HW2 STA521

[Jae Hyun Lee, jl914, jaehyunlee1221]

Due September 12, 2019 10am

Background Reading

Readings: Chapters 3-4, 8-9 and Appendix in Weisberg Applied Linear Regression

This exercise involves the UN data set from `alr3` package. Install `alr3` and the `car` packages and load the data to answer the following questions adding your code in the code chunks. Please add appropriate code to the chunks to suppress messages and warnings as needed once you are sure the code is working properly and remove instructions if no longer needed. Figures should have informative captions. Please switch the output to pdf for your final version to upload to Sakai. **Remove these instructions for final submission**

Exploratory Data Analysis

0. Preliminary read in the data. After testing, modify the code chunk so that output, messages and warnings are suppressed. *Exclude text from final*

```
library(alr3)
```

```
## Loading required package: car

## Loading required package: carData
```

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

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

```
str(UN3)
```

```
## 'data.frame':    210 obs. of  7 variables:
##  $ ModernC  : int  NA NA 49 NA NA NA 51 NA 22 NA ...
##  $ Change   : num  3.88 0.68 1.67 2.37 2.59 3.2 0.53 1.17 -0.45 2.02 ...
##  $ PPgdp    : int  98 1317 1784 NA 14234 739 8461 7163 687 NA ...
##  $ Frate    : int  NA NA 7 42 NA NA 63 44 51 53 ...
##  $ Pop      : num  23897 3167 31800 57 64 ...
##  $ Fertility: num  6.8 2.28 2.8 NA NA 7.2 NA 2.44 1.15 NA ...
##  $ Purban   : int  22 43 58 53 92 35 37 88 67 51 ...
```

```
smry_un3 <- summary(UN3)
na_count <- smry_un3[7,]
na_count
```

```
##        ModernC         Change          PPgdp          Frate            Pop
## "NA's   :58  "  "NA's   :1  "  "NA's   :9  " "NA's   :43  "  "NA's   :2  "
##      Fertility         Purban
## "NA's   :10  "             NA
```

answer: As we can see in outlook of data.frame UN3, there are all quantative variables. Except for variable named Purban, those of variables including ModernC, Change, PPgdp, Frate, Pop, Fertility have at least one missing data.

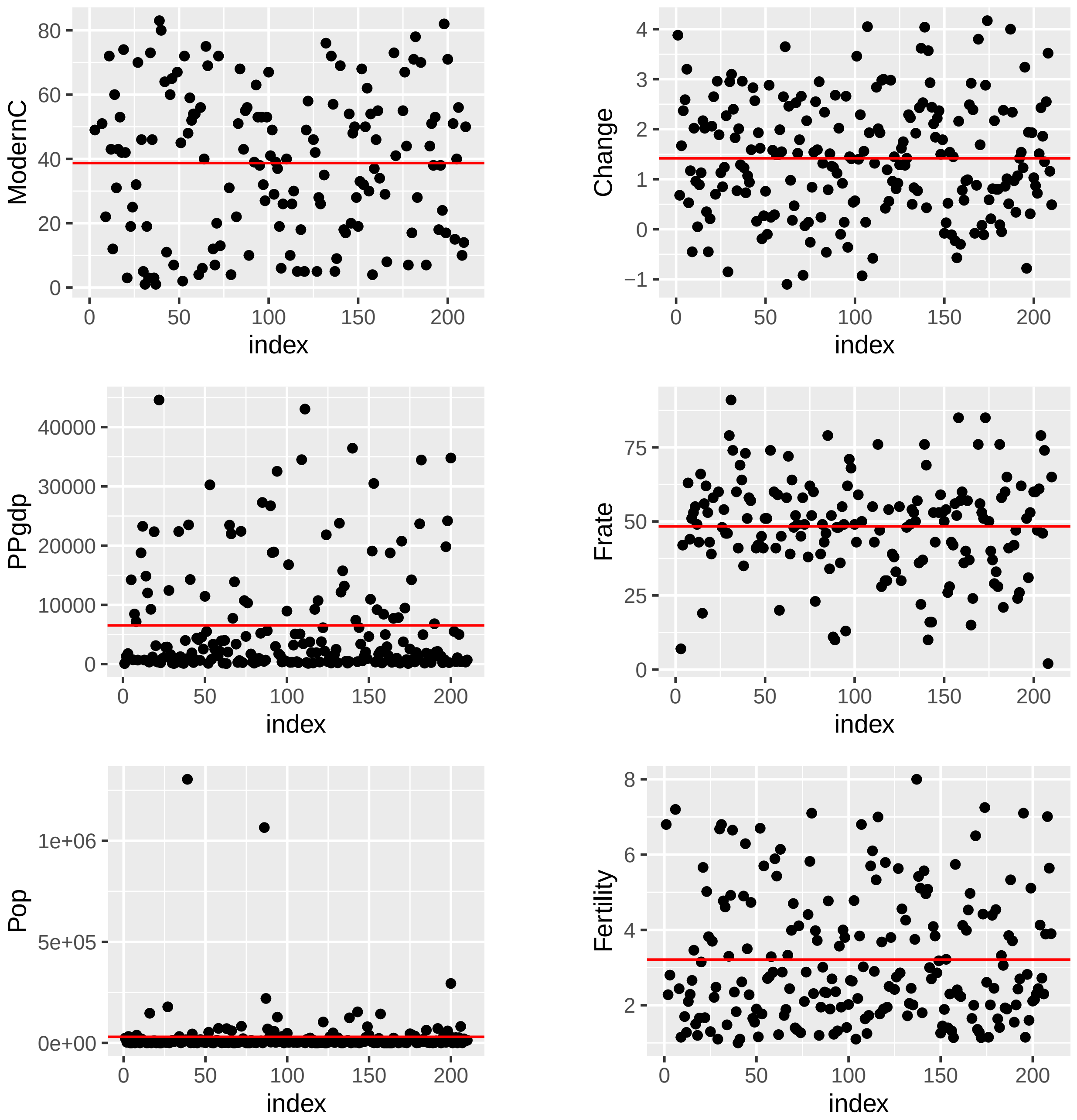
2. What is the mean and standard deviation of each quantitative predictor? Provide in a nicely formatted table.

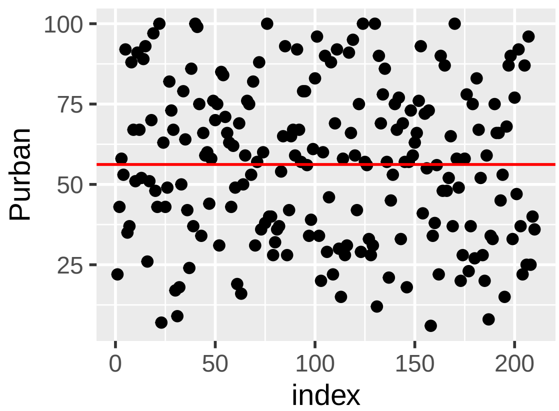
```r
library(knitr)
mn_st_table <- matrix(rep(0,3*length(UN3)),nrow = length(UN3))
for(i in 1:length(UN3)){
  mn_st_table[i,] <- c(colnames(UN3)[i],
                       round(mean(UN3[,i],na.rm = T),3),
                       round(sd(UN3[,i],na.rm=T),3))
}
rownames(mn_st_table) <- 1:length(UN3)
colnames(mn_st_table) <- c("variable","mean","stand deviation")
kable(mn_st_table)
```

variable	mean	stand deviation
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	56.2	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?

```r
library(ggplot2)
for(i in 1:length(UN3)){
  print(ggplot(data=UN3, mapping=aes(x=1:nrow(UN3),y=UN3[,i]))+
          geom_point()+
          geom_hline(yintercept = mean(UN3[,i],na.rm = T),color="red")+
          ylab(colnames(UN3)[i]) + xlab("index"))
}
```


ModernC
80
60
40
20
0
0
50
100
150
200
index
Change
4
3
2
1
0
−1
0
50
100
150
200
index
PPgdp
40000
30000
20000
10000
0
0
50
100
150
200
index
Frate
75
50
25
0
0
50
100
150
200
index
Pop
1e+06
5e+05
0e+00
0
50
100
150
200
index
Fertility
8
6
4
2
0
50
100
150
200
index

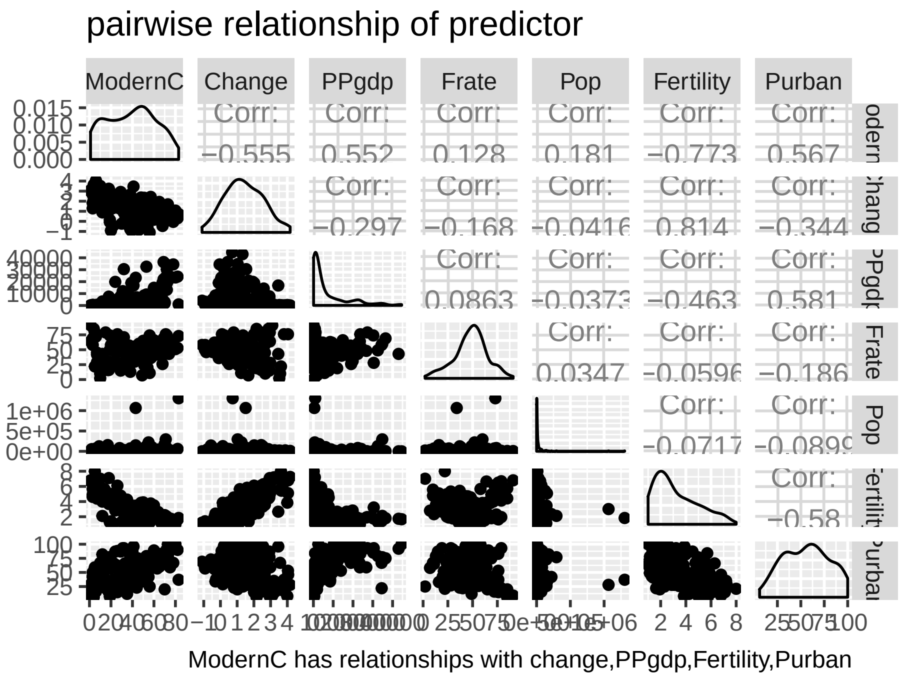


When we inspect scatterplots of predictors, most of them are distributed randomly from their mean. In case of PPgdp, they are skewed right. Thus I think it needs to be transformed. Furthemore, Pop seems to have some potential outliers. Therefore, we should be cautious dealing with Pop variable.

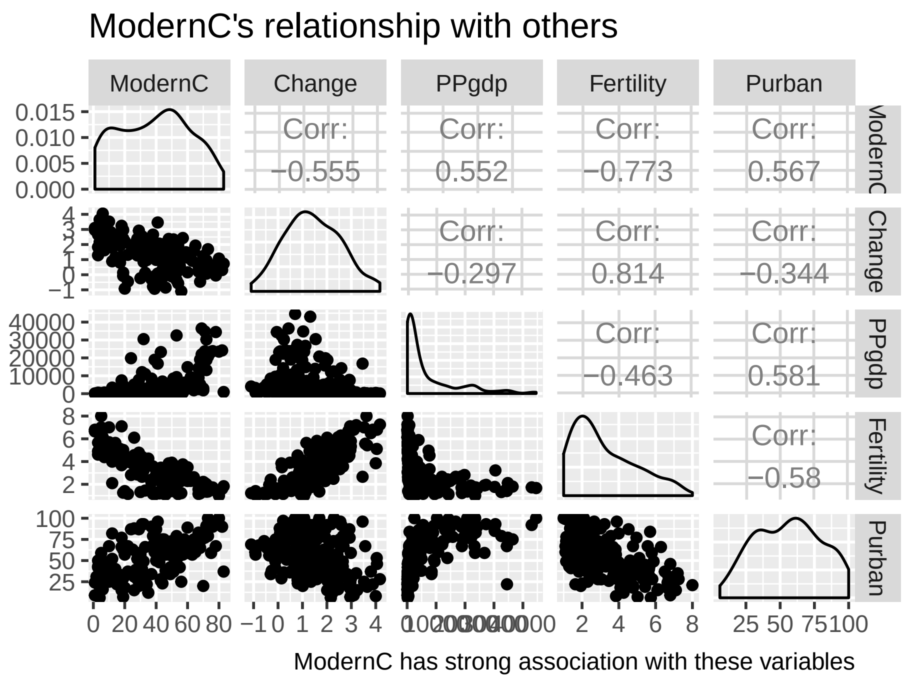
```r
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```r
ggpairs(UN3) +
  labs(title = "pairwise relationship of predictor",
       caption = "ModernC has relationships with change,PPgdp,Fertility,Purban")
```



```r
ggpairs(UN3[,c(1,2,3,6,7)]) +
  labs(title = "ModernC's relationship with others",
       caption = "ModernC has strong association with these variables")
```

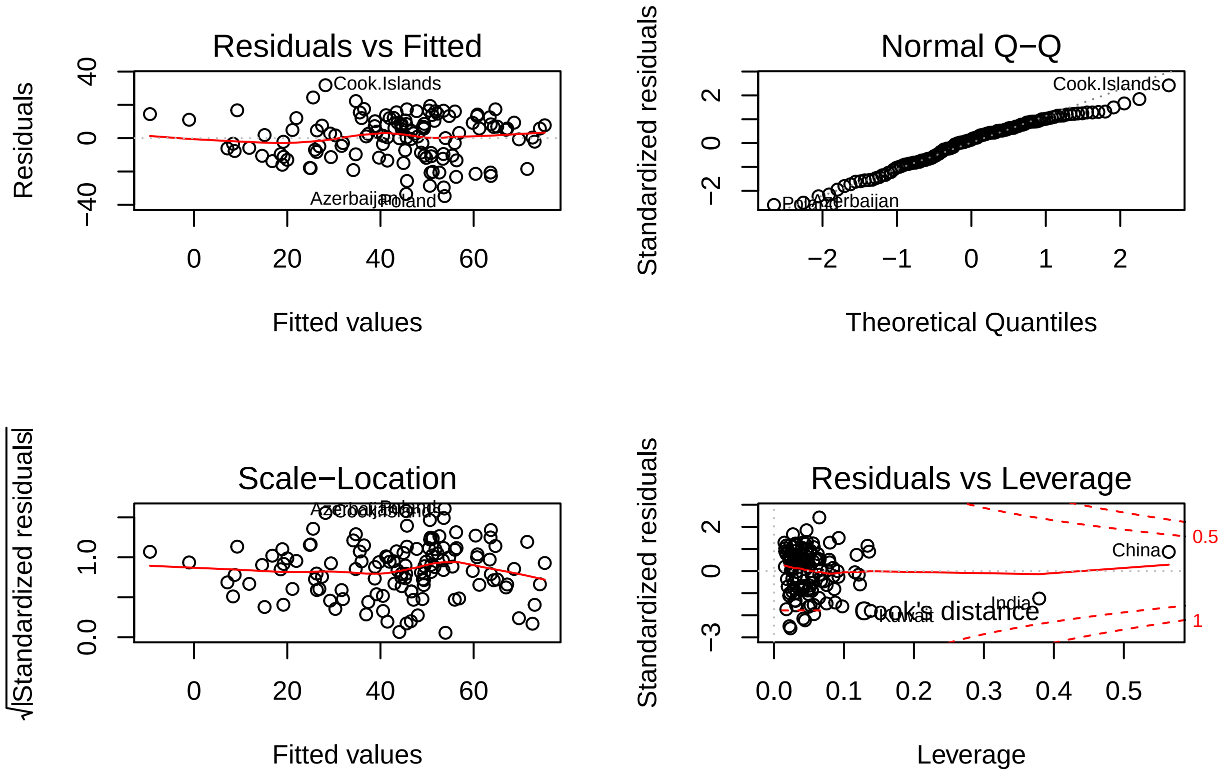


When I see pairwise plot among predictor variables, I can find that ModernC has quite strong relationship with variables named change,PPgdp,Fertility,Purban. Three of them have linear relationship with ModernC. But PPgdp seems to have non-linear relationship with ModernC. I think this phenomenom stem from skewness of PPgdp. Thus I should recheck after taking transformation on PPgdp.

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?

```r
Fm <- lm(ModernC~.,data = UN3)
par(mfrow=c(2,2))
plot(Fm)
```



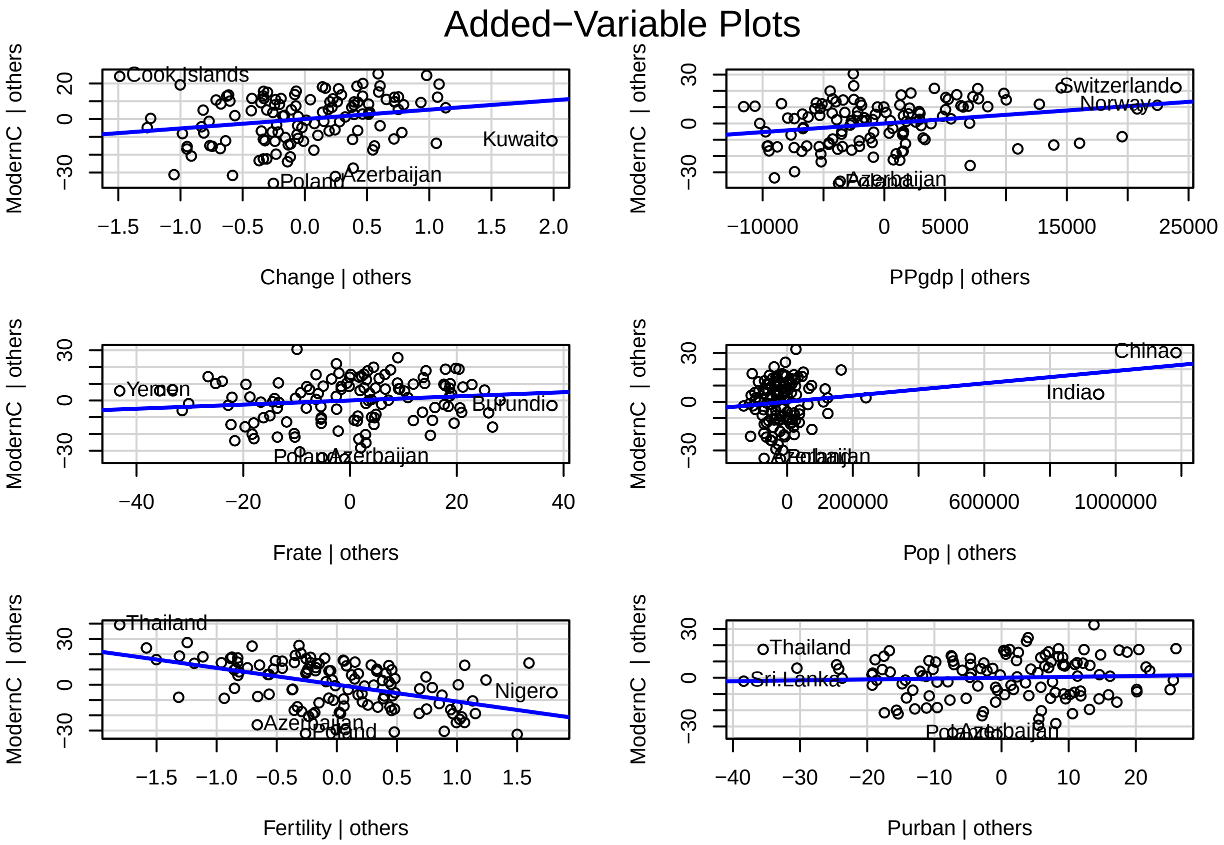
```r
summary(Fm)
```

```
##
## 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
```

When it comes to residual vs fitted value plot, there isn't any violation sign such as non-linear relationship between them. However, although it it not severe, I can find out normality assumption is violated at margin of normal q-qplot. In scale-location plot, they are randomly distributed forming straight band. Thus there is no evidence that homongenuity assumption is violated. But in leverage vs residual plot, there are some potential influencial point. Therefore we should pay attention to those observation. In model fitting, 118 observations are used and 85 observations omitted because of their missingness.

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?

```r
library(car)
avPlots(Fm)
```



Among these variables, it is the one, Pop, which need to be transformed. Because there are some potential influential point. As mentioned, Pop has potential influential point, China and India.

6. Using the multivariate BoxCox `car::powerTransform` or Box-Tidwell `car::boxTidwell` 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.

```r
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
## The following object is masked from 'package:GGally':
##
##     nasa

## The following object is masked from 'package:car':
##
##     recode

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
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                      
```

```
UN <-UN3 %>%
      mutate(Change_add = Change+1.2) %>%
      select(ModernC,Change_add,PPgdp,Frate,Pop,Fertility,Purban)
summary(UN)
```

```
##     ModernC        Change_add        PPgdp           Frate      
##  Min.   : 1.00   Min.   :0.100   Min.   :   90   Min.   : 2.00  
##  1st Qu.:19.00   1st Qu.:1.780   1st Qu.:  479   1st Qu.:39.50  
##  Median :40.50   Median :2.600   Median : 2046   Median :49.00  
##  Mean   :38.72   Mean   :2.618   Mean   : 6527   Mean   :48.31  
##  3rd Qu.:55.00   3rd Qu.:3.470   3rd Qu.: 8461   3rd Qu.:58.00  
##  Max.   :83.00   Max.   :5.370   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
```

```
powerTransform(UN,family="bcPower")
```

```
## Estimated transformation parameters 
##     ModernC  Change_add       PPgdp       Frate         Pop   Fertility 
##  0.87069484  0.93338011 -0.15621030  1.09144998  0.06285445  0.18829460 
##      Purban 
##  0.92703643
```

```
UN2 <- UN %>%
  mutate(logPPgdp = log(PPgdp),
         logPop = log(Pop),
         logFertility = log(Fertility))  %>%
  select(ModernC,Change_add,logPPgdp,Frate,logPop,logFertility,Purban)
```

Checking summary of UN3, I can find out that Chnage variable has minimum negative value -1.1. Thus I decide to add 1.2 on Change. Since ModernC,Frate, Purban, Change_add have optimal value for lamda which is approximately 1, they don't need to be transformed. However, in the case of Pop, PPgdp, and Fertility, they have optimal value for lamda which is approximately 0. Thus they are required to be log transformed.

7. Given the selected transformations of the predictors, select a transformation of the response using `MASS::boxcox` or `car::boxCox` and justify. Do you need to do this if you used `car::powerTransform` above? Explain.

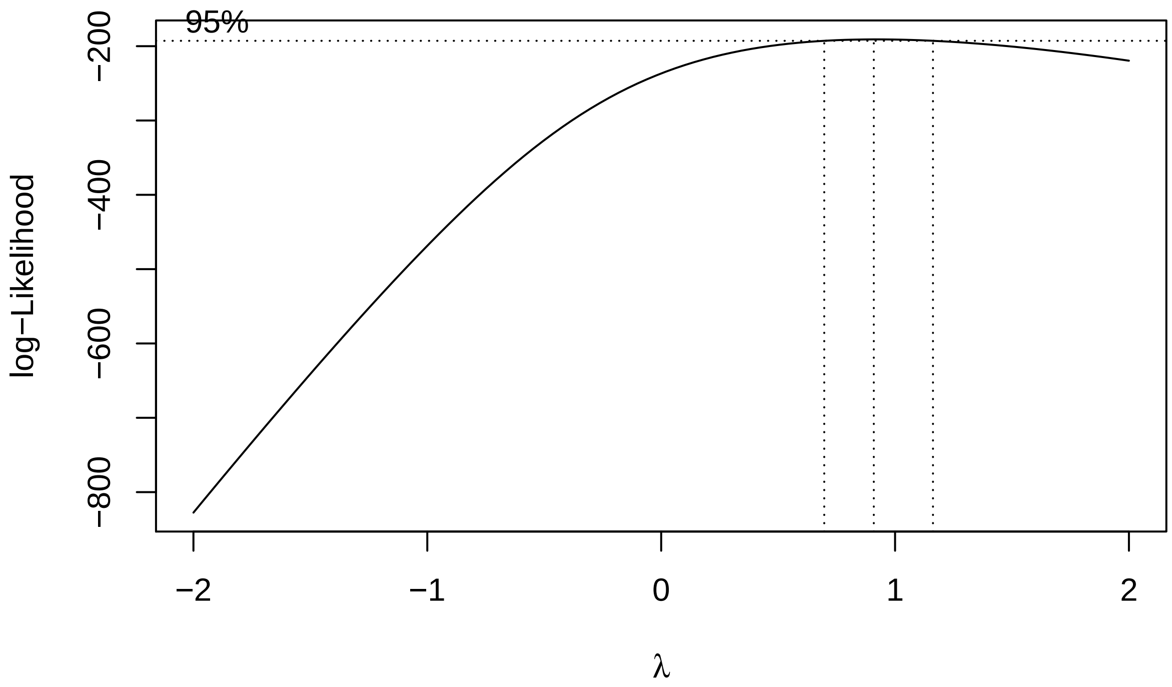
```
library(MASS)
```

```
## 
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
## 
##     select
```

```
## The following object is masked from 'package:alr3':
## 
##     forbes
```

```
boxcox(lm(ModernC~.,data=UN2))
```



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.

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

Summary of Results

10. 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!

11. 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.

Methodology

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

13. Exercise 9.12 from ALR

Using $X^TX = X_{(i)}^T X_{(i)} + x_i x_i^T$ where the subscript (i) means without the ith case, show that

$$(X_{(i)}^T X_{(i)})^{-1} = (X^TX)^{-1} + \frac{(X^TX)^{-1} x_i x_i^T (X^TX)^{-1}}{1 - h_{ii}}$$

where h_{ii} is the ith diagonal element of $H = X(X^TX)^{-1}X^T$ using direct multiplication and simplify in terms of_ h_{ii}.

13. Exercise 9.13 from ALR. Using the above, show

$$\hat{\beta}_{(i)} = \hat{\beta} - \frac{(X^TX)^{-1} x_i e_i}{1 - h_{ii}}$$