HW2 STA521 Fall18

[Dhanasekar Sundararaman, ds448 and Dhanasekar-S]

Due September 23, 2018 5pm

Backgound Reading

Readings: Chapters 3-4 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?

All the variables have missing data and all of them can be termed quantitative.

```
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                      
```

```r
is.na(UN3)
```

```
##                         ModernC Change PPgdp Frate   Pop Fertility Purban
## Afghanistan                TRUE  FALSE FALSE  TRUE FALSE     FALSE  FALSE
## Albania                    TRUE  FALSE FALSE  TRUE FALSE     FALSE  FALSE
## Algeria                   FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Am.Samoa                   TRUE  FALSE  TRUE FALSE FALSE      TRUE  FALSE
## Andorra                    TRUE  FALSE FALSE  TRUE FALSE      TRUE  FALSE
## Angola                     TRUE  FALSE FALSE  TRUE FALSE     FALSE  FALSE
## Antigua.and.Barbuda       FALSE  FALSE FALSE FALSE FALSE      TRUE  FALSE
## Argentina                  TRUE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Armenia                   FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Aruba                      TRUE  FALSE  TRUE FALSE FALSE      TRUE  FALSE
## Australia                 FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Austria                   FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Azerbaijan                FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Bahamas                   FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Bahrain                   FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Bangladesh                FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Barbados                  FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Belarus                   FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Belgium                   FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Belize                    FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Benin                     FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Bermuda                    TRUE  FALSE FALSE  TRUE FALSE     FALSE  FALSE
## Bhutan                    FALSE  FALSE FALSE  TRUE FALSE     FALSE  FALSE
## Bolivia                   FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Bosnia-Herzegovina         TRUE  FALSE FALSE  TRUE FALSE     FALSE  FALSE
## Botswana                  FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Brazil                    FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Brunei                     TRUE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Bulgaria                  FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Burkina.Faso              FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Burundi                   FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Cambodia                  FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Cameroon                  FALSE  FALSE FALSE  TRUE FALSE     FALSE  FALSE
## Canada                    FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Cape.Verde                FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Central.African.Rep       FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Chad                      FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Chile                      TRUE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## China                     FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Hong.Kong                 FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Macao                      TRUE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Colombia                  FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Comoros                   FALSE  FALSE FALSE  TRUE FALSE     FALSE  FALSE
## Congo                      TRUE  FALSE FALSE  TRUE FALSE     FALSE  FALSE
## Cook.Islands              FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Costa.Rica                FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Cote.dIvoire              FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Croatia                    TRUE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Cuba                      FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Cyprus                     TRUE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Czech.Rep                 FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
```

```
## .Congo                  FALSE  FALSE FALSE  TRUE FALSE     FALSE  FALSE
## Denmark                 FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Djibouti                 TRUE  FALSE FALSE  TRUE FALSE     FALSE  FALSE
## Dominica                FALSE  FALSE FALSE FALSE FALSE      TRUE  FALSE
## Dominican.Rep           FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Ecuador                 FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Egypt                   FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## El.Salvador             FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Equatorial.Guinea        TRUE  FALSE FALSE  TRUE FALSE     FALSE  FALSE
## Eritrea                 FALSE  FALSE FALSE  TRUE FALSE     FALSE  FALSE
## Estonia                 FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Ethiopia                FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Fiji                    FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Finland                 FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## France                  FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Fr.Guiana                TRUE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Fr.Polynesia             TRUE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Gabon                   FALSE  FALSE FALSE  TRUE FALSE     FALSE  FALSE
## Gambia                  FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Georgia                 FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Germany                 FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Ghana                   FALSE  FALSE FALSE  TRUE FALSE     FALSE  FALSE
## Greece                   TRUE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Grenada                  TRUE  FALSE FALSE FALSE FALSE      TRUE  FALSE
## Guadeloupe               TRUE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Guam                     TRUE  FALSE  TRUE FALSE FALSE     FALSE  FALSE
## Guatemala               FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Guinea                  FALSE  FALSE FALSE  TRUE FALSE     FALSE  FALSE
## Guinea-Bissau            TRUE  FALSE FALSE  TRUE FALSE     FALSE  FALSE
## Guyana                   TRUE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Haiti                   FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Honduras                FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Hungary                 FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Iceland                  TRUE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## India                   FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Indonesia               FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Iran                    FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Iraq                    FALSE  FALSE  TRUE FALSE FALSE     FALSE  FALSE
## Ireland                  TRUE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Israel                   TRUE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Italy                   FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Jamaica                 FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Japan                   FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Jordan                  FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Kazakhstan              FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Kenya                   FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Kiribati                FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## N.Korea                 FALSE  FALSE FALSE  TRUE FALSE     FALSE  FALSE
## S.Korea                 FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Kuwait                  FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Kyrgyzstan              FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Laos                    FALSE  FALSE FALSE  TRUE FALSE     FALSE  FALSE
## Latvia                  FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Lebanon                 FALSE  FALSE FALSE  TRUE FALSE     FALSE  FALSE
```

```
## Lesotho                  FALSE  FALSE FALSE  TRUE FALSE     FALSE  FALSE
## Liberia                  FALSE  FALSE FALSE  TRUE FALSE     FALSE  FALSE
## Libya                    FALSE  FALSE FALSE  TRUE FALSE     FALSE  FALSE
## Liechtenstein             TRUE   TRUE FALSE  TRUE  TRUE     FALSE  FALSE
## Lithuania                FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Luxembourg                TRUE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Madagascar               FALSE  FALSE FALSE  TRUE FALSE     FALSE  FALSE
## Malawi                   FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Malaysia                 FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Maldives                  TRUE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Mali                     FALSE  FALSE FALSE  TRUE FALSE     FALSE  FALSE
## Malta                     TRUE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Marshall.Is              FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Martinique                TRUE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Mauritania               FALSE  FALSE FALSE  TRUE FALSE     FALSE  FALSE
## Mauritius                FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Mexico                   FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Micronesia                TRUE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Monaco                    TRUE  FALSE FALSE  TRUE FALSE      TRUE  FALSE
## Mongolia                 FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Morocco                  FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Mozambique               FALSE  FALSE FALSE  TRUE FALSE     FALSE  FALSE
## Myanmar                  FALSE  FALSE  TRUE  TRUE FALSE     FALSE  FALSE
## Namibia                  FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Nauru                     TRUE  FALSE FALSE  TRUE  TRUE      TRUE  FALSE
## Nepal                    FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Netherlands              FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Neth.Antilles             TRUE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## New.Caledonia             TRUE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## New.Zealand              FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Nicaragua                FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Niger                    FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Nigeria                  FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## N.Mariana.Is              TRUE  FALSE  TRUE FALSE FALSE     FALSE  FALSE
## Norway                   FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Occ.Palestinian.Terr.     TRUE  FALSE  TRUE FALSE FALSE     FALSE  FALSE
## Oman                     FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Pakistan                 FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Palau                     TRUE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Panama                   FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Papua.New.Guinea         FALSE  FALSE FALSE  TRUE FALSE     FALSE  FALSE
## Paraguay                 FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Peru                     FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Philippines              FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Poland                   FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Portugal                 FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Puerto.Rico              FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Qatar                    FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Moldova                  FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Reunion                  FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Romania                  FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Russia                   FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Rwanda                   FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Saint.Kitts.and.Nevis    FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
```

```
## Saint.Lucia              FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## St.Vincent/Grenadines    FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Samoa                    FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## San.Marino                TRUE  FALSE FALSE FALSE FALSE      TRUE  FALSE
## Sao.Tome.and.Principe     TRUE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Saudi.Arabia             FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Senegal                  FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Serbia.and.Montenegro.    TRUE  FALSE FALSE  TRUE FALSE     FALSE  FALSE
## Seychelles                TRUE  FALSE FALSE  TRUE FALSE     FALSE  FALSE
## Sierra.Leone              TRUE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Singapore                FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Slovakia                 FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Slovenia                  TRUE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Solomon.Islands           TRUE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Somalia                   TRUE  FALSE FALSE  TRUE FALSE     FALSE  FALSE
## South.Africa             FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Spain                    FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Sri.Lanka                FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Sudan                    FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Suriname                  TRUE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Swaziland                FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Sweden                   FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Switzerland              FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Syria                    FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Tajikistan                TRUE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Thailand                 FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Macedonia                 TRUE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Timor-Leste               TRUE  FALSE FALSE  TRUE FALSE     FALSE  FALSE
## Togo                     FALSE  FALSE FALSE  TRUE FALSE     FALSE  FALSE
## Tonga                     TRUE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Trinidad.and.Tobago      FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Tunisia                  FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Turkey                   FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Turkmenistan             FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Tuvalu                    TRUE  FALSE FALSE  TRUE FALSE      TRUE  FALSE
## Uganda                   FALSE  FALSE FALSE  TRUE FALSE     FALSE  FALSE
## Ukraine                  FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## United.Arab.Emirates     FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## United.Kingdom           FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Tanzania                 FALSE  FALSE FALSE  TRUE FALSE     FALSE  FALSE
## USA                      FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## USVirgin.Islands          TRUE  FALSE  TRUE FALSE FALSE     FALSE  FALSE
## Uruguay                   TRUE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Uzbekistan               FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Vanuatu                  FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Venezuela                FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Viet.Nam                 FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Western.Sahara            TRUE  FALSE  TRUE  TRUE FALSE     FALSE  FALSE
## Yemen                    FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
## Zambia                   FALSE  FALSE FALSE  TRUE FALSE     FALSE  FALSE
## Zimbabwe                 FALSE  FALSE FALSE FALSE FALSE     FALSE  FALSE
```

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

```r
library(knitr)
df1 <- c(mean(UN3$ModernC,na.rm = TRUE),sd(UN3$ModernC,na.rm = TRUE))
df2 <- c(mean(UN3$Change,na.rm = TRUE),sd(UN3$Change,na.rm = TRUE))
df3 <- c(mean(UN3$PPgdp,na.rm = TRUE),sd(UN3$PPgdp,na.rm = TRUE))
df4 <- c(mean(UN3$Frate,na.rm = TRUE),sd(UN3$Frate,na.rm = TRUE))
df5 <- c(mean(UN3$Pop,na.rm = TRUE),sd(UN3$Pop,na.rm = TRUE))
df6 <- c(mean(UN3$Fertility,na.rm = TRUE),sd(UN3$Fertility,na.rm = TRUE))
df7 <- c(mean(UN3$Purban,na.rm = TRUE),sd(UN3$Purban,na.rm = TRUE))

df = data.frame(df1,df2,df3,df4,df5,df6,df7)
colnames(df) <- c("ModernC","Change","PPgdp","Frate","Pop","Fertility","Purban")
df <- cbind(Row.Names = c("mean","sd"), df)
df <- t(df)
kable(df)
```

Row.Names	mean	sd
ModernC	38.71711	22.63661
Change	1.418373	1.133133
PPgdp	6527.388	9325.189
Frate	48.30539	16.53245
Pop	30281.87	120676.69
Fertility	3.214000	1.706918
Purban	56.20000	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?

Upon investigating the data visually with GGPlot, I found that there are certain variables that needs to be transformed. The PPgdp and Pop data is skewed and hence needs a transformation.

```r
library(GGally)
```

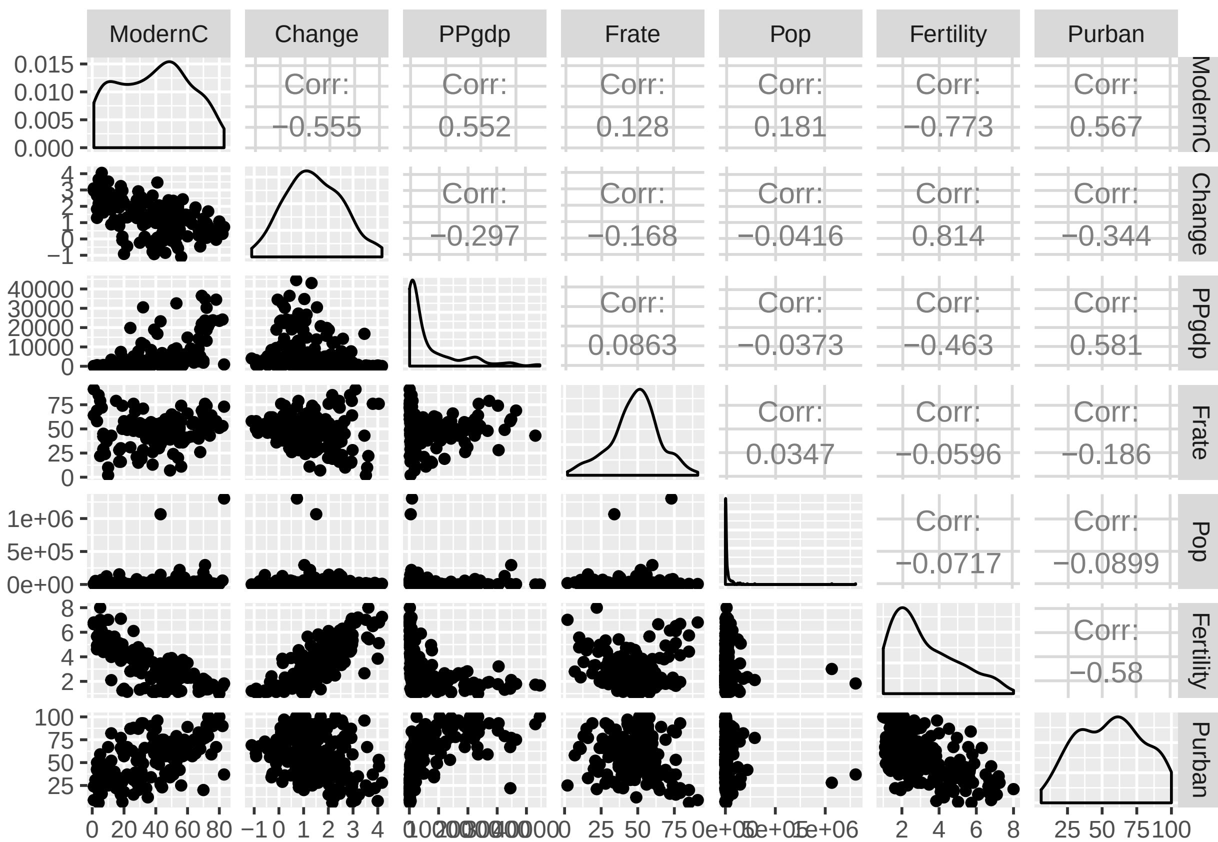
```
## Loading required package: ggplot2
```

```r
ggpairs(UN3,columns <- c(1,2,3,4,5,6,7))
```

```
## Warning: Removed 58 rows containing non-finite values (stat_density).

## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 58 rows containing missing values

## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 60 rows containing missing values

## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 82 rows containing missing values

## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 58 rows containing missing values

## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 60 rows containing missing values

## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 58 rows containing missing values
```

```
## Warning: Removed 58 rows containing missing values (geom_point).

## Warning: Removed 1 rows containing non-finite values (stat_density).

## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 10 rows containing missing values

## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 43 rows containing missing values

## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 2 rows containing missing values

## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 11 rows containing missing values

## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removing 1 row that contained a missing value

## Warning: Removed 60 rows containing missing values (geom_point).

## Warning: Removed 10 rows containing missing values (geom_point).

## Warning: Removed 9 rows containing non-finite values (stat_density).

## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 50 rows containing missing values

## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 11 rows containing missing values

## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 17 rows containing missing values

## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 9 rows containing missing values

## Warning: Removed 82 rows containing missing values (geom_point).

## Warning: Removed 43 rows containing missing values (geom_point).

## Warning: Removed 50 rows containing missing values (geom_point).

## Warning: Removed 43 rows containing non-finite values (stat_density).

## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 43 rows containing missing values

## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 49 rows containing missing values

## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 43 rows containing missing values

## Warning: Removed 58 rows containing missing values (geom_point).

## Warning: Removed 2 rows containing missing values (geom_point).

## Warning: Removed 11 rows containing missing values (geom_point).

## Warning: Removed 43 rows containing missing values (geom_point).

## Warning: Removed 2 rows containing non-finite values (stat_density).

## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 11 rows containing missing values
```

```
## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 2 rows containing missing values

## Warning: Removed 60 rows containing missing values (geom_point).

## Warning: Removed 11 rows containing missing values (geom_point).

## Warning: Removed 17 rows containing missing values (geom_point).

## Warning: Removed 49 rows containing missing values (geom_point).

## Warning: Removed 11 rows containing missing values (geom_point).

## Warning: Removed 10 rows containing non-finite values (stat_density).

## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 10 rows containing missing values

## Warning: Removed 58 rows containing missing values (geom_point).

## Warning: Removed 1 rows containing missing values (geom_point).

## Warning: Removed 9 rows containing missing values (geom_point).

## Warning: Removed 43 rows containing missing values (geom_point).

## Warning: Removed 2 rows containing missing values (geom_point).

## Warning: Removed 10 rows containing missing values (geom_point).
```



Model Fitting

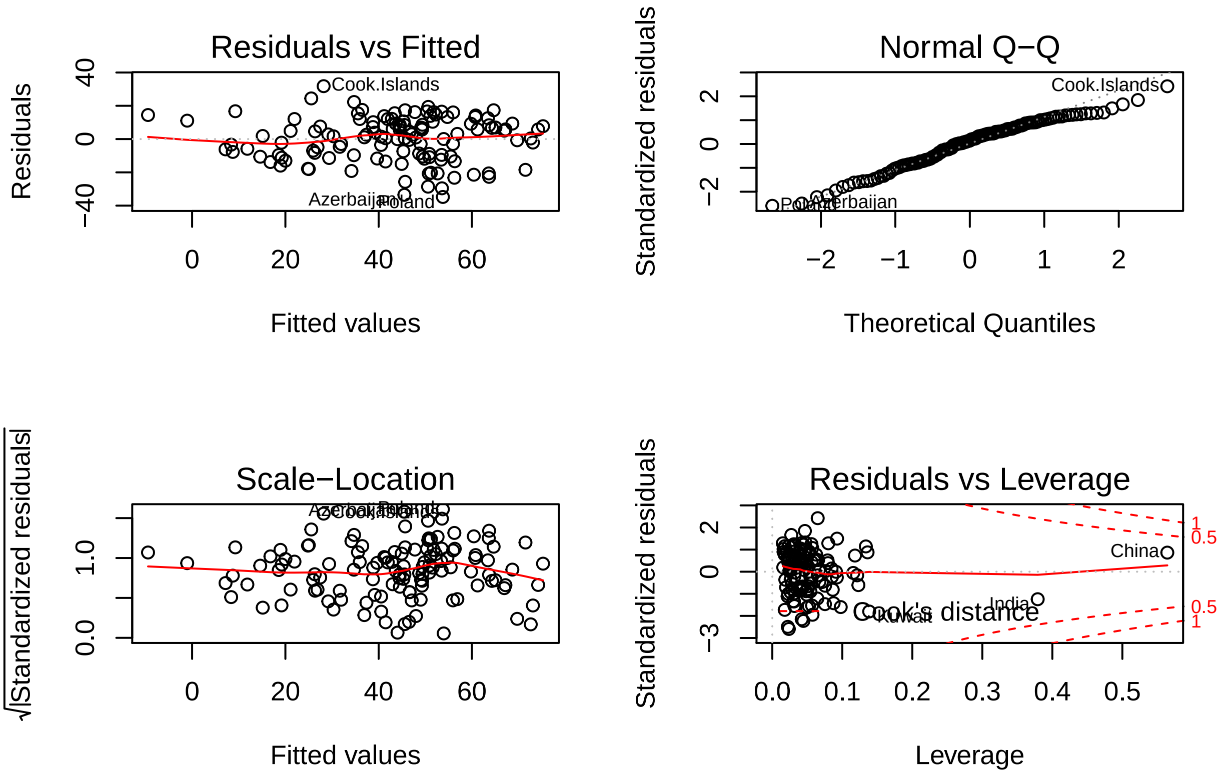
4. Use the `lm()` function to perform a multiple linear regression with `ModernC` as the response and all other variables as the predictors, using the formula `ModernC ~ .`, where the `.` includes all remaining variables in the dataframe. Create diagnostic residual plot from the linear model object and comment on results regarding assumptions. How many observations are used in your model fitting?

The linear model was created with Y as ModernC and X with all the other variables. There were 85 data points missing due to NA, the rest were used in the model.

```
model.lm <-lm(ModernC~., data <-UN3)
summary(model.lm)
```

```
##
## Call:
## lm(formula = ModernC ~ ., data = 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
```

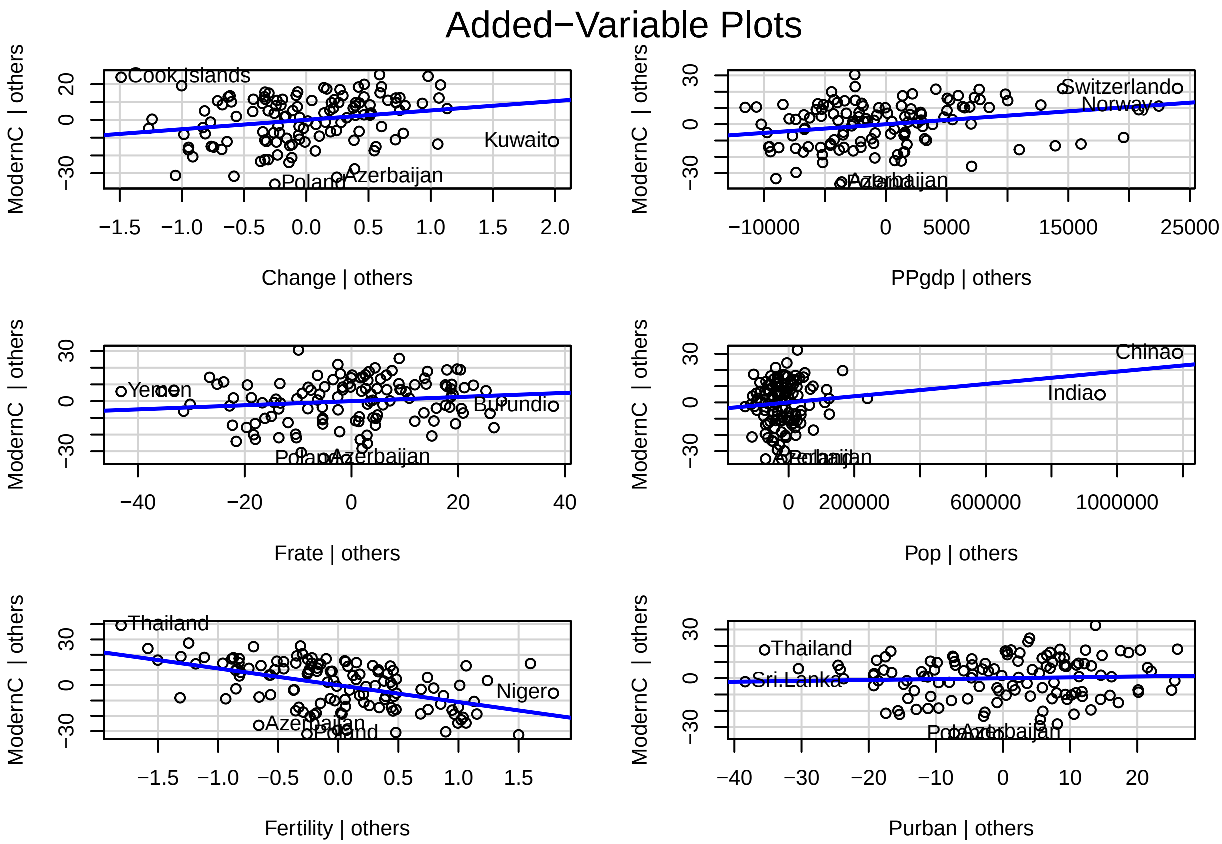
```
par(mfrow=c(2,2))
plot(model.lm)
```



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?

The GGplot and scatterplots suggest that PPgdp and Pop variables require a log transformation. They are skewed to the right and hence a log transformation can make it look better.

```
car::avPlots(model.lm)
```



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.

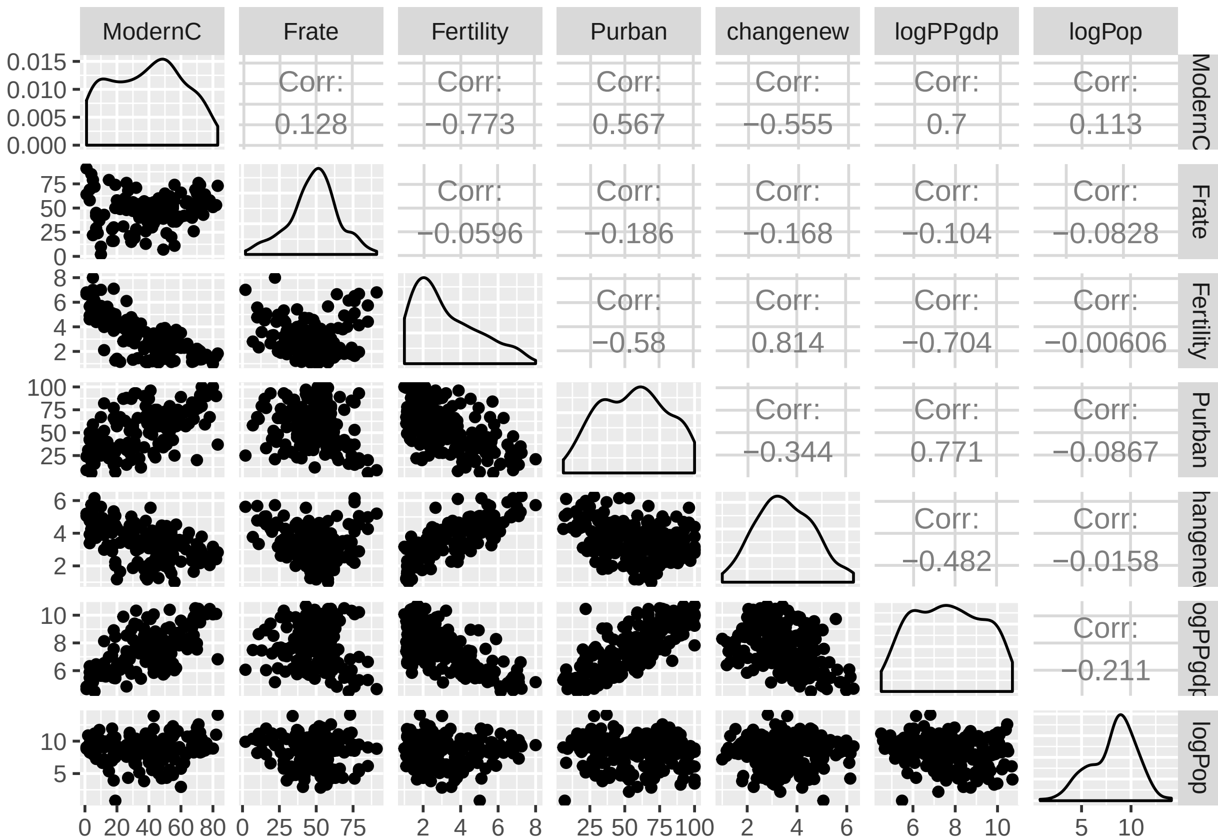
The 'Change' variable has negative values. One way to get rid of this is to subtract all the values from the minimum value and add a constant. PPgdp and Pop ariables are log transformed.

```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
```

```r
UN = UN3 %>%
  mutate(changenew = Change+ 1 - min(Change, na.rm = TRUE),
```

```
        logPPgdp = log(PPgdp),
        logPop = log(Pop)) %>%
        select(-c("Change","Pop","PPgdp"))

ggpairs(UN)
```



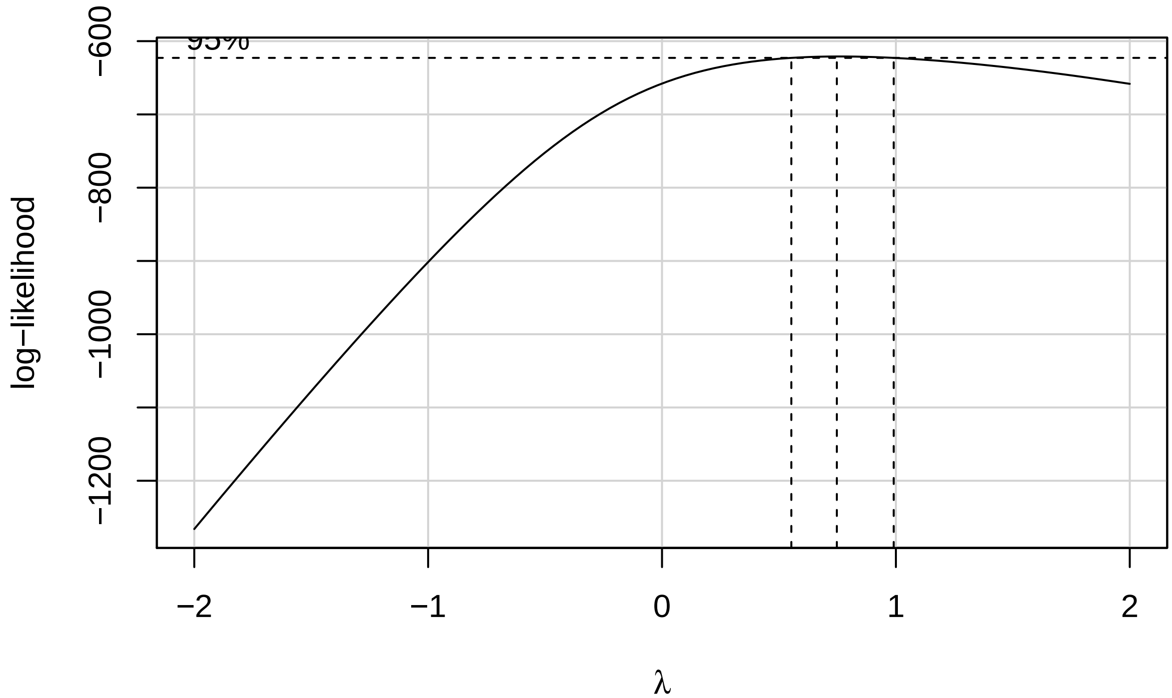

```
boxTidwell(ModernC ~ Pop, ~ + Change + PPgdp + Fertility + Purban + Frate,data = UN3)
```

```
##  MLE of lambda Score Statistic (z) Pr(>|z|)
##        0.63309             -0.5543   0.5794
##
## iterations =  3
```

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

The lambda value for boxcox is around 1, which tells that the model is performing well.

```
model2 <-lm(ModernC ~logPop + changenew + logPPgdp + Fertility + Purban + Frate, data = UN )
boxCox(model2)
```

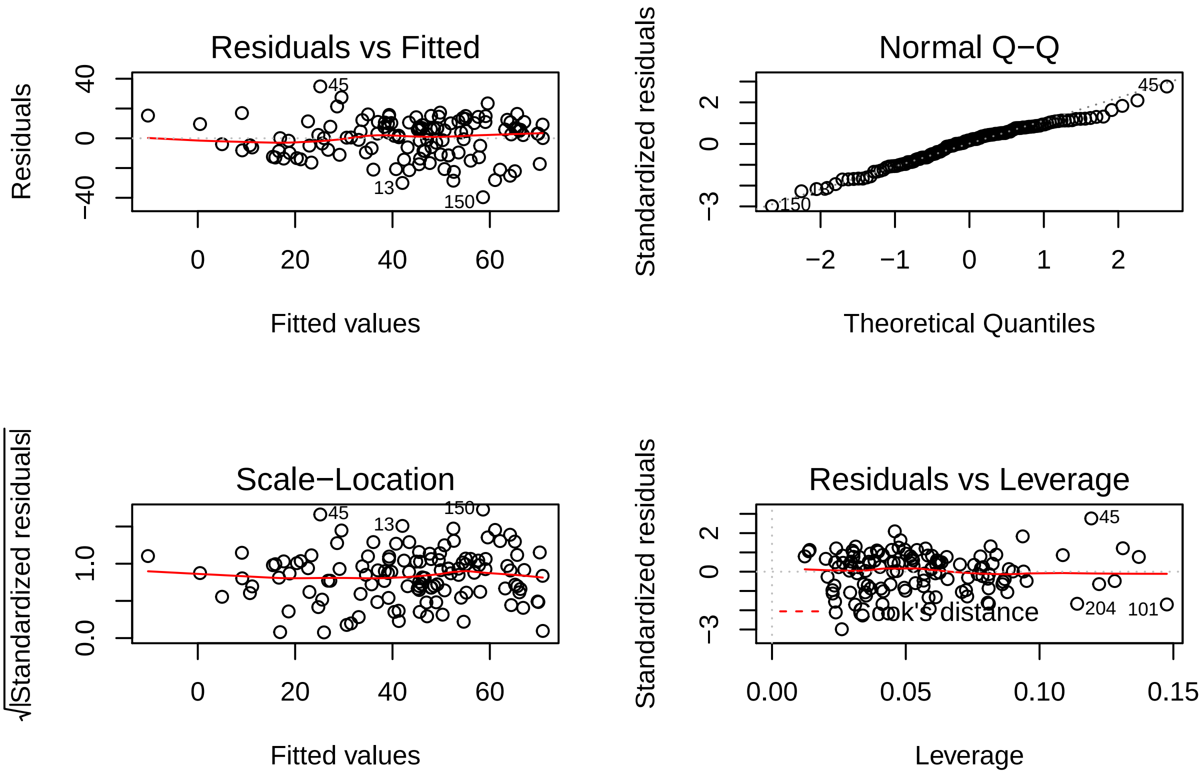


8. Fit the regression using the transformed variables. Provide residual plots and added variables plots and comment. If you feel that you need additional transformations of either the response or predictors, repeat any steps until you feel satisfied.

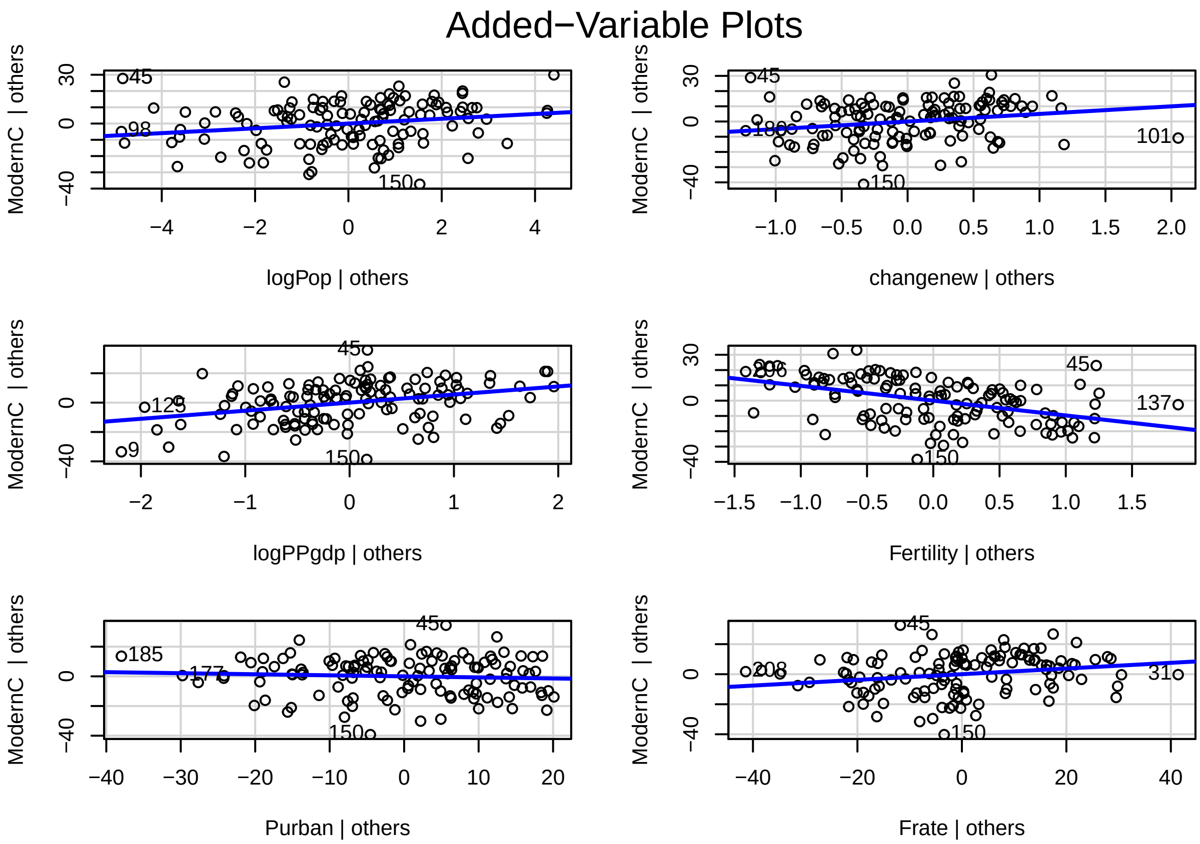
```
summary(model2)
```

```
##
## Call:
## lm(formula = ModernC ~ logPop + changenew + logPPgdp + Fertility +
##     Purban + Frate, data = UN)
##
## 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) -6.36974   14.22449  -0.448 0.655118
## logPop       1.47207    0.62875   2.341 0.020897 *
## changenew    4.99296    2.07709   2.404 0.017781 *
## logPPgdp     5.50728    1.40505   3.920 0.000149 ***
## Fertility   -9.67594    1.76561  -5.480 2.44e-07 ***
## Purban      -0.07077    0.09760  -0.725 0.469829
## Frate        0.18939    0.07711   2.456 0.015500 *
## ---
## 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
```

```r
par(mfrow = c(2,2))
plot(model2)
```




```r
avPlots(model2)
```



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?

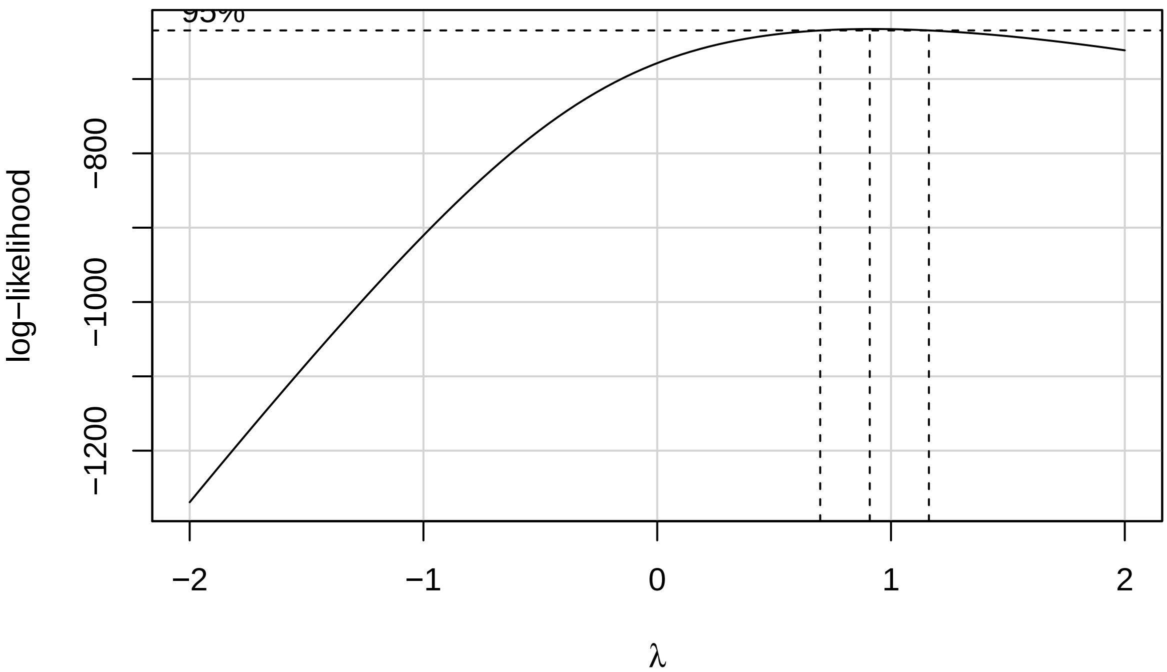
Another variant of the model, with log transformation for variable 'Fertility' was also created and the result was stored in a temporary model which shows a different summary of the model.

```
modeltemp <-lm(ModernC ~logPop + changenew + logPPgdp + log(Fertility) + Purban + Frate, data = UN )
summary(modeltemp)
```

```
##
## Call:
## lm(formula = ModernC ~ logPop + changenew + logPPgdp + log(Fertility) +
##     Purban + Frate, data = UN)
##
## Residuals:
##     Min       1Q  Median      3Q     Max
## -41.235 -11.589   2.498  10.748  31.954
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -19.958066  15.253221  -1.308  0.19326
## logPop           1.595611   0.699289   2.282  0.02430 *
## changenew        2.310274   2.560728   0.902  0.36879
## logPPgdp         6.445713   1.508057   4.274 3.91e-05 ***
## log(Fertility) -18.237639   6.336680  -2.878  0.00475 **
## Purban          -0.007352   0.106591  -0.069  0.94513
## Frate            0.178242   0.083567   2.133  0.03500 *
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.55 on 118 degrees of freedom
##   (85 observations deleted due to missingness)
## Multiple R-squared:  0.5615, Adjusted R-squared:  0.5392
## F-statistic: 25.19 on 6 and 118 DF,  p-value: < 2.2e-16
```

```r
boxCox(modeltemp)
```



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.

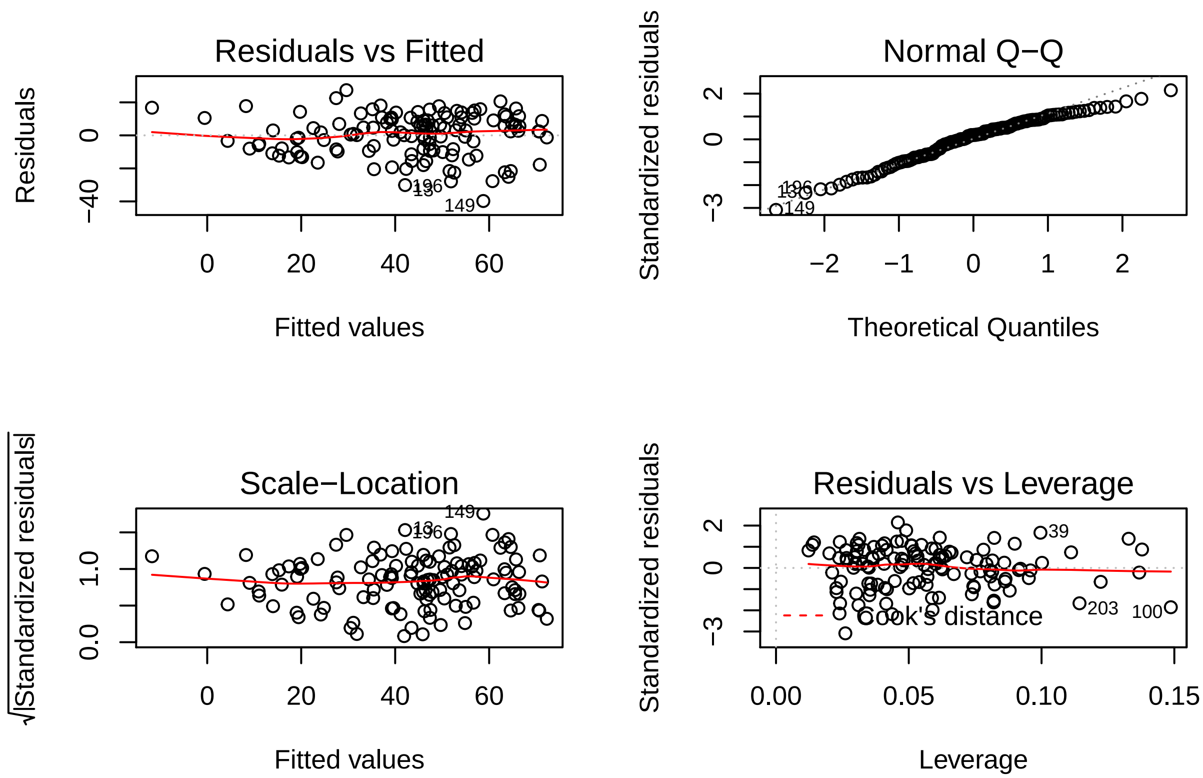
There are some outliers in the data set. Especially data point 45 seems to be a clear outlier. That data point is removed and then the model is refit without that data point and the residual plots are significantly better. The scales in the plots have changed slightly which also changed the cook's distance.

```r
UNs = UN %>% slice(-45)
model3 <-lm(ModernC ~logPop + changenew + logPPgdp + Fertility + Purban + Frate, data = UNs )

summary(model3)
```

```
##
## Call:
## lm(formula = ModernC ~ logPop + changenew + logPPgdp + Fertility +
##     Purban + Frate, data = UNs)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -39.760  -9.209   2.442   9.791  27.380
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -11.24467   13.92178  -0.808  0.42090
## logPop        1.89052    0.62817   3.010  0.00320 **
## changenew     6.11720    2.05580   2.976  0.00355 **
```

```
## logPPgdp       5.43342    1.36492   3.981  0.00012 ***
## Fertility    -10.51515    1.74010  -6.043 1.85e-08 ***
## Purban        -0.08248    0.09488  -0.869  0.38646
## Frate          0.20474    0.07509   2.727  0.00738 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.06 on 117 degrees of freedom
##   (85 observations deleted due to missingness)
## Multiple R-squared:  0.6484, Adjusted R-squared:  0.6304
## F-statistic: 35.96 on 6 and 117 DF,  p-value: < 2.2e-16
```

```r
par(mfrow = c(2,2))
plot(model3)
```



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!

The confint command returns the confidence intervals of the various variables. A 95% confidence interval gives the interval range of a variable 95% of the times. logpop 2.5% confidence interval suggests that 0.64 is the value 2.5% of the times and 3.13 is the value 97.5 % of the times. The same way for all other variables. x <- exp(logPop) gives the value in original units by taking exponent, since we did log transformation.

```r
confint(model3, level = 0.95)
```

```
##                   2.5 %     97.5 %
## (Intercept) -38.81603363 16.3266877
## logPop        0.64646112  3.1345845
```

```
## changenew      2.04579991 10.1885944
## logPPgdp       2.73025854  8.1365822
## Fertility    -13.96131706 -7.0689732
## Purban        -0.27039497  0.1054291
## Frate          0.05602591  0.3534451
```

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

The final model 'model3' has transformed variables 'Change', 'Pop', and 'PPgdp'. Change variable had negative values. It was transformed. 'Pop' and "PPgdp' had skewed scatterplots and hence was log transformed. An outlier was detected and hence was removed. The justification for removal of an outlier is that, that particular country may affect the model and hence the coefficients of all the other countries would have been affected. Hence a single country, even though important has to be removed to leave way for a better model.

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.

$ Y = b0 + b1X$

$eY = b0 + eX.b1$

$eY = b0 + (X^TX)^-1X^TY.(I-H)X$

$Substitute X as (I-H)Xj and simplifying$

$Xj^T(I-H)Y = Xj^Tb0 + Xj^TXj^T(I-H)Xj^-1.(Xj^T.(I-H)Y).(I-H)Xj$

$Taking (I-H)^2 as (I-H) and (I-H)^T as (I-H) and simplifying$

$Xj^T(I-H)Y = Xj^Tb0 + Xj^T(I-H)Y$

$Xj^T.b0 = 0$

$b0 = 0$

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. The residuals of Y was regressed with a lm with all other x's except Xj and Xj is regressed with all other x's. Finally they both are regressed with a lm to compare the coefficients of Xj.

```
e_Y = residuals(lm(ModernC ~changenew + logPPgdp + Fertility + Purban + Frate, data=UN[1:4,]))
e_X1 = residuals(lm(logPop ~ changenew + logPPgdp + Fertility + Purban + Frate, data=UN[1:4,]))
```