HW2 STA521 Fall18

Zhen Han Si, 0854615, szhhan

Due September 23, 2018 5pm

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*

```
options(warn=-1)
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
```

```
sapply(UN3,class)
```

```
##   ModernC    Change     PPgdp     Frate       Pop Fertility    Purba
n
## "integer" "numeric" "integer" "integer" "numeric" "numeric" "integer
"
```

All the variables except Purban have missing values. All the values are quantitative.

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

Import the library needed

```
options(warn=-1)
library(knitr)
library(ggplot2)
library(GGally)
library(MASS)
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:alr3':
##
##     forbes
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following object is masked from 'package:MASS':
##
##     select
```

```
## The following object is masked from 'package:GGally':
##
##     nasa
```

```
## The following object is masked from 'package:car':
##
##     recode
```

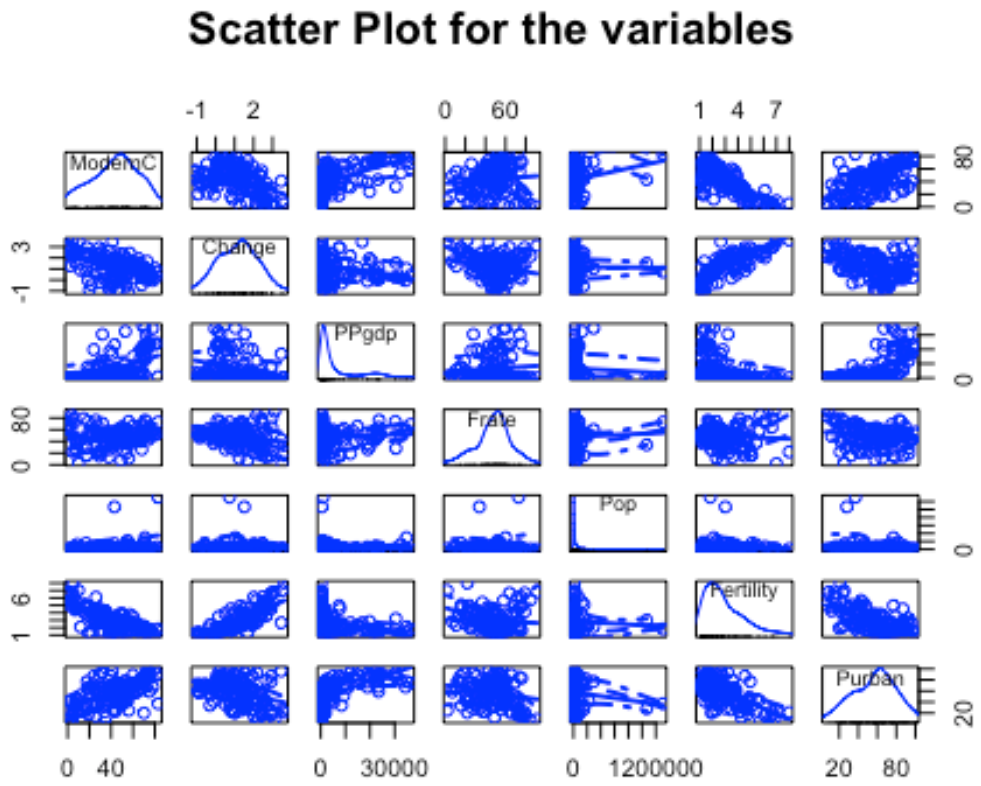
```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

library(car)

average = sapply(UN3,mean,na.rm=TRUE)
sd = sapply(UN3,sd,na.rm=TRUE)
matrix = matrix(c(average,sd),ncol=2,nrow=7)
table = as.data.frame(matrix)
rownames(table) = colnames(UN3)
colnames(table) = c('mean','sd')
kable(table,format='markdown')
```

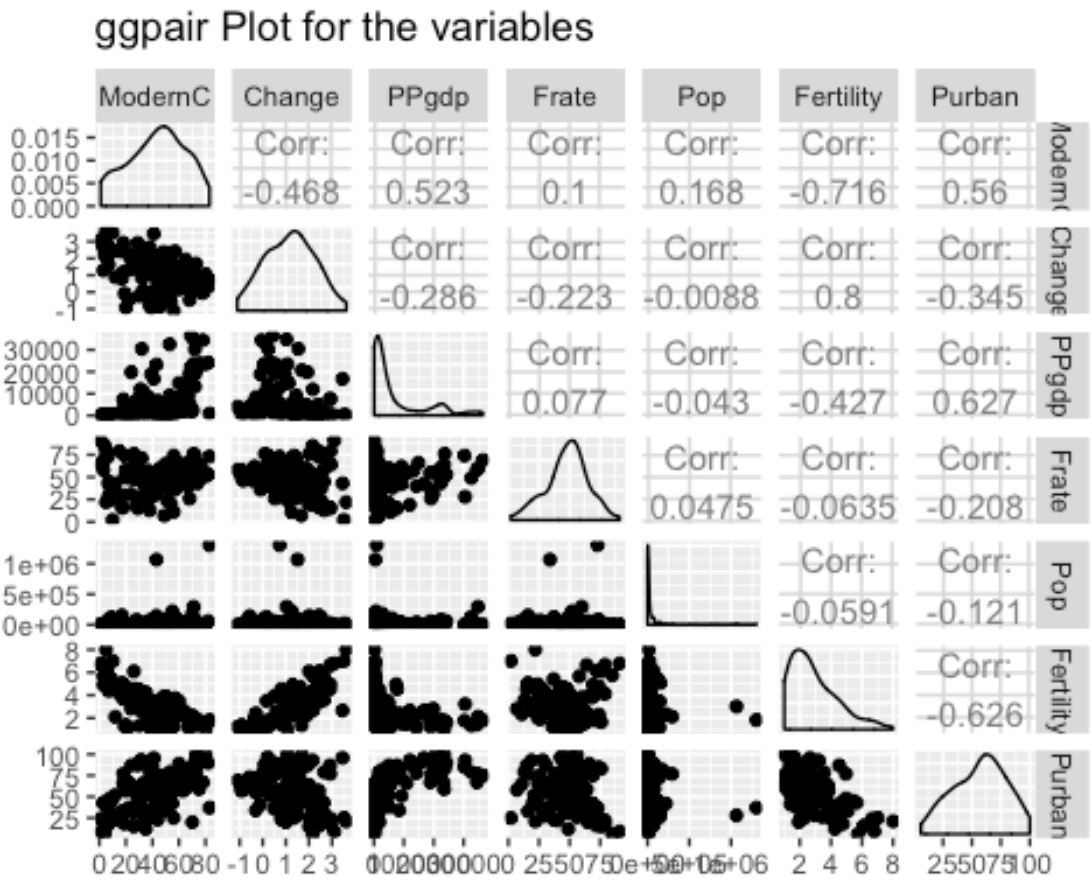
	mean	sd
ModernC	38.717105	2.263661e+01
Change	1.418373	1.133133e+00
PPgdp	6527.388060	9.325189e+03
Frate	48.305389	1.653245e+01
Pop	30281.871428	1.206767e+05
Fertility	3.214000	1.706918e+00
Purban	56.200000	2.410976e+01

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?

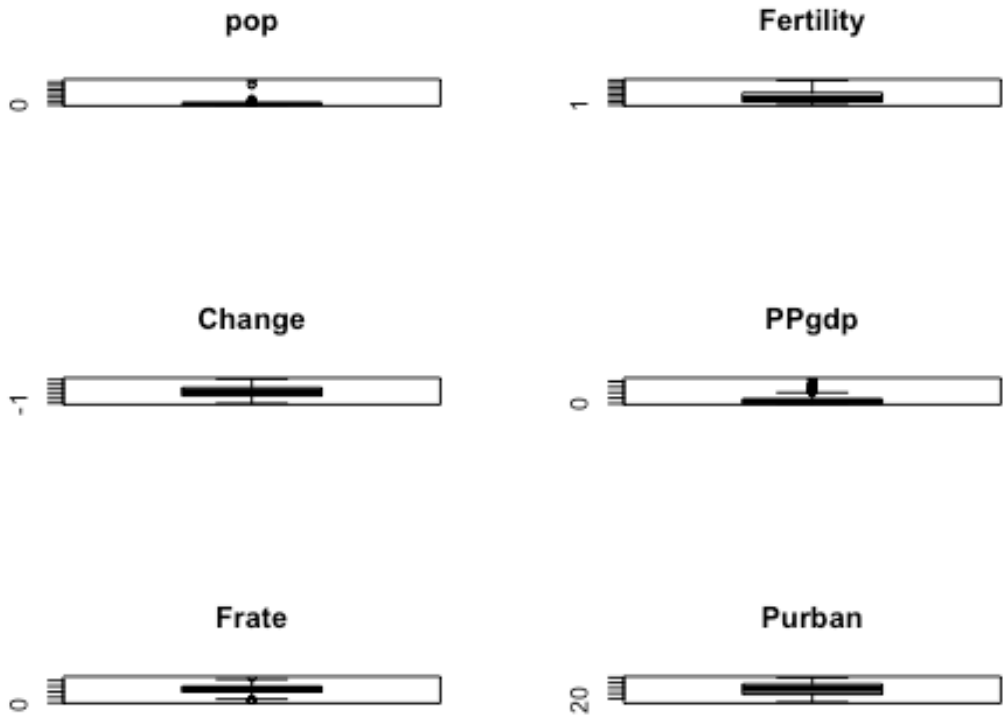
```
scatterplotMatrix(UN3,main='Scatter Plot for the variables')
```




```
UN3_fix = na.omit(UN3)
ggpairs(UN3_fix,title='ggpair Plot for the variables')
```




```
par(mfrow=c(3,2))
boxplot(UN3['Pop'],main='pop')
boxplot(UN3['Fertility'],main='Fertility')
boxplot(UN3['Change'],main='Change')
boxplot(UN3['PPgdp'],main='PPgdp')
boxplot(UN3['Frate'],main='Frate')
boxplot(UN3['Purban'],main='Purban')
```




```
par(mfrow=c(1,1))
```

By looking at the scatterplot and the ggplot, I think obviously PPgdp and Pop need a transformation becasue most of the points are crowded together with several outliers. We cannot match any linearities under that condition. Frate doesnt look like have a great correlation/linearity with the Y variable(ModernC) and other 3 variables: Fertility, Purban and changes looks good (having linearity and correlations).

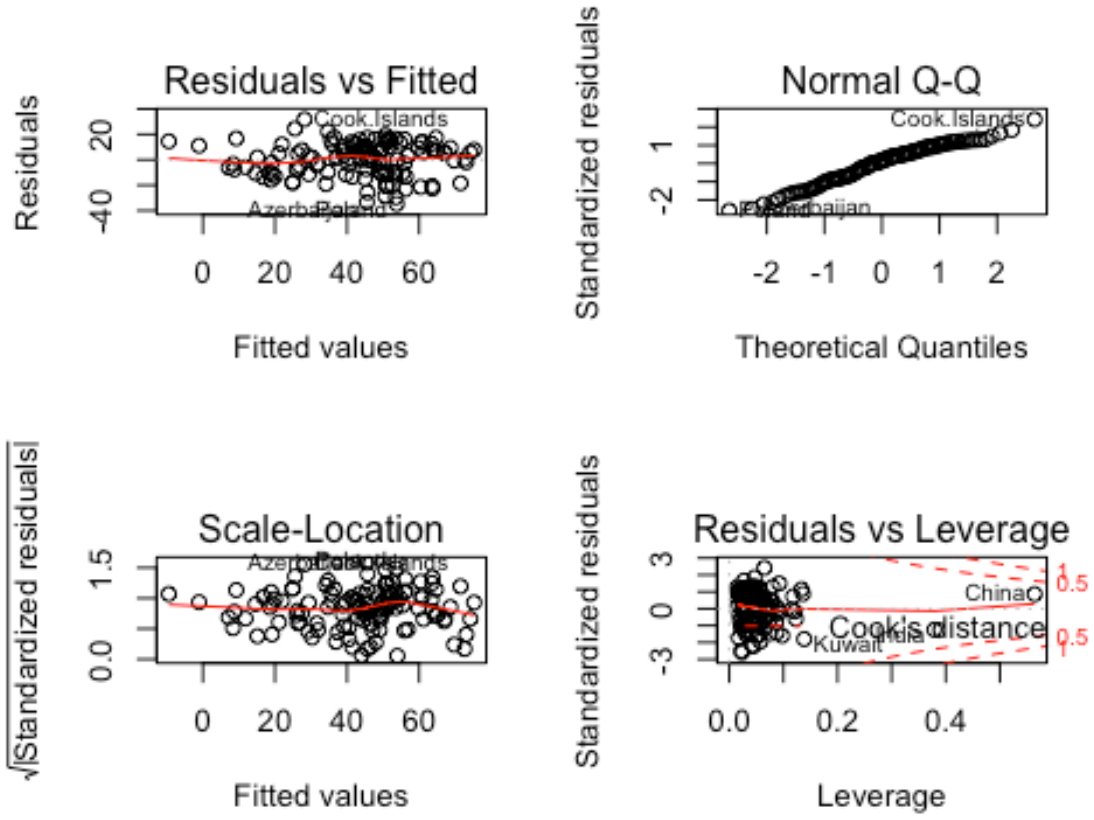
I saw some potential outliers in PPgdp and Pop, we are going to transform these two. I also saw some potential y outliers in Purban, I will check it further later.

Two observations, China and India are obviously from others.

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?

```
model = lm(ModernC ~.,data=UN3)
par(mfrow=c(2,2))
plot(model)
```




```
par(mfrow=c(1,1))
summary(model)

##
## 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
```

By summary, 210-85 = 125 observations are used in the regression.

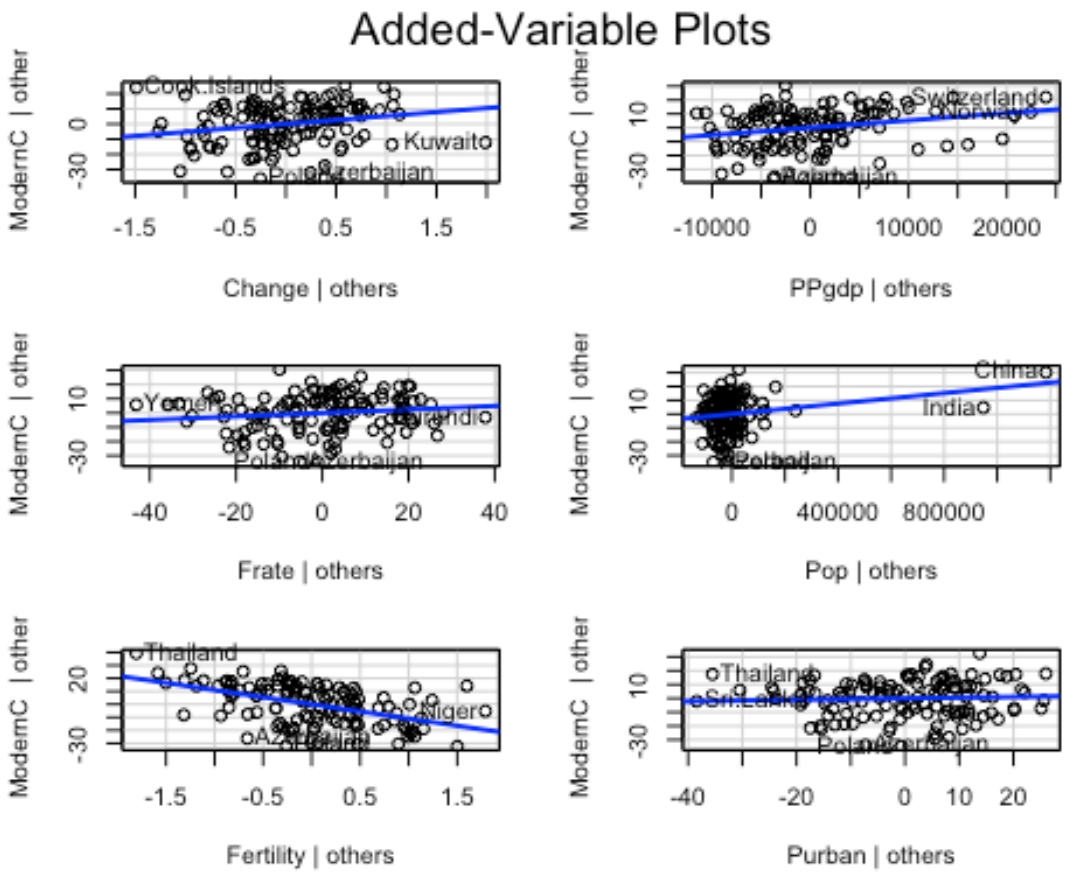
Residuals plots is okay but quite diverse. There are some points having big residuals. but overall, the residual vs fitted plot is a good indication I don't have non-linear relationships since the fitted line is horizontal with no trends.

The normal QQ also looks fine, but a tiny left skewed on large theoretical quantiles. The scale Location plots mentions an equal variance, the line is basically horizontal.

And there's not much outlier, besides China and India having very large leverages.

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?

```
avPlots(model)
```



From the avPlot I am further making sure that I definitely to transform the Pop because most of the points are just crowded together with only two outliers affecting the line. I also want to think about maybe transform the PPgdp even though it's not as bad as the Pop. I will test it in the next questions to see whether transform it or not. But looking at the plot I will definitely transform Pop.

Looking at the Change avPlot, it shows that Kuwait and Cooks island maybe an influential point. And Switzerland and Norway are influnetial to PPgdp, India and China are very influntial to Pop, Thailand is influntial to Fertility.

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.

```
min(UN3_fix['Change'])

## [1] -1.1

UN3_fix['Change'] = UN3_fix['Change'] + 2.1
boxTidwell(ModernC~Pop+PPgdp,~Change+Purban+Fertility+Frate,data=UN3_fi
x)
```

```
##       MLE of lambda Score Statistic (z) Pr(>|z|)
## Pop          0.40749             -0.7874   0.4310
## PPgdp       -0.12921             -1.1410   0.2539
##
## iterations =  4
```

```
powerTransform(as.matrix(UN3_fix)~.,family="bcnPower",data=UN3_fix)
```

```
## Estimated transformation power, lambda
## [1] 0.9999792 0.9992714 0.9999976 0.9999856 0.3251012 0.9993639 0.99
99825
##
## Estimated location, gamma
## [1] 1.000000e-01 1.000000e-01 3.080181e+00 1.000000e-01 1.304196e+06
## [6] 1.000000e-01 1.000000e-01
```

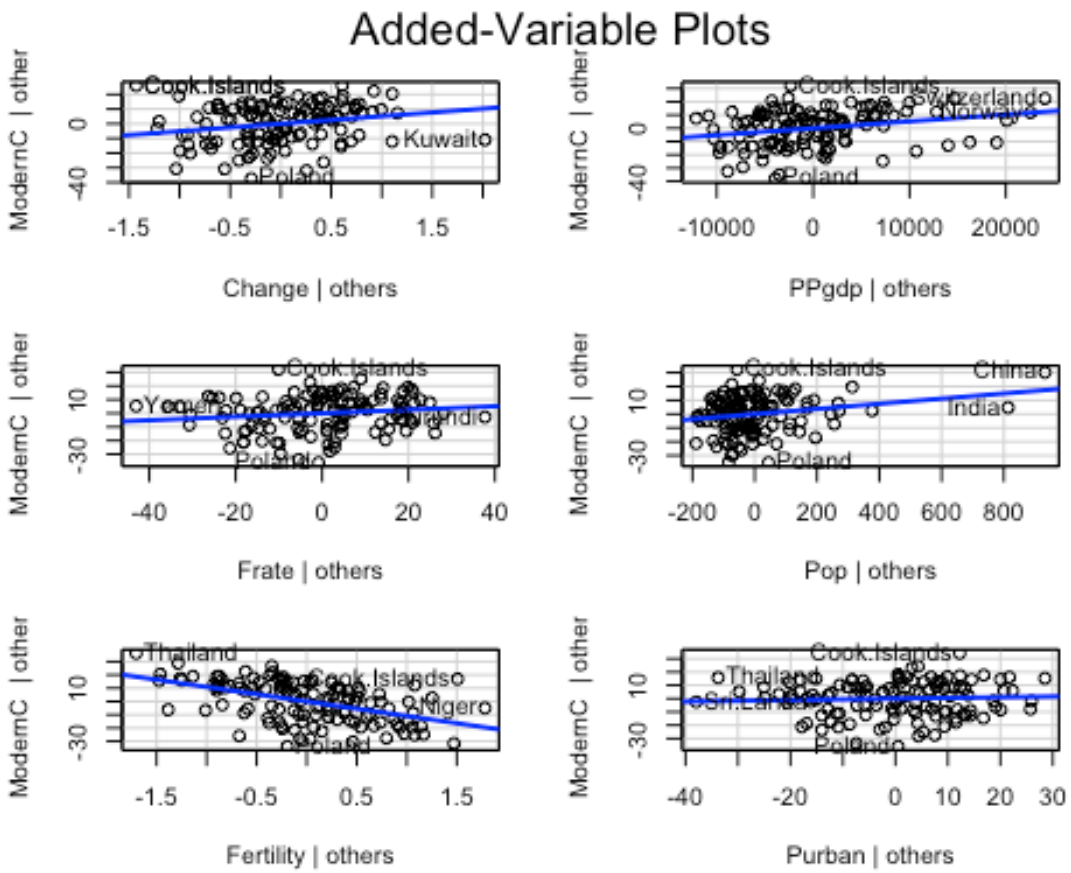
I first transformed the Change to all positive values by adding 2.1 to each entries because the minumum value of Changes is -1.1 so i want to make sure that all values in Changes are above 1.

Then i use the boxTidewell to do the transformation for PPgdp and Pop I discovered that boxTidewill did not suggest us to do any transformations (P value is so large). But by looking at the avplot, I believe at least we have to do a transformation on Pop. So then I try the function powertransform()

By using the powertransform(), it suggests I only have to transform Pop on a degree of 0.33, which is a number between 0.5 and 0. I asked the professor for recommendations and she tells me she will try both of them and she will suggest to use log. I decided to try both log transformation and square root transformation to see which one is better.

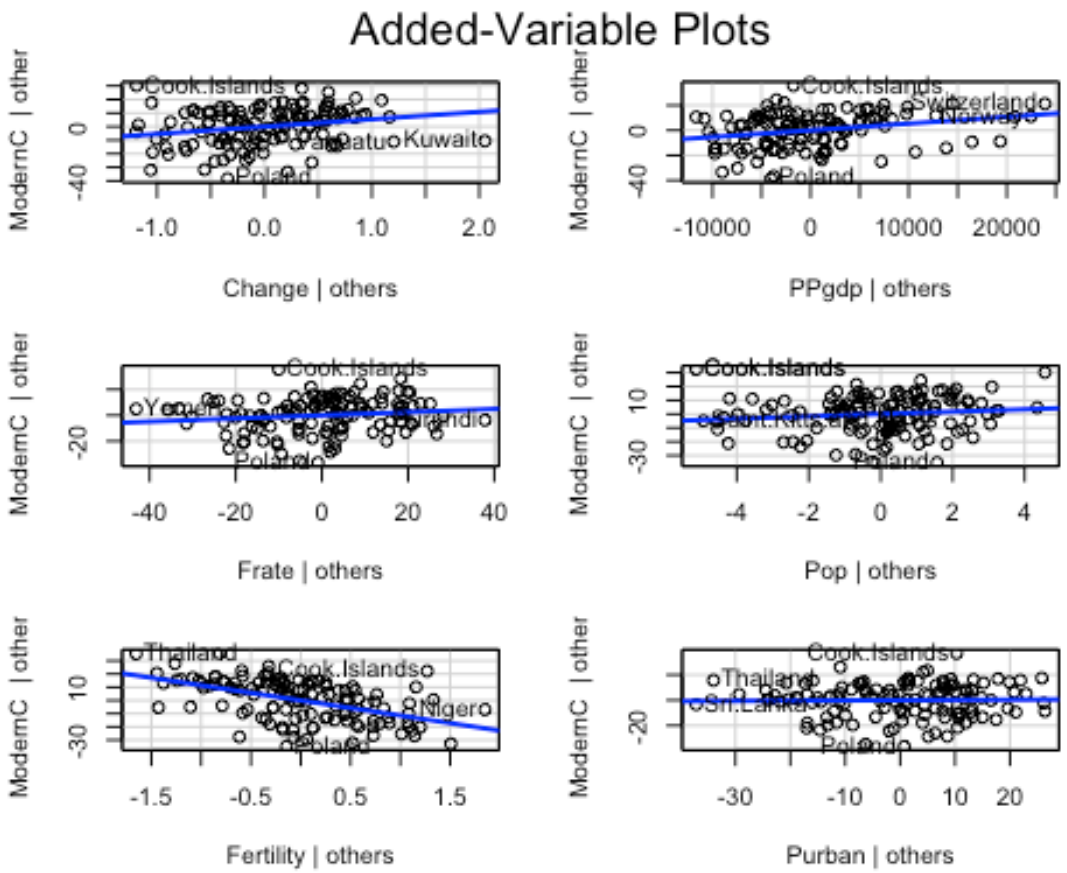
I first tried to see the square root one and see its av plots:

```
UN3_new = UN3_fix
UN3_new['Pop'] = UN3_fix['Pop']^0.5
model2 = lm(ModernC ~.,data=UN3_new)
avPlots(model2)
```

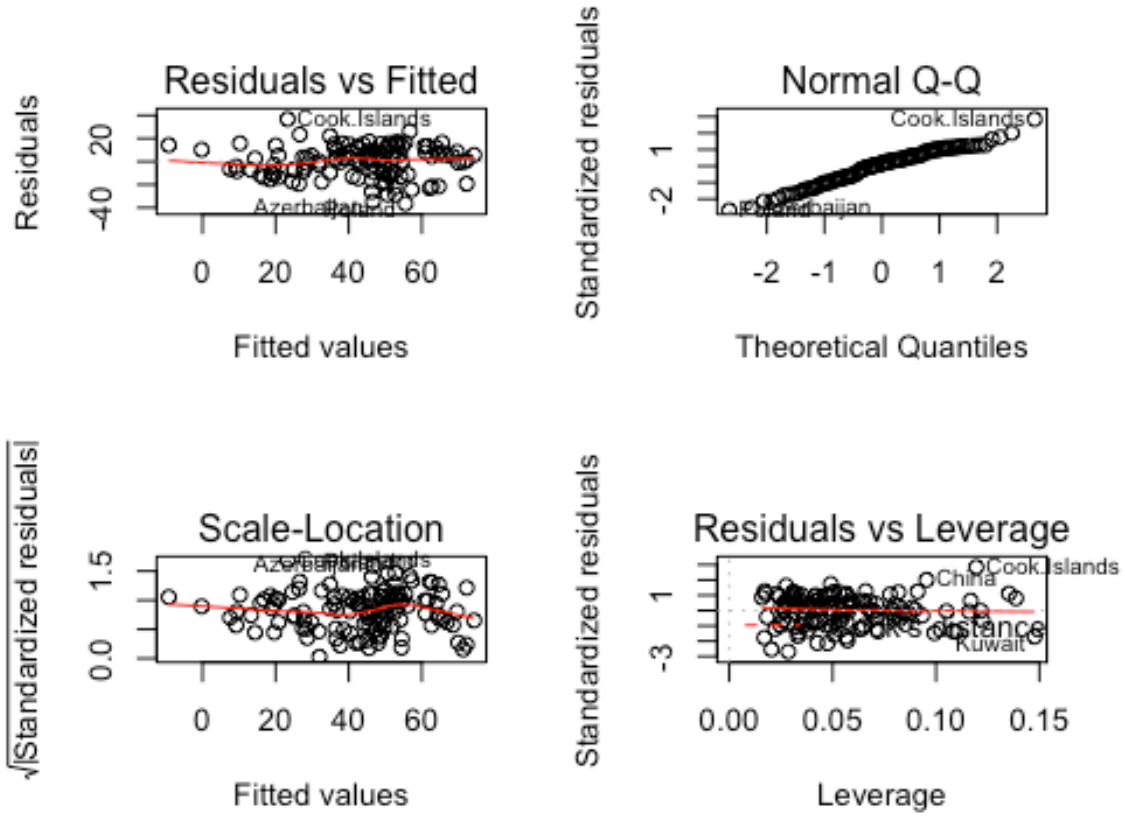


Then I try the log one:

```
UN3_new2 = UN3_fix
UN3_new2['Pop'] = log(UN3_fix['Pop'])
model2_2 = lm(ModernC ~.,data=UN3_new2)
avPlots(model2_2)
```




```
par(mfrow=c(2,2))
plot(model2_2)
```




```
par(mfrow=c(1,1))
```

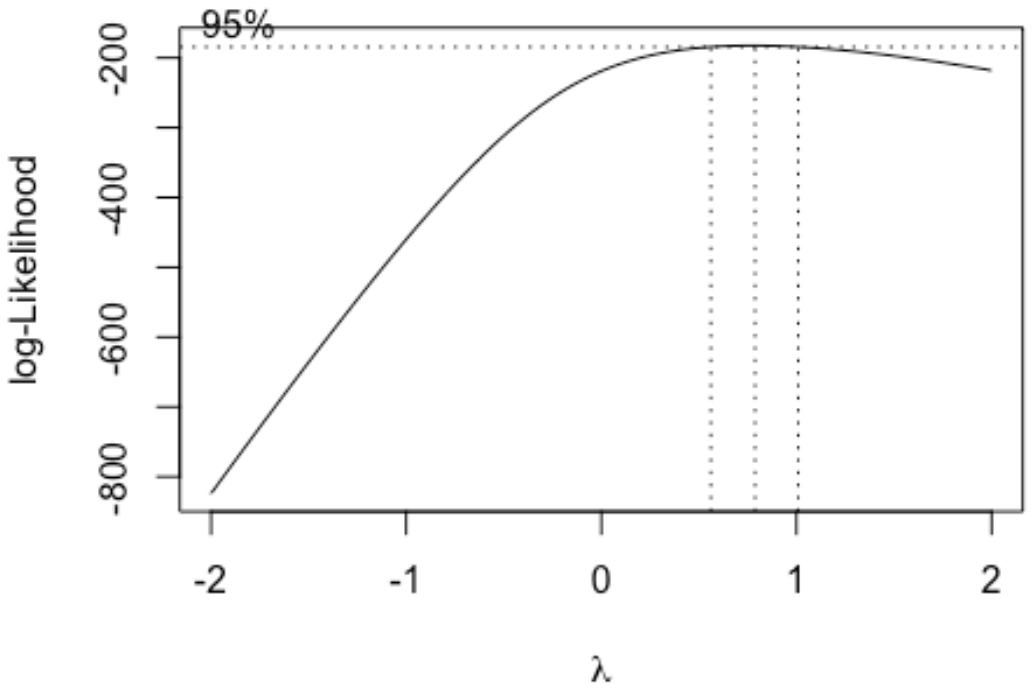
At last I find out the log transformation is better in the new avplots because in the sqrt root transformation pop still looks crowded together.

Therefore, I decided to use the log transformation on Pop. The avPlots for Pop becomes much better.

I also take a look on 4 diagonostic plots and they become better too. For specific, The left skewed problem is better than before. But it looks like there's an outlier problem.

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

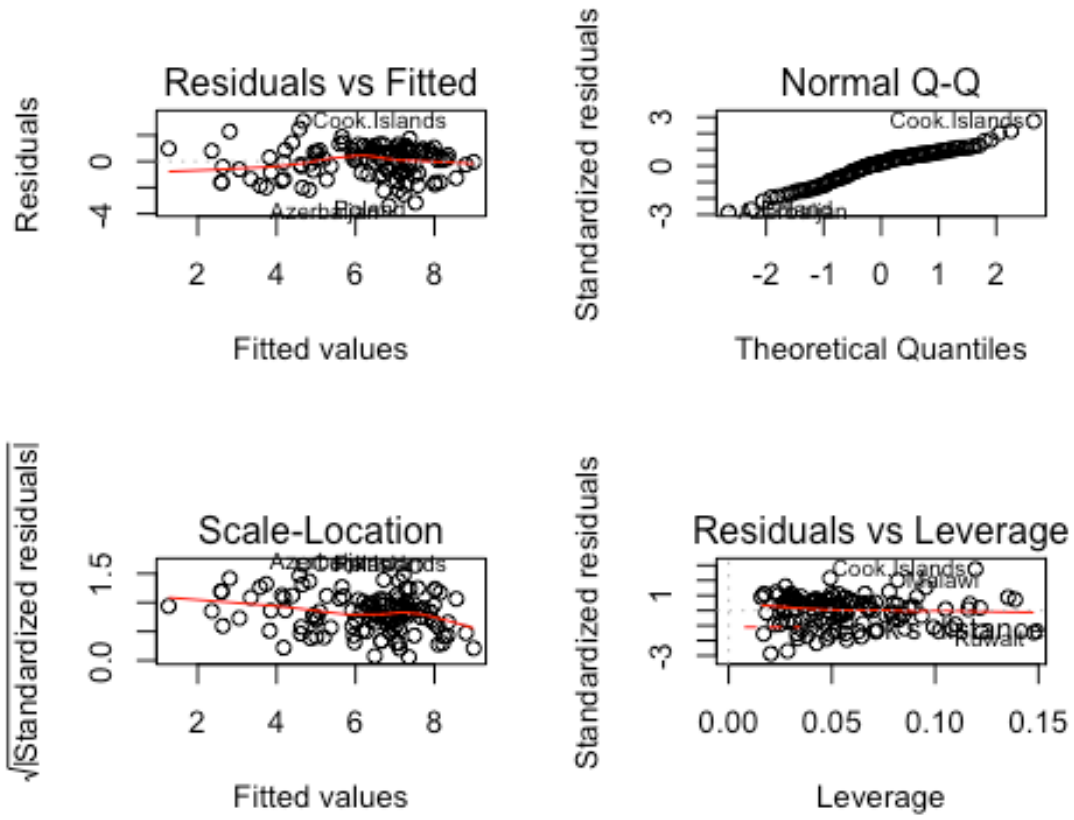
```
boxcox(model2_2)
x= boxcox(model2_2)$x
```



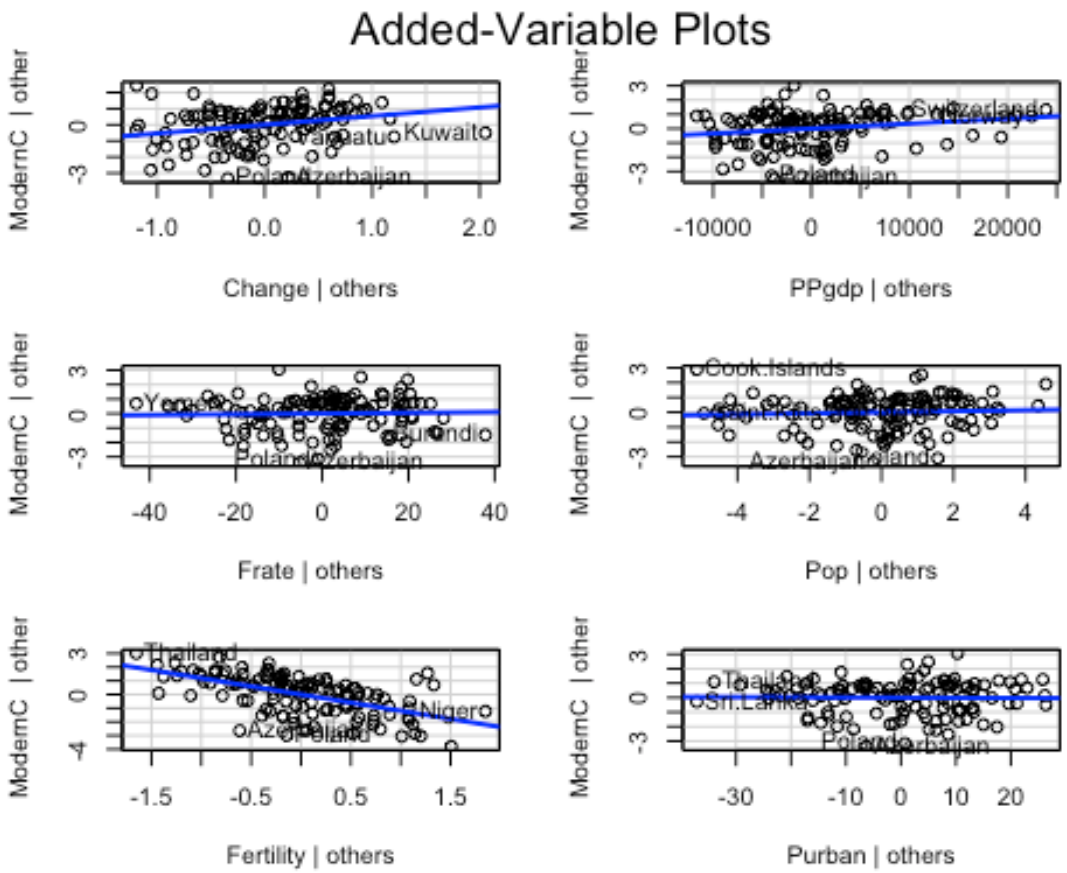

```
y = boxcox(model2_2)$y
table2 = cbind(x,y)
table2=as.data.frame(table2)
head(arrange(table2,desc(y)))

##           x         y
## 1 0.7878788 -182.4408
## 2 0.7474747 -182.4543
## 3 0.8282828 -182.5494
## 4 0.7070707 -182.5982
## 5 0.8686869 -182.7728
## 6 0.6666667 -182.8809
```

Since the interval of lambda includes 1 and I asked the professor, she suggests to not do any transformations to y. But I will still try to fit the y^0.5 to see if it getting anything better.

```
UN3_final = UN3_new2
UN3_final['ModernC'] = UN3_final['ModernC']^0.5
model3 = lm(ModernC ~.,data=UN3_final)
par(mfrow=c(2,2))
plot(model3)
```




```r
par(mfrow=c(1,1))
avPlots(model3)
```




```
summary(model3)

##
## Call:
## lm(formula = ModernC ~ ., data = UN3_final)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3817 -0.8134  0.2133  0.8026  3.0731
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.340e+00  9.631e-01   7.621 6.95e-12 ***
## Change       5.394e-01  1.840e-01   2.931  0.00406 **
## PPgdp        3.469e-05  1.551e-05   2.236  0.02725 *
## Frate        2.652e-03  7.068e-03   0.375  0.70820
## Pop          3.658e-02  5.443e-02   0.672  0.50293
## Fertility   -1.193e+00  1.524e-01  -7.825 2.41e-12 ***
## Purban      -1.369e-03  7.987e-03  -0.171  0.86420
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.191 on 118 degrees of freedom
```

```
## Multiple R-squared:  0.6391, Adjusted R-squared:  0.6208
## F-statistic: 34.83 on 6 and 118 DF,  p-value: < 2.2e-16
```

```
summary(model2_2)
```

```
##
## Call:
## lm(formula = ModernC ~ ., data = UN3_new2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -36.520  -9.676   2.132   9.191  36.723
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.994e+01  1.114e+01   3.585 0.000492 ***
## Change       5.444e+00  2.129e+00   2.557 0.011827 *
## PPgdp        5.384e-04  1.795e-04   3.000 0.003296 **
## Frate        1.226e-01  8.176e-02   1.499 0.136442
## Pop          8.623e-01  6.297e-01   1.369 0.173447
## Fertility   -1.145e+01  1.763e+00  -6.495 2.06e-09 ***
## Purban       2.014e-02  9.240e-02   0.218 0.827855
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.78 on 118 degrees of freedom
## Multiple R-squared:  0.6072, Adjusted R-squared:  0.5872
## F-statistic:  30.4 on 6 and 118 DF,  p-value: < 2.2e-16
```

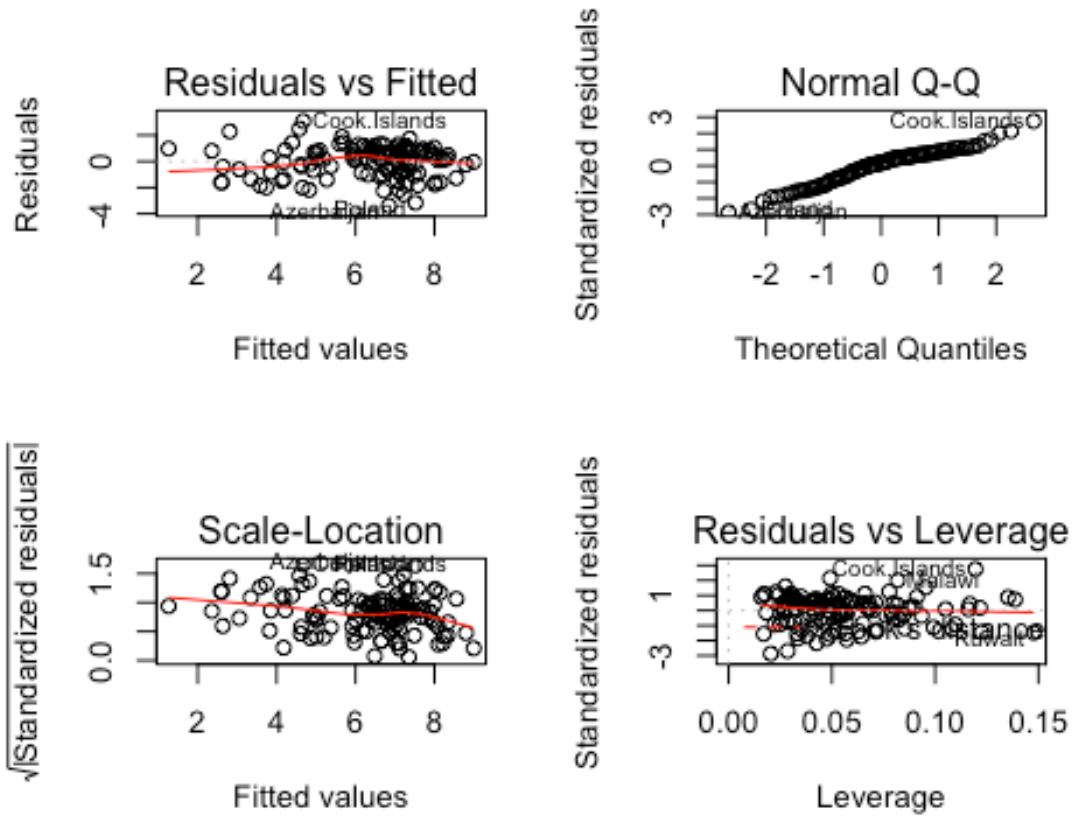
By looking at these plots, I think model2_2 is better then model 3(the model after transforming y)by looking at assumption plots. In the residual fitted plots, model3's residual plots looks not that flat. And the scale location plot looks like there's a decreasing trend.

By looking at avplots, I also think model2_2 is better because the relationship between variables are more clear in model2_2 avplots.

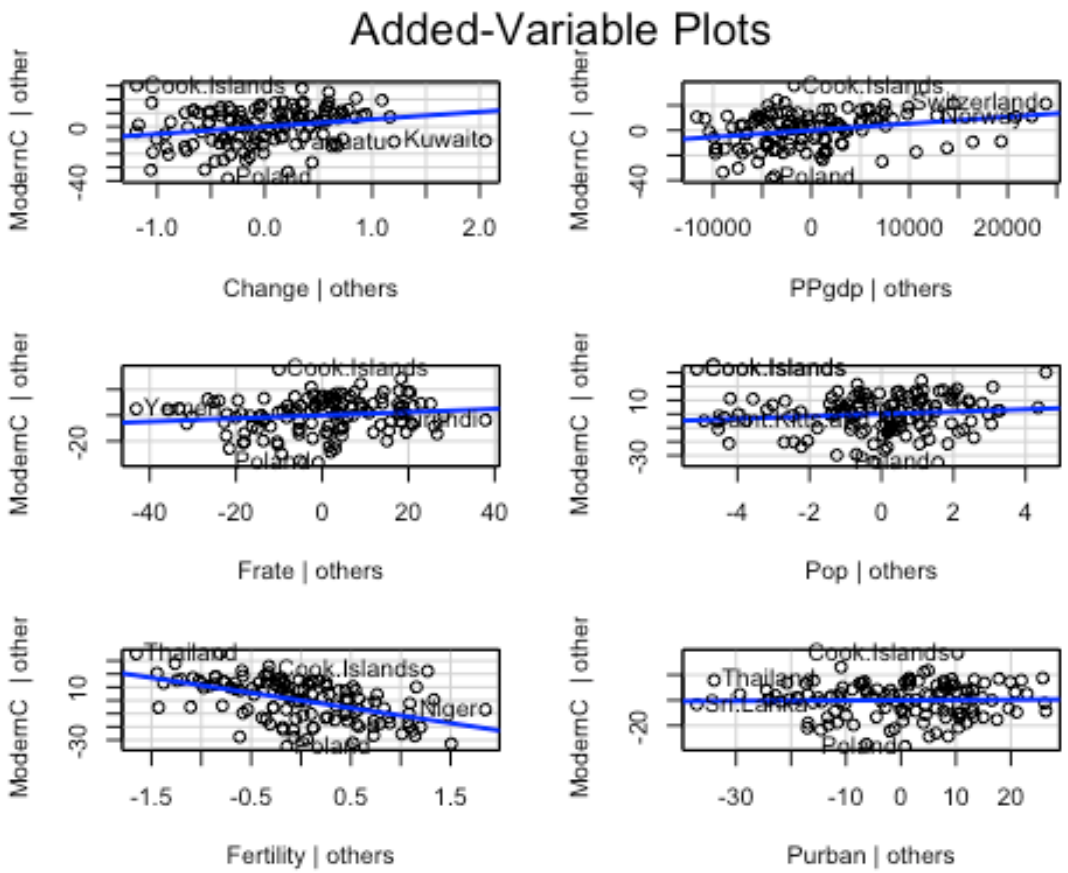
Therefore, I will not transform Y.

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.

```
UN3_final = UN3_new2
UN3_final['ModernC'] = UN3_final['ModernC']
model4 = lm(ModernC ~.,data=UN3_final)
par(mfrow=c(2,2))
plot(model3)
```




```r
par(mfrow=c(1,1))
avPlots(model4)
```

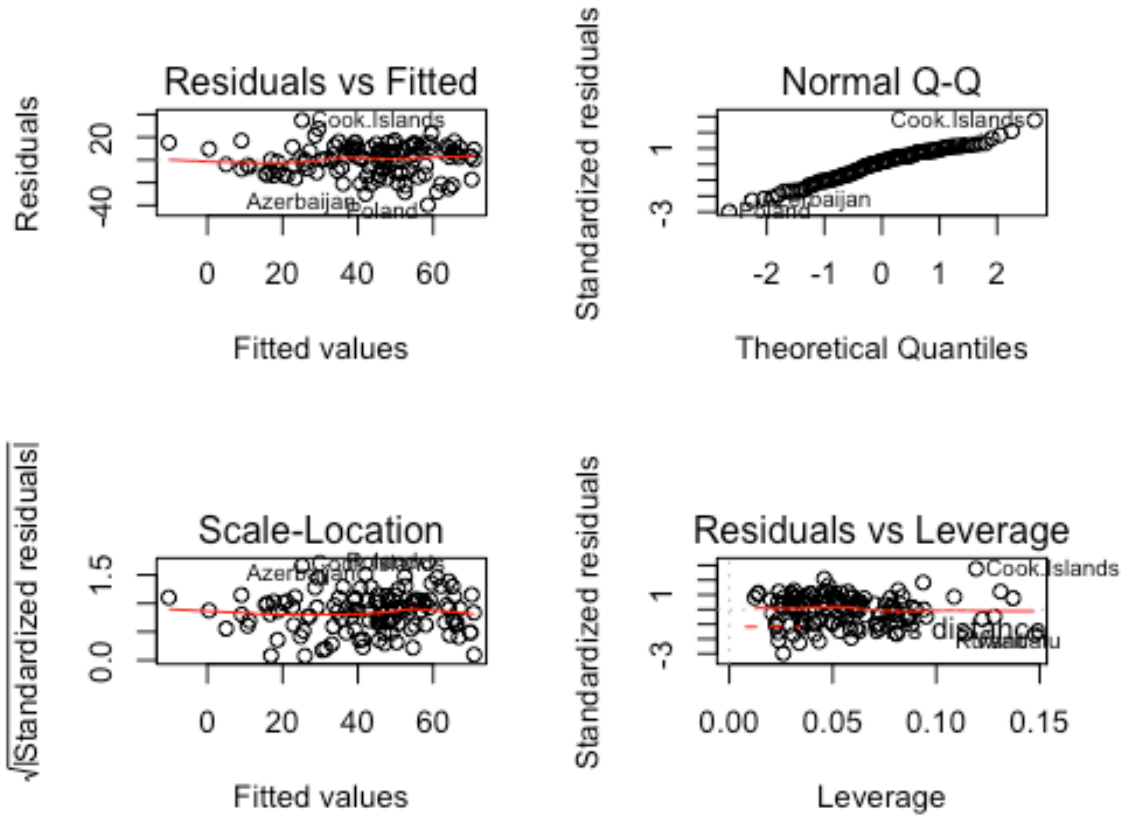



```
summary(model4)

##
## Call:
## lm(formula = ModernC ~ ., data = UN3_final)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -36.520  -9.676   2.132   9.191  36.723
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.994e+01  1.114e+01   3.585 0.000492 ***
## Change       5.444e+00  2.129e+00   2.557 0.011827 *
## PPgdp        5.384e-04  1.795e-04   3.000 0.003296 **
## Frate        1.226e-01  8.176e-02   1.499 0.136442
## Pop          8.623e-01  6.297e-01   1.369 0.173447
## Fertility   -1.145e+01  1.763e+00  -6.495 2.06e-09 ***
## Purban       2.014e-02  9.240e-02   0.218 0.827855
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.78 on 118 degrees of freedom
```

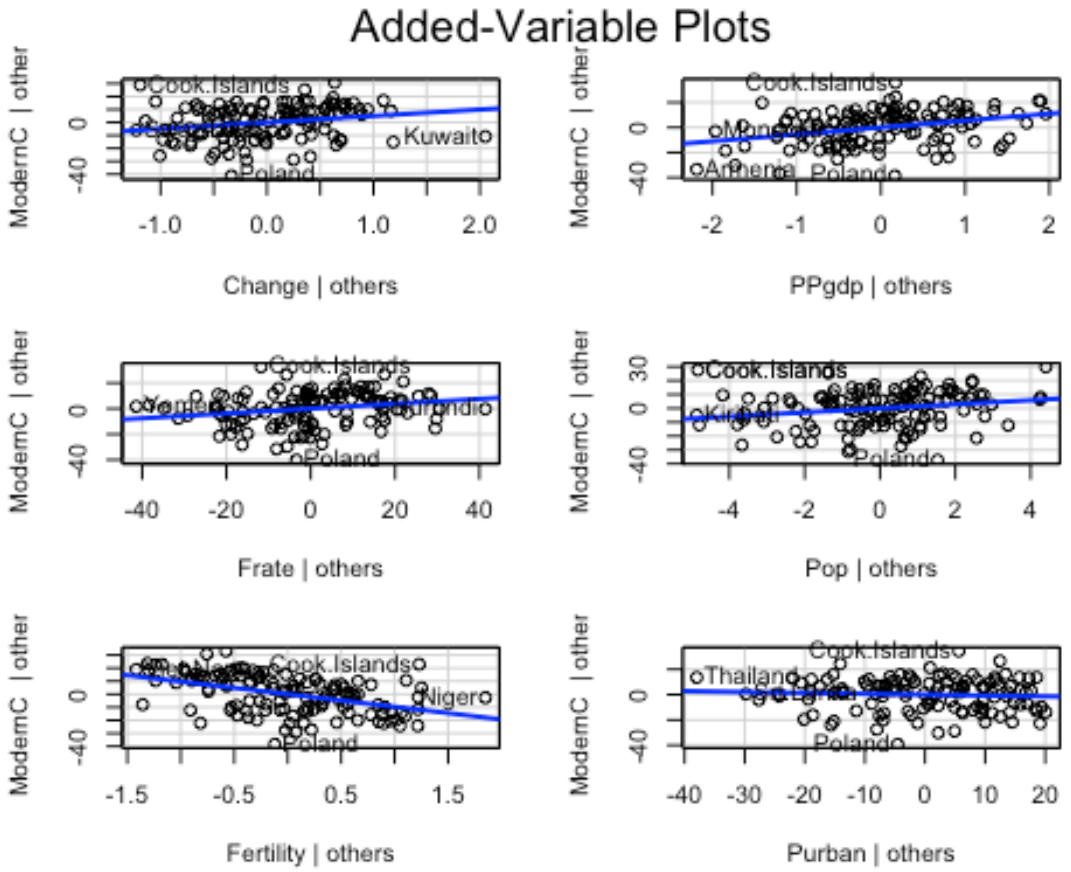
```
## Multiple R-squared:  0.6072, Adjusted R-squared:  0.5872
## F-statistic:  30.4 on 6 and 118 DF,  p-value: < 2.2e-16
```

I am pretty satisfied with all other variables except for the PPgdp, so I decided to check them one more time and I discovered that no further transformation is needed. But I still want to try if PPgdp may need one more transformation. Just have a try:

```
powerTransform(as.matrix(UN3_final)~.,family="bcnPower",data=UN3_final)

## Estimated transformation power, lambda
## [1] 0.9999768 0.9992719 0.9999988 0.9999863 0.9999014 0.9993732 0.99
99956
##
## Estimated location, gamma
## [1] 0.100000 0.100000 1.564386 0.100000 0.100000 0.100000 0.100000

UN3_final2 = UN3_final
UN3_final2['PPgdp'] = log(UN3_final['PPgdp'])
model4_2 = lm(ModernC ~.,data=UN3_final2)
par(mfrow=c(2,2))
plot(model4_2)
```



```
par(mfrow=c(1,1))
avPlots(model4_2)
```



The plots did not look like having a big difference, especially qq plots become even worse, so we look at the summary of the model:

```
summary(model4_2)

##
## Call:
## lm(formula = ModernC ~ ., data = UN3_final2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -39.597  -9.540   2.238  10.024  34.840
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6.36974   14.22449  -0.448 0.655118
## Change       4.99296    2.07709   2.404 0.017781 *
## PPgdp        5.50728    1.40505   3.920 0.000149 ***
## Frate        0.18939    0.07711   2.456 0.015500 *
## Pop          1.47207    0.62875   2.341 0.020897 *
## Fertility   -9.67594    1.76561  -5.480 2.44e-07 ***
```

```
## Purban      -0.07077    0.09760  -0.725 0.469829
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.44 on 118 degrees of freedom
## Multiple R-squared:  0.626,  Adjusted R-squared:  0.6069
## F-statistic: 32.91 on 6 and 118 DF,  p-value: < 2.2e-16
```

```
summary(model4)
```

```
##
## Call:
## lm(formula = ModernC ~ ., data = UN3_final)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -36.520  -9.676   2.132   9.191  36.723
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.994e+01  1.114e+01   3.585 0.000492 ***
## Change       5.444e+00  2.129e+00   2.557 0.011827 *
## PPgdp        5.384e-04  1.795e-04   3.000 0.003296 **
## Frate        1.226e-01  8.176e-02   1.499 0.136442
## Pop          8.623e-01  6.297e-01   1.369 0.173447
## Fertility   -1.145e+01  1.763e+00  -6.495 2.06e-09 ***
## Purban       2.014e-02  9.240e-02   0.218 0.827855
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.78 on 118 degrees of freedom
## Multiple R-squared:  0.6072, Adjusted R-squared:  0.5872
## F-statistic:  30.4 on 6 and 118 DF,  p-value: < 2.2e-16
```

```
summary(model2_2)
```

```
##
## Call:
## lm(formula = ModernC ~ ., data = UN3_new2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -36.520  -9.676   2.132   9.191  36.723
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.994e+01  1.114e+01   3.585 0.000492 ***
## Change       5.444e+00  2.129e+00   2.557 0.011827 *
## PPgdp        5.384e-04  1.795e-04   3.000 0.003296 **
## Frate        1.226e-01  8.176e-02   1.499 0.136442
## Pop          8.623e-01  6.297e-01   1.369 0.173447
```

```
## Fertility   -1.145e+01  1.763e+00  -6.495 2.06e-09 ***
## Purban       2.014e-02  9.240e-02   0.218 0.827855
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.78 on 118 degrees of freedom
## Multiple R-squared:  0.6072, Adjusted R-squared:  0.5872
## F-statistic:  30.4 on 6 and 118 DF,  p-value: < 2.2e-16
```

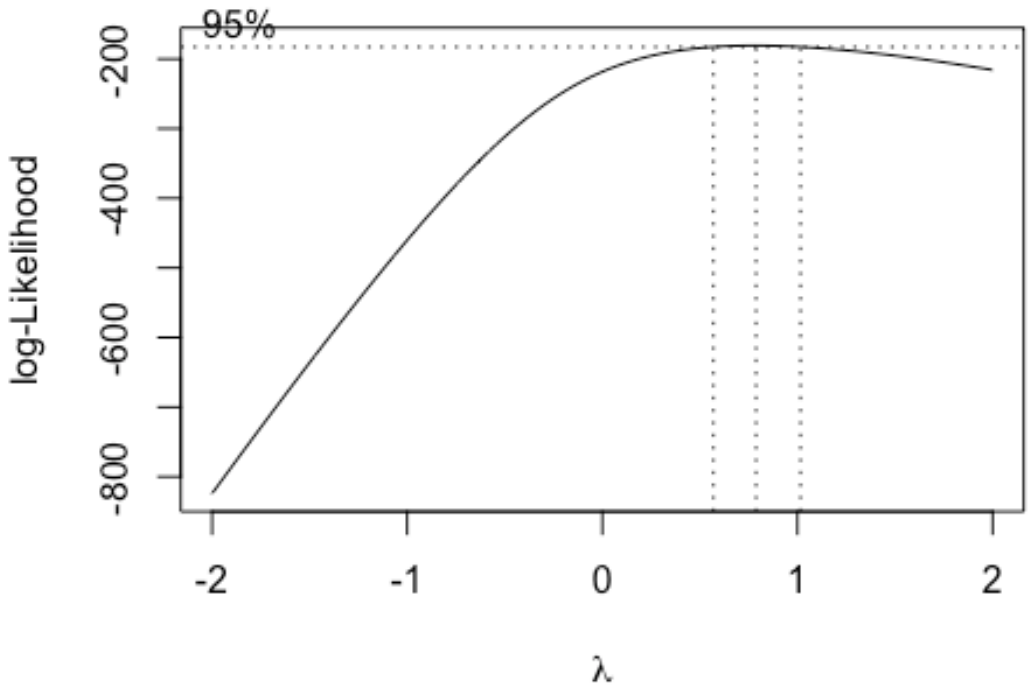
It basically only have very very slightly difference for the model, less then 0.15 changes in adjusted r squared and nearly no changes on significance level on the PPgdp or on the diagnostic plots. So I think do one more transformation is redundant.

Therefore, after several test, I conclude my first model is the best model:

```
summary(model2_2)

##
## Call:
## lm(formula = ModernC ~ ., data = UN3_new2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -36.520  -9.676   2.132   9.191  36.723
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.994e+01  1.114e+01   3.585 0.000492 ***
## Change       5.444e+00  2.129e+00   2.557 0.011827 *
## PPgdp        5.384e-04  1.795e-04   3.000 0.003296 **
## Frate        1.226e-01  8.176e-02   1.499 0.136442
## Pop          8.623e-01  6.297e-01   1.369 0.173447
## Fertility   -1.145e+01  1.763e+00  -6.495 2.06e-09 ***
## Purban       2.014e-02  9.240e-02   0.218 0.827855
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.78 on 118 degrees of freedom
## Multiple R-squared:  0.6072, Adjusted R-squared:  0.5872
## F-statistic:  30.4 on 6 and 118 DF,  p-value: < 2.2e-16
```

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?

```
boxcox(model)
x2= boxcox(model)$x
```




```
y2 = boxcox(model)$y
table3 = cbind(x2,y2)
table3=as.data.frame(table3)
head(arrange(table3,desc(y2)))
```

```
##          x2        y2
## 1 0.7878788 -180.7493
## 2 0.7474747 -180.7851
## 3 0.8282828 -180.8375
## 4 0.7070707 -180.9532
## 5 0.8686869 -181.0420
## 6 0.6666667 -181.2625
```

```
#so there's no transformation for y
UN3_q9 = UN3_fix
UN3_q9['ModernC'] = UN3_q9['ModernC']
model_q9 = lm(ModernC ~.,data=UN3_q9)
boxTidwell(ModernC~Pop+PPgdp,~Change+Purban+Fertility+Frate,data=UN3_q9)
```

```
##       MLE of lambda Score Statistic (z) Pr(>|z|)
## Pop         0.40749             -0.7874   0.4310
## PPgdp      -0.12921             -1.1410   0.2539
##
## iterations =  4
```

```r
powerTransform(as.matrix(UN3_q9)~.,family="bcnPower",data=UN3_q9)
```

```
## Estimated transformation power, lambda
## [1] 0.9999792 0.9992714 0.9999976 0.9999856 0.3251012 0.9993639 0.99
99825
##
## Estimated location, gamma
## [1] 1.000000e-01 1.000000e-01 3.080181e+00 1.000000e-01 1.304196e+06
## [6] 1.000000e-01 1.000000e-01
```

```r
UN3_Q9_2 = UN3_q9
UN3_Q9_2['Pop'] = log(UN3_q9['Pop'])
model_q9 = lm(ModernC ~.,data=UN3_Q9_2)
summary(model_q9)
```

```
##
## Call:
## lm(formula = ModernC ~ ., data = UN3_Q9_2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -36.520  -9.676   2.132   9.191  36.723
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.994e+01  1.114e+01   3.585 0.000492 ***
## Change       5.444e+00  2.129e+00   2.557 0.011827 *
## PPgdp        5.384e-04  1.795e-04   3.000 0.003296 **
## Frate        1.226e-01  8.176e-02   1.499 0.136442
## Pop          8.623e-01  6.297e-01   1.369 0.173447
## Fertility   -1.145e+01  1.763e+00  -6.495 2.06e-09 ***
## Purban       2.014e-02  9.240e-02   0.218 0.827855
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.78 on 118 degrees of freedom
## Multiple R-squared:  0.6072, Adjusted R-squared:  0.5872
## F-statistic:  30.4 on 6 and 118 DF,  p-value: < 2.2e-16
```

```r
summary(model2_2)
```

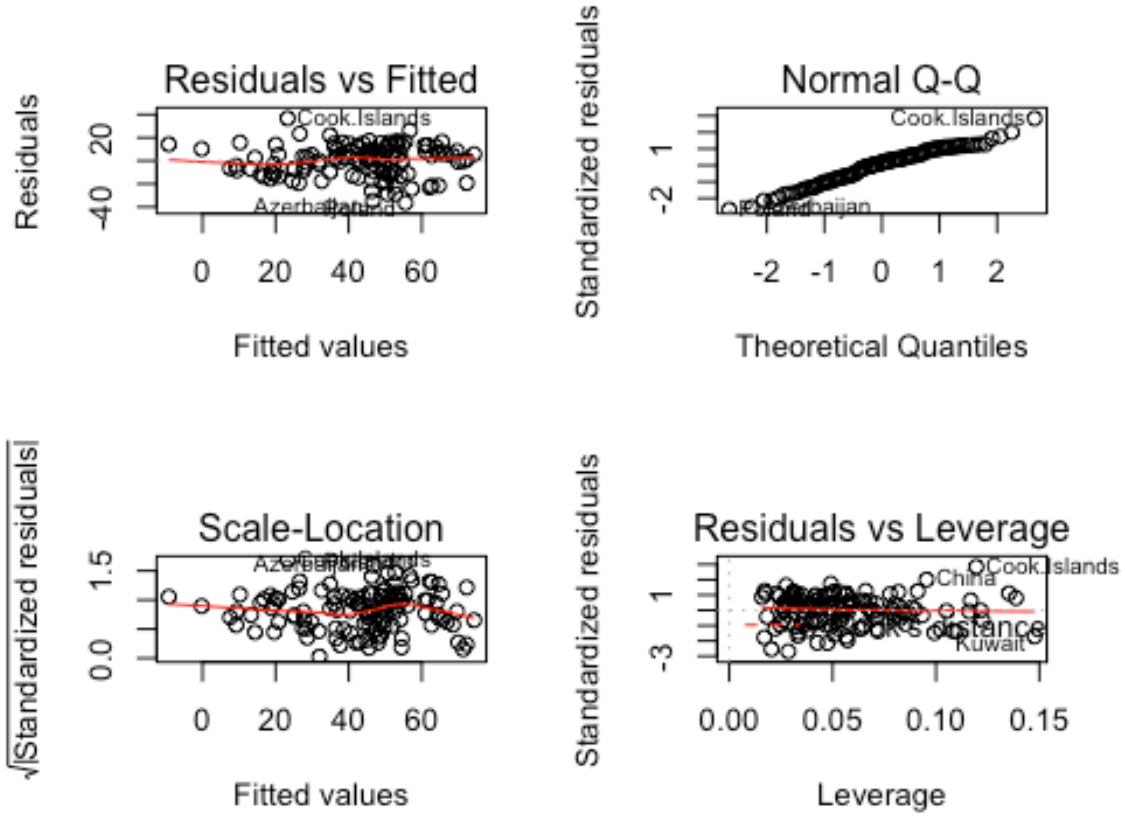
```
##
## Call:
## lm(formula = ModernC ~ ., data = UN3_new2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -36.520  -9.676   2.132   9.191  36.723
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  3.994e+01  1.114e+01   3.585 0.000492 ***
## Change       5.444e+00  2.129e+00   2.557 0.011827 *
## PPgdp        5.384e-04  1.795e-04   3.000 0.003296 **
## Frate        1.226e-01  8.176e-02   1.499 0.136442
## Pop          8.623e-01  6.297e-01   1.369 0.173447
## Fertility   -1.145e+01  1.763e+00  -6.495 2.06e-09 ***
## Purban       2.014e-02  9.240e-02   0.218 0.827855
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.78 on 118 degrees of freedom
## Multiple R-squared:  0.6072, Adjusted R-squared:  0.5872
## F-statistic:  30.4 on 6 and 118 DF,  p-value: < 2.2e-16
```

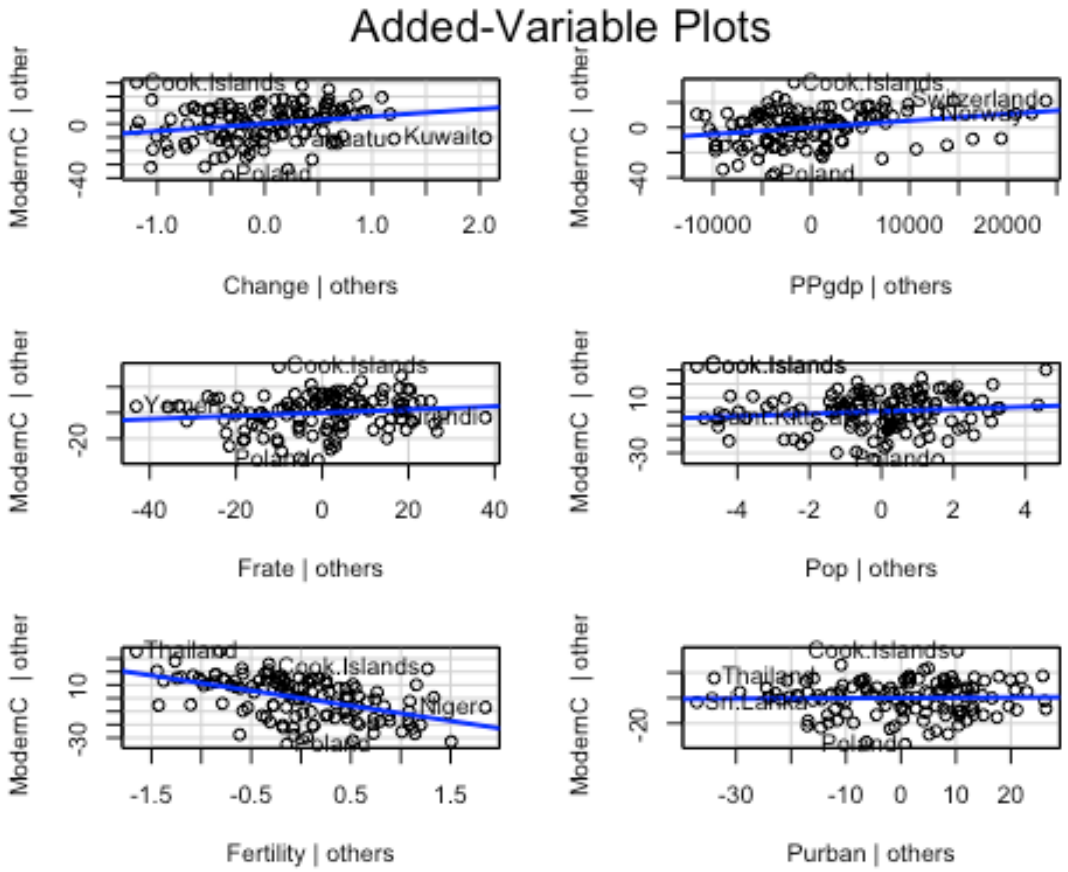
I got the excatly the same model as question 8 because I did not do any transformation to Y as reasons mentioned above.

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.

```
par(mfrow=c(2,2))
plot(model2_2)
```



```
par(mfrow=c(1,1))
avPlots(model2_2)
```

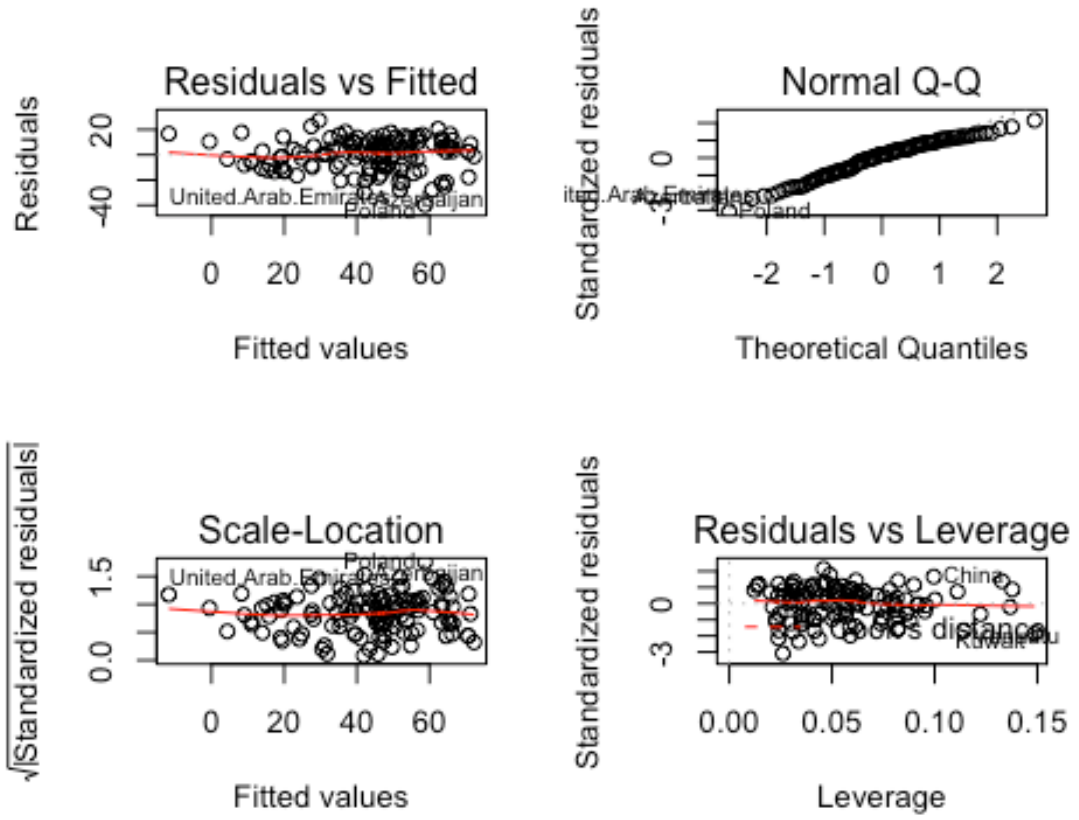


By looking at the data, I think the data point cook islands is an outlier because in the avPlots nearly every plots Cook Island exists as an outlier. so I will remove it to see what happens:

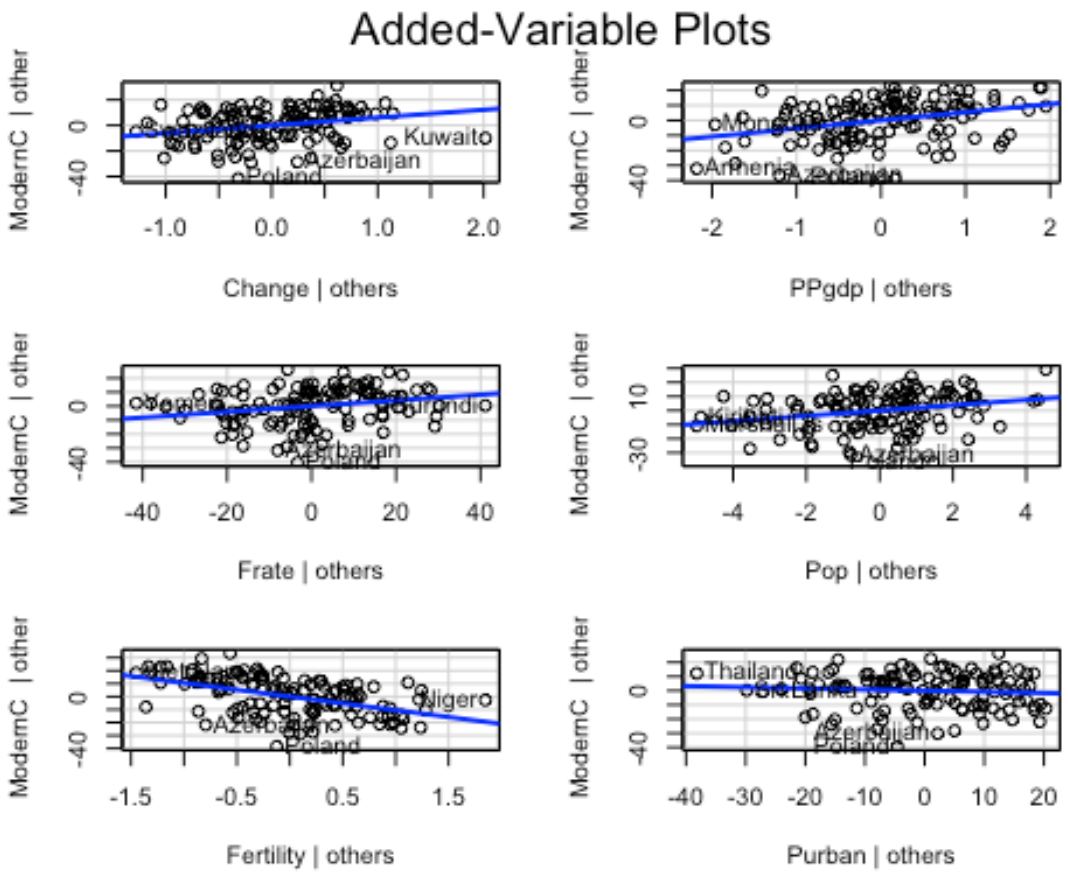
```
UN3_final3 = UN3_final2[-28,]
model2_3 = lm(ModernC ~.,data=UN3_final3)
summary(model2_3)

##
## Call:
## lm(formula = ModernC ~ ., data = UN3_final3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -39.760  -9.209   2.442   9.791  27.380
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -11.24467   13.92178  -0.808  0.42090
## Change        6.11720    2.05580   2.976  0.00355 **
```

```
## PPgdp          5.43342    1.36492   3.981  0.00012 ***
## Frate          0.20474    0.07509   2.727  0.00738 **
## Pop            1.89052    0.62817   3.010  0.00320 **
## Fertility    -10.51515    1.74010  -6.043 1.85e-08 ***
## Purban        -0.08248    0.09488  -0.869  0.38646
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.06 on 117 degrees of freedom
## Multiple R-squared:  0.6484, Adjusted R-squared:  0.6304
## F-statistic: 35.96 on 6 and 117 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(model2_3)
```




```
par(mfrow=c(1,1))
avPlots(model2_3)
```



Actually I do think scale location plot and fitted residual plot is better than before and by looking at the summary the adjusted R squared increased a lot. So I decided to remove Cooks island from my model.

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!

```
coefficients(model2_3)

##  (Intercept)       Change        PPgdp        Frate          Pop
## -11.24467297   6.11719718   5.43342037   0.20473549   1.89052282
##    Fertility       Purban
## -10.51514514  -0.08248294

summary(model2_3)

##
## Call:
## lm(formula = ModernC ~ ., data = UN3_final3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -39.760  -9.209   2.442   9.791  27.380
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -11.24467   13.92178  -0.808  0.42090
## Change        6.11720    2.05580   2.976  0.00355 **
## PPgdp         5.43342    1.36492   3.981  0.00012 ***
## Frate         0.20474    0.07509   2.727  0.00738 **
## Pop           1.89052    0.62817   3.010  0.00320 **
## Fertility   -10.51515    1.74010  -6.043 1.85e-08 ***
## Purban       -0.08248    0.09488  -0.869  0.38646
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.06 on 117 degrees of freedom
## Multiple R-squared:  0.6484, Adjusted R-squared:  0.6304
## F-statistic: 35.96 on 6 and 117 DF,  p-value: < 2.2e-16
```

```
rownames = names(model2_3$coefficients)
ci = matrix(data=NA, nrow = 7, ncol = 2)
for(i in 1:length(rownames)){
  ci[i,] = confint(model2_3,rownames[i],level=0.95)
}
ci = as.data.frame(ci)
rownames(ci) = colnames(UN3_final3)
rownames(ci)[1] = 'Intercept'
colnames(ci) = c('2.5%','97.5%')
rownames(ci)[5] = "Pop (10% increase)"
ci["Pop (10% increase)", ] = ci["Pop (10% increase)", ] * log(110 / 100)
kable(ci,format='markdown')
```

	2.5%	97.5%
Intercept	-38.8160336	16.3266877
Change	2.0457999	10.1885944
PPgdp	2.7302585	8.1365822
Frate	0.0560259	0.3534451
Pop (10% increase)	0.0616143	0.2987578
Fertility	-13.9613171	-7.0689732
Purban	-0.2703950	0.1054291

Interpretation: (a) Change: For each one point increase in the annual population growth rate, the ModernC (Percent of unmarried women using a modern method of contraception.) will increase by 6.117.

(b) Frate: For each one point increase in the Frate (percent of female over age 15 economically active), the ModernC(Percent of unmarried women using a modern method of contraception.) will increase by 0.20474.

(c) Fertility: For each one point increase in Fertility (the expected number of live births per female), the ModernC (Percent of unmarried women using a modern method of contraception.) will decrease by 10.515.

(d) Purban: For each one point increase in Purban (percent of population that is urban), the ModernC (Percent of unmarried women using a modern method of contraception.) will decrease by 0.0825.

(e) Pop: For each 10% increase in Population, the ModernC (Percent of unmarried women using a modern method of contraception.) will increase by 1.89052 * log(1.1) = 0.1802.

(f) PPgdp: For each one point increase in PPgdp (per Capita 2001 GDP), The modernC (Percent of unmarried women using a modern method of contraception.) will increase by 5.433.

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

```
summary(model2_3)

##
## Call:
## lm(formula = ModernC ~ ., data = UN3_final3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -39.760  -9.209   2.442   9.791  27.380
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -11.24467   13.92178  -0.808  0.42090
## Change        6.11720    2.05580   2.976  0.00355 **
## PPgdp         5.43342    1.36492   3.981  0.00012 ***
## Frate         0.20474    0.07509   2.727  0.00738 **
## Pop           1.89052    0.62817   3.010  0.00320 **
## Fertility   -10.51515    1.74010  -6.043 1.85e-08 ***
## Purban       -0.08248    0.09488  -0.869  0.38646
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.06 on 117 degrees of freedom
## Multiple R-squared:  0.6484, Adjusted R-squared:  0.6304
## F-statistic: 35.96 on 6 and 117 DF,  p-value: < 2.2e-16
```

My final model is:

ModernC = -11.24467 + 6.117Change + + 5.433PPgdp + 0.205Frate - 10.515Fertility - 0.0825Purban + 1.89log(Pop)

My final model excludes Cook island because i think it's highly influential and are outliers in all av plots. I also applied log transformation to Pop suggested by powerTransformation() and av plots.

From the summary plot we can see that all variables except Purban are very significant. Change, PPgdp, Frate and log(pop) are positive correlated to ModernC and Fertility, Purban is negatively related to ModernC. Two of there are pretty interesting: First is the relationship between Change and ModernC, which means higher population growth rate correlates to higher Percent of unmarried women using a modern method of contraception, that's kind of contradiction. Similar things happending on log(pop) and ModernC means that larger population implies Percent of unmarried women using a modern method of contraception. Maybe further research needed on that topics. Other coefficients are pretty reasonable.

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.*

Let $H = X_{(j)}(X_{(j)}^T X_{(j)})^{-1} X_{(j)}^T$

$e_Y = \hat{B_0} + \hat{B_1}\, e_{x_j}$

So $(I-H)y = \hat{B_0} + \hat{B_1}\,(I-H)x_j$

since $\hat{B_1} = (X^T X)X^T y$,for here let $X = (I-H)x_j$

then the equation becomes:

$(I-H)y = \hat{B_0}\, 1_n + [x_j^T(I-H)(I-H)x_j]^{-1}[(I-H)x_j]^T(I-H)y(I-H)x_j$

$(I-H)y = \hat{B_0}\, 1_n + [x_j^T(I-H)x_j]^{-1}x_j^T(I-H)y(I-H)x_j$

multiply both sides by x_j^T at the beginning:

$x_j^T(I-H)y = x_j^T\,\hat{B_0}\, 1_n + x_j^T[x_j^T(I-H)x_j]^{-1}x_j^T(I-H)y(I-H)x_j$

$x_j^T(I-H)y = x_j^T 1_n\,\hat{B_0} + x_j^T(I-H)x_j[x_j^T(I-H)x_j]^{-1}x_j^T(I-H)y$

since $[x_j^T(I-H)x_j]^{-1}$ and $x_j^T(I-H)y$ are both a numeric value.

Then:

$x_j^T(I-H)y = \sum x_j\,\hat{B_0} + x_j^T(I-H)y$

Therefore:

$\sum x_j \hat{B_0} = 0$

$\hat{B_0} = 0$

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 a manually constructed added variable plot for one of the predictors in Ex. 10 is the same as the estimate from your model.

```
model6 = lm(ModernC ~ Change + PPgdp + Frate + Pop + Purban,data=UN3_fi
nal3)
UN3_predict = UN3_final3
UN3_predict['prediction'] = predict(model6,UN3_final3)
UN3_predict['e'] = UN3_predict['ModernC'] - UN3_predict['prediction']

model4 = lm(Fertility ~ Change + PPgdp + Frate + Pop + Purban,data=UN3_
final3)
UN3_predict2 = UN3_final3
UN3_predict2['predictionF'] = predict(model4,UN3_final3)
UN3_predict2['eF'] = UN3_predict2['Fertility'] - UN3_predict2['predicti
onF']

e = c(UN3_predict['e'],UN3_predict2['eF'])
e = as.data.frame(e)
model5 = lm(e~eF, data= e)
model5

##
## Call:
## lm(formula = e ~ eF, data = e)
##
## Coefficients:
## (Intercept)           eF
##  -1.114e-14   -1.052e+01
```

By selecting Fertility as my variable tested, Basically what I did is that I first regress the ModernC(Y) on all other variables(Xi) besides Fertility. And I made a prediction for Y and calculate eY at that situation.

Then I regress Fertility on all other X variables and predict for the Fertility and calculate eX at this situation.

At last I regress eY ~ eX and getting the slope coefficient of eX is the same as the coefficients of Fertility in my final model(model2_3).