HW2 STA521 Fall18

Bin Han

NetID: bh193

GitHub: BeanHam

Due September 23, 2018

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)
data(UN3, package="alr3")
library(dplyr)
library(ggplot2)
library(GGally)
library(knitr)
```

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

```
dim(UN3)
```

```
## [1] 210   7
```

```
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
```

(a) There are total of 7 variables, 6 of which have missing values. They are “ModernC”, “Change”, “PPgdp”, “Frate”, “Pop”, and “Fertility”.

(b) Based on the values and explanation from the data documentation, all the variables are quantitative.

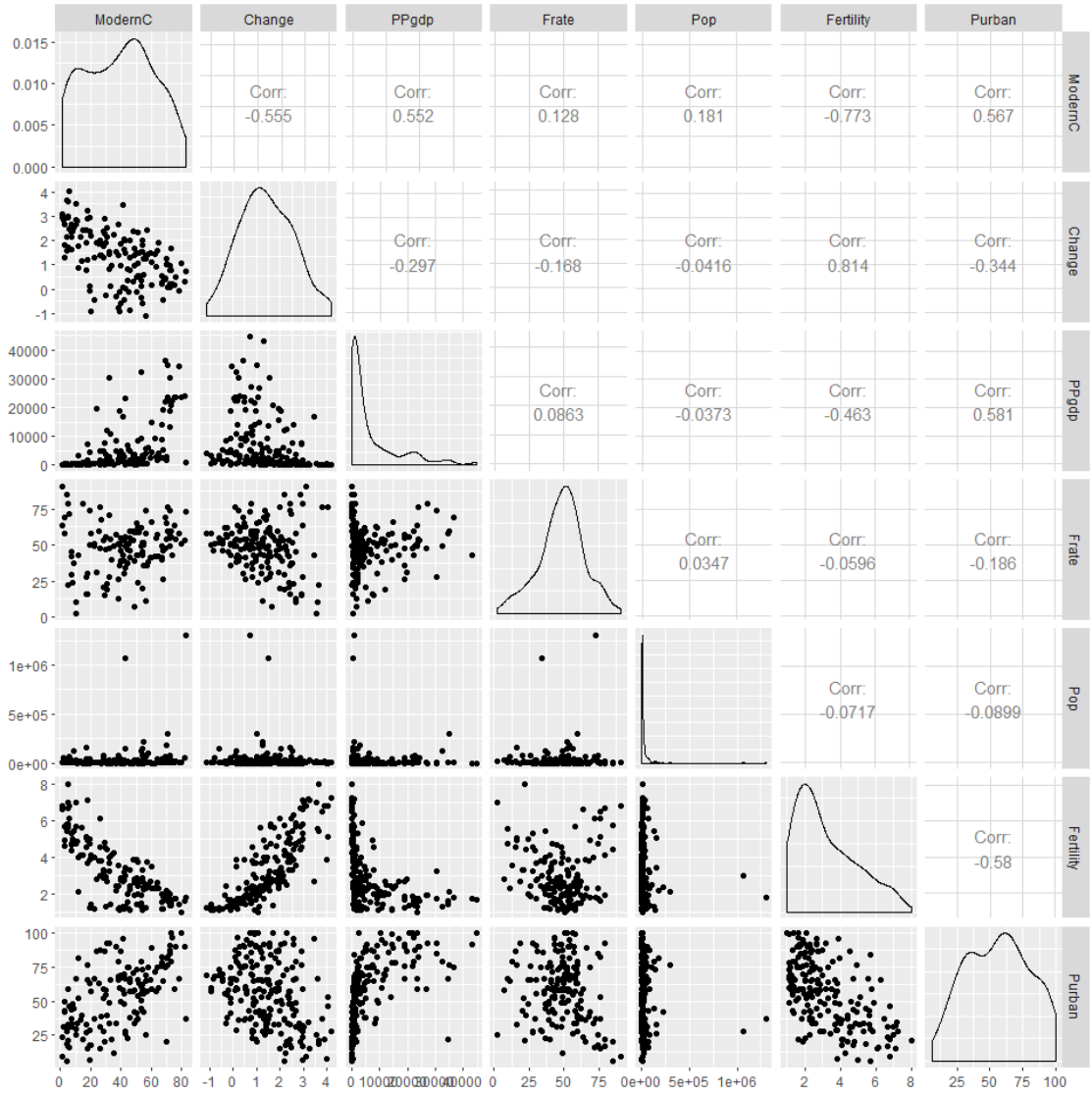
2. What is the mean and standard deviation of each quantitative predictor? Provide in a nicely formatted table.

```
mean <- sapply(UN3, function(x) mean(x, na.rm=TRUE))
sd <- sapply(UN3, function(x) sd(x, na.rm=TRUE))
kable(data.frame(mean, sd), digits = 3)
```

	mean	sd
ModernC	38.717	22.637
Change	1.418	1.133
PPgdp	6527.388	9325.189
Frate	48.305	16.532
Pop	30281.871	120676.694
Fertility	3.214	1.707
Purban	56.200	24.110

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?

```
ggpairs(UN3, columns =1:ncol(UN3))
```



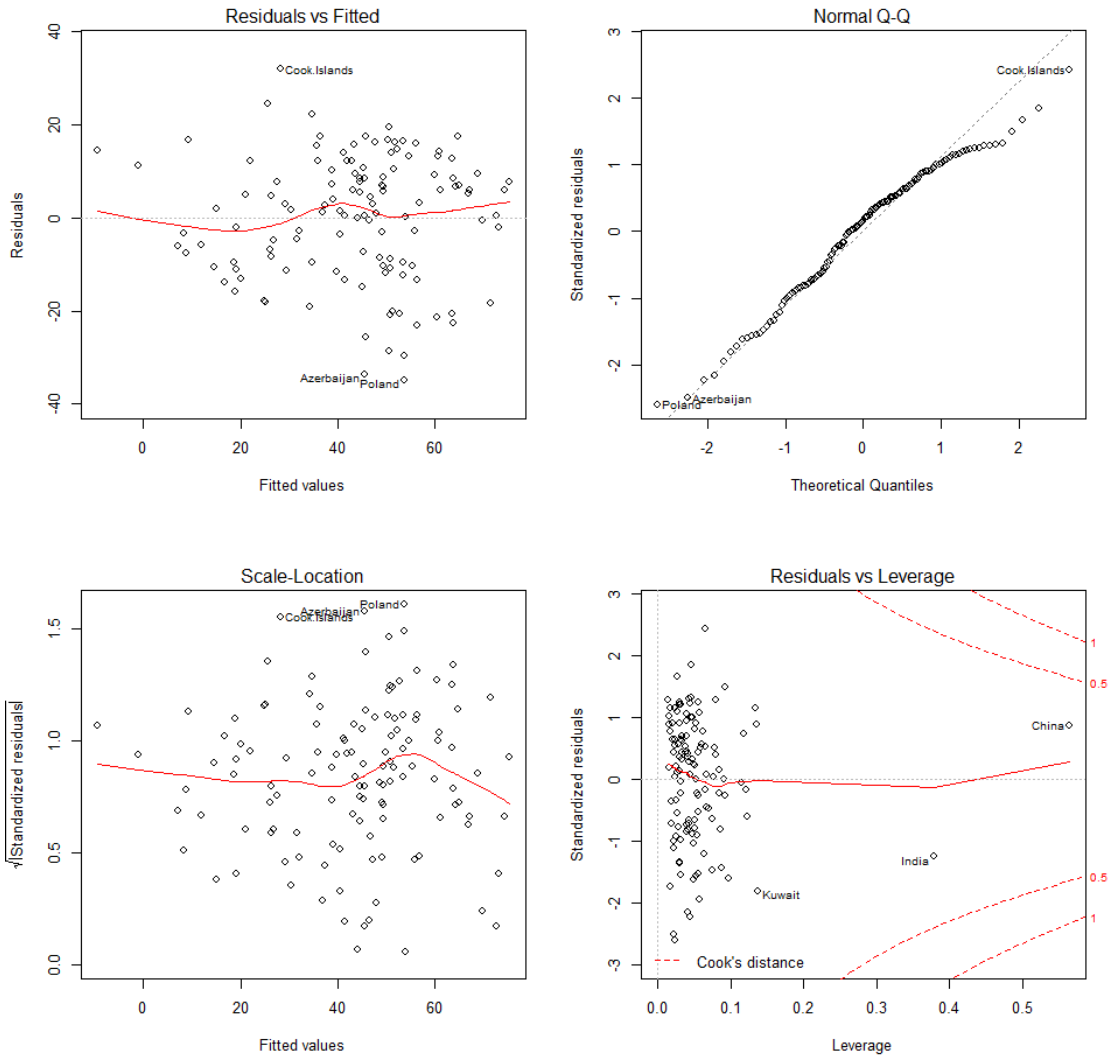
(a) Given that "ModernC" is the response variable, we can see that the correlation between "ModernC" and "Change", "Fertility", and "Purban" are fairly strong and linear. However, for explanatory variable "PPgdp", even though the correlation is strong, the relationship does not appear to be quite linear. There seems to be a quadratic pattern displayed. Some transformation is needed for "PPgdp" to make the relationship more linear.

(b) The variable "Pop" has an extremely right-skewed distribution. It is mainly because it has two observations with extremely high values. Those two values could be potential outliers, which need to be testified in later process.

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?

```
ModernC_lm <- lm(ModernC~., data=UN3)

summary(ModernC_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

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



(a) Based on the summary of linear regression model, the degree of freedom is 118. Calculating backward, n - p -1 = 118, so that we have n = 125. Therefore, 125 observations have been used to fit the model. We can also prove that from the explanation that “85 observations deleted due to missingness”, 210 - 85 = 125.

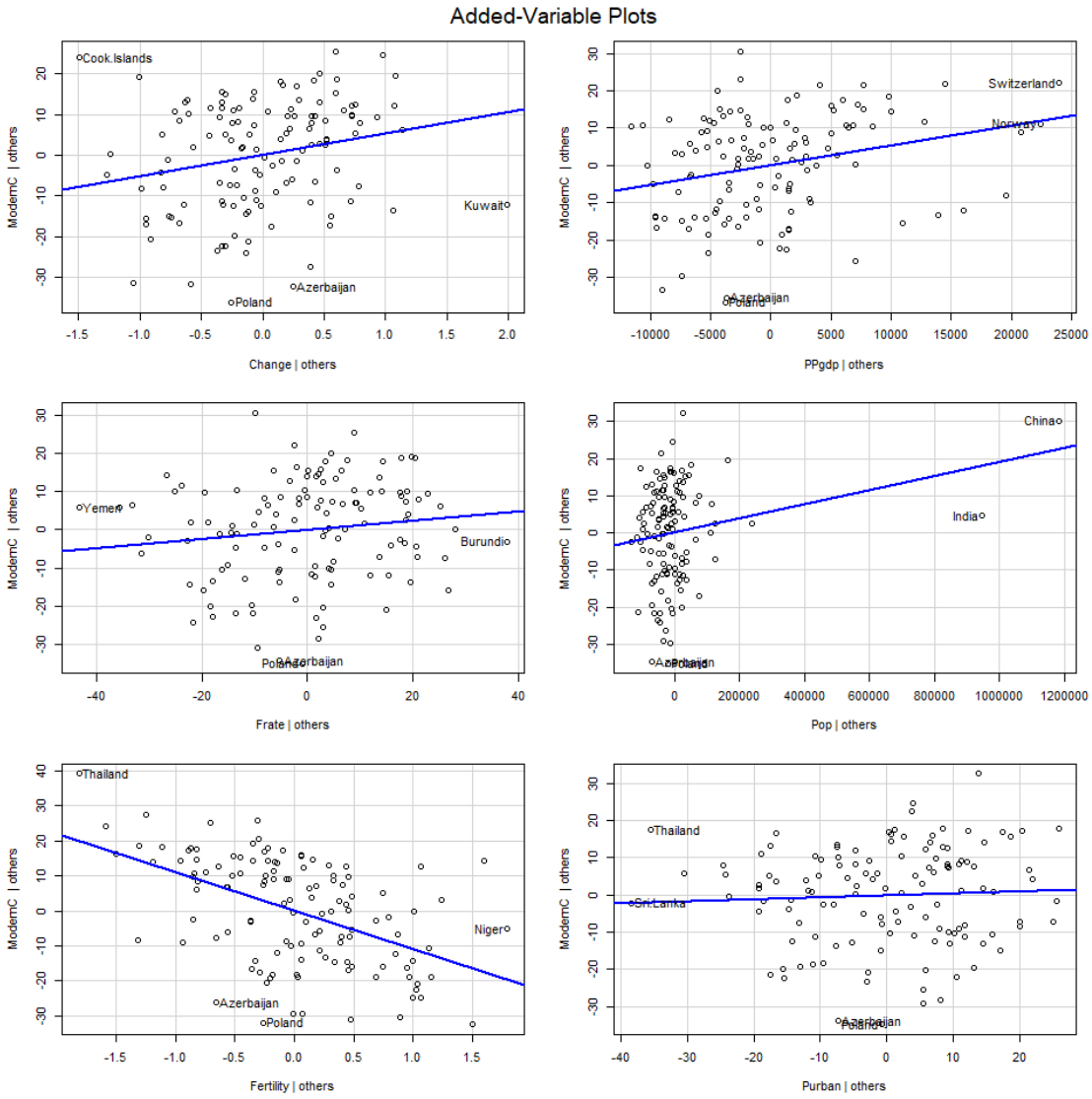
(b) From the diagnostic plots, we can see that the variances of residuals over the fitted values are fairly constant, with some variations displayed. It can be also seen from the Scale-Location graph.

(c) From the normal qqplot, we can see that several points scatter below the normal line on the right side, with most of the observations on the line. Therefore, we may conclude that the distribution is roughly normal.

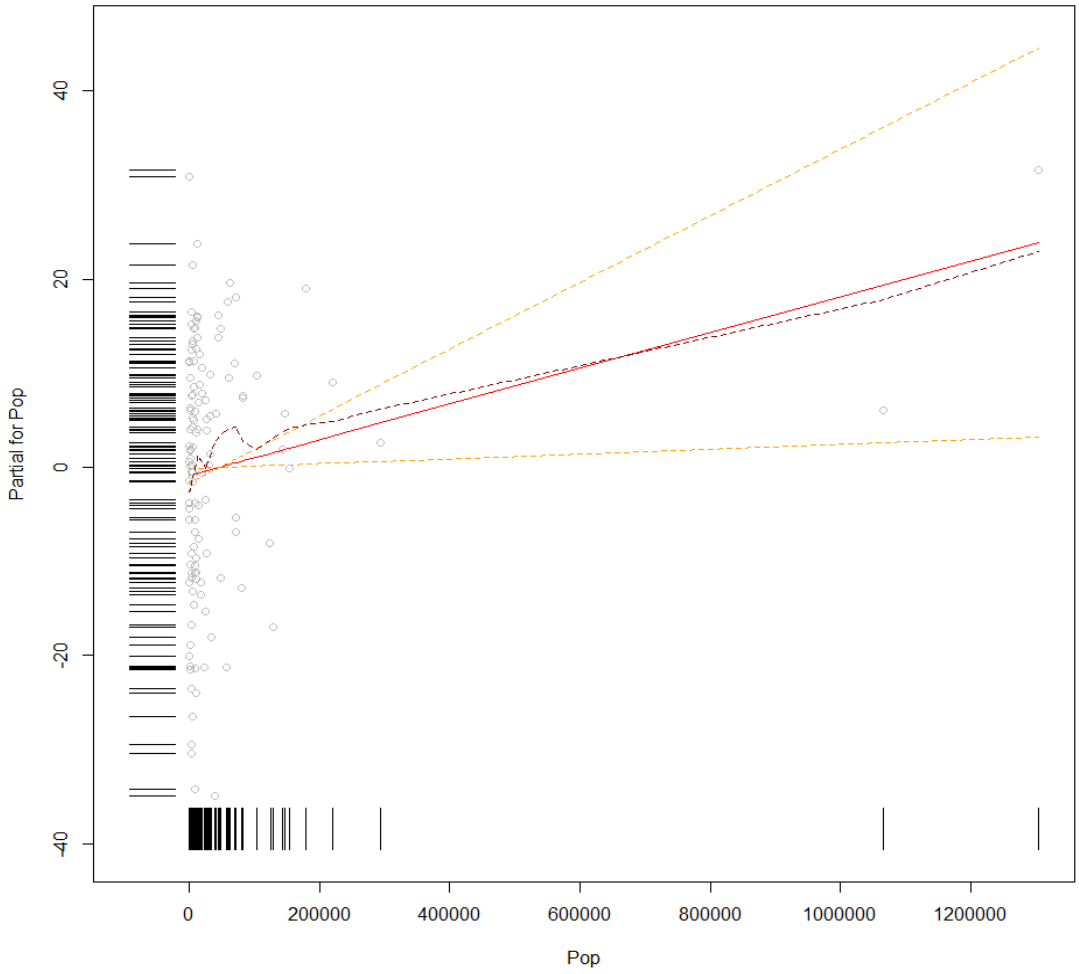
(d) From the leverage plot and Cook’s Distance value, we can see that there are no influential points. However, there are values that have high leverages, such as China and India.

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?

```
car::avPlots(ModernC_lm)
```



```
termplot(ModernC_lm, terms = "Pop",
         partial.resid = T, se=T, rug=T,
         smooth = panel.smooth)
```



(1) The scatter plot between "Pop" and "ModernC" shows that "Pop" needs to be transformed since there are two observations, China and India, that are way far away from other observations. Therefore, we could potentially apply some transformation on the variable "Pop" to make it less skewed and more linear.

(2) As discussed in question 3, the relationship between "ModernC" and "PPgdp" does not seem to be quite linear, with a quadratic pattern shown. Based on the scatter plot above, we can also see that some observations scattered far to the right and down below. Some transformation is also needed on "PPgdp".

(3) There is no actual influential point existing in the dataset, as tested in question (4) with the Cook's Distance Criteria. And also, from the term plot of Pop, we can see China and India are not influential points.

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_nona <- UN3 %>% na.omit()

summary(powerTransform(cbind(PPgdp, Pop, Fertility, Purban, Change, Frate)~.,
data = UN3_nona, family = "bcnPower"))

## bcnPower transformation to Multinormality
##
## Estimated power, lambda
##           Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## PPgdp       -0.1438      -0.144      -0.2391      -0.0486
## Pop          0.0629       0.000      -0.0043       0.1302
## Fertility    0.1786       0.000      -0.0744       0.4317
## Purban       0.7971       1.000      -2.1966       3.7908
## Change       0.0613       1.000      -6.3660       6.4885
## Frate        0.8828       1.000      -0.7289       2.4944
##
## Estimated location, gamma
##           Est gamma Std Err. Wald Lower Bound Wald Upper Bound
## PPgdp        0.1000       NA               NA               NA
## Pop          0.1000       NA               NA               NA
## Fertility    0.1000       NA               NA               NA
## Purban     450.9307       NA               NA               NA
## Change      51.9840       NA               NA               NA
## Frate      180.2682       NA               NA               NA
##
## Likelihood ratio tests about transformation parameters
##                                      LRT df       pval
## LR test, lambda = (0 0 0 0 0 0)  16.17995  6 0.01282001
## LR test, lambda = (1 1 1 1 1 1) 919.36775  6 0.00000000



UN3_xtransform <- UN3 %>%
    mutate(Pop_trans = log(Pop), PPgdp_trans = log(PPgdp))
```

Based on the powerTransform result, we can see that the ideal transformation for the two variables are:

PPgdp_trans = PPgdp^ (-0.144); Pop_trans =log (Pop)

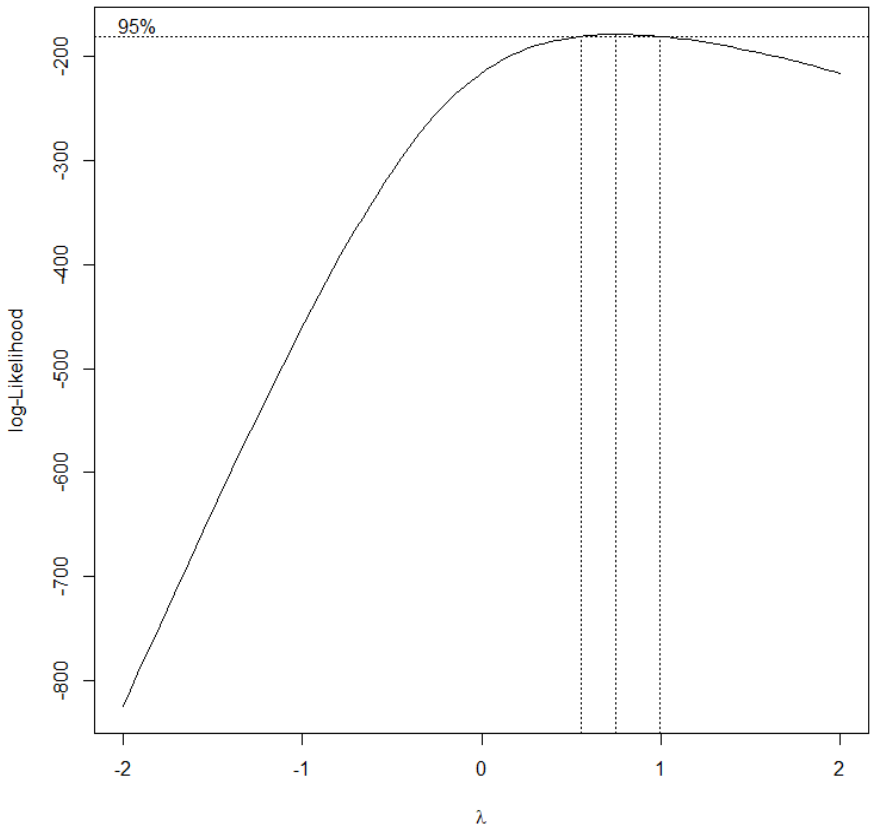
However, considering of interpretability of the variables, I decided to use the log transformation (lambda = 0) which are very close to the ideal transformations:

PPgdp_trans = log (PPgdp); Pop_trans = log (Pop)

Another independent variable “fertility” also has a suggested transformation from the result. However, judging the linearity from the scatter plot between “fertility” and “ModernC”, I decided not to transform it.

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

```
ModernC_xtrans <- lm(ModernC ~ Pop_trans + Change + PPgdp_trans + Frate +
Fertility + Purban, data=UN3_xtransform)
MASS::boxcox(ModernC_xtrans)
```



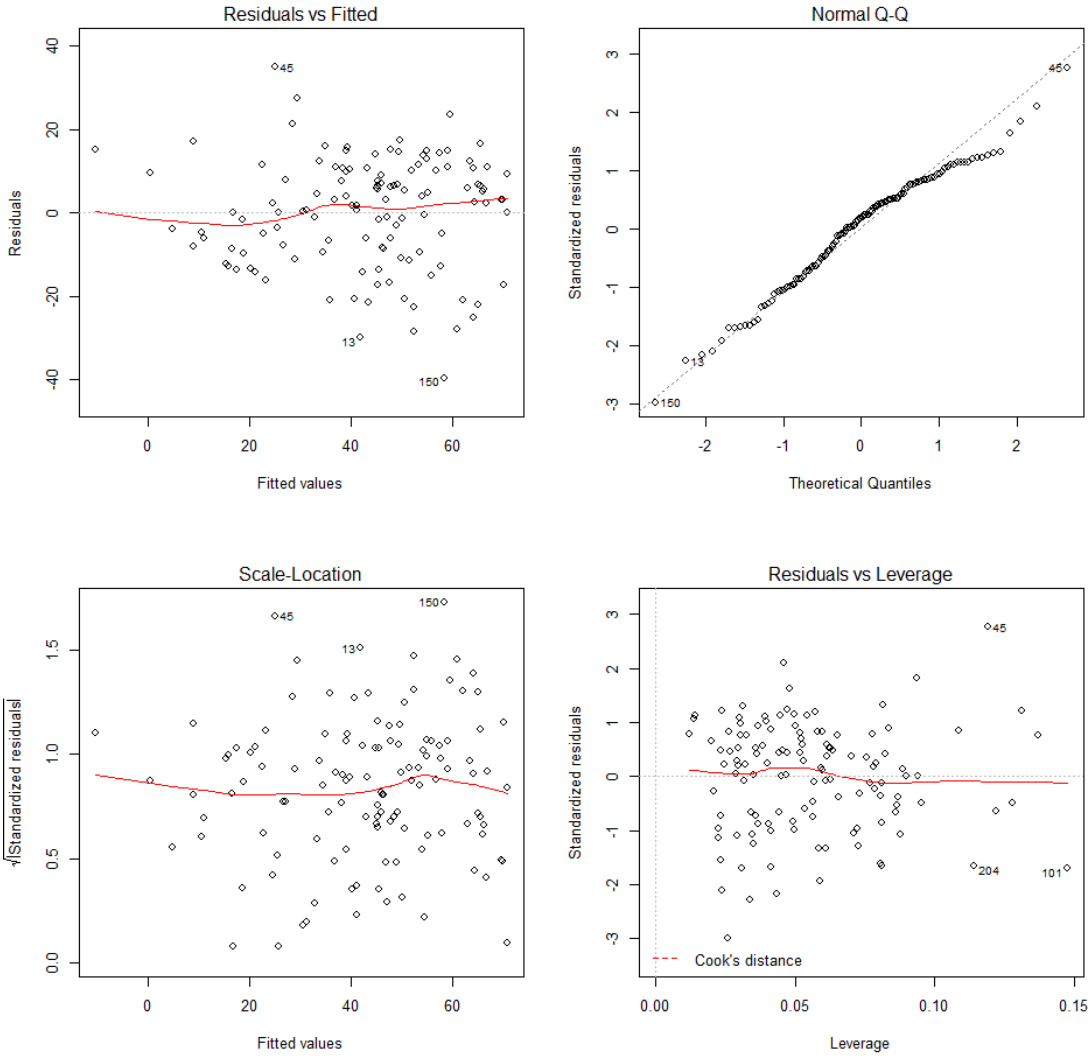
Given the transformation on “Pop” and “PPgdp”, the ideal transformation of response variable covers the range of (0.5, 1). Similarly, taken the easiness of interpretating the result into consideration, I decided to use lambda = 1, which does not transform the response variable.

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.

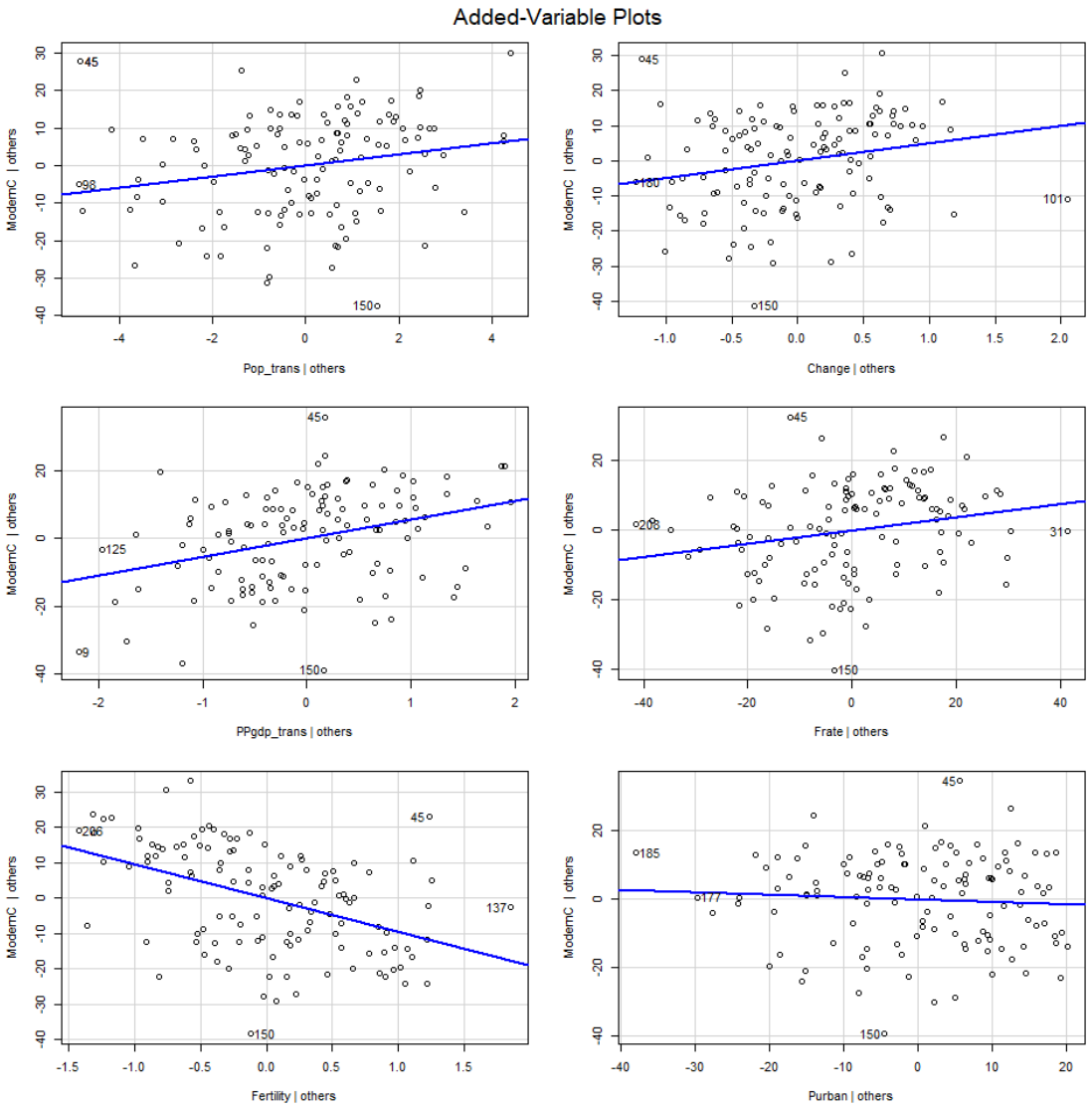
```
summary(ModernC_xtrans)
```

```
##
## Call:
## lm(formula = ModernC ~ Pop_trans + Change + PPgdp_trans + Frate +
##     Fertility + Purban, data = UN3_xtransform)
##
## 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
## Pop_trans    1.47207    0.62875   2.341 0.020897 *
## Change       4.99296    2.07709   2.404 0.017781 *
## PPgdp_trans  5.50728    1.40505   3.920 0.000149 ***
## 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
##   (85 observations deleted due to missingness)
## Multiple R-squared:  0.626,  Adjusted R-squared:  0.6069
## F-statistic: 32.91 on 6 and 118 DF,  p-value: < 2.2e-16
```

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




```
car::avPlots(ModernC_xtrans)
```

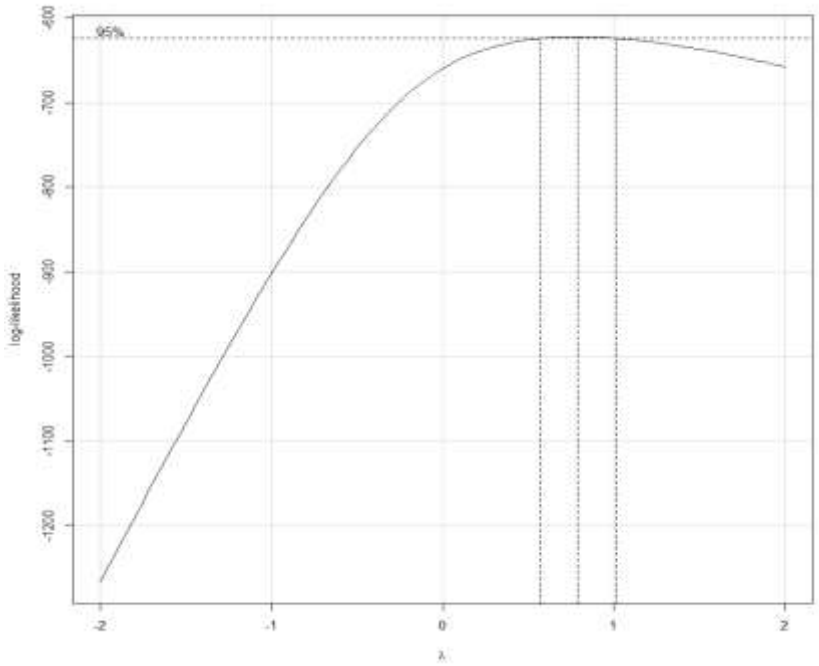


(1) From the residual plot and scale-location plot, we can see that the variance of residual is fairly constant, with some up and down variation. From the normal quantile plot, we can determine that it is roughly normal. Based on the leverage plot, we can conclude that there are still no influential points.

(2) From the added variable plots, we can see that the transformed PPgdp and Pop both look much more linear with ModernC.

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(ModernC_lm)
```




```
UN3_nona <- UN3 %>% na.omit()

summary(powerTransform(cbind(PPgdp, Pop, Fertility, Purban, Change, Frate)~.,
data = UN3_nona, family = "bcnPower"))

## bcnPower transformation to Multinormality
##
## Estimated power, lambda
##           Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## PPgdp       -0.1438      -0.144      -0.2391      -0.0486
## Pop          0.0629       0.000      -0.0043       0.1302
## Fertility    0.1786       0.000      -0.0744       0.4317
## Purban       0.7971       1.000      -2.1966       3.7908
## Change       0.0613       1.000      -6.3660       6.4885
## Frate        0.8828       1.000      -0.7289       2.4944
##
## Estimated location, gamma
##           Est gamma Std Err. Wald Lower Bound Wald Upper Bound
## PPgdp        0.1000       NA               NA               NA
## Pop          0.1000       NA               NA               NA
## Fertility    0.1000       NA               NA               NA
## Purban     450.9307       NA               NA               NA
```

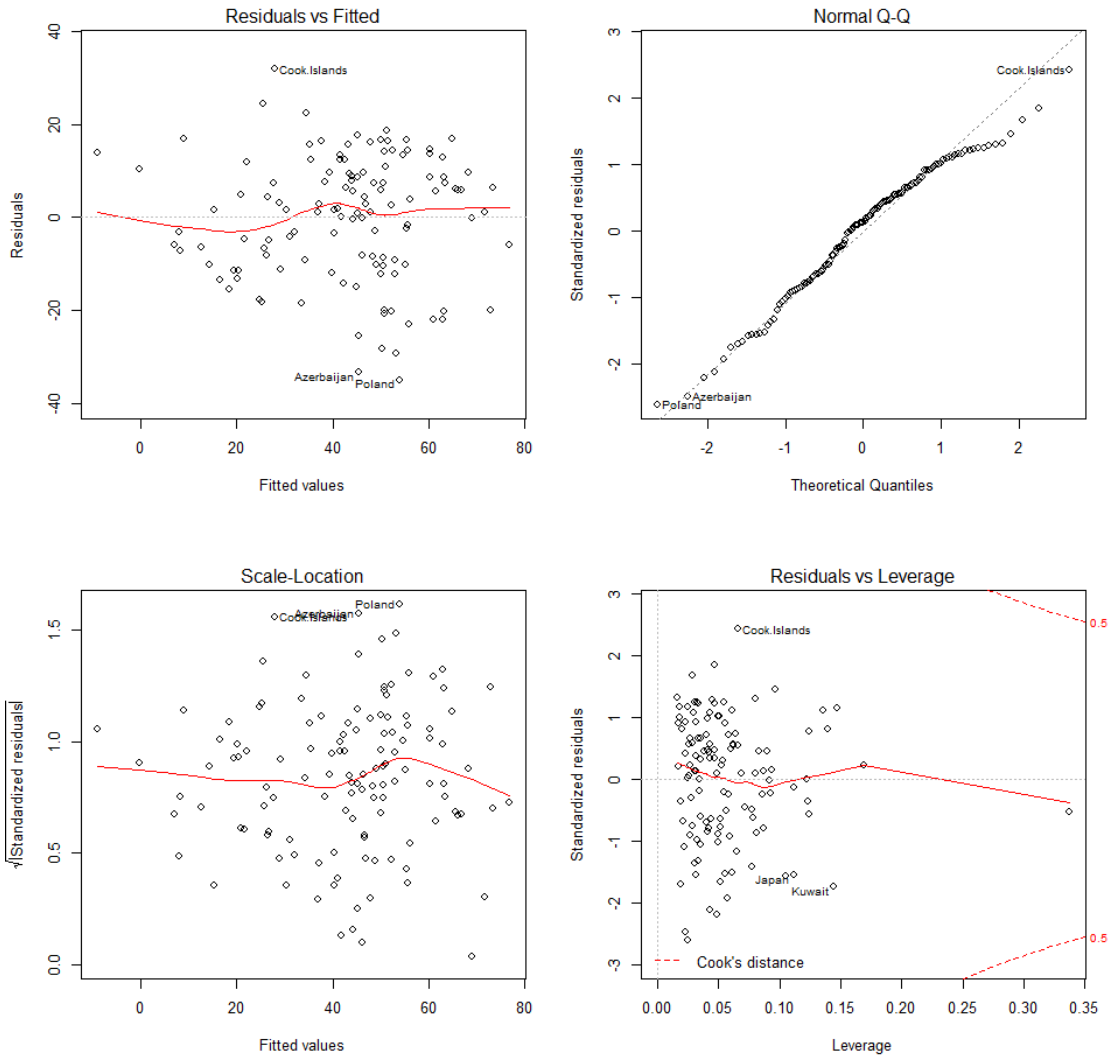
```
## Change      51.9840       NA               NA               NA
## Frate      180.2682       NA               NA               NA
##
## Likelihood ratio tests about transformation parameters
##                                        LRT df       pval
## LR test, lambda = (0 0 0 0 0 0)  16.17995  6 0.01282001
## LR test, lambda = (1 1 1 1 1 1) 919.36775  6 0.00000000

UN3_xtransform <- UN3 %>%
    mutate(Pop_trans = log(Pop), PPgdp_trans = log(PPgdp))
```

No. The result would be the same.

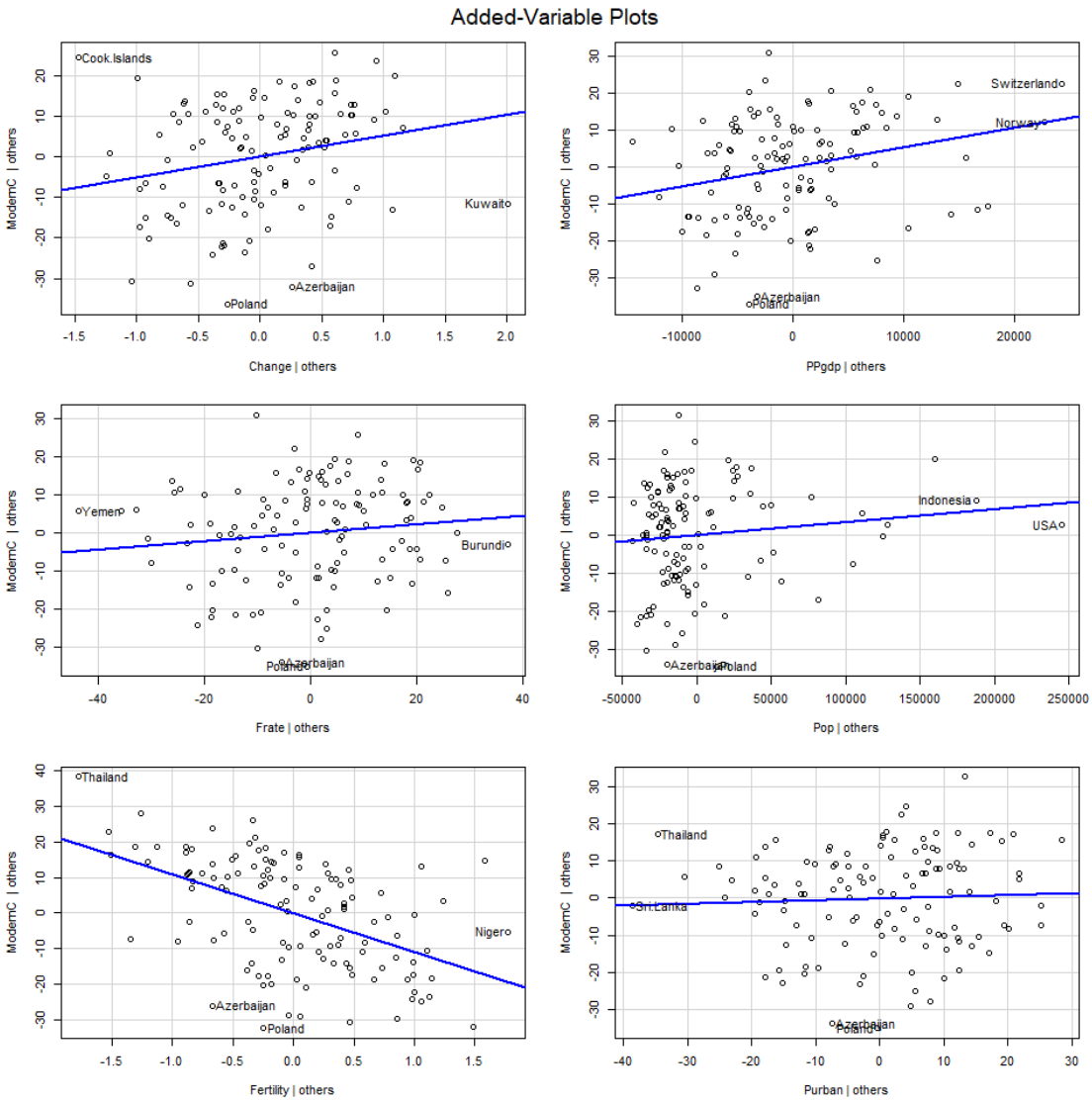
It is because when I started with transforming the response variable, the ideal range of exponent is still (0.5, 1). For the sake of interpretability, I will still choose to use the original response variable, which corresponds to lamda = 1. Afterthen, the process of finding out the transformation for explanatory variables is exactly the same as shown in the previous questions. Therefore, the two models will be the same.

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.

```
## Cook's Distance
rownames(UN3)[cooks.distance(ModernC_lm)>1]

## character(0)



## Bonferonni Correction
abs.ti <- abs(rstudent(ModernC_xtrans))
pval <- 2*(1-pt(max(abs.ti), ModernC_xtrans$df-1))

criteria <- 0.05/210
mean(pval < criteria)

## [1] 0



## Remove observations with higher leverage values.
ModernC_remove <- lm(ModernC~., data=UN3, subset = rownames(UN3)!=
c("China","India"))

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




```
avPlots(ModernC_remove)
```



From previous Leverage Plot and the Cook's Distance method, there are no influential points.

Using Bonferronni Correction, which compares p-value with alpha/n, we fail to reject the null hypothesis, meaning that there are no outliers either.

If instead, we remove the two observations with higher leverage values, China and India, we find out that all the plots in the diagnostic analysis are approximately the same. The variance of residuals is still roughly constant. The normal qqplot indicates that the distribution of response variable is fairly normal.

The only difference is that even though we remove China and India, there are still other observations that will come up with higher leverage values. Therefore, we do not need to remove the two observations. Transform explanatory variables would be enough.

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!

```
table <- data.frame(ModernC_xtrans$coefficients, confint(ModernC_xtrans))
table <- table %>%
    rename(Coefficients = ModernC_xtrans.coefficients, Lower_Bound = X2.5..,
Upper_Bound = X97.5..)
table
```

```
##             Coefficients  Lower_Bound Upper_Bound
## (Intercept)   4.11547111 -24.61538573  32.8463280
## Pop_trans     1.47207436   0.22696989   2.7171788
## Change        4.99295735   0.87974961   9.1061651
## PPgdp_trans   5.50727842   2.72490390   8.2896530
## Frate         0.18939357   0.03669429   0.3420929
## Fertility    -9.67594142 -13.17233431  -6.1795485
## Purban       -0.07076799  -0.26403910   0.1225031
```

Interpretation:

(a) Change: For each one percentage point increase in the annual population growth rate, the percent of unmarried women will increase by 4.993 percentage points.

(b) Frate: For each one percentage point increase in the percent of female over age 15 economically active, the percent of unmarried women will increase by 0.189 percentage point.

(c) Fertility: For each one unit increase in the expected number of live births per female, the percent of unmarried women will decrease by 9.67 percentage points.

(d) Purban: For each one percentage point increase in percent of population that is urban, the percent of unmarried women will decrease by 0.071 percentage point.

(e) Pop_trans: For each 10% increase in Population, the percent of unmarried women will increase by 1.47207 * log(1.1) = 0.14 percentage point.

(f) PPgdp_trans: For each 10% increase in per Capita 2001 GDP, the percent of unmarried women will increase 5.507278 * log(1.1) = 0.5248 percentage point.

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

With all the diagnostic analysis and model selection, I ended up with the model:

$$\text{ModernC} = 4.115 + 4.993\textit{Change} + 0.189\text{Frate} - 9.676\textit{Fertility} - 0.071\text{Purban} + 1.472\textit{log(Pop)} + 5.507\log(\text{PPgdp})$$

Generally, we can see that the percentage of unmarried women is negatively correlated with expected number of live birth per female and the percent of population that is urban, while positively correlated with other variables. Based on US's population composition and developmental plan, the government could take actions on policy-making to influence the population structure of unmarried women.

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.

$$

For the added variable plot, we are regressing the residuals from the overall model excluding one explanatory variable (ex. x1) on the residuals from regressing x1 on all other independent variables.

$$e_y = \hat{\beta}_0 + \hat{\beta}_1 e_{x_3}$$

$$(I-H)Y = \hat{\beta}_0 + \hat{\beta}_1 (I-H)X_3$$

Given that: $\hat{\beta}_1 = (X^TX)^{-1}X^TY$ We substitute $X^{=}(I-H)X_3$, $\sim Y = (I-H)Y$ We have:

$$\begin{aligned}(I-H)Y &= \hat{\beta}_0 + [X_3^T(I-H)(I-H)X_3]^{-1}[(I-H)X_3]^T(I-H)Y(I-H)X_3 \\ &= \hat{\beta}_0 + [X_3^T(I-H)X_3]^{-1}X_3^T(I-H)Y(I-H)X_3\end{aligned}$$

Multiply both sides with X_3^T, we have:

$$X_3^T(I-H)Y = X_3^T\hat{\beta}_0 + X_3^T[X_3^T(I-H)X_3]^{-1}X_3^T(I-H)Y(I-H)X_3$$

Since $[X_3^T(I-H)X_3]^{-1}$ and $X_3^T(I-H)Y$ are both scalar, we have:

$$\begin{aligned}X_3^T(I-H)Y &= X_3^T\hat{\beta}_0 + [X_3^T(I-H)X_3][X_3^T(I-H)X_3]^{-1}X_3^T(I-H)Y \\ &= X_3^T\hat{\beta}_0 + X_3^T(I-H)Y\end{aligned}$$

So we have:

$$X_3^T \hat{\beta}_0 = 0$$

Therefore:

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

$$

14. For multiple regression with more than 2 predictors, say a full model given by `Y ~ X1 + X2 + ... Xp` we create the added variable plot for variable `j` by regressing `Y` on all of the X's except `Xj` to form `e_Y` and then regressing `Xj` on all of the other X's to form `e_X`. Confirm that the slope in a manually constructed added variable plot for one of the predictors in Ex. 10 is the same as the estimate from your model.

```
UN3_xtransform_NoNA <- UN3_xtransform %>% na.omit()
e_Y <- residuals(lm(ModernC ~ Pop_trans + PPgdp_trans + Change + Fertility +
Purban, data=UN3_xtransform_NoNA))


e_X <- residuals(lm(Frate ~ Pop_trans + PPgdp_trans + Change + Fertility +
Purban, data=UN3_xtransform_NoNA))

residual_regreesion<- data.frame(e_Y, e_X)

addedvairable <- lm(e_Y ~ e_X, data=residual_regreesion)
summary(addedvairable)$coef

##                 Estimate Std. Error      t value   Pr(>|t|)
## (Intercept) 2.780438e-16  1.1777067 2.360892e-16 1.00000000
## e_X         1.893936e-01  0.0755267 2.507637e+00 0.01345709

summary(ModernC_xtrans)$coef

##                Estimate  Std. Error    t value     Pr(>|t|)
## (Intercept)  4.11547111 14.50853884  0.2836586 7.771692e-01
## Pop_trans    1.47207436  0.62875419  2.3412557 2.089650e-02
## Change       4.99295735  2.07709205  2.4038209 1.778126e-02
## PPgdp_trans  5.50727842  1.40504647  3.9196415 1.492131e-04
## Frate        0.18939357  0.07711025  2.4561402 1.550017e-02
## Fertility   -9.67594142  1.76561222 -5.4802189 2.444298e-07
## Purban      -0.07076799  0.09759825 -0.7250948 4.698293e-01
```

We can see that for variable "Frate", in the added regression, the estimated coefficien is 0.18939, which is exactly the same as the estimated coefficient of "Frate" in the overall regression.