HW2 STA521 Fall18

[Wei Zhang, wz94 wzhang675]

Due September 19, 2018

Backgound Reading

Readings: Chapters 3-4 in Weisberg Applied Linear Regression

This exercise involves the UN data set from `alr3` package. Install `alr3` and the `car` packages and load the data to answer the following questions adding your code in the code chunks. Please add appropriate code to the chunks to suppress messages and warnings as needed once you are sure the code is working properly and remove instructions if no longer needed. Figures should have informative captions. Please switch the output to pdf for your final version to upload to Sakai. **Remove these instructions for final submission**

Exploratory Data Analysis

0. Preliminary read in the data. After testing, modify the code chunk so that output, messages and warnings are suppressed. *Exclude text from final*

```
library(alr3)
```

```
## Loading required package: car
```

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

```
data(UN3, package="alr3")
help(UN3)
library(car)
```

1. Create a summary of the data. How many variables have missing data? Which are quantitative and which are qualtitative?

```
summary(UN3)
```

```
##     ModernC          Change           PPgdp           Frate      
##  Min.   : 1.00   Min.   :-1.100   Min.   :   90   Min.   : 2.00  
##  1st Qu.:19.00   1st Qu.: 0.580   1st Qu.:  479   1st Qu.:39.50  
##  Median :40.50   Median : 1.400   Median : 2046   Median :49.00  
##  Mean   :38.72   Mean   : 1.418   Mean   : 6527   Mean   :48.31  
##  3rd Qu.:55.00   3rd Qu.: 2.270   3rd Qu.: 8461   3rd Qu.:58.00  
##  Max.   :83.00   Max.   : 4.170   Max.   :44579   Max.   :91.00  
##  NA's   :58      NA's   :1        NA's   :9       NA's   :43     
##       Pop              Fertility         Purban      
##  Min.   :      2.3   Min.   :1.000   Min.   :  6.00  
##  1st Qu.:    767.2   1st Qu.:1.897   1st Qu.: 36.25  
##  Median :   5469.5   Median :2.700   Median : 57.00  
##  Mean   :  30281.9   Mean   :3.214   Mean   : 56.20  
##  3rd Qu.:  18913.5   3rd Qu.:4.395   3rd Qu.: 75.00  
##  Max.   :1304196.0   Max.   :8.000   Max.   :100.00  
##  NA's   :2           NA's   :10
```

```
#We can see from the results that six variables have missing data.
#(All except Purban have missing data). All variables are quantitative.
```

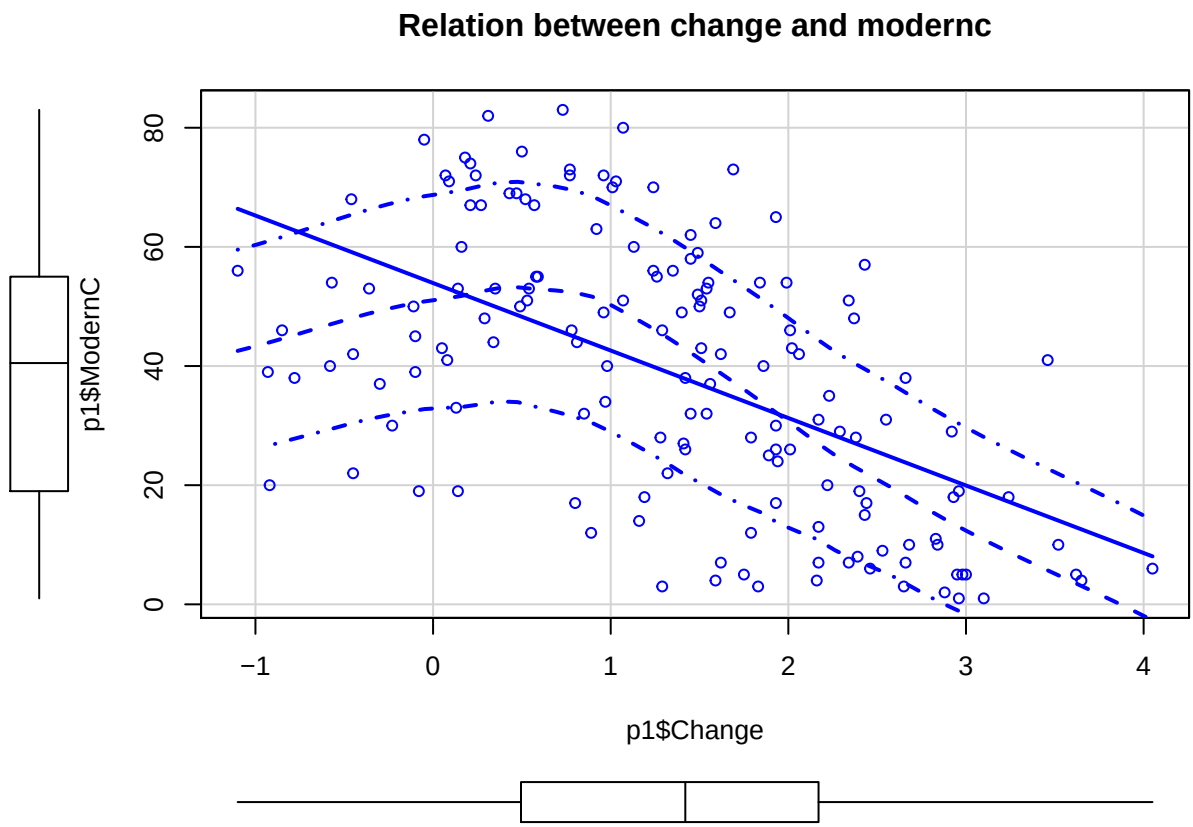
2. What is the mean and standard deviation of each quantitative predictor? Provide in a nicely formatted table.

```
library(knitr)
df1<-data.frame(matrix(ncol = 3, nrow = 0))
df1<-rbind(df1,data.frame(t(c("ModernC",mean(na.omit(UN3$ModernC)),sd(na.omit(UN3$ModernC))))))
df1<-rbind(df1,data.frame(t(c("Change",mean(na.omit(UN3$Change)),sd(na.omit(UN3$Change))))))
df1<-rbind(df1,data.frame(t(c("PPgdp",mean(na.omit(UN3$PPgdp)),sd(na.omit(UN3$PPgdp))))))
df1<-rbind(df1,data.frame(t(c("Frate",mean(na.omit(UN3$Frate)),sd(na.omit(UN3$Frate))))))
df1<-rbind(df1,data.frame(t(c("Pop",mean(na.omit(UN3$Pop)),sd(na.omit(UN3$Pop))))))
df1<-rbind(df1,data.frame(t(c("Fertility",mean(na.omit(UN3$Fertility)),sd(na.omit(UN3$Fertility))))))
df1<-rbind(df1,data.frame(t(c("Purban",mean(na.omit(UN3$Purban)),sd(na.omit(UN3$Purban))))))
varna1<-c("Variable name","mean", "std")
colnames(df1)<-varna1
kable(df1)
```

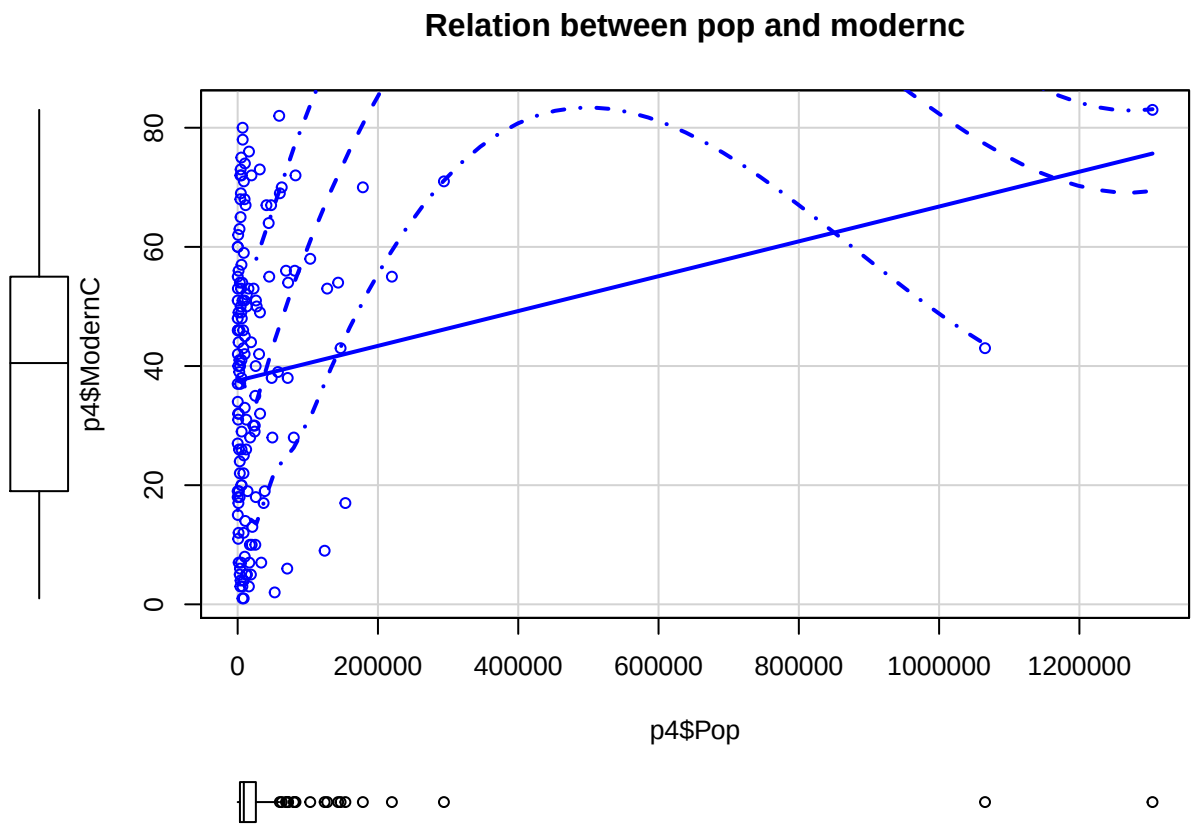
Variable name	mean	std
ModernC	38.7171052631579	22.6366103759673
Change	1.41837320574163	1.13313267030361
PPgdp	6527.38805970149	9325.18855244529
Frate	48.3053892215569	16.5324480416909
Pop	30281.8714278846	120676.694478229
Fertility	3.214	1.70691793716661
Purban	56.2	24.1097570036514

3. Investigate the predictors graphically, using scatterplots or other tools of your choice. Create some plots highlighting the relationships among the predictors. Comment on your findings regarding trying to predict `ModernC` from the other variables. Are there potential outliers, nonlinear relationships or transformations that appear to be needed based on your graphical EDA?

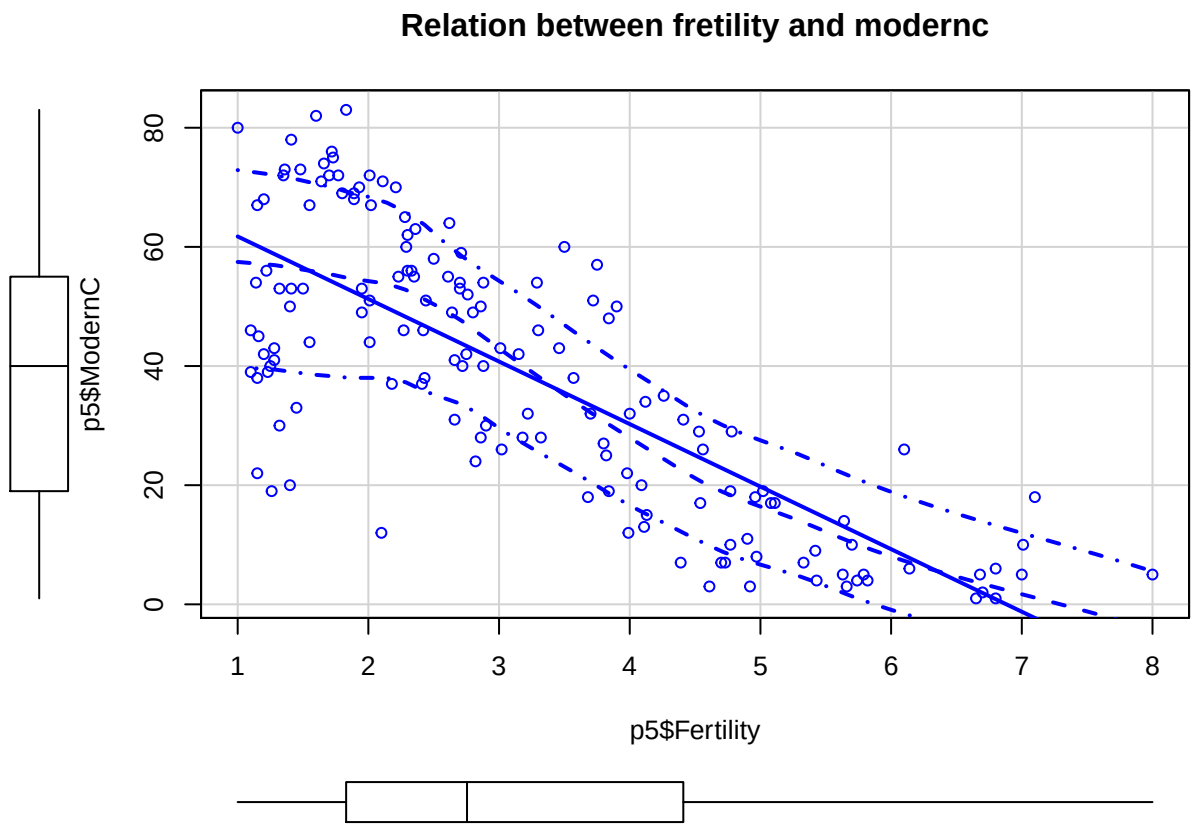
```
p1<-na.omit(subset(UN3,select=c(Change,ModernC)))
scatterplot(p1$Change,p1$ModernC,main="Relation between change and modernc")
```



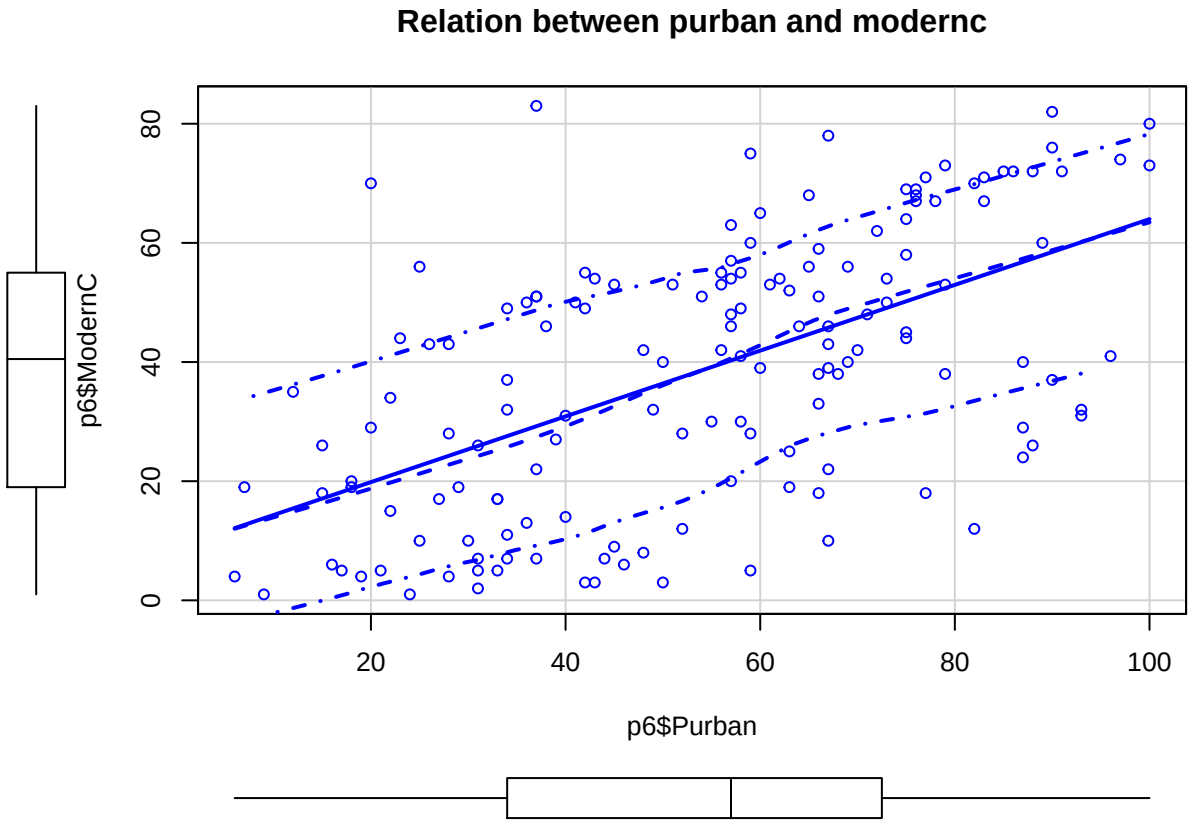

```
#p2<-na.omit(subset(UN3,select=c(PPgdp,ModernC)))
#scatterplot(p2$PPdgp,p2$ModernC,main="Relation between ppgdp and modernc")
p4<-na.omit(subset(UN3,select=c(Pop,ModernC)))
scatterplot(p4$Pop,p4$ModernC,main="Relation between pop and modernc")
```




```
p5<-na.omit(subset(UN3,select=c(Fertility,ModernC)))
scatterplot(p5$Fertility,p5$ModernC,main="Relation between fretility and modernc")
```




```
p6<-na.omit(subset(UN3,select=c(Purban,ModernC)))
scatterplot(p6$Purban,p6$ModernC,main="Relation between purban and modernc")
```

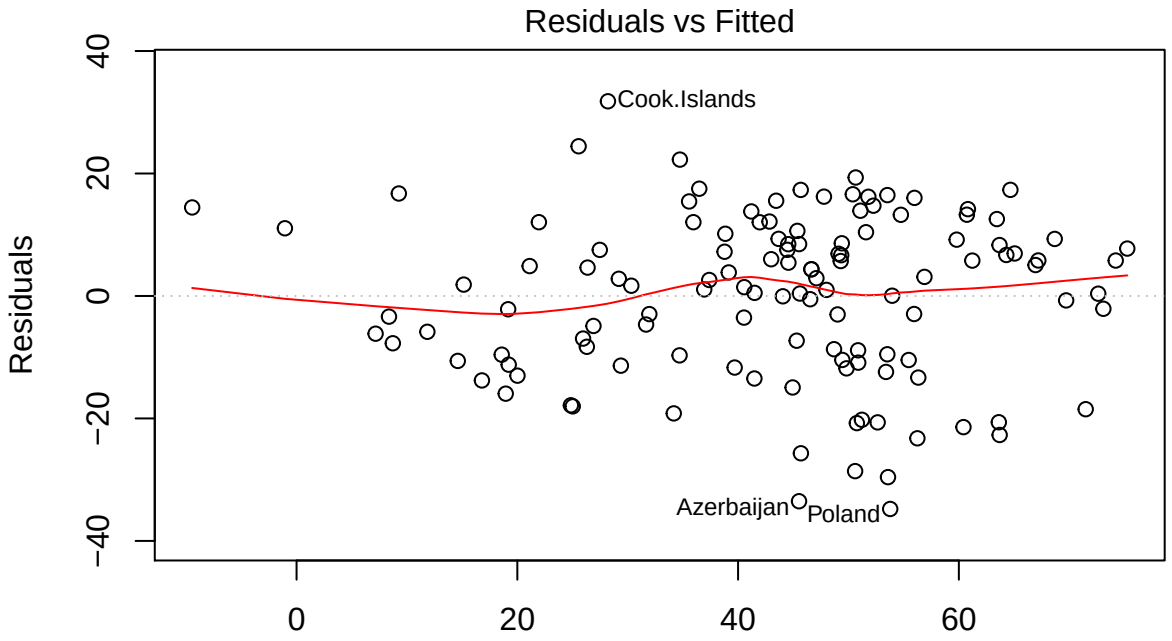


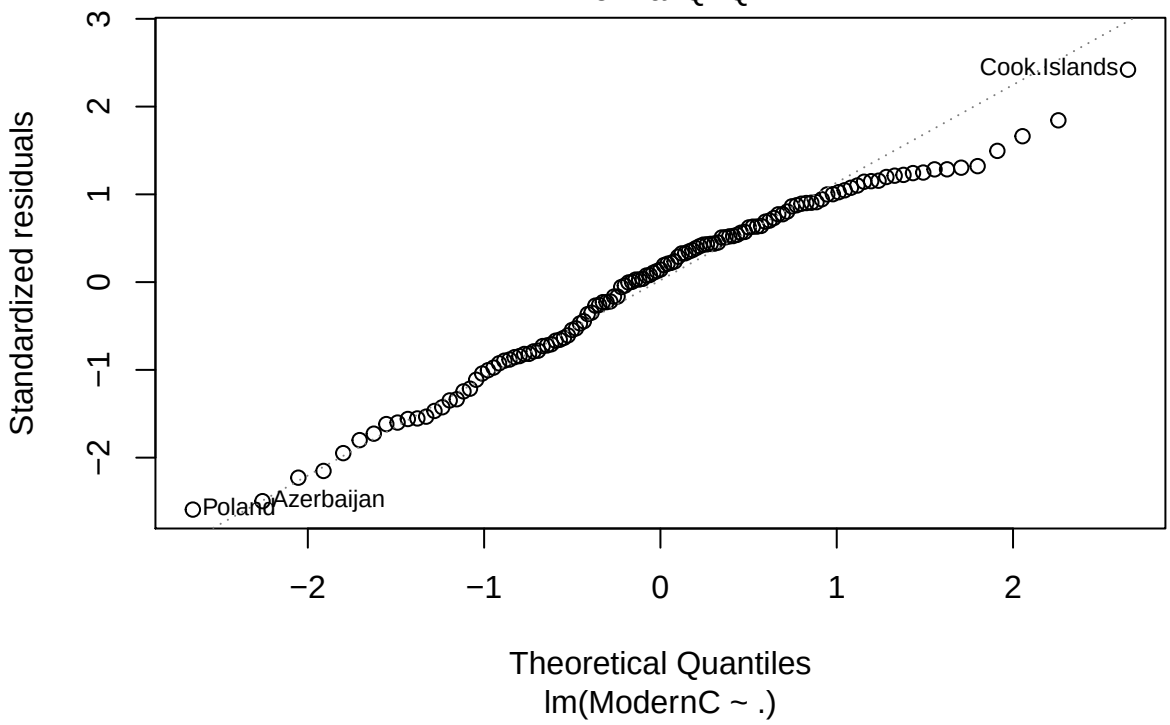

```
#It seems a lot of variables have the potential to explain Modernc.
#we can see from the graph that there might be some outliers in pop
#variable. The realtion of Change and ModernC might not be linear.
#And we may need to transform Pop, Fertility and Change variables.
```

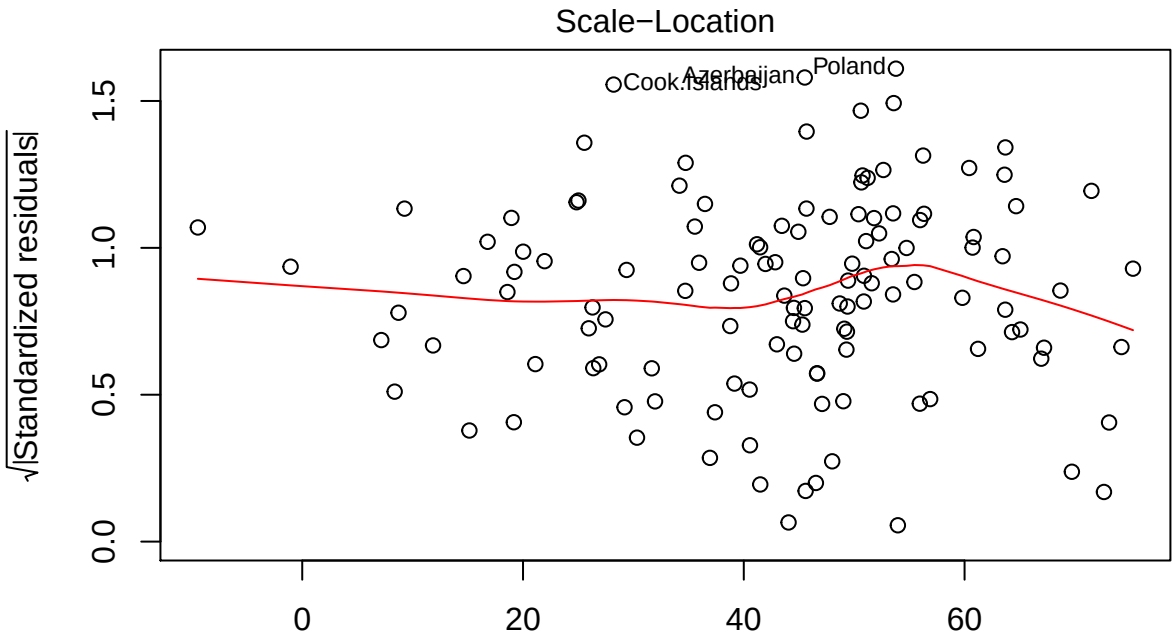
Model Fitting

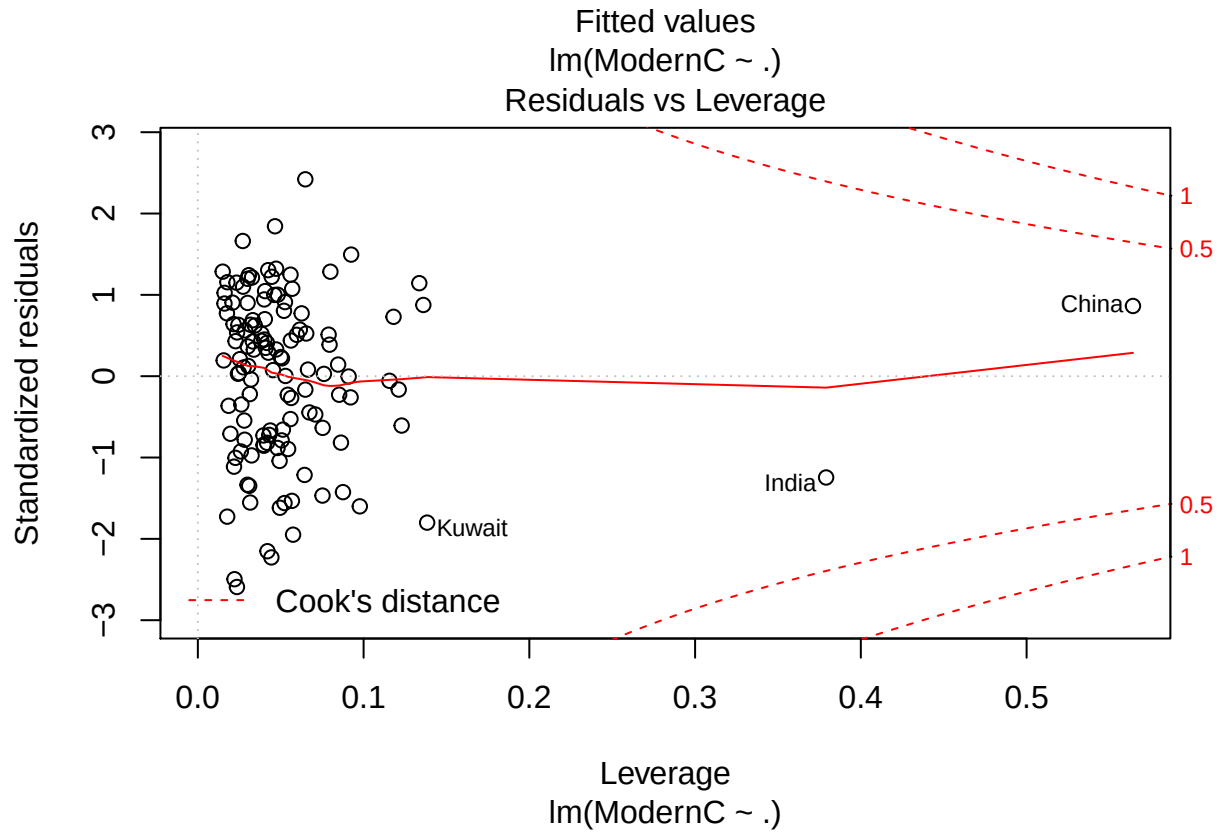
4. Use the `lm()` function to perform a multiple linear regression with `ModernC` as the response and all other variables as the predictors, using the formula `ModernC ~ .`, where the `.` includes all remaining variables in the dataframe. Create diagnostic residual plot from the linear model object and comment on results regarding assumptions. How many observations are used in your model fitting?

```
model1=lm(ModernC~.,data=UN3);
plot(model1)
```


Residuals vs Fitted
Residuals
Fitted values
lm(ModernC ~ .)
Cook.Islands
Azerbaijan
Poland
40
20
0
−20
−40
0
20
40
60


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(ModernC ~ .)
Cook.Islands
Azerbaijan
Poland
3
2
1
0
−1
−2
−2
−1
0
1
2






```
summary(model1)
```

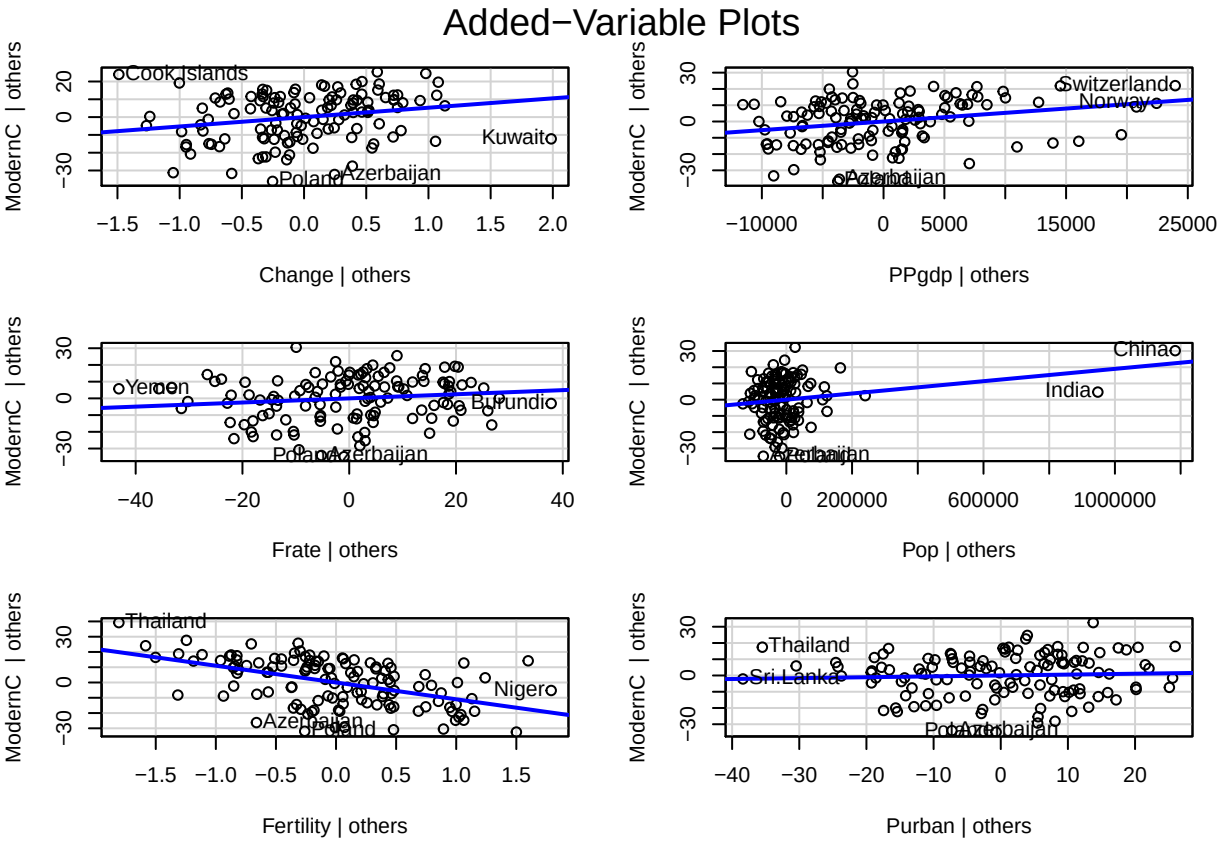
```
##
## Call:
## lm(formula = ModernC ~ ., data = UN3)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -34.781  -9.698   1.858   9.327  31.791
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.529e+01  9.467e+00   5.841 4.69e-08 ***
## Change       5.268e+00  2.088e+00   2.524  0.01294 *
## PPgdp        5.301e-04  1.770e-04   2.995  0.00334 **
## Frate        1.232e-01  8.060e-02   1.529  0.12901
## Pop          1.899e-05  8.213e-06   2.312  0.02250 *
## Fertility   -1.100e+01  1.752e+00  -6.276 5.96e-09 ***
## Purban       5.408e-02  9.285e-02   0.582  0.56134
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.58 on 118 degrees of freedom
##   (85 observations deleted due to missingness)
## Multiple R-squared:  0.6183, Adjusted R-squared:  0.5989
## F-statistic: 31.85 on 6 and 118 DF,  p-value: < 2.2e-16
```

```
#From residual vs Fitted and Scale-location graphs, we can see that
#the the variance seems constant for different values. So the
#assumption that constant variance seems hold.
#From Normal Q-Q graph we find that there are lots of points donot
#lie in the diagonal(the theoretical line), which means the normailty
#assumption might be violated. From the Residuals vs leverage graph
#we find that all points has cook disntance less than .5. So none
#of them are quiet influencial.
#We have 125 observations in the modeling.This is because some obs
#was deleted from original data due to missingness.
```

5. Examine added variable plots `car::avPlot` or `car::avPlots` for your model above. Are there any plots that suggest that transformations are needed for any of the terms in the model? Describe. Is it likely that any of the localities are influential for any of the terms? Which localities? Which terms?

```
car::avPlots(model1)
```




```
#
#Yes,for Fertility, PPgdp, and Pop variables, I think we should
#do transformation. For Pop variable, china and india are influential
#and it seems to be good if we do log or other transformation.
#For Fertility, It seems to have some kind of convexity and I feel
#like the relation is not linear and it seems to have expoential
#relaiton. So I want to transform it as well. Same resaon hold for
#PPgdp, as it seems not linear.
```

6. Using the Box-Tidwell `car::boxTidwell` or graphical methods find appropriate transformations of the predictor variables to be used as predictors in the linear model. If any predictors are negative, you may need to transform so that they are non-negative. Describe your method and the resulting transformations.

```
UN3=na.omit(UN3)
boxTidwell(ModernC~Pop+Fertility+PPgdp,~Change+Purban+Frate,data=UN3)
```

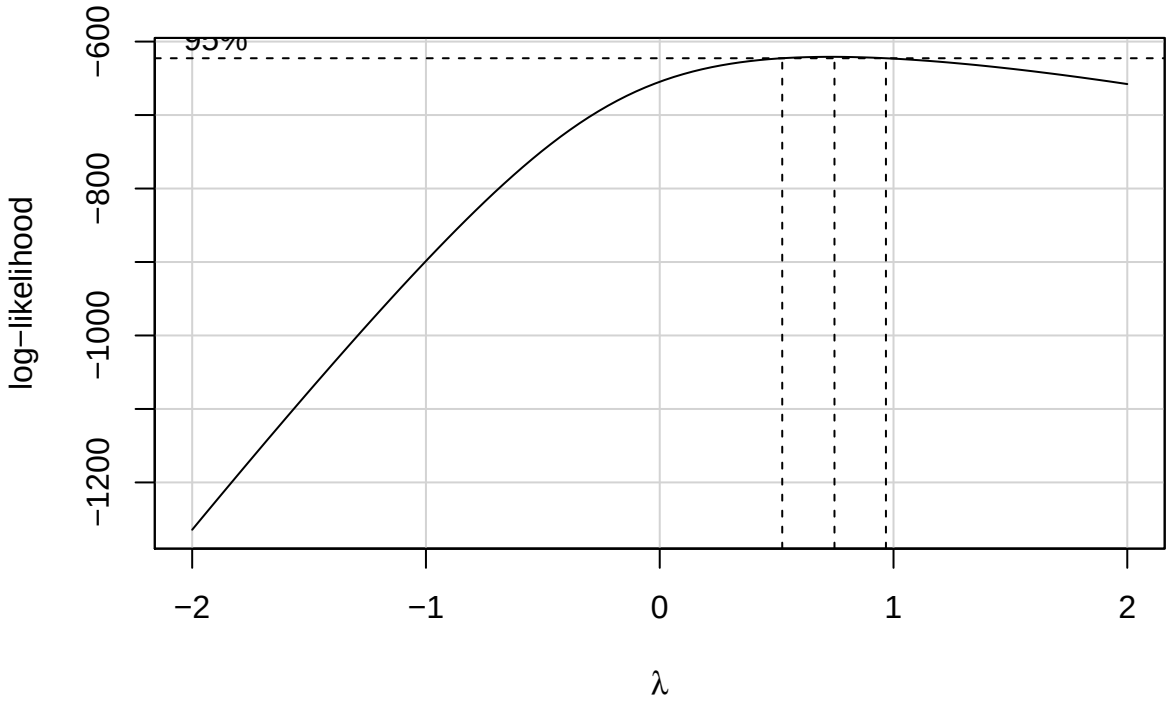
```
##            MLE of lambda Score Statistic (z) Pr(>|z|)
## Pop             0.374984             -0.9042   0.3659
## Fertility       1.346874             -1.7985   0.0721 .
## PPgdp          -0.035767             -1.2324   0.2178
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## iterations =  22
```

```
#As I mentiond before, I think Pop, Fertility and PPgdp are varialbes
#that we need to transform. So I put those in the set to
#transform and put others in the set that we do not need to transform
```

```
#As only change has data that is smaller than zero and I do not think
#We need to transform it. So we do not to care about that.
# From the results we can find that Fertility is significant at 10%
#level and we should transform it with power of 1.346874. However,for
#Pop and PPgdp variable since they are not significant and there is no
#great improvement if we did the transformation. SO I decide to keep
#them unchanged.
```

7. Given the selected transformations of the predictors, select a transformation of the response using `MASS::boxcox` or `car::boxCox` and justify.

```
Fertilityt=UN3$Fertility^1.346874;
bc=boxCox(lm(ModernC~Change+PPgdp+Frate+Pop+Fertilityt+Purban, data=UN3))
```



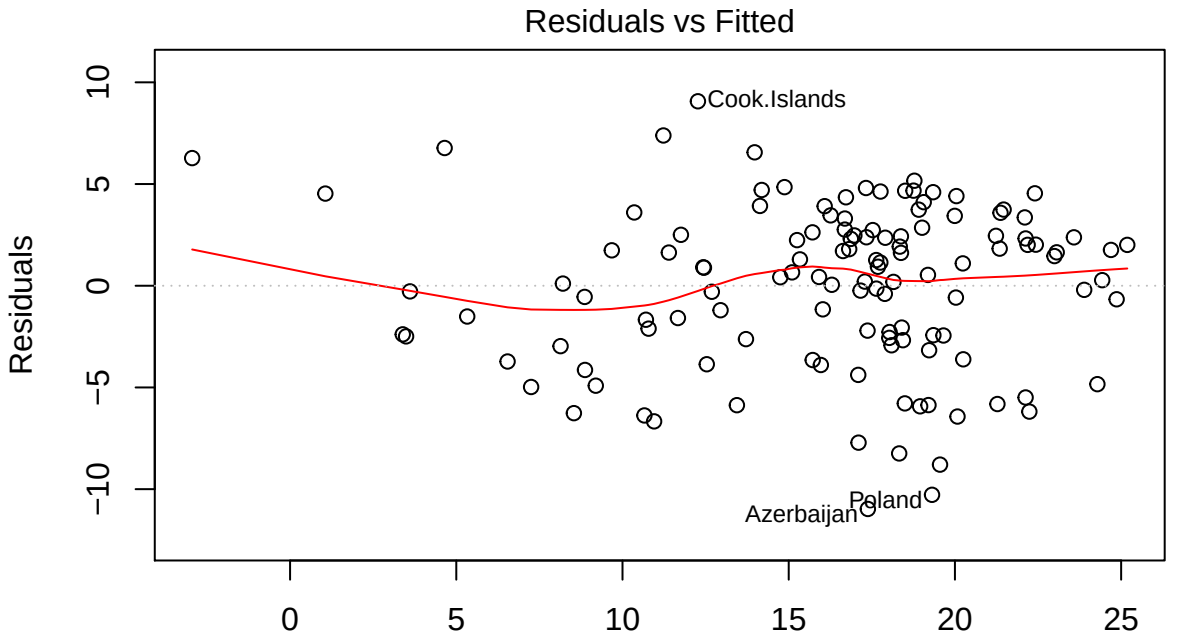

```
with(bc, x[which.max(y)])
```

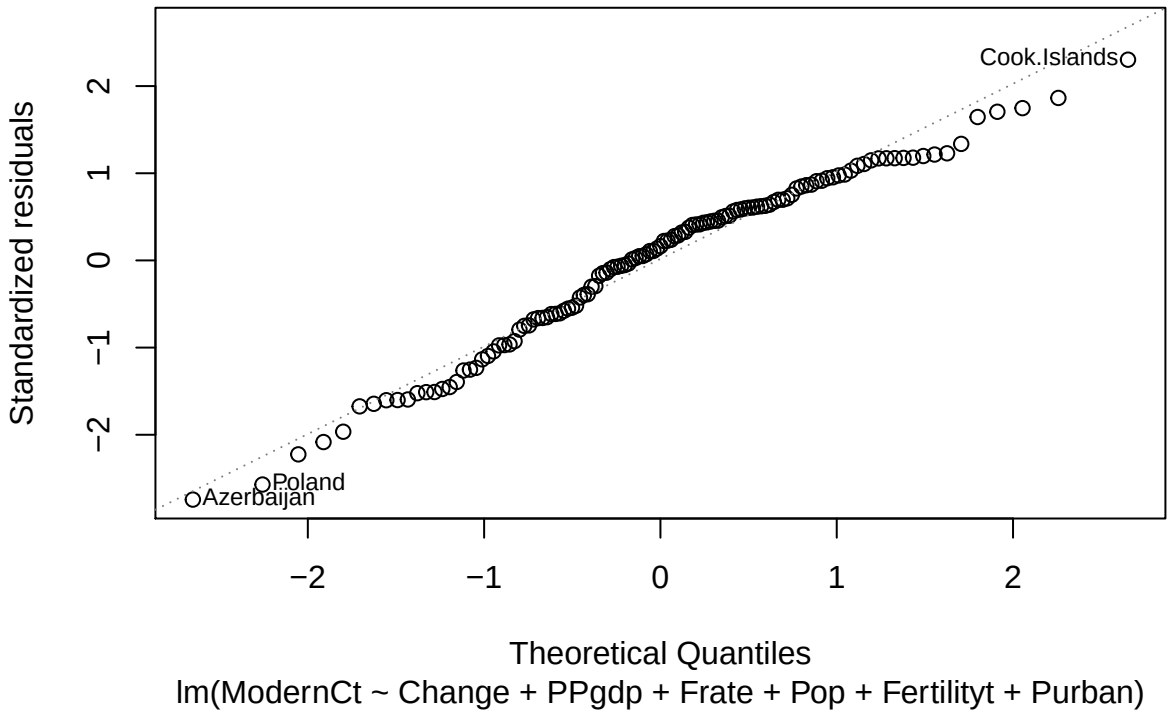
```
## [1] 0.7474747
```

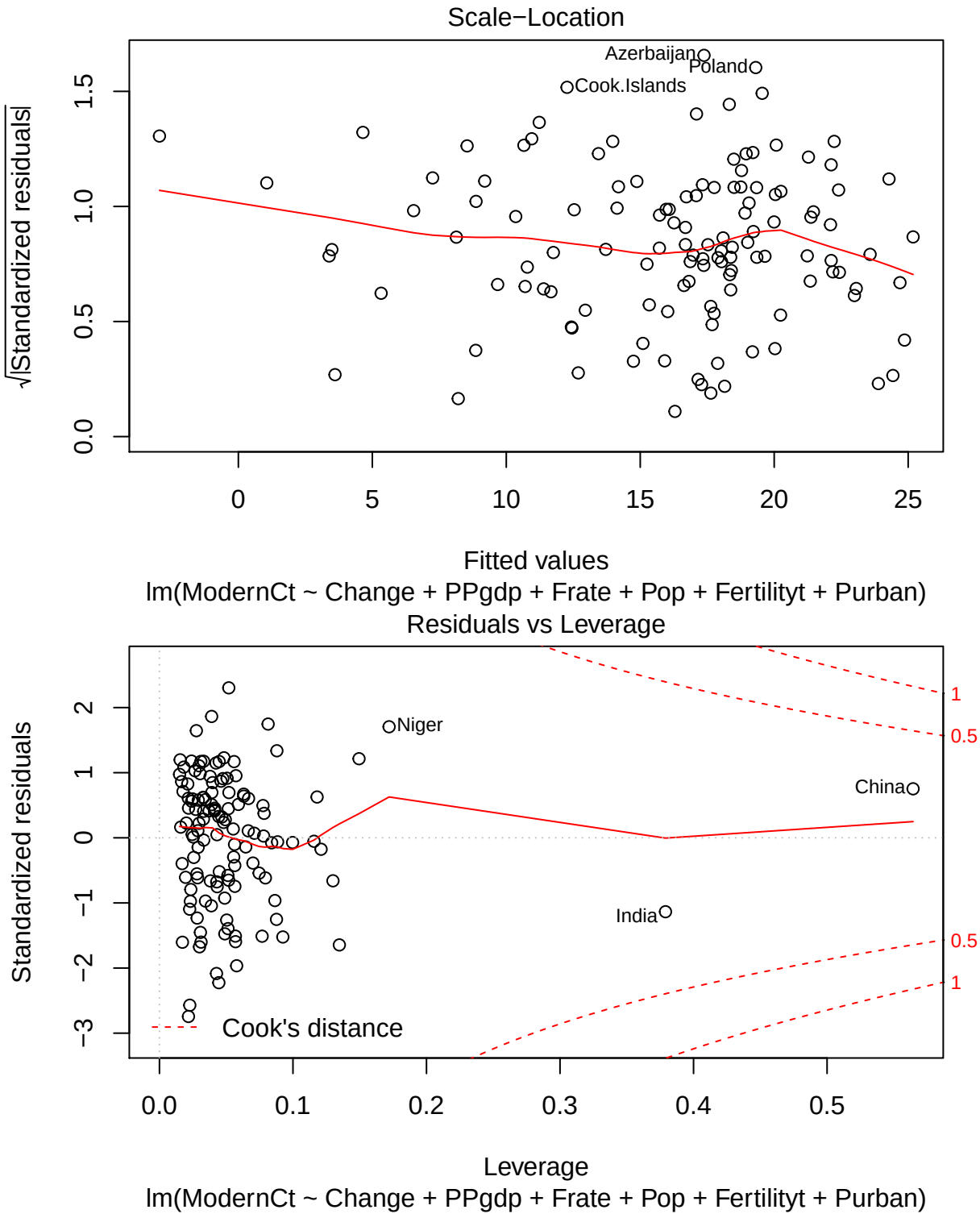
```
#From the this result, we know we should transform response to
#the power of 0.7474747
```

8. Fit the regression using the transformed variables. Provide residual plots and added variables plots and comment. If you feel that you need additional transformations of either the response or predictors, repeat any steps until you feel satisfied.

```
ModernCt=UN3$ModernC^0.7474747;
model2=lm(ModernCt~Change+PPgdp+Frate+Pop+Fertilityt+Purban,data=UN3);
plot(model2)
```


Residuals vs Fitted
Residuals
Cook.Islands
Azerbaijan
Poland
10
5
0
−5
−10
0
5
10
15
20
25
Fitted values
lm(ModernCt ~ Change + PPgdp + Frate + Pop + Fertilityt + Purban)


Normal Q−Q
Standardized residuals
Cook.Islands
Poland
Azerbaijan
2
1
0
−1
−2
−2
−1
0
1
2
Theoretical Quantiles
lm(ModernCt ~ Change + PPgdp + Frate + Pop + Fertilityt + Purban)




```
summary(model2)
```

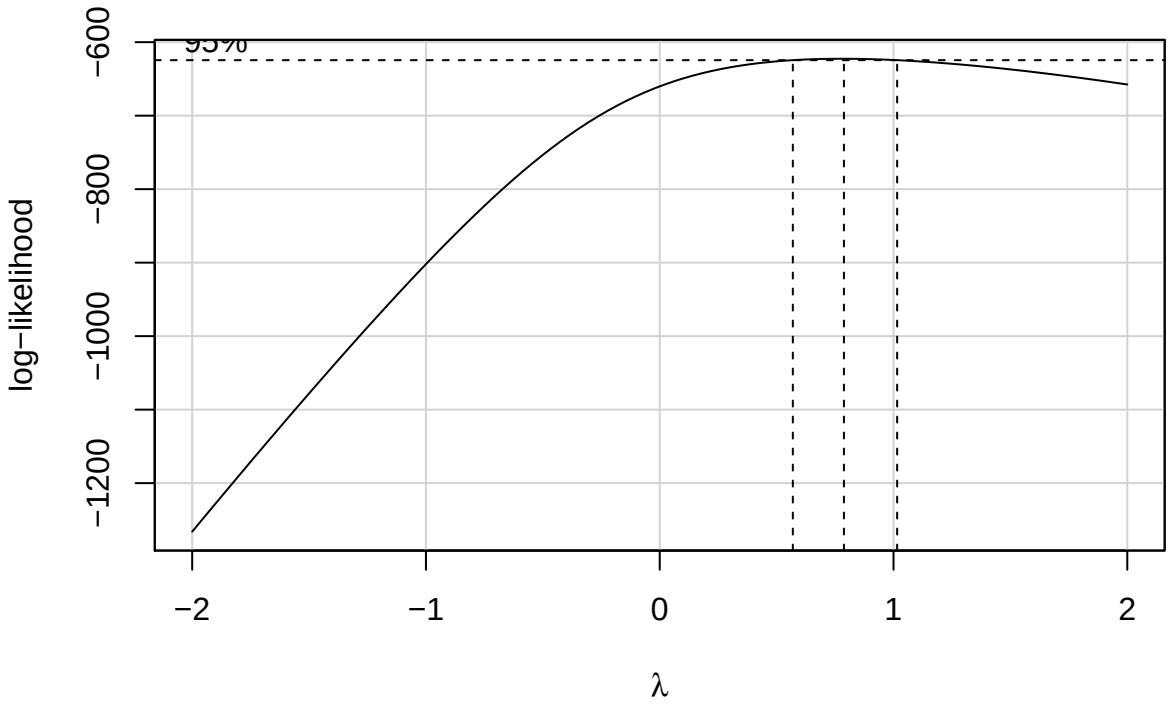
```
##
## Call:
## lm(formula = ModernCt ~ Change + PPgdp + Frate + Pop + Fertilityt +
##     Purban, data = UN3)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.9738  -2.5626   0.6573   2.6248   9.0677
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.890e+01  2.554e+00   7.402 2.16e-11 ***
## Change       1.461e+00  5.831e-01   2.506  0.01356 *
## PPgdp        1.514e-04  5.277e-05   2.870  0.00487 **
## Frate        2.667e-02  2.400e-02   1.111  0.26871
## Pop          5.068e-06  2.444e-06   2.074  0.04029 *
## Fertilityt  -1.702e+00  2.335e-01  -7.292 3.80e-11 ***
## Purban       9.761e-03  2.745e-02   0.356  0.72282
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.044 on 118 degrees of freedom
## Multiple R-squared:  0.6446, Adjusted R-squared:  0.6265
## F-statistic: 35.67 on 6 and 118 DF,  p-value: < 2.2e-16
```

```
#From the transformation,we can see that the r square improved a lot.
#Additionally, then normality assumption seems satisfied in the new
#model.So I am satisfied with this current model.
```

9. Start by finding the best transformation of the response and then find transformations of the predictors. Do you end up with a different model than in 8?

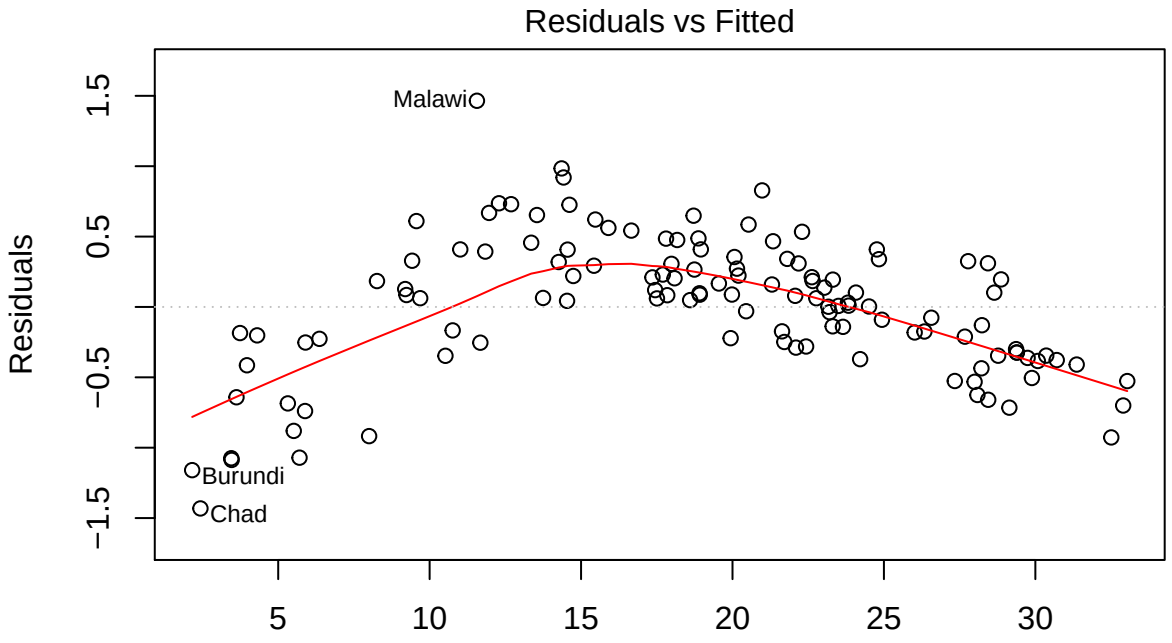
```
bc1=boxCox(lm(ModernC~., data=UN3));
```

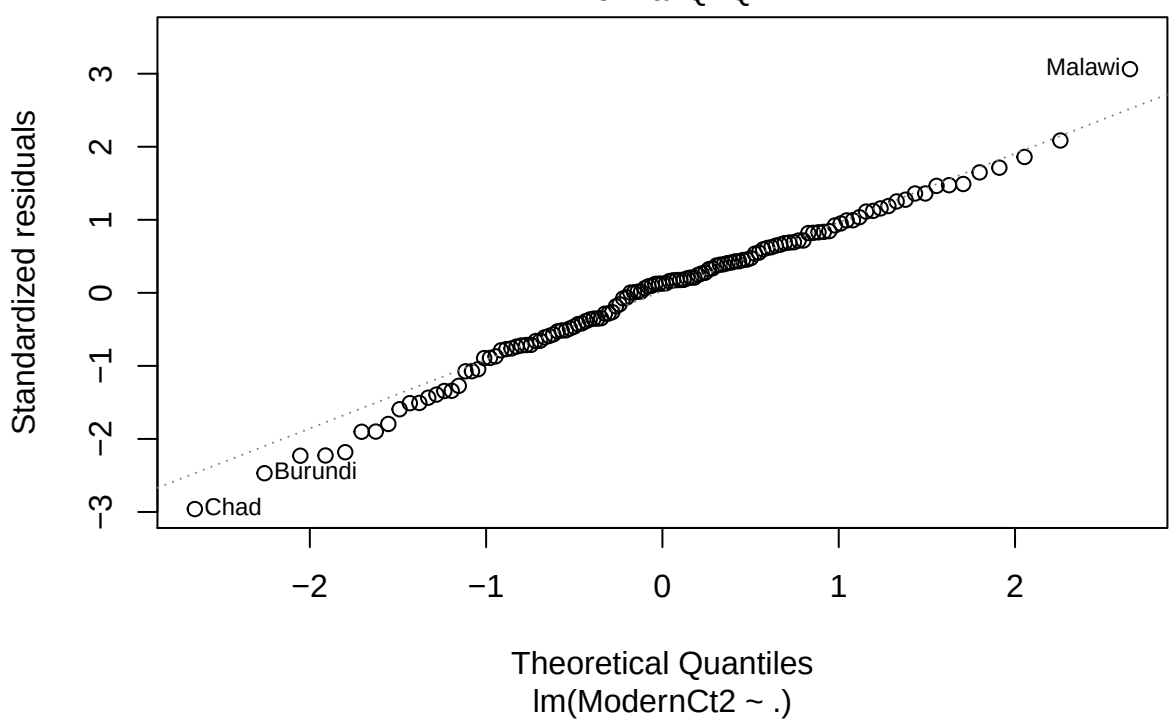


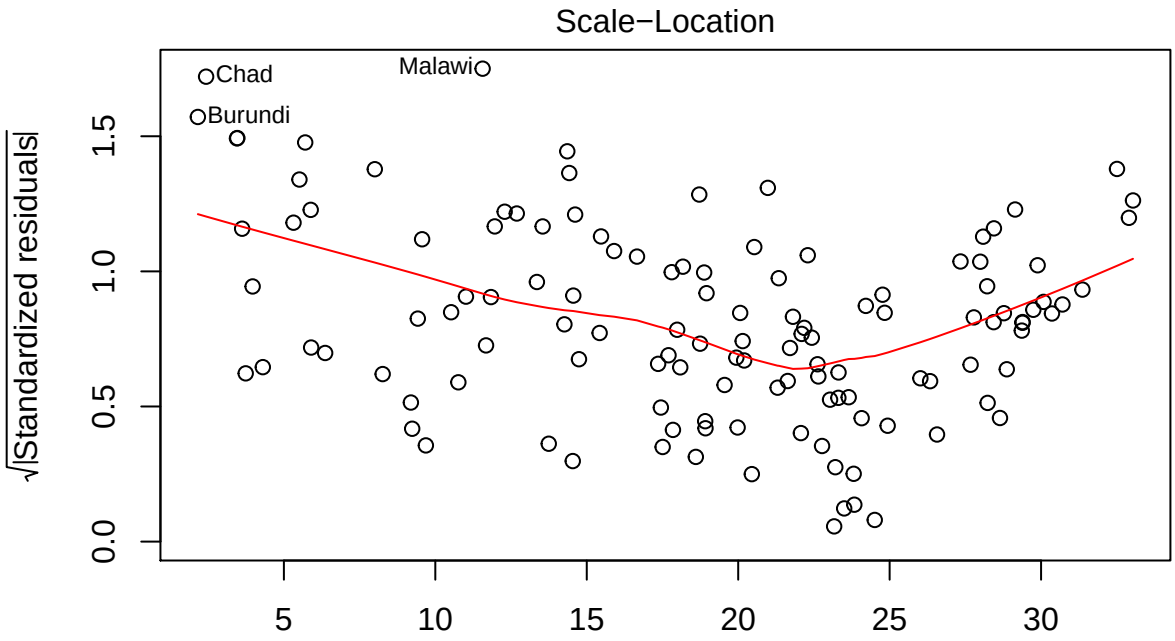

```
k=with(bc1, x[which.max(y)]);
k
```

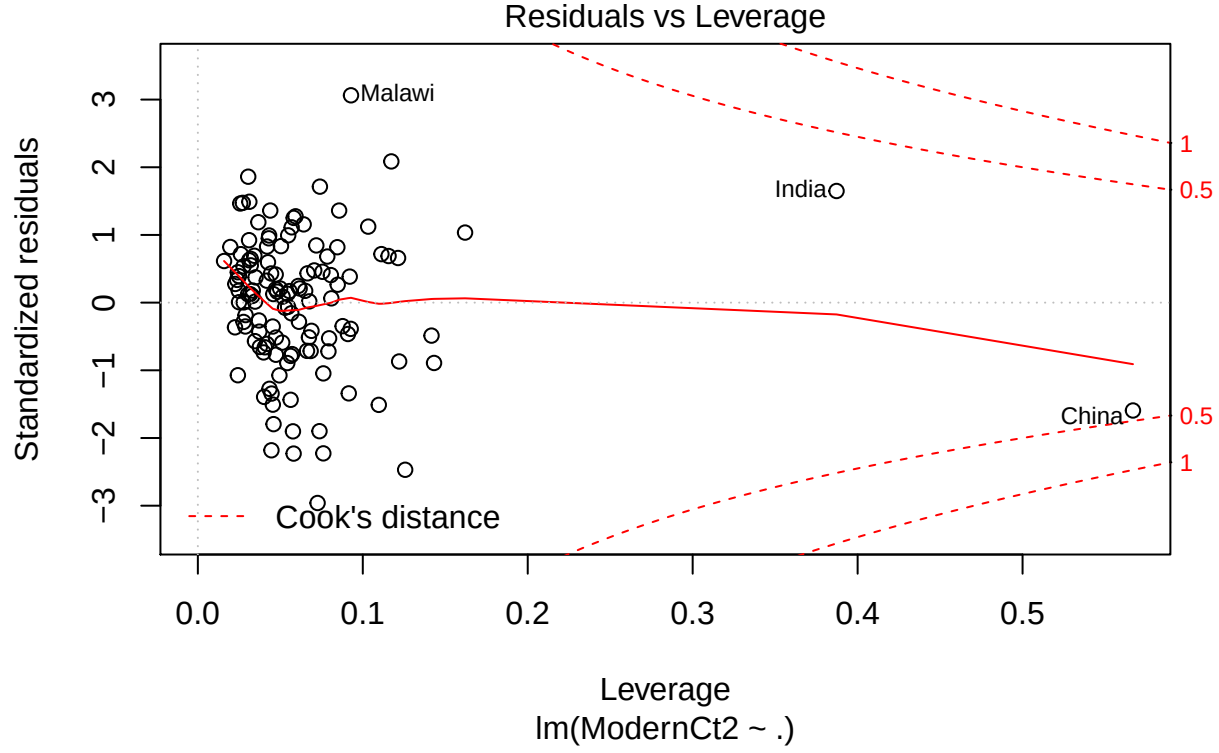
```
## [1] 0.7878788
```

```r
#We can see that we should transform the response to the power of k.
ModernCt2=UN3$ModernC^k;
model3=lm(ModernCt2~.,data=UN3);
plot(model3)
```







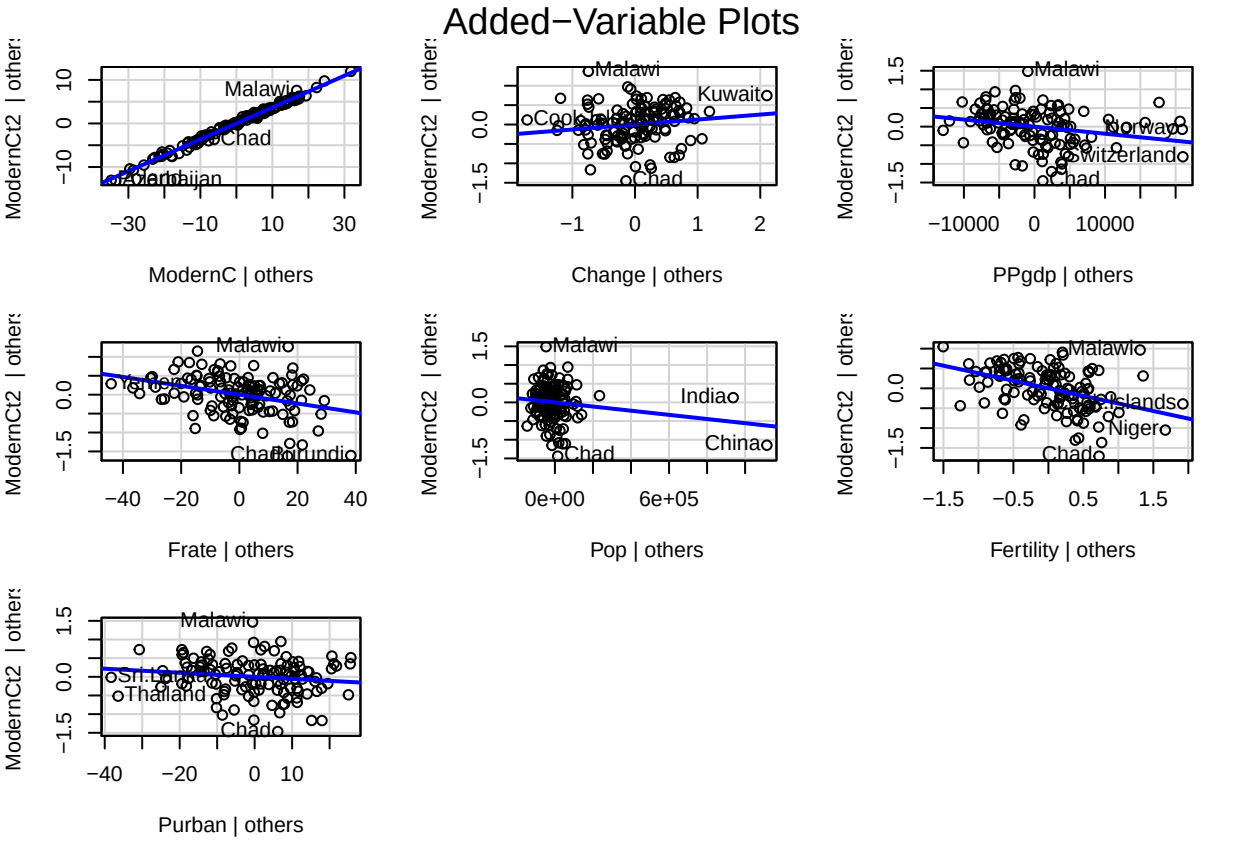



```
summary(model3)
```

```
##
## Call:
## lm(formula = ModernCt2 ~ ., data = UN3)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.43147 -0.29966  0.06214  0.31930  1.46483
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.105e+00  3.974e-01  12.846  < 2e-16 ***
## ModernC      3.654e-01  3.404e-03 107.361  < 2e-16 ***
## Change       1.283e-01  7.925e-02   1.619  0.10808
## PPgdp       -1.901e-05  6.787e-06  -2.801  0.00596 **
## Frate       -1.169e-02  3.010e-03  -3.885  0.00017 ***
## Pop         -5.565e-07  3.105e-07  -1.792  0.07564 .
## Fertility   -3.809e-01  7.484e-02  -5.090 1.38e-06 ***
## Purban      -5.433e-03  3.438e-03  -1.580  0.11675
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5022 on 117 degrees of freedom
## Multiple R-squared:  0.9963, Adjusted R-squared:  0.9961
## F-statistic:  4497 on 7 and 117 DF,  p-value: < 2.2e-16
```

```
car::avPlots(model3)
```



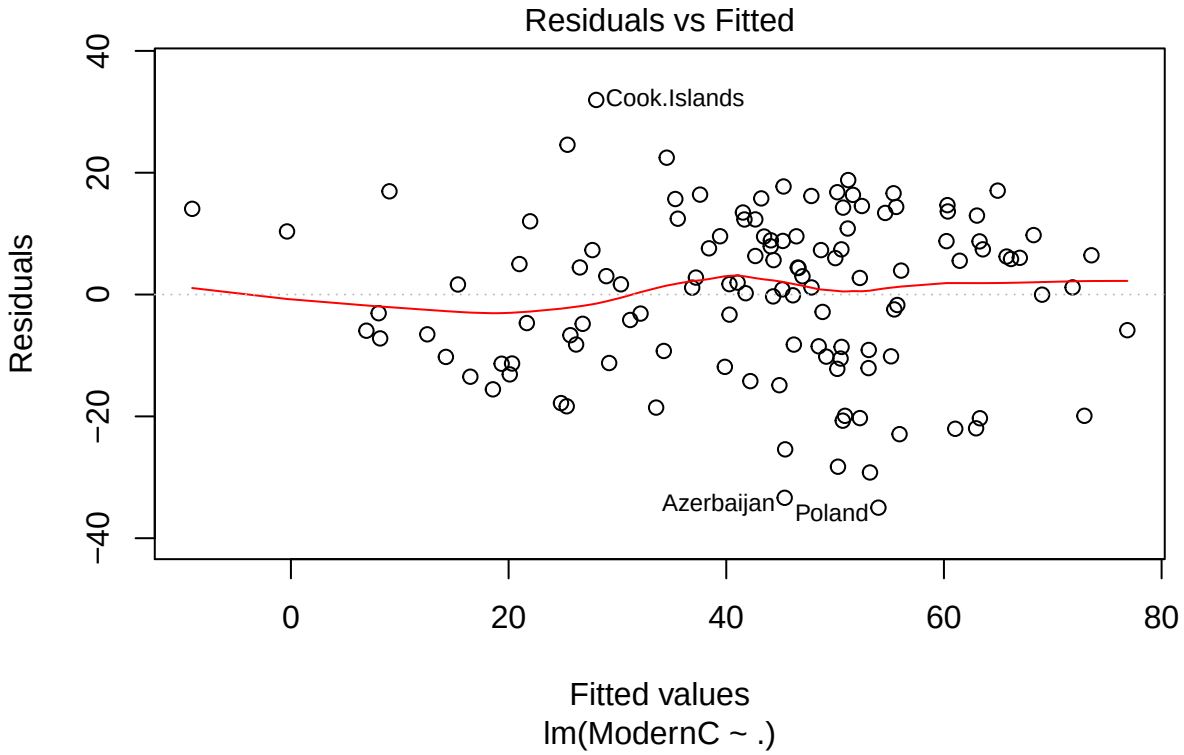

```
#From the avplot,I think for Fertility and PPgdp variable, we should
#do transformation. Both of them seem to have nolinear relation
#So I want to transform them.
boxTidwell(ModernCt2~PPgdp+Fertility,~Frate+Pop+Change+Purban,data=UN3)
```

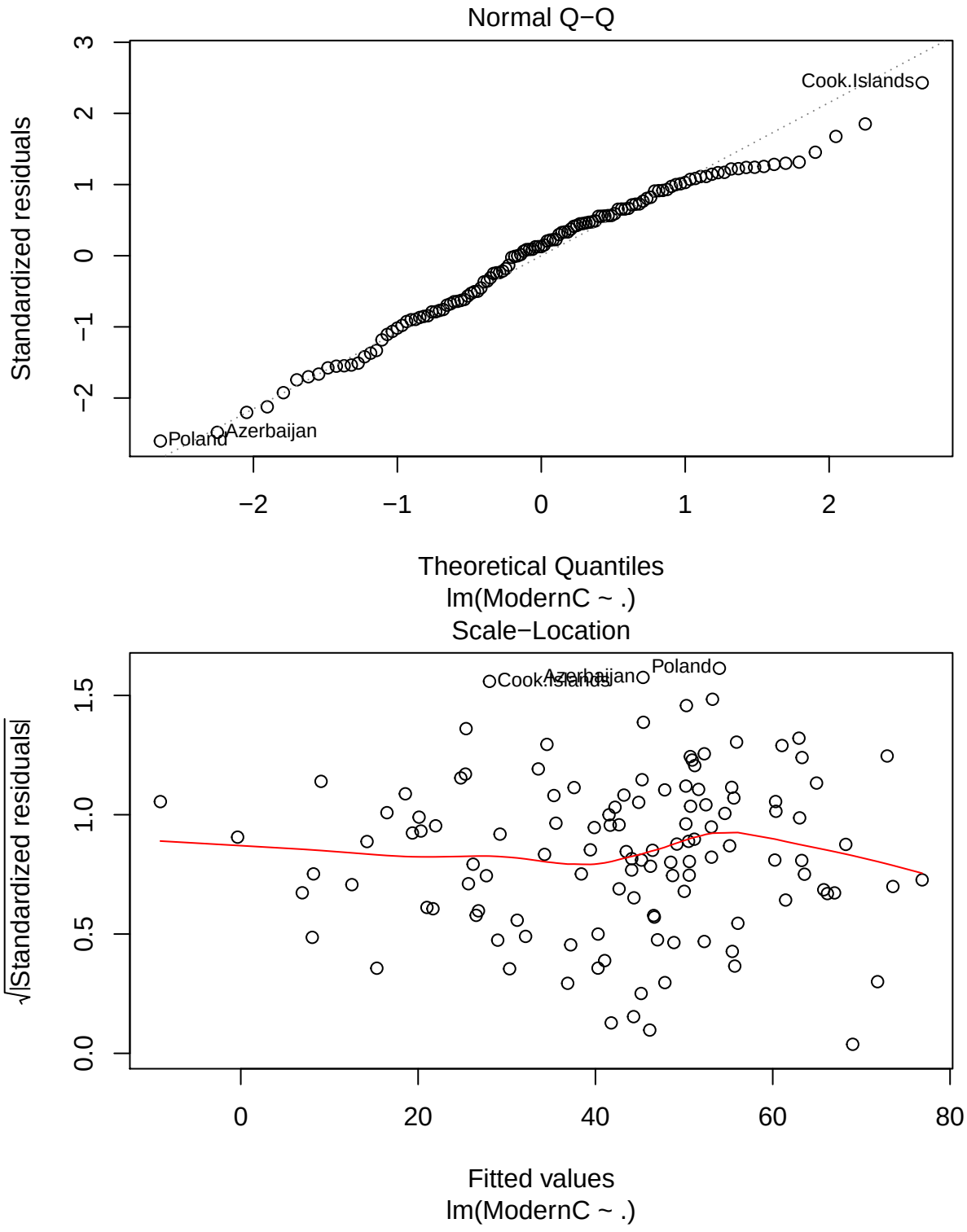
```
##           MLE of lambda Score Statistic (z) Pr(>|z|)
## PPgdp           -0.01184             -1.1004  0.27117
## Fertility        1.38308             -2.2507  0.02441 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## iterations =  18
```

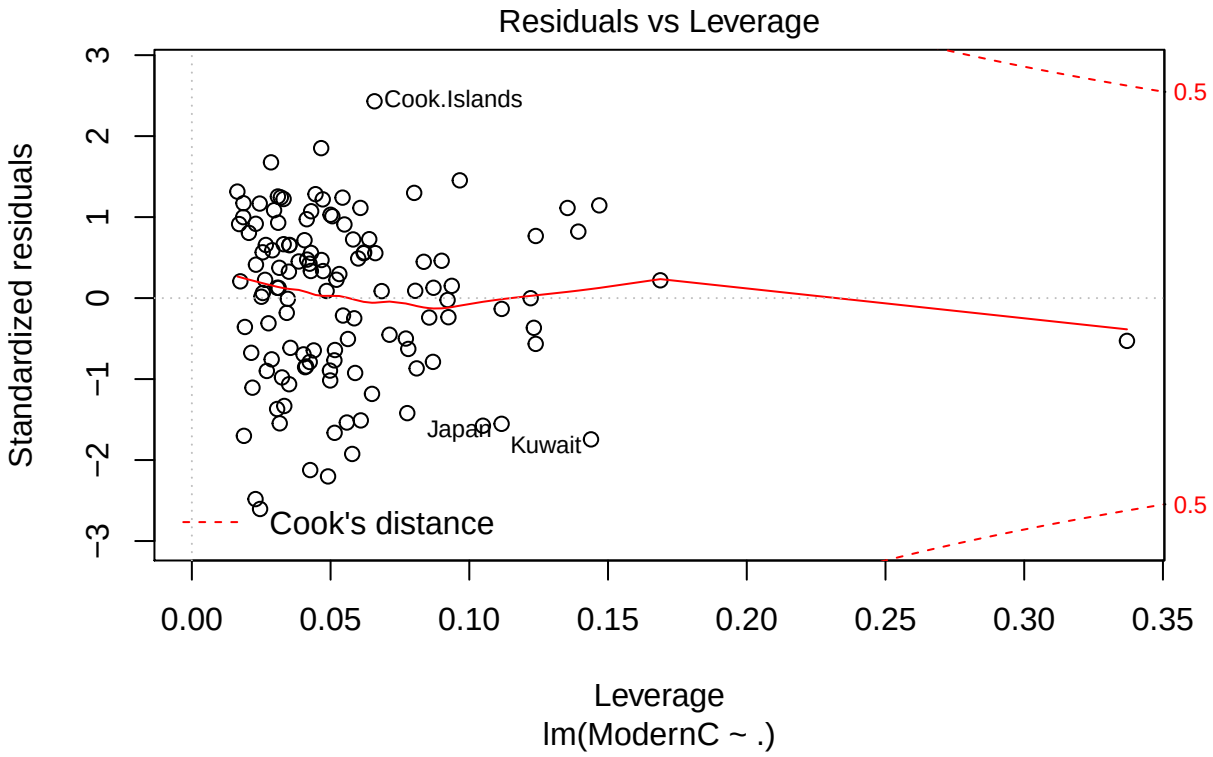
```
#From the results we find that only Fertility we need to transform
#as it is significant. But the power here is differnt from before.
#Also, the power for response is also different. So I got a different
#model.
```

10. Are there any outliers or influential points in the data? Explain. If so, refit the model after removing any outliers and comment on residual plots.

```
#Yes, seems china and indina are influential. So I remove them.
todrop=list("China","India");
UN3=UN3[!(rownames(UN3) %in% todrop), ];
model4=lm(ModernC~.,data=UN3);
plot(model4)
```




Normal Q−Q
Standardized residuals
Cook.Islands
Azerbaijan
Poland
−2
−1
0
1
2
3
Theoretical Quantiles
lm(ModernC ~ .)
Scale−Location
√|Standardized residuals|
Cook.Islands
Azerbaijan
Poland
0.0
0.5
1.0
1.5
0
20
40
60
80
Fitted values
lm(ModernC ~ .)

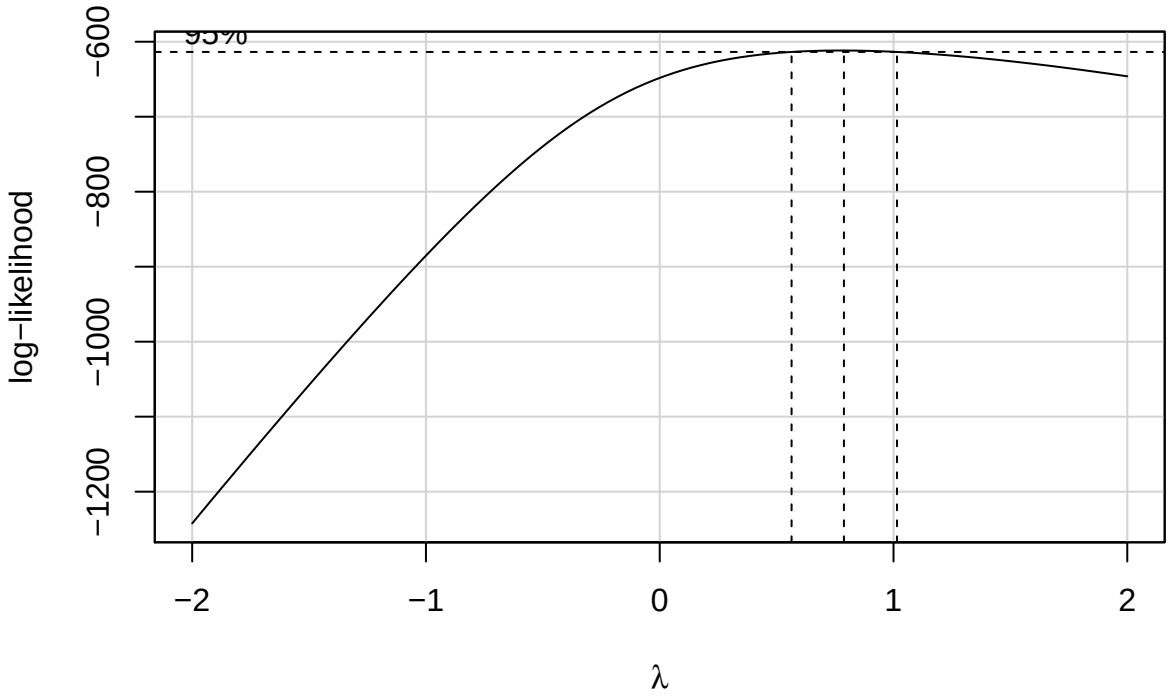



```
summary(model4)
```

```
##
## Call:
## lm(formula = ModernC ~ ., data = UN3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -34.981  -9.701   1.708   9.564  31.946
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.583e+01  9.602e+00   5.815 5.46e-08 ***
## Change       5.169e+00  2.096e+00   2.467  0.01510 *
## PPgdp        5.284e-04  1.792e-04   2.949  0.00385 **
## Frate        1.104e-01  8.188e-02   1.349  0.18012
## Pop          3.419e-05  2.702e-05   1.266  0.20821
## Fertility   -1.095e+01  1.758e+00  -6.226 7.89e-09 ***
## Purban       4.847e-02  9.344e-02   0.519  0.60494
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.6 on 116 degrees of freedom
## Multiple R-squared:  0.6128, Adjusted R-squared:  0.5928
## F-statistic:  30.6 on 6 and 116 DF,  p-value: < 2.2e-16
```

```
#After remove two influential points,
#From residual vs Fitted and Scale-location graphs, we can see that
#the the variance seems constant for different values. So the
#assumption that constant variance still holds.
```

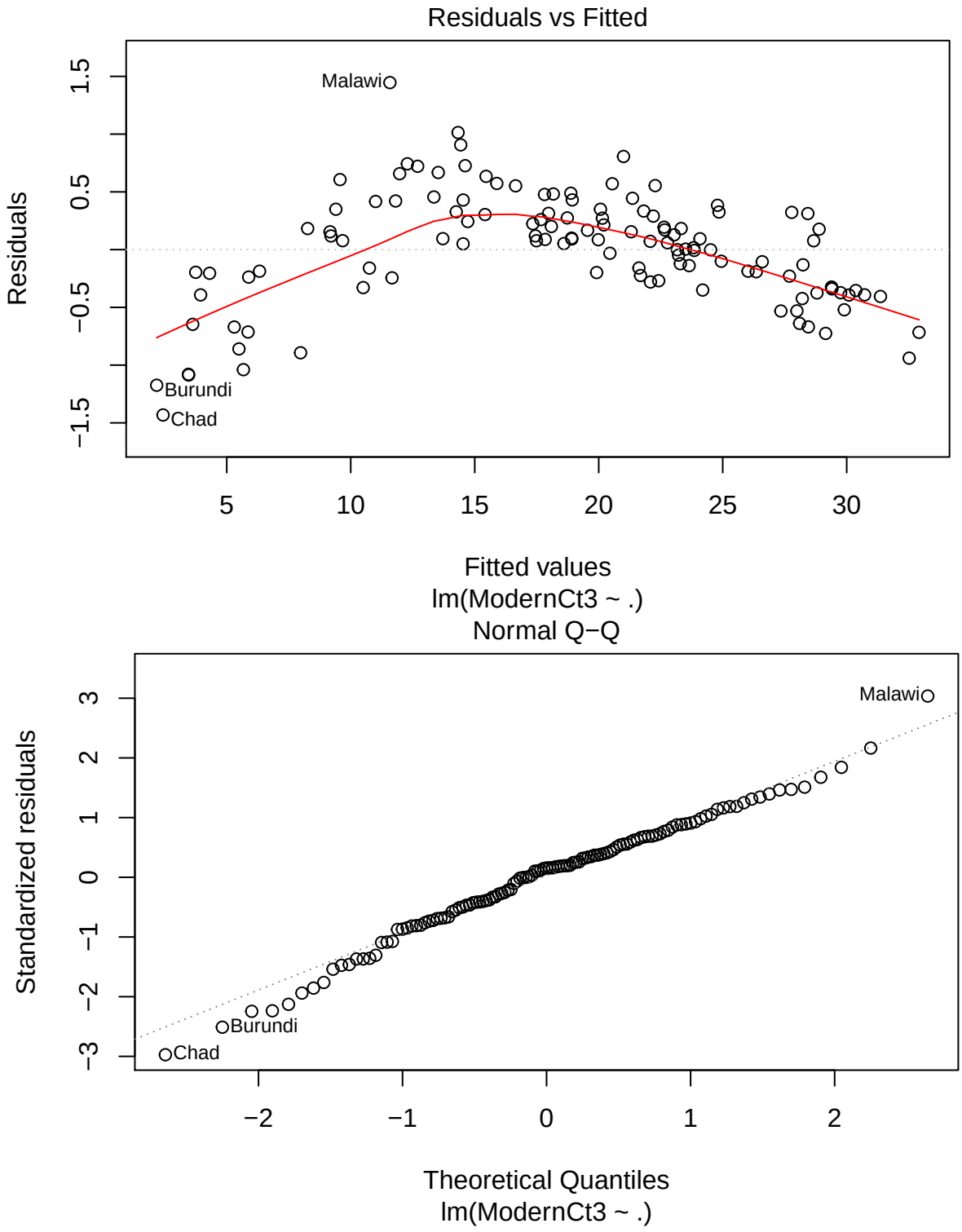
```
#From Normal Q-Q graph we can see it is getting better but there are still lots of points donot lie in
#which means the normailty assumption still may be violated. From the
#Residuals vs leverage graph we find that all points has cook
#disntance less than .5. So none of them are quiet influencial.
bc2=boxCox(lm(ModernC~., data=UN3));
```

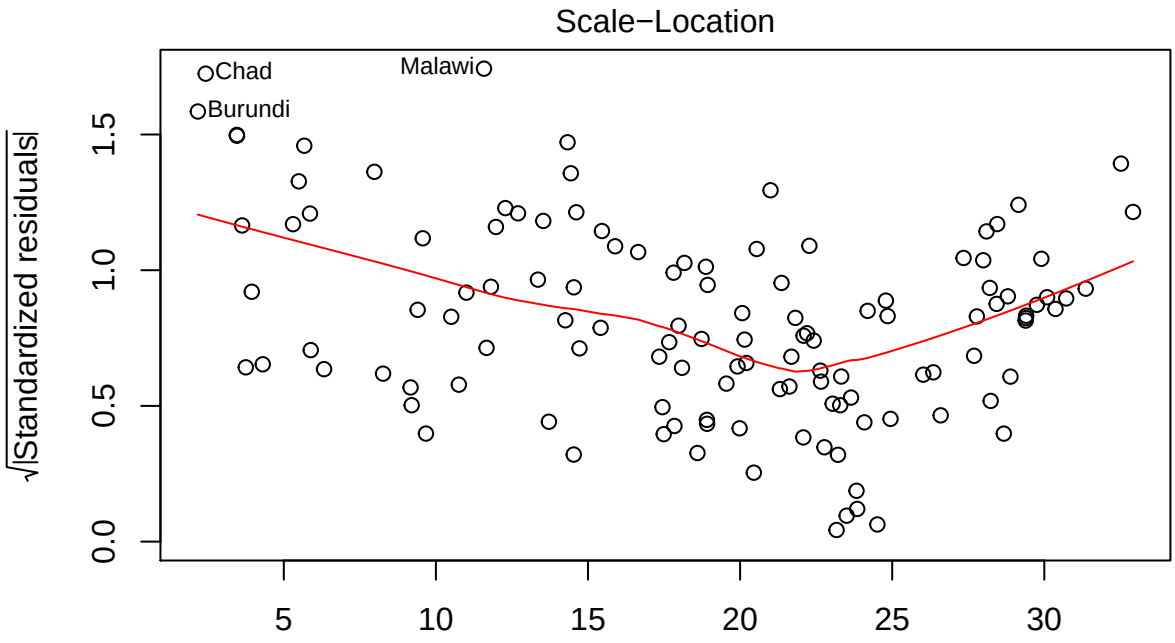


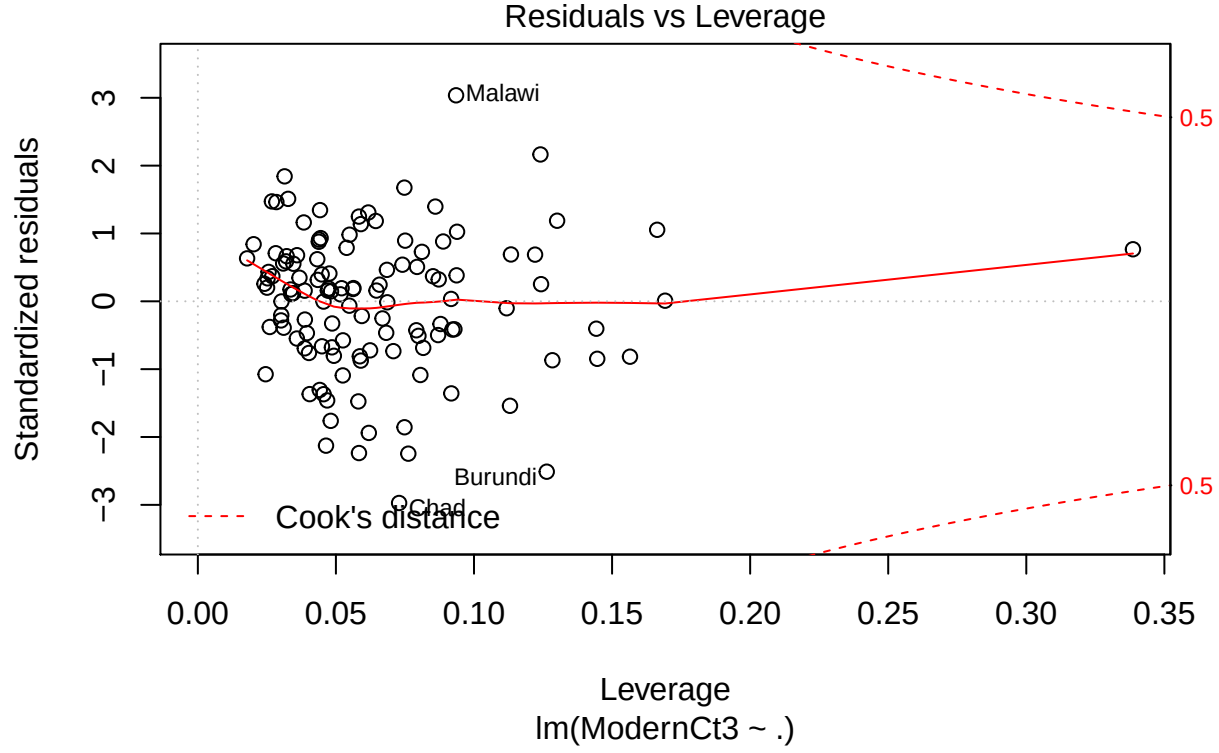

```
k=with(bc2, x[which.max(y)]);
k
```

```
## [1] 0.7878788
```

```
ModernCt3=UN3$ModernC^k;
model4=lm(ModernCt3~.,data=UN3);
plot(model4)
```


Residuals vs Fitted
Residuals
Malawi
Burundi
Chad
1.5
0.5
−0.5
−1.5
5
10
15
20
25
30
Fitted values
lm(ModernCt3 ~ .)
Normal Q−Q
Standardized residuals
Malawi
Burundi
Chad
3
2
1
0
−1
−2
−3
−2
−1
0
1
2
Theoretical Quantiles
lm(ModernCt3 ~ .)



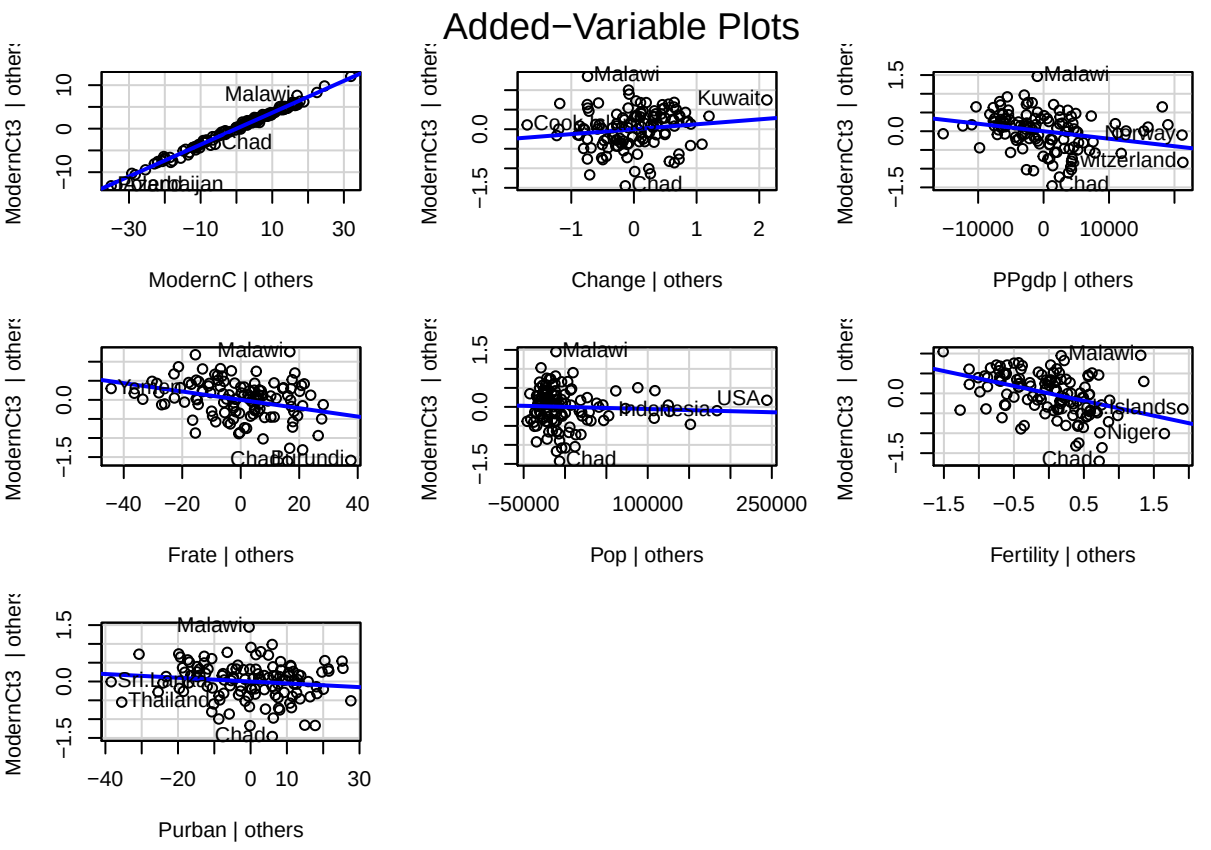



```
summary(model4)
```

```
##
## Call:
## lm(formula = ModernCt3 ~ ., data = UN3)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.43173 -0.30220  0.07659  0.31811  1.44606
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.005e+00  4.013e-01  12.472  < 2e-16 ***
## ModernC      3.660e-01  3.415e-03 107.198  < 2e-16 ***
## Change       1.260e-01  7.907e-02   1.594 0.113790
## PPgdp       -2.003e-05  6.832e-06  -2.932 0.004069 **
## Frate       -1.094e-02  3.035e-03  -3.604 0.000464 ***
## Pop         -5.635e-07  1.000e-06  -0.563 0.574335
## Fertility   -3.738e-01  7.469e-02  -5.005 2.03e-06 ***
## Purban      -4.986e-03  3.440e-03  -1.449 0.149946
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5002 on 115 degrees of freedom
## Multiple R-squared:  0.9963, Adjusted R-squared:  0.9961
## F-statistic:  4426 on 7 and 115 DF,  p-value: < 2.2e-16
```

```
car::avPlots(model4)
```

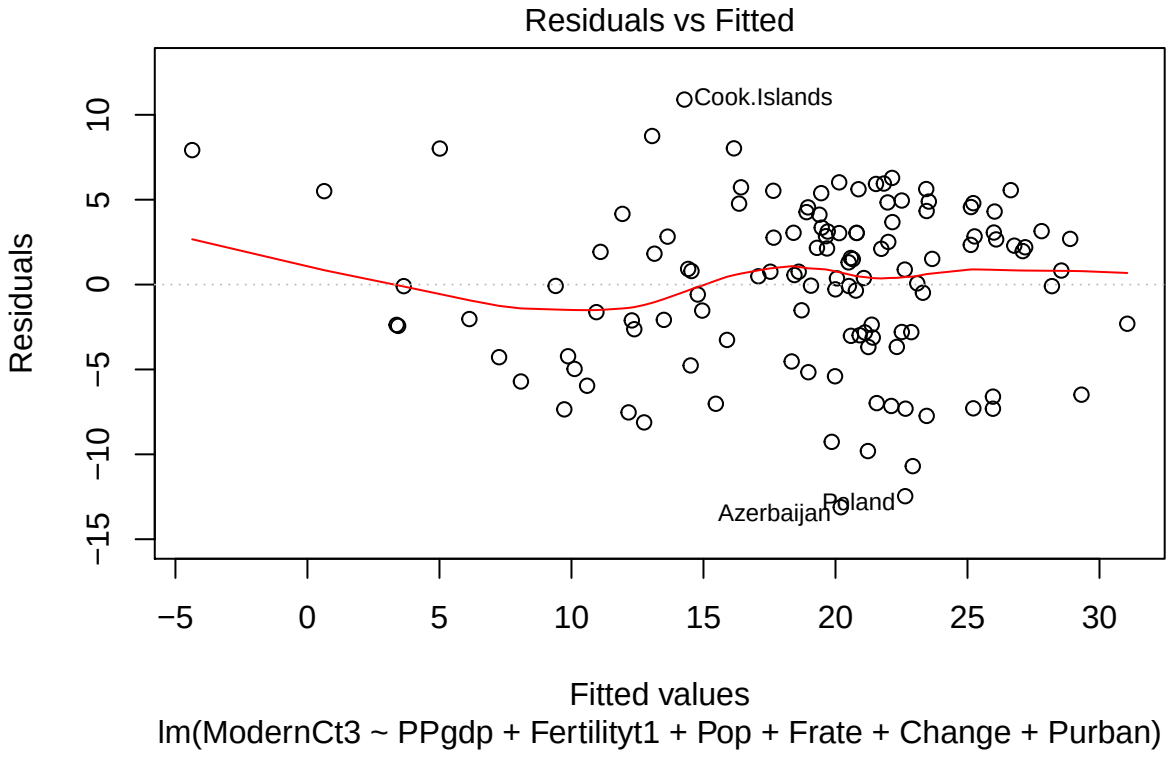


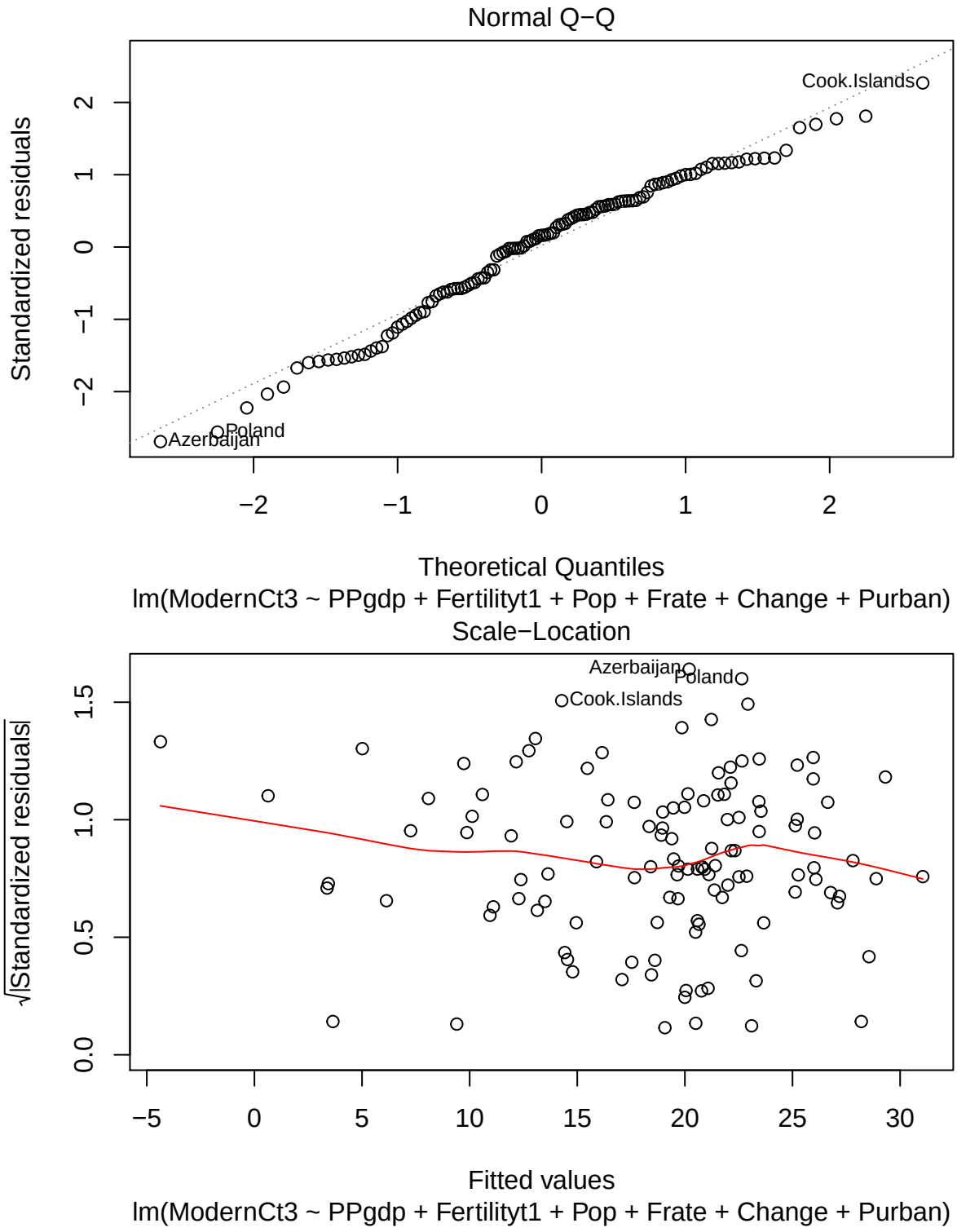

```
#Looks like PPgdp and Fertility needs transformation;
boxTidwell(ModernCt3~PPgdp+Fertility,~Pop+Frate+Change+Purban,data=UN3)
```

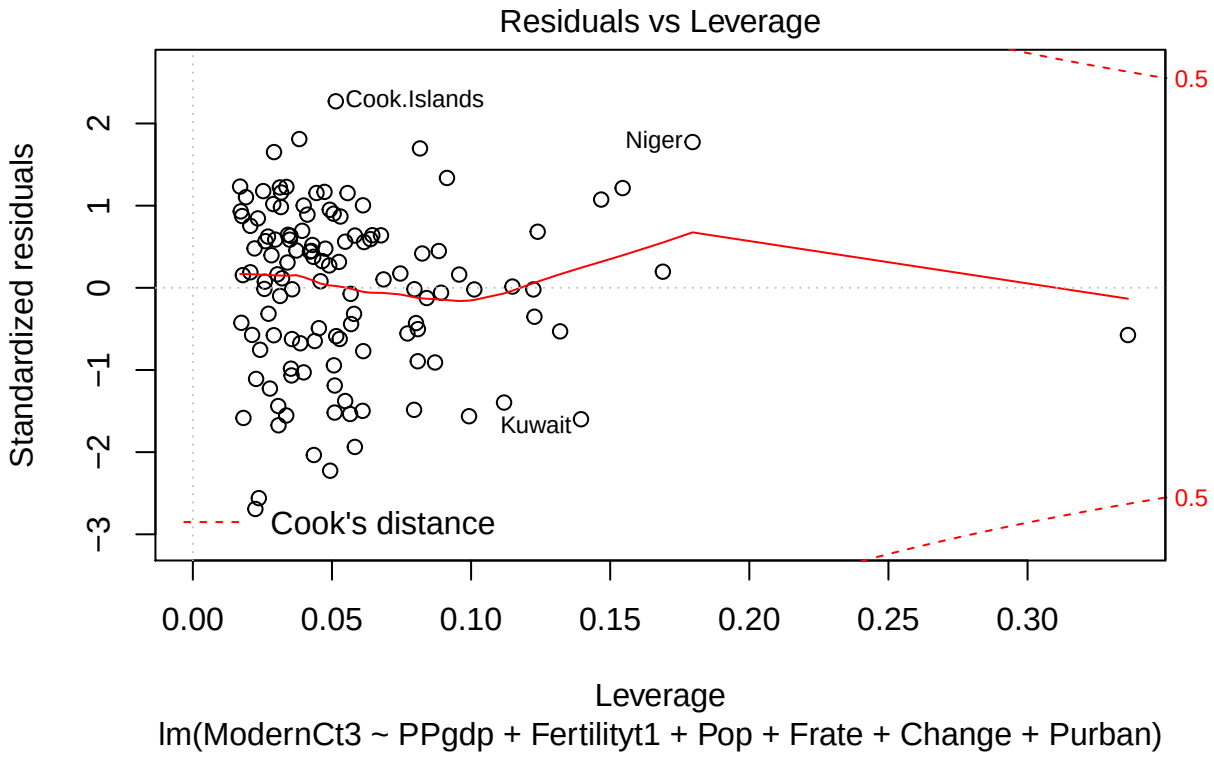
```
##           MLE of lambda Score Statistic (z) Pr(>|z|)
## PPgdp         -0.026004             -1.2534  0.21007
```

```
## Fertility      1.399604             -2.3307  0.01977 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## iterations =  18
```

```
Fertilityt1=UN3$Fertility^1.399604;
model5=lm(ModernCt3~PPgdp+Fertilityt1+Pop+Frate+Change+Purban,data=UN3);
plot(model5)
```




Normal Q−Q
Standardized residuals
Cook.Islands
Poland
Azerbaijan
−2
−1
0
1
2
Theoretical Quantiles
lm(ModernCt3 ~ PPgdp + Fertilityt1 + Pop + Frate + Change + Purban)
Scale−Location
√|Standardized residuals|
Azerbaijan
Poland
Cook.Islands
0.0
0.5
1.0
1.5
−5
0
5
10
15
20
25
30
Fitted values
lm(ModernCt3 ~ PPgdp + Fertilityt1 + Pop + Frate + Change + Purban)




```
summary(model5)
```

```
##
## Call:
## lm(formula = ModernCt3 ~ PPgdp + Fertilityt1 + Pop + Frate +
##     Change + Purban, data = UN3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.1163  -3.0031   0.7571   3.1360  10.8978
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.165e+01  3.128e+00   6.921 2.63e-10 ***
## PPgdp        1.887e-04  6.506e-05   2.901  0.00445 **
## Fertilityt1 -1.805e+00  2.532e-01  -7.128 9.30e-11 ***
## Pop          1.191e-05  9.786e-06   1.216  0.22628
## Frate        3.098e-02  2.967e-02   1.044  0.29855
## Change       1.667e+00  7.053e-01   2.363  0.01978 *
## Purban       1.167e-02  3.359e-02   0.347  0.72892
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.929 on 116 degrees of freedom
## Multiple R-squared:  0.6378, Adjusted R-squared:  0.6191
## F-statistic: 34.05 on 6 and 116 DF,  p-value: < 2.2e-16
```

```
#model 5 is my final model. We can see from the residual vs Fitted and
#Scale-location graphs, the the variance seems constant for different
#values. For Normal Q-Q, normality seems still be a issue but it is
```

```
#getting better. For the last graph, we can see all has small cook's
#distance which means most of points are not influential.
```

Summary of Results

11. For your final model, provide summaries of coefficients with 95% confidence intervals in a nice table with interpretations of each coefficient. These should be in terms of the original units!

```
v1=confint(model5,'PPgdp',level=0.95)
v2=confint(model5,'Fertilityt1',level=0.95)
v1=rbind(v1,v2)
v3=confint(model5,'Pop',level=0.95)
v1=rbind(v1,v3)
v4=confint(model5,'Frate',level=0.95)
v1=rbind(v1,v4)
v5=confint(model5,'Change',level=0.95)
v1=rbind(v1,v5)
v6=confint(model5,'Purban',level=0.95)
v1=rbind(v1,v6)
v1
```

```
##                      2.5 %        97.5 %
## PPgdp         5.986713e-05  3.175715e-04
## Fertilityt1 -2.306550e+00 -1.303475e+00
## Pop         -7.478216e-06  3.128738e-05
## Frate       -2.778258e-02  8.974541e-02
## Change       2.697931e-01  3.063489e+00
## Purban      -5.486297e-02  7.820269e-02
```

```
#PS: if in original unit, the 95% for Fertility is:
-2.306550^(1/1.399604)     -1.303475^(1/1.399604)
```

```
## [1] -3.025386
```

```
#interperations:
#If PPgdp increase 1 other hold same, then resoponse^0.787878 will
#increase between the confidence interval of PPgdp, and that happens
#for 95% of all the cases.Same for Pop, Frate, Change, and Purban. For
#Fertility,if Fertility increase 1, than resoponse^0.787878
#will increase between the (original unit)confidence interval of Fertilityt1 with 95%
#confidence level.
```

12. Provide a paragraph summarizing your final model and findings suitable for the US envoy to the UN after adjusting for outliers or influential points. You should provide a justification for any case deletions in your final model

```
#My final model almost satisfied the assumption of linear regression
#And the we find that expect Fertility, all other variables have
#positive relation with the response ModernC. So the US envoy might
#approximate the level of ModernC by several indicators. As China and
#India are countries with super big population, so that the ModernC
#are hard to capture in those two countries. So I delete them.
```

Methodology

13. Prove that the intercept in the added variable scatter plot will always be zero. *Hint: use the fact that if H is the project matrix which contains a column of ones, then $1_n^T(I-H)=0$. Use this to show that the sample mean of residuals will always be zero if there is an intercept.*

Answer: As Y~x_1+x_2... (xi not in right side)
X_i~x_1+x_2... (xi not in right side)
$\hat{e_Y}=(I-H)y$
$\hat{e_{xi}}=(I-H)x_i$
now the regression form is$\hat{e_Y}$~$\hat{e_{xi}}$
Let's calculate $\hat{\beta_0}*\mathbf{1}$
by defintion it should be $[I-\hat{e_{xi}}*(\hat{e_{xi}}'\hat{e_{xi}})^{-1}\hat{e_{xi}}']\hat{e_y}$
$=[I-\frac{\hat{e_{xi}}*\hat{e_{xi}}'}{\sum e_{xi}^2}]e_y$
$=[I-H]y-[\frac{(I-H)'x_ix_i'(I-H)}{\sum e_{xi}^2}]e_y$
As (I-H) is idempotent
$=[I-H]y-[\frac{(I-H)\sum x_i^2}{\sum e_{xi}^2}](I-H)y$
$=[I-H]y-\frac{(I-H)y\sum x_i^2}{\sum e_{xi}^2}$
by hint multiply by 1_n^T (This hint means theoretical the sum of residule euqals to zero)
$=1_n^T[I-H]y-\frac{1_n^T(I-H)y\sum x_i^2}{\sum e_{xi}^2}$
$=0*y-\frac{0*\sum x_i^2}{\sum e_{xi}^2}$
$=0_n^T$
so $\hat{\beta_0}$should be zero to make $\hat{\beta_0}\mathbf{1}$ to be zero. DONE!

14. For multiple regression with more than 2 predictors, say a full model given by `Y ~ X1 + X2 + ... Xp` we create the added variable plot for variable `j` by regressing `Y` on all of the `X`'s except `Xj` to form `e_Y` and then regressing `Xj` on all of the other X's to form `e_X`. Confirm that the slope in the manually constructed added variable plot for one of the predictors in Ex. 10 is the same as the estimate from your model.

Answer:

```
# if we regress Modernc on all 6 variables
model4=lm(ModernCt3~Pop+PPgdp+Purban+Fertility+Frate+Change,data=UN3);
summary(model4);
```

```
##
## Call:
## lm(formula = ModernCt3 ~ Pop + PPgdp + Purban + Fertility + Frate +
##     Change, data = UN3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.1335  -3.6338   0.8996   3.3539  12.0190
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.544e+01  3.532e+00   7.203 6.38e-11 ***
## Pop          1.195e-05  9.939e-06   1.203  0.23160
## PPgdp        1.734e-04  6.591e-05   2.631  0.00968 **
## Purban       1.275e-02  3.437e-02   0.371  0.71126
## Fertility   -4.380e+00  6.468e-01  -6.773 5.50e-10 ***
```

```
## Frate        2.948e-02  3.012e-02   0.979  0.32976
## Change       2.018e+00  7.709e-01   2.618  0.01003 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.004 on 116 degrees of freedom
## Multiple R-squared:  0.6268, Adjusted R-squared:  0.6075
## F-statistic: 32.47 on 6 and 116 DF,  p-value: < 2.2e-16
```

```
#we can see the coefficient for change is 2.018e+00;
#regress without Change
model7=lm(ModernCt3~Pop+PPgdp+Purban+Fertility+Frate,data=UN3)
ey=resid(model7)
model8=lm(Change~Pop+PPgdp+Purban+Fertility+Frate,data=UN3)
ex=resid(model8)
model9=lm(ey~ex)
summary(model9)
```

```
##
## Call:
## lm(formula = ey ~ ex)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.1335  -3.6338   0.8996   3.3539  12.0190
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.014e-17  4.417e-01   0.000  1.00000
## ex           2.018e+00  7.548e-01   2.674  0.00854 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.899 on 121 degrees of freedom
## Multiple R-squared:  0.05579,    Adjusted R-squared:  0.04798
## F-statistic: 7.149 on 1 and 121 DF,  p-value: 0.008537
```

```
#we can see the slope is 2.018e+00; which is same as I calcualte before.
```