HW2 STA521 Fall18

[Your Name Here, netid and github username here]

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*

```
library(alr3)
```

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

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

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

```
## Loading required package: ggplot2
```

```
library(dplyr)
```

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

```
## 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(outliers)
'%!in%' <- function(x,y)!('%in%'(x,y))
```

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

```
print(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
```

```
paste(length(colnames(is.na(UN3))),c('varibles have missing data, they are:'))
```

```
## [1] "7 varibles have missing data, they are:"
```

```
paste(colnames(is.na(UN3)),collapse=', ')
```

```
## [1] "ModernC, Change, PPgdp, Frate, Pop, Fertility, Purban"
```

```
print('All the variables are quantitative')
```

```
## [1] "All the variables are quantitative"
```

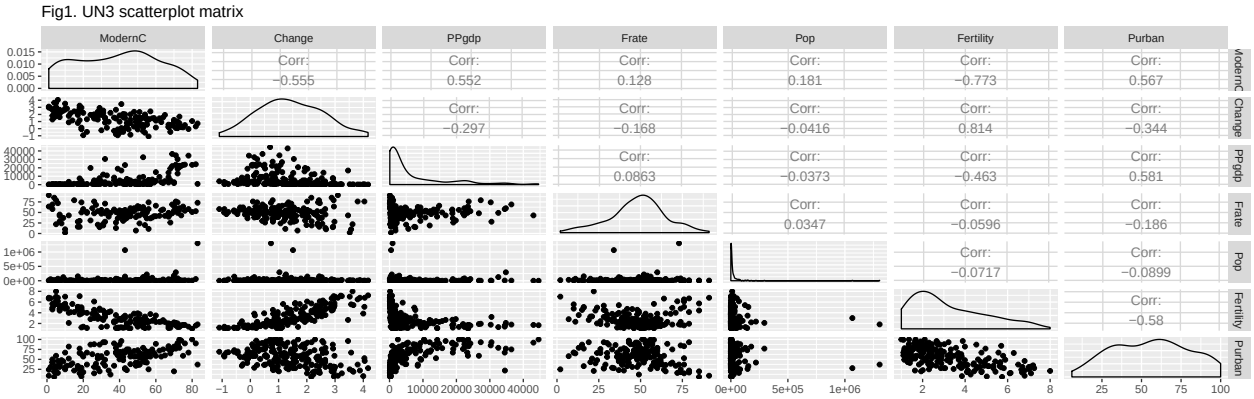
2. What is the mean and standard deviation of each quantitative predictor? Provide in a nicely formatted table.

```
mnstd_chart=data.frame(Predictor=colnames(UN3),Mean=c(''),Standatd_devation=c(''),stringsAsFactors = F)
for (i in 1:length(colnames(UN3))){
quan_mean=mean(UN3[,i],na.rm = T)
quan_std=sd(UN3[,i],na.rm = T)
mnstd_chart$Mean[i] = as.character(quan_mean)
mnstd_chart$Standatd_devation[i]= as.character(quan_std)
}
print(mnstd_chart)
```

```
##   Predictor             Mean Standatd_devation
## 1   ModernC 38.7171052631579  22.6366103759673
## 2    Change 1.41837320574163  1.13313267030361
## 3     PPgdp 6527.38805970149  9325.18855244529
## 4     Frate 48.3053892215569  16.5324480416909
## 5       Pop 30281.8714278846  120676.694478229
## 6 Fertility            3.214  1.70691793716661
## 7    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?

```
library(GGally)
gp=ggpairs(UN3,title=c("Fig1. UN3 scatterplot matrix"))
print(gp,progress=F)
```



As it shows in the scatter plot by using ggpairs, the relationship between PPdgp, Pop and ModernC are obviously unlinear. In Pop vs others plots, the two high dots seem to be outliers because of the huge deviation from the rest samples. In PPdgp vs others plots, most of the samples seem accumulate at the bottom line, indicating the need for transformation

Model Fitting

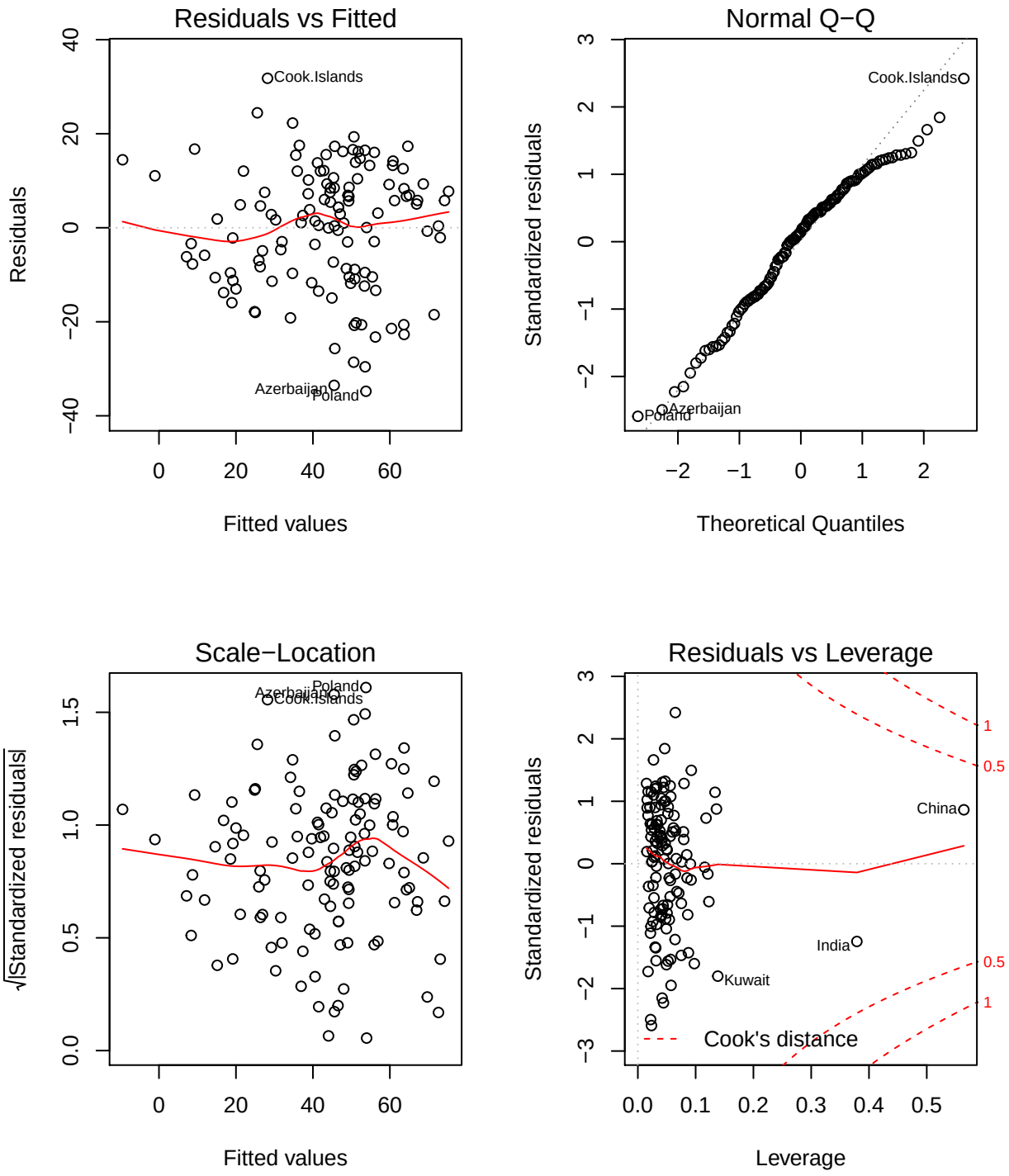
4. Use the `lm()` function to perform a multiple linear regression with `ModernC` as the response and all other variables as the predictors, using the formula `ModernC ~ .`, where the `.` includes all remaining variables in the dataframe. Create diagnostic residual plot from the linear model object and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

```
###Remove Na
UN3_rn=na.omit(UN3)
mc_lm=lm(ModernC~.,data=UN3_rn)
print(summary(mc_lm))
```

```
##
## Call:
## lm(formula = ModernC ~ ., data = UN3_rn)
##
## 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
## Multiple R-squared:  0.6183, Adjusted R-squared:  0.5989
## F-statistic: 31.85 on 6 and 118 DF,  p-value: < 2.2e-16
```

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



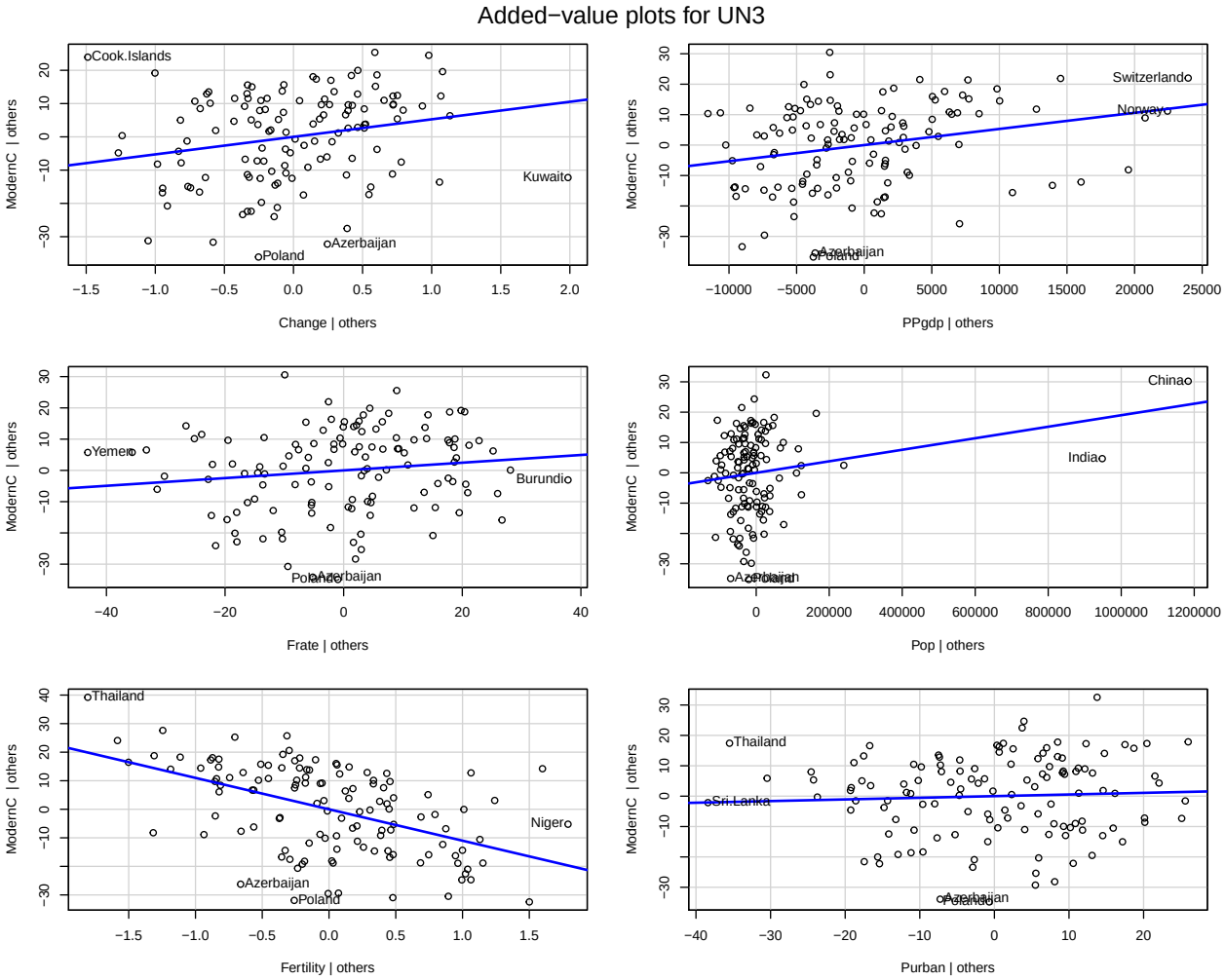

```
paste(as.character(nobs(mc_lm)), c("Observations are used in this model fitting"))
```

```
## [1] "125 Observations are used in this model fitting"
```

Remove all the Na to make sure the stability of modeling. From the Residules vs Fitted plot, we can see the fluctuate in the middle of the trand, indicating the non-constant varience in this model. From the Q-Q plot, the upper right part shows a big deviation from diagonal, which means the data are not fully normal distributed. The fluctuation in the Scale-location plot implies the non-constant varience in the model. The residules vs leverage plot shows seveal observations might have a hugh influence on the fitting model (e.g. China, India). To sum up, the model does not fit the data perfectly

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?

```
mc_lm=lm(ModernC~.,data=UN3_rn)
car::avPlots(mc_lm,main='Added-value plots for UN3')
```



It is easy to notice that Pop need to be transformed. Because in the av plot, all the samples are stacked at X=0, and China and India are too influential in this plot compared to others. Change might need to be transformed although the distribution seems fine, the 'Cook.Islands','Kuwait','Poland' and 'Azerbaijan' might be influential. 'PPgdp' also needs transformed, for the samples are assembled in the left side of the plot.

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.

```
summary(UN3_rn)
```

```
##     ModernC          Change           PPgdp           Frate      
##  Min.   : 1.00   Min.   :-1.100   Min.   :   90   Min.   : 2.00  
##  1st Qu.:28.00   1st Qu.: 0.340   1st Qu.:  687   1st Qu.:39.00  
##  Median :45.00   Median : 1.260   Median : 2077   Median :49.00  
##  Mean   :43.27   Mean   : 1.182   Mean   : 6613   Mean   :48.11  
##  3rd Qu.:58.00   3rd Qu.: 1.940   3rd Qu.: 7724   3rd Qu.:58.00  
##  Max.   :83.00   Max.   : 3.620   Max.   :36445   Max.   :91.00  
##       Pop            Fertility         Purban      
##  Min.   :     19   Min.   :1.000   Min.   :  6.00  
##  1st Qu.:   3443   1st Qu.:1.700   1st Qu.: 40.00  
##  Median :   8877   Median :2.500   Median : 58.00  
##  Mean   :  46060   Mean   :2.876   Mean   : 56.98  
##  3rd Qu.:  31510   3rd Qu.:3.750   3rd Qu.: 75.00  
##  Max.   :1304196   Max.   :8.000   Max.   :100.00
```

```
UN3_rn$Change=UN3_rn$Change+2.5
summary(UN3_rn)
```

```
##     ModernC          Change           PPgdp           Frate      
##  Min.   : 1.00   Min.   :1.400   Min.   :   90   Min.   : 2.00  
##  1st Qu.:28.00   1st Qu.:2.840   1st Qu.:  687   1st Qu.:39.00  
##  Median :45.00   Median :3.760   Median : 2077   Median :49.00  
##  Mean   :43.27   Mean   :3.682   Mean   : 6613   Mean   :48.11  
##  3rd Qu.:58.00   3rd Qu.:4.440   3rd Qu.: 7724   3rd Qu.:58.00  
##  Max.   :83.00   Max.   :6.120   Max.   :36445   Max.   :91.00  
##       Pop            Fertility         Purban      
##  Min.   :     19   Min.   :1.000   Min.   :  6.00  
##  1st Qu.:   3443   1st Qu.:1.700   1st Qu.: 40.00  
##  Median :   8877   Median :2.500   Median : 58.00  
##  Mean   :  46060   Mean   :2.876   Mean   : 56.98  
##  3rd Qu.:  31510   3rd Qu.:3.750   3rd Qu.: 75.00  
##  Max.   :1304196   Max.   :8.000   Max.   :100.00
```

```
car::boxTidwell(ModernC~Pop+PPgdp+Change,other.x=~Frate+Fertility+Purban,data=UN3_rn)
```

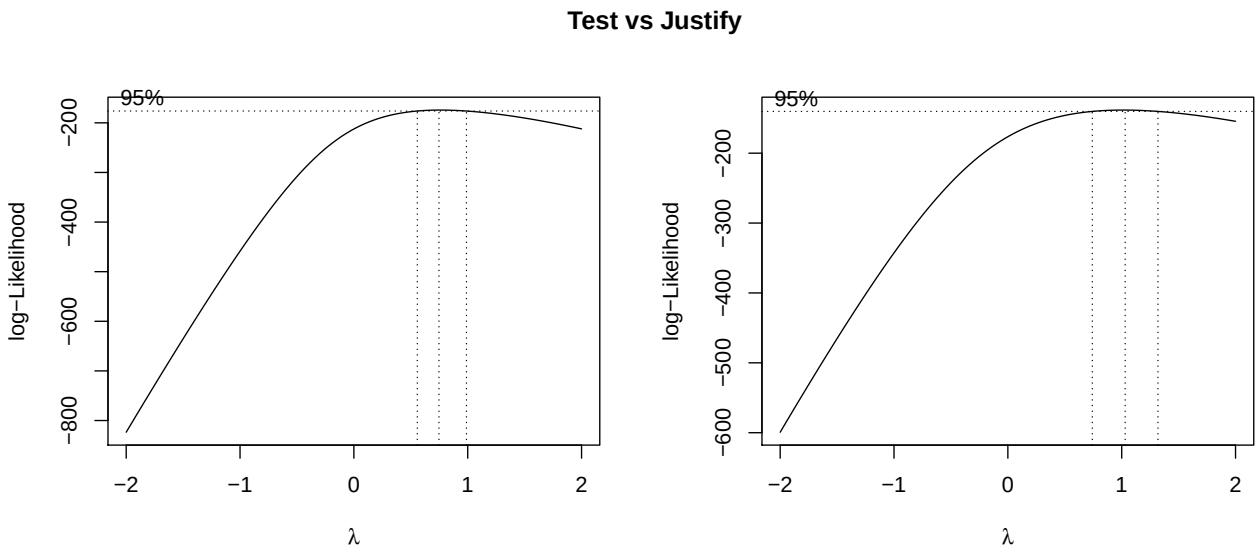
```
## Warning in boxTidwell.default(y, X1, X2, max.iter = max.iter, tol = tol, :
## maximum iterations exceeded
```

```
##        MLE of lambda Score Statistic (z) Pr(>|z|)
## Pop          0.41151             -0.6557 0.512016
## PPgdp       -0.11625             -1.0153 0.309949
## Change      -1.65451             -2.8932 0.003813 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## iterations =  26
```

The method to make the predictor non-negative is to firstly find the minimun value of the observations. Avoid changing the power of the original observation, I choose to add a constant to the columns containing negative value. (Adding 2.5 to 'Change'). The result of boxTidwell describes the calculated lamda for the targeting transformation. Pop's lamda is 0.41, so I will take Popˆ0.4 as the transformation. (Althought the most appropriate transformation will be (Popˆ0.4 -1)/0.4, Popˆ0.4 will not change the power of the equation). PPgdp's lamda is -0.11, which is close to 0, so I choose log PPgdp to transform. Change's lamda is -1.7, so I use Changeˆ(-1.7) to transform the data.

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

```
par(mfrow=c(1,2))
MASS::boxcox(lm(ModernC~Fertility+I(Pop^0.4)+log(PPgdp)+I(Change^(-1.7))+Frate+Purban ,data=UN3_rn),)
MASS::boxcox(lm(ModernC^0.75~Fertility+I(Pop^0.4)+log(PPgdp)+I(Change^(-1.7))+Frate+Purban ,data=UN3_rn)
title("Test vs Justify",outer = T,line=-1.5)
```



The boxcox plot shows that the lamda of response is approximately 0.75(left side). After transform the response using the lamda, the fitted lamda is very close to 1.

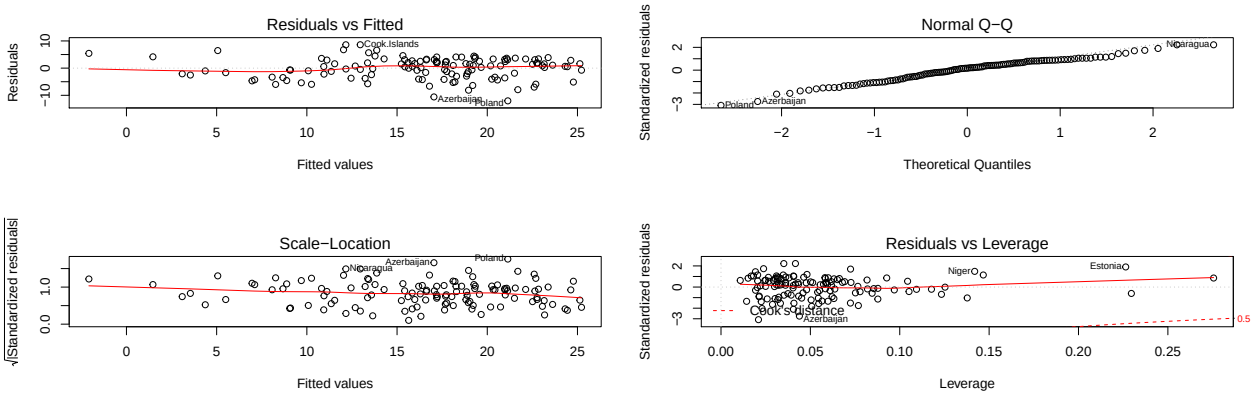
8. Fit the regression using the transformed variables. Provide residual plots and added variables plots and comment. If you feel that you need additional transformations of either the response or predictors, repeat any steps until you feel satisfied.

```
lmc_lm=lm(ModernC^0.75~I(Pop^0.4)+log(PPgdp)+I(Change^(-1.7))+Frate+Fertility+Purban ,data=UN3_rn)
summary(lmc_lm)
```

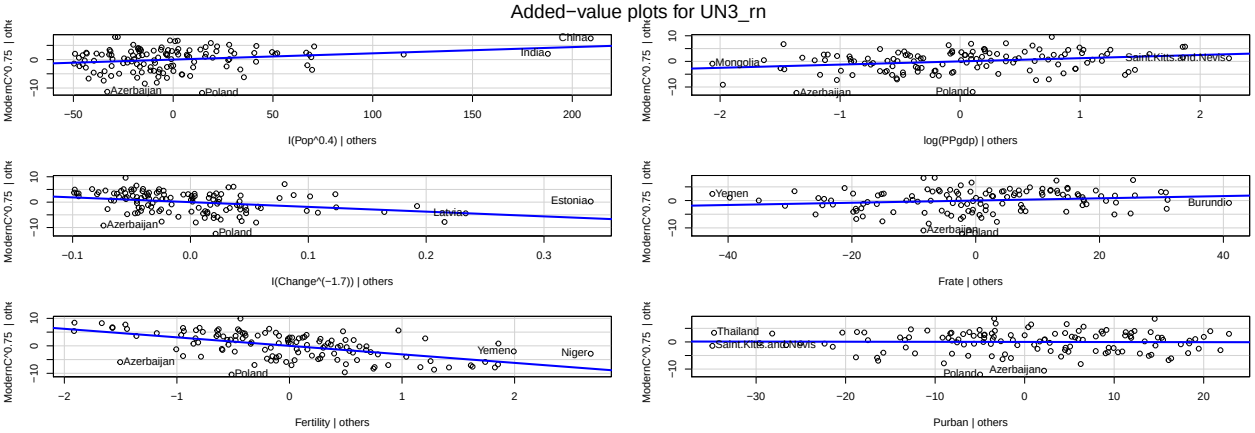
```
##
## Call:
## lm(formula = ModernC^0.75 ~ I(Pop^0.4) + log(PPgdp) + I(Change^(-1.7)) +
##     Frate + Fertility + Purban, data = UN3_rn)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.0393  -2.5433   0.6588   2.9692   8.5984
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       15.359009   4.199378   3.657 0.000382 ***
## I(Pop^0.4)         0.021962   0.008895   2.469 0.014986 *
## log(PPgdp)         1.251381   0.413578   3.026 0.003045 **
## I(Change^(-1.7)) -18.702369   5.169414  -3.618 0.000438 ***
## Frate              0.040700   0.022281   1.827 0.070280 .
## Fertility         -3.100139   0.418440  -7.409 2.09e-11 ***
## Purban            -0.004046   0.027960  -0.145 0.885204
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 3.944 on 118 degrees of freedom
## Multiple R-squared:   0.67,  Adjusted R-squared:  0.6532
## F-statistic: 39.92 on 6 and 118 DF,  p-value: < 2.2e-16
```

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



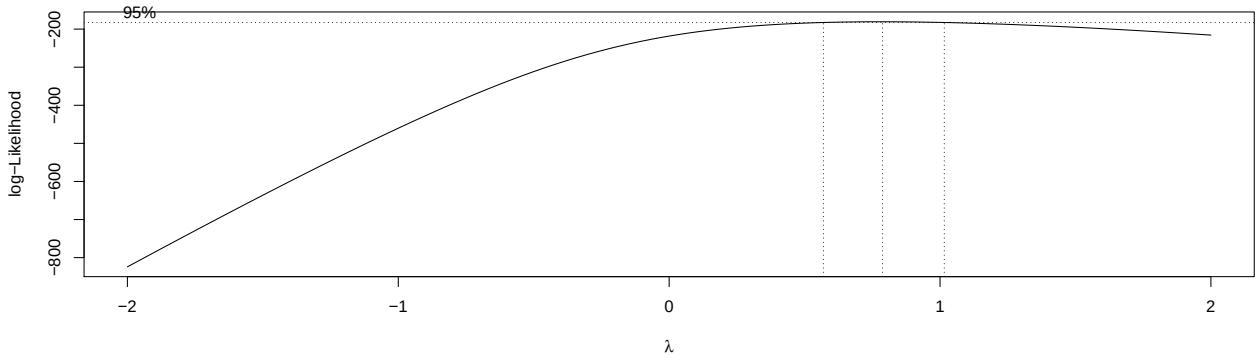

```
car::avPlots(lmc_lm,main='Added-value plots for UN3_rn')
```



From the summary of the linear model, the significance of the predictor was enhanced compared with the one before transformation. Although the residules vs fitted and scales-location plot still indicate the non-constant varience of the predictor samples, the trend is milder than the untransformed one. The Q-Q plot fits the diagnol better than the previous one. The leveage for most of samples get slightly bigger as a more average distribution, but there are still a few samples having big leverge that influence the model largrly.

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?

```
MASS::boxcox(mc_lm)
```



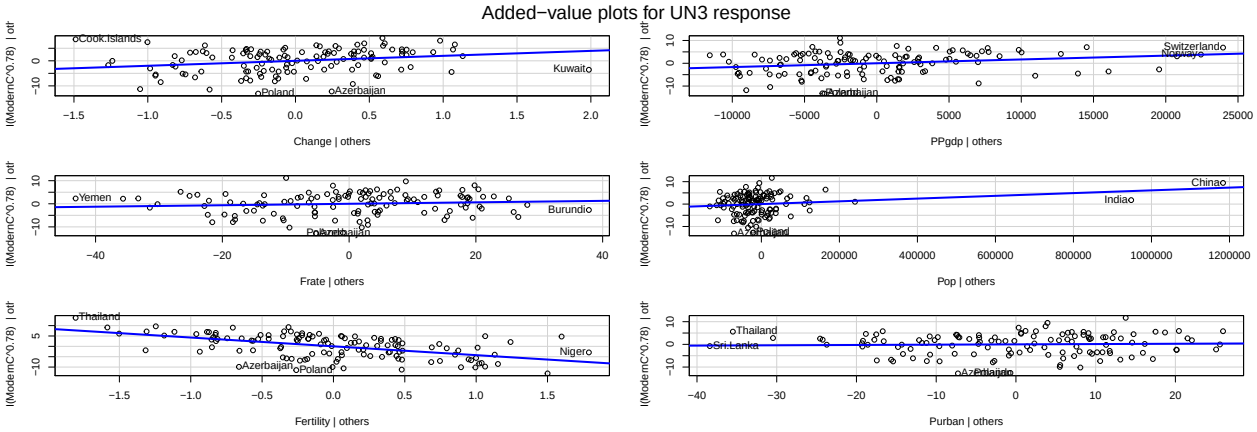

```
powerTransform(mc_lm)
```

```
## Estimated transformation parameter
##        Y1
## 0.7789722
```

```
rt_lm=lm(I(ModernC^0.78)~.,data = UN3_rn)
summary(rt_lm)
```

```
##
## Call:
## lm(formula = I(ModernC^0.78) ~ ., data = UN3_rn)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.7057  -3.6080   0.7829   3.3217  11.5201
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.963e+01  3.432e+00   5.718 8.28e-08 ***
## Change       1.981e+00  7.382e-01   2.684  0.00833 **
## PPgdp        1.674e-04  6.258e-05   2.676  0.00852 **
## Frate        3.161e-02  2.850e-02   1.109  0.26969
## Pop          6.124e-06  2.904e-06   2.108  0.03712 *
## Fertility   -4.249e+00  6.197e-01  -6.857 3.43e-10 ***
## Purban       1.358e-02  3.283e-02   0.414  0.67983
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.802 on 118 degrees of freedom
## Multiple R-squared:  0.632,  Adjusted R-squared:  0.6133
## F-statistic: 33.77 on 6 and 118 DF,  p-value: < 2.2e-16
```

```
car::avPlots(rt_lm,main='Added-value plots for UN3 response')
```




```
summary(UN3_rn)
```

```
##     ModernC          Change          PPgdp           Frate      
##  Min.   : 1.00   Min.   :1.400   Min.   :   90   Min.   : 2.00  
##  1st Qu.:28.00   1st Qu.:2.840   1st Qu.:  687   1st Qu.:39.00  
##  Median :45.00   Median :3.760   Median : 2077   Median :49.00  
##  Mean   :43.27   Mean   :3.682   Mean   : 6613   Mean   :48.11  
##  3rd Qu.:58.00   3rd Qu.:4.440   3rd Qu.: 7724   3rd Qu.:58.00  
##  Max.   :83.00   Max.   :6.120   Max.   :36445   Max.   :91.00  
##       Pop            Fertility         Purban      
##  Min.   :     19   Min.   :1.000   Min.   :  6.00  
##  1st Qu.:   3443   1st Qu.:1.700   1st Qu.: 40.00  
##  Median :   8877   Median :2.500   Median : 58.00  
##  Mean   :  46060   Mean   :2.876   Mean   : 56.98  
##  3rd Qu.:  31510   3rd Qu.:3.750   3rd Qu.: 75.00  
##  Max.   :1304196   Max.   :8.000   Max.   :100.00
```

```
car::boxTidwell(ModernC^0.78~Pop+Change+PPgdp,other.x=~Fertility+Purban+Frate,data=UN3_rn)
```

```
## Warning in boxTidwell.default(y, X1, X2, max.iter = max.iter, tol = tol, :
## maximum iterations exceeded
```

```
##        MLE of lambda Score Statistic (z) Pr(>|z|)
## Pop          0.41136             -0.6299 0.528741
## Change      -1.42342             -3.0100 0.002612 **
## PPgdp       -0.22450             -1.0206 0.307452
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## iterations =  26
```

```
rt_mc_lm=lm(ModernC^0.78~I(Pop^0.41)+I(Change^(-1.5))+log(PPgdp)+Frate+Purban+Fertility,data = UN3_rn)
summary(rt_mc_lm)
```

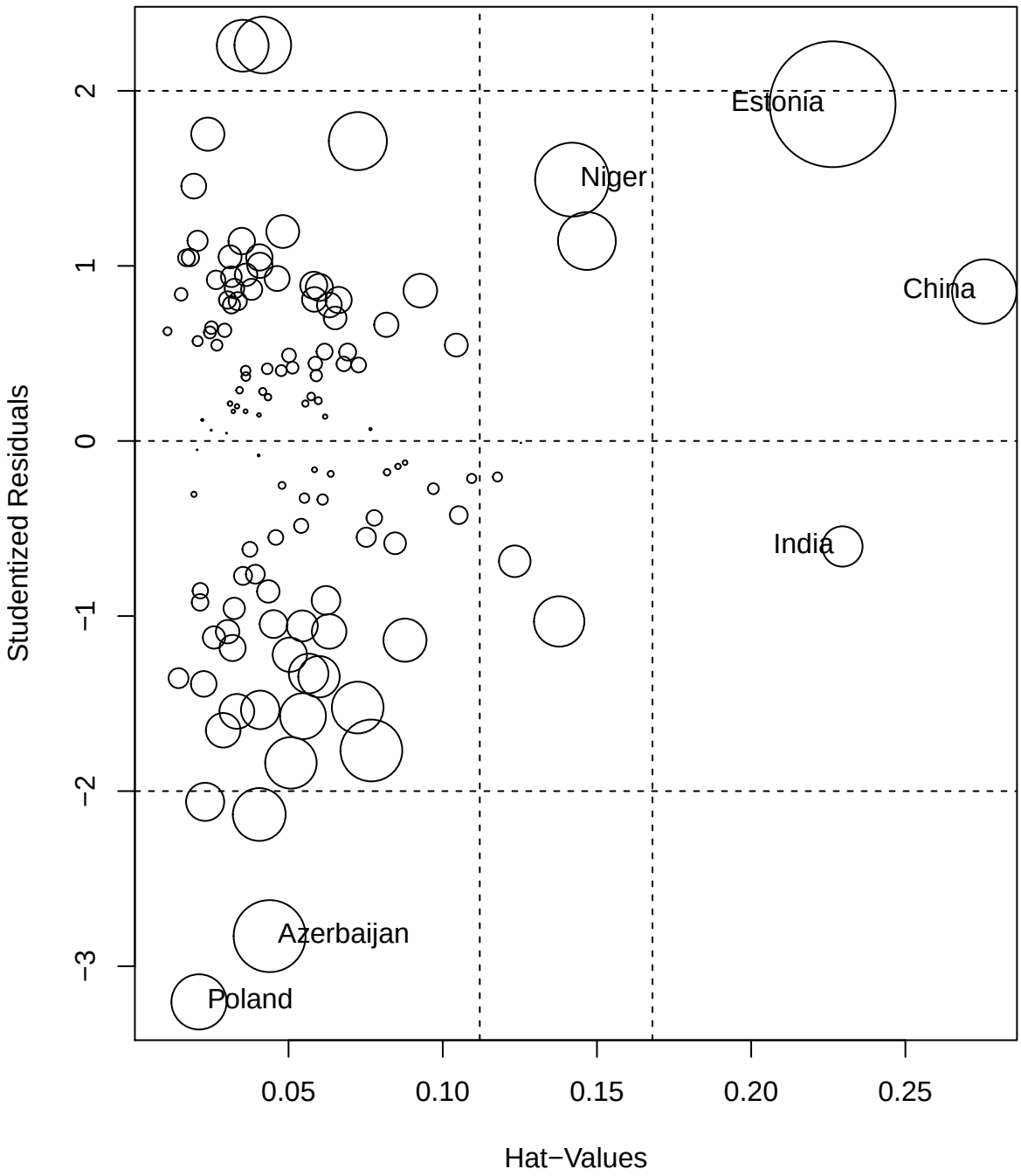
```
##
## Call:
## lm(formula = ModernC^0.78 ~ I(Pop^0.41) + I(Change^(-1.5)) +
##     log(PPgdp) + Frate + Purban + Fertility, data = UN3_rn)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.8415  -2.9057   0.7771   3.4085  10.0410
```

```
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       17.420803   4.930147   3.534 0.000586 ***
## I(Pop^0.41)        0.022550   0.008979   2.511 0.013382 *
## I(Change^(-1.5)) -20.692365   5.727246  -3.613 0.000446 ***
## log(PPgdp)         1.475008   0.476795   3.094 0.002469 **
## Frate              0.050054   0.025771   1.942 0.054486 .
## Purban            -0.004833   0.032337  -0.149 0.881448
## Fertility         -3.579767   0.490773  -7.294 3.76e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.559 on 118 degrees of freedom
## Multiple R-squared:  0.6684, Adjusted R-squared:  0.6515
## F-statistic: 39.64 on 6 and 118 DF,  p-value: < 2.2e-16
```

The transformation of response is very similar (lamda=0.75 vs 0.78). Also, the lamdas for predictor are vert similar than starting from predictor(Pop: 0.41 vs 0.41, Change -1.41 vs. -1.65, PPdgp -0.22 vs -0.12)

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.

```
influencePlot(lmc_lm)
```

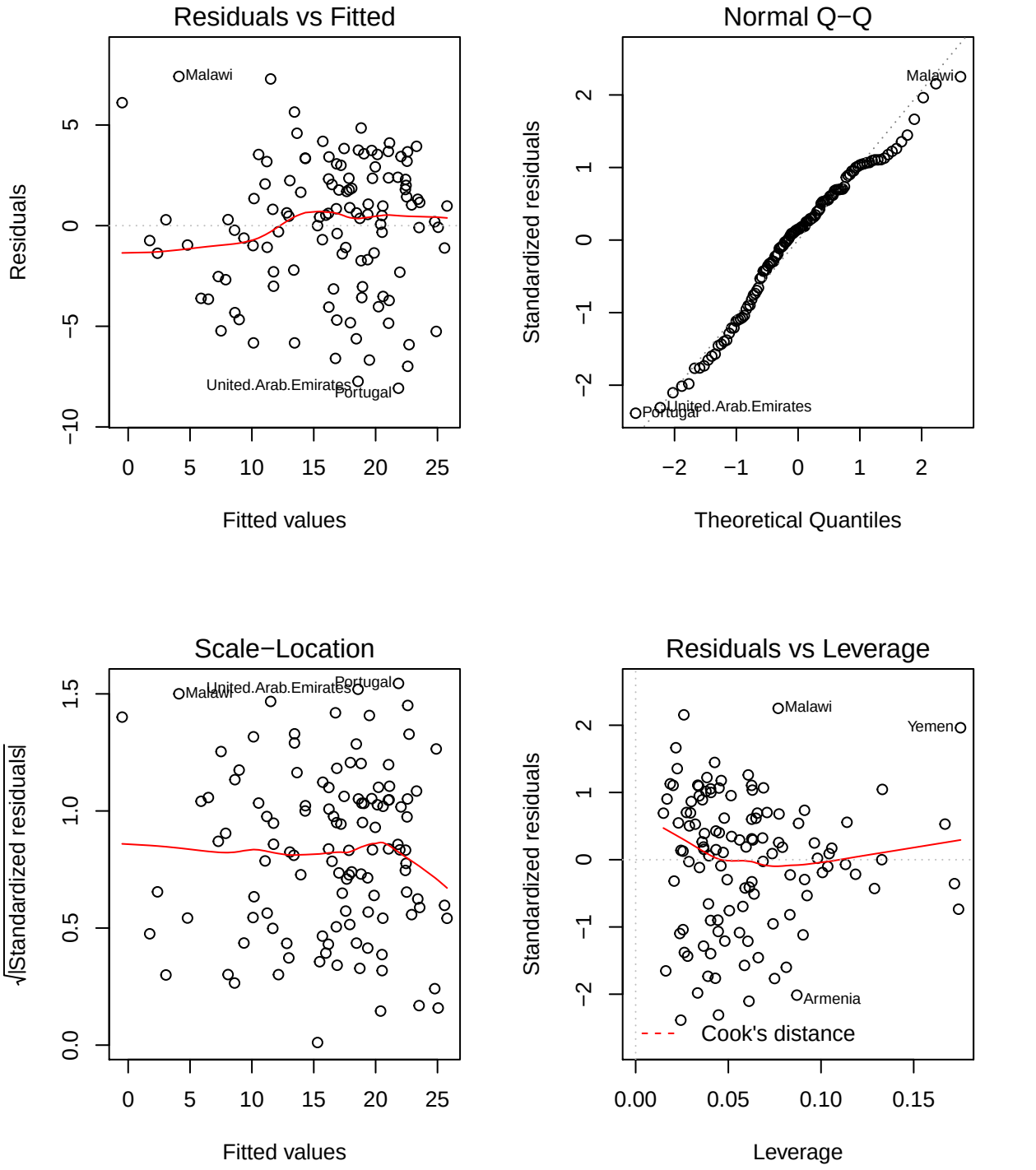



```
##                StudRes        Hat      CookD
## Azerbaijan -2.8281447 0.04386235 0.04948280
## China       0.8532853 0.27559021 0.03966175
## Estonia     1.9233587 0.22640624 0.15120815
## India      -0.6025659 0.22957710 0.01554036
## Niger       1.4927363 0.14191859 0.05210528
## Poland     -3.2039484 0.02096861 0.02912179
```

```
UN3_rn_ro2=UN3_rn[which(rownames(UN3_rn) %!in% c("Azerbaijan","China","India","Poland","Niger","Estonia
lmc_lm_ro2=lm(ModernC^0.75~Fertility+I(Pop^0.4)+log(PPgdp)+I(Change^(-1.7))+Frate+Purban ,data=UN3_rn_r
summary(lmc_lm_ro2)
```

```
##
## Call:
## lm(formula = ModernC^0.75 ~ Fertility + I(Pop^0.4) + log(PPgdp) +
##     I(Change^(-1.7)) + Frate + Purban, data = UN3_rn_ro2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.0793 -2.2927  0.5012  2.3507  7.4130
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       20.45887    3.88720   5.263 7.07e-07 ***
## Fertility         -3.87104    0.40079  -9.659 2.37e-16 ***
## I(Pop^0.4)         0.02515    0.01043   2.412  0.01751 *
## log(PPgdp)         1.06259    0.37115   2.863  0.00503 **
## I(Change^(-1.7)) -27.34034    5.20773  -5.250 7.49e-07 ***
## Frate              0.04402    0.01984   2.219  0.02855 *
## Purban            -0.01482    0.02474  -0.599  0.55031
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.428 on 110 degrees of freedom
## Multiple R-squared:  0.7442, Adjusted R-squared:  0.7303
## F-statistic: 53.35 on 6 and 110 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(lmc_lm_ro2)
```




```
title=(c("residual plots after removing outliers"))
```

From the residule plots of the designed model, “Cook.Islands”,“Nicaragua”,“Azerbaijan”,“Poland” are the outiers. Shown by the influence plot,“China”,“India”,“Niger”,“Estonia” have a huge influence on the result of the model. Therefore, removal of those samples is not able to change the non-constant varience shown by the Residule vs. Fitted and Scale-location plot. The Q-Q plot has a little improvement on the right top and left bottom side, which can be explained easily by the removing of the outliers. The leverage plot has a more even distribution after removing the influential samples.

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!

```
lmc_lm_ro2_rpb=lm(ModernC^0.75~Fertility+I(Pop^0.4)+log(PPgdp)+I(Change^(-1.7))+Frate,data=UN3_rn_ro2)
anova(lmc_lm_ro2_rpb,lmc_lm_ro2)
```

```
## Analysis of Variance Table
##
## Model 1: ModernC^0.75 ~ Fertility + I(Pop^0.4) + log(PPgdp) + I(Change^(-1.7)) +
##     Frate
## Model 2: ModernC^0.75 ~ Fertility + I(Pop^0.4) + log(PPgdp) + I(Change^(-1.7)) +
##     Frate + Purban
##   Res.Df    RSS Df Sum of Sq     F Pr(>F)
## 1    111 1296.7
## 2    110 1292.5  1    4.2177 0.359 0.5503
```

```
coef_md=confint(lmc_lm_ro2_rpb,level=0.95)
Pop=(coef_md[c("I(Pop^0.4)"),])^(-0.4)
PPgdp=exp(coef_md[c("log(PPgdp)"),])
Change=as.numeric(as.complex(coef_md[c("I(Change^(-1.7))"),])^1.7)-2.5
```

```
## Warning: imaginary parts discarded in coercion
```

```
ori_uni=as.data.frame(t(data.frame(Pop,PPgdp,Change)))
coef_md_ori=rbind(coef_md,ori_uni)
coef_md_ori_uni=coef_md_ori[c("(Intercept)","Fertility","Frate","Pop","PPgdp","Change"),]
print(coef_md_ori_uni)
```

```
##                      2.5 %      97.5 %
## (Intercept)  12.722372390 28.07846454
## Fertility    -4.590393575 -3.04568284
## Frate         0.008400196  0.08486685
## Pop           8.928905388  3.44746273
## PPgdp         1.415903117  4.49813696
## Change      276.502765908 69.80375057
```

```
summary(lmc_lm_ro2_rpb)
```

```
##
## Call:
## lm(formula = ModernC^0.75 ~ Fertility + I(Pop^0.4) + log(PPgdp) +
##     I(Change^(-1.7)) + Frate, data = UN3_rn_ro2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.9476 -2.4622  0.5413  2.3366  7.4014
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       20.40042    3.87474   5.265 6.93e-07 ***
## Fertility         -3.81804    0.38977  -9.796  < 2e-16 ***
## I(Pop^0.4)         0.02476    0.01038   2.386  0.01872 *
## log(PPgdp)         0.92572    0.29166   3.174  0.00195 **
## I(Change^(-1.7)) -27.24269    5.19012  -5.249 7.43e-07 ***
## Frate              0.04663    0.01929   2.417  0.01728 *
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.418 on 111 degrees of freedom
## Multiple R-squared:  0.7434, Adjusted R-squared:  0.7318
## F-statistic: 64.32 on 5 and 111 DF,  p-value: < 2.2e-16
```

I made another model without Purban because in every step, Purban does not have a significant coefficient in the summary of every linear model. Anova was used to test whether Purban has an effect on the final result. For the Pr(>F) equals to 0.55, we fail to reject the H0, therefore we can assume that Purban does not affect the result. Therefore I remove Purban as the predictor.

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

    The designed model is shown as below$ModernC^{0.78} = 20.40+0.05Frate-3.82Fertility+0.02Pop^{0.4}+0.92log(PPgdp)-2$

    85 cases were deleted because of the missing value, while other 8 cases were deleted because they are outliers or influential points. 1 predictor was removed from the final designed model for the effect it makes cannot provide significant predictor compared to the rest of the predictors. The intercept is 20.4, while 1 unit of Frate provides 0.05 unit increase in the ModernCˆ0.78, 1 unit of Fertility provides 3.82 decrease. 1 unit of Popˆ0.4 provides 0.02 increase in ModernCˆ0.78, 1 unit of Changeˆ(-1.7) provides 27.24 decrease. 1 unit of log(Pop) provides 0.92 unit increase of ModernC

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.*

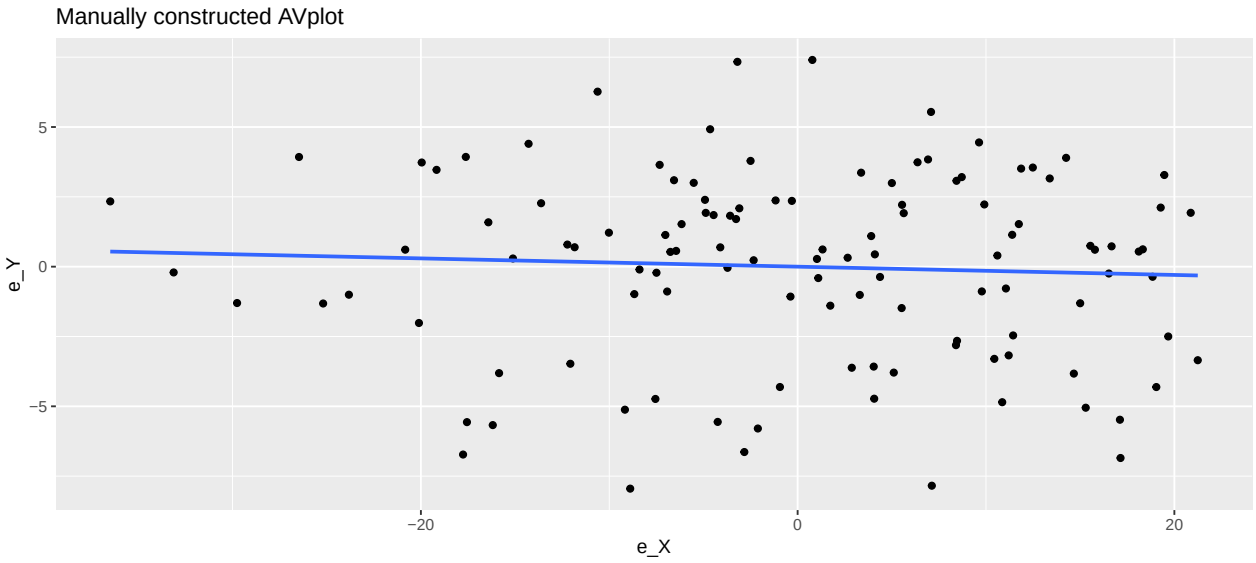
In the added variable scatter plot, for the second variable, we will perform $\hat{e}_2 \sim \hat{e}_1$In which we have $\hat{e}_2 = (1-H)Y$, where $H = X$

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.

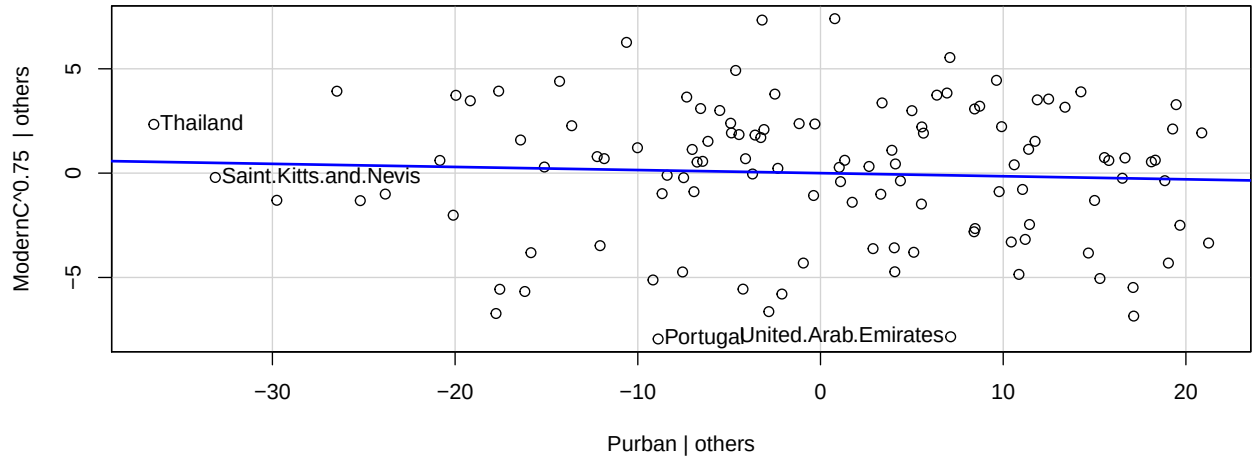
```
e_Y=residuals(lmc_lm_ro2_rpb)
e_X=residuals(lm(Purban~Fertility+I(Popˆ0.4)+log(PPgdp)+I(Changeˆ(-1.7))+Frate,data=UN3_rn_ro2))
e_Y_X=lm(e_Y~e_X)
df_ex14 = data.frame(Original_coef = lmc_lm_ro2$coefficients["Purban"],
                avPlot_coef = e_Y_X$coefficients["e_X"], row.names = "Coeffs")
print(df_ex14)
```

```
##        Original_coef avPlot_coef
## Coeffs   -0.01482352 -0.01482352
```

```
df = data.frame(e_Y = e_Y, e_X1 = e_X)
ggplot(data=df, aes(x = e_X, y = e_Y)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) + ggtitle("Manually constructed AVplot")
```




```
car::avPlots(lmc_lm_ro2,~Purban, main="Automatically constructed AVplot")
```



As the table shows, the two factors have the same coefficient. As the plots shows, the manually constructed plots have the same slope with the automatically generated one.