# HW2 STA521 Fall18

*Jiayi Ding, jd402, jiayid]*

*Due September 23, 2018 5pm*

## 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")
help(UN3)
library(car)
library(ggplot2)
library(GGally)
library(dplyr)
```

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

```
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                      
```

```
for(c in 1: dim(UN3)[2]){
  print (any(is.na(UN3[,c])))
}
```

```
## [1] TRUE
## [1] TRUE
## [1] TRUE
## [1] TRUE
## [1] TRUE
## [1] TRUE
## [1] FALSE
```

```
?UN3
```

###Answer: all variables except Purban have missing values. From the summary and descriptions of the dataset, it seems all of the variables are quantitative.

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

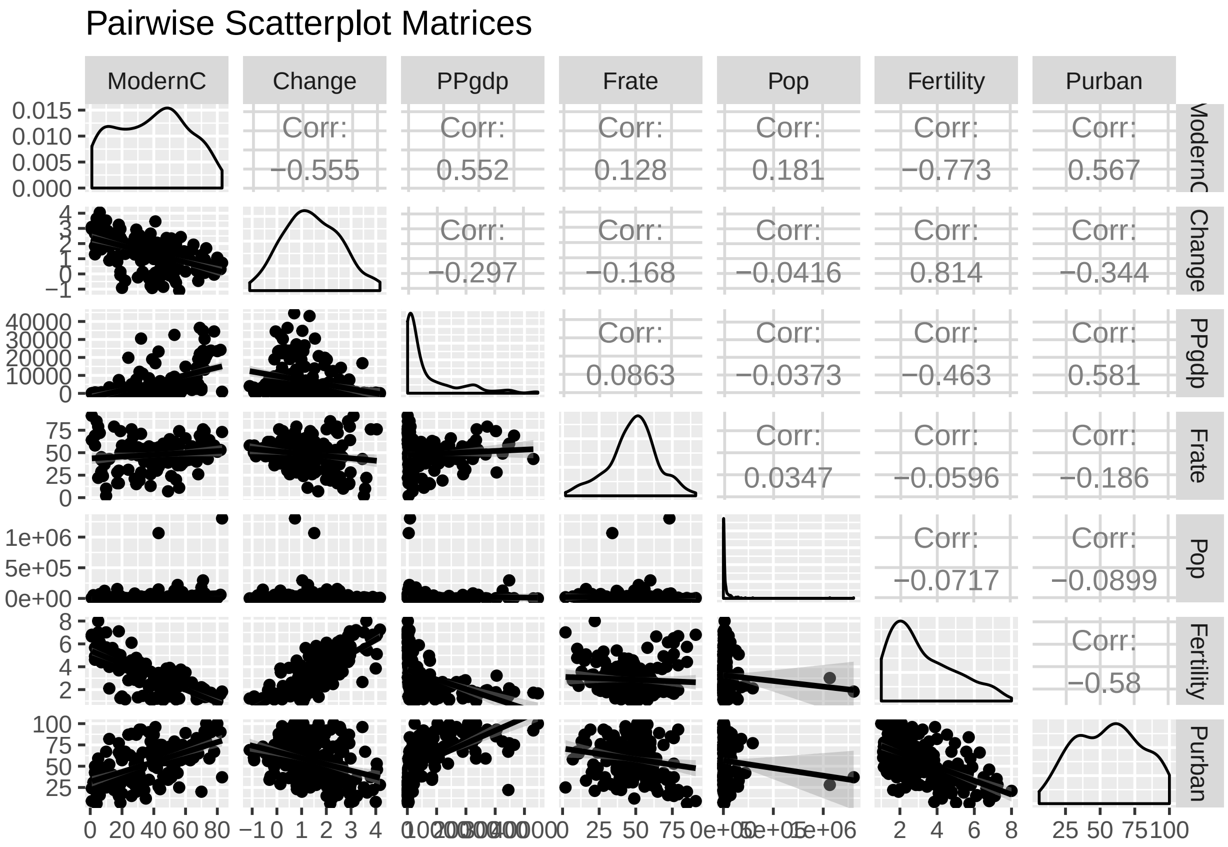
```
variable = colnames(UN3)
mean = sapply(UN3, function(x) mean(x, na.rm=TRUE))
sd = sapply(UN3, function(x) sd(x, na.rm = TRUE))

df<- data.frame(variable, mean, sd)
knitr:: kable(df, row.names = FALSE,
              caption = c("Mean and Standard Deviations of variables"))
```

Table 1: Mean and Standard Deviations of variables

| variable | mean | sd |
|:---|---:|---:|
| ModernC | 38.717105 | 2.263661e+01 |
| Change | 1.418373 | 1.133133e+00 |
| PPgdp | 6527.388060 | 9.325189e+03 |
| Frate | 48.305389 | 1.653245e+01 |
| Pop | 30281.871428 | 1.206767e+05 |
| Fertility | 3.214000 | 1.706918e+00 |
| Purban | 56.200000 | 2.410976e+01 |

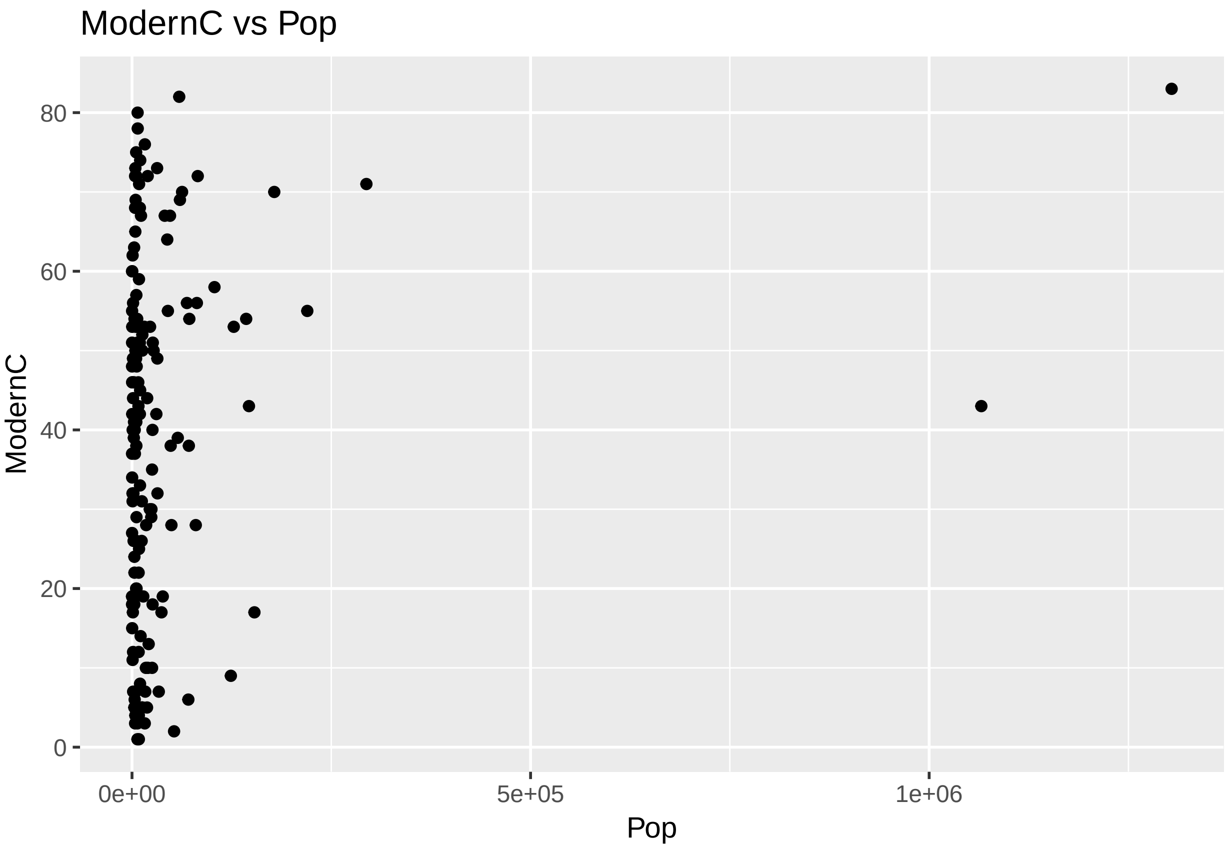
3. Investigate the predictors graphically, using scatterplots or other tools of your choice. Create some plots highlighting the relationships among the predictors. Comment on your findings regarding trying to predict `ModernC` from the other variables. Are there potential outliers, nonlinear relationships or transformations that appear to be needed based on your graphical EDA?

```
pm = ggpairs(UN3, title = "Pairwise Scatterplot Matrices", axisLabels = "show",
             columnLabels = colnames(UN3),lower = list(continuous = "smooth"))
print(pm)
```



Pairwise Scatterplot Matrices

```r
ggplot(UN3, aes(Pop, ModernC)) + geom_point() + ggtitle("ModernC vs Pop")
```

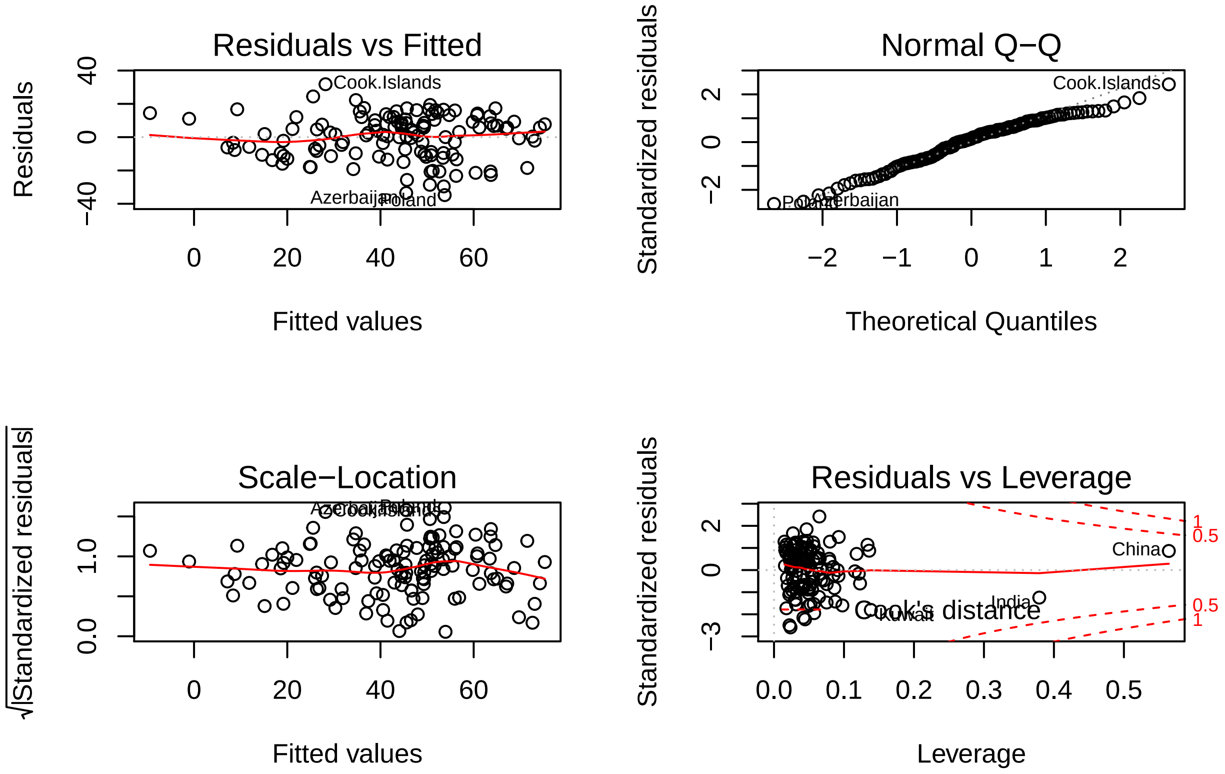


###Answer: Based on the plots, we find that Change, Fertility, and Purban with strong correlation with ModernC are great candidates for predicting ModernC. For variable PPgdp, its relationship with ModernC are not quite linear, and the quadratic pattern may suggest some kind of transformations. Frate and Pop have weak correlation with ModernC, which, however dosen't mean we should simply jump to conclusions that Frate or Pop don't have any predictive power on ModernC. Interestingly, Pop has a strongly skewed distribution, it might be the effects of the two potential outliers.

## 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<- na.omit(UN3)
lr<-lm(ModernC~.,Un3)
par(mfrow=c(2,2))
plot(lr)
```



```r
summary(lr)
```

```
##
## 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
## Multiple R-squared:  0.6183, Adjusted R-squared:  0.5989
## F-statistic: 31.85 on 6 and 118 DF,  p-value: < 2.2e-16
```

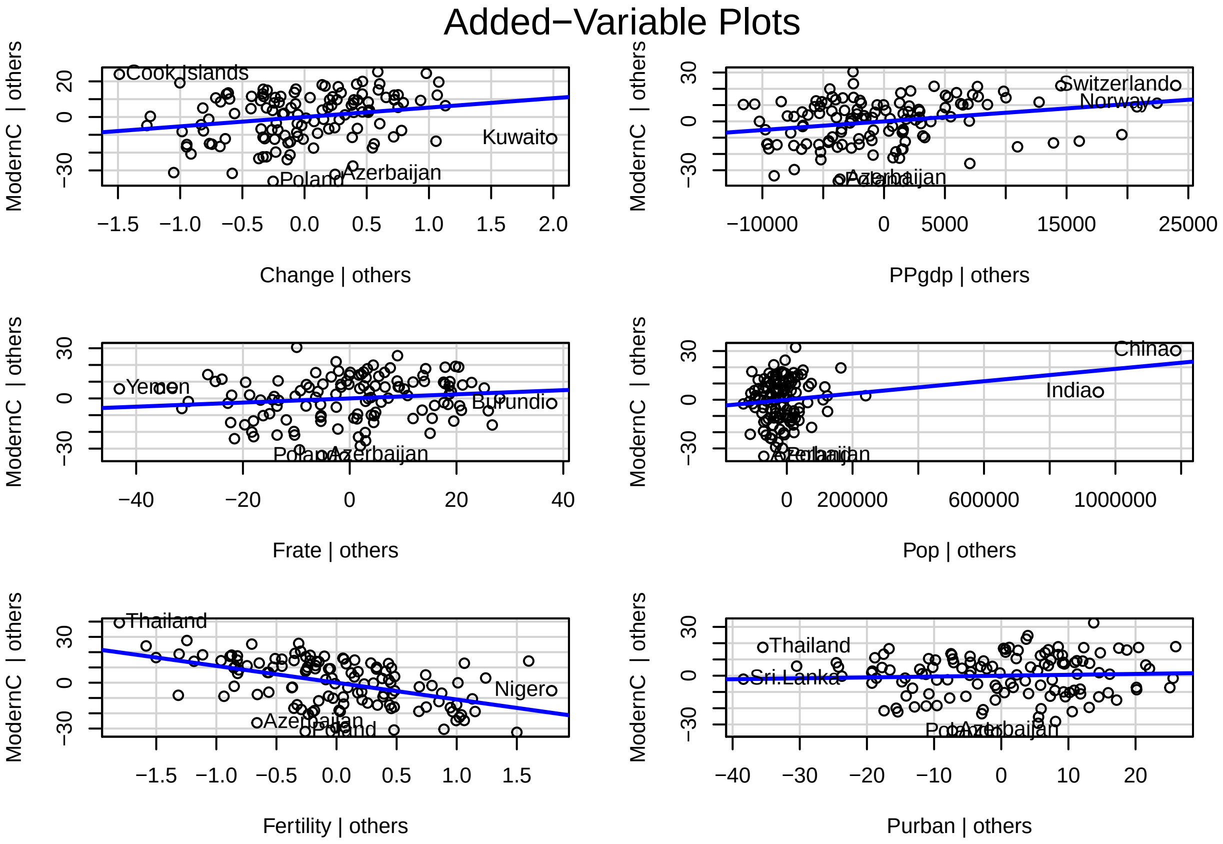
```r
nobs(lr)
```

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

###Answer: We can observe from the plots that residuals are not showing any obvious pattern and approximately centered around the 0 line, which is a support for our constant variance assumption. It might seem to be more variance in the center of the plot, but i think overall it looks fine. The Q-Q plot tells us that the residuals are approximately normallly distributed, except those with large theoretical quatiles seem to have a thinner tail. Though China and India and Kuwait have high leverage, the cook distance doesn't give us evidence for these points being influential. 125 observations are being used in model fitting.

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(lr)
```



###Answer: As we can see from the plot of ModernC and Pop, China and India are extremely far from the others. There is a huge difference between the populations of these two countries and the others, so a log transformation may be helpful in closing such "gap". PPgdp might also need log transformation to distribute more evenly.

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.

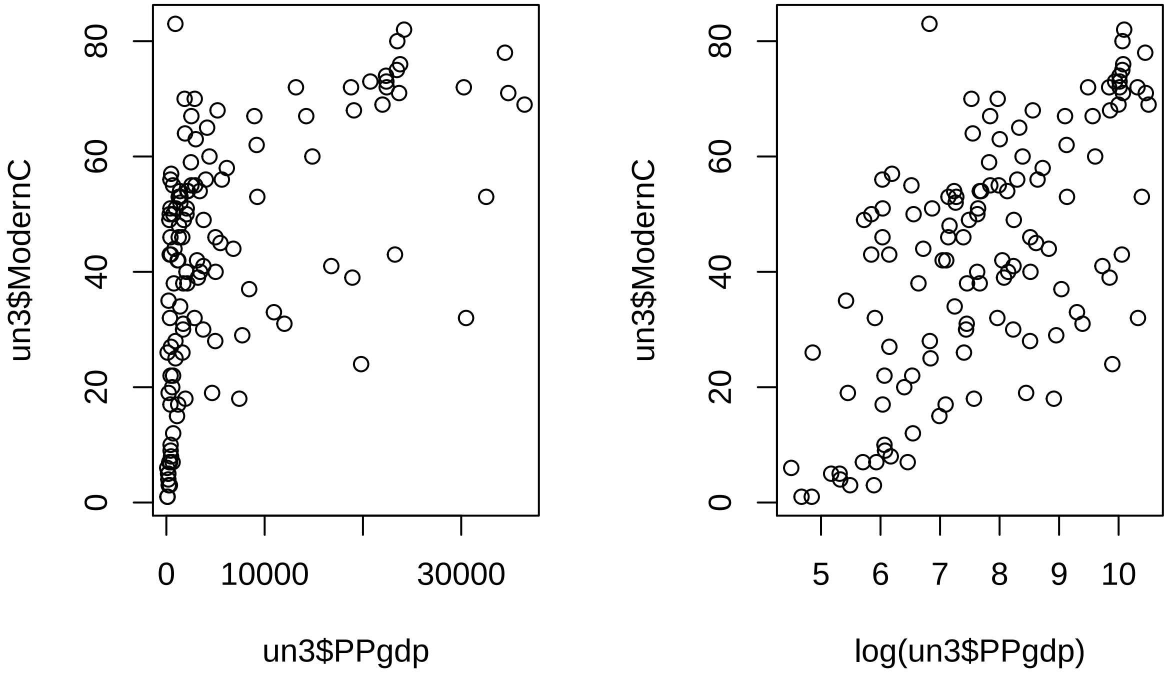
```
un3<- Un3
un3$Change = Un3$Change - min(Un3$Change) + 1
range(un3$Change)
```

```
## [1] 1.00 5.72
```

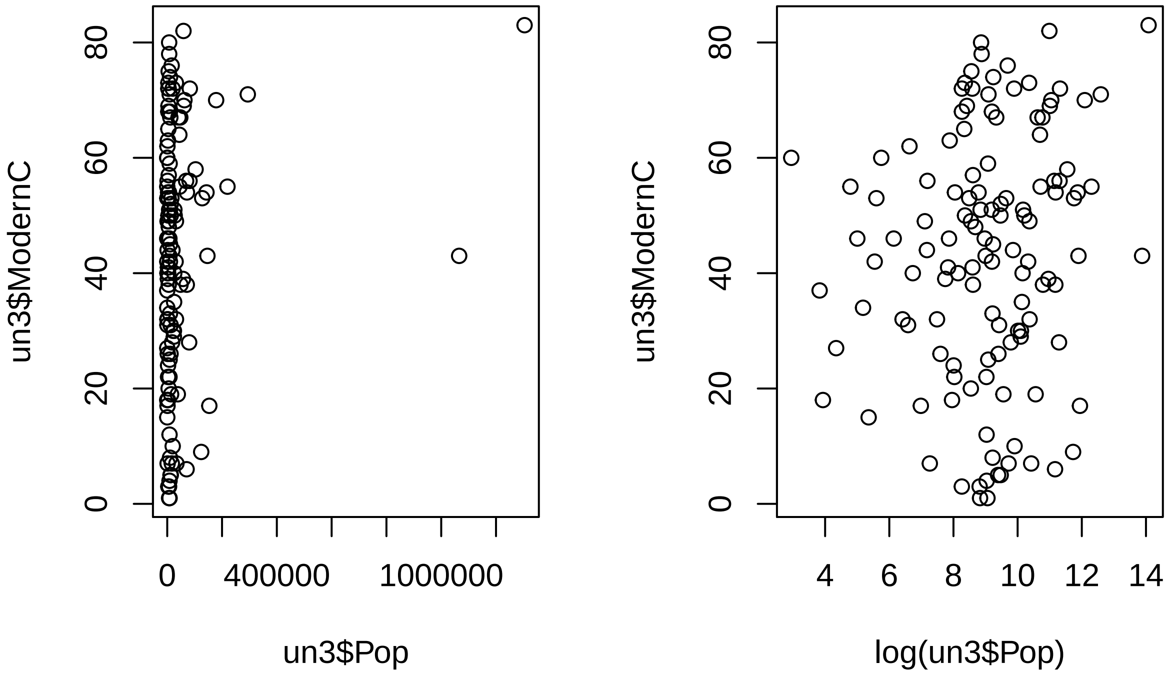
```
boxTidwell(ModernC~Pop+PPgdp, other.x = ~Frate+Fertility+Change+Purban,
           data=un3, max.iter = 100)
```

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

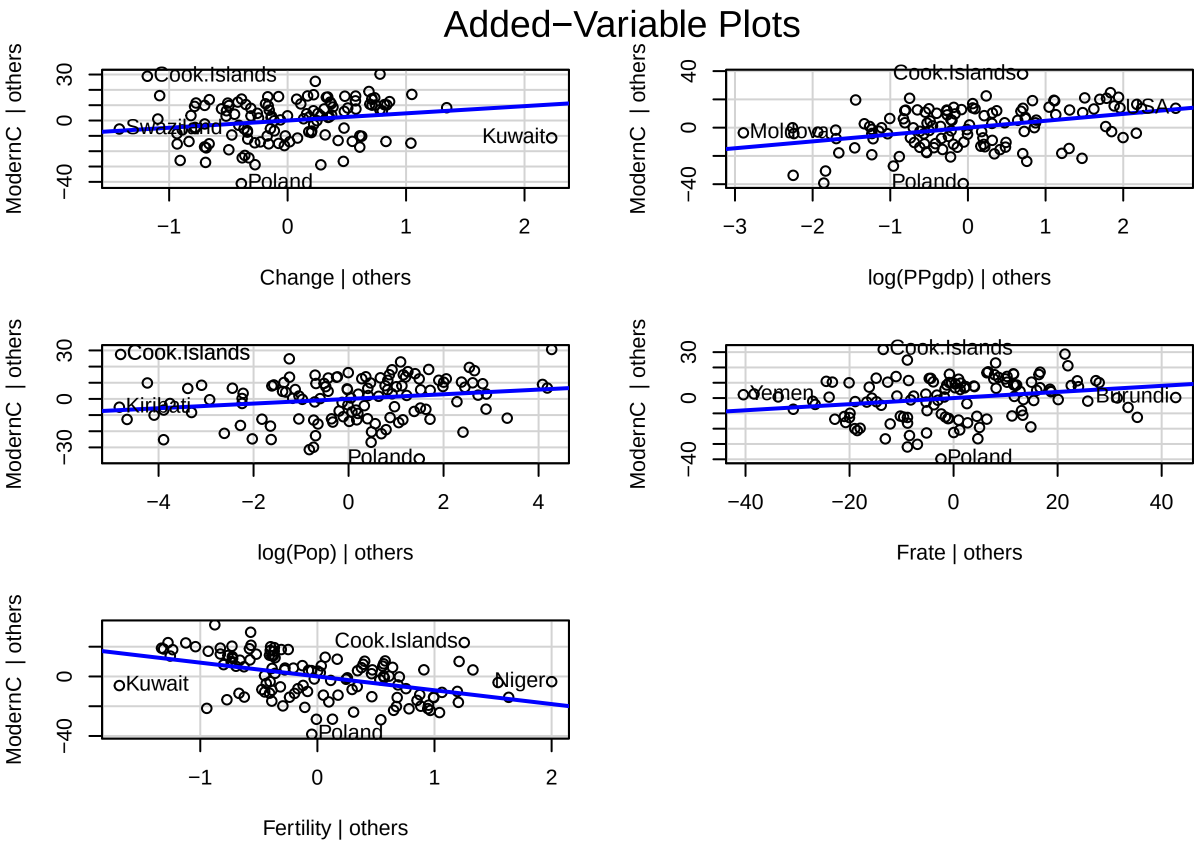
```
par(mfrow=c(1,2))
plot(un3$PPgdp, un3$ModernC)
plot(log(un3$PPgdp), un3$ModernC)
```



```
plot(un3$Pop, un3$ModernC)
plot(log(un3$Pop), un3$ModernC)
```



```r
lr2<- lm(ModernC ~ Change + log(PPgdp) + log(Pop) + Frate + Fertility, un3)
car:: avPlots(lr2)
```

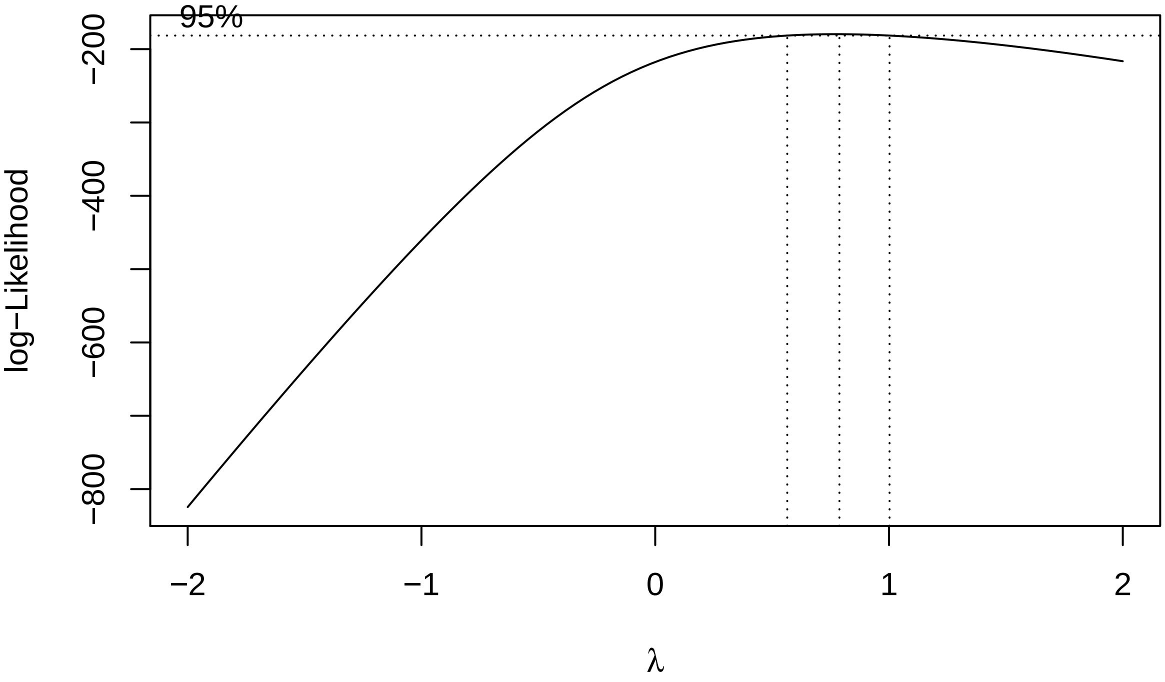


###Answer: From the avplots, “Purban” has little influence on “ModernC”. Thus, “Purban” will be excluded in the future model fitting.Based on previous analysis, I would like to transform predictors PPgdp and Pop. Since the minimal value Change is -1.1 negative, we will first convert it to non-negative by adding

2.1. Based on the Box-Tidewell results, we can roughly say that MLE (approximately equal to 0.5) of Pop suggests a square root transformation of Pop, while MLE(approximately equal to 0) of PPgdp suggests a log transformation of PPgdp. Though from the p-value perspective, it's not significant enough to reject the null $H_0 : \lambda = 1$ for Pop and PPgdp. Yet considering previous analysis and our background knowledge of the GDP per capita and Population among different countries, we can benefit from a log transformation on these two variables, as the linear relationship between Pop and PPgdp and ModernC becomes clearer from the plots.

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

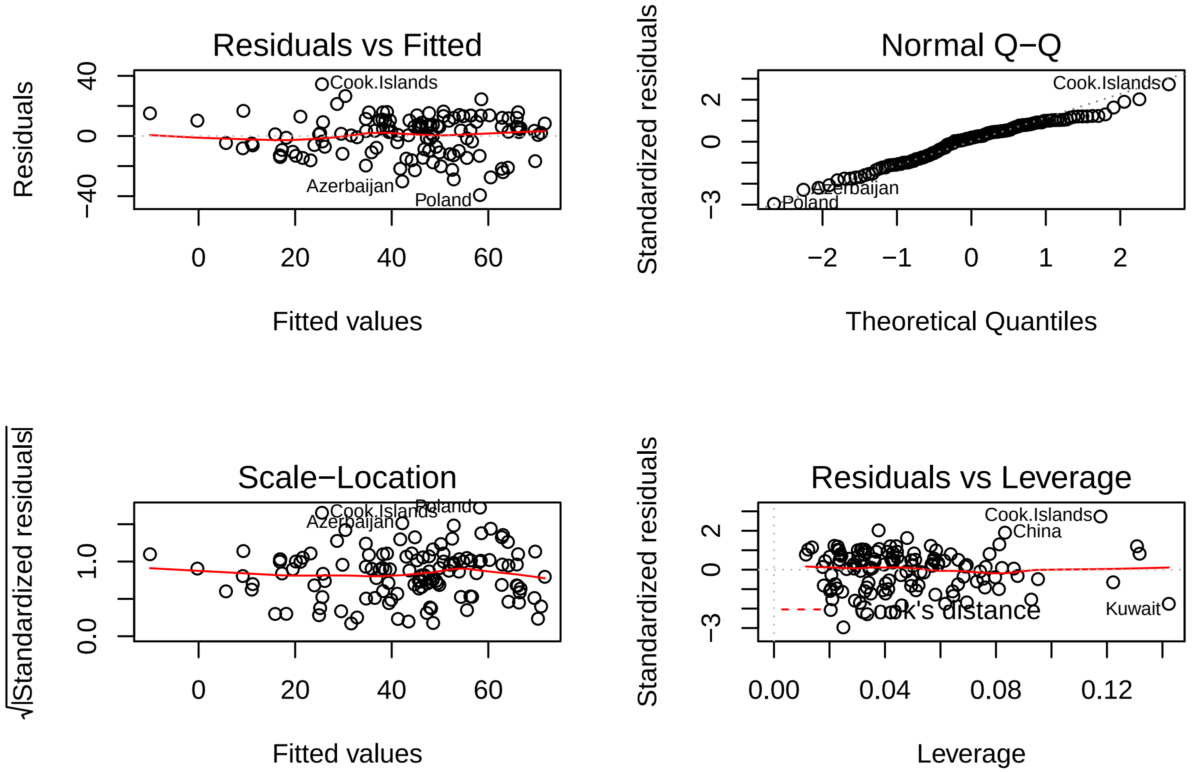
```
MASS::boxcox(lr2)
```



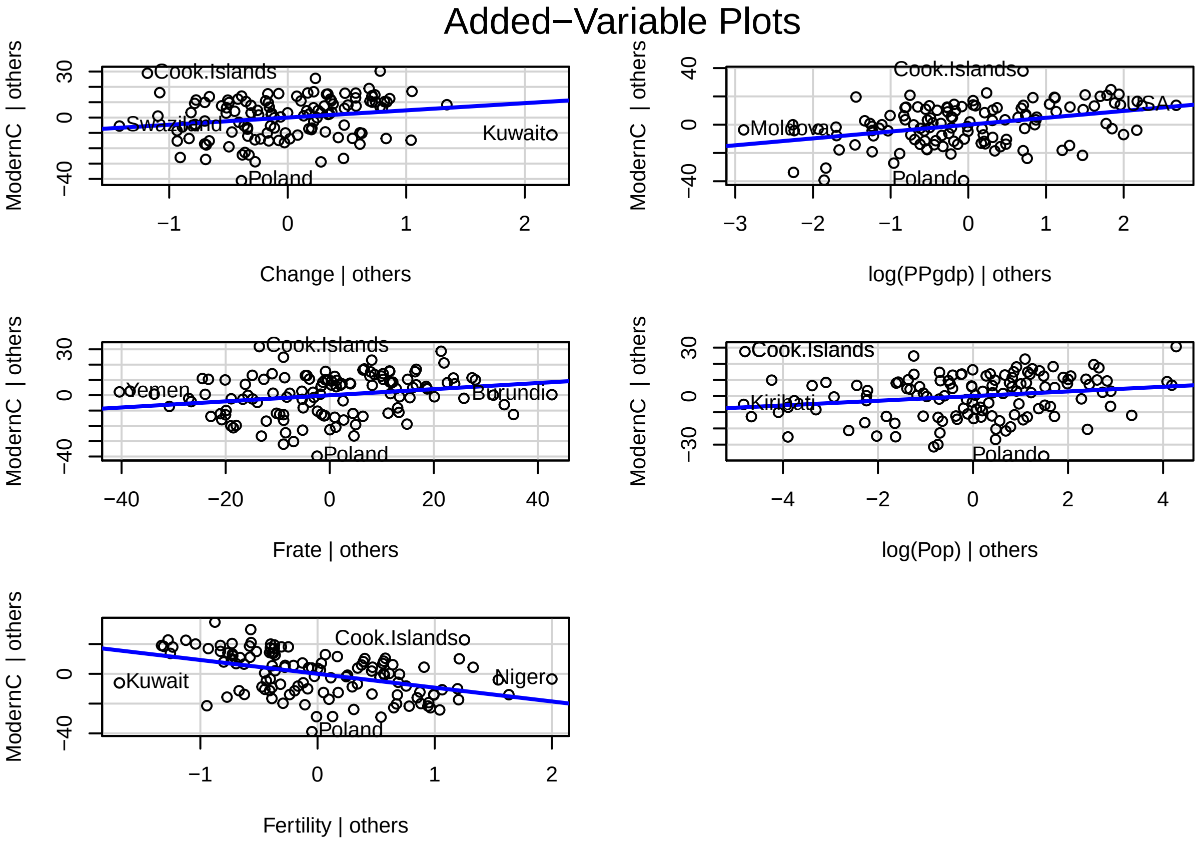
###Answer: After transforming Pop and PPgdp, we can see from the boxcox plot that a 95% CI for lambda is approximately (0.5, 1). For interpretability and simplicity, I would like to use lambda = 1, which does not transform the response variable.

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.

```
lr2<- lm(ModernC ~ Change+log(PPgdp)+Frate+log(Pop)+Fertility, un3)
par(mfrow=c(2,2))
plot(lr2)
```



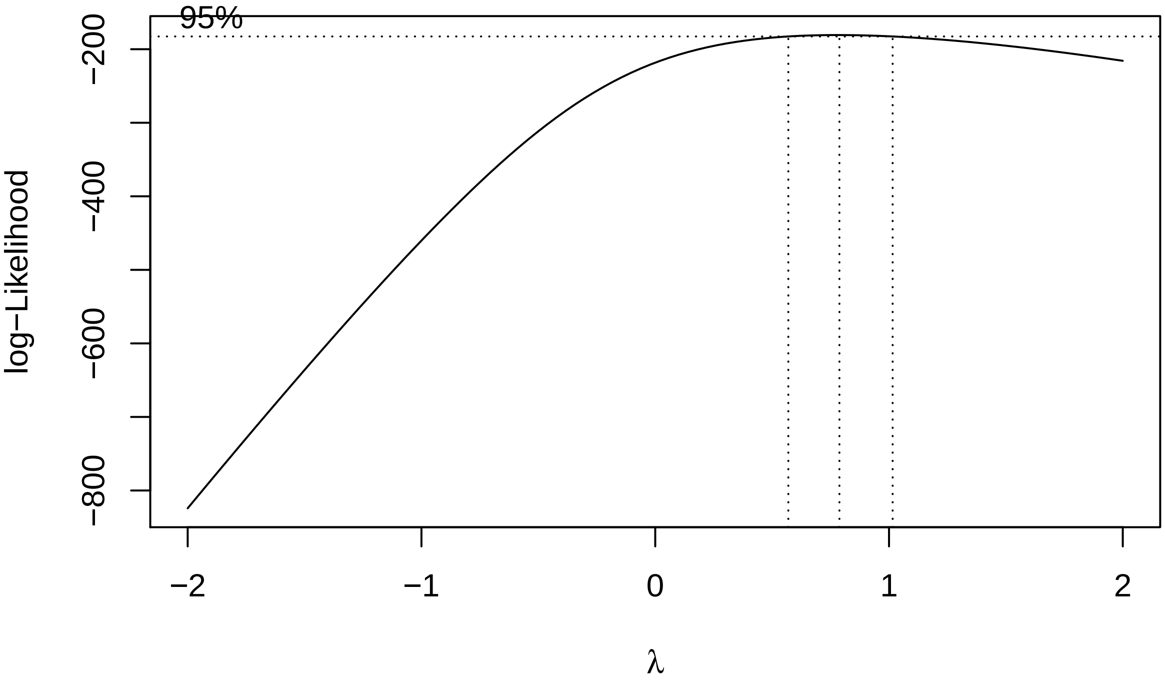
```
car::avPlots(lr2)
```



###Answer: The residuals are randomly and evenly distributed above and below the 0 line, condirming our assumption about constance variance and that the residuals are independent. The qq-plot generally indicates normally distributed standardized residuals.The log transformation reduced the leverage of previous worrisome high leverge points such as China and India. The added variable plots fit better than previously.

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?

```
MASS::boxcox(lr)
```



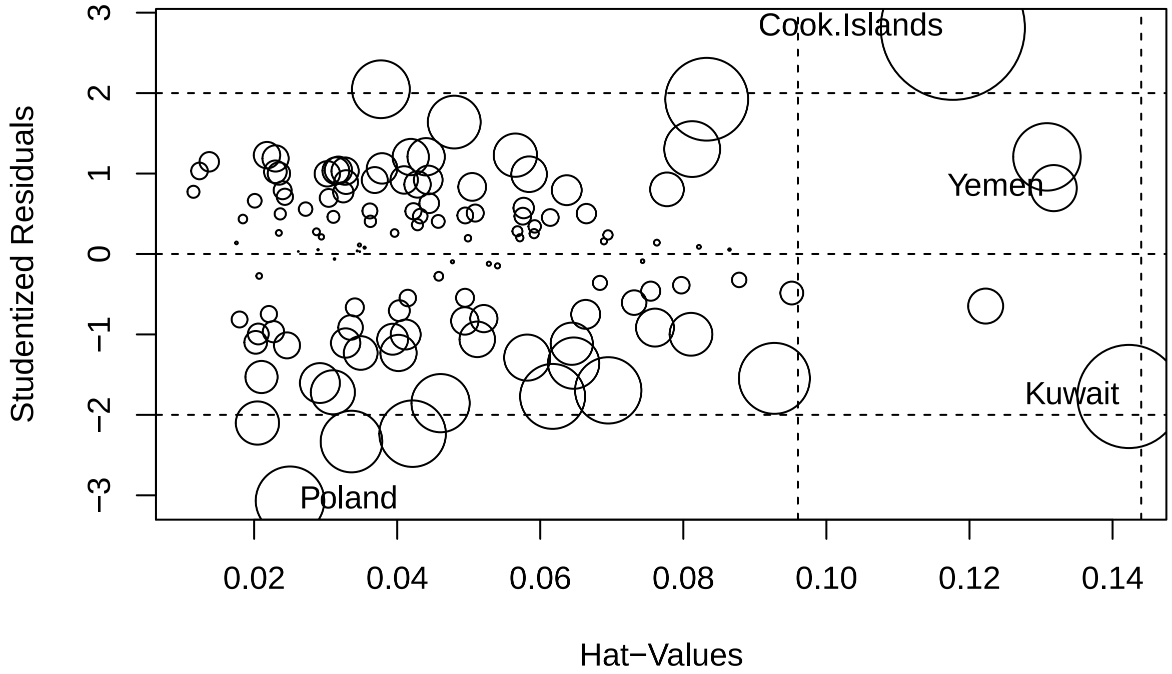
###Answer: Again with 1 in the interval, the Boxcox suggests no transformation for ModernC. Thus, boxTidwell will be doing the same job as before, and we get the same model.

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.

```
pval <- 2*(1-pt(abs(rstudent(lr2)), lr2$df-1))
criteria <- 0.05/nrow(un3)
mean(pval < criteria)
```

```
## [1] 0
```

```
influencePlot(lr2)
```



```
##                  StudRes        Hat      CookD
## Cook.Islands  2.8111289 0.11766863 0.16601679
## Kuwait       -1.7715300 0.14229068 0.08524080
## Poland       -3.0677469 0.02501099 0.03758018
## Yemen         0.8190358 0.13177110 0.01701545
```

###Answer: Based on Bonferroni correction, there is no outlier in the model. Though Cook.island, Yemen, Kuwait and Potland with relatively big influence on the model, the Cook Distances of whic are all within a reasonable range. Thus, no influncial observation.

## 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(lr2$coefficients, confint(lr2))
table <- table %>%
    rename(Coefficients = lr2.coefficients, "2.5%" = X2.5.., "97.5%" = X97.5..)
table
```

```
##             Coefficients         2.5%      97.5%
## (Intercept)    -5.763208 -33.82430137 22.2978848
## Change          4.697757   0.67272868  8.7227854
## log(PPgdp)      4.859362   2.71662357  7.0021013
## Frate           0.199546   0.04969767  0.3493943
## log(Pop)        1.441225   0.20155957  2.6808895
## Fertility      -9.278421 -12.59507558 -5.9617672
```

###Answer: Change: For every percent increase in annual population growth rate, the expected percentage of unmarried women using modern contraception increases by 4.70%.

PPgdp: When the 2001 GDP increases by 10%, the expected percentage of unmarried women using modern contraception increases by 0.46%.

Frate: For every percent increase in females over 15 who are economically active, the expected percentage of unmarried women using modern contraception increases by 0.2%

Pop: When the population increases by 10%, the expected percentage of unmarried women using modern contraception increases by 0.14%

Fertility: When the number of live births per female increases by 1, the expected percentage of unmarried women using modern contraception decreases by 9.28%.

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

```
summary(lr2)
```

```
##
## Call:
## lm(formula = ModernC ~ Change + log(PPgdp) + Frate + log(Pop) +
##     Fertility, data = un3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -39.276  -9.928   2.572  10.253  34.442
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.76321   14.17155  -0.407  0.68498
## Change       4.69776    2.03274   2.311  0.02255 *
## log(PPgdp)   4.85936    1.08214   4.491 1.65e-05 ***
## Frate        0.19955    0.07568   2.637  0.00949 **
## log(Pop)     1.44122    0.62606   2.302  0.02307 *
## Fertility   -9.27842    1.67499  -5.539 1.85e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.42 on 119 degrees of freedom
## Multiple R-squared:  0.6243, Adjusted R-squared:  0.6085
## F-statistic: 39.55 on 5 and 119 DF,  p-value: < 2.2e-16
```

###Answer: Our final model is: $ModernC = -5.7632 + 4.6978(Change) + 4.8594(log(PPgdp)) + 0.1996(Frate) + 1.4412(log(Pop)) - 9.2784(Fertility)$\$. The model intends to predict the percentage of women using modern contraception based on factors: population, annual population growth rate, per capita GDP, percent of females over age 15 economically active, and fertility(expected number of live births per female). Intuitively, fertility has a negative relationship with the use of Modern Contraception, i.e. more number of live births per female the less use of modern contraception. While the other factors all have a positive relationship with the use of modern contraception. For example, the higher percentage of females over 15 economically active, the more likely for them to use the modern contraception, probably due to the affordability of these modern contraception for them. Another finding is that countries with higher GDP's also have higher percentage of female using modern contraception. This makes sense as more developed a country is, it will have more developed technology and more well-established health system, which enables female in these countries more accessible for modern contraception. These findings are helpful when thinking about which countries one should devote more sources in helping and how. We excluded 85 observations due to missing values in certain datafield. But we did not delete any outlier or influential observation.

## 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.*

$$\begin{aligned}
e_i &= Y_i - \hat{Y}_i \\
e_{(Y)} &= (I-H)Y \\
H &= X(X^TX)^{-1}X^T \\
\beta_1 &= (X^TX)^{-1}X^TY \\
X &= (I-H)X_i \\
Y &= (I-H)Y
\end{aligned}$$

Thus we have the followings:

$$\begin{aligned}
(1-H)Y &= \hat{\beta}_0 I + \hat{\beta}_1(I-H)X_i \; (1-H)Y \\
&= \hat{\beta}_0 I + [X_i^T(I-H)^T(I-H)X_i]^{-1}((I-H)X_i)^T(I-H)Y(I-H)X_i \\
&= \hat{\beta}_0 I + (X_i^T(I-H)X_1)^{-1}X_i^T(I-H)Y(I-H)X_i \\
X_i^T(1-H)Y &= X_i^T\hat{\beta}_0 I + X_i(I-H)X_i^T(X_i^T(I-H)X_i)^{-1}X_i^T(I-H)YX_i^T\hat{\beta}_0 I \\
&= \sum_{j=1}^{n} X_{ij}\hat{\beta}_0 + X_i^T(I-H)Y \\
\sum_{j=1}^{n} x_{ij}\hat{\beta}_0 &= 0
\end{aligned}$$

Due to the fact that $\sum_{j=1}^{n} x_{ij}$ is a constant, we know $\hat{\beta}_0 = 0$.

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.

```r
e_Y <- residuals(lm(ModernC ~ Change + log(PPgdp) + log(Pop) + Frate, data=un3))
e_X <- residuals(lm(Fertility ~ Change + log(PPgdp) + log(Pop) + Frate, data=un3))
res<- data.frame(e_Y, e_X)
av <- lm(e_Y ~ e_X, data=res)
av$coef["e_X"]
```

```
##       e_X
## -9.278421
```

```r
lr2$coef["Fertility"]
```

```
## Fertility
## -9.278421
```

###Answer: As we can tell from the results that the coefficient of $e_X$ is the same as coefficient of "Fertility" in previous model.