PSTAT 126 Final Project: Option 1

Brandon Hang, Michael Kwok, Warner Lew

June 9, 2019

Abstract

In this project, we aim to find the best model to predict the number of physicians in a single county. We looked at a few different models that all had multiple variables and transformations. Various methods of regression were used to generate two optimal models.

Problem and Motivation

The main problem of this project is finding the right variables and transformations to accurately predict the number of physicians in a county. The CDI dataset comes with numerous variables that may or may not be useful. Many of these variables are not normally distributed, nor are they related in ways that we initially expected. Once a useful model is obtained, we can answer questions such as what factors, such as crime, poverty, and elderly population, are correlated with a higher requirement of physicians. Such model can help urban planners better understand and suit their communities' needs. It can also help physicians identify which locations have a high demand for their services.

Data

The dataset describes the county demographic information from 440 of the most populous counties from 1990-1992 in the US. The variables tested include: number of professional physicians (Physicians), land area in square miles (LandArea), estimated population (TotalPop), total number of serious crimes (Crimes), percentage of adult population with a bachelor's degree (Bachelor), population of people 65 or older (Pop65), per capita income (IncPerCap), total personal income (PersonalInc), geographical region (Region), and percentage of population with income below poverty level (Poverty).

Questions of Interest

In our models, we are looking specific predictors that best estimate the number of physicians in a county. The relationships between predictors also play a large part in selecting a model. We also run diagnostics to look for violations of the assumptions of a linear model, and we look for the ideal adjustments for predictor variables if adjustment is necessary.

Regression Methods

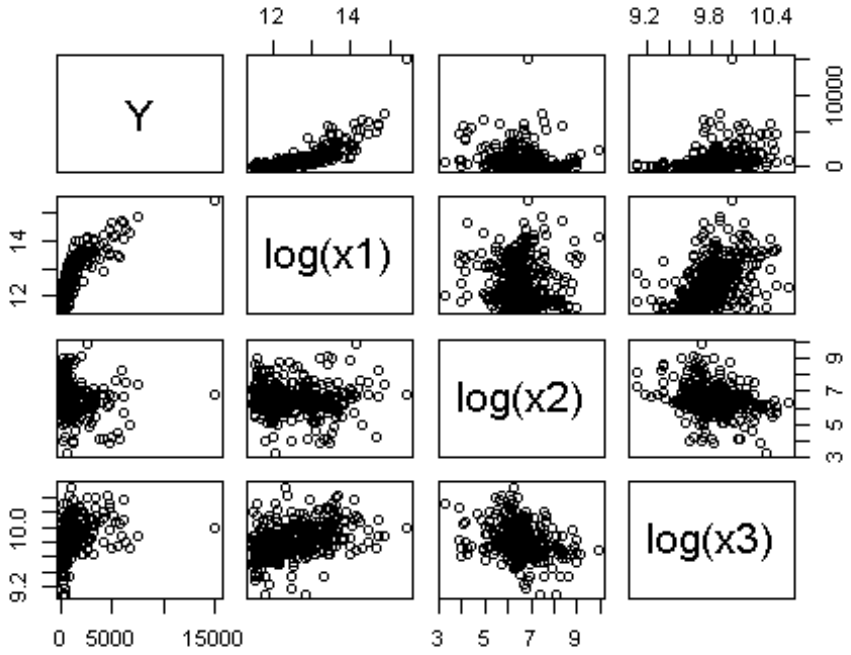
We used the pairs plot to analyze the relationships between all variables in our models. Residual vs fitted value plots and QQ-plots were used to assess the diagnostics (linearity, independence, normality, and equal variance) of our linear models. We used the Box-Cox method to obtain the optimal transformations of our response variables and PowerTransform method to check for any suggested predictor transformations. A non-constant variance test was also run to see if weight adjustments are necessary for the model. We ran ANOVA F-tests and used best subset regression to evaluate models (using measures like AIC, BIC, Cp values, R-squared, and adjusted-R-squared) and determine which variables are useful for prediction. We also used Cook's distance, hat value, and

studentized plots to identify influential points (high leverage and outliers) in the data so we can remove them and improve the accuracy of our models.

Regressional Analysis

1.

```
library(leaps)

## Warning: package 'leaps' was built under R version 3.5.3

library(alr4)

## Warning: package 'alr4' was built under R version 3.5.3

## Loading required package: car

## Warning: package 'car' was built under R version 3.5.3

## Loading required package: carData

## Warning: package 'carData' was built under R version 3.5.2

## Loading required package: effects

## Warning: package 'effects' was built under R version 3.5.3

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

CDI <- readRDS("C:/Users/Michael/Downloads/CDI.rds")
attach(CDI)

#1a
Y <- CDI$Physicians
x1 <- CDI$TotalPop
x2 <- CDI$LandArea
x3 <- CDI$IncPerCap
#we expect to see positive relationships between physicians and each
predictor
pairs(Y~log(x1)+log(x2)+log(x3))
```



1a)

We expect to see positive correlation between the response and all 3 predictors. Total population should be positively correlated with the number of physicians, because as population increases, the number of physicians should also increase to take care of the larger population. Land Area should also be positively correlated with number of physicians because increasing land area may provide space for more physicians to provide service in the county. Since counties with larger per capita income should be more able to afford physician care, so increasing per capita income may indicate larger numbers of physicians in the county. As for relationships between predictors, land area and total population may be positively correlated since an increasing number of people could live in a larger land area.

Observing the relationships between each variable in the scatterplot matrix reveals that our intuition differs from the actual results. The logarithm of total population appears to have a linear relationship with the number of physicians when controlling the effects of other predictors. Income per capita also appears to have a slight linear relationship with the number of physicians as well as with the logarithm of total population. Land area does not appear to be linearly related to any of the other variables.

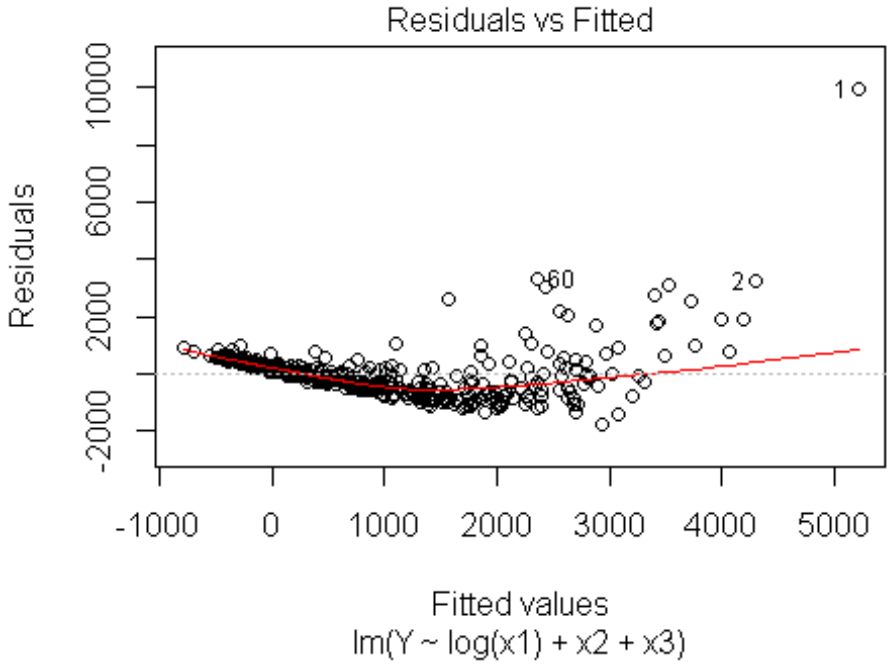
```
#1b
cdi.lm <- lm(Y~log(x1)+x2+x3)
summary(cdi.lm)

##
## Call:
## lm(formula = Y ~ log(x1) + x2 + x3)
```

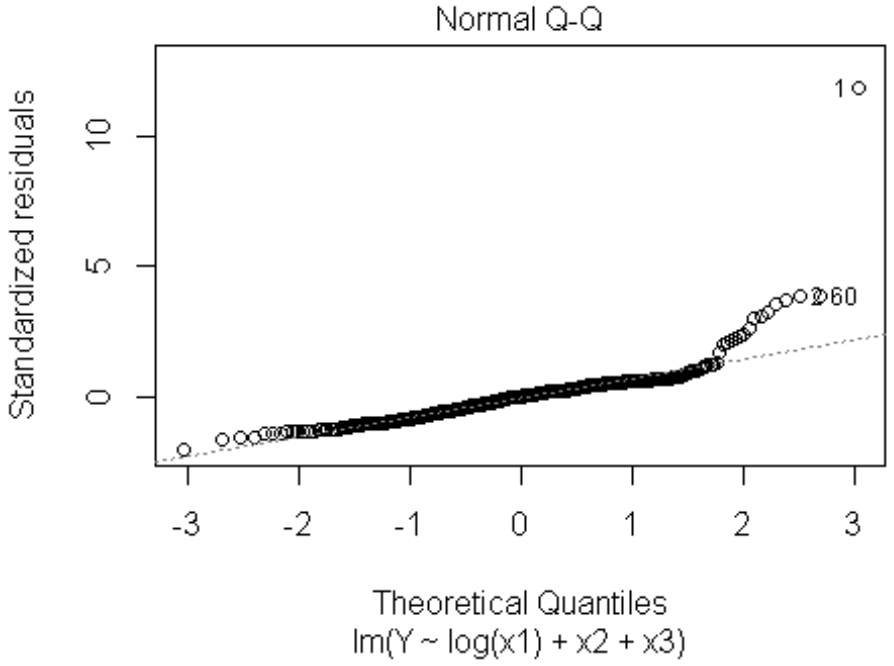
```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1739.9  -495.4    -5.4   375.4  9938.9
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.706e+04  7.060e+02 -24.165   <2e-16 ***
## log(x1)      1.427e+03  6.293e+01  22.683   <2e-16 ***
## x2          -5.488e-02  2.865e-02  -1.916   0.0561 .
## x3           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
```

1b) For every 1% increase in total pop, we have a 1427*ln(1.01)=14.199 increase in number of physicians when all other factors are held constant. For every 1 square mile increase in land area, we have a .05488 decrease in number of physicians when all other factors are held constant. For every $1 increase in per capita income, we have a .01285 increase in number of physicians.

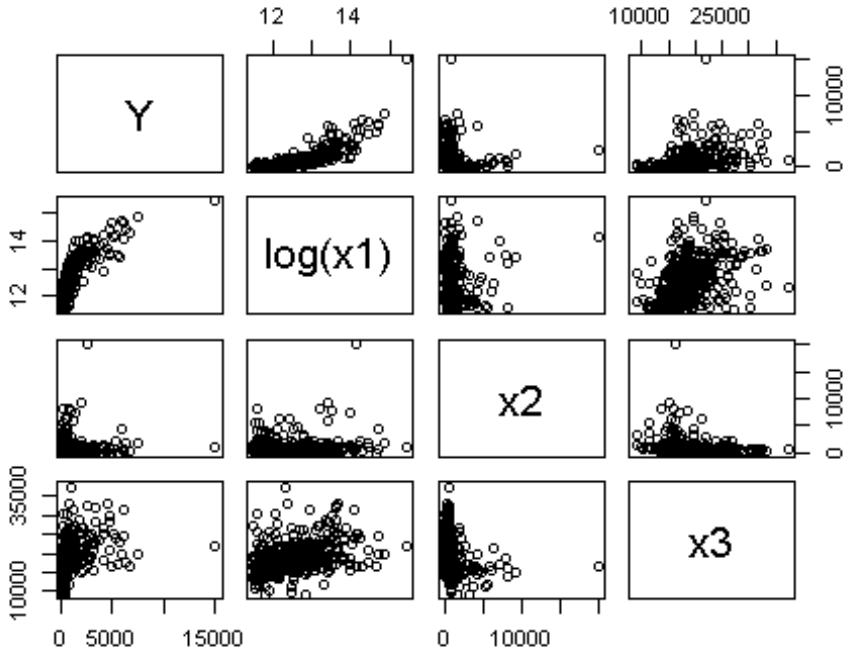
The adjusted R^2 value is 0.6175, meaning 61.75% of the variability in the number of physicians is explained by the multiple regression model with all the predictors collectively.

```
#1c

#diagnostics
plot(cdi.lm,which=c(1,2))
```

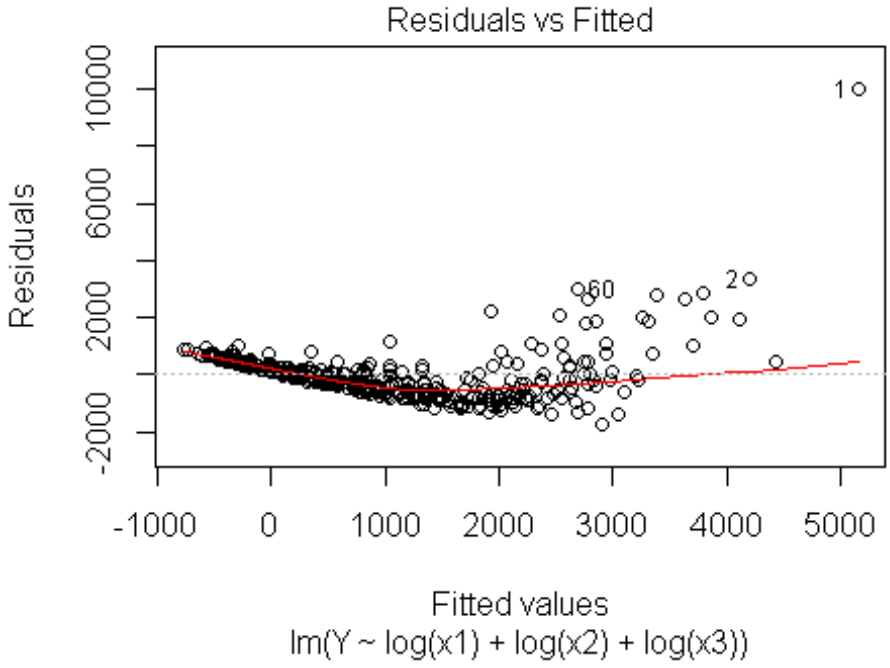


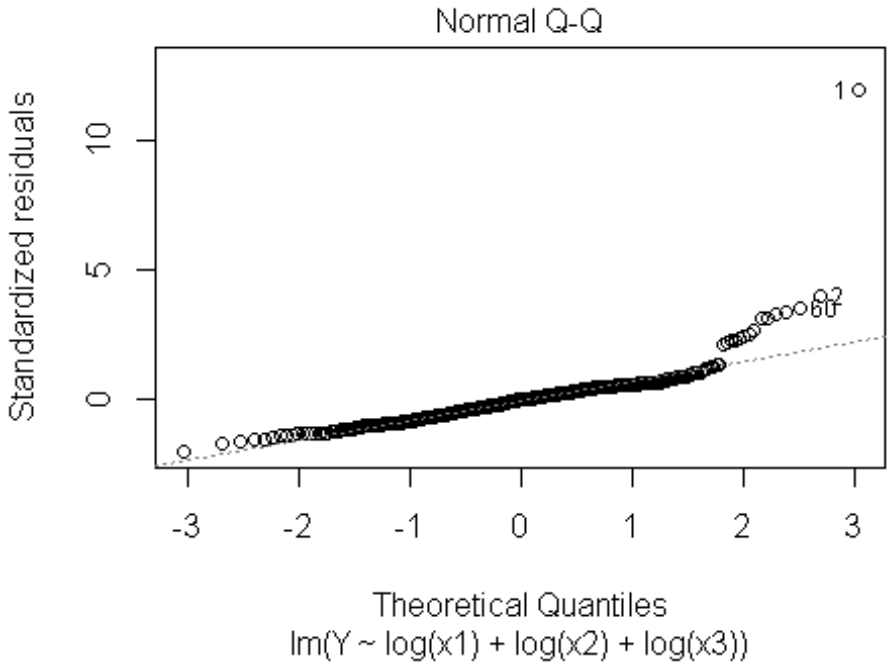



```
pairs(Y~log(x1)+x2+x3)
```



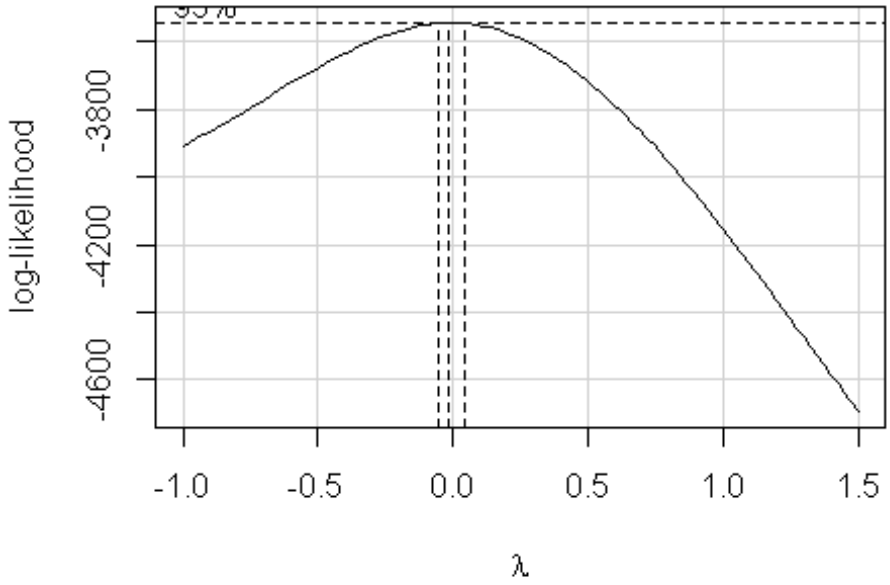

```
#LINE violations (nonconstant variance, nonlinearity, non-normality)
#solution: check powerTransform (for predictors), then boxcox (for response)

#powerTransform
cdi.pt <- powerTransform(cbind(x1,x2,x3)~1)
summary(cdi.pt) #suggests transforming all predictors

## bcPower Transformations to Multinormality
##    Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## x1   -0.5589        -0.5      -0.6941      -0.4236
## x2   -0.0080         0.0      -0.0727       0.0567
## x3   -0.3156        -0.5      -0.5978      -0.0334
##
## Likelihood ratio test that transformation parameters are equal to 0
##  (all log transformations)
##                                  LRT df       pval
## LR test, lambda = (0 0 0) 81.81138  3 < 2.22e-16
##
## Likelihood ratio test that no transformations are needed
##                                  LRT df       pval
## LR test, lambda = (1 1 1) 1714.683  3 < 2.22e-16

cdi.lmX <- lm(Y~log(x1)+log(x2)+log(x3))
plot(cdi.lmX,which=c(1,2))
```



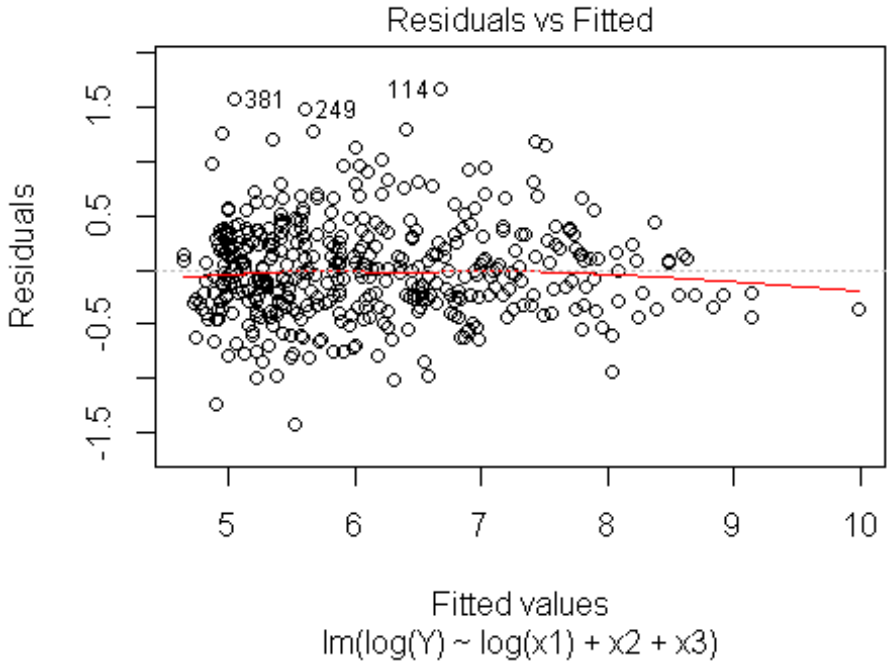


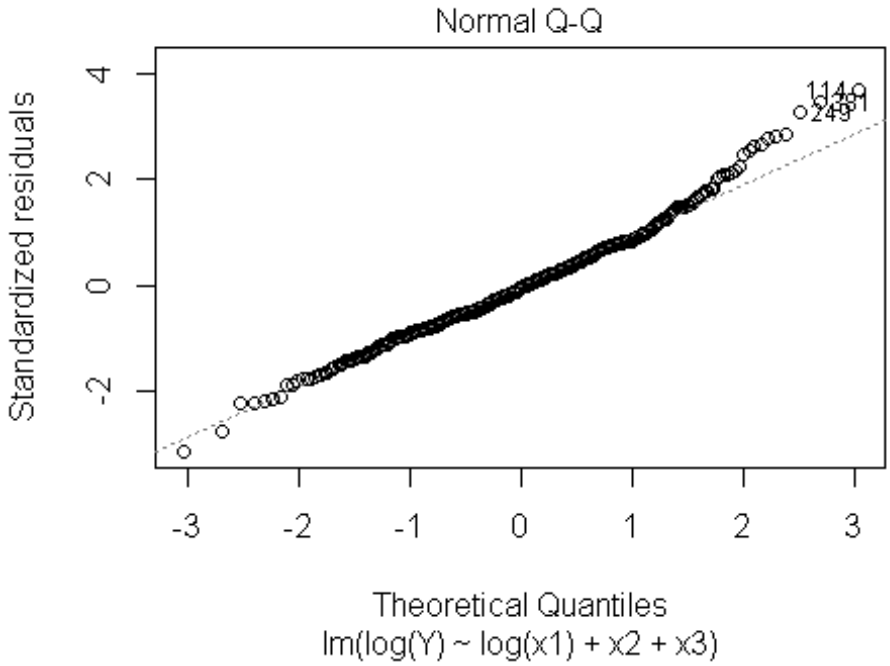

```
#improvements from log-transforming x2 and x3 are not visible/significant,
not worth the change
```

```r
#boxCox
cdi.bc <- boxCox(Y~log(x1)+x2+x3,lambda = c(-1, -.5, 0, .5, 1, 1.5))
```

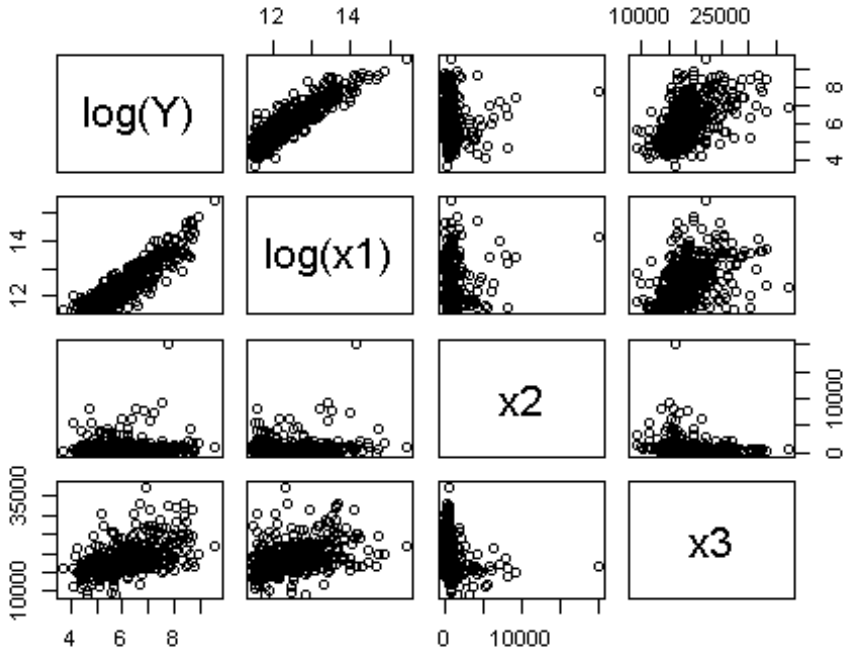



```r
summary(cdi.bc) #suggests transforming response

##   Length Class  Mode
## x 100    -none- numeric
## y 100    -none- numeric

cdi.lmY <- lm(log(Y)~log(x1)+x2+x3)
plot(cdi.lmY,which=c(1,2))
```






```
pairs(log(Y)~log(x1)+x2+x3)
```




```
summary(cdi.lmY)

##
## Call:
## lm(formula = log(Y) ~ log(x1) + x2 + x3)
##
## 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(x1)      1.255e+00  3.323e-02  37.780  < 2e-16 ***
## x2          -2.980e-05  1.513e-05  -1.970   0.0495 *
## x3           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
```

1c)

For the original model, none of the conditions for linear regression are met. By observing the residuals versus fits plot, we see that the points are not randomly scattered about e=0 nor are they contained in a horizontal band about e=0, thus the model is not linear nor does

it have constant variance. By observing the QQ-plot, we can see that the data is not normally distributed because the points are not in a linear formation.
By using the Box-Cox method, we conclude that a logarithmic transformation of the number of physicians adequately accounts for non-normality, non-constant variance, and linearity.

For every 1% increase in total pop, we have a 1.01^1.255= 1.0126 increase in the expected median number of physicians when all other factors are held constant. For every 1 square mile increase in land area, we have a e^(-2.98e-05)=1 decrease in the expected median number of physicians when all other factors are held constant. For every $1 increase in per capita income, we have a e^(3.531e-05)=1 increase in the expected median number of physicians when all other factors are held constant.

The adjusted R^2 value is 0.8328, so 83.28% of the variability in the logarithm of the number of physicians is explained by the regression model with all the predictors collectively.

```
#1d
confint(cdi.lmY, level=.95) #confidence intervals
```

```
##                      2.5 %        97.5 %
## (Intercept) -1.087549e+01 -9.410109e+00
## log(x1)      1.189973e+00  1.320591e+00
## x2          -5.952709e-05 -6.439696e-08
## x3           2.295433e-05  4.766106e-05
```

```
summary(cdi.lmY) #test linear relationship between predictors and response at
alpha = .01
```

```
##
## Call:
## lm(formula = log(Y) ~ log(x1) + x2 + x3)
##
## 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(x1)      1.255e+00  3.323e-02  37.780  < 2e-16 ***
## x2          -2.980e-05  1.513e-05  -1.970   0.0495 *
## x3           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
```

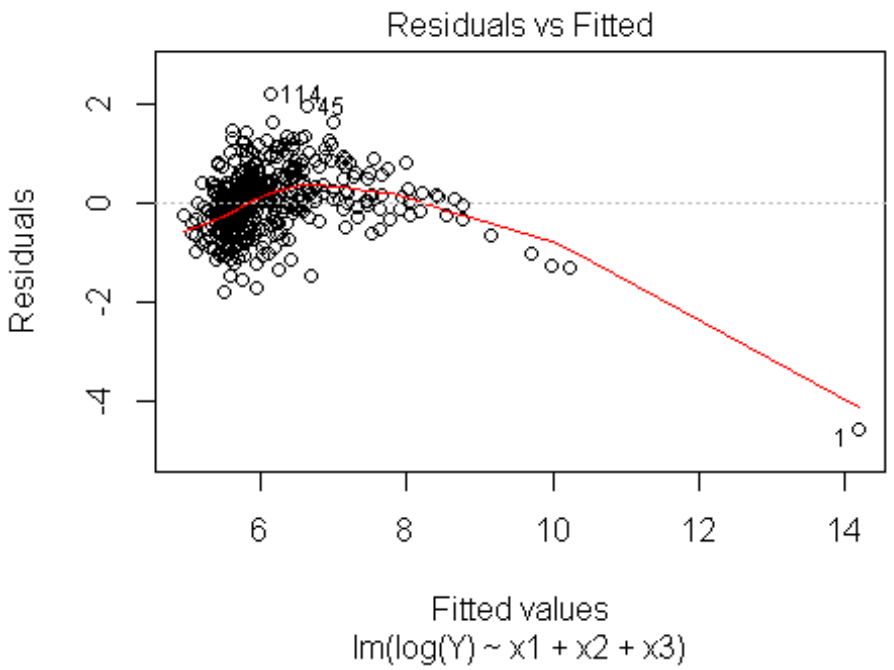
1d)
With 95% confidence, the expected median change in the number of physicians due to a 1% increase in total population lies within the interval: (1.01^1.190, 1.01^1.321)

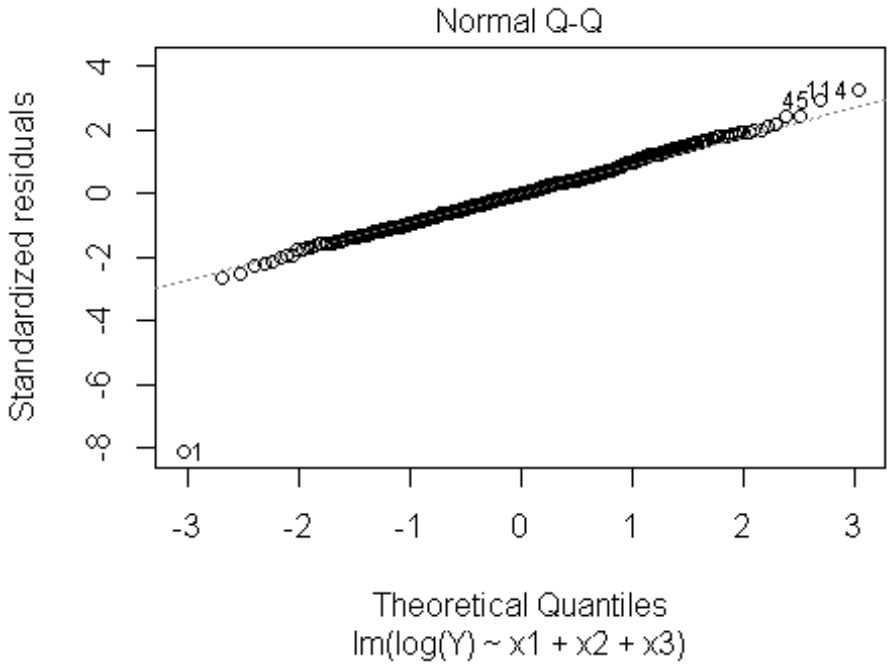
With 95% confidence, the expected median change in the number of physicians due to a 1 square mile increase in land area lies within the interval: (e^-5.953e-05, e^-6.440e-08)

With 95% confidence, the expected median change in the number of physicians due to a $1 increase in per capita income lies within the interval: (e^2.295e-05, e^4.766e-05)

Based on the summary function's global F-test, we have a test-statistic of 705.2 (null distribution of F with 3, 421 degrees of freedom) and p-value of 2.2e-16 < .01 = alpha when testing the null hypothesis that there are no useful predictors in the model (all predictor coefficients are 0 so that there is no linear relationship between any predictor and the transformed response when holding all other predictors constant). Hence, we reject the null hypothesis and accept the alternate hypothesis that there exists a linear relationship between at least one of the predictors and the transformed response while holding all other predictors constant.

```
#1e
cdi.lmY1 <- lm(log(Y)~x1+x2+x3) #no log of totalPop
plot(cdi.lmY1,which=c(1,2)) #check for higher or lower variance (relative to
taking log of totalPop)
```






```r
ncvTest(cdi.lmY,~log(x1)+x2+x3) #fail to reject null hypothesis of constant variance
```

```
## Non-constant Variance Score Test
## Variance formula: ~ log(x1) + x2 + x3
## Chisquare = 7.22885, Df = 3, p = 0.06495
```

1e)

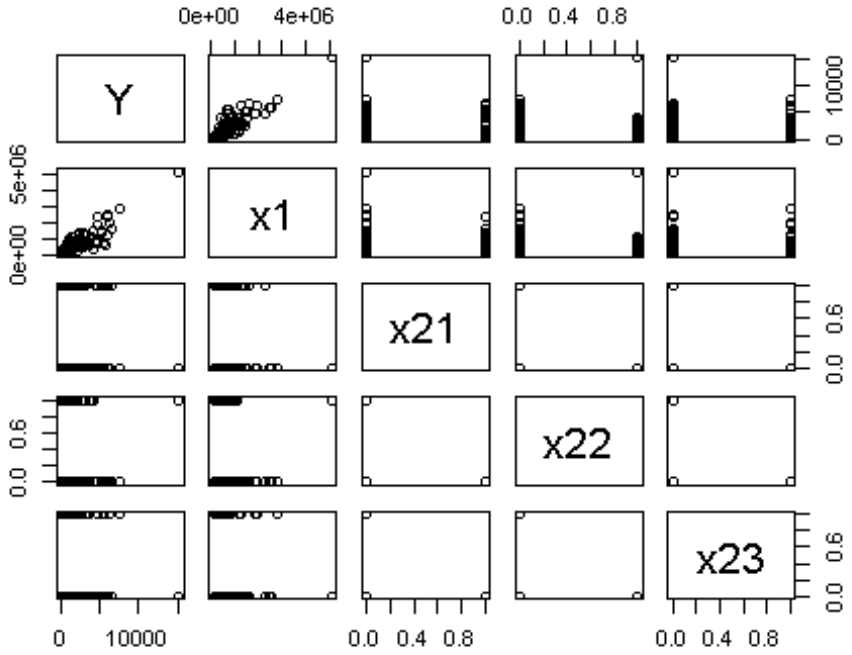
From comparing the residuals vs fits plots of two nearly identical models with TotalPop and the logarithm of TotalPop, we see that the variance decreases when we use log(TotalPop) instead of TotalPop. This is seen through the residuals vs fits plot: the range of the y-axis is smaller for the model with log(TotalPop), hence there is lower variance. We also conclude that the model's variance is constant using the ncv function since we observe a p-value greater than alpha of .05. Hence, we fail to reject the null hypothesis that the variance is constant.

1f)

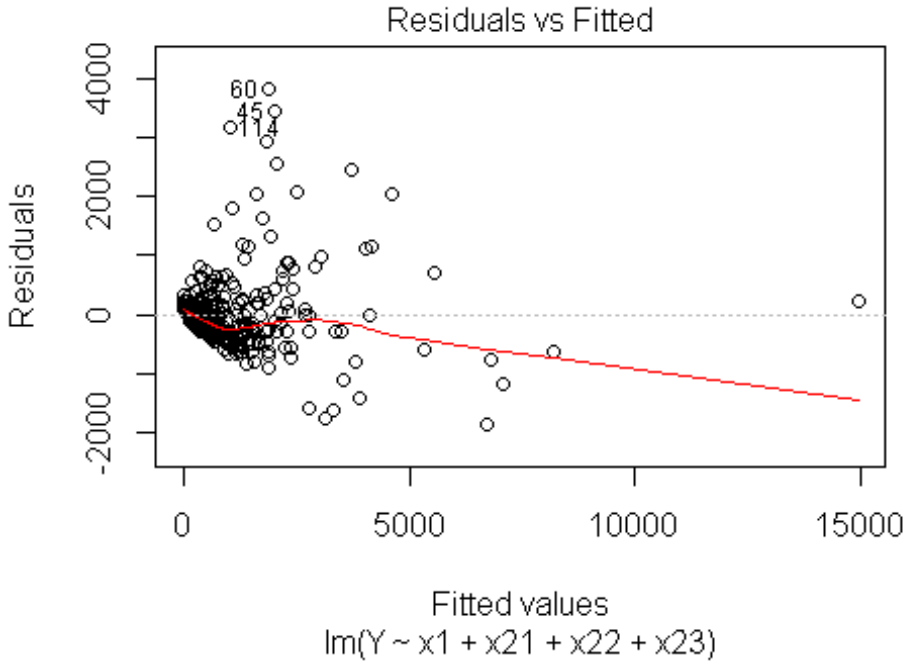
We choose our optimal model to be log(Physicians) = log(TotalPop) + LandArea + IncomePerCapita. It was interesting to see that normality, variance, and linearity were not significantly improved upon transforming LandArea and IncomePerCapita.

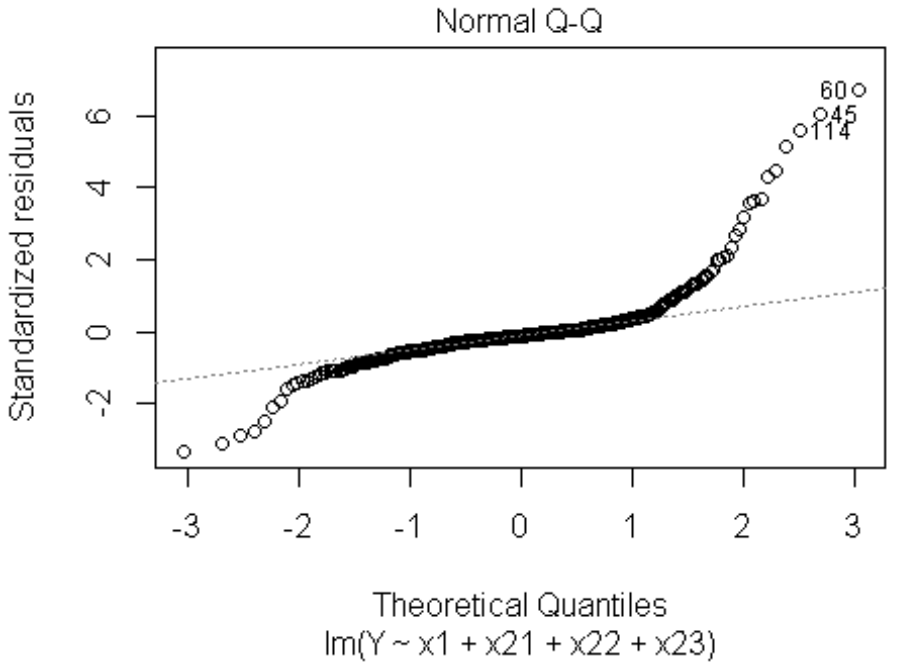
2.

```
CDI <- readRDS("C:/Users/Michael/Downloads/CDI.rds")

#2a
#original
Y <- CDI$Physicians
x1 <- CDI$TotalPop
x2 <- CDI$Region

x21 <- ifelse(as.character(x2)=='1',1,0)
x22 <- ifelse(as.character(x2)=='2',1,0)
x23 <- ifelse(as.character(x2)=='3',1,0)

cdi.lm1 <- lm(Y~x1+x21+x22+x23)
pairs(Y~x1+x21+x22+x23)
```




```
plot(cdi.lm1,which=c(1,2))
```

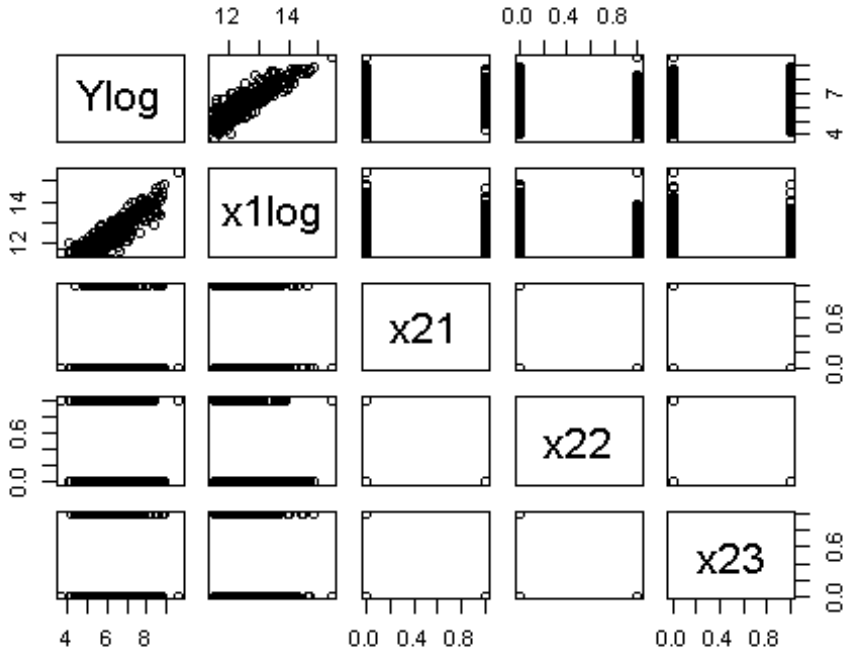




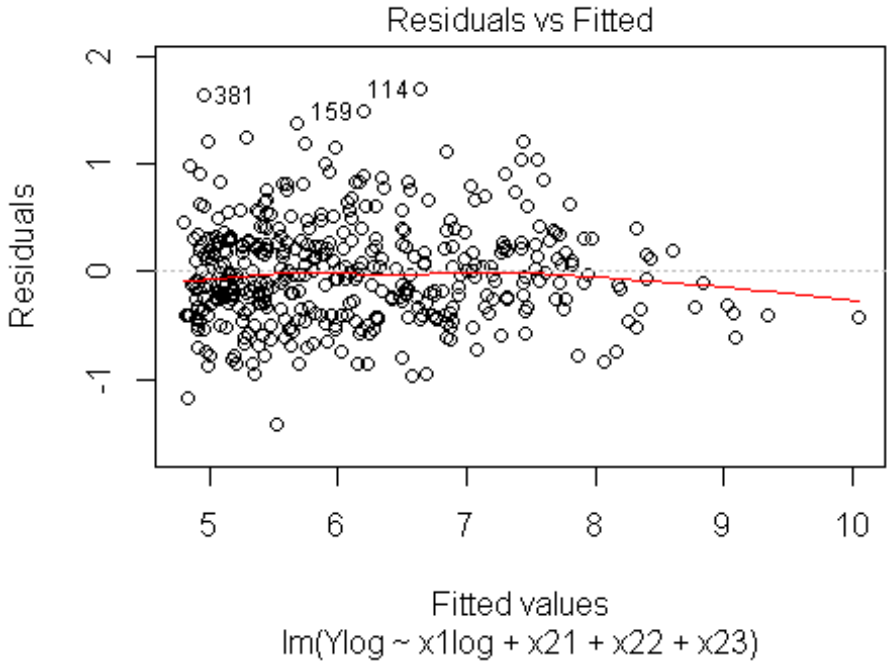

```
summary(cdi.lm1)

##
## Call:
```

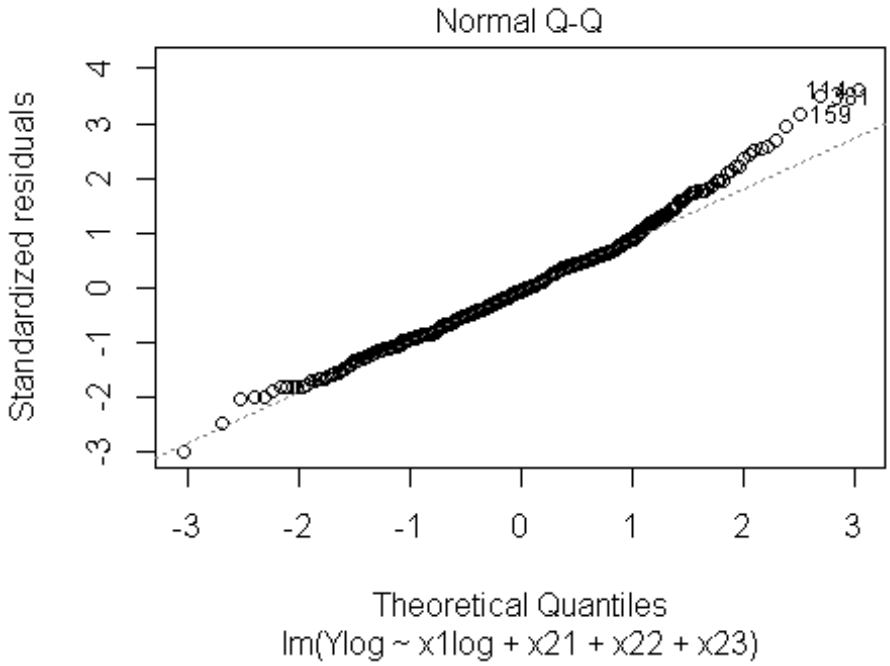
```
## lm(formula = Y ~ x1 + x21 + x22 + x23)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1845.2  -215.1   -67.5    96.0  3809.8
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.154e+02  7.214e+01  -4.373 1.55e-05 ***
## x1           2.958e-03  6.496e-05  45.542  < 2e-16 ***
## x21          2.156e+02  8.733e+01   2.469   0.0139 *
## x22          1.541e+02  8.694e+01   1.773   0.0770 .
## x23          1.507e+02  8.151e+01   1.848   0.0653 .
## ---
## 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
```

```
#check transformation
Ylog <- log(CDI$Physicians)
x1log <- log(CDI$TotalPop)

cdi.lm2 <- lm(Ylog~x1log+x21+x22+x23)

pairs(Ylog~x1log+x21+x22+x23)
```




```
plot(cdi.lm2,which=c(1,2))
```






```
summary(cdi.lm2)

##
## Call:
```

```
## lm(formula = Ylog ~ x1log + x21 + x22 + x23)
##
## 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) -10.54137    0.39652 -26.585   <2e-16 ***
## x1log         1.33167    0.03107  42.862   <2e-16 ***
## x21           0.12888    0.07256   1.776   0.0764 .
## x22           0.01655    0.07269   0.228   0.8200
## x23           0.10421    0.06791   1.535   0.1257
## ---
## 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

#improved indeed (log of phys and totalPop)
```

2a)

We applied a logarithmic transformation to the number of physicians and total population. By transforming these two variables, we are able to satisfy the assumptions of linear regression.

```
#2b
x3 <- CDI$PersonalInc
#new model
cdi.lm3 <- lm(Ylog~x1log+x21+x22+x23+x3)
summary(cdi.lm3)

##
## Call:
## lm(formula = Ylog ~ x1log + x21 + x22 + x23 + x3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.41560 -0.32621 -0.02898  0.26279  1.68414
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.088e+01  6.752e-01 -16.113   <2e-16 ***
## x1log        1.360e+00  5.578e-02  24.387   <2e-16 ***
## x21          1.300e-01  7.264e-02   1.790   0.0742 .
## x22          1.913e-02  7.286e-02   0.263   0.7930
## x23          1.039e-01  6.796e-02   1.529   0.1271
## x3          -2.712e-06  4.376e-06  -0.620   0.5358
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4751 on 419 degrees of freedom
## Multiple R-squared:  0.819,  Adjusted R-squared:  0.8169
## F-statistic: 379.3 on 5 and 419 DF,  p-value: < 2.2e-16

#log(Y1) = -10.88+1.36*x1+.13-.000002712*x3
#log(Y2) = -10.88+1.36*x1+.01913-.000002712*x3
#log(Y3) = -10.88+1.36*x1+.1039-.000002712*x3
#log(Y4) = -10.88+1.36*x1+.000002712*x3
```

2b)
E(lnY1) = (-10.88+.13) + 1.36x1 - .000002712x3 = *-10.75 + 1.36x1 - .000002712x3*
E(lnY2) = (-10.88+.01913) + 1.36x1 - .000002712x3 = *-10.86 + 1.36x1 - .000002712x3*
E(lnY3) = (-10.88+.1039) + 1.36x1-.000002712x3 = *-10.78 + 1.36x1-.000002712x3*
E(lnY4) = *-10.88 + 1.36x1 + .000002712x3*
Where Y = number of physicians, x1 = total population, x3 = personal income

These are parallel regression models because the coefficients of the predictors are the same for each equation, so they all have the same slope with different intercepts. Hence, these four regression lines are all parallel to each other.

```
#2c
#Is region significant?
cdi.lm4 <- lm(Ylog~x1log+x3)
cdi.lm5 <- lm(Ylog~x1log+x21+x22+x23+x3)
anova(cdi.lm4, cdi.lm5)

## Analysis of Variance Table
##
## Model 1: Ylog ~ x1log + x3
## Model 2: Ylog ~ x1log + x21 + x22 + x23 + x3
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    422 95.742
## 2    419 94.579  3    1.1632 1.7177 0.1627

#fail to reject reduced model, hence use reduced model (region has no
significant effect on number of physicians; don't include region)
#cdi.lm4 is new model now
```

2c) Upon using the ANOVA function to compare the model excluding region and including region, we see that the p-value is greater than .05, so we fail to reject the null hypothesis. Thus, we fail to conclude that there is a linear relationship between region and number of physicians while holding total population constant.

```
#2d
x4 <- CDI$Pop65
x5 <- CDI$Crimes
x6 <- CDI$Bachelor
x7 <- CDI$Poverty
```

```r
cdi.lm6 <- lm(Ylog~x1log+x3+x4+x5+x6+x7)
summary(cdi.lm6)
```

```
##
## Call:
## lm(formula = Ylog ~ x1log + x3 + x4 + x5 + x6 + x7)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.40378 -0.23261  0.02105  0.21865  1.32080
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.093e+01  5.404e-01 -20.224  < 2e-16 ***
## x1log        1.224e+00  4.578e-02  26.747  < 2e-16 ***
## x3          -2.726e-06  4.291e-06  -0.635    0.526
## x4           4.223e-02  5.044e-03   8.372 8.64e-16 ***
## x5          -1.696e-07  6.419e-07  -0.264    0.792
## x6           4.630e-02  3.147e-03  14.711  < 2e-16 ***
## x7           3.816e-02  4.673e-03   8.166 3.80e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3836 on 418 degrees of freedom
## Multiple R-squared:  0.8823, Adjusted R-squared:  0.8806
## F-statistic: 522.4 on 6 and 418 DF,  p-value: < 2.2e-16
```

```r
#no bueno: x3(personal income) x5(crimes) due to t-test
#thus, relevant predictors are pop65, bachelor, poverty
cdi.lm7 <- lm(Ylog~x1log+x4+x6+x7)
anova(cdi.lm4, cdi.lm7)
```

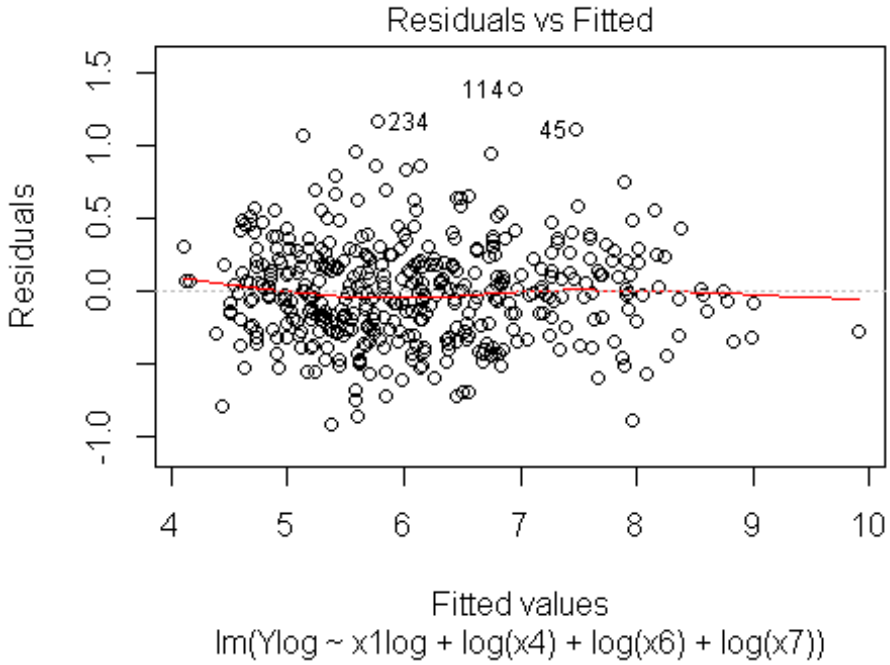
```
## Analysis of Variance Table
##
## Model 1: Ylog ~ x1log + x3
## Model 2: Ylog ~ x1log + x4 + x6 + x7
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    422 95.742
## 2    420 61.638  2    34.104 116.19 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

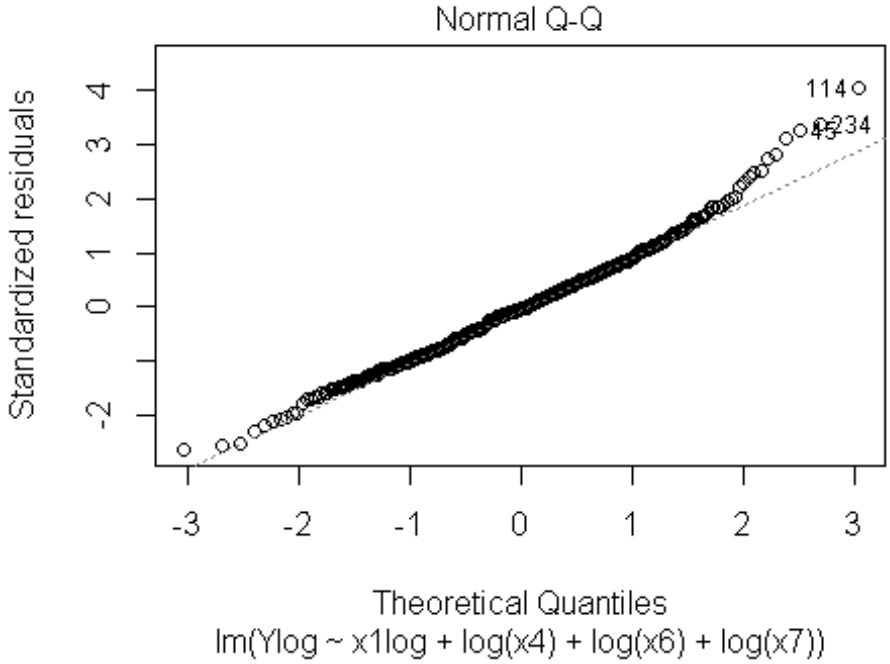
```r
#F-test says reject reduced for full mode
#cdi.lm7 is new model now

cdi.pt1 <- powerTransform(cbind(x4,x6,x7)~1) #check to see if we want
transformations for the new predictors
summary(cdi.pt1) #says to take log transformations
```

```
## bcPower Transformations to Multinormality
##    Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
```

```
## x4    0.0021           0      -0.1778       0.1820
## x6    0.0849           0      -0.1372       0.3071
## x7    0.1295           0      -0.0093       0.2684
##
## Likelihood ratio test that transformation parameters are equal to 0
##  (all log transformations)
##                                  LRT df    pval
## LR test, lambda = (0 0 0) 4.037954  3 0.25739
##
## Likelihood ratio test that no transformations are needed
##                                  LRT df       pval
## LR test, lambda = (1 1 1) 340.4401  3 < 2.22e-16
```

```
cdi.lm8 <- lm(Ylog~x1log+log(x4)+log(x6)+log(x7))
plot(cdi.lm8,which=c(1,2)) #we see that the normality got worse, so we do not
do any transformations
```






```r
#regsubsets test
mod.reg<-regsubsets(cbind(x1log,x3,x4,x5,x6,x7),Ylog)
summary.reg<-summary(mod.reg)
summary.reg$which
```

```
##   (Intercept) x1log    x3    x4    x5    x6    x7
## 1        TRUE  TRUE FALSE FALSE FALSE FALSE FALSE
## 2        TRUE  TRUE FALSE FALSE FALSE  TRUE FALSE
## 3        TRUE  TRUE FALSE FALSE FALSE  TRUE  TRUE
## 4        TRUE  TRUE FALSE  TRUE FALSE  TRUE  TRUE
## 5        TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE
## 6        TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
```

```
summary.reg$rsq#2
```

```
## [1] 0.8166034 0.8474865 0.8622058 0.8820612 0.8823147 0.8823343
```

```
summary.reg$cp#4
```

```
## [1] 230.505055 122.794847  72.505305   3.970053   5.069829   7.000000
```

```
summary.reg$adjr2#4
```

```
## [1] 0.8161698 0.8467636 0.8612239 0.8809380 0.8809103 0.8806453
```

```
summary.reg$bic#4
```

```
## [1] -708.7401 -781.0570 -818.1391 -878.2152 -873.0773 -867.0962
```

2d) From our testing, we conclude that the only additional useful predictors are Pop65, Bachelor, and Poverty. Each of these predictors have a significant effect on the number of physicians in the country. When comparing the new model to the old model, the resulting p-value is less than .05, so we reject the original model in favor of the new one.

```
#2e

#outlier
outlierTest(cdi.lm7)
```

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

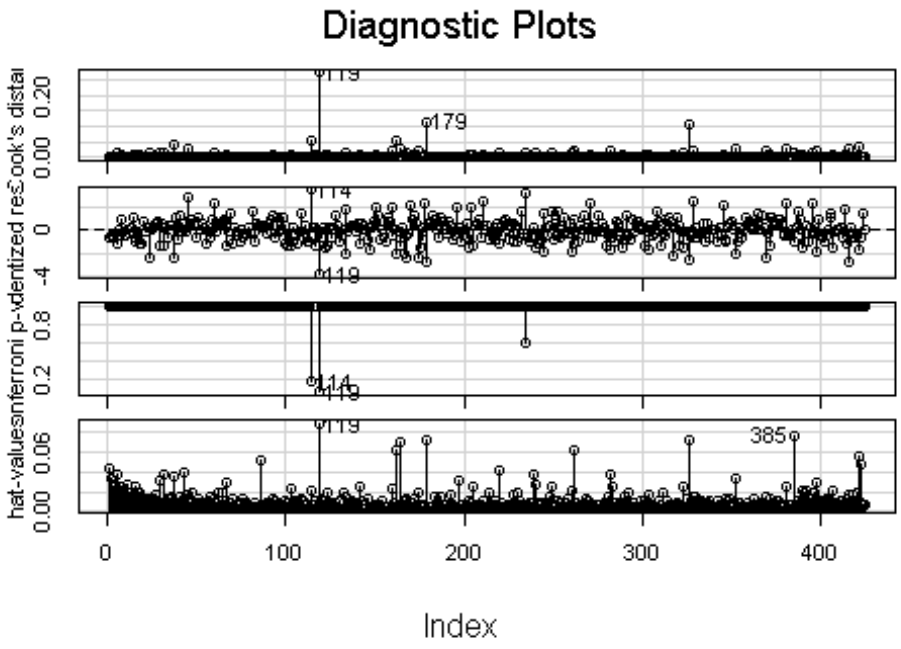
```
#high leverage
cdi.hats <- hatvalues(cdi.lm7)
n <- length(CDI$Physicians)
which(cdi.hats > 3*sum(cdi.hats)/n)
```

```
##   1   5  31  43  86 119 162 164 179 220 239 261 282 326 385 421 422
##   1   5  31  43  86 119 162 164 179 220 239 261 282 326 385 421 422
```

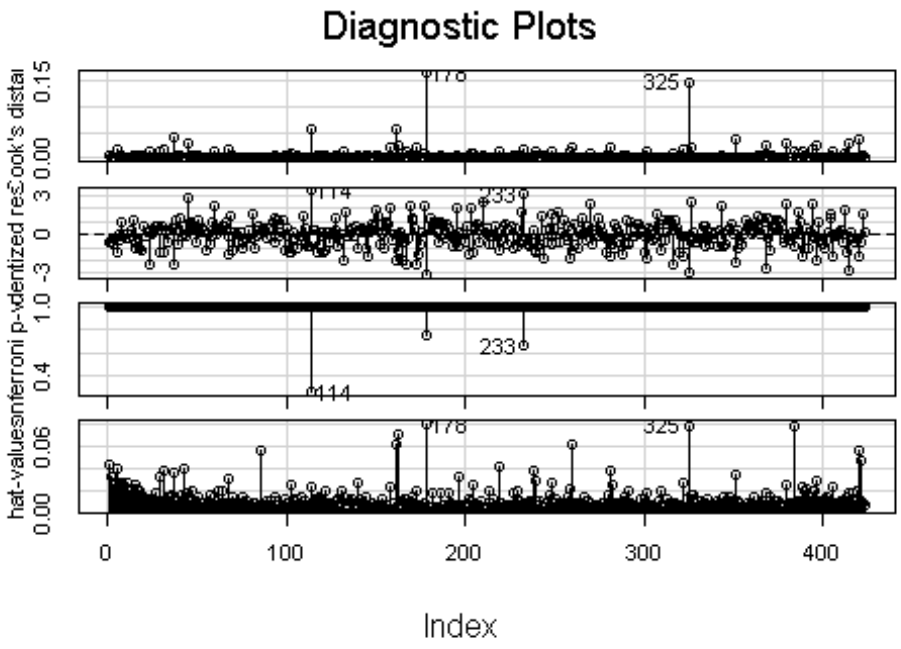
```
#high influential
cdi.cooks <- cooks.distance(cdi.lm7)
which(cdi.cooks > 4/(n-4-1)) #n-p-1
```

```
##   5  23  29  31  37  45  60  67 103 114 119 133 159 162 164 168 174 179
##   5  23  29  31  37  45  60  67 103 114 119 133 159 162 164 168 174 179
```

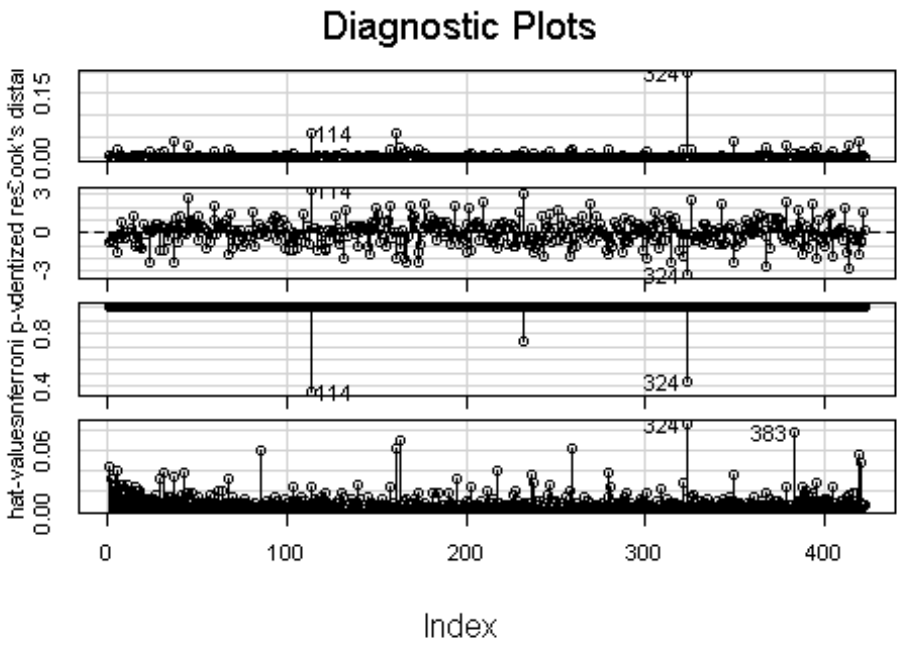
```
## 220 234 240 249 260 261 282 303 323 326 328 352 369 381 385 395 398 407
## 220 234 240 249 260 261 282 303 323 326 328 352 369 381 385 395 398 407
## 416 421
## 416 421

influenceIndexPlot(cdi.lm7)
```



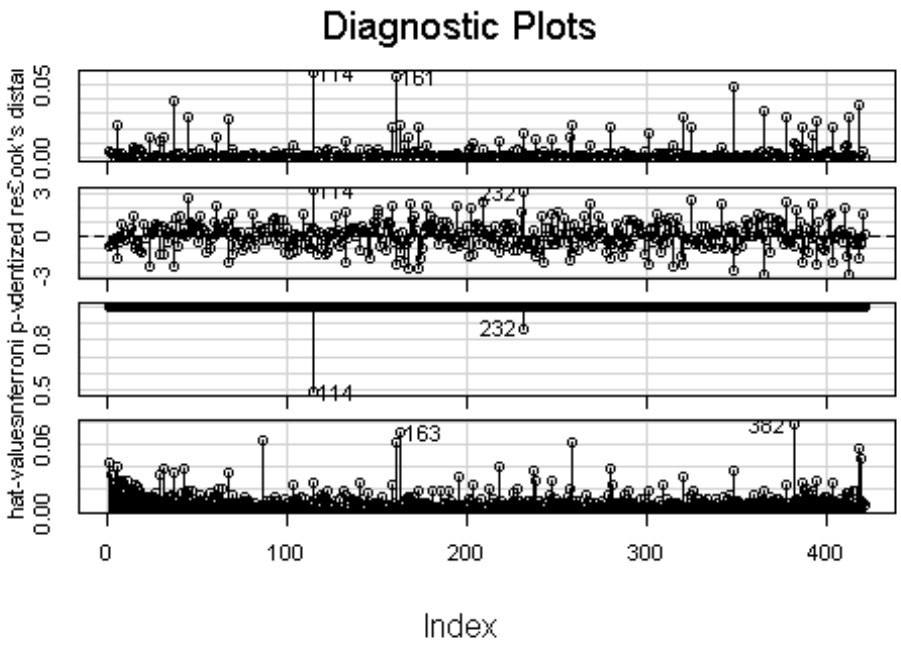

```
#remove 119 first

CDI.NEW <- CDI[-c(119),]

Ylog1 <- log(CDI.NEW$Physicians)
x1log1 <- log(CDI.NEW$TotalPop)
x41 <- CDI.NEW$Pop65
x61 <- CDI.NEW$Bachelor
x71 <- CDI.NEW$Poverty
cdi.lm9 <- lm(Ylog1~x1log1+x41+x61+x71)

influenceIndexPlot(cdi.lm9)
```




```
#remove 178 next

CDI.NEW1 <- CDI.NEW[-c(178),]

Ylog2 <- log(CDI.NEW1$Physicians)
x1log2 <- log(CDI.NEW1$TotalPop)
x42 <- CDI.NEW1$Pop65
x62 <- CDI.NEW1$Bachelor
x72 <- CDI.NEW1$Poverty
cdi.lm10 <- lm(Ylog2~x1log2+x42+x62+x72)

influenceIndexPlot(cdi.lm10)
```




```
#remove 324 next

CDI.NEW2 <- CDI.NEW1[-c(324),]

Ylog3 <- log(CDI.NEW2$Physicians)
x1log3 <- log(CDI.NEW2$TotalPop)
x43 <- CDI.NEW2$Pop65
x63 <- CDI.NEW2$Bachelor
x73 <- CDI.NEW2$Poverty
cdi.lm11 <- lm(Ylog3~x1log3+x43+x63+x73)

influenceIndexPlot(cdi.lm11)
```



2e)
In our model, the influential points are: 119 and 178 and 324 as shown by Cook's distance diagnostics (removing them one by one). These points show up on both the hat-values plot and studentized plot. A point that shows up on the hat-values plot means that it has high leverage, so the xi value is far away from all the other predictor values. Hence, the fitted Yi value plays a big role in predicting the actual Yi value. A point that shows up identified on the studentized plot means that it has a large studentized residual, hence the Yi for that data point is an outlier.

Since these points show up on all of these diagnostic plots, we remove them all. While it is possible to remove more points, we have determined 119, 178, and 324 to be the most influential (since they are all clearly both high-leverage and outlier points). Thus, we decided to remove only these three points since removing more points would be unnecessary.

2f)
We choose our optimal model to be log(Physicians) = log(TotalPop) + Pop65 + Bachelor + Poverty. It was interesting to see that normality got worse upon transforming the predictors. It was also interesting to decide when to stop removing outliers.

Conclusion

Our findings show that the given base models were insufficient in predicting the number of physicians in a county. We applied many transformations to make the model applicable for

linear regression. Our best model obtained through testing predicts the logarithm of Physicians using the logarithm of total population, percentage of the population 65 or older, percentage of the population with a bachelor's degree, and percentage of the population with income below the poverty level as predictors. The equation for this model is:

log(Physicians)~log(TotalPop)+Pop65+Bachelor+Poverty

This model is not entirely normal, so it does not perfectly predict the response. However, it is mostly accurate and we believe this model is the best for predicting the number of physicians in a given county.