PSTAT126 - Final Project

QirongHe

5/25/2019

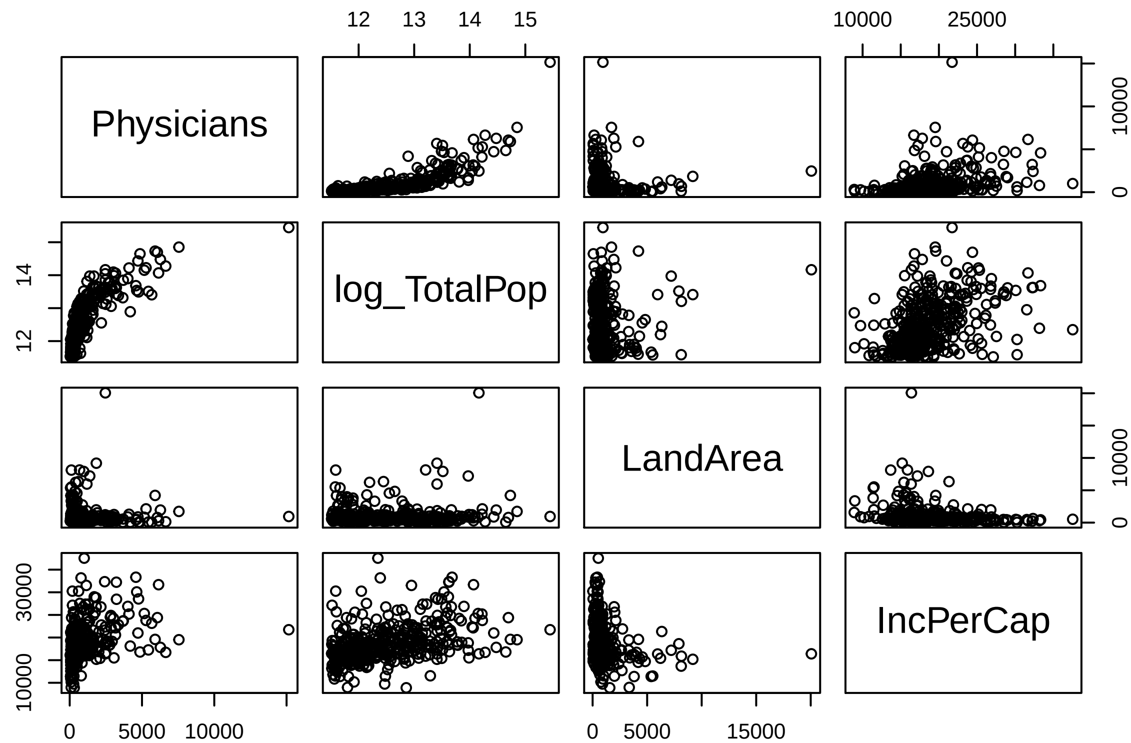
Problem 1

c) Do diagnostic checks to assess whether or not the linear regression assumptions seem to hold. If the model assumptions do not hold in your view, investigate possible transformations for predictors and/or response. Once suitable transformations are found, repeat b) for this new model and use this model for the remainder of Part I. Otherwise, move on to d).

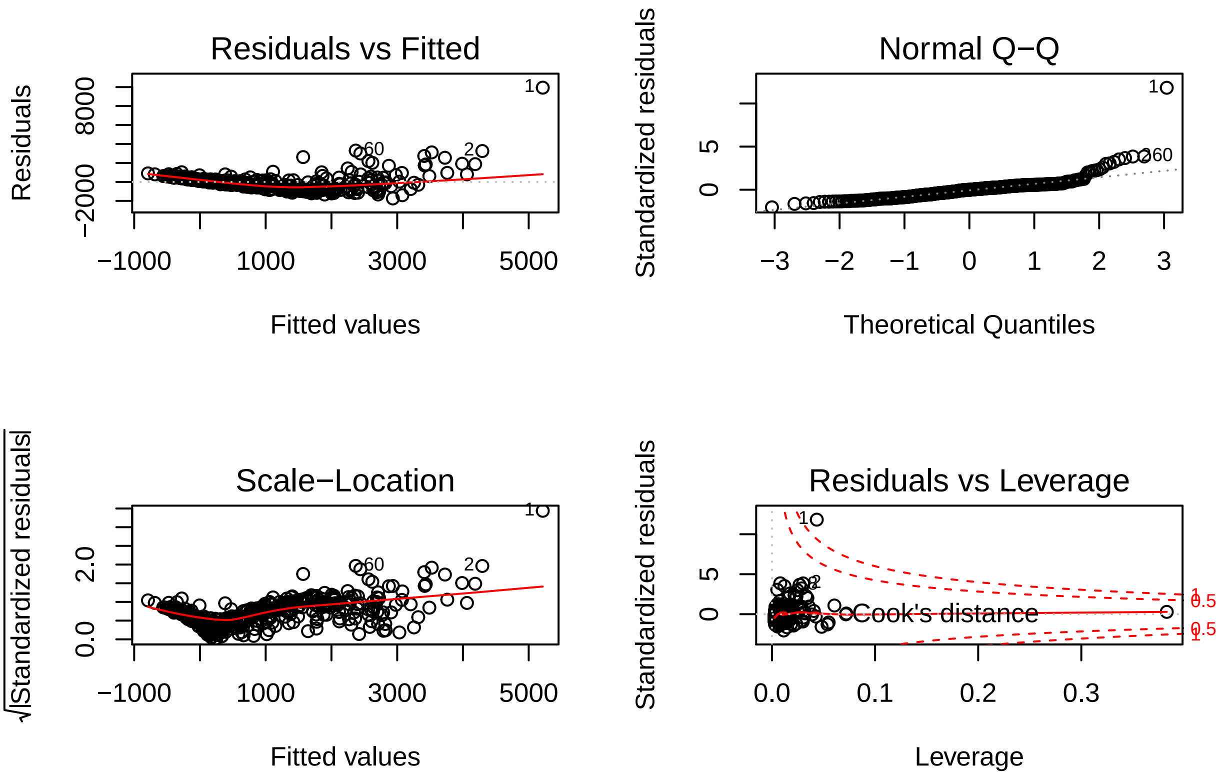
```r
library(readr)
CDI <- readRDS(file = '/Users/cheriehe/CDI.rds')
head(CDI)
```

```
##      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
```

```r
attach(CDI)
CDI$log_TotalPop = log(TotalPop)
pairs(CDI[c('Physicians','log_TotalPop','LandArea','IncPerCap')])
```



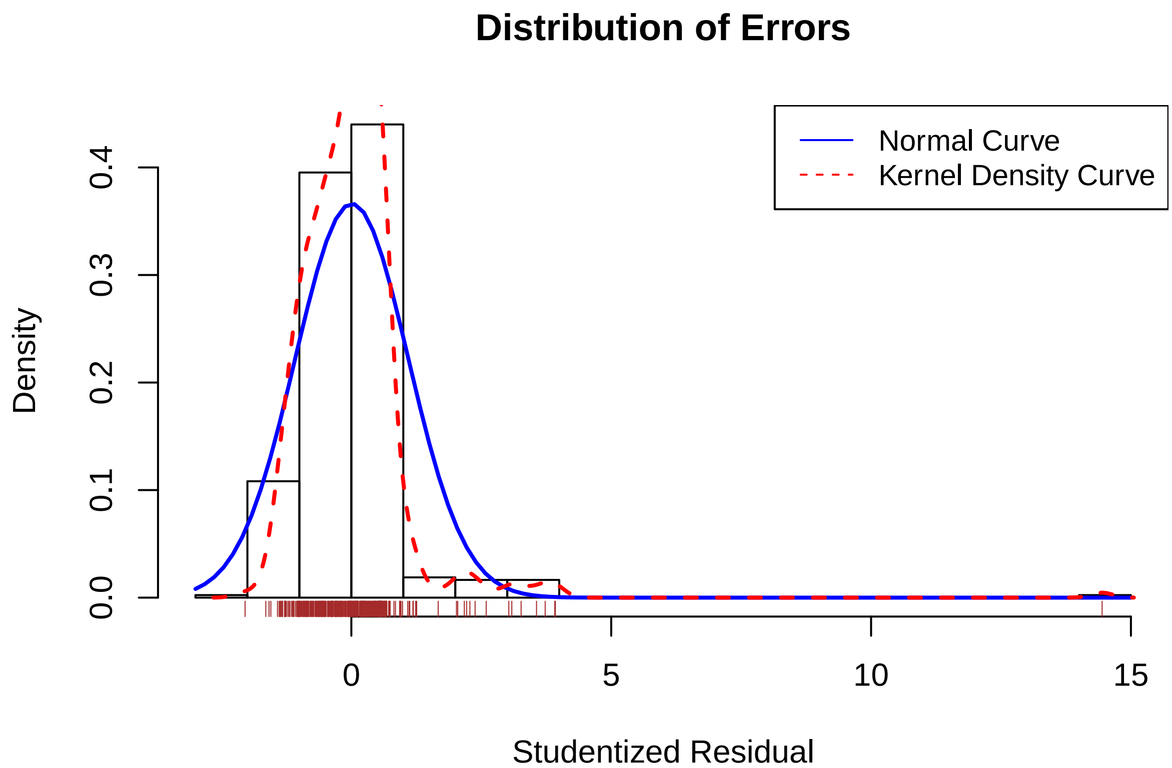

```
CDI.lm <- lm(Physicians~log(TotalPop)+LandArea+IncPerCap)
# do the diagnostic checks
par(mfrow = c(2,2))
plot(CDI.lm)
```



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 linaer pattern and don't bounce randomly around the 0-line. Ee 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 serve skewed, so it doesn't meet the normality assumption.

```
residplot <- function(CDI.lm, nbreaks=20) {
  z <- rstudent(CDI.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.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 meet 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.
- An observation with a high leverage value has an unusual combination of predictor values. That is, it's an outlier in the predictor space. The dependent variable value isn't used to calculate an observation's leverage.
- An influential observation is an observation that has a disproportionate impact on the determination of the model parameters. Influential observations are identified using a statistic called Cook's distance, or Cook's D.

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

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.

```r
library(car)
```

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

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

```
## bcPower Transformations to Multinormality
##           Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
##             -6.3082       -6.31      -7.7811      -4.8352
## LandArea    -0.0080        0.00      -0.0727       0.0567
## IncPerCap   -0.3166       -0.50      -0.5989      -0.0344
##
## Likelihood ratio test that transformation parameters are equal to 0
##  (all log transformations)
##                                 LRT df       pval
## LR test, lambda = (0 0 0) 63.94477  3 8.4377e-14
##
## Likelihood ratio test that no transformations are needed
##                                 LRT df       pval
## LR test, lambda = (1 1 1) 987.8428  3 < 2.22e-16
```

The columns labeled "Wald Lower Bound" and "Wald Upper Bound"" are the boundaries of 95% confidence intervals for the maximum likelhood power estimates. Intervals of LandArea contains 0, so we do the log transformation for LandArea. The likelihood ratio tests inidicate that using log transformations for IncPerCap is not appropriate, neither should we use no transformations.
As the above list indicates, I include the power of -0.5 for IncPerCap in the model and test whether it is useful.

```r
testTransform(pt, lambda = c(0, 0, -0.5))
```

```
##                                    LRT df       pval
## LR test, lambda = (0 0 -0.5) 61.41224  3 2.9343e-13
```

New fitted model:

```r
CDI_1.lm<-lm(Physicians~log(TotalPop)+log(LandArea)+ IncPerCap+I(IncPerCap^(-0.5)))
summary(CDI_1.lm)
```

```
##
## 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
```

```
confint(CDI_1.lm,level = 0.95)
```

```
##                             2.5 %        97.5 %
## (Intercept)         -2.504744e+04 -1.561971e+04
## log(TotalPop)        1.314980e+03  1.557324e+03
## log(LandArea)       -2.711777e+02 -6.949001e+01
## IncPerCap            2.086435e-03  1.555649e-01
## I(IncPerCap^(-0.5)) -7.428812e+03  8.099742e+05
```

From the summary we can see that the p-value of term $IncPerCap^{-0.5}$ is large and from the confidence interval we can see that the confidence interval of term $IncPerCap^{-0.5}$ contains 0, so we assume that term $IncPerCap^{-0.5}$ is not siginifiant. We will not include this term in our model.

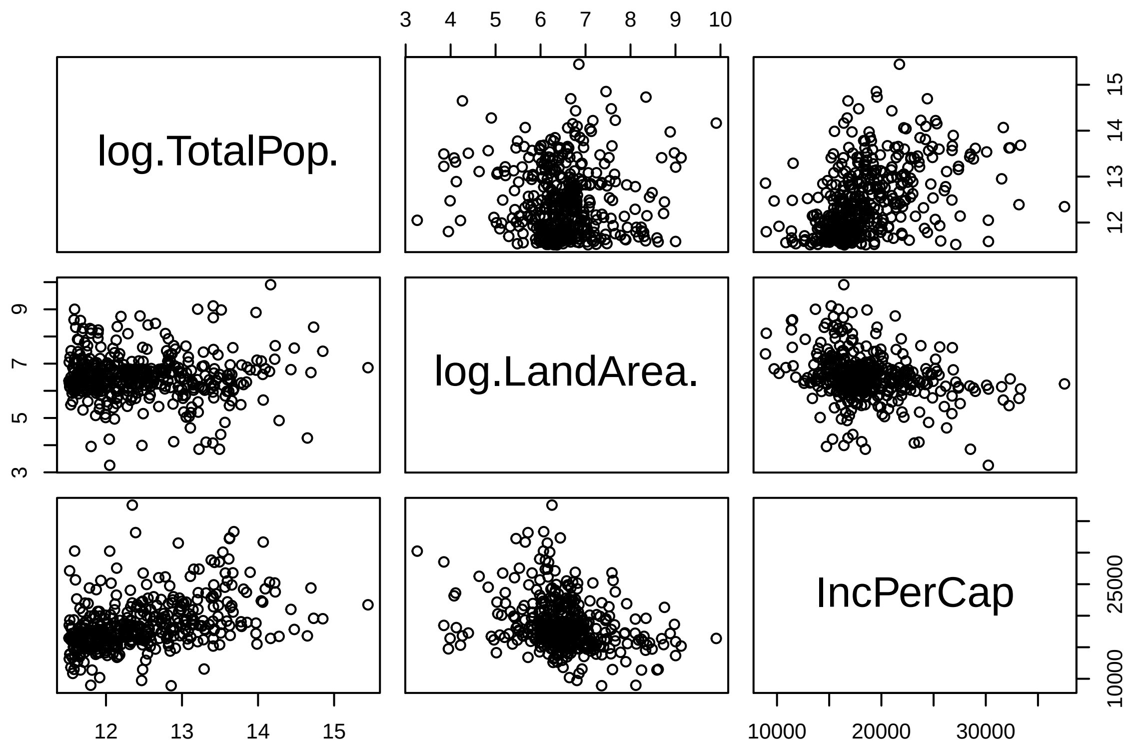
New our model is *Physicians = log(TotalPop) + log(LandArea) + IncPerCap*

```
testTransform(pt, lambda = c(0, 0, 1))
```

```
##                                   LRT df       pval
## LR test, lambda = (0 0 1) 137.2918  3 < 2.22e-16
```

We can not reject the null hyphothsis.

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



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

```
CDI_2.lm<-lm(Physicians~log(TotalPop)+log(LandArea)+ IncPerCap)
```

Conduct test:

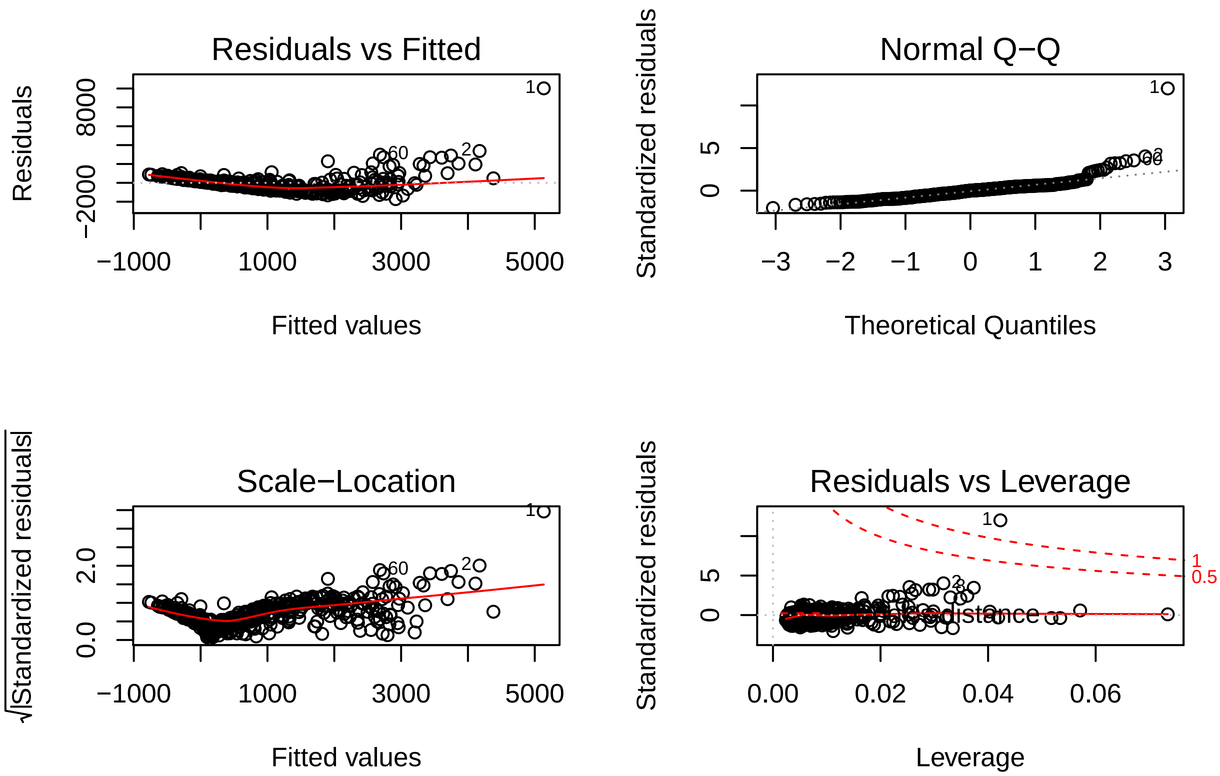
```
summary(CDI_2.lm)
```

```
##
## Call:
## lm(formula = Physicians ~ log(TotalPop) + log(LandArea) + IncPerCap)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1724.2  -487.1     4.8   381.2 10018.4
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -1.594e+04  7.634e+02 -20.885  < 2e-16 ***
## log(TotalPop)  1.426e+03  6.163e+01  23.142  < 2e-16 ***
## log(LandArea) -1.614e+02  5.126e+01  -3.149  0.00176 **
## IncPerCap      7.103e-03  1.200e-02   0.592  0.55413
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 853.4 on 421 degrees of freedom
## Multiple R-squared:  0.6257, Adjusted R-squared:  0.6231
## F-statistic: 234.6 on 3 and 421 DF,  p-value: < 2.2e-16
```

From the summary we can see that the p-value of IncPerCap is large, so model still seems don't work well.

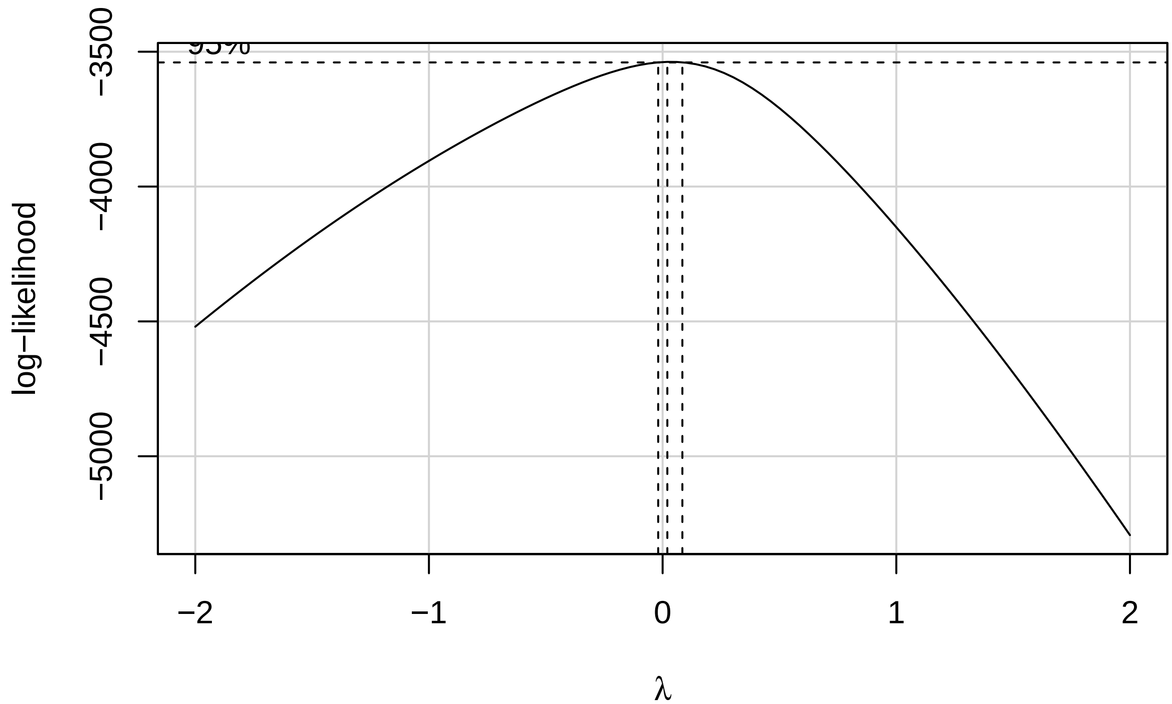
```r
par(mfrow = c(2,2))
plot(CDI_2.lm)
```



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

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

```r
bc <- boxCox(CDI_1.lm, 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*

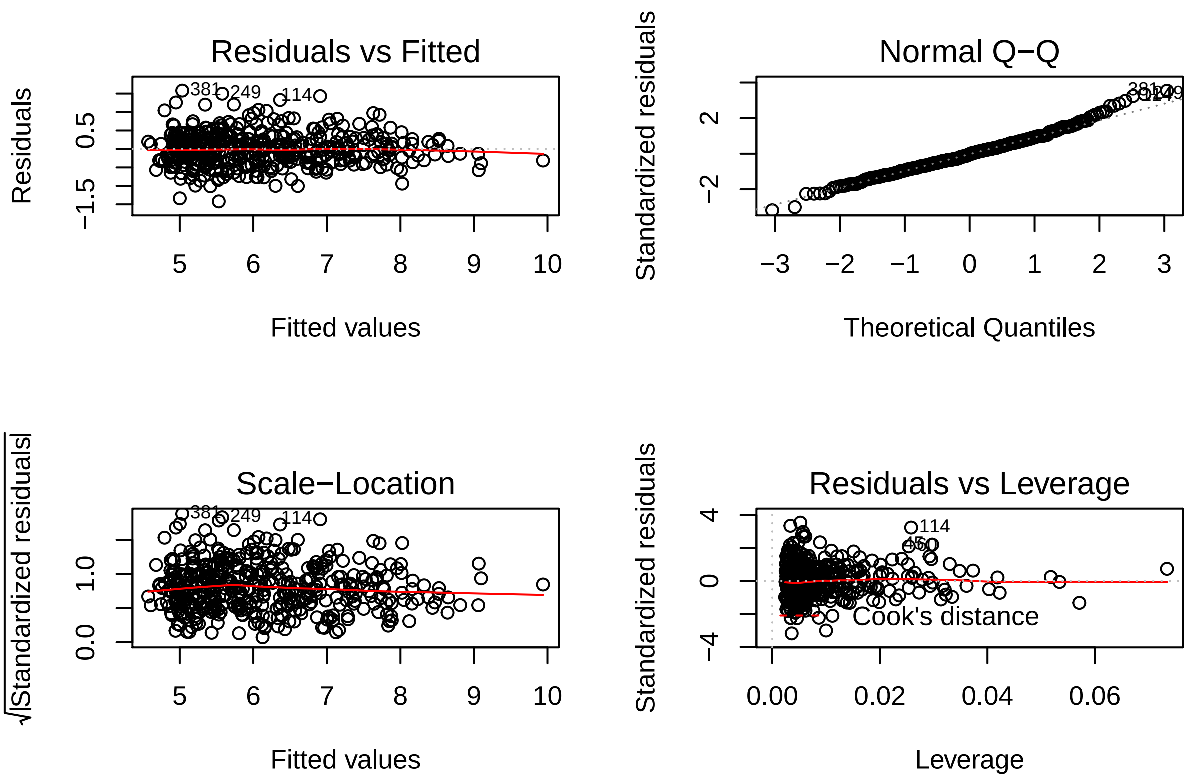
```
CDI_new.lm<-lm(log(Physicians)~log(TotalPop)+log(LandArea)+ IncPerCap)
summary(CDI_new.lm)
```

```
##
## 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
```

Now, all of the predictors seems useful. Now, we do the diagnostic checks

```r
par(mfrow = c(2,2))
plot(CDI_new.lm)
```



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. This suggests that the variances of the error terms are equal. Also the points in the Scale-Location graph (bottom left) random band around a horizontal line, so our constant variance hold well.

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

d) Using your fitted model, compute 95% confidence intervals for each of the coefficients in the model, and provide an interpretation for each. Conduct a test for the existence of a linear relationship between the predictors and response at $\alpha = 0.01$. Give the null and alternative hypotheses (defining any notation that you use), value of the test statistic and its null distribution, the p-value or critical value, and your decision.

The confidence interval:

```
confint(CDI_new.lm,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 varianle 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.

The null hypothesis and alternative for each test is:

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

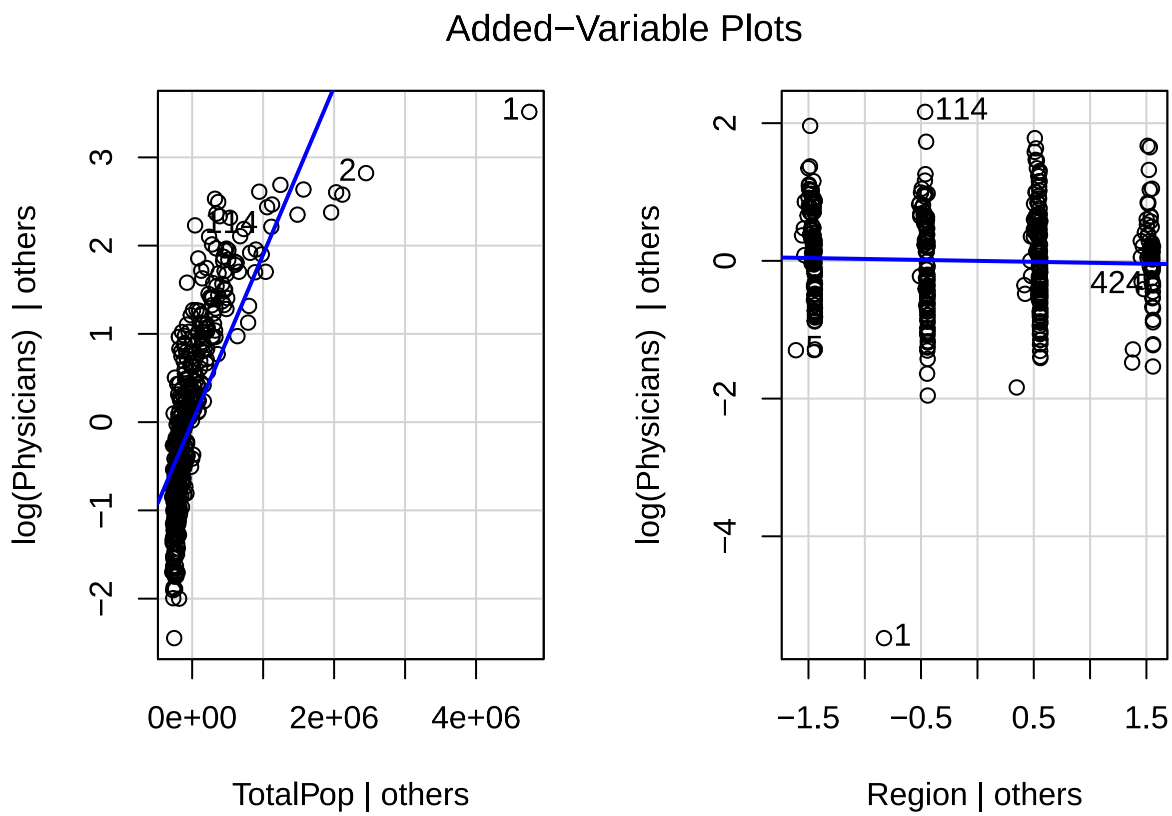
$H_0 : \beta_2 = 0$ vs $\beta_2 \neq 0$

$H_0 : \beta_3 = 0$ vs $\beta_2 \neq 0$

Problem 2

c) Does the geographic region have a significant effect on the number of physicians in a county? Explain your answer. If geographic region is not important, remove it from the model from now on.

```
CDI.lm2<-lm(log(Physicians)~TotalPop + Region)
avPlots(CDI.lm2,data = CDI)
```



From the added-variable plot we can see that the plot for TotalPop after Region shows that TotalPop is still correlated with log(Physicians) even after accounting for the effects of Region. However, that of Region after TotalPop shows that it is not useful when TotalPop is already in the model. So, we think geographic region is not important, and we'll remove it from the model from now on.

d) Use model selection techniques from class, build on your current model by selecting relevant predictors from Pop65, Crimes, Bachelor, Poverty, and PersonalInc. Perform a partial F-test to assess whether the improvement from adding these predictors compared to the first model is statistically significant at $\alpha = 0.05$.

```
CDI.0<-lm(log(Physicians)~1,data = CDI)
CDI.selfrom<-lm(log(Physicians) ~ Pop65 + Crimes + Bachelor + Poverty + PersonalInc)
```

Forward stepwise selection

```
step(CDI.0,scope = list(lower=CDI.0, upper=CDI.selfrom),
     direction ="forward")
```

```
## Start:  AIC=89.88
## log(Physicians) ~ 1
##
##               Df Sum of Sq    RSS      AIC
## + PersonalInc  1   289.439 233.19 -251.107
```

```
## + Crimes       1   173.811 348.81  -79.959
## + Bachelor     1   116.443 406.18  -15.248
## <none>                     522.62   89.879
## + Pop65        1     0.303 522.32   91.633
## + Poverty      1     0.014 522.61   91.868
##
## Step:  AIC=-251.11
## log(Physicians) ~ PersonalInc
##
##            Df Sum of Sq    RSS     AIC
## + Bachelor  1    32.291 200.90 -312.45
## + Poverty   1     2.935 230.25 -254.49
## <none>                  233.19 -251.11
## + Pop65     1     0.821 232.37 -250.60
## + Crimes    1     0.162 233.02 -249.40
##
## Step:  AIC=-312.46
## log(Physicians) ~ PersonalInc + Bachelor
##
##           Df Sum of Sq    RSS     AIC
## + Poverty  1   18.6389 182.26 -351.84
## + Pop65    1    9.8149 191.08 -331.74
## + Crimes   1    2.8559 198.04 -316.54
## <none>                 200.90 -312.46
##
## Step:  AIC=-351.84
## log(Physicians) ~ PersonalInc + Bachelor + Poverty
##
##          Df Sum of Sq    RSS     AIC
## + Pop65   1   15.1537 167.10 -386.73
## <none>                182.26 -351.84
## + Crimes  1    0.1266 182.13 -350.13
##
## Step:  AIC=-386.73
## log(Physicians) ~ PersonalInc + Bachelor + Poverty + Pop65
##
##          Df Sum of Sq    RSS     AIC
## <none>                167.10 -386.73
## + Crimes  1   0.35767 166.75 -385.64


##
## Call:
## lm(formula = log(Physicians) ~ PersonalInc + Bachelor + Poverty +
##     Pop65, data = CDI)
##
## Coefficients:
## (Intercept)  PersonalInc     Bachelor      Poverty        Pop65
##   3.142e+00    7.372e-05    6.306e-02    5.642e-02    5.097e-02
```

```
#use AIC by default
```

AIC=-386.73 is the smallest. We get the best model which is: *log(Physicians) ~ PersonalInc + Bachelor + Poverty + Pop65*

Backward stepwise selection

```
step(CDI.selfrom,scope = list(lower=CDI.0, upper=CDI.selfrom),
     direction ="backward")
```

```
## Start:  AIC=-385.64
## log(Physicians) ~ Pop65 + Crimes + Bachelor + Poverty + PersonalInc
##
##               Df Sum of Sq    RSS     AIC
## - Crimes       1     0.358 167.10 -386.73
## <none>                     166.75 -385.64
## - Pop65        1    15.385 182.13 -350.13
## - Poverty      1    20.154 186.90 -339.15
## - Bachelor     1    62.637 229.38 -252.10
## - PersonalInc  1    64.379 231.12 -248.88
##
## Step:  AIC=-386.73
## log(Physicians) ~ Pop65 + Bachelor + Poverty + PersonalInc
##
##               Df Sum of Sq    RSS     AIC
## <none>                     167.10 -386.73
## - Pop65        1    15.154 182.26 -351.84
## - Poverty      1    23.978 191.08 -331.74
## - Bachelor     1    62.329 229.43 -254.01
## - PersonalInc  1   186.180 353.28  -70.55

##
## Call:
## lm(formula = log(Physicians) ~ Pop65 + Bachelor + Poverty + PersonalInc)
##
## Coefficients:
## (Intercept)        Pop65     Bachelor      Poverty  PersonalInc
##   3.142e+00    5.097e-02    6.306e-02    5.642e-02    7.372e-05
```

AIC=-386.73 The model we will choose is that *log(Physicians) ~ Pop65 + Bachelor + Poverty + PersonalInc*

Stepwise stepwise selection

```
step(CDI.0,scope = list(lower=CDI.0, upper=CDI.selfrom),
     direction ="both")
```

```
## Start:  AIC=89.88
## log(Physicians) ~ 1
##
##               Df Sum of Sq    RSS      AIC
## + PersonalInc  1   289.439 233.19 -251.107
## + Crimes       1   173.811 348.81  -79.959
## + Bachelor     1   116.443 406.18  -15.248
## <none>                     522.62   89.879
## + Pop65        1     0.303 522.32   91.633
```

```
## + Poverty      1     0.014 522.61   91.868
##
## Step:  AIC=-251.11
## log(Physicians) ~ PersonalInc
##
##               Df Sum of Sq    RSS      AIC
## + Bachelor     1    32.291 200.90 -312.455
## + Poverty      1     2.935 230.25 -254.491
## <none>                     233.19 -251.107
## + Pop65        1     0.821 232.37 -250.605
## + Crimes       1     0.162 233.02 -249.402
## - PersonalInc  1   289.439 522.62   89.879
##
## Step:  AIC=-312.46
## log(Physicians) ~ PersonalInc + Bachelor
##
##               Df Sum of Sq    RSS     AIC
## + Poverty      1    18.639 182.26 -351.84
## + Pop65        1     9.815 191.08 -331.74
## + Crimes       1     2.856 198.04 -316.54
## <none>                     200.90 -312.46
## - Bachelor     1    32.291 233.19 -251.11
## - PersonalInc  1   205.286 406.18  -15.25
##
## Step:  AIC=-351.84
## log(Physicians) ~ PersonalInc + Bachelor + Poverty
##
##               Df Sum of Sq    RSS     AIC
## + Pop65        1    15.154 167.10 -386.73
## <none>                     182.26 -351.84
## + Crimes       1     0.127 182.13 -350.13
## - Poverty      1    18.639 200.90 -312.46
## - Bachelor     1    47.995 230.25 -254.49
## - PersonalInc  1   200.020 382.28  -39.03
##
## Step:  AIC=-386.73
## log(Physicians) ~ PersonalInc + Bachelor + Poverty + Pop65
##
##               Df Sum of Sq    RSS     AIC
## <none>                     167.10 -386.73
## + Crimes       1     0.358 166.74 -385.64
## - Pop65        1    15.154 182.26 -351.84
## - Poverty      1    23.978 191.08 -331.74
## - Bachelor     1    62.329 229.43 -254.01
## - PersonalInc  1   186.180 353.28  -70.55
```

```
##
## Call:
## lm(formula = log(Physicians) ~ PersonalInc + Bachelor + Poverty +
##     Pop65, data = CDI)
##
## Coefficients:
## (Intercept)  PersonalInc     Bachelor      Poverty        Pop65
##   3.142e+00    7.372e-05    6.306e-02    5.642e-02    5.097e-02
```

AIC=-386.73 The best model is: *log(Physicians) ~ PersonalInc + Bachelor + Poverty + Pop65*

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

Do the F-test for the null model and the model we choose by using predictor selection techniques.

```
choose.lm<-lm(log(Physicians) ~ PersonalInc + Bachelor + Poverty + Pop65)
anova(CDI.lm2,choose.lm)
```

```
## Analysis of Variance Table
##
## Model 1: log(Physicians) ~ TotalPop + Region
## Model 2: log(Physicians) ~ PersonalInc + Bachelor + Poverty + Pop65
##   Res.Df    RSS Df Sum of Sq     F    Pr(>F)
## 1    422 240.77
## 2    420 167.10  2    73.668 92.58 < 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 is selceted model is useful compare to the null model.