HW2 STA521 Fall18

[Yiwei Gong yg140 ywgej9]

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?

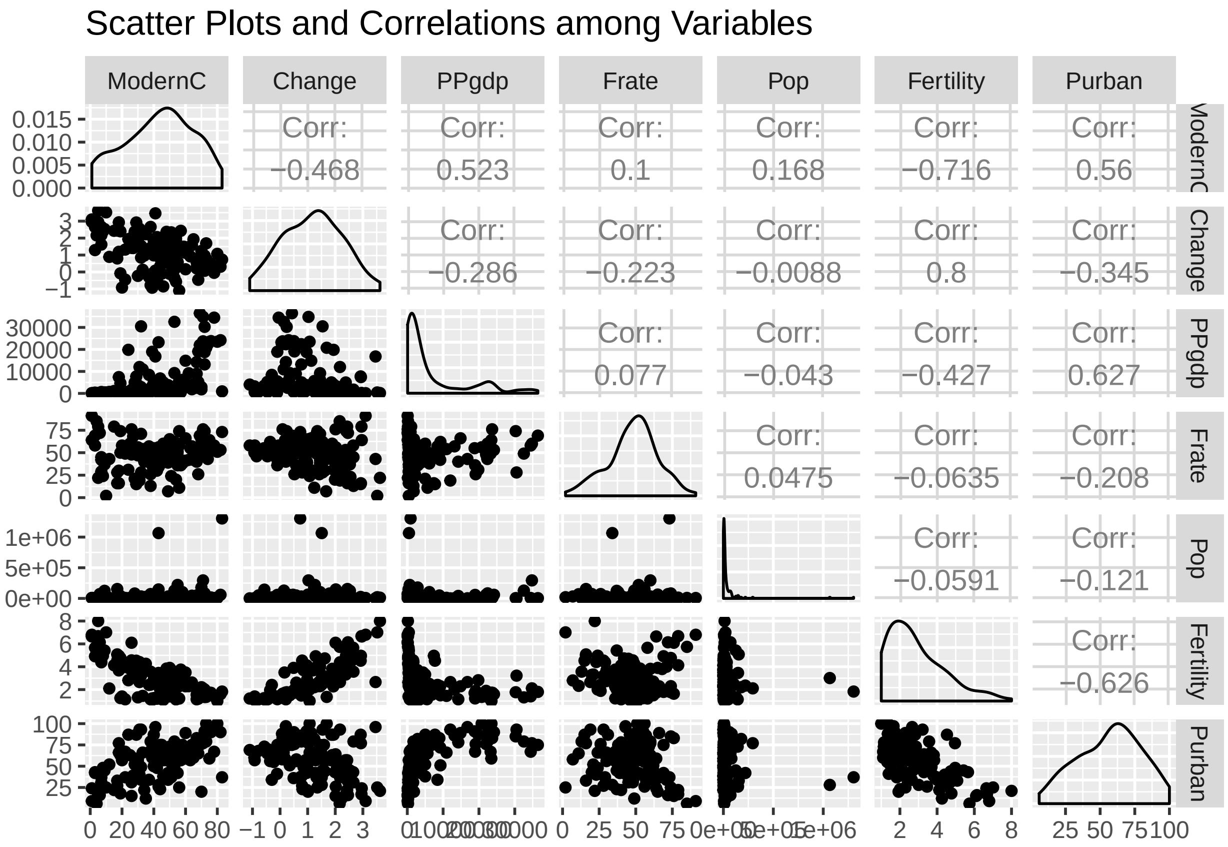
```
##     ModernC          Change           PPgdp            Frate      
##  Min.   : 1.00   Min.   :-1.100   Min.   :   90   Min.   : 2.00  
##  1st Qu.:19.00   1st Qu.: 0.580   1st Qu.:  479   1st Qu.:39.50  
##  Median :40.50   Median : 1.400   Median : 2046   Median :49.00  
##  Mean   :38.72   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                      
```

Answer: There are 7 variables having missing data. ModernC, Change, PPgdp, Frate, Pop, Fertility and Purban all are quantitative.

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

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



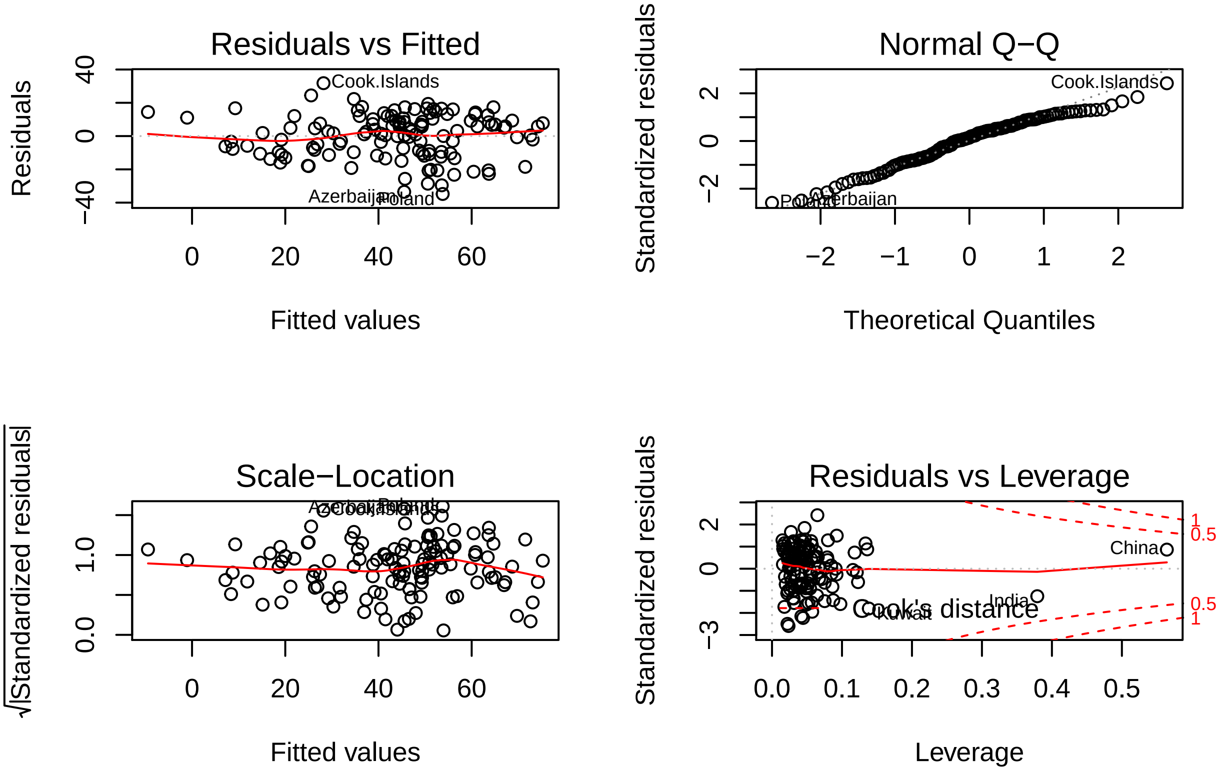
Answer: There seem downward linear trendings in ModernC when Change and Fertility increase, but ModernC seems to increase together with Purban, and when PPgdp increases, ModernC seems to increase exponentially. Two points are suspicious to be outliers. Transformation of Pop may be required as the range of Pop is considerably large.

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?

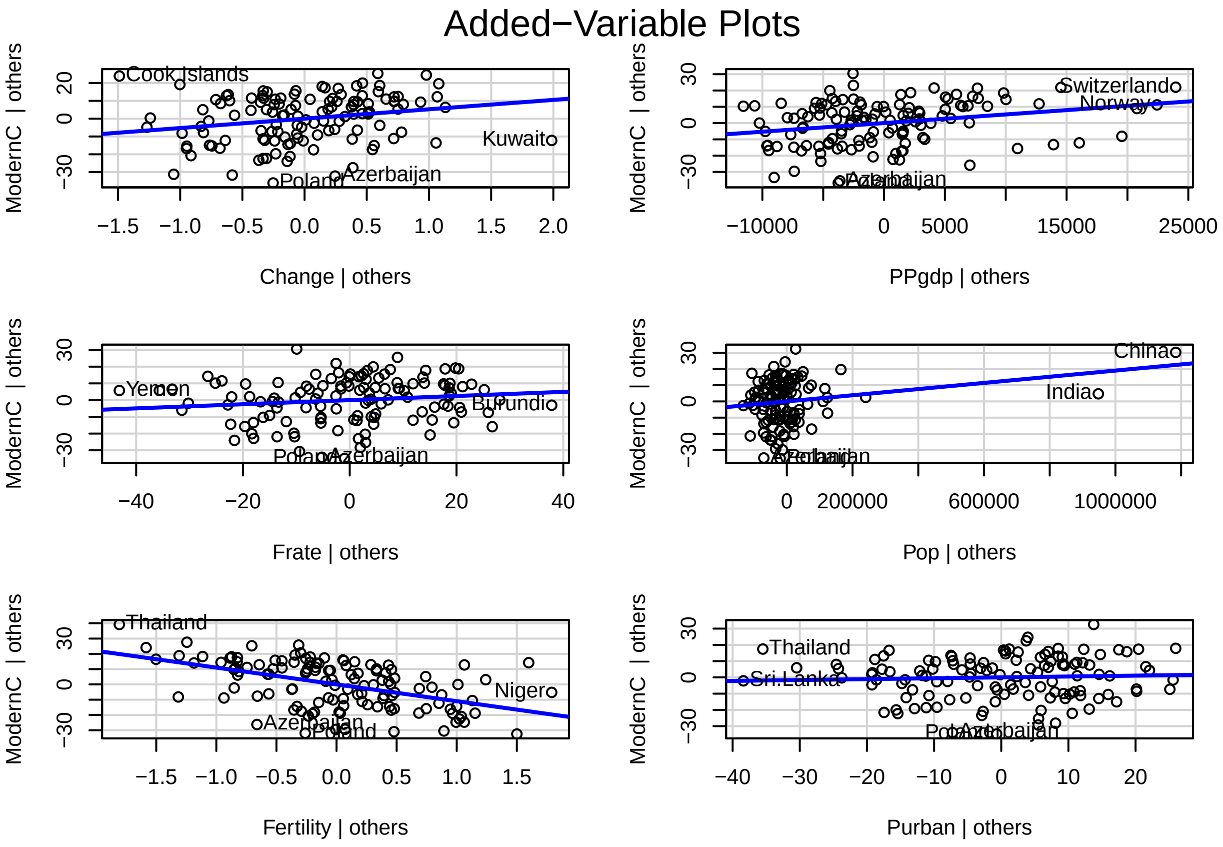
```
##
## Call:
## lm(formula = ModernC ~ Change + PPgdp + Frate + Pop + Fertility +
##     Purban, data = UN.nna)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -34.781  -9.698   1.858   9.327  31.791
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  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
```

```
## [1] 125
```



The Residuals vs Fitted plot suggests constant and 0 expectation of residuals, though fluctuated in the middle, and the Scale-Location plot shows possible violation of constant variance. The Normal QQ plot fits well in the middle, though the point of Poland seems strange. The Residuals vs Leverage suggests there is no highly influential point. 210 observations exist, but only 125 observations are used in this model fitting, since they don't have missing values.

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?



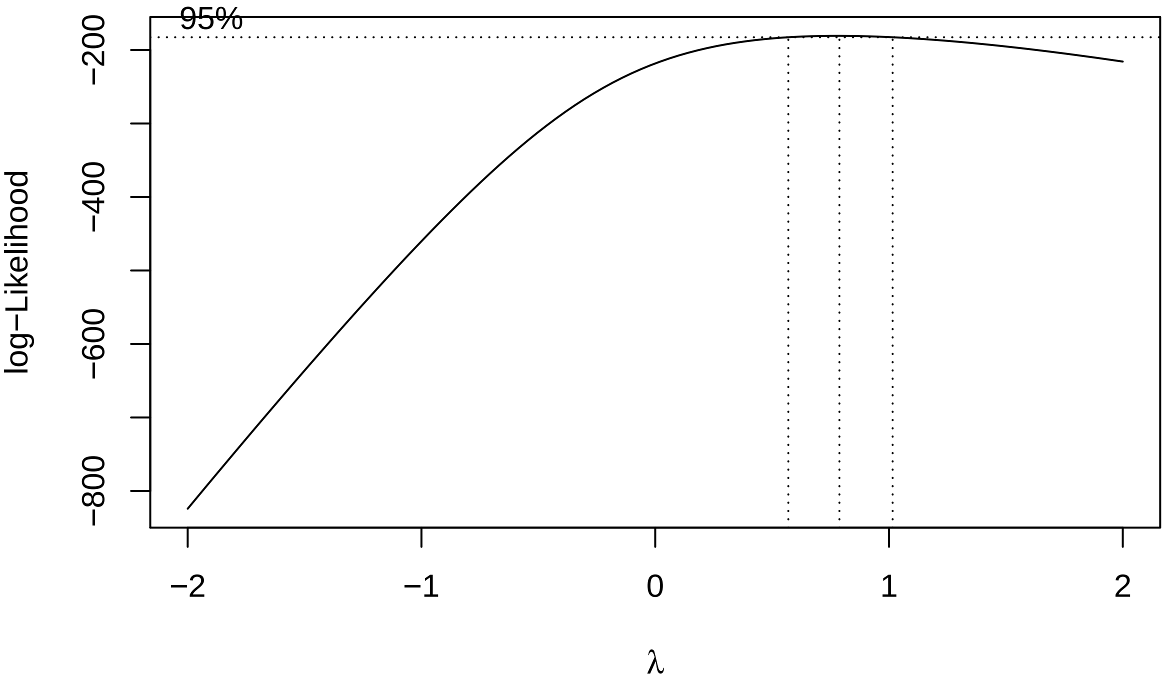
Answer: Transformation seems needed for Pop, as most of the points cluster together except two. PPgdp might need transformation as well. China and India in terms of Pop seem influential, though the Residual vs Leverage does not suggest that.

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.

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

Answer: When boxTidwell is applied to PPgdp and Pop, MLE are closer to 0, so may be a log transformation, but there does not seem significant evidence for transformation, since both p-values are pretty large. However, the scatterplots shows that (log(Pop), ModernC) and (log(PPgdp), ModernC) illustrate clearer linear relationship. Therefore, log(Pop) and log(PPgdp) seem proper candidates for transformation.

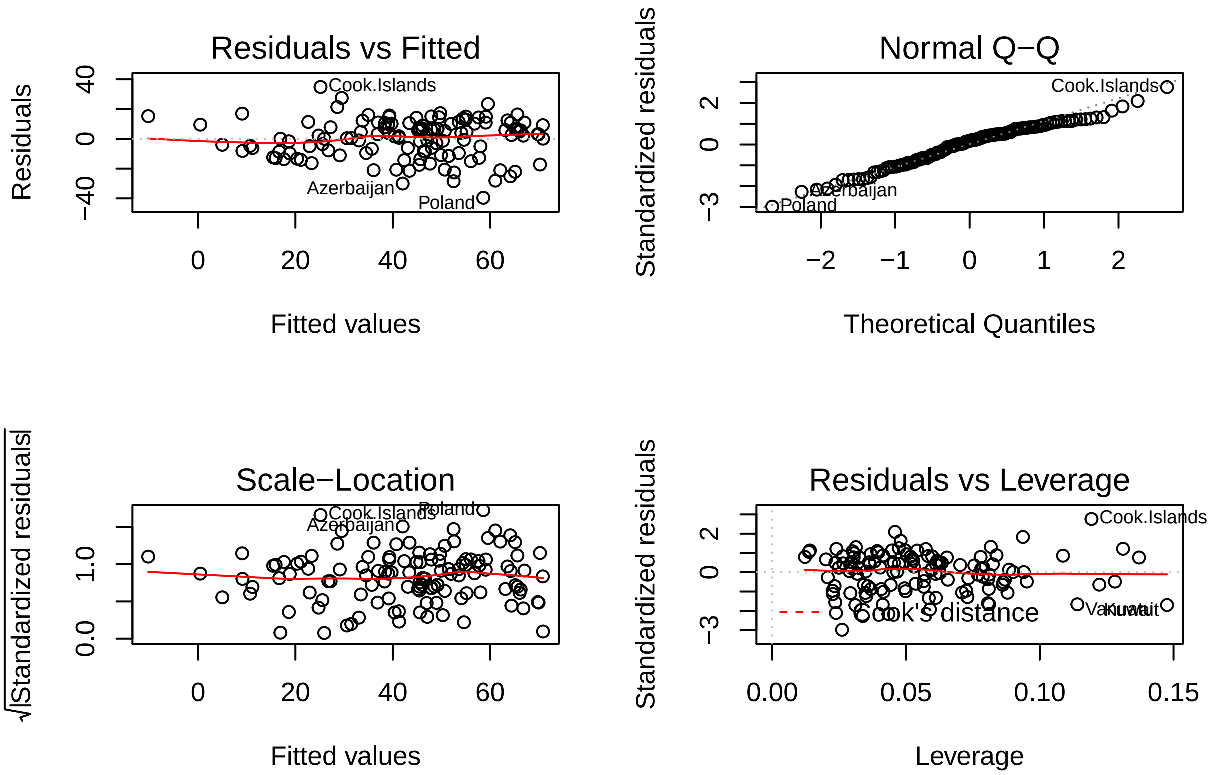
7. Given the selected transformations of the predictors, select a transformation of the response using `MASS::boxcox` or `car::boxCox` and justify.

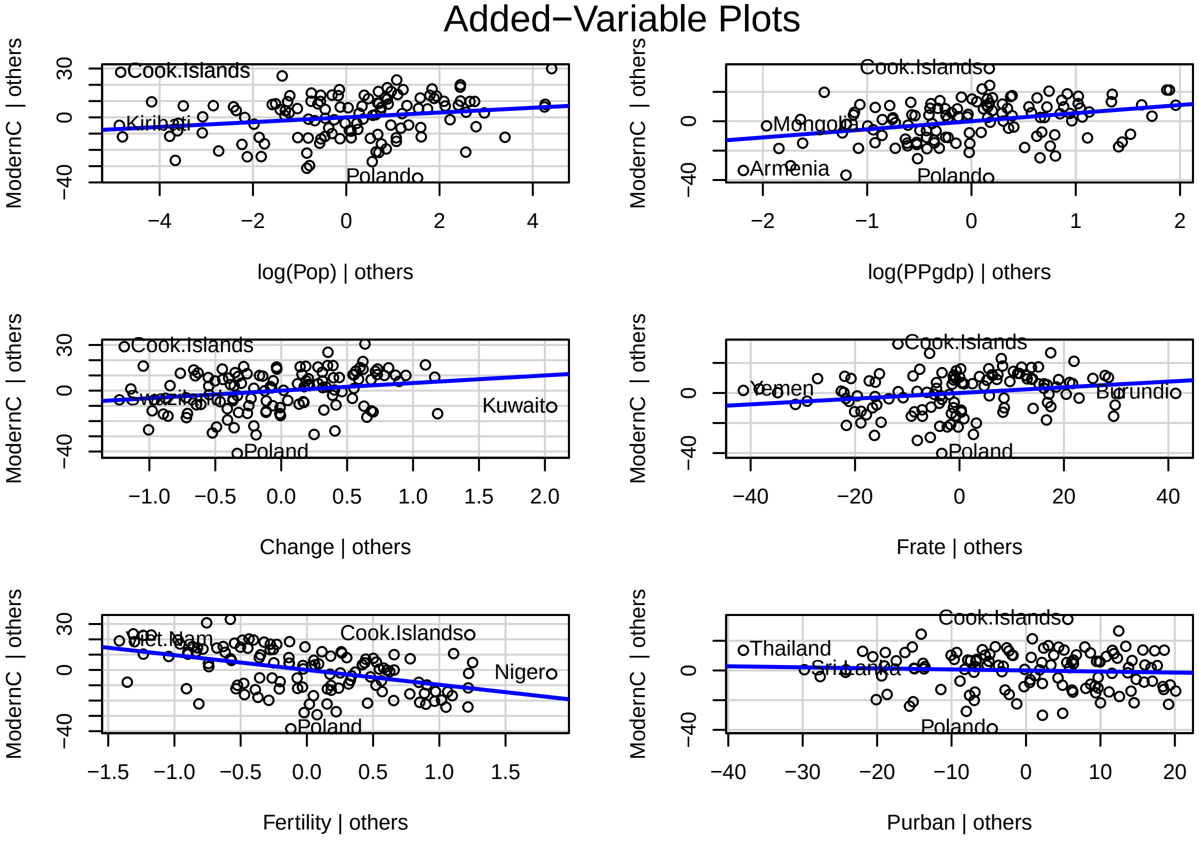


Answer: The Boxcox plot suggests there might be some transformation changing ModernC's power to some number close to 0.8. However, for simplicity and interpretation, there may not be any transformation of Y required since 1 is also in the range for available λ.

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.

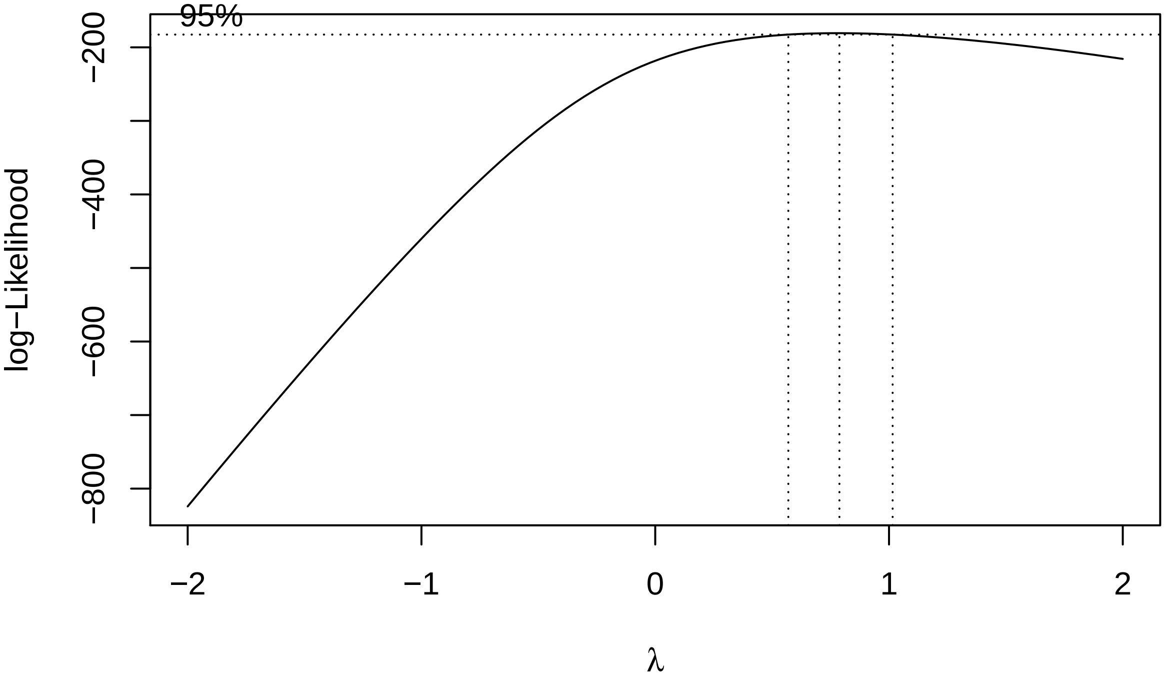
```
##
## Call:
## lm(formula = ModernC ~ log(Pop) + log(PPgdp) + Change + Frate +
##     Fertility + Purban, data = UN.nna)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -39.597  -9.540   2.238  10.024  34.840
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.11547   14.50854   0.284 0.777169
## log(Pop)     1.47207    0.62875   2.341 0.020897 *
## log(PPgdp)   5.50728    1.40505   3.920 0.000149 ***
## Change       4.99296    2.07709   2.404 0.017781 *
## Frate        0.18939    0.07711   2.456 0.015500 *
## 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
```





Answer: After refitting the model with log(Pop) and log(PPgdp), there is improvement in the Scale-Location plot. There seems no potential highly influential point, after the transformation.

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?



Answer: The Boxcox suggests no transformation for ModernC since again, 1 is in the interval. Therefore, the following steps for boxTidwell will be the same as before (Question 6 to Q8). Thus there is no difference between these two procedures.

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.

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

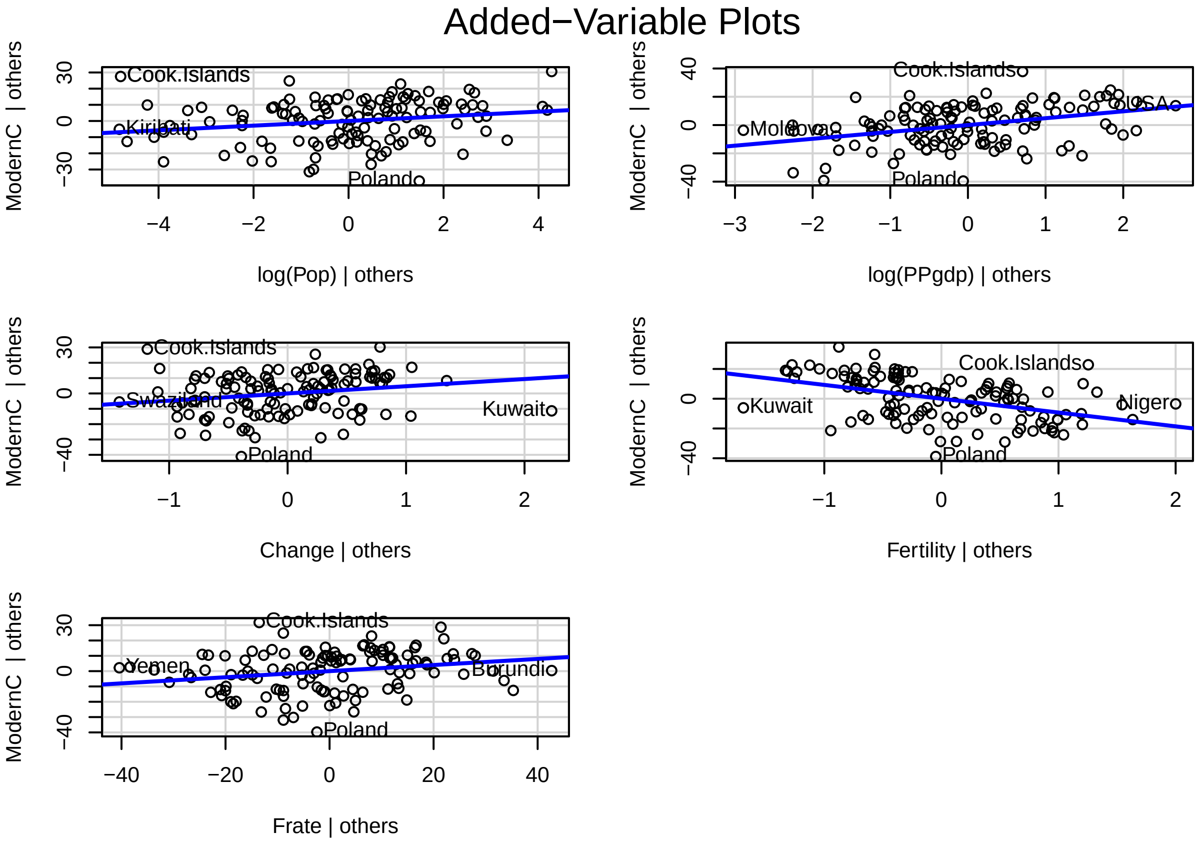
Answer: The Residuals vs Fitted, Normal Q-Q, and Scale-Location plots suggest there are three potential outliers, which are Poland, Azerbajian, and Cook Island. However, Bonferroni test suggests that there is no outlier.

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!

```
##
## Call:
## lm(formula = ModernC ~ log(Pop) + log(PPgdp) + Change + Fertility +
##     Frate, data = UN.nna)
##
```

```
## 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
## log(Pop)     1.44122    0.62606   2.302  0.02307 *
## log(PPgdp)   4.85936    1.08214   4.491 1.65e-05 ***
## Change       4.69776    2.03274   2.311  0.02255 *
## Fertility   -9.27842    1.67499  -5.539 1.85e-07 ***
## Frate        0.19955    0.07568   2.637  0.00949 **
## ---
## 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
```



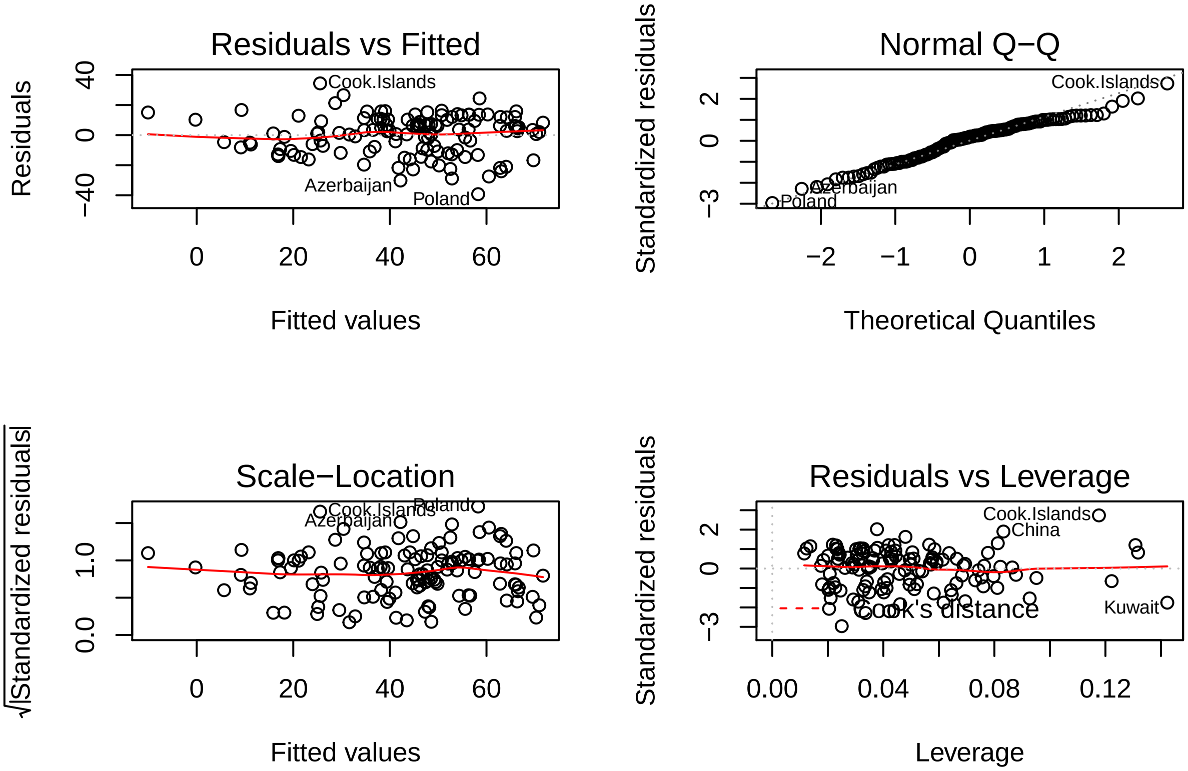
	2.5 %	97.5 %	interpretation
(Intercept)	-24.568950122334	32.7731131377968	The base value of ModernC without any predictor
log(Pop)	0.201559572990944	2.68088948644444	10% increase will increase ModernC by 0.137%
log(PPgdp)	2.71662357046803	7.00210125983713	10% increase will increase ModernC by 0.461%
Change	0.67272868376971	8.72278538727312	1% increase will increase ModernC by 4.698%
Fertility	-12.5950755844239	-5.96176721019508	1% increase will decrease ModernC by 9.278%
Frate	0.0496976732857353	0.349394345426476	1 unit increase will increase ModernC by 0.200%

Answer: The summary suggests that the transformed model satisfies

$$ModernC = 4.102 + 1.441 log(Pop) + 4.859 log(PPgdp) + 4.698 Change - 9.278 Fertility + 0.200 Frate$$

This means, 10% increase in Pop will lead to ModernC's increase by 1.441*log1.1 percent, which is 0.137%, and 10% increase in PPgdp will lead to 0.461 (4.859*log(1.1)) percent in ModernC. 1 unit increase in Change and Fertility will increase ModernC by 4.698% and 0.200% respectively, while 1 percent increase in Frate will decrease ModernC by 9.278 percents.

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



According to Cook's distance in Residuals vs Leverage plot, there is no point with this distance over 1. Therefore, I don't think there is any influential point so no deletion, same model as in Q11.

Methodology

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

Answer:

$$\begin{aligned}
e_Y &= \hat{\beta}_0 + \hat{\beta}_1 e_{x_i} \\
\mathbf{1}_n^T e &= \mathbf{1}_n^T (I-H)Y \qquad \text{times row vector 1 on both sides} \\
&= [\mathbf{1}_n^T (I-H)]Y \\
&= 0 * Y \qquad \text{using hint} \\
&= 0 \qquad (1)
\end{aligned}$$

$$e_Y = (I-H)Y = \hat{\beta}_0 + \hat{\beta}_1 \underbrace{e_{x_i}}_{(I-H)X_i} \qquad (2)$$

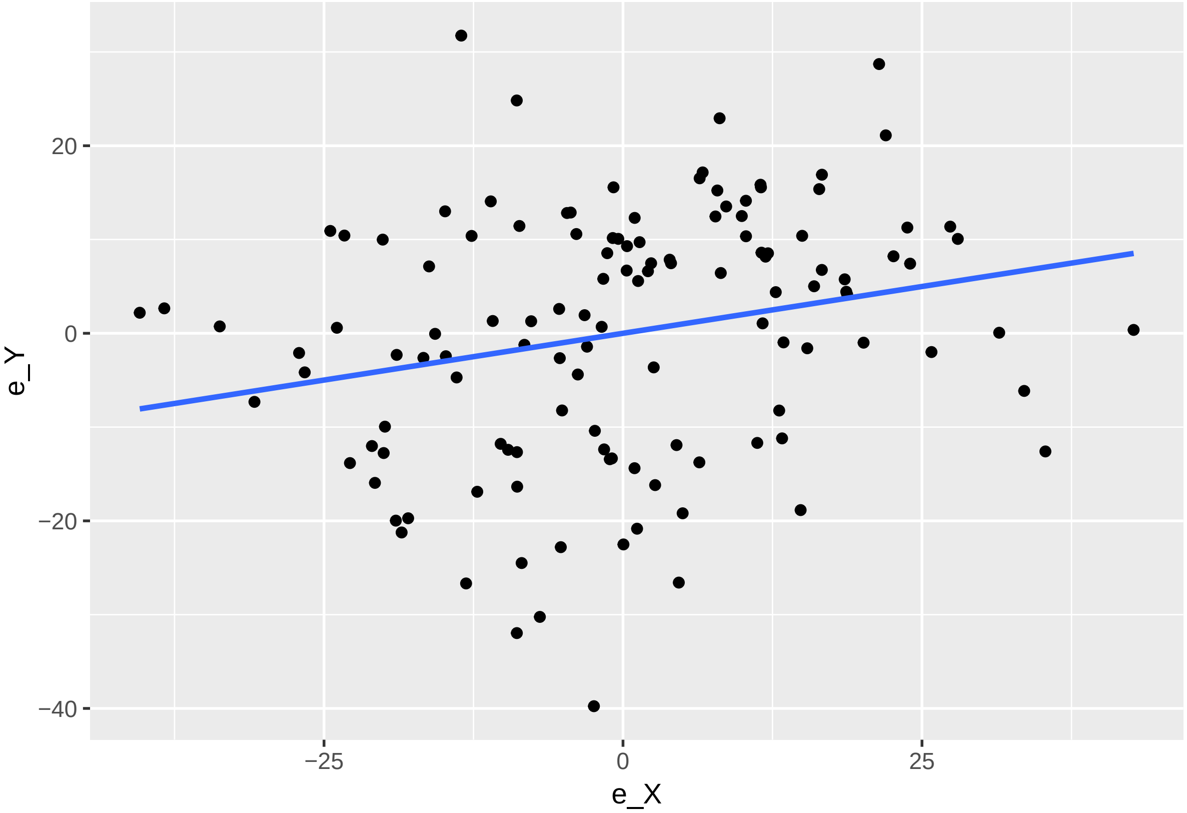
$$\begin{aligned}
(1) \implies 0 &= \mathbf{1}_n^T (I-H)Y \\
&= \mathbf{1}_n^T [\hat{\beta}_0 + \hat{\beta}_1 e_{x_i}] \qquad \text{by (2)} \\
&= \mathbf{1}_n^T \hat{\beta}_0 + \mathbf{1}_n^T \hat{\beta}_1 (I-H) x_i \\
&= \mathbf{1}_n^T \hat{\beta}_0 + \hat{\beta}_1 [\mathbf{1}_n^T (I-H)] x_i
\end{aligned}$$

$$\implies 0 = \mathbf{1}_n^T \hat{\beta}_0 + 0 \qquad \text{by hint}$$

$$\implies \mathbf{1}_n^T \hat{\beta}_0 = 0 \implies \hat{\beta}_0 = 0$$

Therefore, the intercept of avplots are always 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.



```
##    Frate
## 0.199546
```

	Estimate	t value
Original	0.199546	2.636806
Partial	0.199546	2.680756

Two regressions give the same coefficients, though tiny different t-values, which may come from the change in degree of freedom.