HW2 STA521 Fall18

Lingyun Shao, ls362, Stveshawn

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(UN3) #summary
```

```
##     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
```

```
(n=nrow(UN3)) #sample size
```

```
## [1] 210
```

```
(p=ncol(UN3)) #number of variables
```

```
## [1] 7
```

```
colSums(is.na(UN3)) #number of missing values for each variable
```

```
##   ModernC    Change     PPgdp     Frate       Pop Fertility    Purban 
##        58         1         9        43         2        10         0
```

```
(cl = sapply(UN3, class)) #check the class of each variable
```

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

```
sapply(UN3,FUN = function(x) {unique(x) %>% length()}) #check number of unique values
```

```
##   ModernC    Change     PPgdp     Frate       Pop Fertility    Purban 
##        74       160       196        63       207       154        86
```

There are 210 samples and 7 variables in this dataset. As the summary indicates, 6 variables (**MordernC**, **Change**, **PPgdp**, **Frate**, **Pop**, **Fertility**) have NA's and only one variable (**Purban**) does not have any

NA.

According to the summary, there are 3 variables (**Pop**, **PPgdp**, **Fertility**) whose means are much larger than their medians. That's a sign of positive skewness and we might need something like log transformation in follow-up model building. The classes of each variable are either "integer" or "numeric". Besides, the numbers of unique values for each variable are large. Their names do not suggest any information of being qualitative either. Therefore, I think all 7 variables should be quantitative.

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

```
UN3.stat = cbind(
  apply(UN3, 2, FUN=function(x){mean(x, na.rm=TRUE)}),
  apply(UN3, 2, FUN=function(x){sd(x, na.rm=TRUE)}))
colnames(UN3.stat)=c('Mean','Std')
#the results are rounded to 2 digits
UN3.stat %>%
  kable(format = 'latex',
        caption = 'Mean and Standard Deviation of Quantitative Varaibles',
        digits = 2) %>%
  kable_styling(latex_options = c('striped', 'hold_position'), font_size=12)
```

Table 1: Mean and Standard Deviation of Quantitative Varaibles

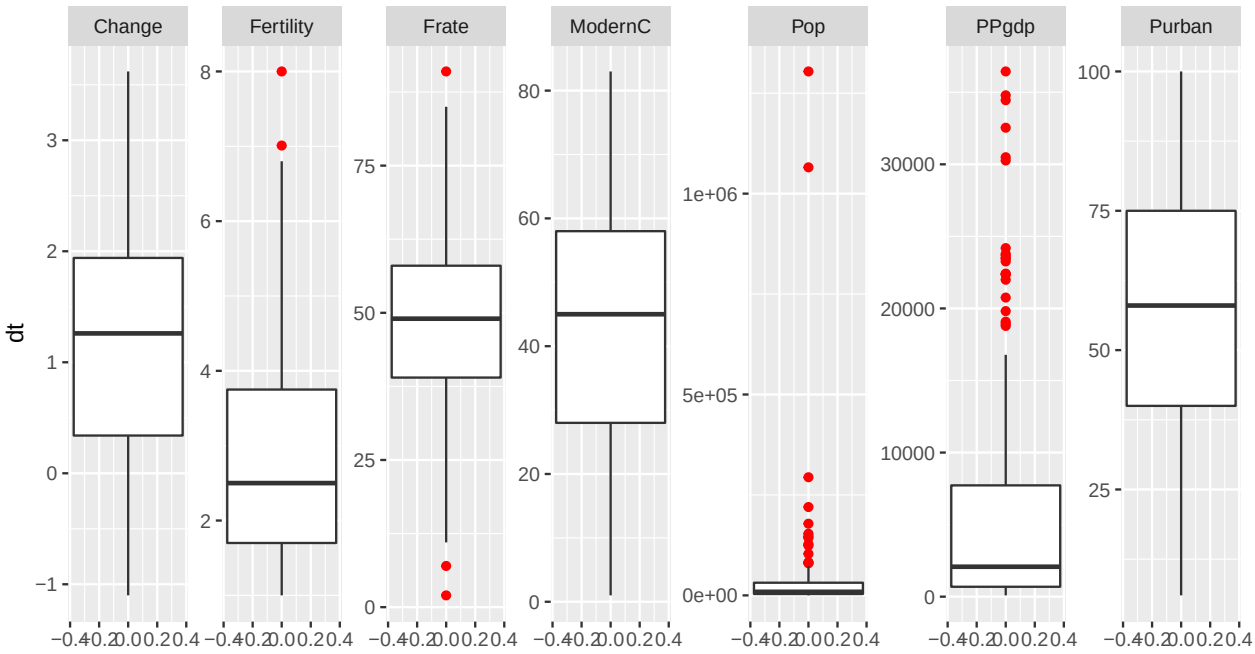
	Mean	Std
ModernC	38.72	22.64
Change	1.42	1.13
PPgdp	6527.39	9325.19
Frate	48.31	16.53
Pop	30281.87	120676.69
Fertility	3.21	1.71
Purban	56.20	24.11

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?

```
#deleting all cases containing NA's
UN3x = UN3 %>% na.omit()

#create dataframe for boxplots
UN3.byname = data.frame(
  dt = unlist(UN3x),
  name = rep(names(UN3x), each = nrow(UN3x))
)
```

```r
#boxplots for each variable
p1 = ggplot(UN3.byname, aes(y = dt)) +
    geom_boxplot(outlier.color = 2) +
    facet_wrap(name ~ ., scales = 'free', ncol = 7) +
    labs(title = '(a) Boxplots for Each Variable') +
    theme(plot.title = element_text(hjust = 0.5))

#ggpairs for original data
p2 = ggpairs(UN3x, columns = c(2:7,1),
             lower = list(continuous = wrap("smooth", method = "lm", size = 0.2))) +
     theme_bw() +
     labs(title = '(b) Scatterplot Matrices') +
     theme(plot.title = element_text(hjust = 0.5))

#log-transformation
UN3.log = UN3x %>%
  mutate(PPgdp.log = log(PPgdp)) %>%
  mutate(Pop.log = log(Pop)) %>%
  mutate(Fertility.log = log(Fertility)) %>%
  mutate(PPgdp = NULL) %>%
  mutate(Pop = NULL) %>%
  mutate(Fertility = NULL)
#ggpairs for log-transformed data
p3 = ggpairs(UN3.log, columns = c(2,5,3,6,7,4,1),
             lower = list(continuous = wrap("smooth", method = "lm", size = 0.2))) +
     theme_bw() +
     labs(title = '(c) Scatterplot Matrices after Log-transformation') +
     theme(plot.title = element_text(hjust = 0.5))

#scatterplot for 3 highly correlated predictors
p4 = ggplot(data = UN3x, aes(x = Fertility, y = PPgdp))+
     geom_point(aes(color = Purban, size = ModernC)) +
     scale_size_continuous(range = c(1,15)) +
    labs(title = '(d) Scatter Plot of Highly-correlated Predictors') +
    theme(plot.title = element_text(hjust = 0.5))

p1 #boxplots for each variable
```

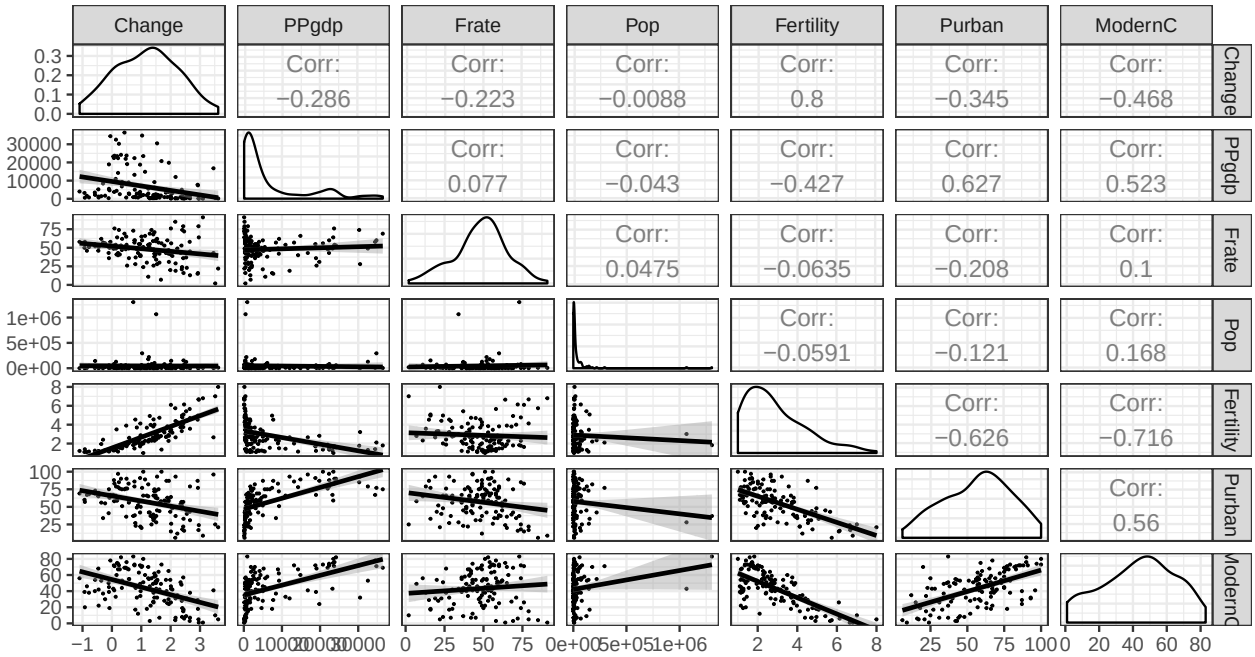
(a) Boxplots for Each Variable



First, I created boxplots for each variable to detect the presence of potential outliers. As is shown in plot (a), there are potential outliers colored red in **Fertility**, **Frate**, **Pop** and **PPgdp**. I metioned, in the summary part, that **Pop** and **PPgdp** are heavily positive-skewed and **Fertility** is also slightly positive-skewed. These boxplots displayed their skewness. Besides, variables like 'population', 'gdp' and 'income' are often log-transformed in practice researches. Therefore, I tried log-transforming these 3 predictors and did some relevant analyses.

```
p2 #ggpairs for original data
```

(b) Scatterplot Matrices

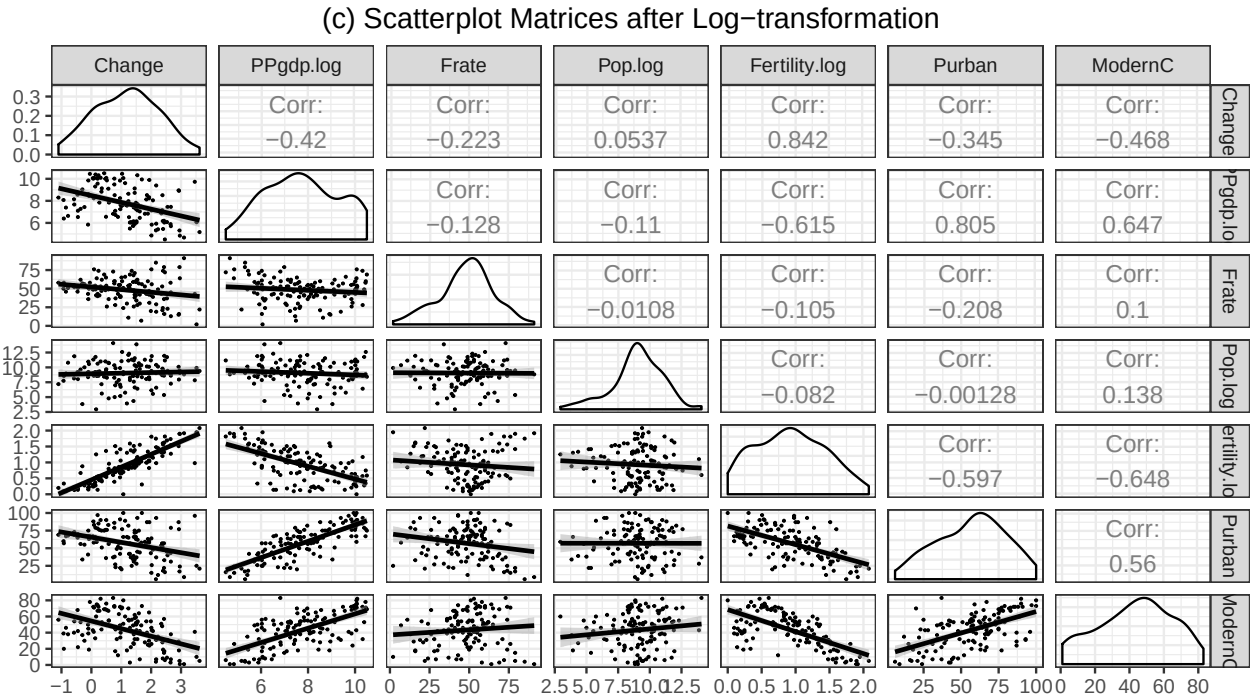


Next, I created scatterplot matrices for original data using ggpairs. In plot (b), the kernal density curves suggest, as mentioned before, there is heavy positive skewness in **PPgdp** and **Pop** and slight positive skewness

in **Fertility**. We can see that, in the scatterplots regarding predictor **Pop**, two points are very far away from the others and they might be outliers. The fitted lines show that these two points have ridiculously huge impact on the standard error.

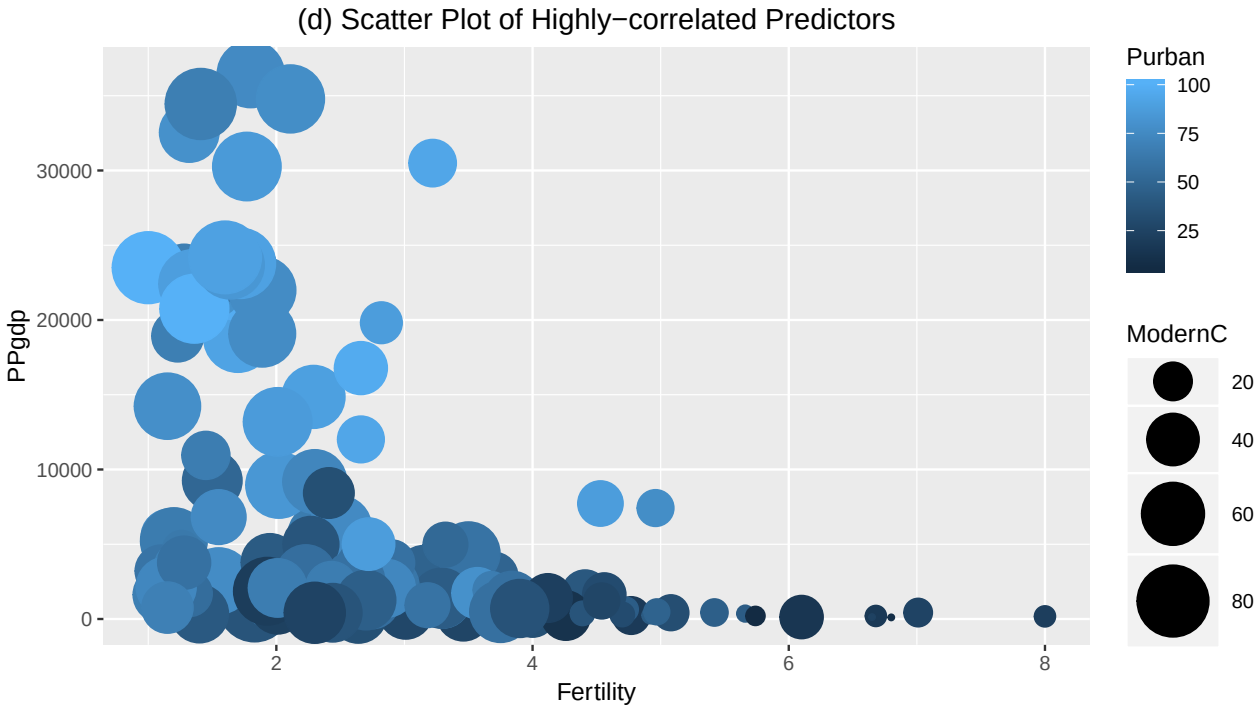
In terms of predicting **ModernC**, only the scatterplots regarding **Fertility**, **Purban** seemed to suggest linear relationships. Scatterplots regarding **Change**, **PPgdp** displayed non-linear patterns. Scatterplots regarding **Frate** and **Pop** has not obvious relationships. There are also some plots suggesting non-linear relationships between predictors, such as **PPgdp-Fertility** and **PPgdp-Purban**.

```
p3 #ggpairs for log-transformed data
```



Then I did log-transformation to **PPgdp**, **Pop** and **Fertility** to see the difference. After log-transformation, the skewness was largely reduced and the correlation between **PPgdp** and **ModernC** also increased dramatically. The previously non-linear pattern in **ModernC-PPgdp** changed into a linear one. There is still no clear relationship between **ModernC** and **Pop**, so maybe **Pop** is not a good predictor. For **Fertility**, the relationship after transformation seemed like a quadratic form, so maybe we should seek a better transformation for it.

```
p4 #scatterplot for 3 highly correlated predictors
```



Last, I chose 3 most correlated predictors with **ModernC** in plot (c) and created the scatterplot for them to display their relationships with our response variable **ModernC**. As plot (d) shows, **ModernC** tends to be large when **Fertility** is small, **PPgdp** is large and **Purban** is 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?

```
UN3.lm = lm(data = UN3, ModernC ~ .)
summary(UN3.lm)
```

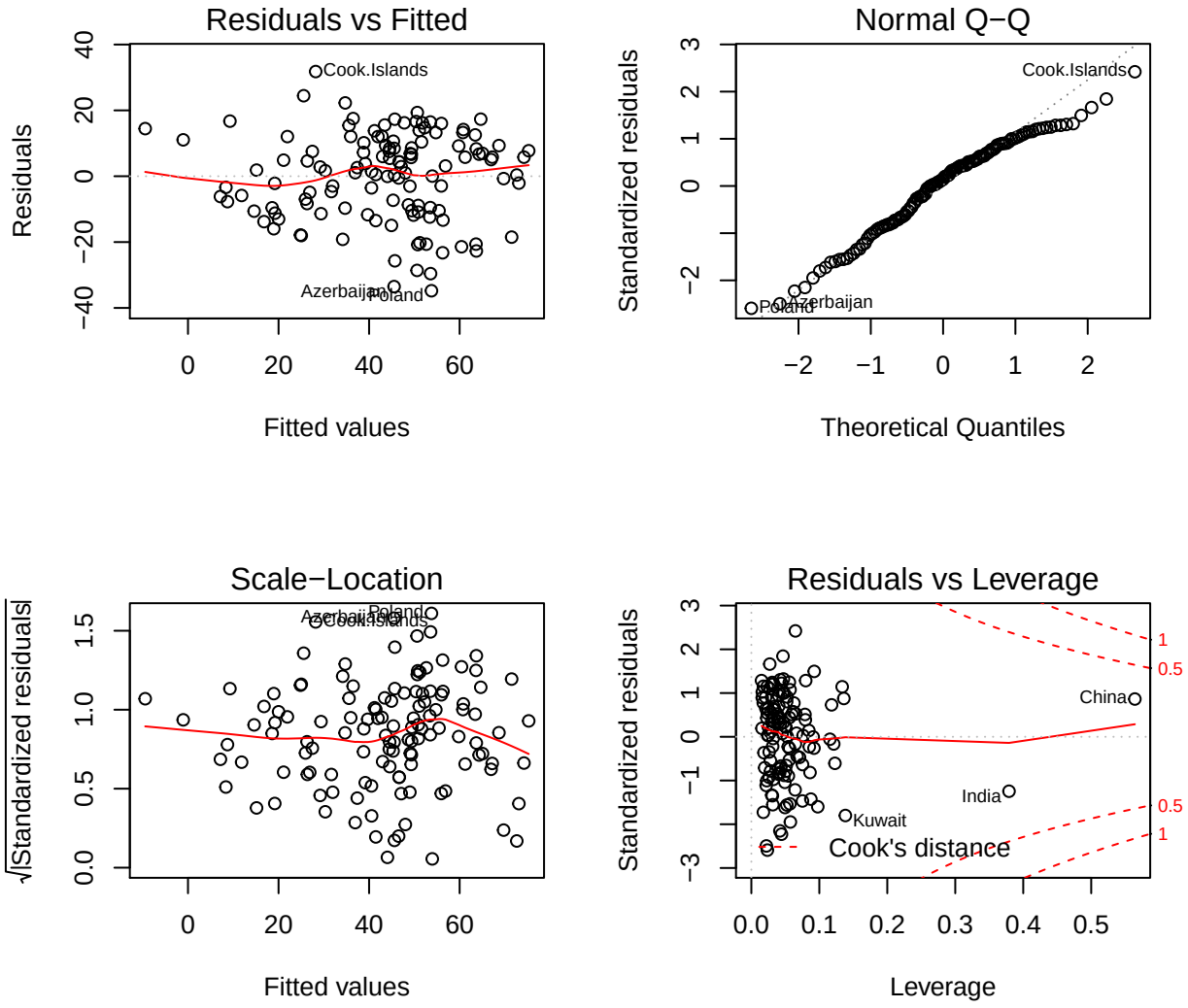
```
##
## Call:
## lm(formula = ModernC ~ ., data = UN3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -34.781  -9.698   1.858   9.327  31.791
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.529e+01  9.467e+00   5.841 4.69e-08 ***
## Change       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
```

First, according to the summary of linear regression, we found that **Frate** and **Purban** are not significant at the level of 0.05 in the t-test for coefficients, so possibly they are not very effective in explaining the response. The adjusted R-squared is 0.5989, meaning that the model can explain about 59.89% of the total variance in the response variable. The result of F-test indicates the effectiveness of the model as a whole (at least one non-zero coefficient).

85 observations are deleted due to missingness, so 125 observations are used in the model fitting.

```
par(mfrow=c(2,2))
plot(UN3.lm)
```



As the **Residuals vs Fitted** plot shows, the residuals are evenly distributed around zero. Also the residuals do not suggest an obvious sign of fan-like shape or non-linear curve. In general, I think there is no obvious clue of heteroscedasticity.

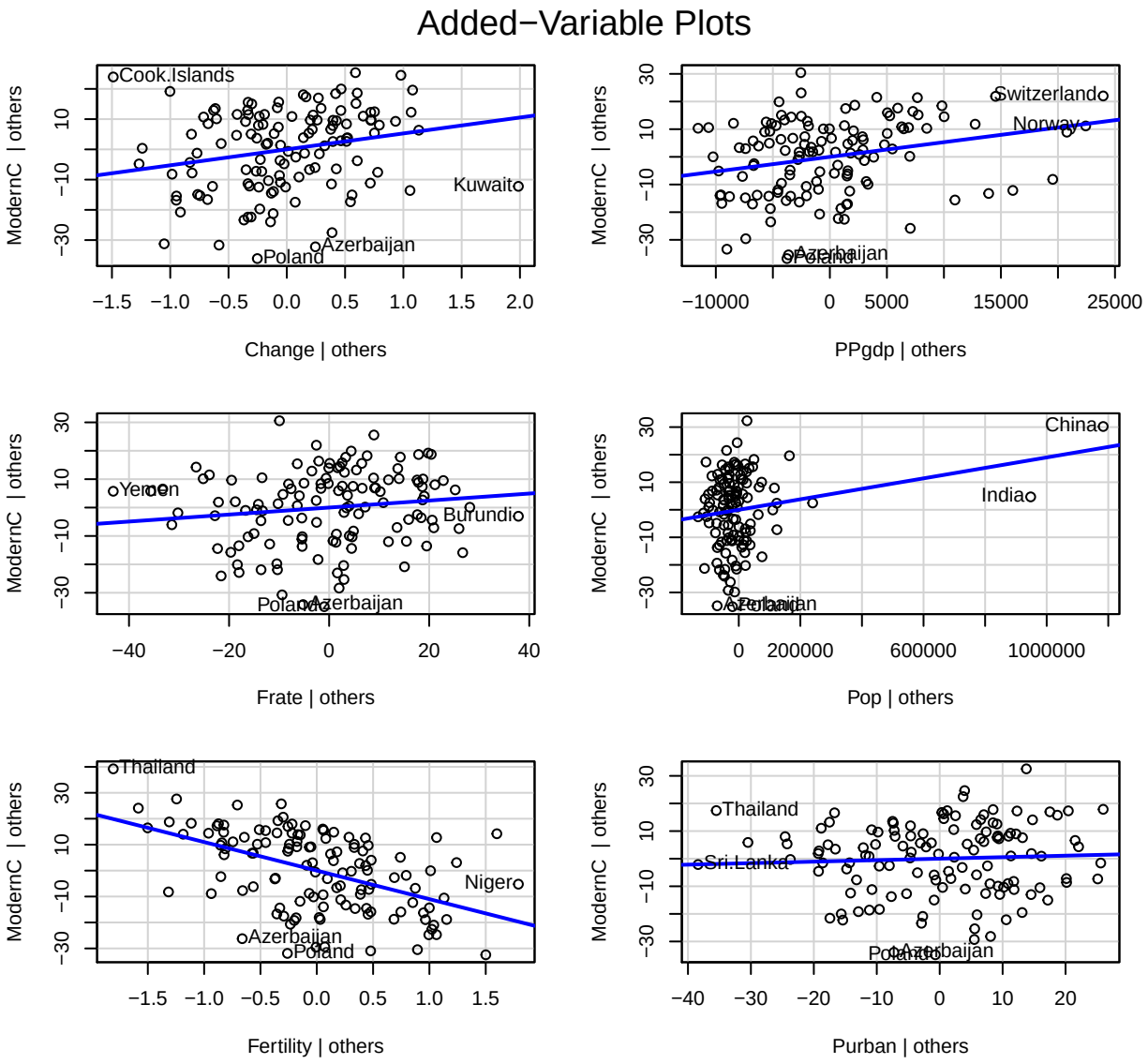
Normal Q-Q plot suggests a little departure from normal distribution in large quantiles, and the shape indicates a potential left-skewness of the residual's distribution. But in general the fitting of quantiles is not too bad. Thus, the assumption of normality may still be considered to hold.

The **Scale-Location** plot shows that the residuals change a little with the increase of fitted values, but I think the shift is not enough to claim the presence of heteroscedasticity. Therefore, this plot also tells us something good regarding linear model assumptions.

The **Residuals vs Leverage** plot indicates that there is no obvious outlier with a Cook's distance larger than 0.5. Yet we need to pay attention to the case China and India since they have large leverage and tend to have great impact. Besides, Kuwait has a comparatively large Cook's distance.

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(UN3.lm)
```



As the **Added-Variable Plots** shows, the points in **ModernC-PPgdp** and **ModernC-Pop** plots are distributed very unevenly in the direction of x-axis, with most points concentrated at the small values and a few points having extremely large values, so we may need to transform **PPgdp** and **Pop** or exclude some potential outliers to make them more fitted to our assumption of normality. Besides, there are clear non-linear patterns in these two plots, so we might want to transform them and do some other analyses.

As for influential points, we can observe that on the right side of **ModernC-Change** plot, there is a point very far away from others. It is very likely that this point has pulled the fitted line down to fit itself. Besides, for the **ModernC-Pop** plot, we see two point with a huge deviation from the their centers and the fitted line has undoubtedly been affected by these two influential points. To identify these three influential localities, I manually did the filtering by calculating the values on x-axis of their plot as below.

```
Change.lm = lm(UN3, formula = Change ~ PPgdp + Frate + Pop + Fertility + Purban)
Pop.lm = lm(UN3, formula = Pop ~ PPgdp + Frate + Change + Fertility + Purban)
Change.lm %>% .$residuals %>% .[.>1.5] #filtering out influential
```

```
##   Kuwait
## 1.899074
```

```
Pop.lm %>% .$residuals %>% .[.>600000] #filtering out influential
```

```
##     China     India
## 1201931.7  971124.4
```

There are 2 potential influential localities in the term of **Pop**, which are China and India. Also there is 1 potential influential locality in the term of **Change**, which is Kuwait.

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

```
UN3.positive = UN3 %>%
  mutate(Change = Change +2) #make all predictors positive

boxTidwell(data = UN3.positive, ModernC ~ Pop+ PPgdp,
           other.x = ~ Change + Frate + Fertility + Purban)
```

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

Based on previous graphical analyses, we have already reached that **PPgdp** and **Pop** are heavily right-skewed and their added-variable plots displayed clear patterns of non-linear relationship. So I believe we need to transform these two predictors to proper forms while keeping the other predictors unchanged. Therefore, I used `boxTidwell` in R to find the optimal transformation.

First of all, we know from the summary only predictor **Change** has negative values and the minimum is -1.1. I set a start point, 2 for this predictor, then making all the predictors positive by adding 2 to each case in the term of **Change**.

Since we are only seeking for transformations of **PPgdp** and **Pop**, we would keep other predictors unchanged and put them in the parameter `other.x`.

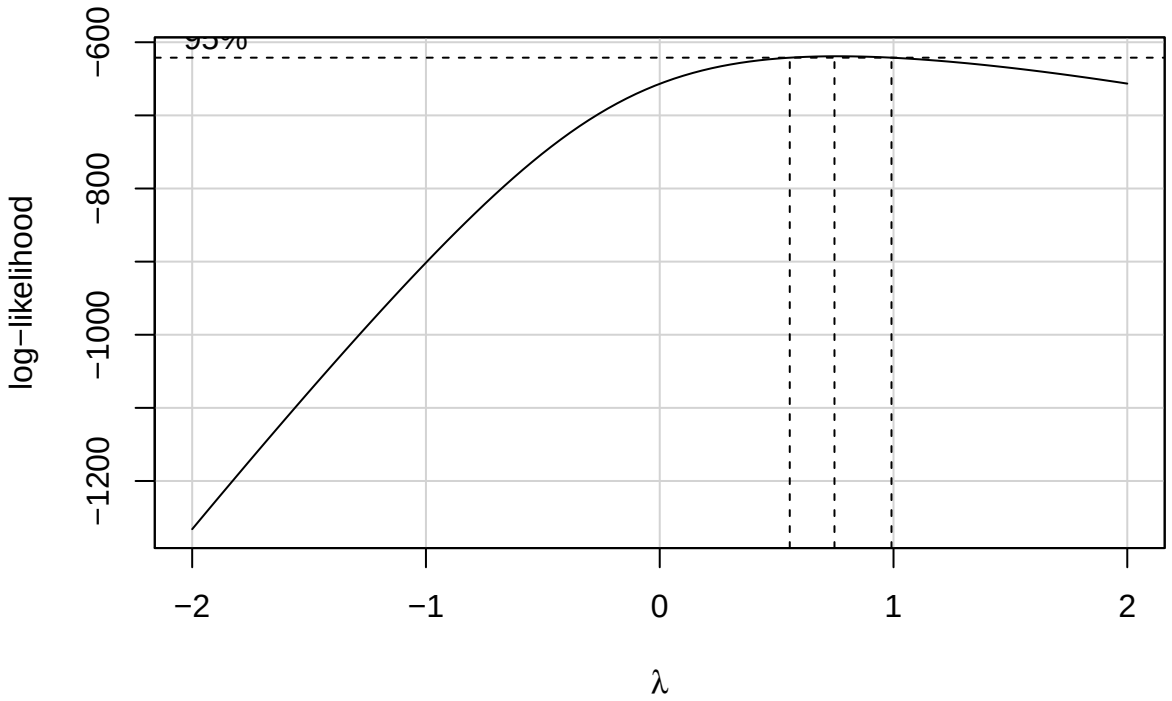
As the result of `boxTidwell` shows, MLE's of lambda are '0.41' and '-0.13'. Usually, to make the transformation have explicit practical meaning, we would choose to round $\lambda's$ to be multiples of 0.5. Therefore I chose $\lambda = 0$ for **PPgdp** and $\lambda = 0.5$ for **Pop**. **Change**, **Frate**, **Fertility** and **Purban** remained unchanged. So far our model would be

$$ModernC \sim Change + Frate + Fertility + Purban + \log(PPgdp) + \sqrt{Pop}$$

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

```
UN3.trans = UN3 %>%
  mutate(PPgdp.trans = log(PPgdp)) %>%
  mutate(Pop.trans = (Pop^(0.5)-1)/0.5)

UN3.trans.lm=lm(data=UN3.trans, ModernC ~ Change + Fertility +
                  Frate + Purban + PPgdp.trans + Pop.trans)
```

```
boxCox(UN3.trans.lm)
```



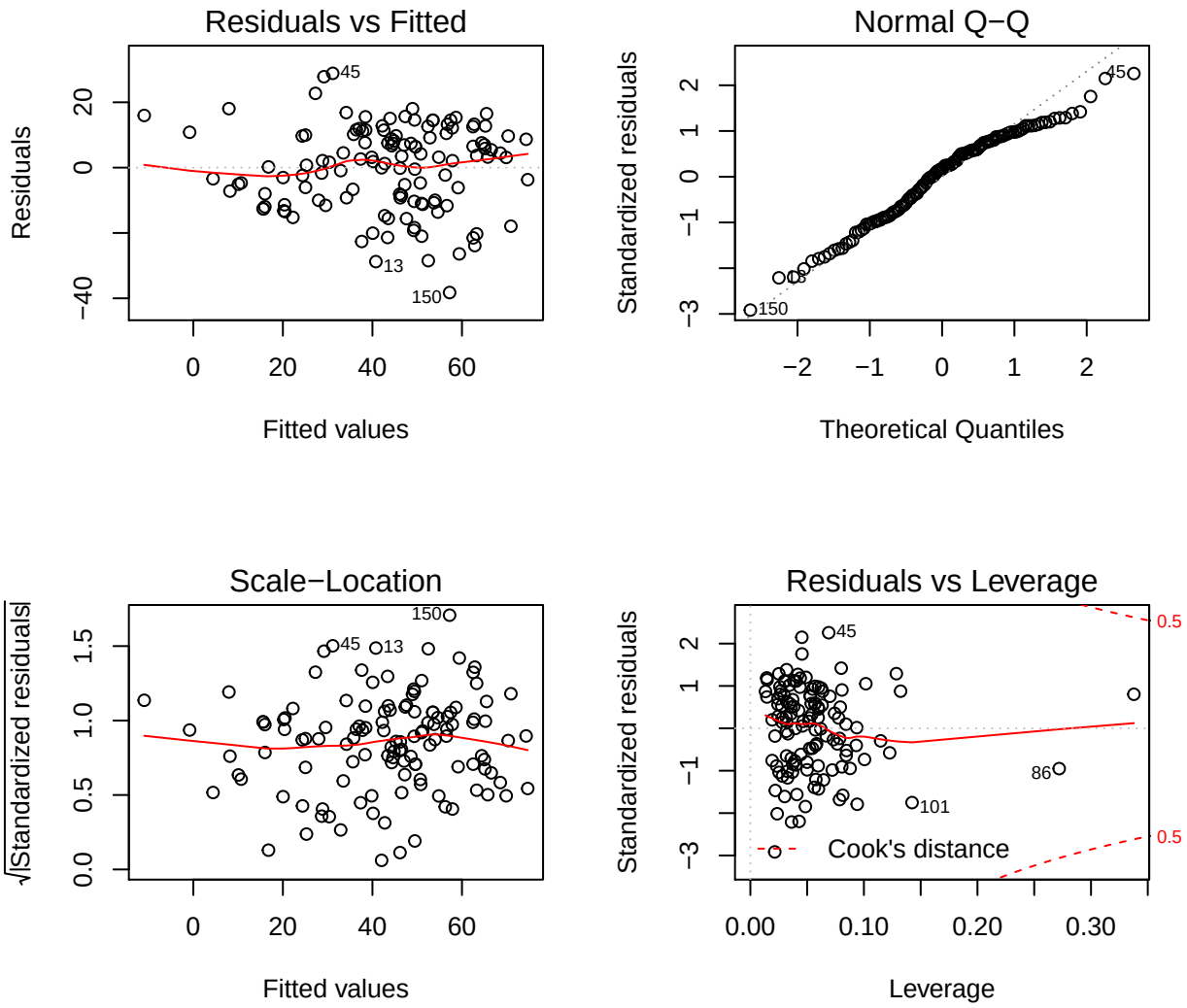
Since the CI of λ is very close to 1, to make the response variable have explicit practical meaning, I chose $\lambda = 1$ and kept the response variable unchanged.

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.

Since we did not do any transformation for response variable, our model would still be

$$ModernC \sim Change + Frate + Fertility + Purban + \log(PPgdp) + \sqrt{Pop}$$

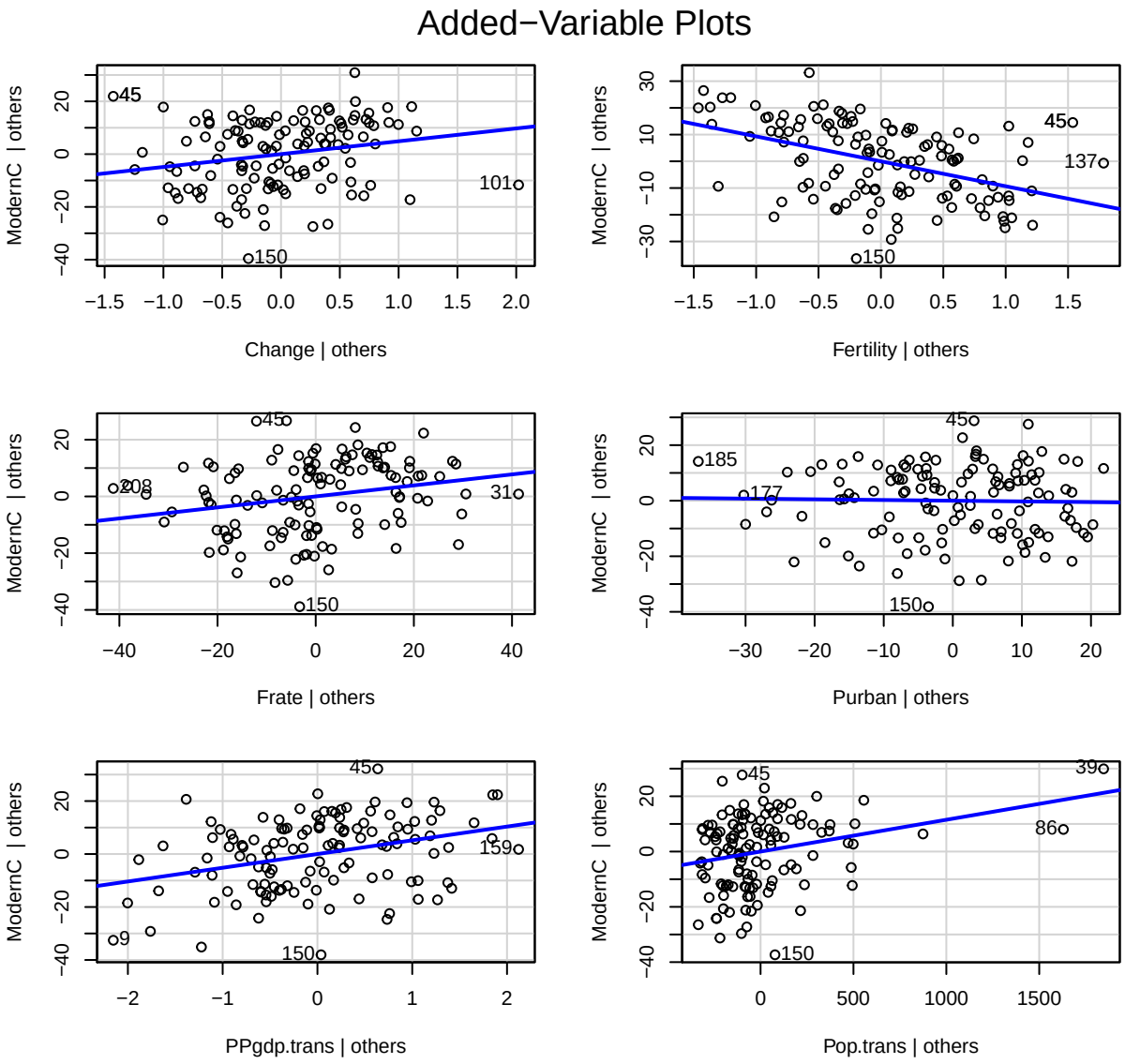
```
par(mfrow=c(2,2))
plot(UN3.trans.lm)
```



Residuals are evenly distributed around 0 and no obvious sign of fan-like or non-linear pattern is observed in **Residuals vs Fitted** plot. The previous deviation in **Normal Q-Q** plot is slightly improved and the residuals can be considered as from normal distribution. The **Scale-Location** plot shows that the square-root of standardized residuals are almost unvarying at the value of 1, which indicates homoscedasticity. There is no outlier that has a Cook's distance larger than 0.5 judging from the **Residuals vs Leverage** plot.

Generally speaking, the assumptions for linear models all seem valid in this model.

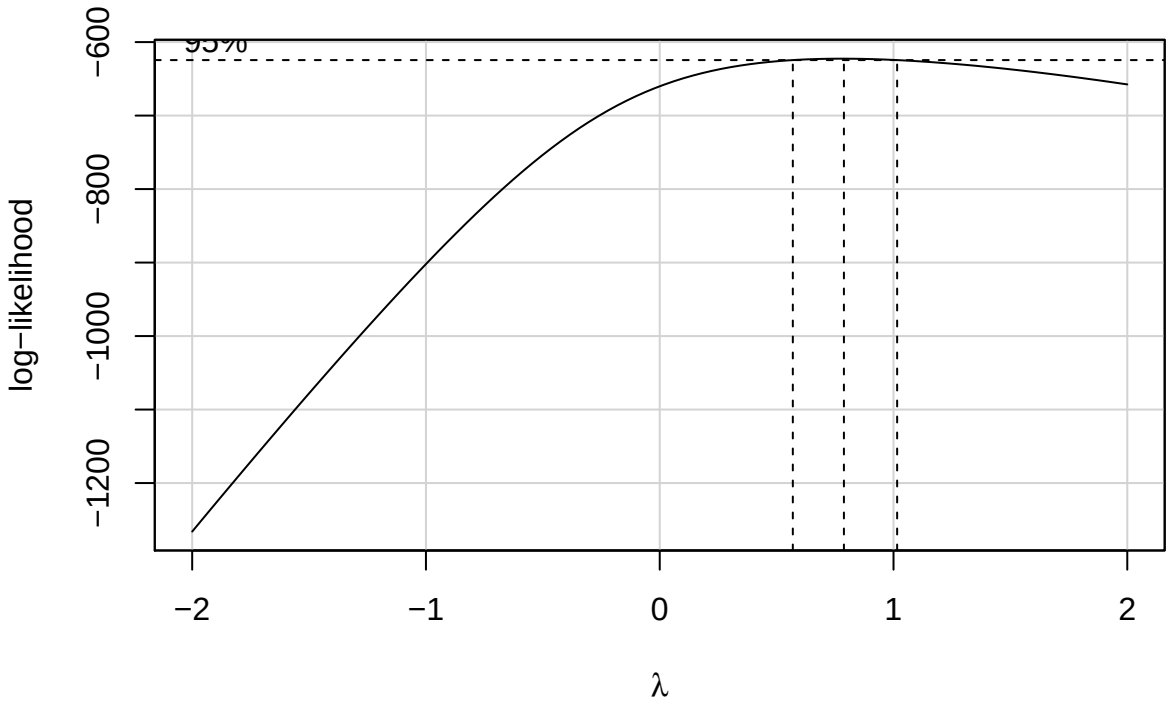
```
avPlots(UN3.trans.lm)
```



From the added-variable plots, we can find that the previous non-linear patterns of **ModernC-PPgdp** and **ModernC-Pop** are improved and changed to linear patterns. The change is particularly significant for the log-transformation of **PPgdp**. **ModernC-Pop** plot still has 2 strong influential points which may affect the result of our model. There is still a influential point on the right of **ModernC-Change**. The other plots look good, with points evenly distributed and showing linear patterns.

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(UN3.lm)
```




```
boxTidwell(data = UN3.positive, ModernC ~ Pop+ PPgdp,
           other.x = ~ Change+Frate+Fertility+Purban)
```

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

I started with using boxcox for my predictors and still get $\lambda = 1$ for response variable. Then I used `boxTidwell` and get the same result. Still, $\lambda_{Pop} \approx 0.5$, $\lambda_{PPgdp} \approx 0$. Our model is still

$$ModernC \sim Change + Frate + Fertility + Purban + \log(PPgdp) + \sqrt{Pop}$$

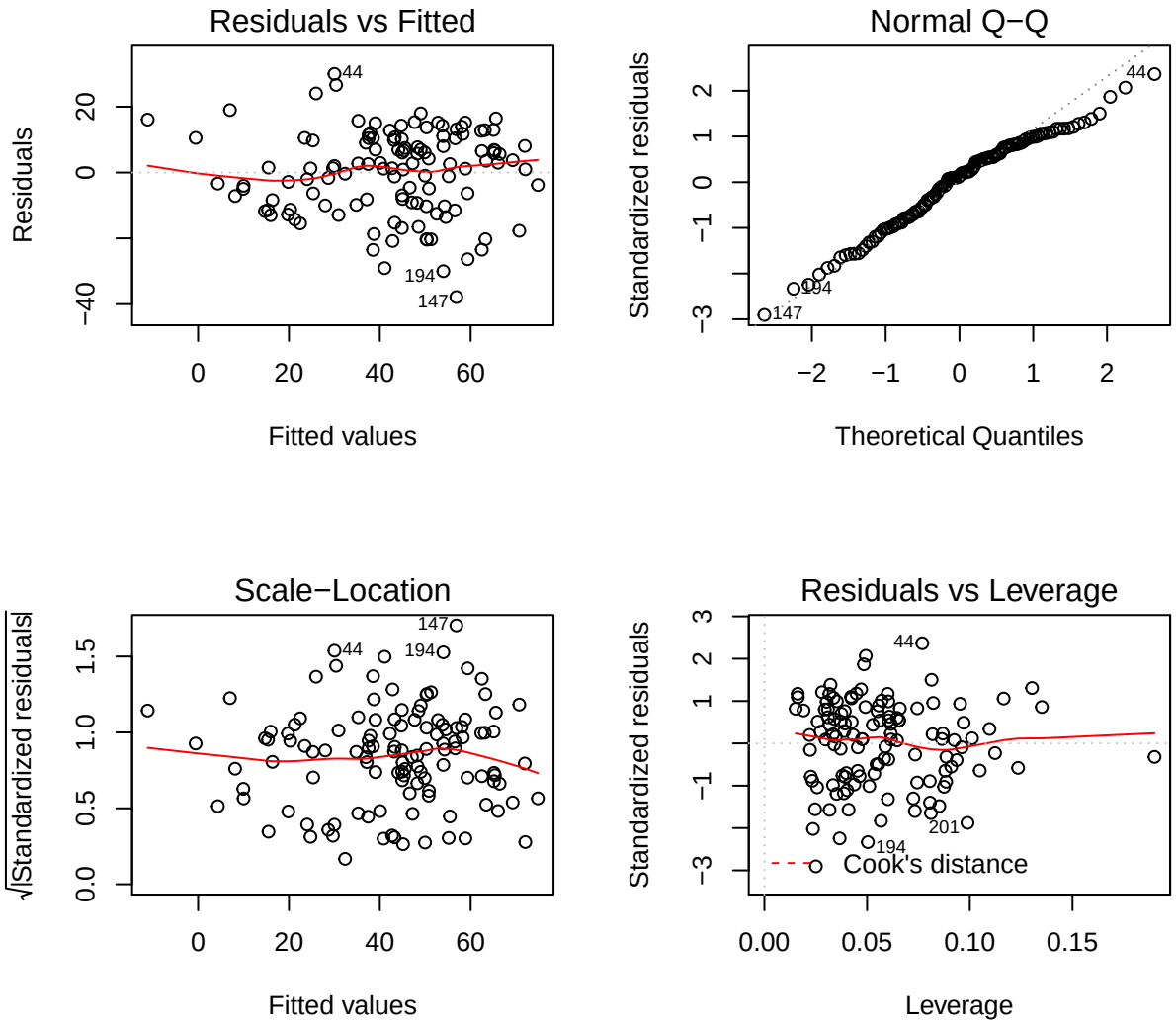
By change the order of finding transformation of predictors and response, I ended up with the same model as in 8.

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.

As previously discussed, there are 2 influential points (China and India) with very large leverage in the term of **Pop** and 1 influential point (Kuwait) in **Change**, and they might lead to large deviation of our model. I excluded these cases and tried refitting the model.

```
#exclude 3 influential points
UN3.trans2 = UN3.trans %>%
  mutate(name = rownames(UN3)) %>%
  filter(!(name %in% c('China','India','Kuwait')))

UN3.trans2.lm=lm(data=UN3.trans2, ModernC ~ Change + Fertility
              + Frate + Purban + PPgdp.trans + Pop.trans)
```

```
par(mfrow=c(2,2))
plot(UN3.trans2.lm)
```



After removing the influential points, the first 3 residual plots did not show much difference and still showed validity of linear model's assumptions. After removing those points with large leverages, those remaining cases apparently have smaller Cook's distances.

```
summary(UN3.trans2.lm)
```

```
##
## Call:
## lm(formula = ModernC ~ Change + Fertility + Frate + Purban +
##     PPgdp.trans + Pop.trans, data = UN3.trans2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -37.847  -9.671   1.747  10.340  29.980
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) 12.895604  12.244450   1.053 0.294467
## Change       6.038124   2.145641   2.814 0.005755 **
## Fertility   -9.881136   1.771014  -5.579 1.63e-07 ***
## Frate        0.195401   0.076951   2.539 0.012443 *
## Purban      -0.017363   0.097009  -0.179 0.858265
## PPgdp.trans  5.208602   1.353807   3.847 0.000196 ***
## Pop.trans    0.010786   0.005737   1.880 0.062613 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.19 on 115 degrees of freedom
##   (85 observations deleted due to missingness)
## Multiple R-squared:  0.6389, Adjusted R-squared:  0.6201
## F-statistic: 33.91 on 6 and 115 DF,  p-value: < 2.2e-16
```

After removing influential points, the model's Adjusted Squared-R surely increased since these points are not well fitted in the previous model. I notice that the P-value for **Pop.trans** has increased from 0.003 to 0.06, which means that this predictor becomes more likely to be 0. So it becomes more likely that this predictor does not have a significant predicting effect. I think the explanation for this might be that **Pop** itself is not a good predictor in the first place judging from its correlation with our response. Only because of the presence of these two extreme points, there is some mistaken relationship. The P-value for **Change** has decreased, showing that the deletion does bring improvement.

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!

```
Coef = UN3.trans2.lm %>% #summary
  summary() %>%
  .$coefficients
Conf = confint(UN3.trans2.lm) #CI
Summary = cbind(Coef, Conf)
Summary %>%
  kable(format = 'latex',
        caption = 'Summary for Each Coefficient',
        digits = 3) %>%
  kable_styling(latex_options = c('striped', 'hold_position'), font_size=12)
```

Table 2: Summary for Each Coefficient

	Estimate	Std. Error	t value	Pr(>|t|)	2.5 %	97.5 %
(Intercept)	12.896	12.244	1.053	0.294	-11.358	37.150
Change	6.038	2.146	2.814	0.006	1.788	10.288
Fertility	-9.881	1.771	-5.579	0.000	-13.389	-6.373
Frate	0.195	0.077	2.539	0.012	0.043	0.348
Purban	-0.017	0.097	-0.179	0.858	-0.210	0.175
PPgdp.trans	5.209	1.354	3.847	0.000	2.527	7.890
Pop.trans	0.011	0.006	1.880	0.063	-0.001	0.022

For **Intercept**, the estimated value is 12.896, which means, **ModernC** should be 12.896 when all predictors are 0. Its P-value is 0.294. So we can't reject the H_0 : *The intercept is* 0. The 95% CI is (-11.358, 37.150), which is basically because it has a large standard error.

For **Change**, the estimated coefficient is 6.038, which means, with all other predictors unchanged, 1 unit increase in **Change** will result in 6.038 units of increase in **ModernC**. This coefficient is significantly non-zero. The 95% CI is (1.788, 10.288).

For **Fertility**, the estimated coefficient is -9.881, which means, with all other predictors unchanged, 1 unit increase in **Fertility** will result in 9.881 units of decrease in **ModernC**. This coefficient is significantly non-zero. The 95% CI is (-13.389, -6.373).

For **Frate**, the estimated coefficient is 0.195, which means, with all other predictors unchanged, 1 unit increase in **Frate** will result in 0.195 unit of increase in **ModernC**. This coefficient is significantly non-zero. The 95% CI is (0.043, 0.348).

For **Purban**, the estimated coefficient is -0.017, which means, with all other predictors unchanged, 1 unit increase in **Purban** will result in 0.017 unit of decrease in **ModernC**. We can not reject the H_0 that this coefficient is zero. The 95% CI is (-0.210, 0.175).

For **PPgdp**, the estimated coefficient is 5.209. Since we have log-transformed it, so 10% increase in **PPgdp** will result in $\beta_{PPgdp}\log(1.1) \approx 0.497$ unit of increase in **ModernC**. This coefficient is significantly non-zero. The 95% CI of the incease in **ModernC** regarding a increase of 10% in **PPgdp** is $\log(1.1) * CI = (0.241, 0.752)$.

For **Pop**, the estimated coefficient is 0.011. Since we have transformed its predictor in to squar-root, so 1 increase in **PPgdp** will result in 0.011 unit of increase in **ModernC** and every k units of increase in **PPgdp** will result in $0.011 * \sqrt{k}$ unit of increase in **ModernC**. We can not reject the H_0 that this coefficient is zero at 0.05 level of significance. The 95% CI of the increase in **ModernC** regarding a increase of $\sqrt{k}$ units in **Pop** is $\sqrt{k} * CI = \sqrt{k}(-0.001, 0.022)$.

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

Our final model to predict **ModernC** from some relevant predictors is

$$ModernC = 12.896 + 6.038Change - 9.881Fertility + 0.195Frate - 0.017Purban + 5.209\log(PPgdp) + 0.011\sqrt{Pop}$$

This model can explain about 62% variance in **ModernC**. We did some transformations to **PPgdp** and **Pop** to alleviate the huge imbalance in these variables to make them fitted to linear model assumptions. 85 cases are deleted due to missingness. 'China' and 'India' are deleted because their extremely large values of **Pop** exert too strong effects on contribution of prediction in this predictor, which will dramatically decrease the robustness of our model. 'Kuwait' is deleted because it has extremely large **Change**, which might also affect our model's robustness.

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.*

Proof 1:

For linear models taking intercepts into consideration, we can obtain the OLS estimator by minimizing the loss $(Y - X\beta)^{\mathrm{T}}(Y - X\beta)$ w.r.t β. Taking the first derivative of the loss w.r.t β and setting it to 0, we have

$$\frac{d(Y-X\beta)^{\mathrm{T}}(Y-X\beta)}{d\beta} = X^{\mathrm{T}}(Y-X\beta)$$
$$X^{\mathrm{T}}(Y-X\hat{\beta}) = X^{\mathrm{T}}\hat{\mathbf{e}} = \mathbf{0}$$

If our model allows for an intercept, then there will be a column of ones in X, a row of ones in X^{T}. Therefore,

$$\mathbf{1}^{\mathrm{T}}\hat{\mathbf{e}} = 0 \tag{13.1}$$

which shows that the sample mean of residuals will always be 0 if there is an intercept.

Then we want to verify that the intercept in the added variable plot will always be 0. Say we are doing the avplot for Y and X_i, where Y is the response variable and X_i is the i-th column of our matrix X, and $X_{(i)}$ stands for the matrix after removing X_i.

Based on (13.1), we know that

$$\mathbf{1}^{\mathrm{T}}\hat{\mathbf{e}}_Y = 0,\ \mathbf{1}^{\mathrm{T}}\hat{\mathbf{e}}_{X_i} = 0$$

where $\hat{\mathbf{e}}_Y$ is the residuals of regressing Y on $X_{(i)}$ and $\hat{\mathbf{e}}_{X_i}$ is the residuals of regressing X_i on $X_{(i)}$.

For any linear model allowing for an intercept, we have

$$\mathbf{1}^{\mathrm{T}}\hat{\mathbf{e}} = \mathbf{1}^{\mathrm{T}}(I-H)Y = 0 \Rightarrow \mathbf{1}^{\mathrm{T}}Y = \mathbf{1}^{\mathrm{T}}HY$$

So the fitted line will always pass through the center of gravity (mean) of X and Y, which is $(\frac{1}{n}\mathbf{1}^{\mathrm{T}}X, \frac{1}{n}\mathbf{1}^{\mathrm{T}}Y)$ since

$$(\frac{1}{n}\mathbf{1}^{\mathrm{T}}X)\hat{\beta} = \frac{1}{n}\mathbf{1}^{\mathrm{T}}X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y = \frac{1}{n}\mathbf{1}^{\mathrm{T}}HY = \frac{1}{n}\mathbf{1}^{\mathrm{T}}Y$$

The means of $\hat{\mathbf{e}}_Y$ and $\hat{\mathbf{e}}_{X_i}$ have been verified to be 0. Therefore, the fitted line of $\hat{\mathbf{e}}_Y$ regressing on $\hat{\mathbf{e}}_{X_i}$ will pass through the center of gravity of its data, which is $(\frac{1}{n}\mathbf{1}^T\hat{\mathbf{e}}_{X_i}, \frac{1}{n}\mathbf{1}^T\hat{\mathbf{e}}_Y) = (0,0)$.

Proof 2:

Suppose $\hat{\mathbf{e}}_{X_i}$ is the residuals of regressing X_i on $X_{(i)}$, where X_i is the i-th column of data $X_{n\times p}$ and $X_{(i)}$ is the matrix after removing the i-th column X_i from X. Similarly, $\hat{\mathbf{e}}_Y$ is the residuals of regressing Y on $X_{(i)}$.

By the knowledge of OLS estimator, we know that

$$\hat{\mathbf{e}}_{X_i} = X_i - X_{(i)}(X_{(i)}^{\mathrm{T}}X_{(i)})^{-1}X_{(i)}^{\mathrm{T}}X_i = (I-H)X_i$$
$$\hat{\mathbf{e}}_Y = Y - X_{(i)}(X_{(i)}^{\mathrm{T}}X_{(i)})^{-1}X_{(i)}^{\mathrm{T}}Y = (I-H)Y$$

where I is the $n\times n$ identity matrix and $H = X_{(i)}(X_{(i)}^{\mathrm{T}}X_{(i)})^{-1}X_{(i)}^{\mathrm{T}}$ is a projection matrix.

When we regress $\hat{\mathbf{e}}_Y$ on $\hat{\mathbf{e}}_{X_i}$, we have:

$$\hat{\mathbf{e}}_Y = \hat{\beta}_0\mathbf{1} + \hat{\beta}_1\hat{\mathbf{e}}_{X_i} \tag{13.2}$$

By the knowledge of OLS estimator, we know that

$$\begin{aligned}\hat{\beta} &= (\hat{\mathbf{e}}_{X_i}^{\mathrm{T}}\hat{\mathbf{e}}_{X_i})^{-1}\hat{\mathbf{e}}_{X_i}^{\mathrm{T}}\hat{\mathbf{e}}_Y \\ &= (X_i^{\mathrm{T}}(I-H)^2X_i)^{-1}X_i^{\mathrm{T}}(I-H)^2Y \\ &= (X_i^{\mathrm{T}}(I-H)X_i)^{-1}X_i^{\mathrm{T}}(I-H)Y \end{aligned} \tag{13.3}$$

By plugging (13.3) into (13.2) and pre-multiply X_i^{T}, we have

$$\begin{aligned} X_i^{\mathrm{T}}\hat{\mathbf{e}}_Y &= X_i^{\mathrm{T}}\hat{\beta}_0\mathbf{1} + X_i^{\mathrm{T}}\hat{\beta}_1\hat{\mathbf{e}}_{X_i} \\ X_i^{\mathrm{T}}(I-H)Y &= X_i^{\mathrm{T}}\hat{\beta}_0\mathbf{1} + X_i^{\mathrm{T}}(X_i^{\mathrm{T}}(I-H)X_i)^{-1}X_i^{\mathrm{T}}(I-H)Y(I-H)X_i \qquad (\textit{rearranging the scalars}) \\ X_i^{\mathrm{T}}(I-H)Y &= X_i^{\mathrm{T}}\hat{\beta}_0\mathbf{1} + [X_i^{\mathrm{T}}(I-H)Y][(X_i^{\mathrm{T}}(I-H)X_i)^{-1}][X_i^{\mathrm{T}}(I-H)X_i] \\ X_i^{\mathrm{T}}(I-H)Y &= X_i^{\mathrm{T}}\hat{\beta}_0\mathbf{1} + X_i^{\mathrm{T}}(I-H)Y \\ 0 &= X_i^{\mathrm{T}}\hat{\beta}_0\mathbf{1} \end{aligned}$$

$X_i^{\mathrm{T}}\hat{\beta}_0\mathbf{1}^{\mathrm{T}} = (\sum_{j=1}^{n} x_{ji})\hat{\beta}_0 = 0$, where x_{ji} is the element in the j-th row and i-column of X. Therefore, we have $\hat{\beta}_0 = 0$, that is the intercept in the added variable scatter plot 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.

```
#manually construct the avplot for Change
UN3.trans2 = na.omit(UN3.trans2)
Change.lm = lm(UN3.trans2, formula = Change ~ PPgdp.trans + Pop.trans +
                 Frate + Fertility + Purban)
ModernC.lm = lm(UN3.trans2, formula = ModernC ~ PPgdp.trans + Pop.trans +
                  Frate + Fertility + Purban)
Change.res = Change.lm$residuals
ModernC.res = ModernC.lm$residuals
Change.av.lm = lm(ModernC.res ~ Change.res)
res=rbind(
  summary(Change.av.lm)$coefficients['Change.res',],
  summary(UN3.trans2.lm)$coefficients['Change',]
)
rownames(res) = c('Coef in avPlots', 'Coef in full model')
kable(res, format = 'latex',
      caption = 'Coef of Change in avPlots and full model') %>%
  kable_styling(latex_options = c('striped', 'hold_position'), font_size=12)
```

Table 3: Coef of Change in avPlots and full model

	Estimate	Std. Error	t value	Pr(>|t|)
Coef in avPlots	6.038124	2.100464	2.874661	0.0047861
Coef in full model	6.038124	2.145641	2.814135	0.0057547

As is shown in the table above, the estimates of coefficient of Change in manually constructed added variable plot is the same as the estimate from my model in Ex. 10.

However, the standard errors and P-values are different. It is due to different degree of freedom in these two models. The sample size of non-NA data is 122 for our final model. The df in my full model is 122-7=115, while the df in the model for added variable plot is 122-2=120. The difference in df will result in different $\hat{\sigma}'s$, subsequently different t values. We will use t-distributions of different df's as well.