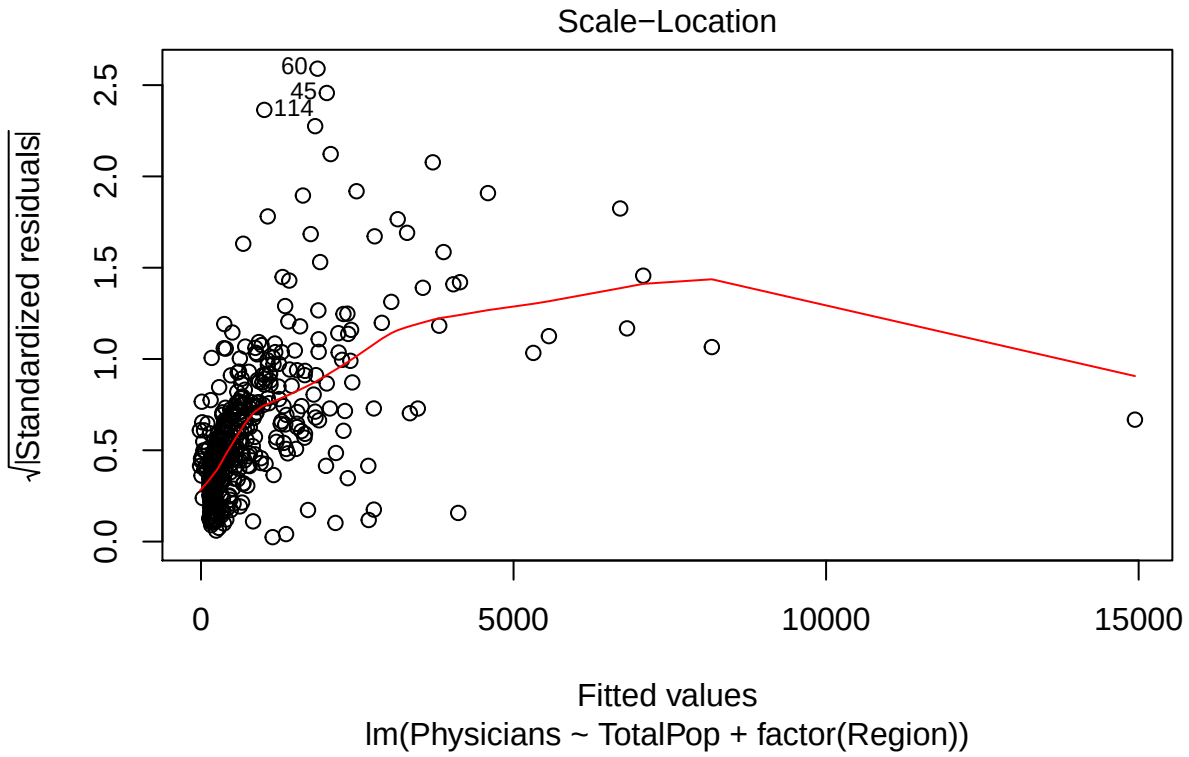
```
plot(CDI_II.lm, which=2)
```



The normality is violated because the standardized residuals are not evenly spread around the reference line and the plot has a clear heavy-tail pattern.

Assumption 3 - Constant Variance: Scale-Location Plot

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



The constant variance is violated because the residuals spread more and more narrowly as the the fitted values increase.
We will now check the outliers.

Testing for outliers

```
outlierTest(CDI_II.lm)
```

```
##     rstudent unadjusted p-value Bonferonni p
## 60  7.091041         5.6970e-12   2.4212e-09
## 45  6.308684         7.1619e-10   3.0438e-07
## 114 5.803531         1.2827e-08   5.4516e-06
## 47  5.343316         1.5014e-07   6.3810e-05
## 43  4.611526         5.3173e-06   2.2599e-03
## 14  4.408130         1.3266e-05   5.6379e-03
```

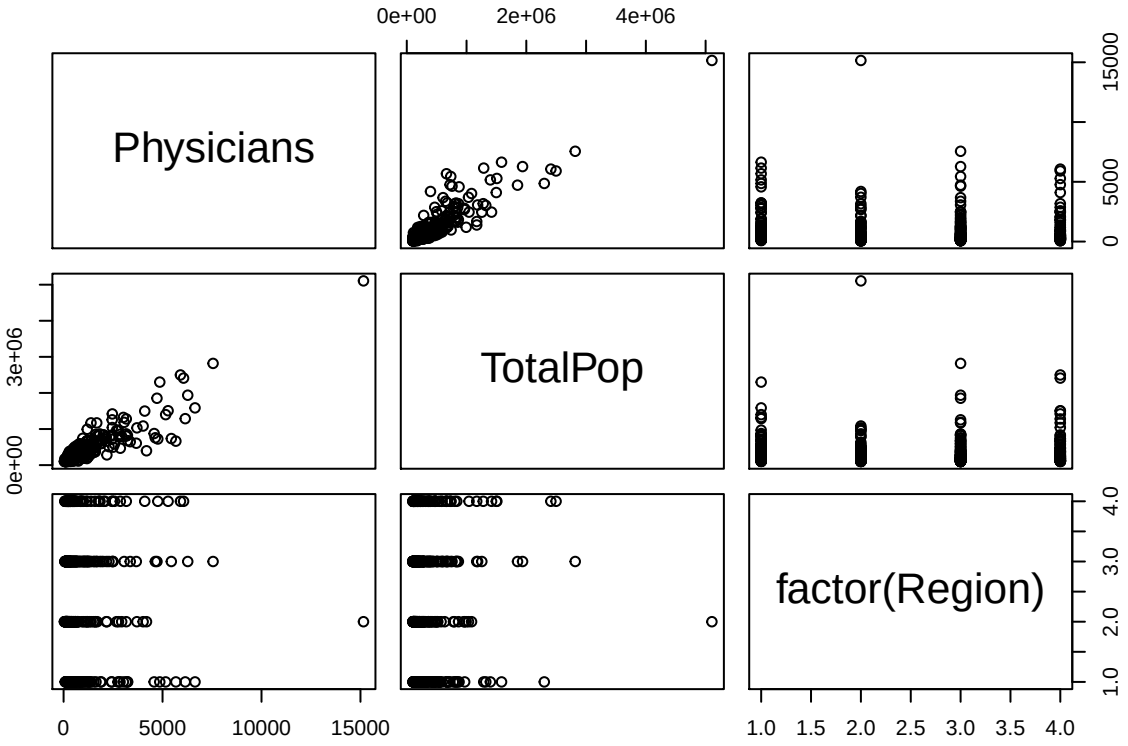
We found 6 outliers in our dataset.
Since our model does not satisfy any of the linear regression assumptions and outliers exist in our data, we will try some transformations for our model.

Transformation

We will try some transformations for predictors **TotalPop** and **Region** first.

Transformating Predictors

```r
#Scatter Matrix Plot
pairs(~Physicians+TotalPop+factor(Region))
```



Some observations:
1. The range for **TotalPop** is very large, from 0 to 4×10^6. Therefore, logarithm is likely to be appropriate for it.
2. Since **Region** is a categorical variable, we will not apply any transformations on it.

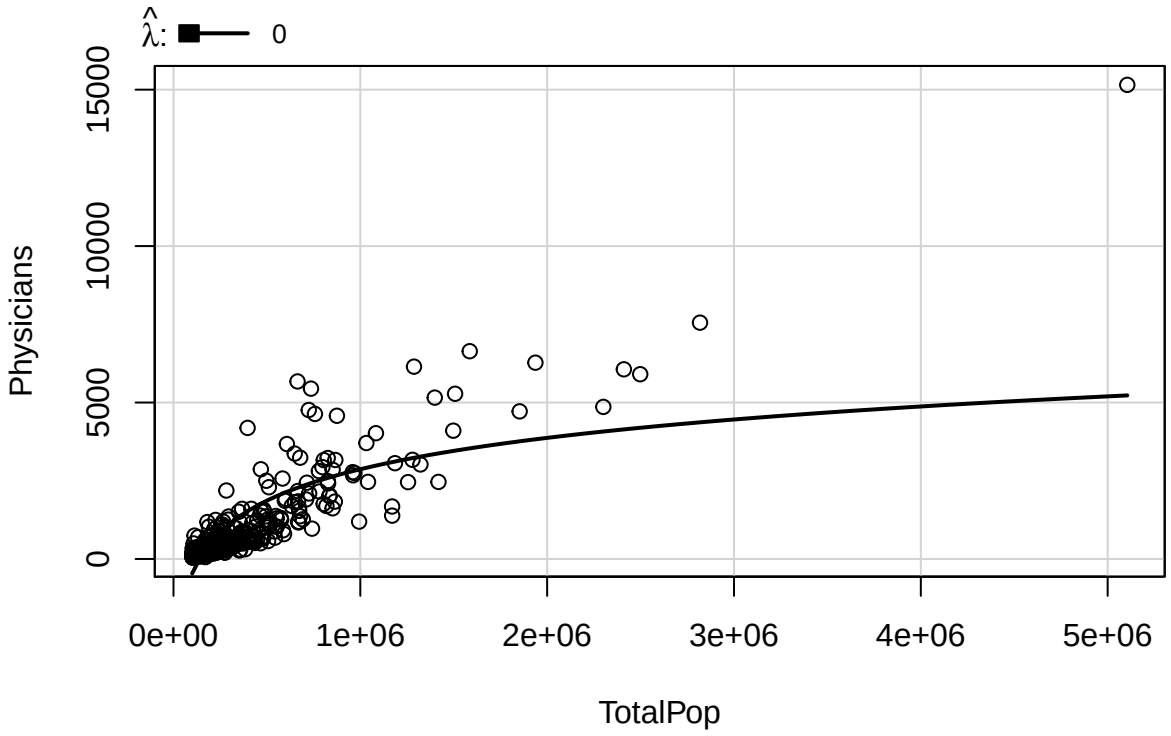
Because of observation 2, we will not transform **Region**. Therefore, we only need to try some transformations for **TotalPop**.

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

The 95% confidence interval contain -0.5. However, since power of -0.5 is hard to interpret and the range for **TotalPop** is broad, we will just apply logarithm for **TotalPop**.
Before doing that, we will check a two-dimentional scatter plot to make sure that logarithm is a appropriate transformtion.

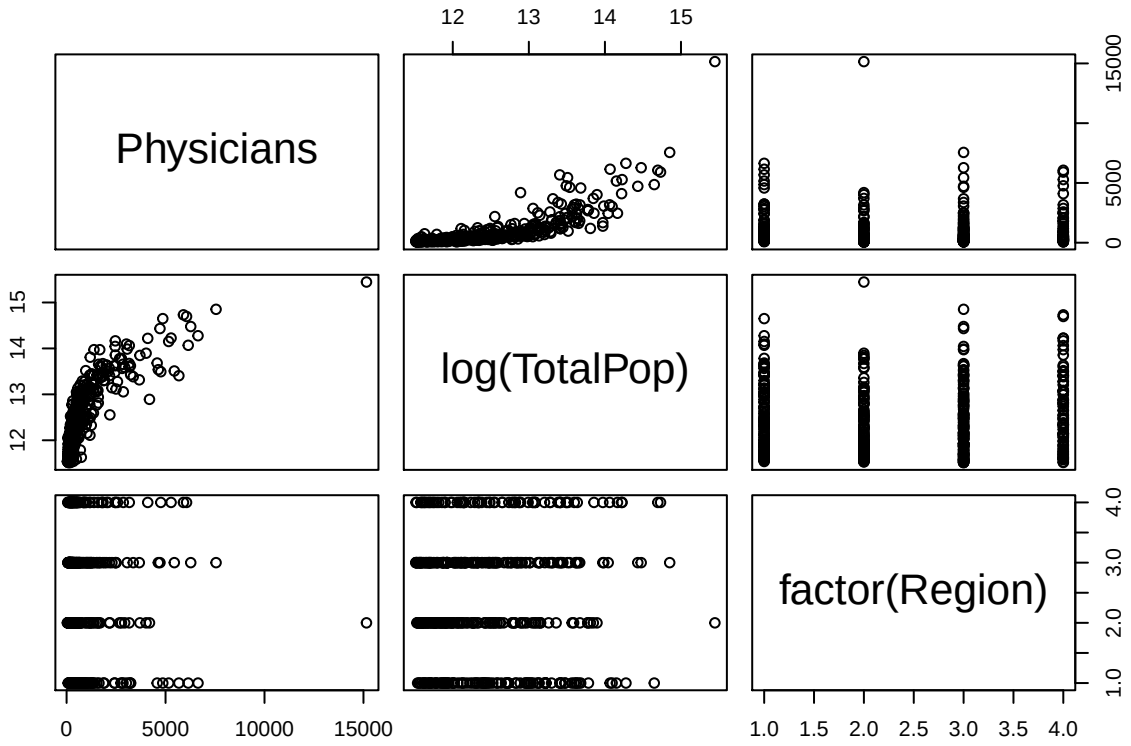
```
invTranPlot(Physicians~TotalPop, lambda = c(0), optimal = F, xlab  = "TotalPop", ylab = "Physicians")
```




```
##   lambda       RSS
## 1      0 315813779
```

From above plot, we find that **TotalPop** with logarithm transformation fits in a good way. Therefore, we will take this transformation. And we will look the matrix plot again after applying transformation to **TotalPop**.

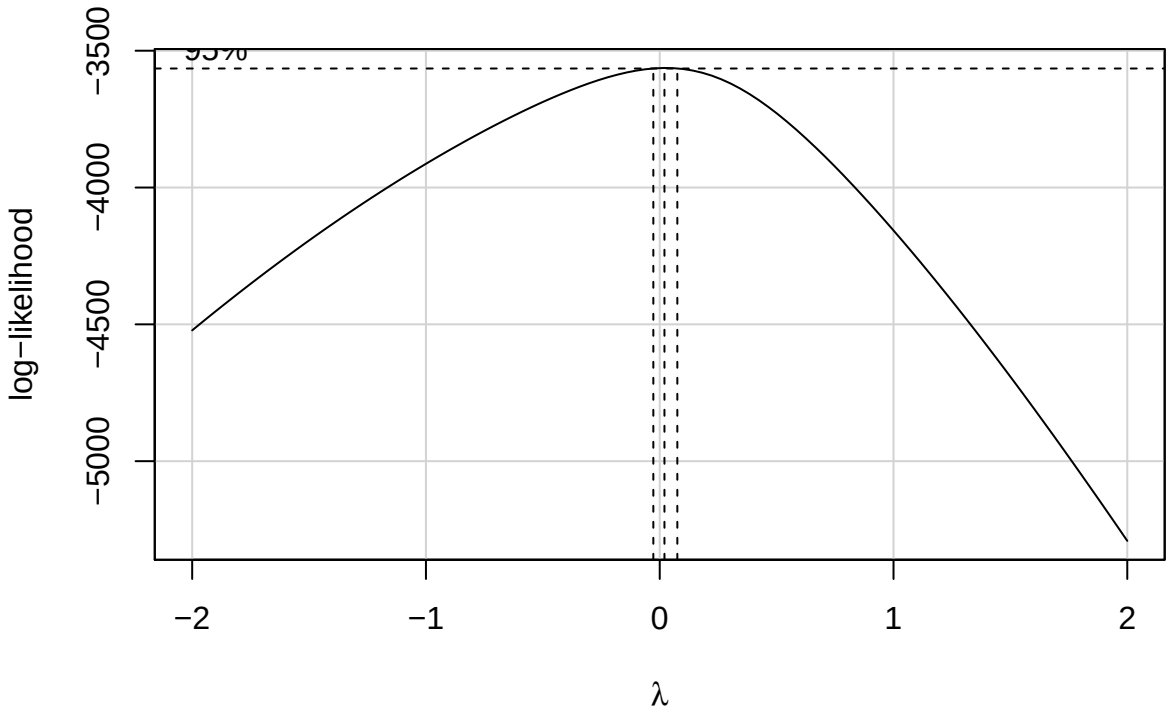
```
pairs(~Physicians+log(TotalPop)+factor(Region))
```



Now we will find a proper transformation for response **Physicians** using the already transformed predcitors. We will use BoxCox method for **Physicians**.

Transforming Response

```
bc = boxCox(Physicians~log(TotalPop)+factor(Region))
```



```
lambda.opt = bc$x[which.max(bc$y)]
lambda.opt
```

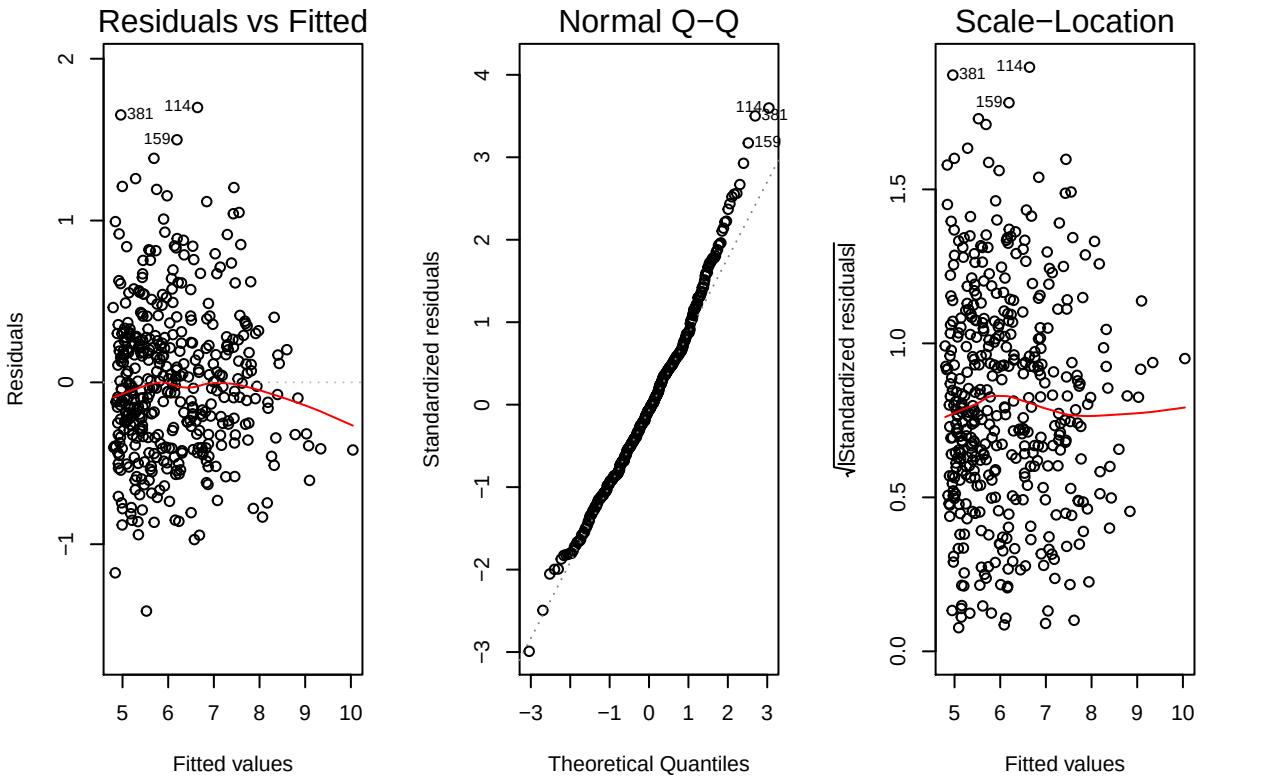
```
## [1] 0.02020202
```

The best optional for lambda is 0.02020202. But since the power of 0.02020202 is hard to interpret and 0 in between our 95% confidence interval, we will choose log transformation here to transform response **Physicians**.
For now, we are done with all transformations. And before we move on, we will check diagnostics and scatter matrix plot again.

$$log(Physicians) \sim log(TotalPop) + Region$$

Check Diagnostics After Transformations

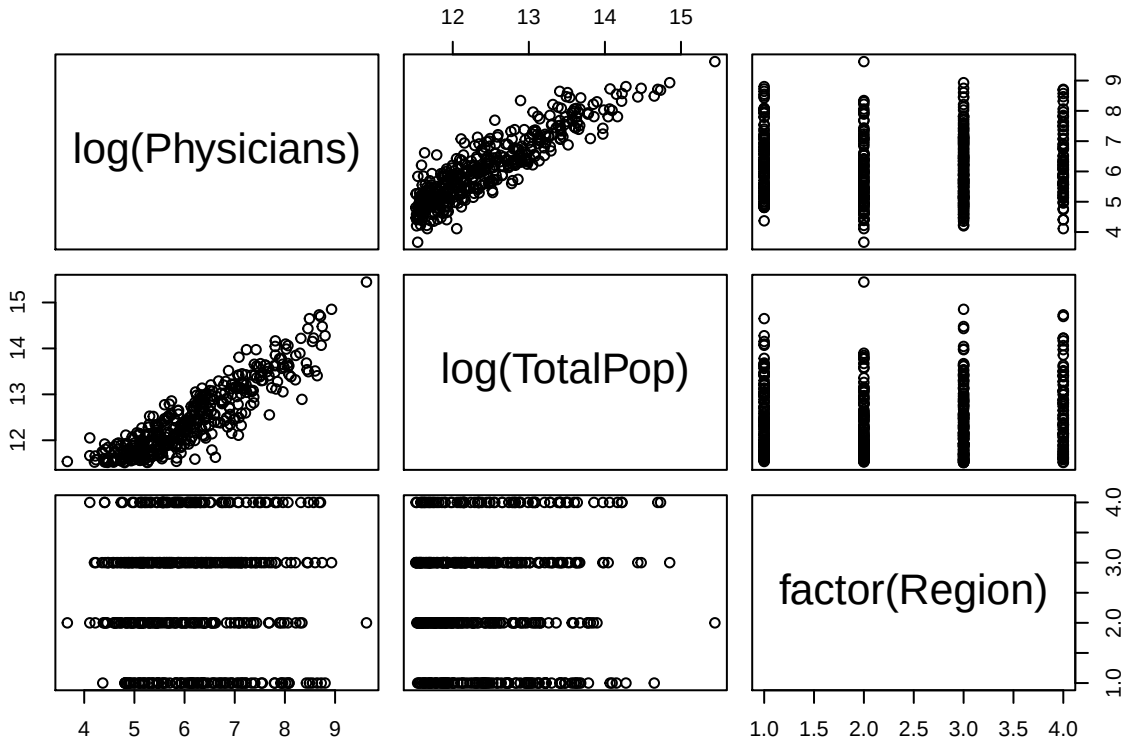
```
CDI_II.lm2 <- lm(log(Physicians)~log(TotalPop)+factor(Region))
par(mfrow = c(1, 3))
for(i in 1:3){
  plot(CDI_II.lm2, which = i)
  }
```



We have already refitted the new model using appropriate transformations as **CDI_II.lm2**. We will first visualize our data by the scatter matrix plot.

Scatter Matrix Plot After Transformations

```
pairs(~log(Physicians)+log(TotalPop)+factor(Region))
```




```
summary(CDI_II.lm2)
```

```
##
## Call:
## lm(formula = log(Physicians) ~ log(TotalPop) + factor(Region))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.41306 -0.32447 -0.02955  0.26244  1.69867
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -10.41248    0.39315 -26.485   <2e-16 ***
## log(TotalPop)     1.33167    0.03107  42.862   <2e-16 ***
## factor(Region)2  -0.11233    0.06699  -1.677   0.0943 .
## factor(Region)3  -0.02468    0.06187  -0.399   0.6902
## factor(Region)4  -0.12888    0.07256  -1.776   0.0764 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4748 on 420 degrees of freedom
## Multiple R-squared:  0.8189, Adjusted R-squared:  0.8171
## F-statistic: 474.7 on 4 and 420 DF,  p-value: < 2.2e-16
```

From the summary of our new model, we notice that the estimated differences between Region1 and Region2, Region3, Region4 separately are all very small. To see this directly, we will write one equation for each region.

Region 1

$$E[log(Physicians)] = -10.41248 + 1.33167 log(TotalPop)$$

Region 2

$$E[log(Physicians)] = -10.52481 + 1.33167log(TotalPop)$$

Region 3

$$E[log(Physicians)] = -10.43716 + 1.33167log(TotalPop)$$

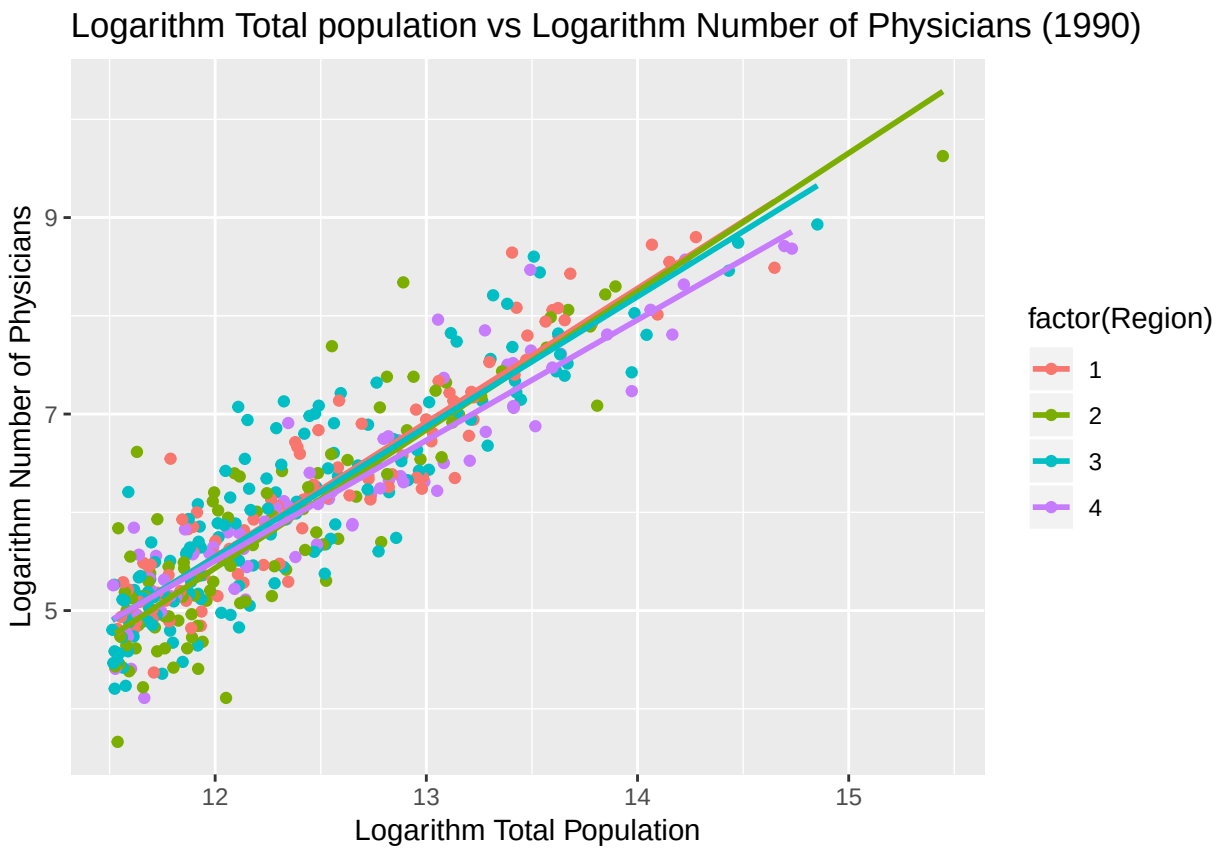
Region 4

$$E[log(Physicians)] = -10.54136 + 1.33167log(TotalPop)$$

We can see that the above four equations are very similar to each other given different regions. The intercepts are slightly different and the estimated coefficients are the same for **log(TotalPop)**. Since the four mean equations have same slope, they are parallel to each other, which is the reason our regression model is called parallel regression model. We could obtain same conclusion from the following plot.

Plot log(TotalPop) vs log(Physicians)

```
ggplot(CDI, aes(log(TotalPop), log(Physicians), color=factor(Region)))+
  geom_point() +
  labs(x = 'Logarithm Total Population',
       y = 'Logarithm Number of Physicians',
       title = 'Logarithm Total population vs Logarithm Number of Physicians (1990)') +
  geom_smooth(method=lm, se=FALSE)
```



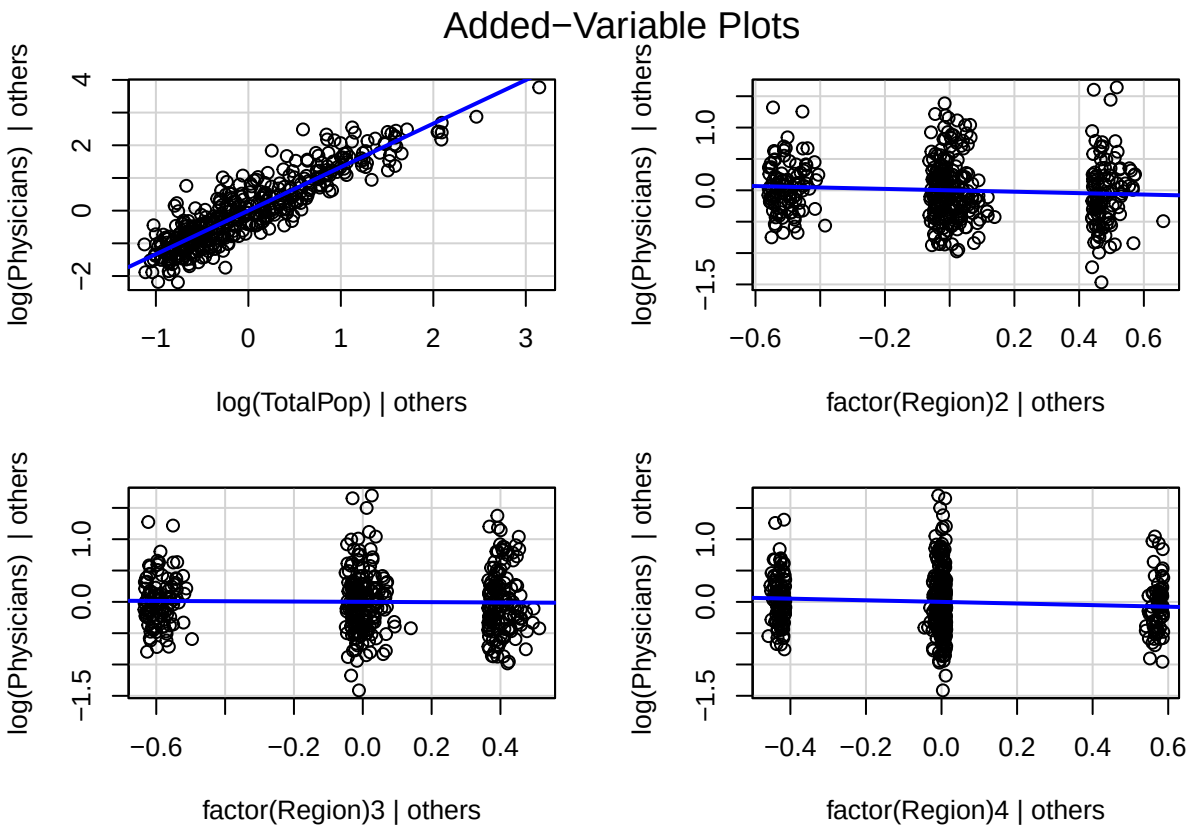
From the summary we know that the p-value of three dummary variables of Region is 0.0943,0.6902 and 0.0764, respectively, which are large. However, we are not 100% sure that Region is not useful in this model because we don't keep other variable constant.

Are Region(4 levels) useful predictors given that the other predictors are in the model?

```
# The model without Region
CDI_without<-lm(log(Physicians)~log(TotalPop))
#Compare
anova(CDI_II.lm2,CDI_without)
```

```
## Analysis of Variance Table
##
## Model 1: log(Physicians) ~ log(TotalPop) + factor(Region)
## Model 2: log(Physicians) ~ log(TotalPop)
##   Res.Df    RSS Df Sum of Sq     F Pr(>F)
## 1    420 94.666
## 2    423 95.848 -3    -1.182 1.748 0.1565
```

```
avPlots(CDI_II.lm2, id=FALSE)
```



We conduct F-test for the model have Region in and without Region in and we can see the p-value is 0.1565 > 0.05. So we fail to reject the null hypothesis that $\beta_2 = \beta_3 = \beta_4 = 0$. Also, form the added-variable plot we can see that the AV plot for each level of Region is a flat horizontal line, Region2,3,4 after log(TotalPop) shows that it is not useful when log(TotalPop) is already in the model. So we assume that Region is not important for the model.

Intuitively, from the Logarithm Total population vs Logarithm Number of Physicians (1990) plot, we can't

tell the difference among the four levels of Region, and we can see that the intercept of the regression line in the four levels of Region almost the same. Our conclusion from F-test is reasonable.

We want to build on our current model by selecting relevant predictors from Pop65, Crimes, Bachelor, Poverty, and PersonalInc.

```
CDI_c<-lm(log(Physicians) ~ log(TotalPop))
CDI.selfrom<-lm(log(Physicians) ~ log(TotalPop) + Pop65 + Crimes + Bachelor + Poverty + PersonalInc)
```

Forward stepwise selection

```
step(CDI_c,scope = list(lower=CDI_c, upper=CDI.selfrom),
     direction ="forward")
```

```
## Start:  AIC=-628.97
## log(Physicians) ~ log(TotalPop)
##
##               Df Sum of Sq    RSS     AIC
## + Bachelor     1   16.1403 79.707 -705.33
## + Pop65        1    1.1733 94.674 -632.20
## + Poverty      1    0.7187 95.129 -630.16
## <none>                     95.848 -628.97
## + Crimes       1    0.2865 95.561 -628.24
## + PersonalInc  1    0.1055 95.742 -627.43
##
## Step:  AIC=-705.33
## log(Physicians) ~ log(TotalPop) + Bachelor
##
##               Df Sum of Sq    RSS     AIC
## + Poverty      1    7.6927 72.015 -746.47
## + Pop65        1    7.2481 72.459 -743.85
## <none>                     79.707 -705.33
## + PersonalInc  1    0.3556 79.352 -705.23
## + Crimes       1    0.0174 79.690 -703.43
##
## Step:  AIC=-746.47
## log(Physicians) ~ log(TotalPop) + Bachelor + Poverty
##
##               Df Sum of Sq    RSS     AIC
## + Pop65        1   10.3770 61.638 -810.60
## <none>                     72.015 -746.47
## + Crimes       1    0.1992 71.815 -745.65
## + PersonalInc  1    0.1093 71.905 -745.11
##
## Step:  AIC=-810.6
## log(Physicians) ~ log(TotalPop) + Bachelor + Poverty + Pop65
##
##               Df Sum of Sq    RSS     AIC
## <none>                     61.638 -810.60
## + PersonalInc  1  0.132438 61.505 -809.51
## + Crimes       1  0.083341 61.554 -809.17
```

```
##
## Call:
## lm(formula = log(Physicians) ~ log(TotalPop) + Bachelor + Poverty +
```

```
##     Pop65)
##
## Coefficients:
##   (Intercept)  log(TotalPop)       Bachelor        Poverty          Pop65
##     -10.50625        1.18840         0.04632        0.03820        0.04226
```

```
#use AIC by default
```

AIC=-810.6 is the smallest.

We get the best model which is: *log(Physicians) ~ log(TotalPop) + Bachelor + Poverty + Pop65*

Backward stepwise selection

```
step(CDI.selfrom,scope = list(lower=CDI_c, upper=CDI.selfrom),
     direction ="backward")
```

```
## Start:  AIC=-807.58
## log(Physicians) ~ log(TotalPop) + Pop65 + Crimes + Bachelor +
##     Poverty + PersonalInc
##
##               Df Sum of Sq    RSS     AIC
## - Crimes       1     0.010 61.505 -809.51
## - PersonalInc  1     0.059 61.554 -809.17
## <none>                     61.495 -807.58
## - Poverty      1     9.811 71.306 -746.67
## - Pop65        1    10.311 71.806 -743.70
## - Bachelor     1    31.839 93.334 -632.26
##
## Step:  AIC=-809.51
## log(Physicians) ~ log(TotalPop) + Pop65 + Bachelor + Poverty +
##     PersonalInc
##
##               Df Sum of Sq    RSS     AIC
## - PersonalInc  1     0.132 61.638 -810.60
## <none>                     61.505 -809.51
## - Pop65        1    10.400 71.905 -745.11
## - Poverty      1    10.507 72.012 -744.48
## - Bachelor     1    32.400 93.905 -631.67
##
## Step:  AIC=-810.6
## log(Physicians) ~ log(TotalPop) + Pop65 + Bachelor + Poverty
##
##            Df Sum of Sq    RSS     AIC
## <none>                  61.638 -810.60
## - Pop65     1    10.377 72.015 -746.47
## - Poverty   1    10.822 72.459 -743.85
## - Bachelor  1    32.320 93.958 -633.43
```

```
##
## Call:
## lm(formula = log(Physicians) ~ log(TotalPop) + Pop65 + Bachelor +
##     Poverty)
##
## Coefficients:
##   (Intercept)  log(TotalPop)          Pop65       Bachelor        Poverty
```

```
##     -10.50625        1.18840        0.04226        0.04632        0.03820
```

AIC=-810.6

The model we will choose is that *log(Physicians) ~ log(TotalPop) + Pop65 + Bachelor + Poverty*

Stepwise stepwise selection

```
step(CDI_c,scope = list(lower=CDI_c, upper=CDI.selfrom),
     direction ="both")
```

```
## Start:  AIC=-628.97
## log(Physicians) ~ log(TotalPop)
##
##               Df Sum of Sq    RSS     AIC
## + Bachelor     1   16.1403 79.707 -705.33
## + Pop65        1    1.1733 94.674 -632.20
## + Poverty      1    0.7187 95.129 -630.16
## <none>                     95.848 -628.97
## + Crimes       1    0.2865 95.561 -628.24
## + PersonalInc  1    0.1055 95.742 -627.43
##
## Step:  AIC=-705.33
## log(Physicians) ~ log(TotalPop) + Bachelor
##
##               Df Sum of Sq    RSS     AIC
## + Poverty      1    7.6927 72.015 -746.47
## + Pop65        1    7.2481 72.459 -743.85
## <none>                     79.707 -705.33
## + PersonalInc  1    0.3556 79.352 -705.23
## + Crimes       1    0.0174 79.690 -703.43
## - Bachelor     1   16.1403 95.848 -628.97
##
## Step:  AIC=-746.47
## log(Physicians) ~ log(TotalPop) + Bachelor + Poverty
##
##               Df Sum of Sq    RSS     AIC
## + Pop65        1   10.3770 61.638 -810.60
## <none>                     72.015 -746.47
## + Crimes       1    0.1992 71.815 -745.65
## + PersonalInc  1    0.1093 71.905 -745.11
## - Poverty      1    7.6927 79.707 -705.33
## - Bachelor     1   23.1143 95.129 -630.16
##
## Step:  AIC=-810.6
## log(Physicians) ~ log(TotalPop) + Bachelor + Poverty + Pop65
##
##               Df Sum of Sq    RSS     AIC
## <none>                     61.638 -810.60
## + PersonalInc  1     0.132 61.505 -809.51
## + Crimes       1     0.083 61.554 -809.17
## - Pop65        1    10.377 72.015 -746.47
## - Poverty      1    10.822 72.459 -743.85
## - Bachelor     1    32.320 93.958 -633.43
```

```
##
```

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

AIC= -810.6

The best model is: *log(Physicians) ~ log(TotalPop) + Bachelor + Poverty + Pop65*

We get the same model for these three model selection results.

We conduct a F-test for the first model and the model we chose by using predictor selection techniques to see which one is better.

```
choose.lm<-lm(log(Physicians) ~ log(TotalPop) + Bachelor + Poverty + Pop65)
anova(CDI_c,choose.lm)
```

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

The p-value is less than 2.2×10^{-16} which is very small so we can reject the null hypothesis and assume that the selected model is useful compare to the first model.

Influential Points

Our model looks pretty good now, but let's check for influential points. This means we need to look at points have large residuals (outlier in Y) and/or large leverage (outlier in x).

```
outlierTest(choose.lm)
```

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

Since our Bonferonni p-value is more than 0.05, observation 119, Hidalgo County, is an outlier in our dataset. We may need to collect more data if we do not want this point to be considered as an outlier. How would our model compare if we deleted Hidalgo County from our dataset?

```
#new dataframe df, which does not contain Hidalgo
df<-CDI[CDI$County!="Hidalgo",]
df.lm<-lm(log(df$Physicians) ~ log(df$TotalPop) + df$Bachelor + df$Poverty + df$Pop65)
df.lm$coefficients
```

```
##      (Intercept) log(df$TotalPop)      df$Bachelor       df$Poverty
##     -10.57702864       1.19014307       0.04686952       0.04307390
##         df$Pop65
##       0.04215304
```

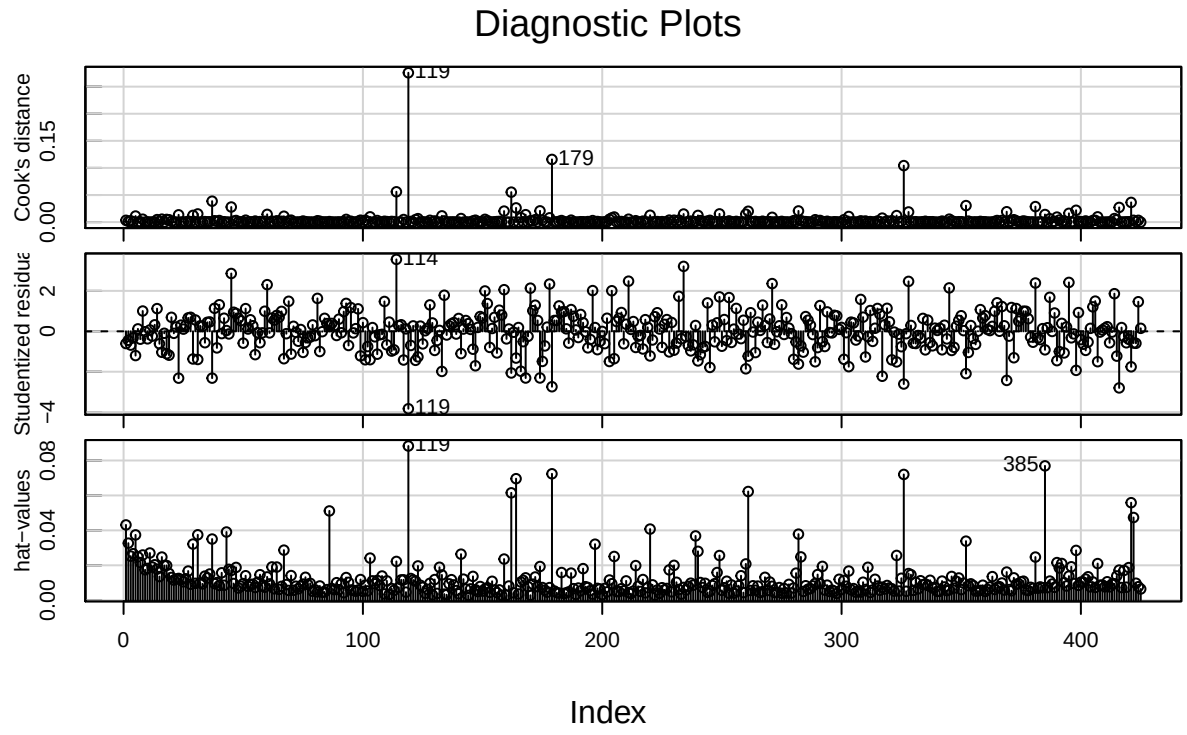
```
choose.lm$coefficients
```

```
##   (Intercept) log(TotalPop)      Bachelor       Poverty         Pop65
##  -10.50624787    1.18840363    0.04632223    0.03820213    0.04226417
```

The coefficient for Poverty was affected the most by removing Hidalgo. We will keep this outlier in our dataset. If we do not want this county to be considered an outlier, we would need to incorporate more information.

Let's look for influential points using a different function.

```
influenceIndexPlot(choose.lm, vars=c('Cook','Studentized','hat'))
```



As we can see from the first plot, observations 119 and 179 have the largest Cook's distance, so they are influential points. They are influential because in the second plot, we can see that they have pretty large residuals. Also, the last plot shows us hat-values, which means leverage. Observations 119 and 179 have large leverage. Therefore, they are influential points. It is worth noting that observation 385 has large leverage but it is not influential because its residual is not that big.

Summary

To summarize, we began a new, second investigation on the model

$$Physicians \sim TotalPop + Region \tag{1}$$

and found that this model was unsatisfactory because its diagnostic plots violate model assumptions. so we made some transformations and came up with

$$log(Physicians) \sim log(TotalPop) + Region \tag{2}$$

We picked this model because its diagnostic plots did not violate model assumptions anymore.

However, we were not sure if **Region** was needed. So we performed a F-test between the model without **Region** and the model with **Region**. It turns out that it is not an important predictor to include. We still had other predictors left to consider, though, such as **Pop65**, **Crimes**, **Bachelor**, **Poverty**, and **PersonalInc**. So we used stepwise selection to find the model with the lowest AIC. As a result, we got

$$log(Physicians) \sim log(TotalPop) + Bachelor + Poverty + Pop65 \tag{3}$$

The last thing we needed to do was to check for any influential points, and we found that observations 119 and 179 had high influence on our model.

What was interesting was that our chosen model from our first investigation on predicting the number of physicians based only on the predictors **log(TotalPop)**, **LandArea**, and **IncPerCap** resulted in us choosing the model

$$log(Physicians) \sim log(TotalPop) + log(LandArea) + IncPerCap \tag{4}$$

which is very different from our final model in the second investigation (where we considered more possible predictors)

$$log(Physicians) \sim log(TotalPop) + Bachelor + Poverty + Pop65 \tag{5}$$

Only **log(TotalPop)** stayed as a predictor in both models, which makes sense because it's common sense that the more people in a county, the more physicians the county needs. We believe that **IncPerCap** and **Poverty** are probably highly correlated with each other, so we only needed to include one of them in our final model. Also, it's interesting to note that **LandArea** is not as good of a predictor compared to **Bachelor**, **Poverty**, and **Pop65**. This may be because some counties have big land areas, but fewer physicians on average or some have small land areas, but more physicians on average.

Our final model makes sense, though, because a county with more graduates means more physicans on average (since being a physician requires at least a bachelor's degree). Additionally, less wealthier counties may tend to have less physicians because not everyone could afford to see a doctor. Finally, the size of the elderly population is an important predictor to include because they need to see the doctor more often, so physicians are in demand. Therefore, **Bachelor**, **Poverty**, and **Pop65** are great predictors to include.

```
summary(choose.lm)
```

```
##
## Call:
## lm(formula = log(Physicians) ~ log(TotalPop) + Bachelor + Poverty +
##     Pop65)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.37983 -0.23391  0.01604  0.22198  1.32960
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -10.506248   0.318585 -32.978  < 2e-16 ***
## log(TotalPop)   1.188404   0.026658  44.579  < 2e-16 ***
## Bachelor        0.046322   0.003121  14.840  < 2e-16 ***
## Poverty         0.038202   0.004449   8.587  < 2e-16 ***
## Pop65           0.042264   0.005026   8.409 6.52e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3831 on 420 degrees of freedom
## Multiple R-squared:  0.8821, Adjusted R-squared:  0.8809
## F-statistic: 785.3 on 4 and 420 DF,  p-value: < 2.2e-16
```

Conclusion

Our model tells us that the number of professionally active nonfederal physicians during 1990 are related with many factors. Specifically, more population during 1990 stands for more physicians. However, for some counties, the more land area it has, the less physicians it had. For other counties, the opposite is true. This is consistent with our first guess. Per capita income has a very weak relationship with the number of physicians but it could useful when we use it together with total population and land area to predict the number of physicians. Similarly, per capita income, percent of adults with a bachelor degree, percent of adults in poverty and percent of adults over 65 are also weakly associated with the number of physicians but play important roles in our model; they could predict the number of physicians well together with total population. Besides, our second model, choose.lm, is better our first model, CDI_new.redu, since the second model has a higher adjusted R^2. This means that more variability in the number of physicians are explained by the linear relationship with all predictors in second model than the first model. But in both models, we find that total population has strongest relatonship with number of physicians.