HW2 STA521 Fall18

Min Chul Kim (NetID mk408, Github ID minchel93)

Due September 23, 2018 5pm

Exploratory Data Analysis

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

```
#Summary of Data
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
```

```
#Additional Missing Data
help(UN3)

#Data Type of Predictors
sapply(UN3, class)
```

```
##    ModernC     Change      PPgdp      Frate        Pop Fertility     Purban 
## "integer" "numeric" "integer" "integer" "numeric" "numeric" "integer"
```

Comment:
According to the summary function, there are 58 data missing from ModernC, 1 missing from Change, 9 missing from PPgdp, 43 from Frate, 2 missing from Pop, 10 missing from Fertility, and 0 missing from Purban.

All the predictors are quantitative, according to the R description of each predictor.

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

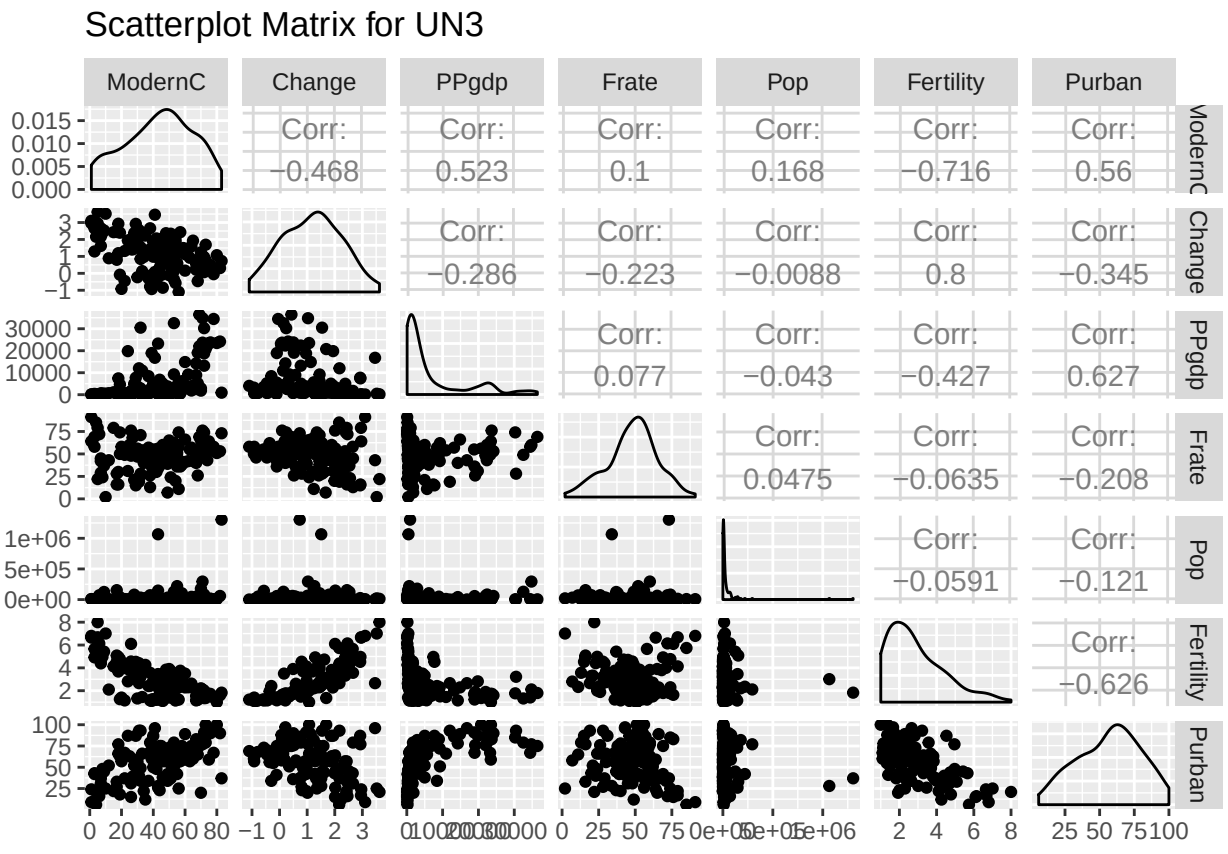
```
#Calculating the Mean and SD, w/ NA's Removed from Predictors
kable(rbind(Mean = (sapply(UN3, mean, na.rm = TRUE)), SD = (sapply(UN3, sd, na.rm = TRUE))), format = "
```

	ModernC	Change	PPgdp	Frate	Pop	Fertility	Purban
Mean	38.71711	1.418373	6527.388	48.30539	30281.87	3.214000	56.20000

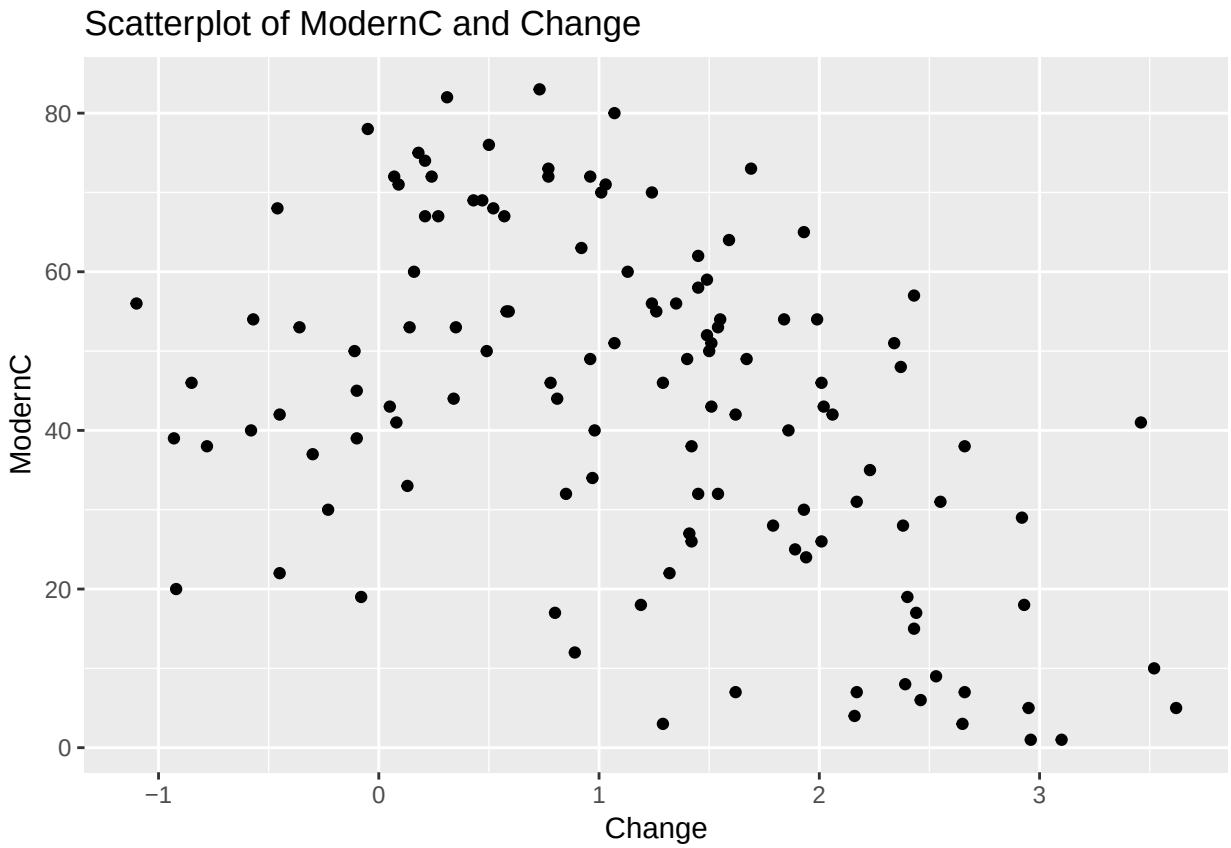
	ModernC	Change	PPgdp	Frate	Pop	Fertility	Purban
SD	22.63661	1.133133	9325.189	16.53245	120676.69	1.706918	24.10976

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?

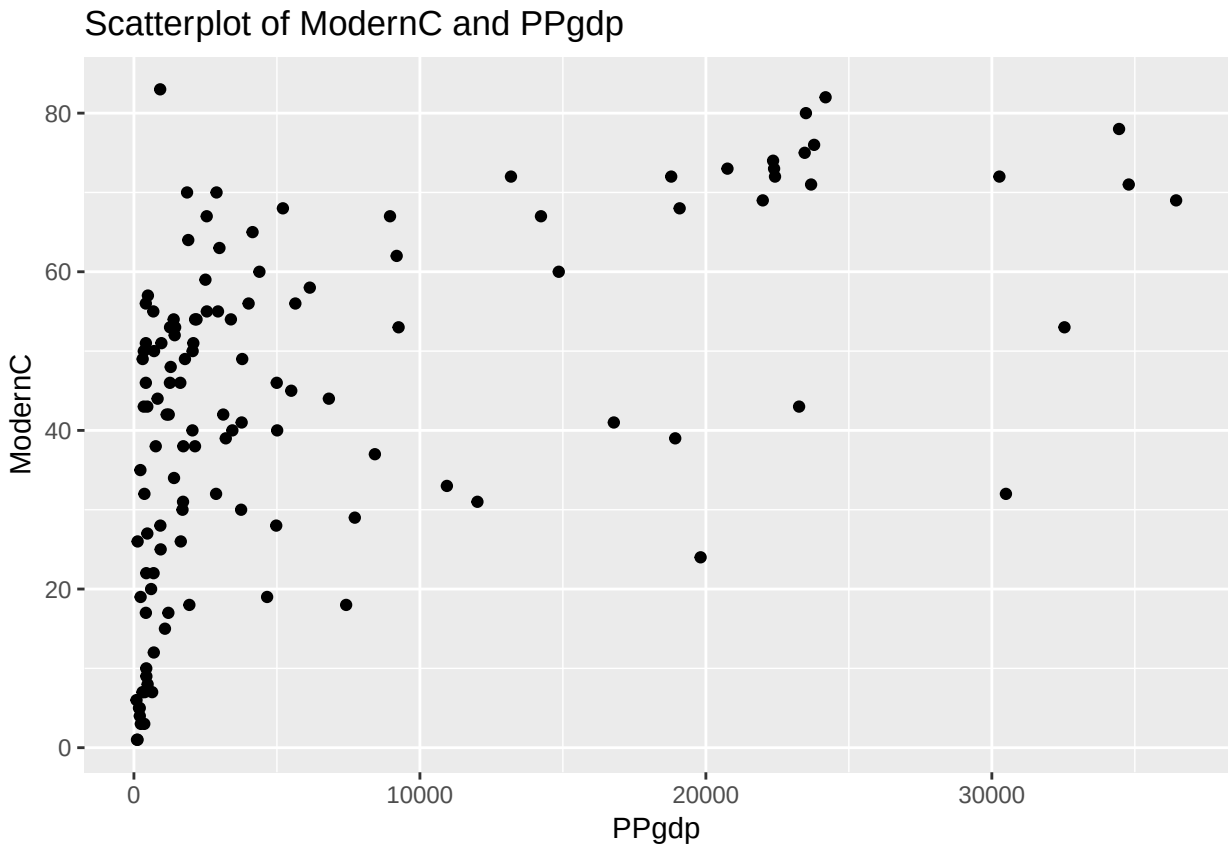
```
#Removing NA's from Data
UN3_No_NA = na.omit(UN3)

#Defining Predictors
ModernC = UN3_No_NA$ModernC
Change = UN3_No_NA$Change
PPgdp = UN3_No_NA$PPgdp
Frate = UN3_No_NA$Frate
Pop = UN3_No_NA$Pop
Fertility = UN3_No_NA$Fertility
Purban = UN3_No_NA$Purban

#Plots
ggpairs(UN3_No_NA, progress = FALSE, title = "Scatterplot Matrix for UN3")
```



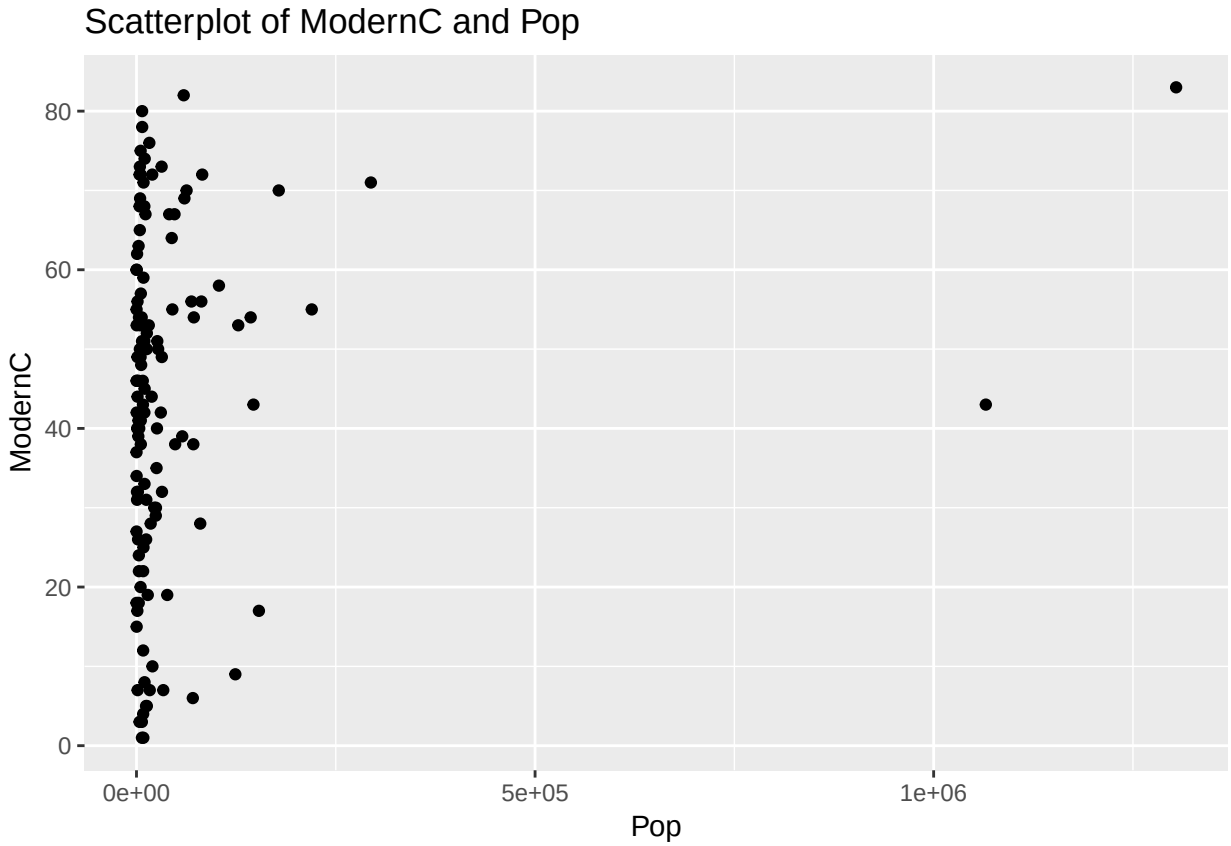

```
ggplot(UN3_No_NA, aes(Change, ModernC)) + geom_point() + ggtitle("Scatterplot of ModernC and Change")
```




```
ggplot(UN3_No_NA, aes(PPgdp, ModernC)) + geom_point() + ggtitle("Scatterplot of ModernC and PPgdp")
```




```
ggplot(UN3_No_NA, aes(Pop, ModernC)) + geom_point() + ggtitle("Scatterplot of ModernC and Pop")
```



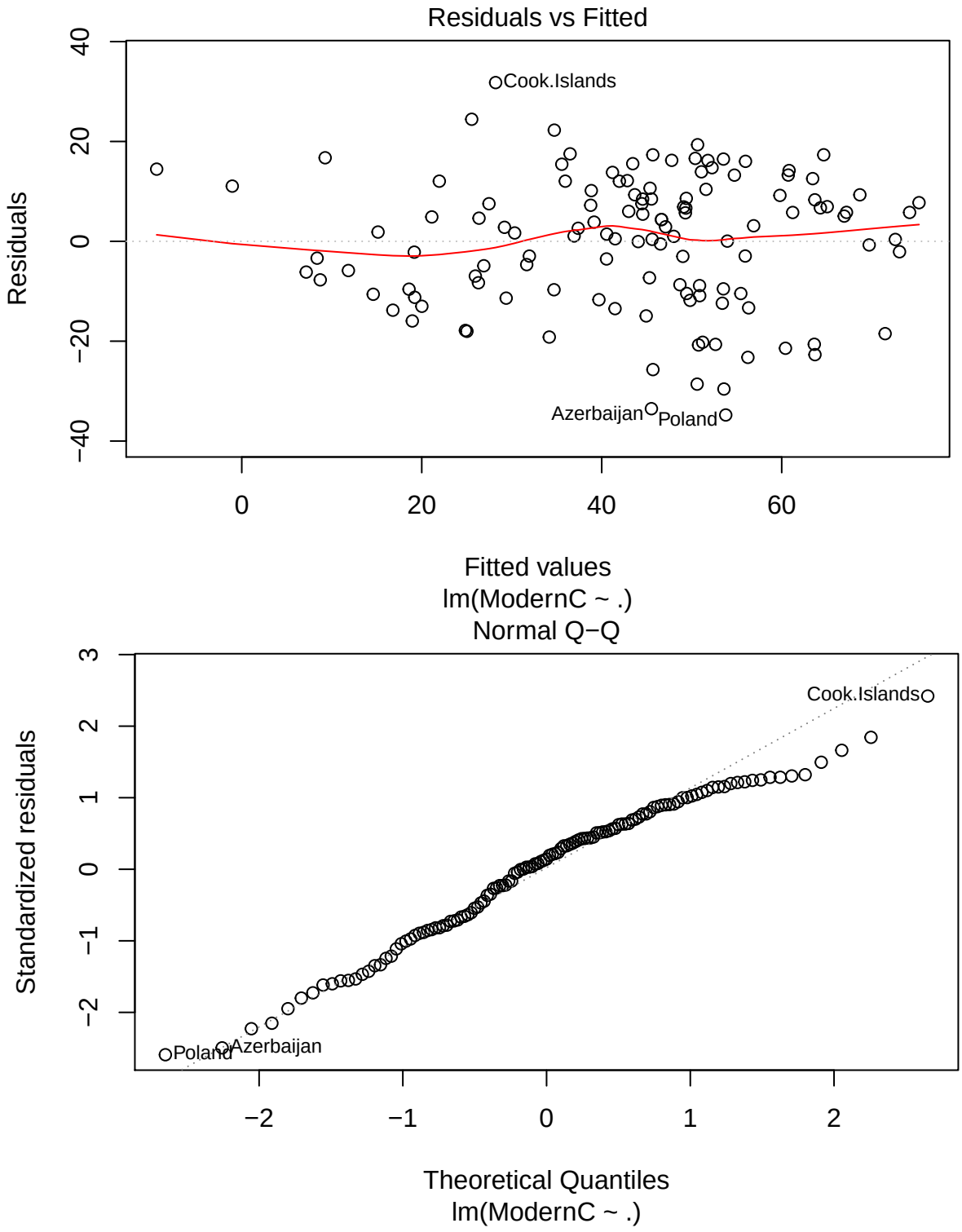
Comment:
By visual inspection, I was not able to identify potential outliers. However, I was able to observe that ModernC and PPgdp, PPgdp and Fertility, and PPgdp and Purban all have noticeable non-linear relationships, which may demand some transformations.

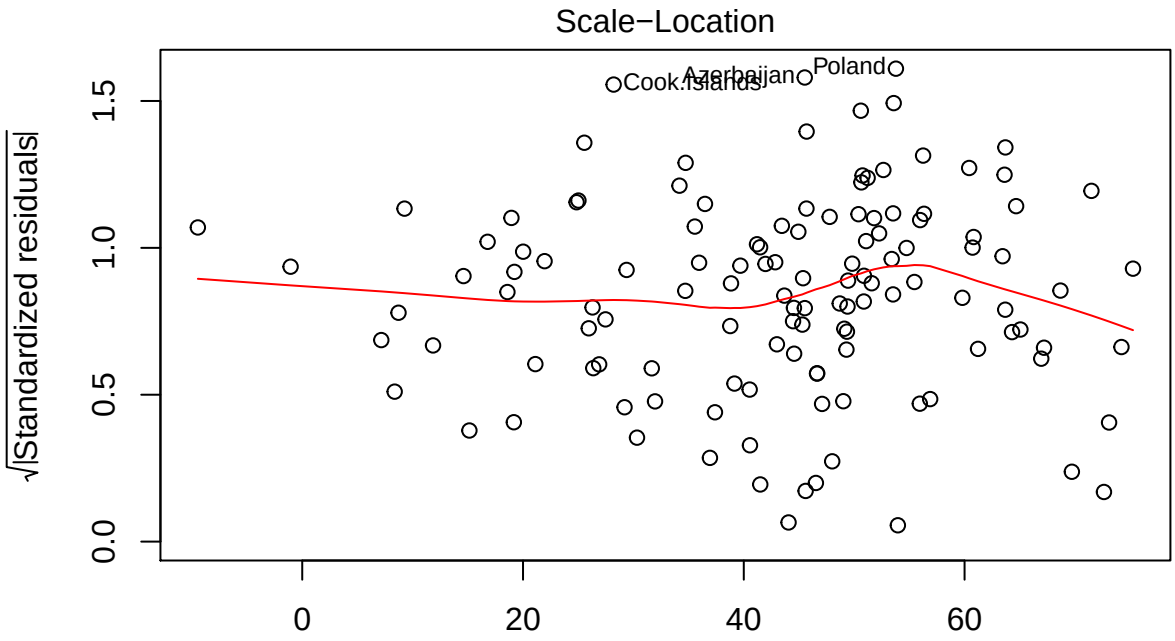
As for the findings regarding predicting ModernC using other variables, I concluded that the predictor Fertility would be the best pre-transformed variable to predict the response, since Fertility has the most linear relationship, among the predictors, with "ModernC." I also concluded that PPgdp would be the worst pre-transformed variable to predict the response, since PPgdp has the most noticeable non-linear relationship, among the predictors, with ModernC.

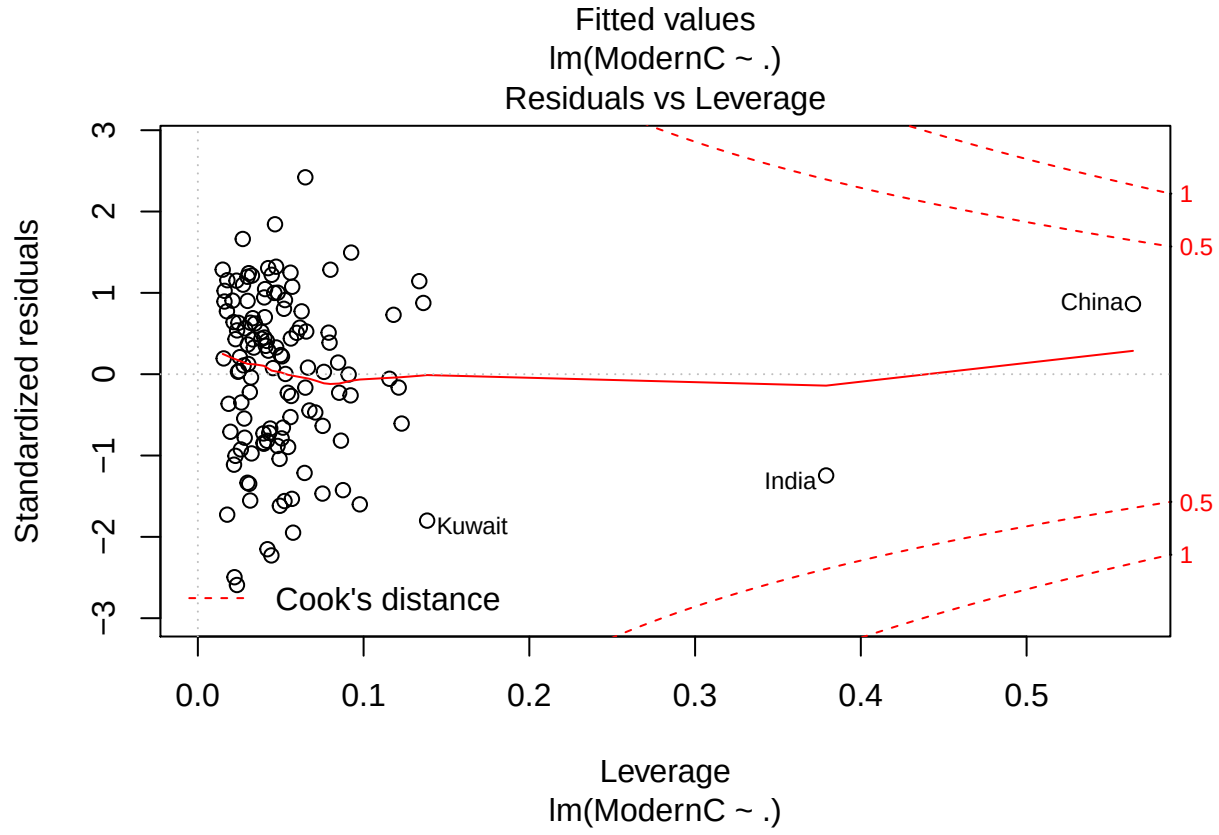
Model Fitting

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?

```
#Multiplie Regression
LM_ModernC = lm(ModernC ~ ., data = UN3_No_NA)

#Multiple Regression Plot
plot(LM_ModernC)
```


Residuals vs Fitted
Residuals
Fitted values
lm(ModernC ~ .)
Cook.Islands
Azerbaijan
Poland
Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(ModernC ~ .)
Cook.Islands
Poland
Azerbaijan






```
#Observations Used in Multiple Regression
kable(nobs(LM_ModernC), col.names = "Number of Observations", format = "markdown")
```

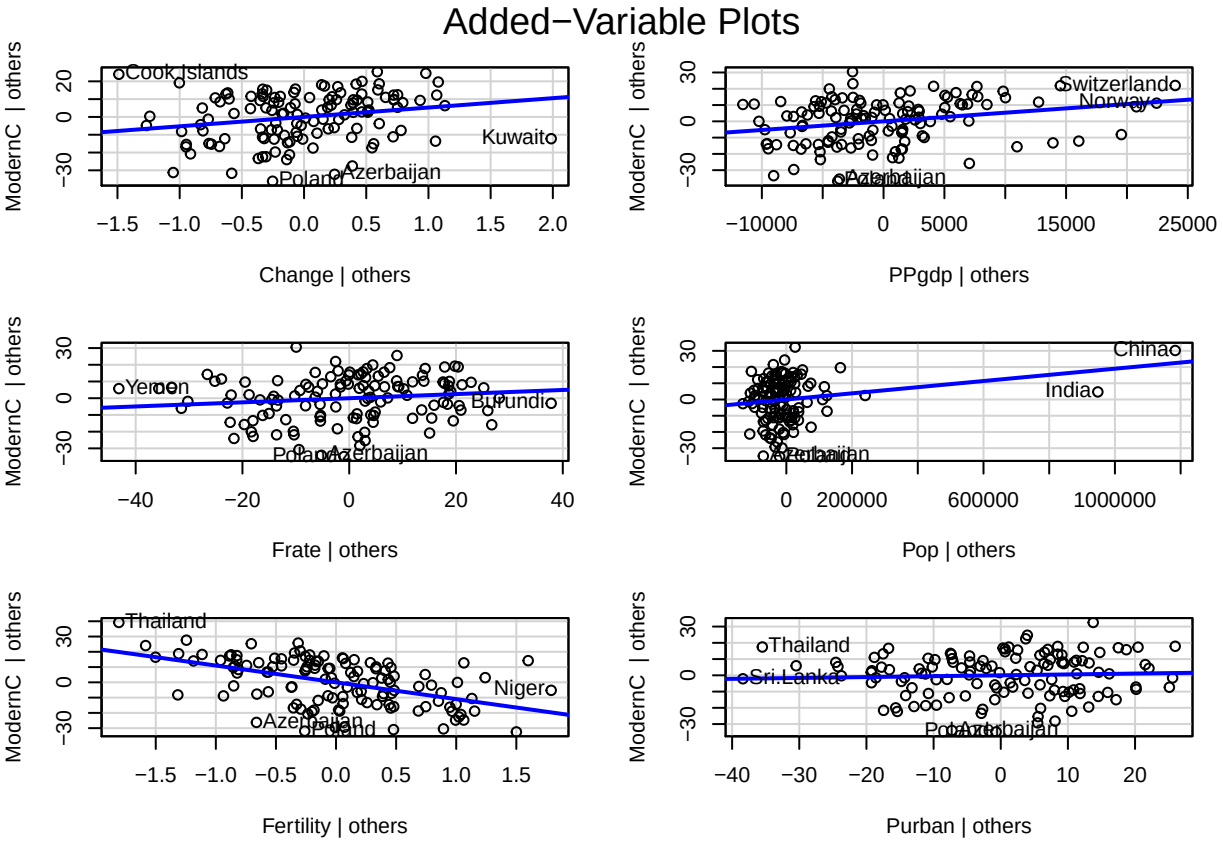
Number of Observations
125

Comment:
According to the residual plot, the constructed linear regression model does not seem to suffer from major heteroskedasticity, which implies that the model is approximately following, if not perfectly following, the constant variance assumption of the OLS model. The model, however, is judged to be suffering from heavy-tails (not normal), according to the Normal Q-Q plot. In terms of normality, this is not an ideal output, but it is deemed to be fixable.

The observations used in the model fitting are 125.

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?

```
#Added Variable Plot
avPlots(LM_ModernC)
```



Comment:
After visual inspection, I would like to suggest transforming the predictors Pop and PPgdp. The predictor "Pop" is too clustered around one section of the line, which I believe would become more spread out after a transformation. The predictor "PPgdp," on the other hand, has a large x-scale, which hints that transformation would significantly change the plot and perhaps provide a better result.

As for the question about any of localities being influential for any of the terms, I would say China and India are potential influential localities in ModernC|Others ~ Pop|Others graph. The presence of the two points is judged - subjectively - to be creating a positive relationship between the response and the predictor.

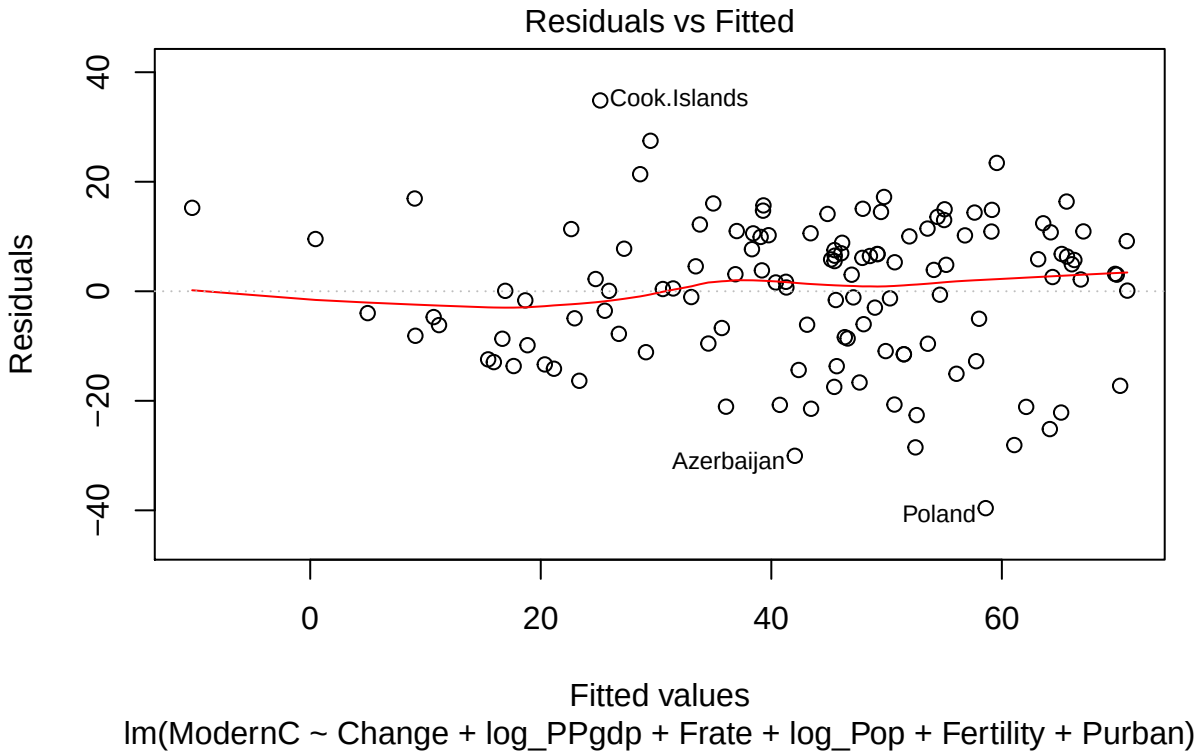
6. Using the Box-Tidwell `car::boxTidwell` or graphical methods find appropriate transformations of

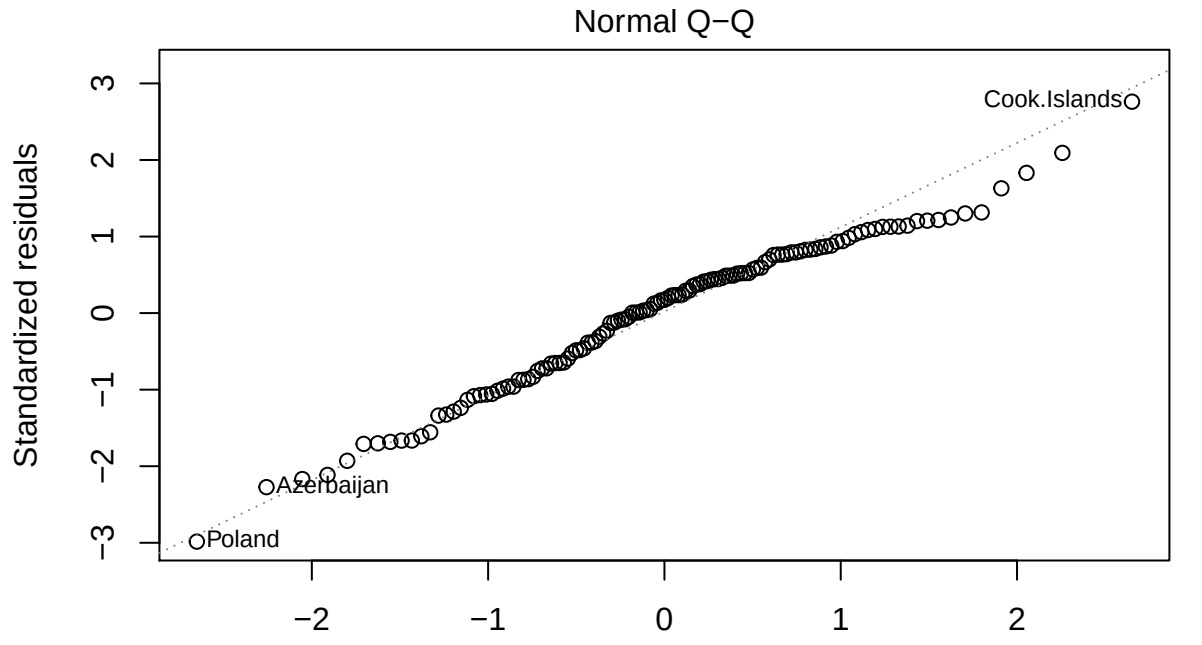
the predictor variables to be used as predictors in the linear model. If any predictors are negative, you may need to transform so that they are non-negative. Describe your method and the resulting transformations.

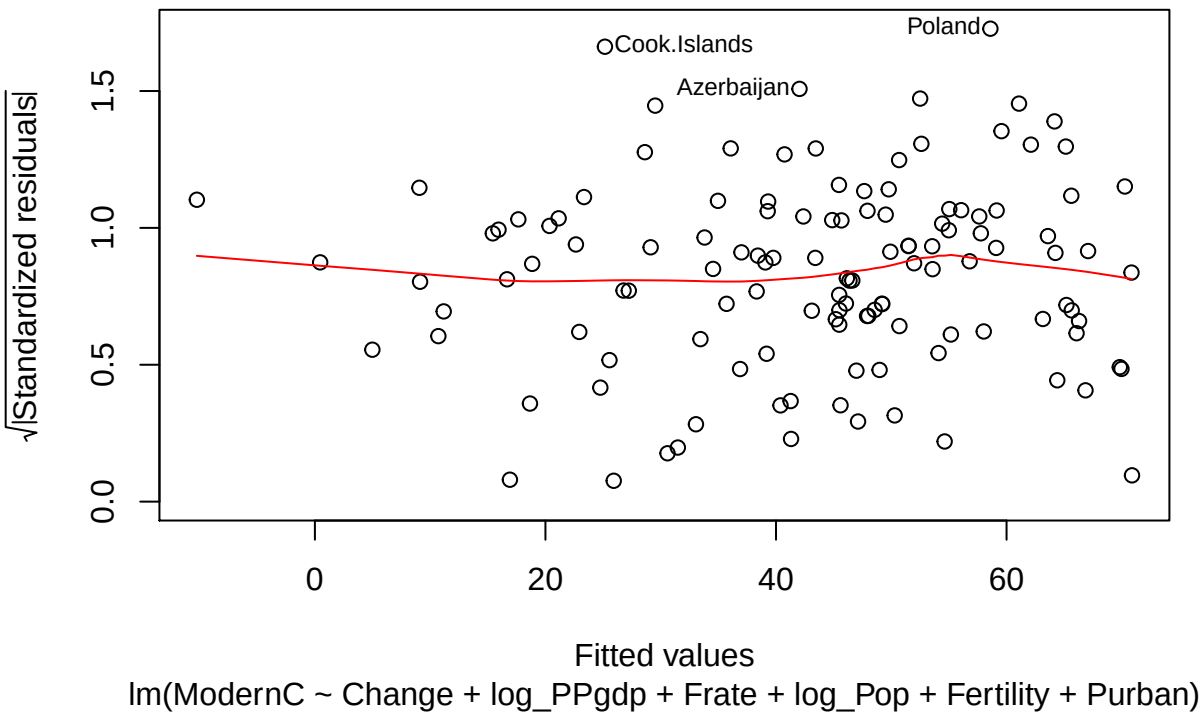
```
#BoxTidwell
bT.noTransformation = boxTidwell(ModernC ~ Pop + PPgdp, other.x = ~ Change + Frate + Fertility + Purban
bT.noTransformation
```

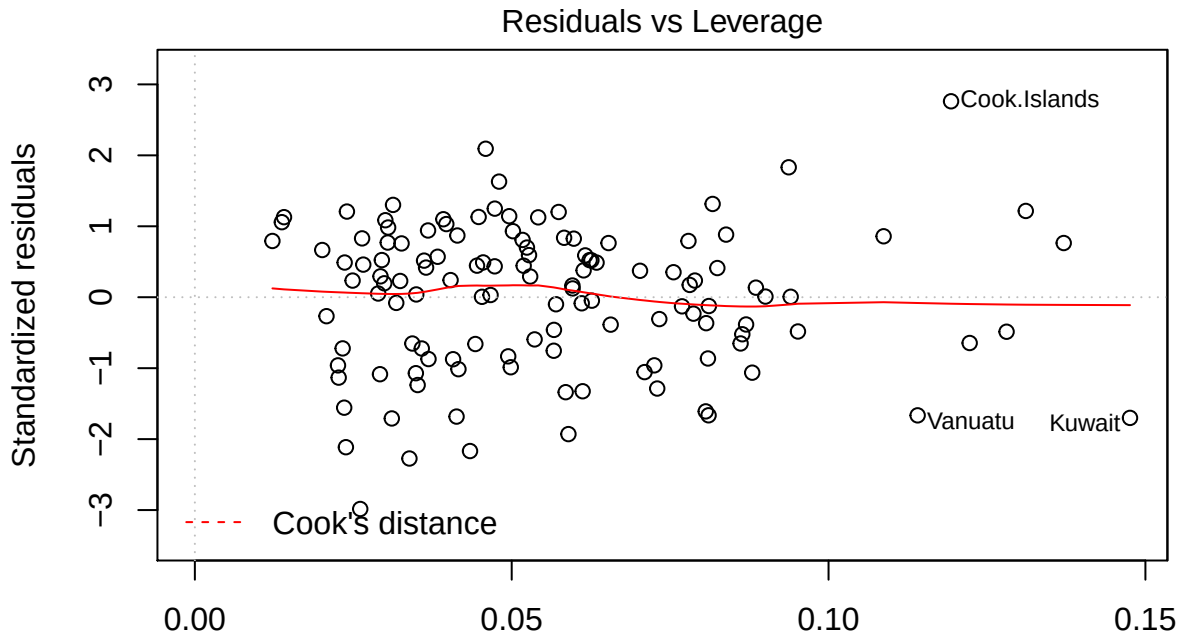
```
##       MLE of lambda Score Statistic (z) Pr(>|z|)
## Pop         0.40749             -0.7874   0.4310
## PPgdp      -0.12921             -1.1410   0.2539
##
## iterations =  4
```

```
#Data Frame with Pop and PPgdp Log Transformed
UN3_No_NA_Log = UN3_No_NA %>%
  rownames_to_column("Country") %>%
  mutate("log_PPgdp" = log(PPgdp), "log_Pop" = log(Pop)) %>%
  column_to_rownames("Country")

#Multiple Regression with Log.PPgdp and Log.Pop
LM_ModernC_Transformed = lm(ModernC ~ Change + log_PPgdp + Frate + log_Pop + Fertility + Purban, data =

#Plot for Multiple Regression with Log.PPgdp and Log.Pop
plot(LM_ModernC_Transformed)
```

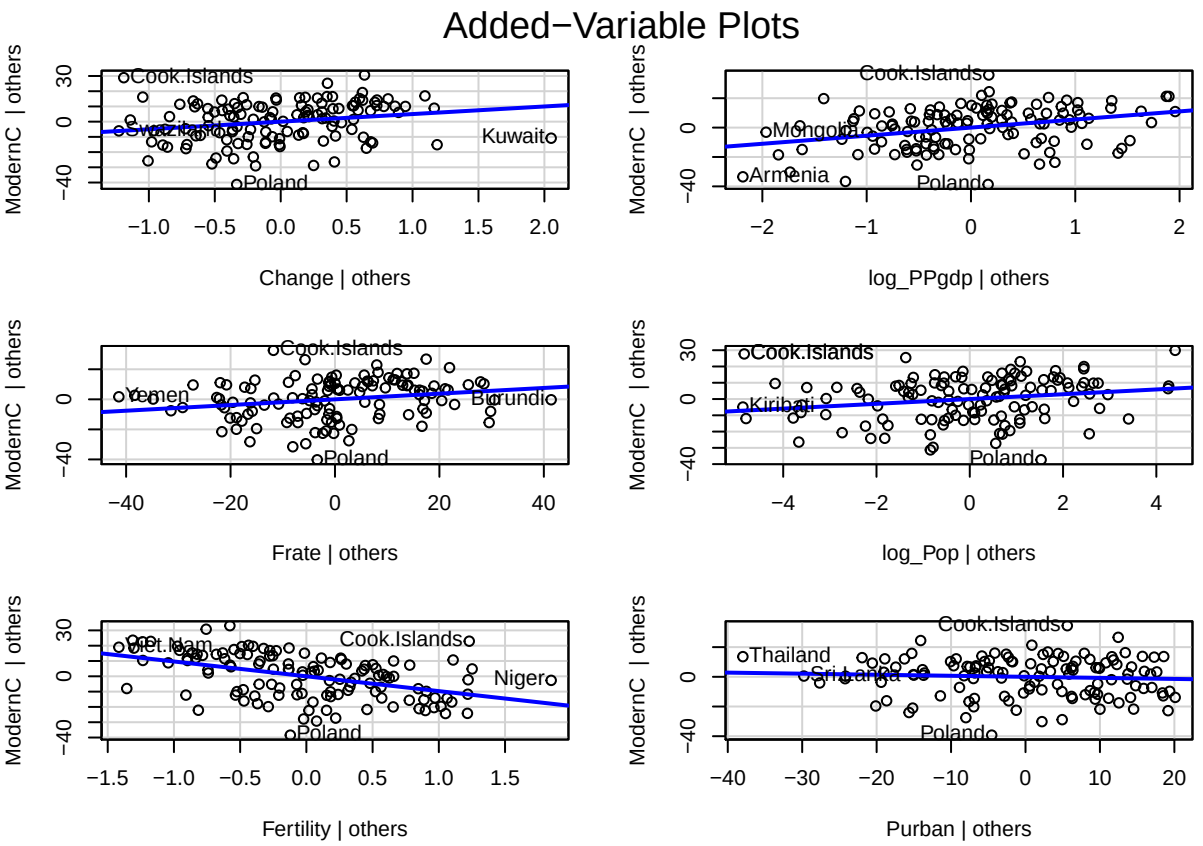



Normal Q−Q
Standardized residuals
Cook.Islands
Azerbaijan
Poland
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(ModernC ~ Change + log_PPgdp + Frate + log_Pop + Fertility + Purban)


Scale−Location
√|Standardized residuals|
Cook.Islands
Poland
Azerbaijan
0.0
0.5
1.0
1.5
0
20
40
60
Fitted values
lm(ModernC ~ Change + log_PPgdp + Frate + log_Pop + Fertility + Purban)




```
#AVPlot for Multiple Regression with Log.Ppgdp and Log.Pop
avPlots(LM_ModernC_Transformed)
```

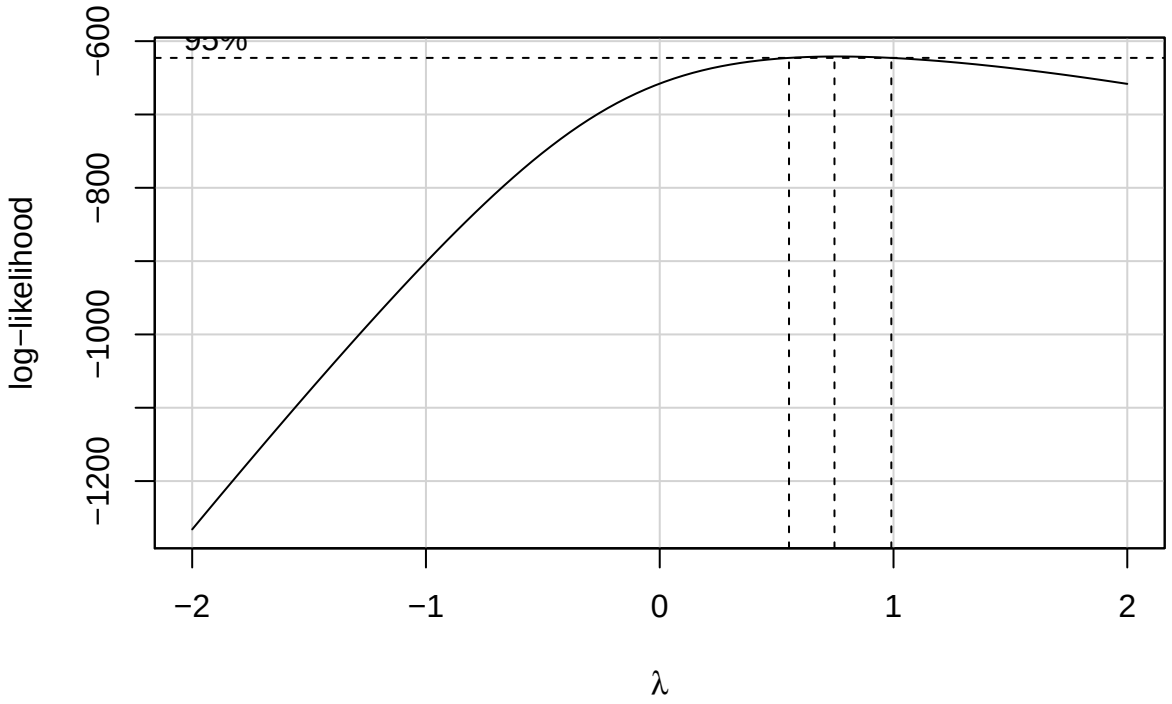


Comment:
After running boxTidwell for predictors Pop and PPgdp, I obtained outputs that suggested statistical insignificance of doing predictors' transformations. However, to not to overlook suggestions about transformations I had made in #3 and #5, I log-transformed the predictors, to see if I really end up with the outputs provided by boxTidwell. The results were different from what boxTidwell provided. Log transforming the predictors PPgdp and Pop actually altered the plots, providing avPlots with more linear trend, less pronounced residuals vs. fitted plot, and less pronounced residuals vs. leverage plot. Since log-transformations ended up improving many aspects of the linear model, despite the insignificance message from boxTidwell, I concluded to log-transform PPgdp and Pop.

I log-transformed Pop and PPgdp because the scatterplot matrix in #3 showed non-linear relationships between each of the aforementioned predictors and the response.

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

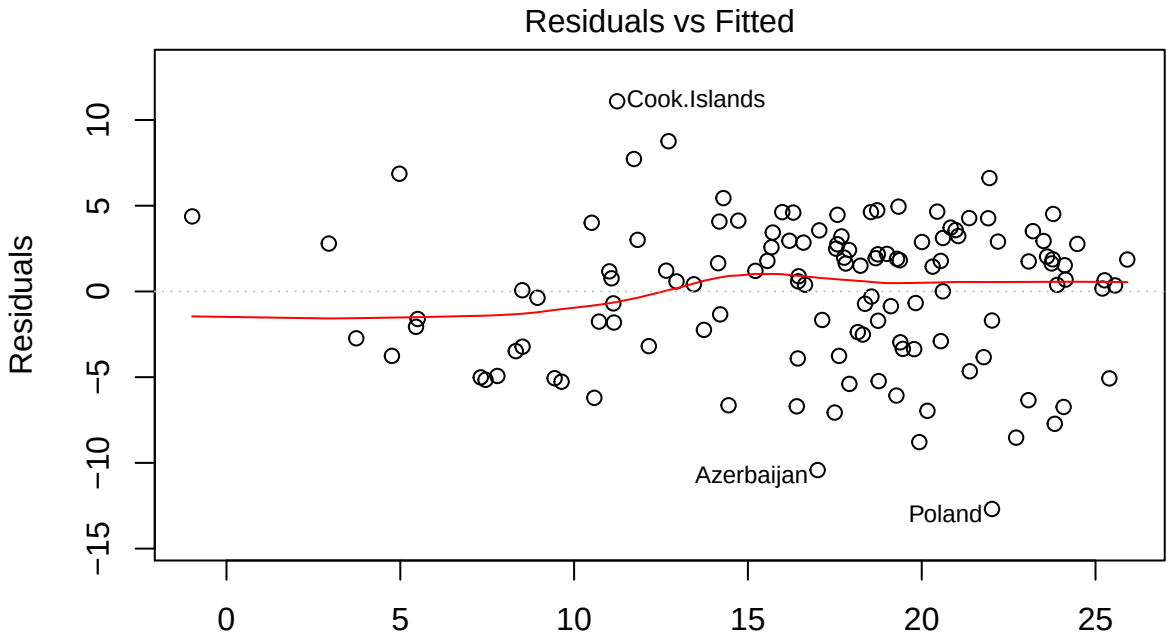
```
#BoxCox Plot for Predictor-Transformed Multiple Regression
boxCox(LM_ModernC_Transformed)
```

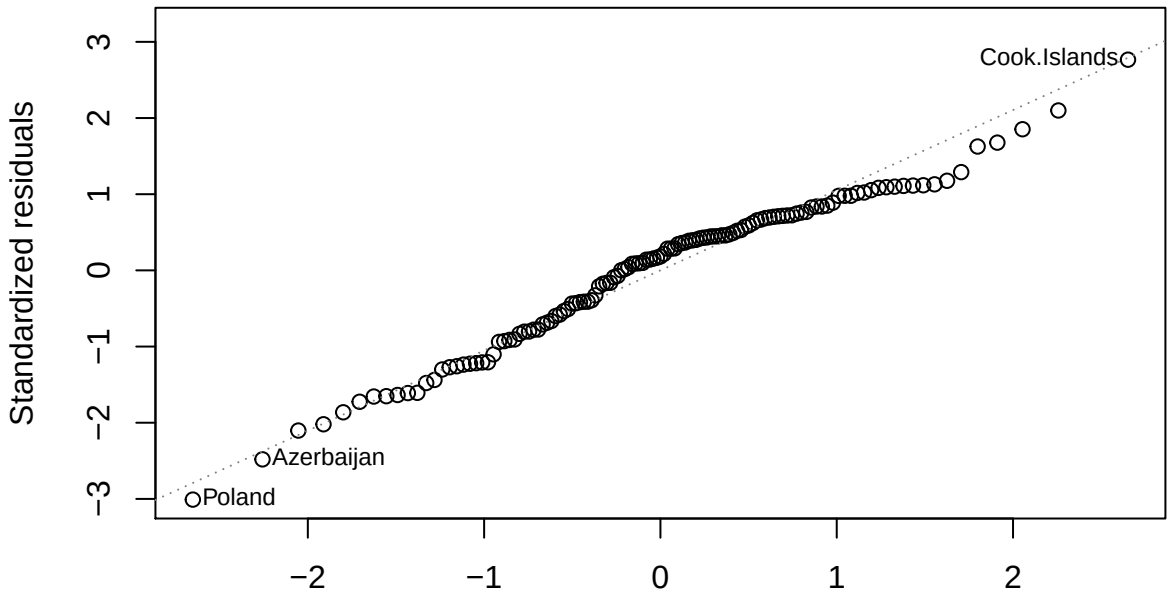


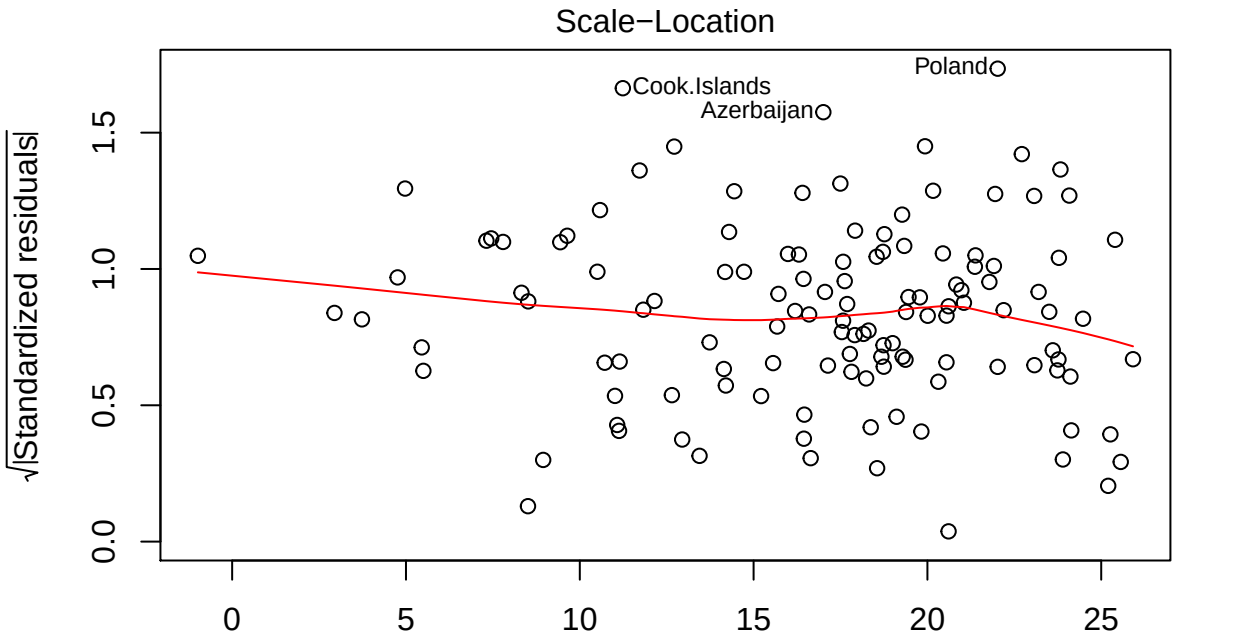

```
#Lambda for Optimal Response Transformation
powerTransform(LM_ModernC_Transformed)
```

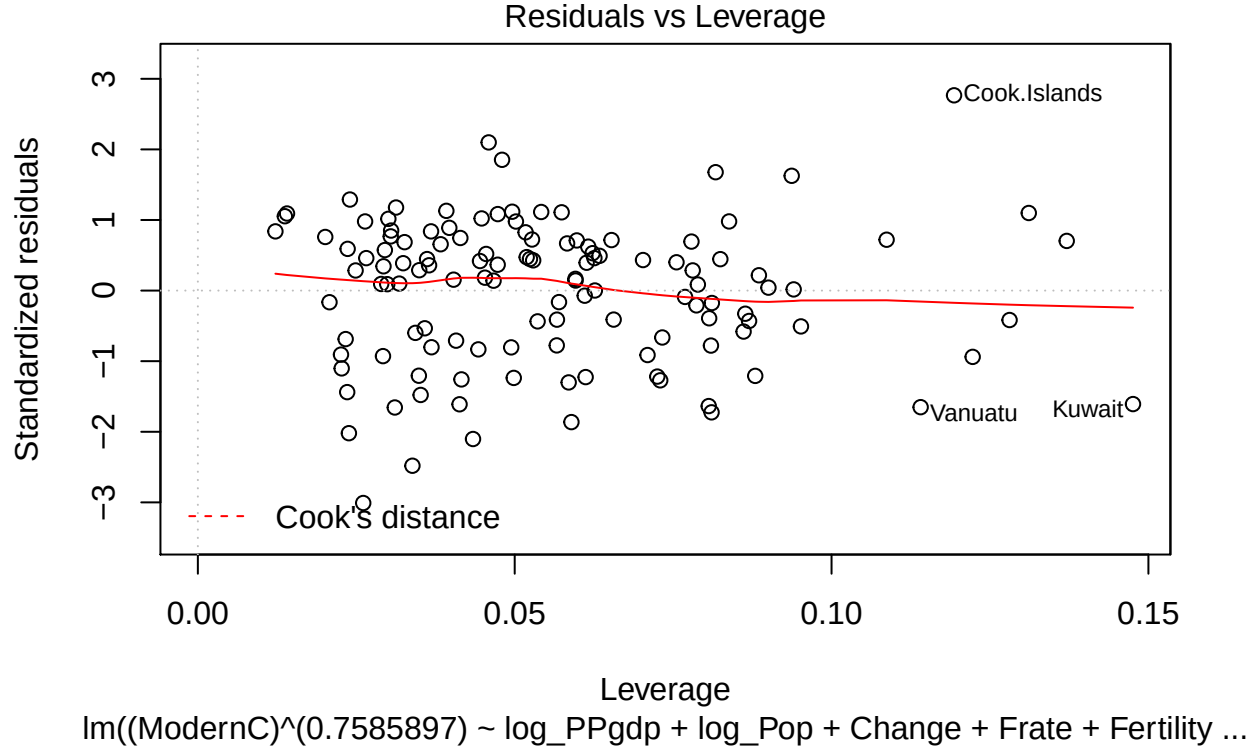
```
## Estimated transformation parameter
##        Y1
## 0.7585897
```

```
#Response-Transformed Multiple Regression
boxcox.transformed.lm = lm((ModernC)^(.7585897) ~ log_PPgdp + log_Pop + Change + Frate + Fertility + Pu

#Comparison of Response Transformed Multiple Regression to Predictors Transformed Regression
plot(boxcox.transformed.lm)
```

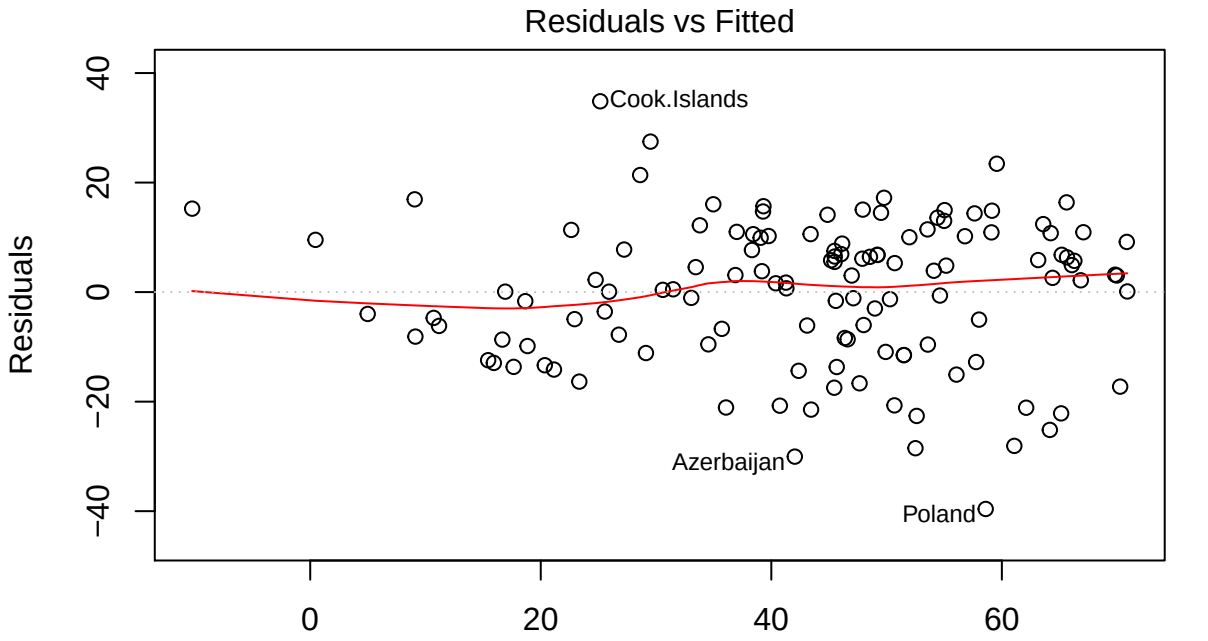

Residuals vs Fitted
Cook.Islands
Azerbaijan
Poland
Residuals
10
5
0
−5
−10
−15
0
5
10
15
20
25
Fitted values
lm((ModernC)^(0.7585897) ~ log_PPgdp + log_Pop + Change + Frate + Fertility ...

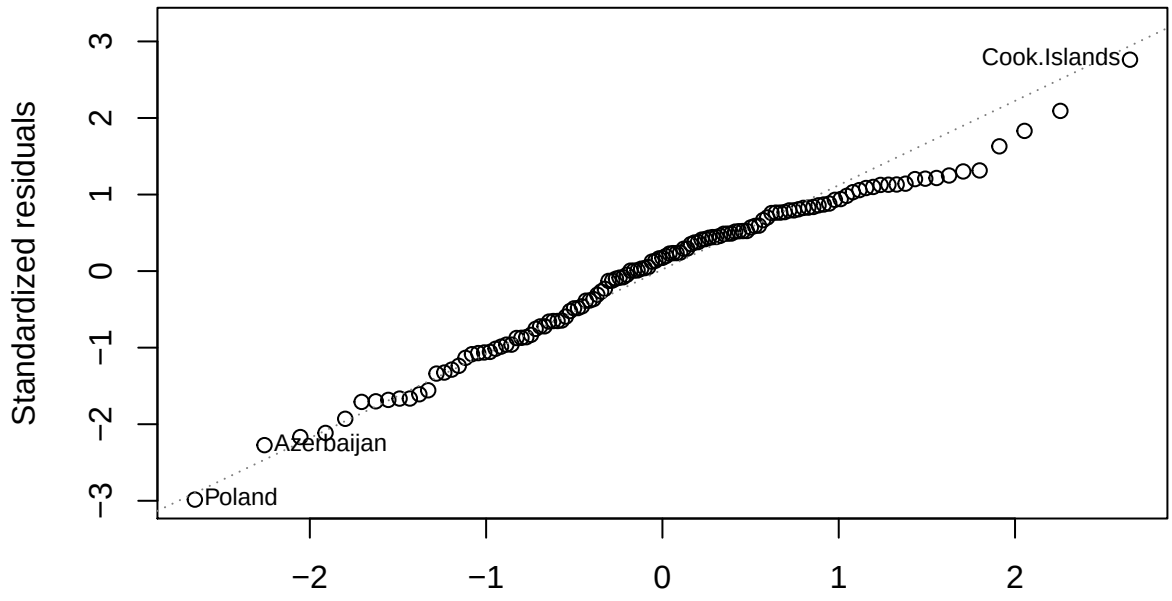

Normal Q−Q
Cook.Islands
Azerbaijan
Poland
Standardized residuals
3
2
1
0
−1
−2
−3
−2
−1
0
1
2
Theoretical Quantiles
lm((ModernC)^(0.7585897) ~ log_PPgdp + log_Pop + Change + Frate + Fertility ...

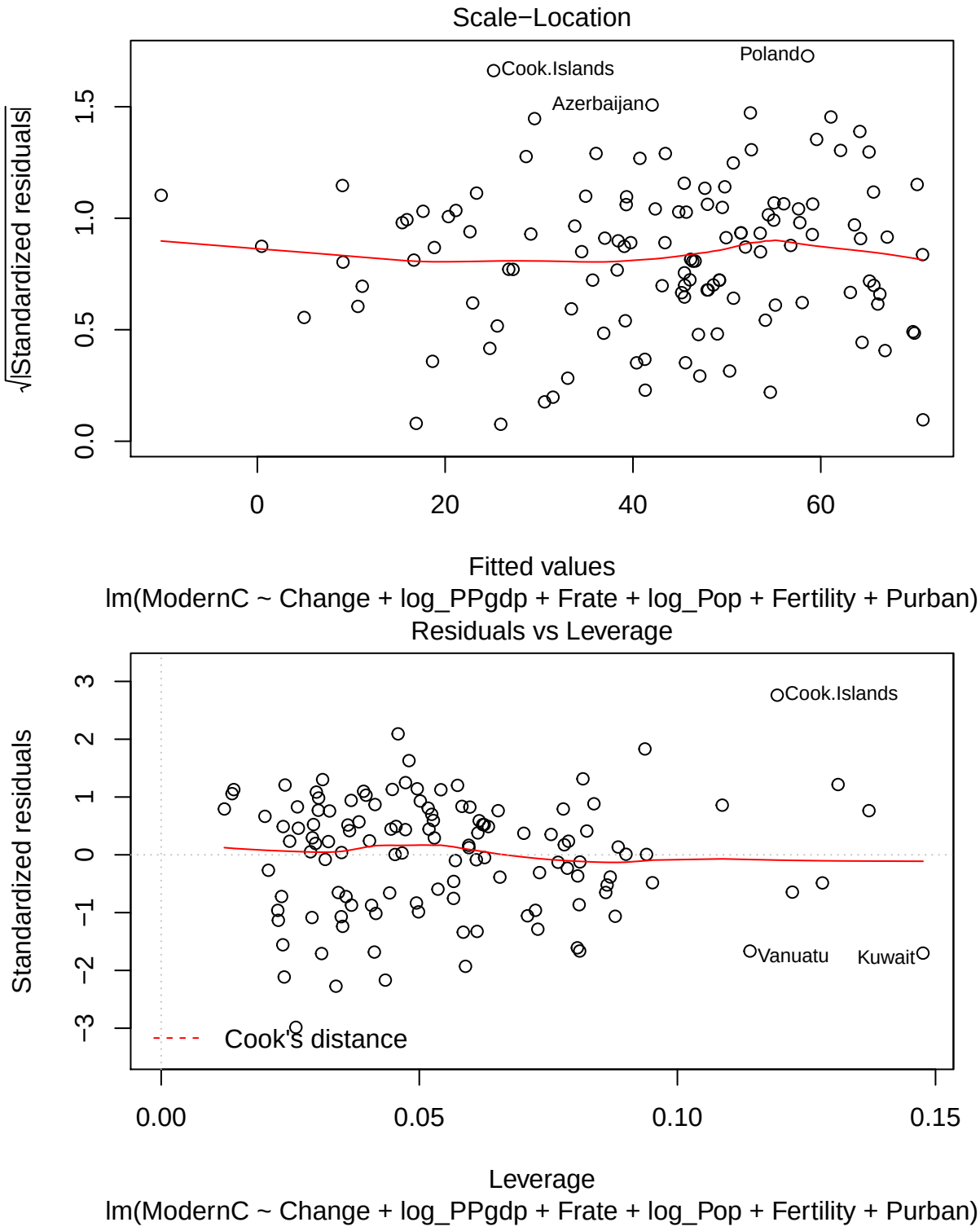





```
plot(LM_ModernC_Transformed)
```


Residuals vs Fitted
Residuals
Cook.Islands
Azerbaijan
Poland
40
20
0
−20
−40
0
20
40
60
Fitted values
lm(ModernC ~ Change + log_PPgdp + Frate + log_Pop + Fertility + Purban)


Normal Q−Q
Standardized residuals
Cook.Islands
Azerbaijan
Poland
3
2
1
0
−1
−2
−3
−2
−1
0
1
2
Theoretical Quantiles
lm(ModernC ~ Change + log_PPgdp + Frate + log_Pop + Fertility + Purban)



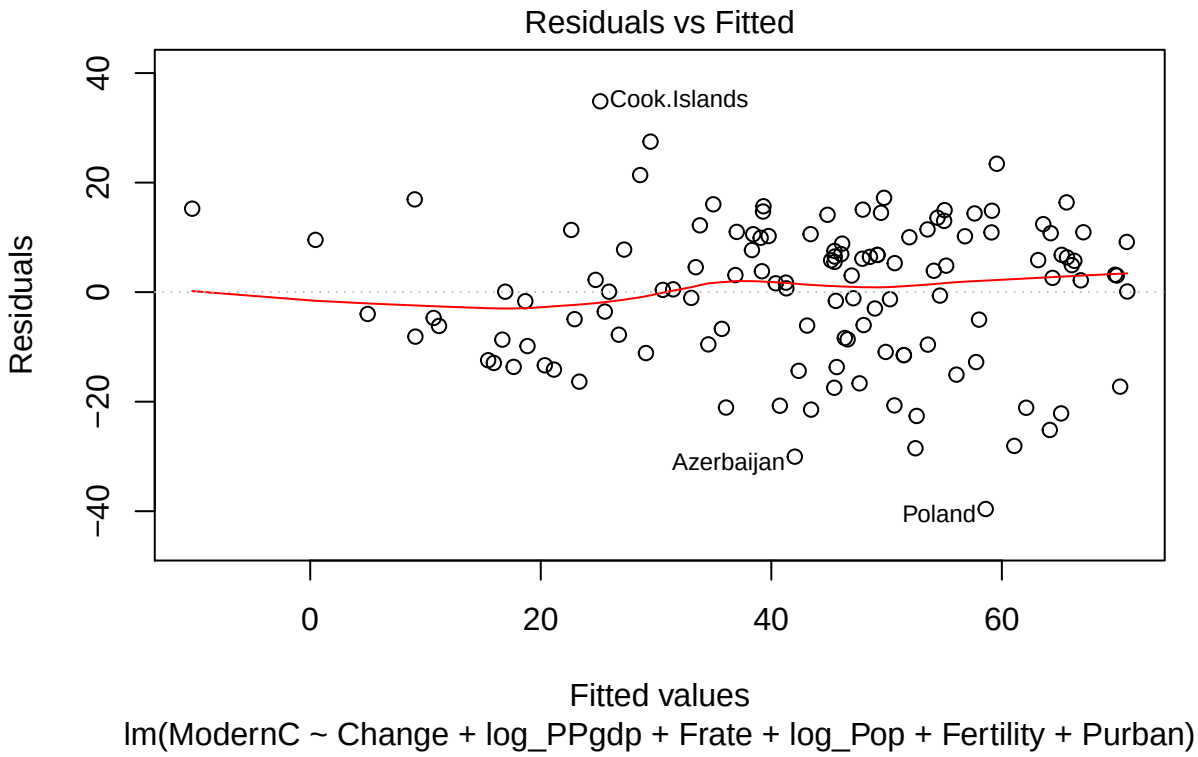
Comment:
BoxCox transformation suggested applying a power of .7585897 to the response variable, which I have implemented by constructing a new linear regression model. However, I noticed that the residuals plots generated from the response-transformed model were not significantly different from those generated from predictors-only transformed model, compelling me to conclude that the response transformation is not being helpful in improving the model. Thus, I elected not to transform the response.

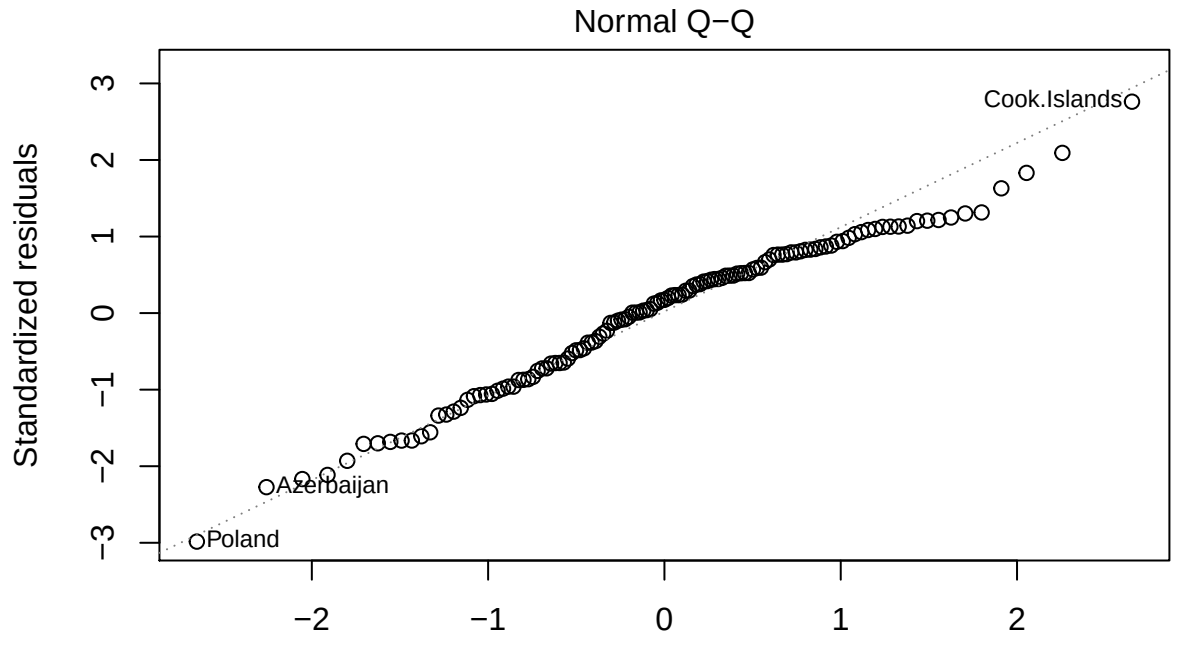
8. Fit the regression using the transformed variables. Provide residual plots and added variables plots and comment. If you feel that you need additional transformations of either the response or predictors, repeat any steps until you feel satisfied.

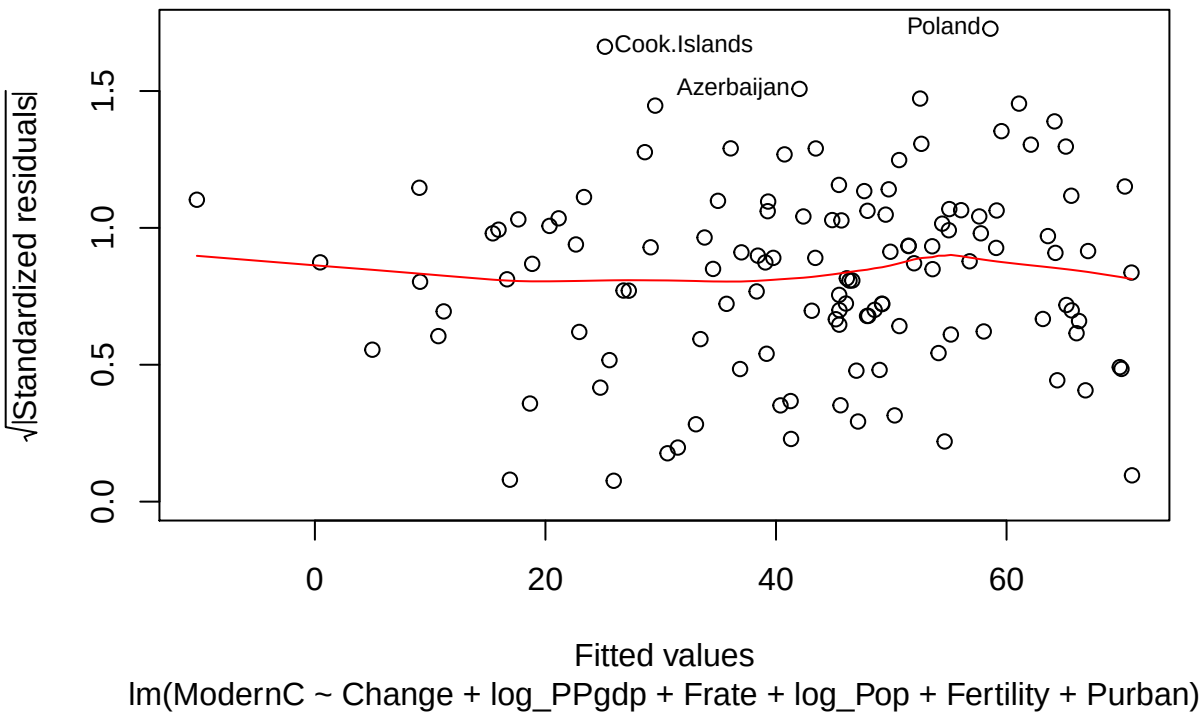
```
#Regression Using the Transformed Variables
LM_ModernC_Transformed
```

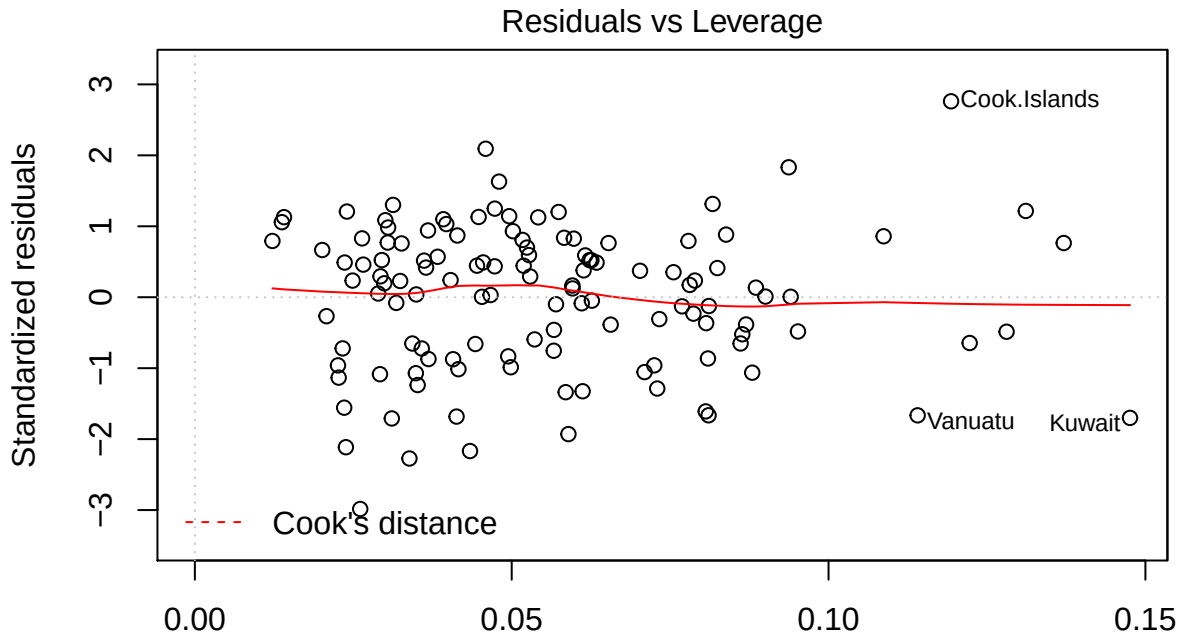
```
##
## Call:
## lm(formula = ModernC ~ Change + log_PPgdp + Frate + log_Pop +
##     Fertility + Purban, data = UN3_No_NA_Log)
##
## Coefficients:
## (Intercept)       Change    log_PPgdp        Frate      log_Pop
##     4.11547      4.99296      5.50728      0.18939      1.47207
##   Fertility       Purban
##    -9.67594     -0.07077
```

```
#Residual Plot
plot(LM_ModernC_Transformed)
```

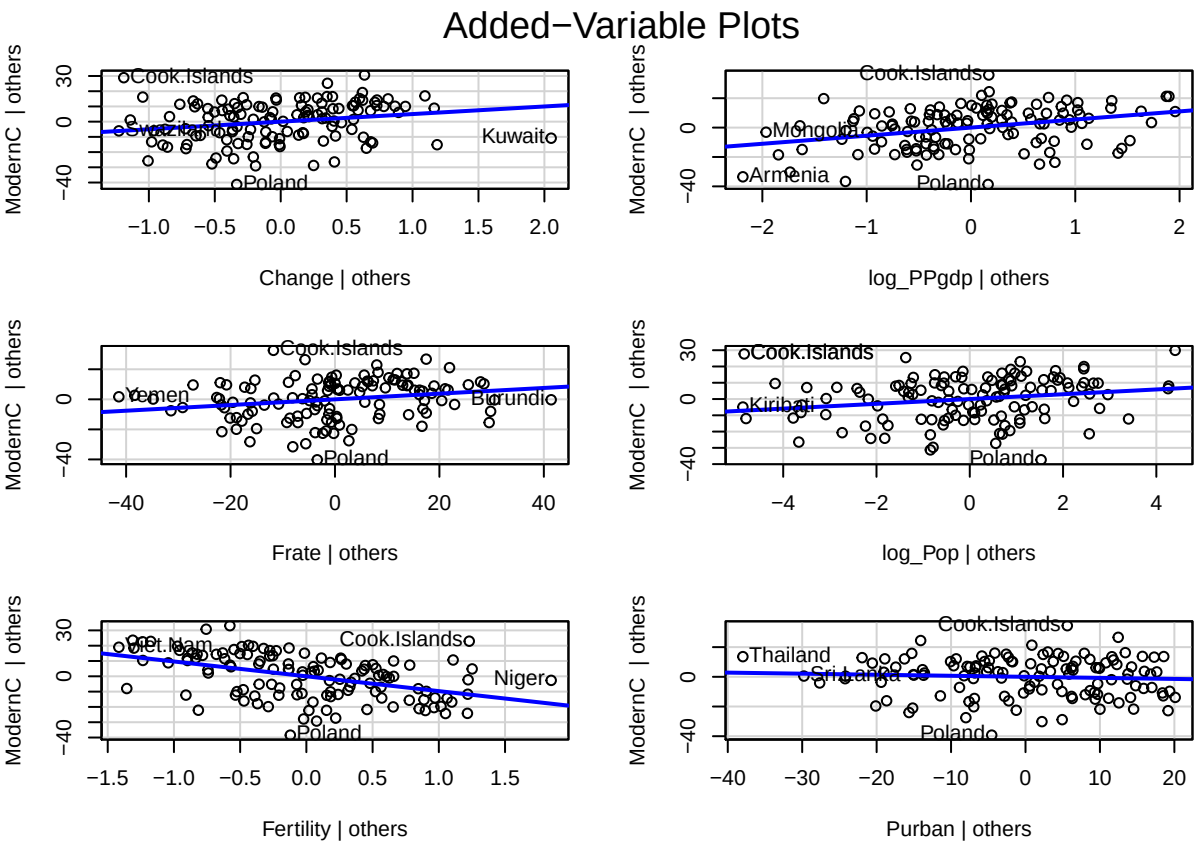



Normal Q−Q
Standardized residuals
Cook.Islands
Azerbaijan
Poland
−3
−2
−1
0
1
2
3
−2
−1
0
1
2
Theoretical Quantiles
lm(ModernC ~ Change + log_PPgdp + Frate + log_Pop + Fertility + Purban)


Scale−Location
√|Standardized residuals|
Cook.Islands
Poland
Azerbaijan
0.0
0.5
1.0
1.5
0
20
40
60
Fitted values
lm(ModernC ~ Change + log_PPgdp + Frate + log_Pop + Fertility + Purban)




```
#Added Variables Plot
avPlots(LM_ModernC_Transformed)
```



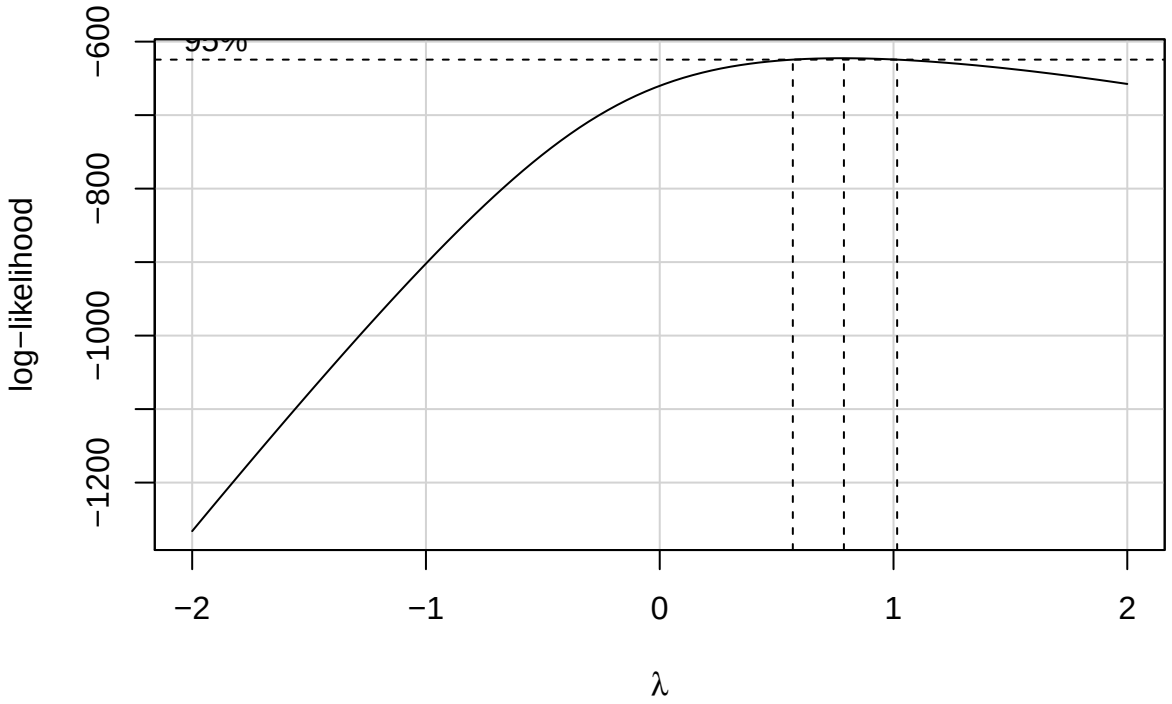
Note:
Please note that since I elected only to transform the predictors, my regression of the transformed variables for this problem is identical to the regression model from #6.

Comment: After the transformations had been done to the original regression model, I observed improvements in residuals plots. The red curve on "Residual vs. Fitted plot" became less pronounced; the right-tail of Normal Q-Q plot bent toward the normal line; the red curve on "Scale-Location" became less pronounced; and the red curve "Residuals vs. Leverage" became flatter. These evidences support that transformations done to the predictors improved the model, at least in terms of residuals.

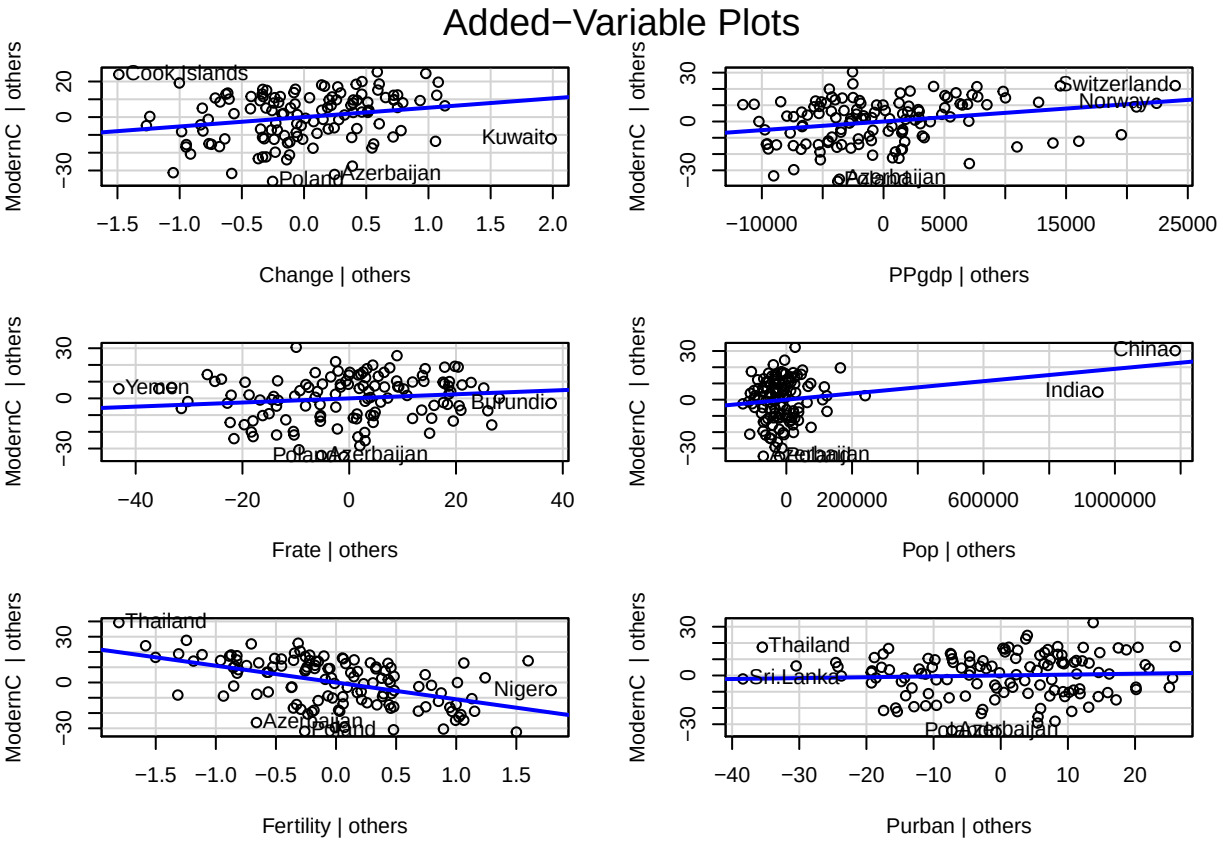
As for the added-variable plot, the transformations also improved it. Pop and PPgdp became more spread out and arguably more linear after the transformations.

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
boxCox(LM_ModernC)
```




```
#AVPlot to Detect Predictors that Need Transformations
avPlots(LM_ModernC)
```




```
#BoxTidwell
boxTidwell(ModernC ~ Pop + PPgdp, other.x = ~ Change + Frate + Purban + Fertility, data = UN3_No_NA)
```

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

Comment:
Unlike the previous problem, the boxCox function for this question did not suggesting me to transform the response, by including 1 in its 95% confidence interval - inclusion of 1 suggests that the power of the response is not statistically different from 1. As for BoxTidwell, it informed me, once again, to not to transform predictors, by providing relatively high p-values - I selected PPgdp and Pop as the predictors to transform after looking at the avPlot of linear model.

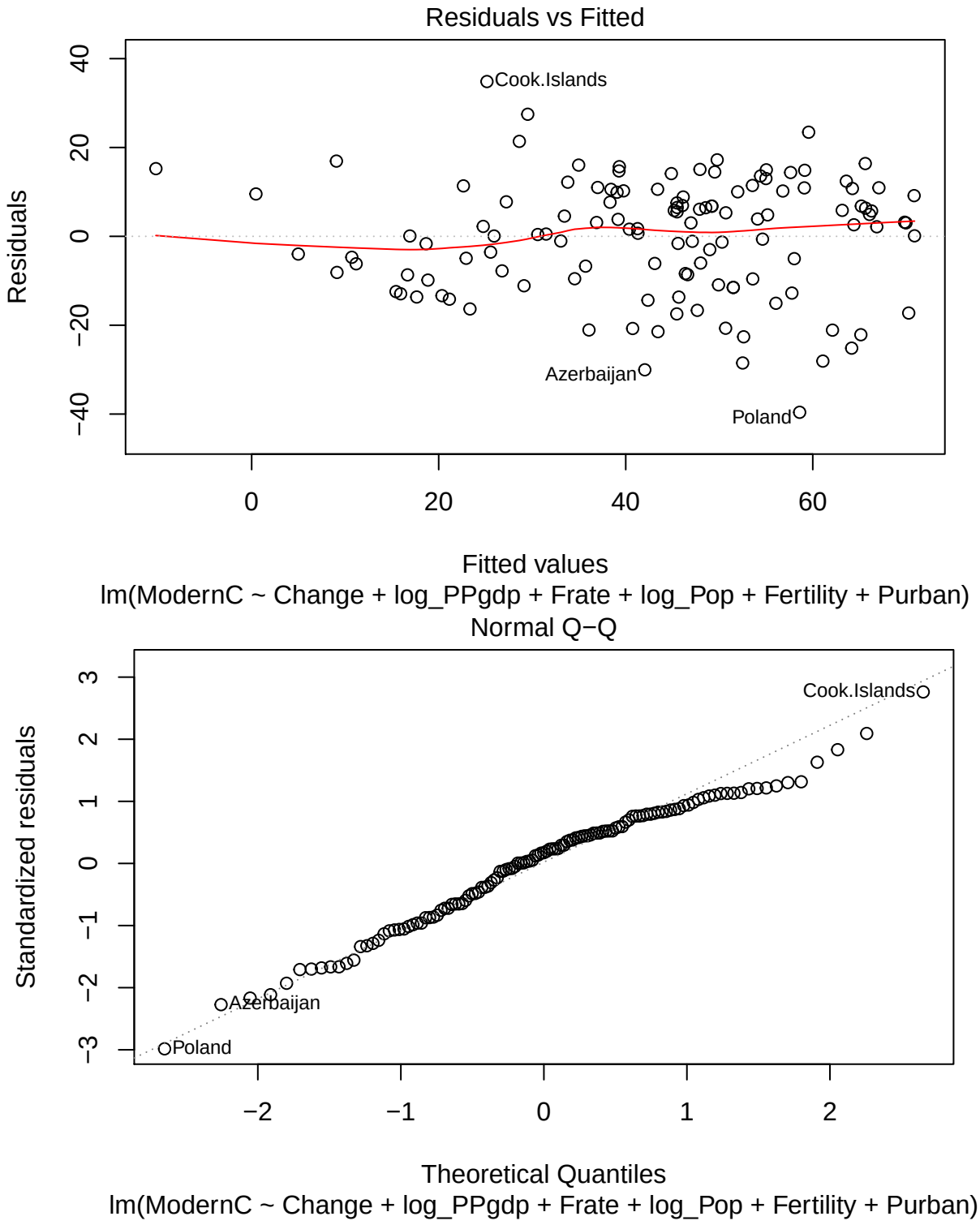
At the end, using the same logics as #6, I ended up with the same model in #8 - the post-boxCox model for this problem was identical to the linear model used for boxTidwell in #6, which meant I should ultimately end up with the same result as #6 (and therefore #7 and #8).

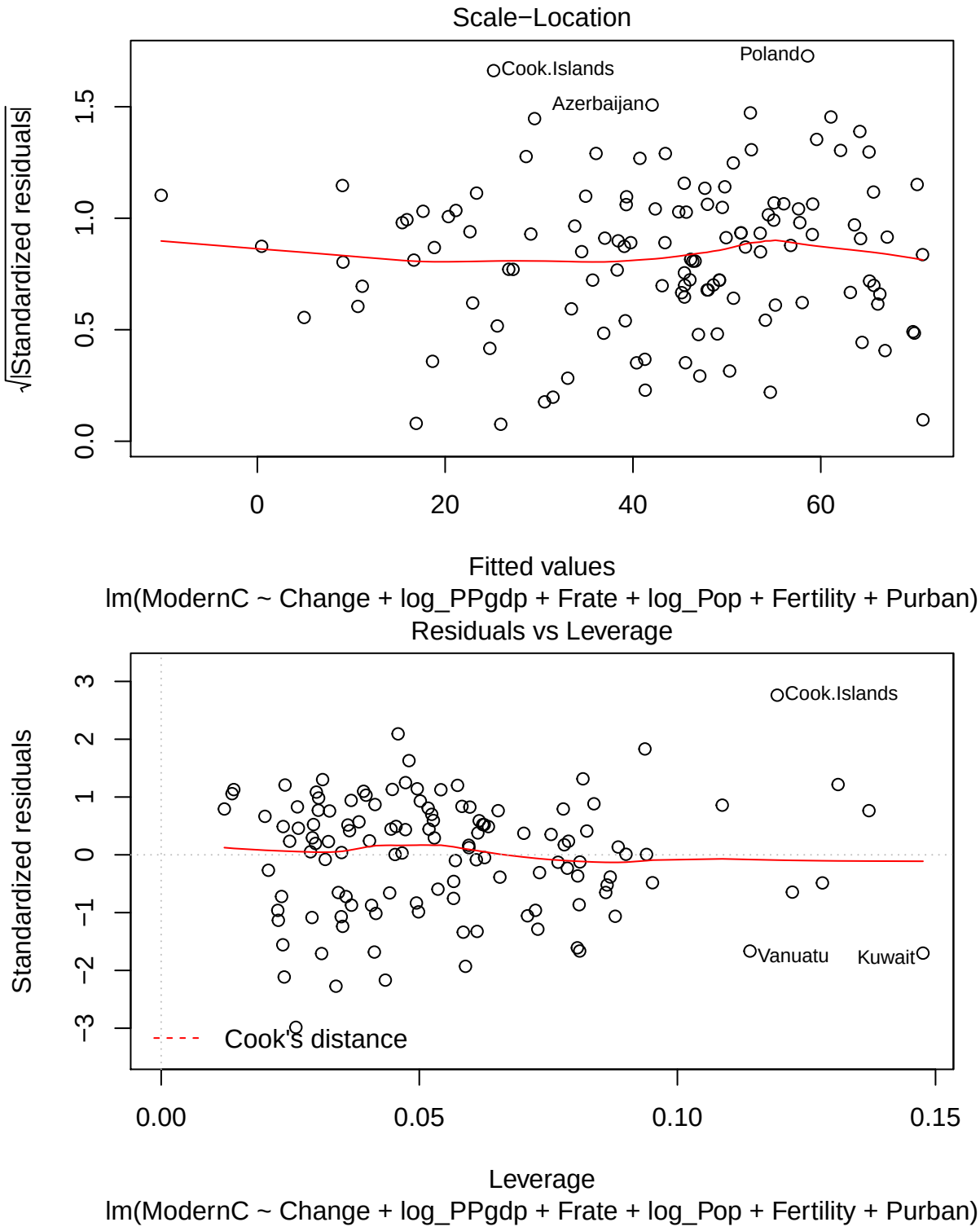
10. Are there any outliers or influential points in the data? Explain. If so, refit the model after removing any outliers and comment on residual plots.

```
#Bonferroni Correction for Outliers
pval = 2*(1 - pt(abs(rstudent(LM_ModernC_Transformed)), LM_ModernC_Transformed$df - 1))
rownames(UN3_No_NA_Log)[pval < .05/nrow(UN3_No_NA_Log)]
```

```
## character(0)
```

```r
#Cook's Distance for Influential Points
plot(LM_ModernC_Transformed)
```





Comment:
According to Bonferroni correction, there are no outliers in the data, and therefore no data need to be removed. As for influential points, cook's distance was absent on the residuals plot, which suggests that no data are influential points.

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!

```
#Multiple Regression with Purban Removed
lm3 = lm(ModernC ~ Change + log_Pop + log_PPgdp + Frate + Fertility, data = UN3_No_NA_Log)

#Justification of Removing Purban
anova(lm3, LM_ModernC_Transformed)
```

```
## Analysis of Variance Table
##
## Model 1: ModernC ~ Change + log_Pop + log_PPgdp + Frate + Fertility
## Model 2: ModernC ~ Change + log_PPgdp + Frate + log_Pop + Fertility +
##     Purban
##   Res.Df   RSS Df Sum of Sq      F Pr(>F)
## 1    119 21420
## 2    118 21325  1    95.016 0.5258 0.4698
```

```
#Summary of Regression with Purban Removed
summary(lm3)
```

```
##
## Call:
## lm(formula = ModernC ~ Change + log_Pop + log_PPgdp + Frate +
##     Fertility, data = UN3_No_NA_Log)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -39.276  -9.928   2.572  10.253  34.442
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.10208   14.47959   0.283  0.77744
## Change       4.69776    2.03274   2.311  0.02255 *
## log_Pop      1.44122    0.62606   2.302  0.02307 *
## log_PPgdp    4.85936    1.08214   4.491 1.65e-05 ***
## Frate        0.19955    0.07568   2.637  0.00949 **
## Fertility   -9.27842    1.67499  -5.539 1.85e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.42 on 119 degrees of freedom
## Multiple R-squared:  0.6243, Adjusted R-squared:  0.6085
## F-statistic: 39.55 on 5 and 119 DF,  p-value: < 2.2e-16
```

```
#Table of 95% Confidence Intervals & Estimates
confint_predictors = as.data.frame(confint(lm3))
confint_predictors["log_Pop", ] = confint_predictors["log_Pop", ] * log(1.1)
confint_predictors["log_PPgdp", ] = confint_predictors["log_PPgdp", ] * log(1.1)
rownames(confint_predictors) = c("(Intercept)", "Change", "Pop (with 10% Increase)", "PPgdp (with 10% In

confint_predictors = confint_predictors %>%
   rownames_to_column("Pred") %>%
  mutate(Estimate = c(4.10208, 4.69776, 1.44122, 4.85936, 0.19955, -9.27842)) %>%
```

```
   column_to_rownames("Pred")

kable(confint_predictors, format = "markdown", digits = 2)
```

	2.5 %	97.5 %	Estimate
(Intercept)	-24.57	32.77	4.10
Change	0.67	8.72	4.70
Pop (with 10% Increase)	0.02	0.26	1.44
PPgdp (with 10% Increase)	0.26	0.67	4.86
Frate	0.05	0.35	0.20
Fertility	-12.60	-5.96	-9.28

Comment:
For the final model, I elected to remove Purban. I ran anova and observed that adding Purban to the linear model is not statistically significant.

Coefficients Interpretation: I attempted including the interpretation of each coefficient in the table, but due to the bad PDF output, I decided to provide them here.

Intercept: When all the other predictors are equal to 0, ModernC will be equal to 4.10208. This, however, is meaningless, because a number of predictors cannot be 0 practically. For example, population being 0 is theoretically possible but practically near impossible.

Change: All else equal, 10% increase in Change will increase ModernC by roghly .47.

Pop: All else equal, 10% increase in Pop will increase ModernC by roughly .14.

PPgdp: All else equal, 10% increase in PPgdp will increase ModernC by roghly .46.

Frate: All else equal, 10% increase in Frate will increase ModernC by .02.

Fertility: All else equal, 10% increase in Fertility will decrease ModernC by .93.

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

Paragraph:
After all the statistical analyses performed on the given data set, we have made a number of discoveries. The first finding is, there are no outliers and no influential points in the supplied data set, based on proven statistical methods. The second is, the data about percent of population that is urban, in 2001, was not helpful in predicting percent of unmarried women using a modern method of contraception, with all the other predictors present in the model; the predictor was removed in the final model. The third is, the relationships between ModernC and PPgdp and between ModernC and Pop were not linear, which means that the changes in PPgdp and Pop do not always change ModernC in a constant manner. Our last finding is, Fertility is seemed to have the highest non-neglegible effect on the ModernC, which may require some more study on it.

Please note that all these statements are based on the supplied data set with missing values removed.

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.*

Prove:

$\hat{\beta_0} = 0$

Note:

X_j is defined to be the j^{th} column of X

Proof:

$$\hat{e}_{(Y)} = \hat{\beta_0} + \hat{\beta_1} e_{X_j}$$

Note that $\hat{e}_{(Y)} = (I - H)Y$, $\hat{e}_{X_j} = (I - H)X_j$ and $\hat{\beta_0} = \vec{1}_{nX1}\hat{\beta_0}$. Hereafter, assume $\vec{1}_{nX1} = \vec{1}$

$$\hat{e}_{(Y)} = \hat{\beta_0} + \hat{\beta_1} e_{X_j}$$

$$(I - H)Y = \vec{1}\hat{\beta_0} + \hat{\beta_1}(I - H)X_j$$

Note that $\hat{\beta_1} = [X_j^T(I-H)^T(I-H)X_j]^{-1}X_j^T(I-H)^TY$ because the predictor for the coefficient is $(I-H)X_j$

$$(I - H)Y = \vec{1}\hat{\beta_0} + [X_j^T(I - H)^T(I - H)X_j]^{-1}X_j^T(I - H)^TY(I - H)X_j$$

Recall that $(I - H)^T = (I - H)$ and $(I - H)^2 = (I - H)$

$$(I - H)Y = \vec{1}\hat{\beta_0} + [X_j^T(I - H)X_j]^{-1}X_j^T(I - H)Y(I - H)X_j$$

Multiplying X_j^T to both sides of the equation

$$X_j^T(I - H)Y = X_j^T\vec{1}\hat{\beta_0} + X_j^T[X_j^T(I - H)X_j]^{-1}X_j^T(I - H)Y(I - H)X_j$$

Here, note that $[X_j^T(I - H)X_j]^{-1}$ and $X_j^T(I - H)Y$ are scalars, which means they can be moved within the equation.

$$X_j^T(I - H)Y = X_j^T\vec{1}\hat{\beta_0} + X_j^T(I - H)X_j[X_j^T(I - H)X_j]^{-1}X_j^T(I - H)Y$$

Recall that $X_j^T(I - H)X_j[X_j^T(I - H)X_j]^{-1} = I$

$$X_j^T(I - H)Y = X_j^T\vec{1}\hat{\beta_0} + IX_j^T(I - H)Y$$

$$X_j^T(I-H)Y = X_j^T\vec{1}\hat{\beta}_0 + X_j^T(I-H)Y$$

Subtracting $X_j^T(I-H)Y$ from both sides, we obtain:

$$\vec{0} = X_j^T\vec{1}\hat{\beta}_0$$

$$\Rightarrow 0 = \sum_{i=1}^{n} X_{i,j}\hat{\beta}_0$$

$$\Rightarrow 0 = \hat{\beta}_0$$

$\therefore$ Intercept in the added variable scatter plot will always be zero

Prove:
$\frac{1}{n}\vec{1}\hat{e}_Y = 0$

Given:
$1_n^T(I-H) = 0$

Proof:

$$e_Y = (I-H)Y$$

Recall that $(I-H)Y = Y - \hat{\beta}_0 - \sum_{i=1}^{p}\hat{\beta}_i X_i$ since the projection matrix H has a colum of ones

.
Multiply $\vec{1}^T$ to both sides of the equation

$$\vec{1}^T e_Y = \vec{1}^T(I-H)Y$$

Note that $1_n^T(I-H) = 0$ is given

$$\vec{1}^T e_Y = 0(Y)$$

$$\vec{1}^T e_Y = 0$$

$$\frac{1}{n}\vec{1}^T e_Y = \frac{0}{n}$$

$$\frac{1}{n}\vec{1}^T e_Y = 0$$

$\therefore$ The sample mean of residuals will always be zero

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 on e_X
e_Y = residuals(lm3)
e_X = residuals(lm(Purban ~ Change + log_Pop + log_PPgdp + Frate + Fertility, data = UN3_No_NA_Log))

#Slope Comparison
lm_e_Y_on_e_X = lm(e_Y ~ e_X)
LM_ModernC_Transformed$coefficients["Purban"]
```

```
##      Purban
## -0.07076799
```

```
x = data.frame(e_X = lm(e_Y ~ e_X)$coefficients["e_X"], LM_ModernC_Transformed$coefficients["Purban"], i

kable(x, col.names = c("e_X", "Model"), format = "markdown")
```

	e_X	Model
Coefficients	-0.070768	-0.070768

```
#AVPlots
avPlots(lm_e_Y_on_e_X)
```