HW2 STA521 Fall18

Zixi Wang, zw152, BillyWangwzx

Due September 23, 2018 5pm

Backgound Reading

Readings: Chapters 3-4 in Weisberg Applied Linear Regression

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(GGally)
```

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

```
library(alr3)
```

```
## Loading required package: car
## Loading required package: carData
```

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

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                      
```

There are total 7 varialbes and 6 them have missing data. All of them are quantitative.

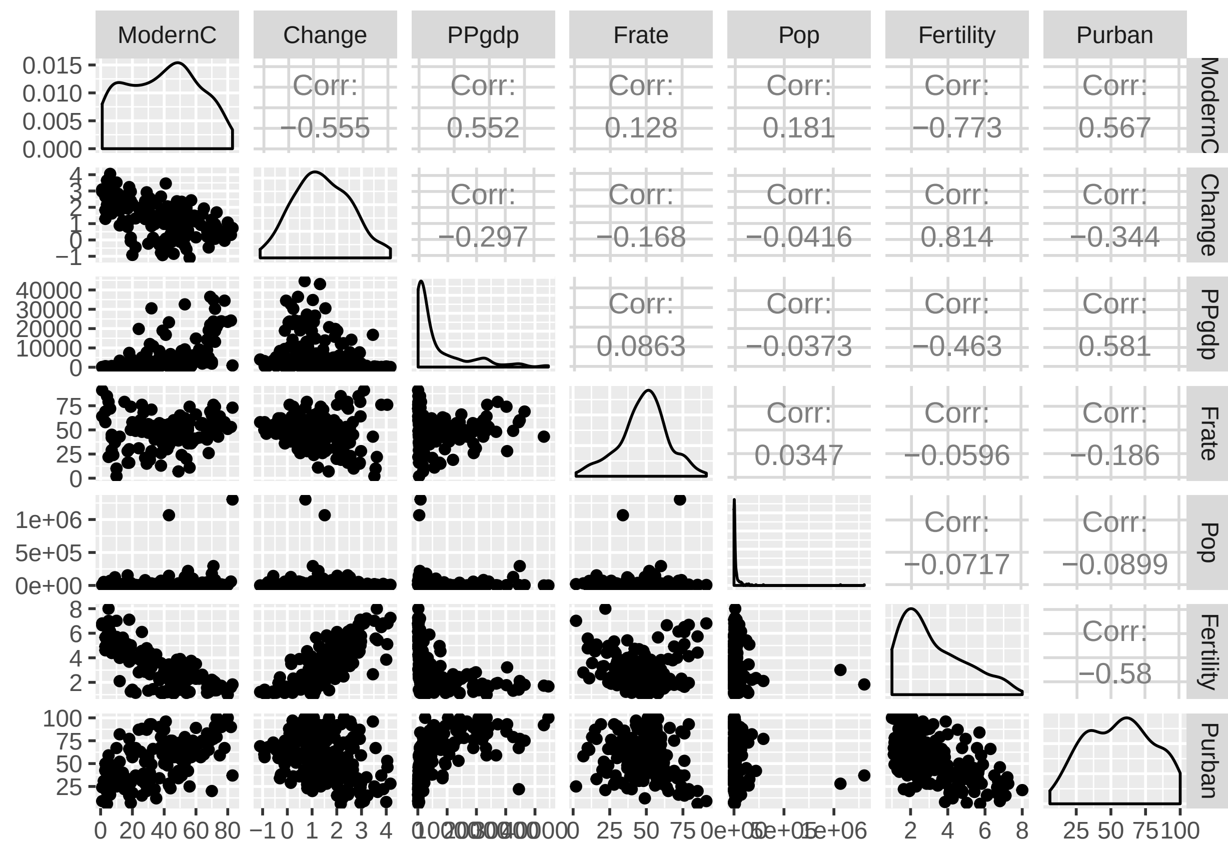
2. What is the mean and standard deviation of each quantitative predictor? Provide in a nicely formatted table.

```r
means <- colMeans(UN3,na.rm=TRUE)
sds <- sqrt(apply(UN3,2,function(x){var(x,na.rm = TRUE)}))
kable(rbind(means,sds))
```

	ModernC	Change	PPgdp	Frate	Pop	Fertility	Purban
means	38.71711	1.418373	6527.388	48.30539	30281.87	3.214000	56.20000
sds	22.63661	1.133133	9325.189	16.53245	120676.69	1.706918	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?

```r
ggpairs(UN3)
```



There is s quite obvious linear relationship between 'ModernC' and 'Change', 'Fertility', 'Purban'. It seems that 'Frate' can't explain anything about 'ModrenC'. We need to do some transformations on the 'PPgdp' and 'Pop' as the scales for these variables are so large that they doesn't show some linear relationship with 'ModernC'. And there are two countries, China and India, that have population seems to be portential outliers.

Model Fitting

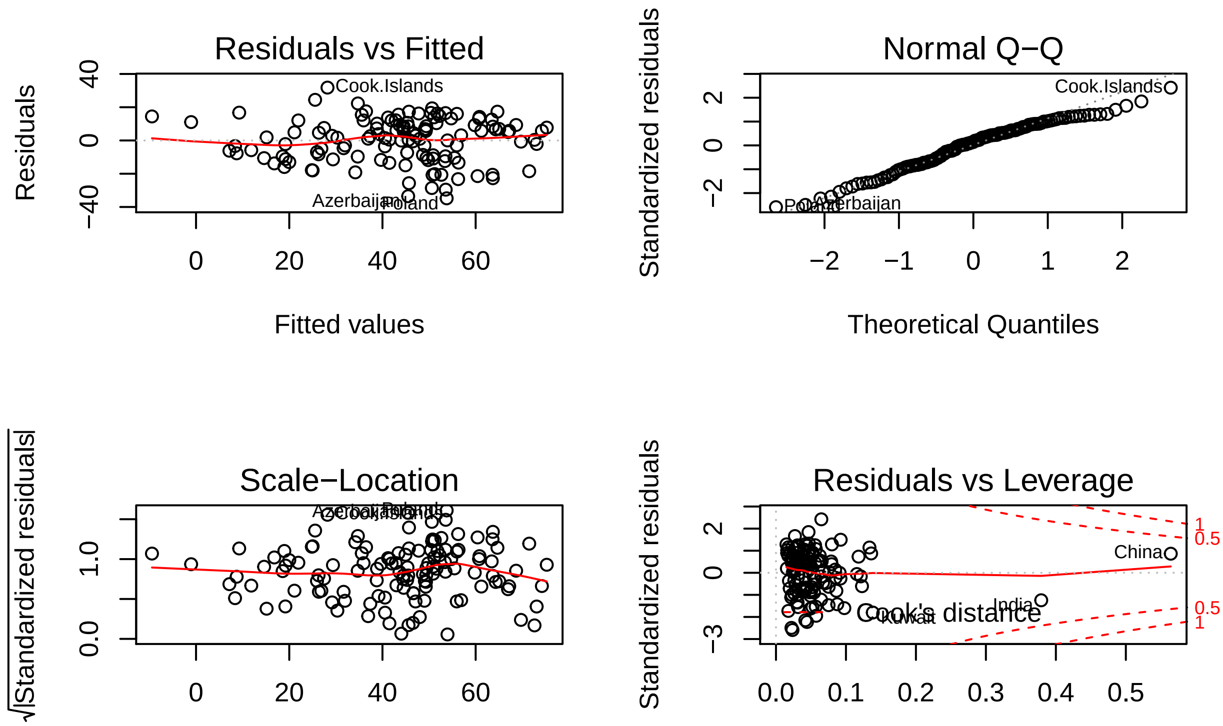
4. Use the `lm()` function to perform a multiple linear regression with `ModernC` as the response and all other variables as the predictors, using the formula `ModernC ~ .`, where the `.` includes all remaining

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?

```
modernc_lm<-lm(ModernC~.,data=UN3,na.action = na.omit)
summary(modernc_lm)
```

```
##
## Call:
## lm(formula = ModernC ~ ., data = UN3, na.action = na.omit)
##
## 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
```

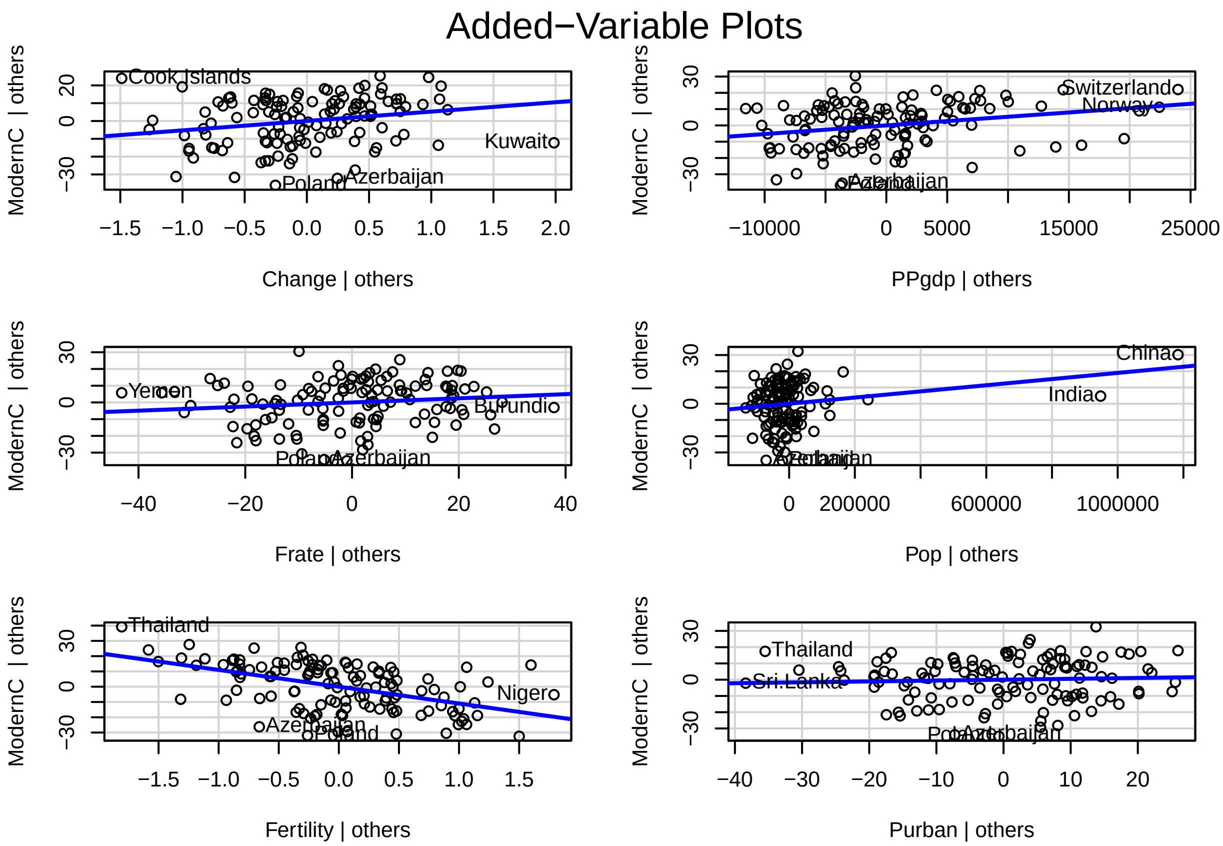
```
par(mfrow=c(2,2))
plot(modernc_lm)
```



The standardized residuals seem to follow normal distribution and don't vary with increase of $\hat{y}$. There are two high influential points because their population is much larger than other countries but they don't have large Cook's distance and shoud not be considered as outliers.

5. Examine added variable plots `car::avPlot` or `car::avPlots` for your model above. Are there any plots that suggest that transformations are needed for any of the terms in the model? Describe. Is it likely that any of the localities are influential for any of the terms? Which localities? Which terms?

```
avPlots(modernc_lm)
```



In the Pop term, we can see that China and India are high influential.

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.

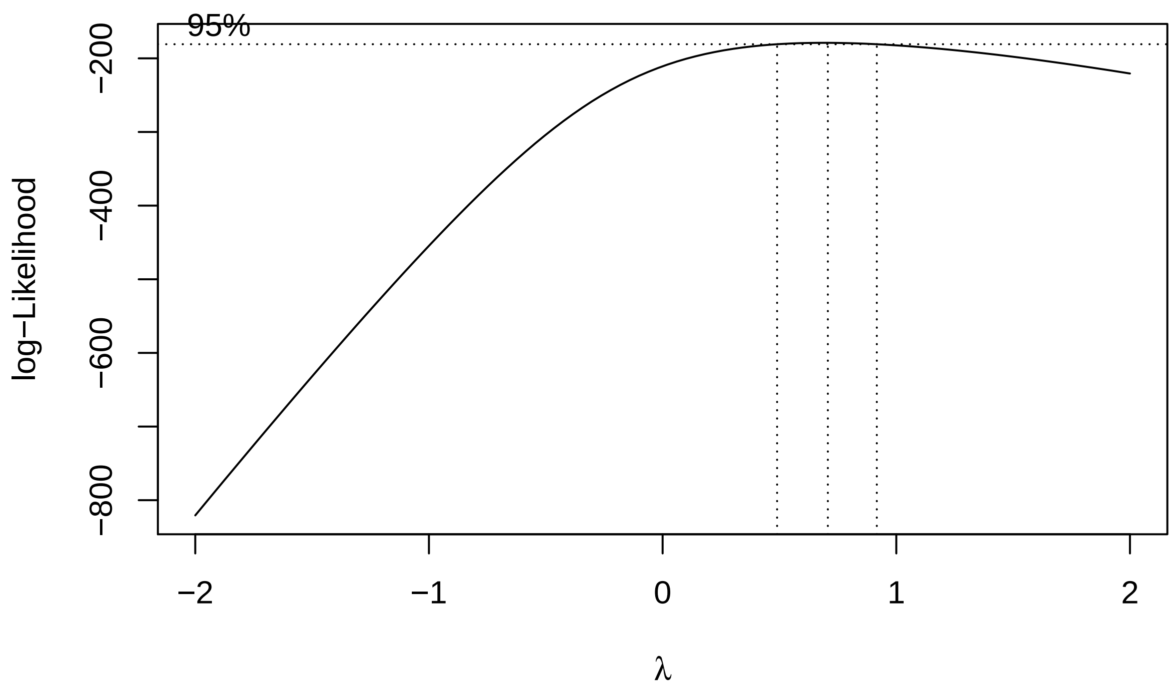
```
tran_predictor <- boxTidwell(ModernC ~ PPgdp + Pop,~Change+Frate+Purban, data = UN3, na.action = na.excl
boxTidwell(ModernC ~ PPgdp + Pop,~Change+Frate+Purban, data = UN3, na.action = na.exclude)
```

```
##       MLE of lambda Score Statistic (z) Pr(>|z|)
## PPgdp      -0.40887             -2.2634  0.02361 *
## Pop         0.32008             -1.2935  0.19582
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## iterations =  5
```

Since we only need to transform Pop and PPgdp and they are nonegative, we don't need to make it nonegetive. According to the result above, we might transform PPgdp to $\frac{1}{\sqrt{PPgdp}}$ and Pop to $log(Pop)$

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

```
UN3['logPop'] <- log(UN3$Pop)
UN3['PPgdp_trans'] <- 1/sqrt(UN3$PPgdp)
modernc_lm_pre_tran <- lm(ModernC~Change+Frate+Fertility+Purban+logPop+PPgdp_trans,data = UN3)
MASS::boxcox(modernc_lm_pre_tran)
```



As the plot shows above, we don't need to do a transformation on the response.

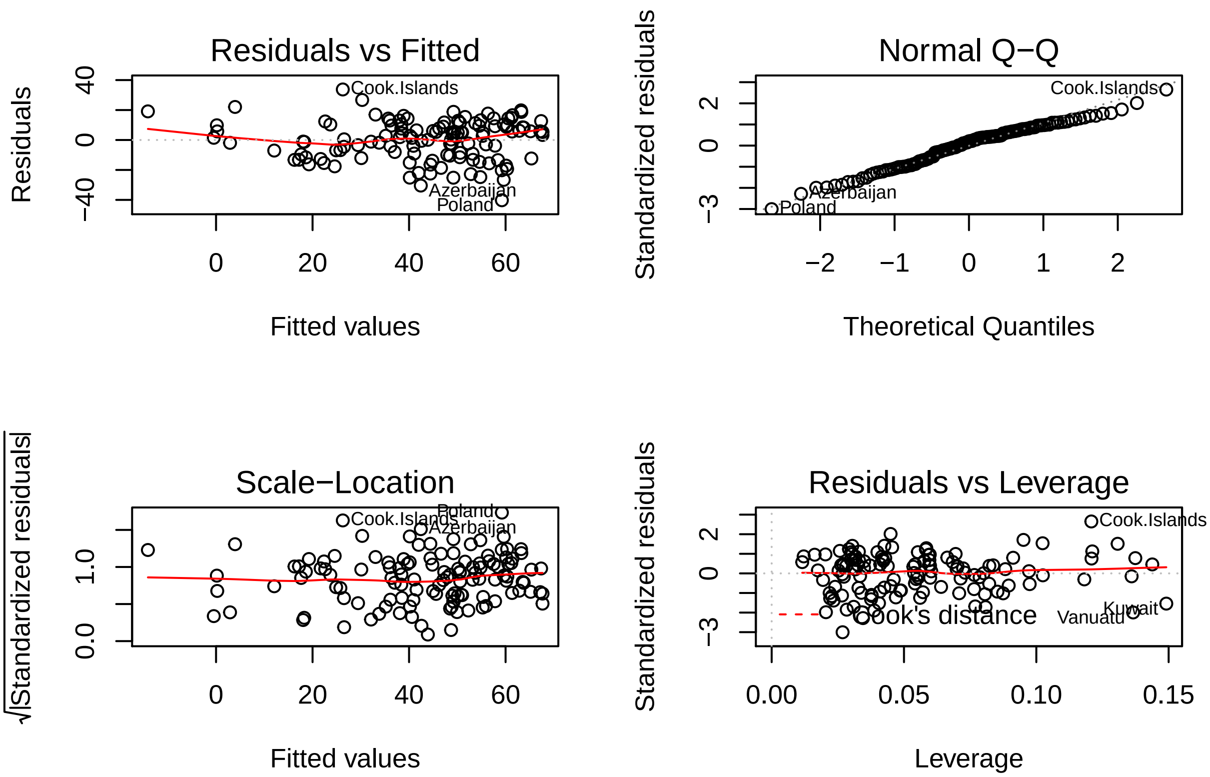
8. Fit the regression using the transformed variables. Provide residual plots and added variables plots and comment. If you feel that you need additional transformations of either the response or predictors, repeat any steps until you feel satisfied.

```
summary(modernc_lm_pre_tran)
```

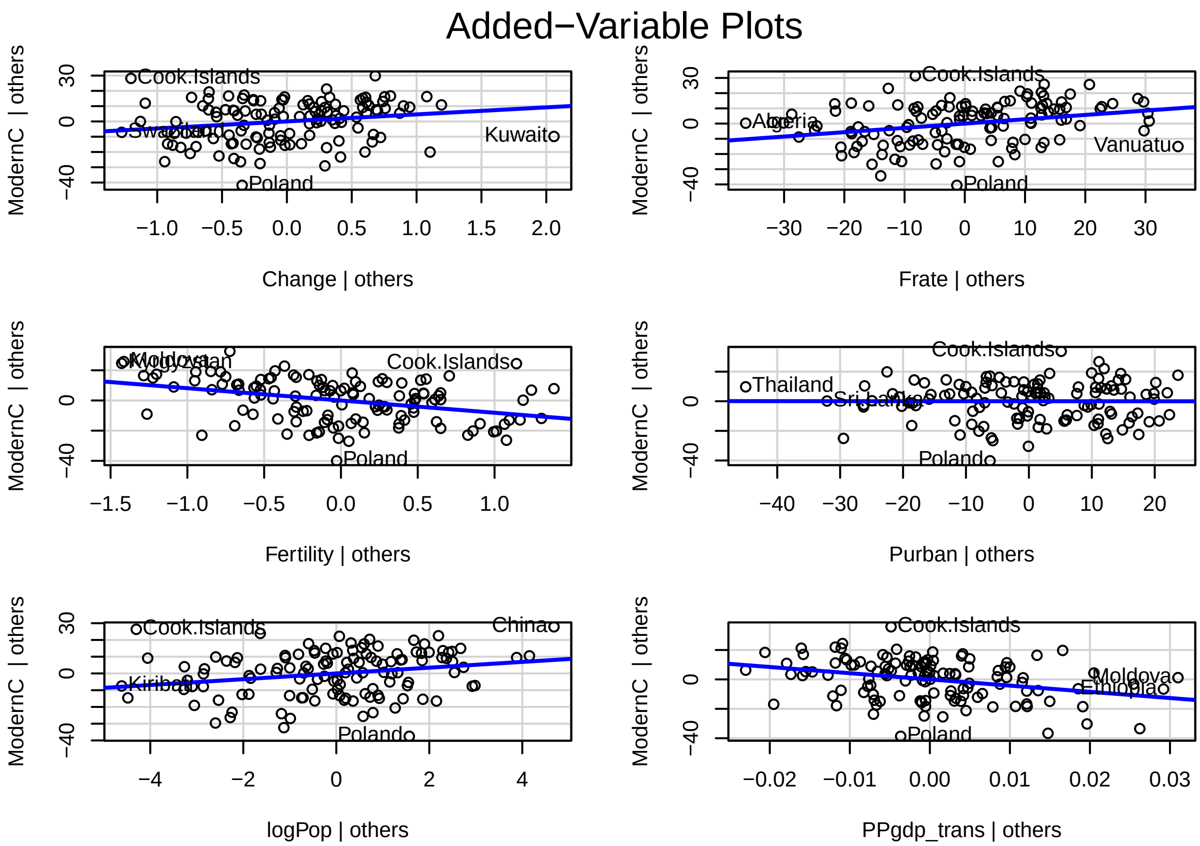
```
##
## Call:
## lm(formula = ModernC ~ Change + Frate + Fertility + Purban +
##     logPop + PPgdp_trans, data = UN3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -40.232  -9.904   2.745   9.632  33.734
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.379e+01  1.089e+01   4.022 0.000102 ***
## Change       4.603e+00  2.103e+00   2.188 0.030621 *
## Frate        2.839e-01  8.255e-02   3.439 0.000808 ***
## Fertility   -8.138e+00  1.942e+00  -4.191 5.38e-05 ***
## Purban      -1.856e-03  9.032e-02  -0.021 0.983638
## logPop       1.721e+00  6.571e-01   2.619 0.009981 **
## PPgdp_trans -4.224e+02  1.180e+02  -3.579 0.000501 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 13.57 on 118 degrees of freedom
##   (85 observations deleted due to missingness)
## Multiple R-squared:  0.6187, Adjusted R-squared:  0.5993
## F-statistic: 31.91 on 6 and 118 DF,  p-value: < 2.2e-16
```

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



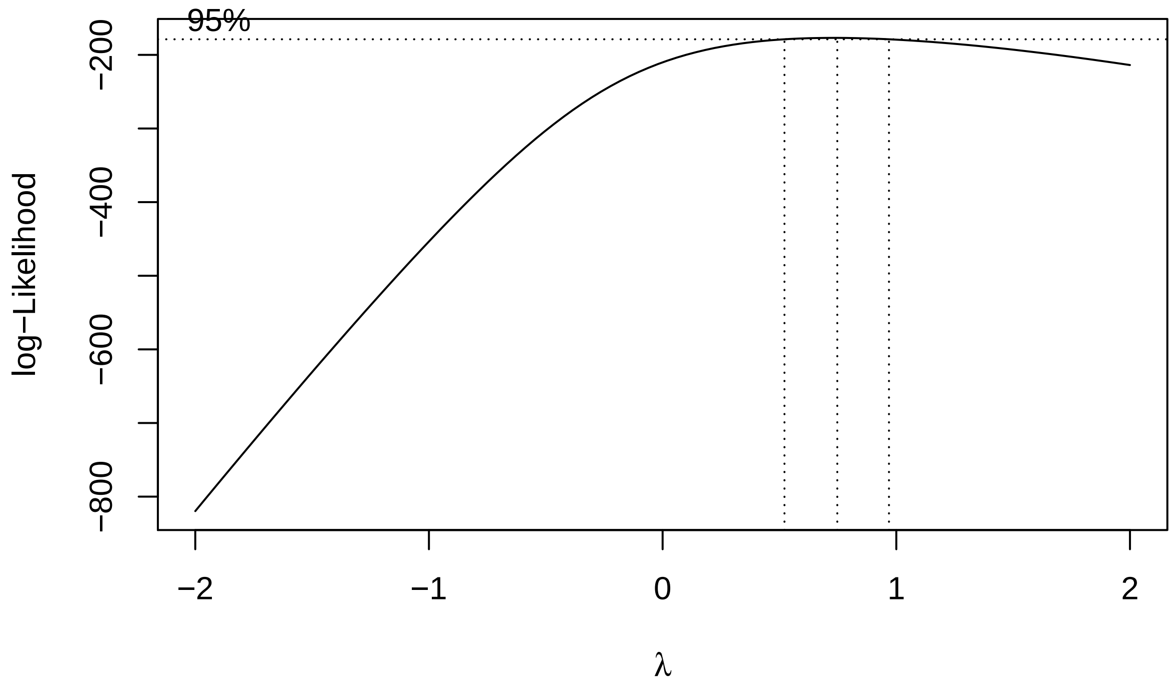

```r
avPlots(modernc_lm_pre_tran)
```



There seems no problem with various plot.

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(moderncc_lm)
```



We see that if we apply boxcox to the response first, we don't need to transform response. So the result would be same as doing transformation of the predictors first.

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 is no any outlier or influential point after the transformation.

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!

	estimate	2.5 %	97.5 %	Interpretations
(Intercept)	43.792	22.231	65.354	
Change	4.603	0.437	8.768	increasing 1 unit of change would increase response by 4.602828 unit
Frate	0.284	0.12	0.447	increasing 1 unit of Frate would increase response by 0.283878 unit
Fertility	-8.138	-11.983	-4.293	increasing 1 unit of Fertility would decrease response by -8.138039 unit
Purban	-0.002	-0.181	0.177	increasing 1 unit of Purban would decrease response by -0.001856 unit
logPop	1.721	0.42	3.022	increasing 10% of Pop would increase response by 1.720949*log(1.1)
PPgdp_trans	-422.375	-656.054	-188.695	increasing 10% of PPgdp would decrease response by -422.374698*(1-1/sc

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

ModernC is propotional to change, $\frac{1}{\sqrt{PPgdp}}$, Frate, $log(POP)$, Fertility and Purban. Pop, Frate, change,PPgdp have positive effect on the ModernC while Fertility, Purban have negetive effect on the ModernC. Small, developed countries have larger ModernC than large, developing coutries.

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.*

$1_n^T e\_Y = 1_n^T(Y - \hat{Y}) = 1_n^T(Y - X\hat{\beta}) = 1_n^T(Y - X(X^TX)X^TY) = 1_n^T(I - X(X^TX)X^T)Y = 1_n^T(I - H)Y = 0$

similarly we can get $1_n^T e\_X = 0$

If we do a regression on e_Y based on e_X,

$\hat{\beta}_0 = e\bar{\_}Y - \hat{\beta}_1 e\bar{\_}X = 1_n^T e\_Y - \hat{\beta}_1 1_n^T e\_X = 0 - 0 = 0$

The intercept in the added variable scatter plot will always be zero.

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

```
beta_of_full_model<-modernc_lm_pre_tran$coef[-1]
slope_av<-c()
UN3_new<-na.omit(UN3[-c(3,5)])
for(i in 2:7){
  X <- cbind(1,as.matrix(UN3_new[-c(1,i)]))
  H <- X%*%solve(t(X)%*%X)%*%t(X)

  e_Y <- (diag(1,nrow(UN3_new))-H)%*%UN3_new$ModernC
  e_X <- (diag(1,nrow(UN3_new))-H)%*%UN3_new[[i]]
  slope_av<-c(slope_av,sum(e_Y*e_X)/sum(e_X**2))
}
beta_vs<-cbind(beta_of_full_model,slope_av)
kable(beta_vs)
```

	beta_of_full_model	slope_av
Change	4.6028285	4.6028285
Frate	0.2838776	0.2838776
Fertility	-8.1380387	-8.1380387
Purban	-0.0018561	-0.0018561
logPop	1.7209488	1.7209488
PPgdp_trans	-422.3746981	-422.3746981

let $X_j = (x_1, x_2, ..., x_{j-1}, x_{j+1}, ..., x_p)$, $X = (x_j, X_j)$

$e\_Y = Y - X_j(X_j^TX_j)^{-1}X_j^TY = (I - X_j(X_j^TX_j)^{-1}X_j^T)Y$

$e\_X = (I - X_j(X_j^TX_j)^{-1}X_j^T)x_j$

$\hat{\beta}_j^* = \frac{e\_X^Te\_Y}{e\_X^Te\_X} = \frac{x_j^T(I - X_j(X_j^TX_j)^{-1}X_j^T)Y}{x_j^T(I - X_j(X_j^TX_j)^{-1}X_j^T)x_j}$

$\hat{\beta} = (X^TX)^{-1}X^TY = (\begin{pmatrix} x_j^T \\ X_j^T \end{pmatrix} \begin{pmatrix} x_j & X_j \end{pmatrix})^{-1} \begin{pmatrix} x_j^T \\ X_j^T \end{pmatrix} Y = \begin{pmatrix} x_j^T x_j & x_j^T X_j \\ X_j^T x_j & X_j^T X_j \end{pmatrix}^{-1} \begin{pmatrix} x_j^T \\ X_j^T \end{pmatrix} Y$

As we only care about the first entry of $\hat{\beta}$, we only need the first row of$\begin{pmatrix} x_j^T x_j & x_j^T X_j \\ X_j^T x_j & X_j^T X_j \end{pmatrix}^{-1}$, donate this by h_1

$h_1 = \begin{pmatrix} \frac{1}{x_j^T x_j} + \frac{1}{(x_j^T x_j)^2} x_j^T X_j A X_j^T x_j & -\frac{1}{x_j^T x_j} x_j^T X_j A \end{pmatrix}$

where $A = (X_j^T X_j - X_j^T x_j x_j^T X_j / x_j^T x_j)^{-1}$

$(X_j^T X_j - X_j^T x_j x_j^T X_j / x_j^T x_j)A = I$

$\frac{X_j^T x_j x_j^T X_j}{x_j^T x_j} A = X_j^T X_j A - I$

$\hat{\beta}_j = h_1 \begin{pmatrix} x_j^T \\ X_j^T \end{pmatrix} Y = \begin{pmatrix} \frac{1}{x_j^T x_j} + \frac{1}{(x_j^T x_j)^2} x_j^T X_j A X_j^T x_j & -\frac{1}{x_j^T x_j} x_j^T X_j A \end{pmatrix} \begin{pmatrix} x_j^T \\ X_j^T \end{pmatrix} Y = ((\frac{1}{x_j^T x_j} + \frac{1}{(x_j^T x_j)^2} x_j^T X_j A X_j^T x_j) x_j^T - \frac{1}{x_j^T x_j} x_j^T X_j A X_j^T) Y$

$= (\frac{x_j^T}{x_j^T x_j} + \frac{1}{xj^T x_j} x_j^T X_j A X_j^T (\frac{x_j x_j^T}{x_j^T x_j} - I)) Y$

$C = \frac{x_j^T}{x_j^T x_j} + \frac{1}{xj^T x_j} x_j^T X_j A X_j^T (\frac{x_j x_j^T}{x_j^T x_j} - I)$

$D = \frac{e\_X^T e\_Y}{e\_X^T e\_X} = \frac{x_j^T (I - X_j (X_j^T X_j)^{-1} X_j^T)}{x_j^T (I - X_j (X_j^T X_j)^{-1} X_j^T) x_j}$

$\therefore \hat{\beta}_j^* = DY$, $\hat{\beta}_j = CY$

$(x_j^T x_j - x_j^T X_j (X_j^T X_j)^{-1} X_j^T x_j) C$
$= x_j^T + x_j^T X_j A X_j^T (\frac{x_j x_j^T}{x_j^T x_j} - I) - \frac{x_j^T H x_j x_j^T}{x_j^T x_j} - x_j^T X_j (X_j^T X_j)^{-1} \frac{X_j x_j x_j^T X_j A}{x_j^T x_j} X_j^T (\frac{x_j x_j^T}{x_j^T x_j} - I)$
$= x_j^T + x_j^T X_j A X_j^T (\frac{x_j x_j^T}{x_j^T x_j} - I) - \frac{x_j^T H x_j x_j^T}{x_j^T x_j} - x_j^T X_j (X_j^T X_j)^{-1} (X_j^T X_j A - I) X_j^T (\frac{x_j x_j^T}{x_j^T x_j} - I)$
$= x_j^T + x_j^T X_j A X_j^T (\frac{x_j x_j^T}{x_j^T x_j} - I) - \frac{x_j^T H x_j x_j^T}{x_j^T x_j} - x_j^T X_j A X_j^T (\frac{x_j x_j^T}{x_j^T x_j} - I) + x_j^T H (\frac{x_j x_j^T}{x_j^T x_j} - I)$
$= x_j^T - x_j^T H$

where $H = X_j (X_j^T X_j)^{-1} X_j^T$

$\therefore C = \frac{x_j^T - x_j^T H}{(x_j^T x_j - x_j^T X_j (X_j^T X_j)^{-1} X_j^T x_j)} = D$

$\hat{\beta}_j^* = \hat{\beta}_j$

The slope of added variable plot is equal to the coefficient of full model.