HW2 STA521 Fall18

Billy Jiang xj35 jiangxiaoyuww

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.

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

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
```

```
missingvalue = sapply(UN3, function(x) sum(is.na(x)))
mode = sapply(UN3, function(x) class(x))
f = rbind(missingvalue,mode)
row.names(f) = c('missing_value', 'mode')
kable(f)
```

	ModernC	Change	PPgdp	Frate	Pop	Fertility	Purban
missing_value	58	1	9	43	2	10	0
mode	integer	numeric	integer	integer	numeric	numeric	integer

Ans: From our data, there are six variables with missing data. It seems that all of them are quantitive.

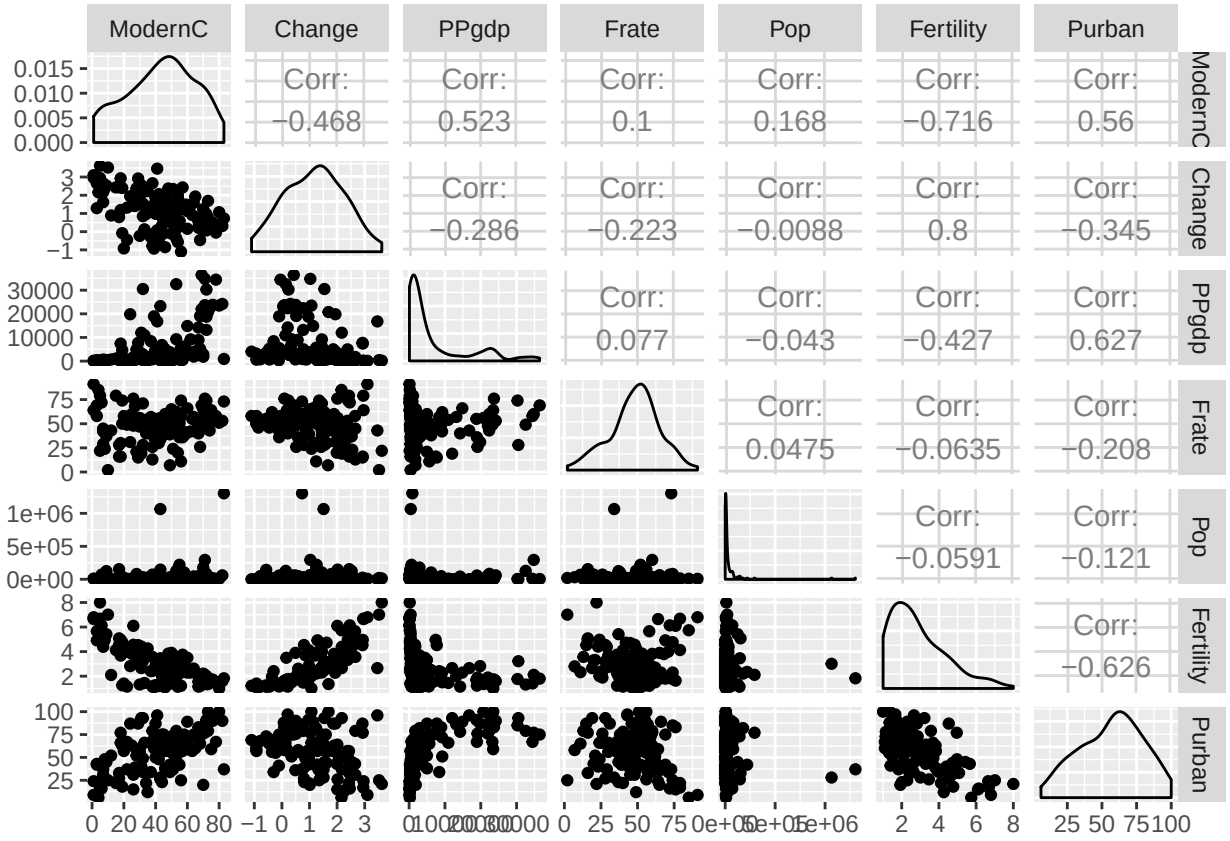
2. What is the mean and standard deviation of each quantitative predictor? Provide in a nicely formatted table.

```
tempm = sapply(UN3, function(x) mean(x, na.rm = TRUE))
tempsd = sapply(UN3, function(x) sd(x, na.rm = TRUE))
total = rbind(tempm,tempsd)
rownames(total) = c("mean", "sd")
total = t(total)
kable(total)
```

	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?

```
library(GGally)
UN3final = UN3[complete.cases(UN3),]
ggpairs(UN3final)
```



Ans: Countries with one or more missing variables are excluded because they are not included in our final model. From our plots, we see that PPgdp, Pop and Purban are strongly and positively correlated with MordenC. Fertility and Change, on the other hand, are negatively correlated with MordenC. There are seem to have two very influential points for Pop. The relationship between PPgdp and ModernC, Fertility and ModernC seem to be nonlinear in some way, one way to solve this problem to log transform our variables.

Model Fitting

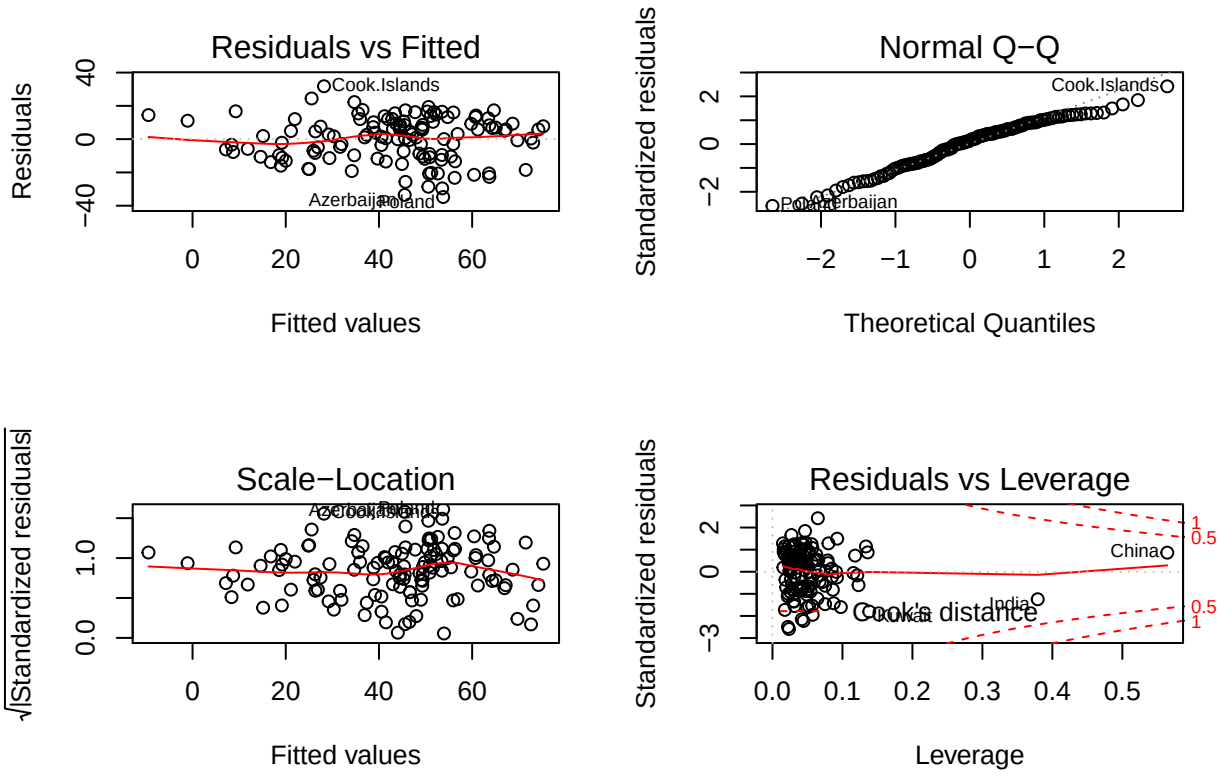
4. Use the `lm()` function to perform a multiple linear regression with `ModernC` as the response and all other variables as the predictors, using the formula `ModernC ~ .`, where the `.` includes all remaining variables in the dataframe. Create diagnostic residual plot from the linear model object and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

```
lmo = lm(ModernC~.,data = UN3)
summary(lmo)
```

```
##
## 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
```

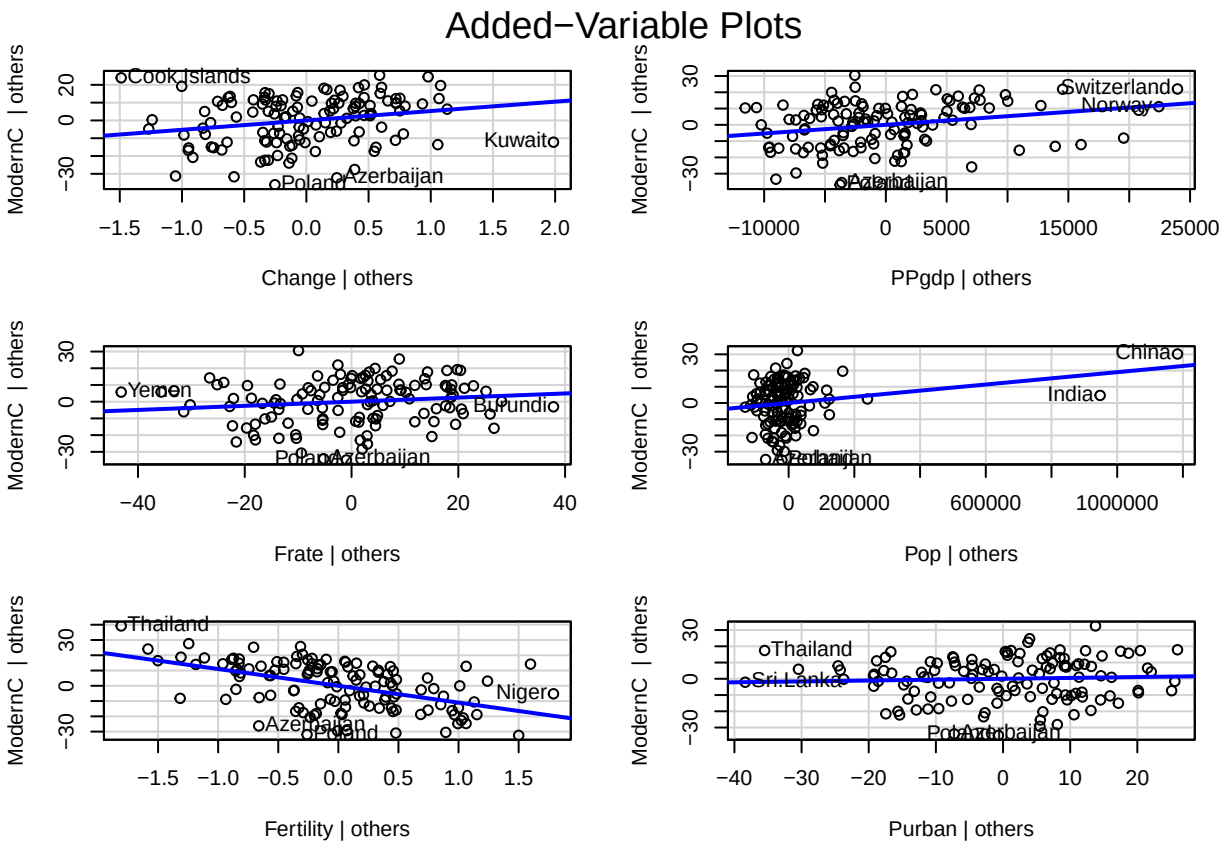
```
par(mfrow=c(2,2))
plot(lmo)
```



Ans: With 85 observations deleted due to missingness, there are 125 observations used in the model fitting.The Normal Q-Q suggests that our sample is heavily tailed. Minor heteroscedastic trend is also present. From the fourth plot, we see taht China, India and Kuwait seem to be particularly influential.

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(lm(formula = ModernC ~ ., data = UN3))
```



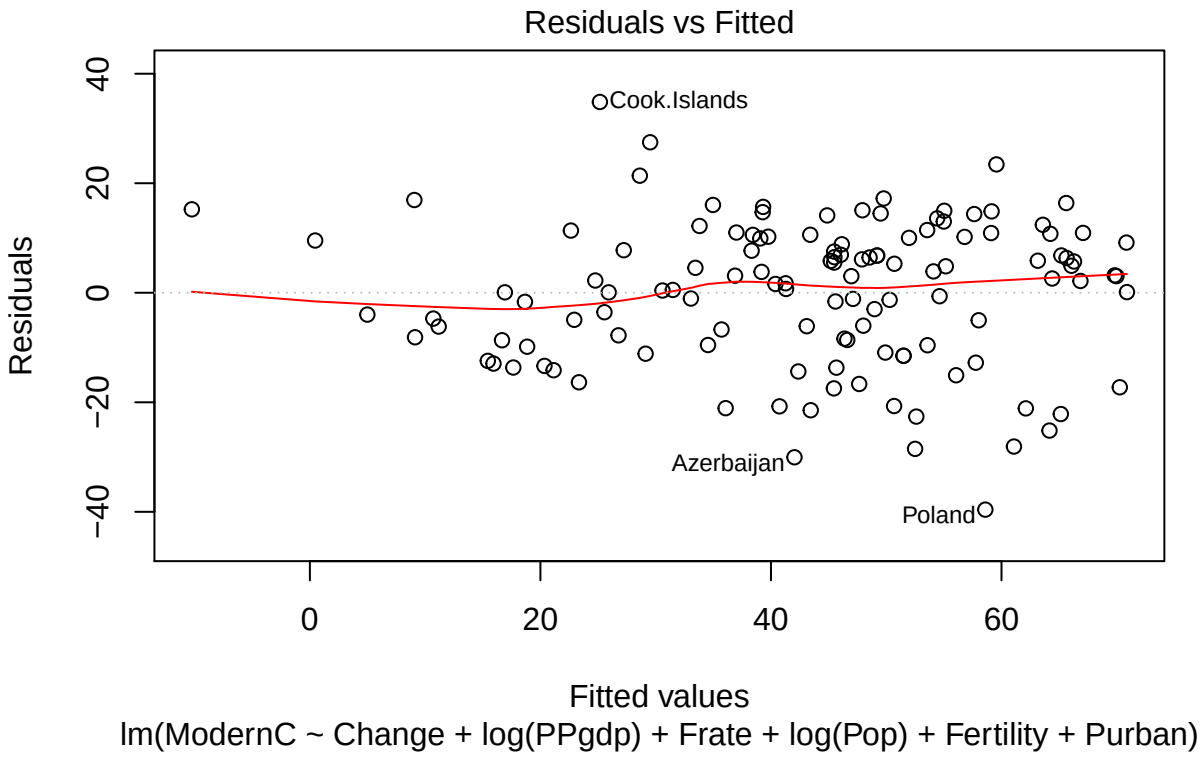
Ans: From the plots presented, there are not any plots that suggest the strong need of transformation; however, it never hurts to compare with the log transformation graph. From pop|others plot, we see that China and India seem to be influential in determining the slope of the fitted line. Other points seem to be clustered. Thailand also might pull the slope a little bit up for the Fertility graph. From problem 3, we also see that PPgdp, Pop seem to suggest log transformations.

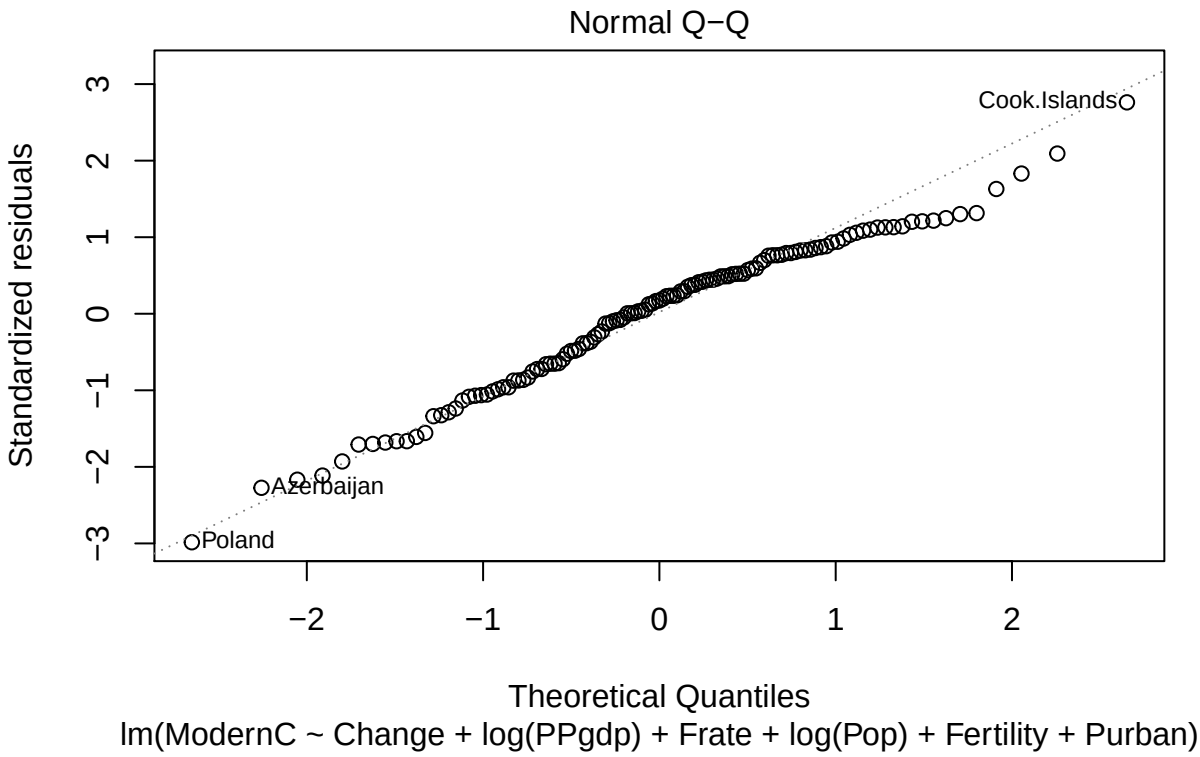
6. Using the Box-Tidwell `car::boxTidwell` or graphical methods find appropriate transformations of the predictor variables to be used as predictors in the linear model. If any predictors are negative, you may need to transform so that they are non-negative. Describe your method and the resulting transformations.

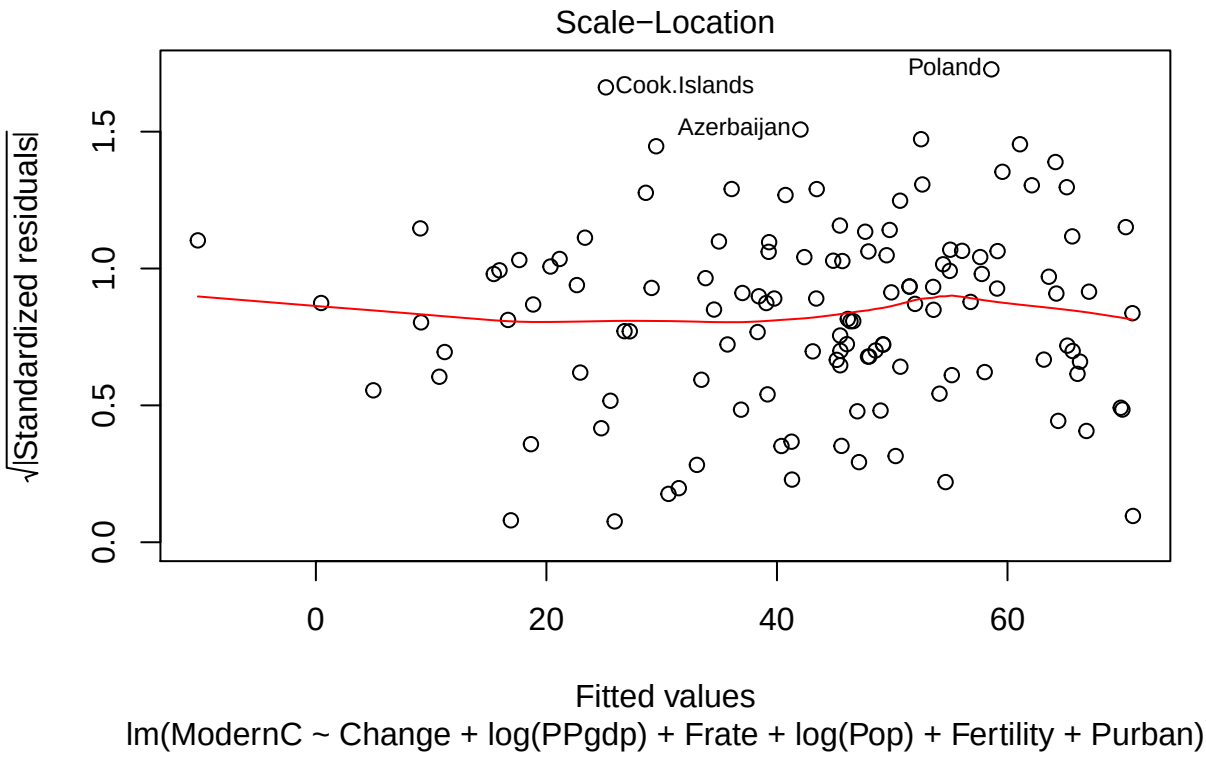
```
library(alr3)
library(dplyr)
final = UN3[complete.cases(UN3), ]
car::boxTidwell(ModernC ~ PPgdp + Pop, other.x = ~Change + Fertility + Purban + Frate, data = final)
```

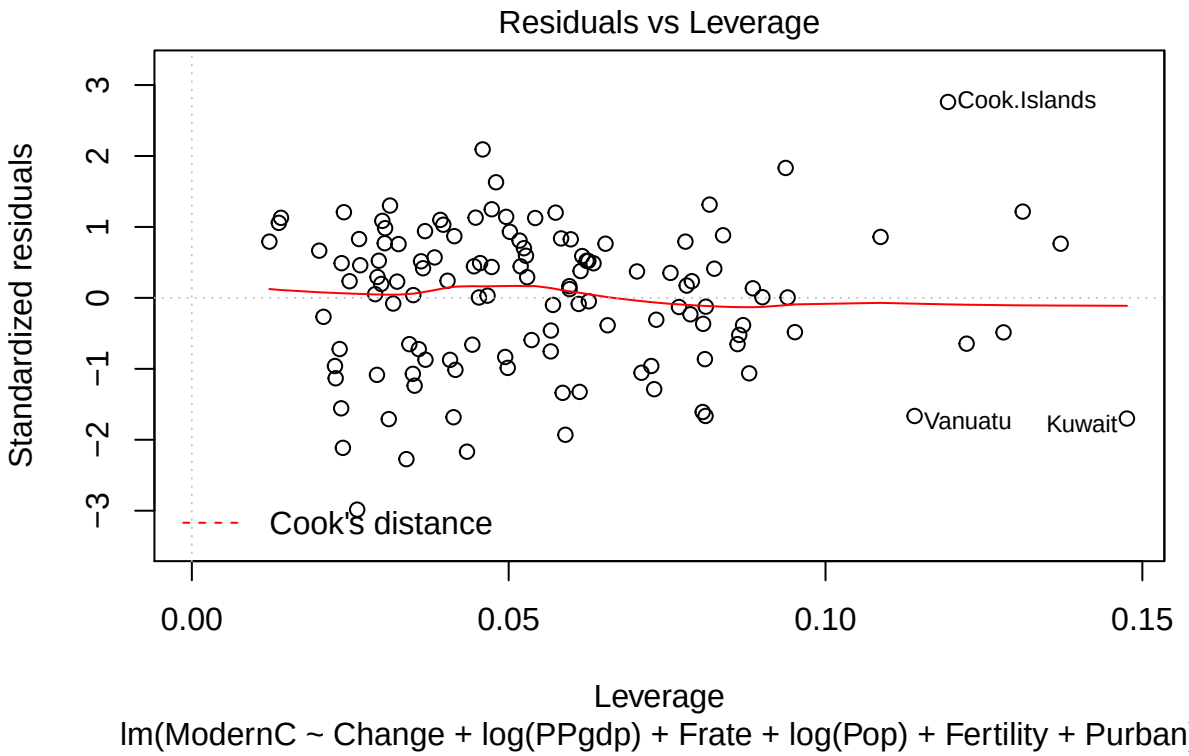
```
##       MLE of lambda Score Statistic (z) Pr(>|z|)
## PPgdp      -0.12921             -1.1410   0.2539
## Pop         0.40749             -0.7874   0.4310
##
## iterations =  4
```

```
lmtransform = lm(formula = ModernC ~ Change + log(PPgdp) + Frate + log(Pop) + Fertility + Purban, final)
plot(lmtransform)
```


Residuals vs Fitted
Residuals
Fitted values
lm(ModernC ~ Change + log(PPgdp) + Frate + log(Pop) + Fertility + Purban)
Cook.Islands
Azerbaijan
Poland
40
20
0
−20
−40
0
20
40
60


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(ModernC ~ Change + log(PPgdp) + Frate + log(Pop) + Fertility + Purban)
Cook.Islands
Azerbaijan
Poland
−3
−2
−1
0
1
2
3

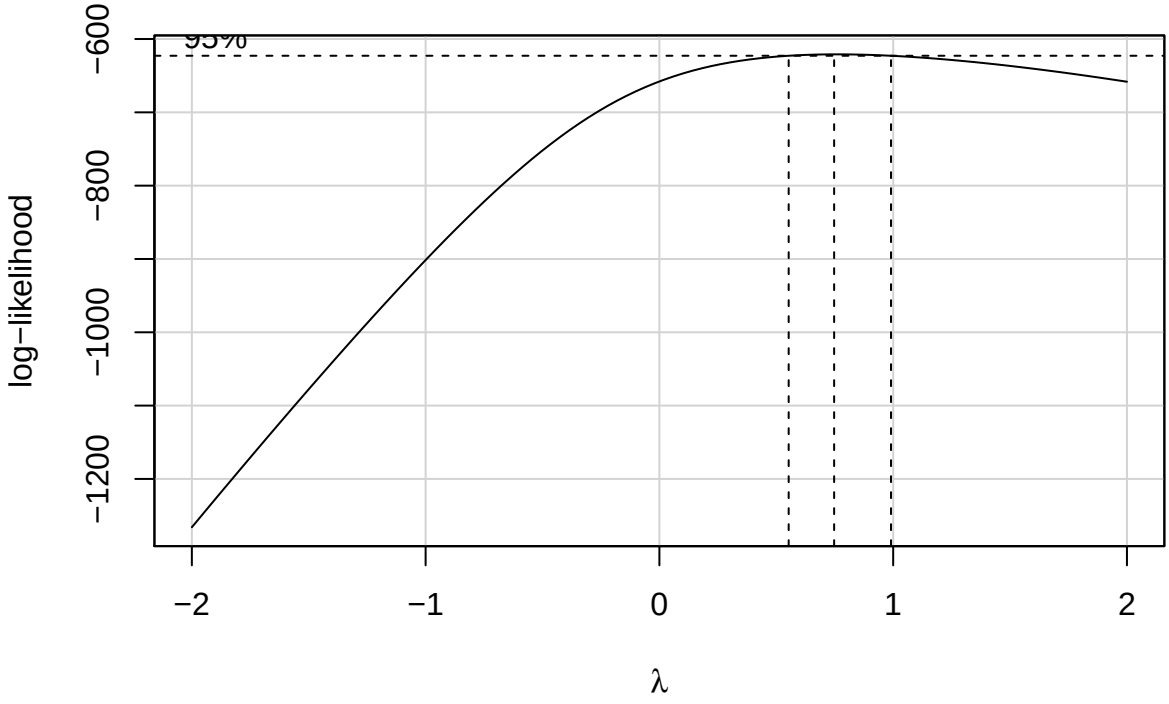

Scale−Location
Cook.Islands
Poland
Azerbaijan
√|Standardized residuals|
0.0
0.5
1.0
1.5
0
20
40
60
Fitted values
lm(ModernC ~ Change + log(PPgdp) + Frate + log(Pop) + Fertility + Purban)



Ans: From our test, it seems that nothing should be transformed; however we log transform PPgdp and Pop from the nonlinearities we detected in problem 3. The residual plot after transformation seems better as variances became more constant and less influential points. In other literatures, PPgdp and Pop are often log transformed, which can serve as some sort of prior belief.

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

```
car::boxCox(lmtransform)
```




```
powerTransform(lmtransform)
```

```
## Estimated transformation parameter
##        Y1
## 0.7585897
```

Ans: From the boxCox plot, it seems that lambda is around 0.75, with the 95 CI covering 0.5 to 1. There is not much point in transforming our response variable, we can just choose lambda to be 1, which is within the CI.

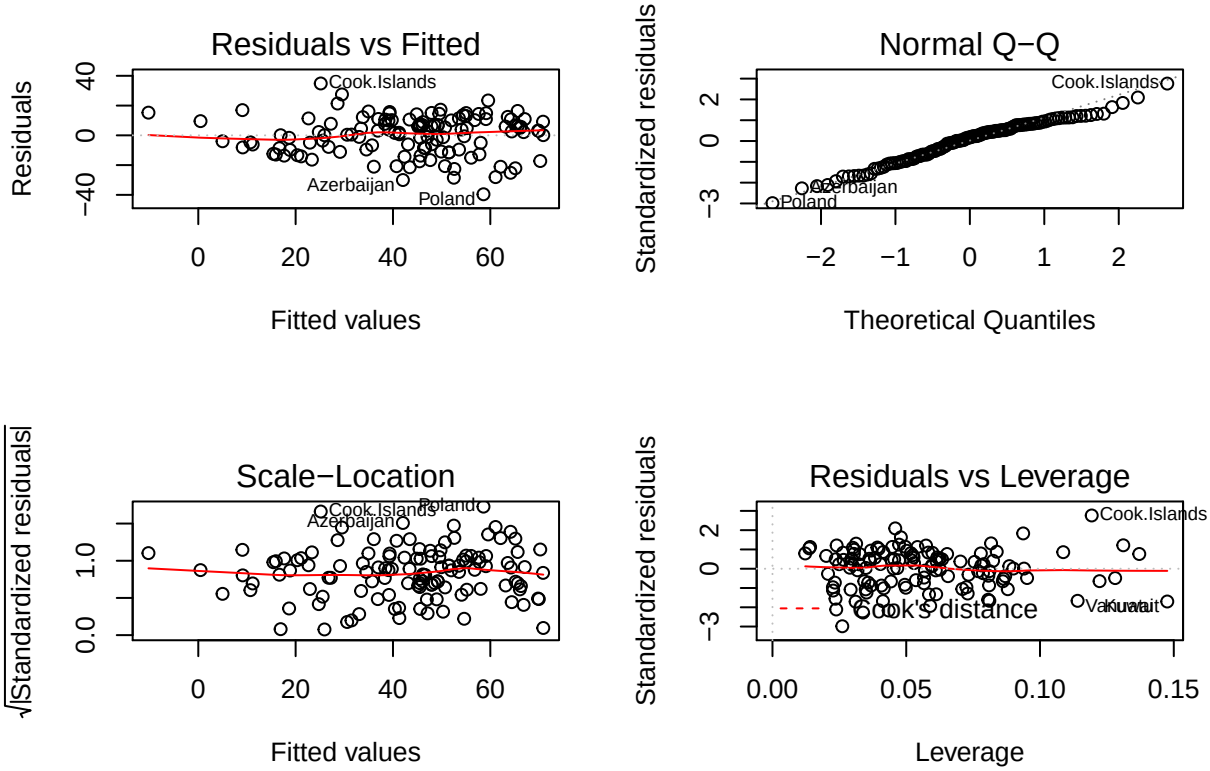
8. Fit the regression using the transformed variables. Provide residual plots and added variables plots and comment. If you feel that you need additional transformations of either the response or predictors, repeat any steps until you feel satisfied.

```
summary(lmtransform)
```

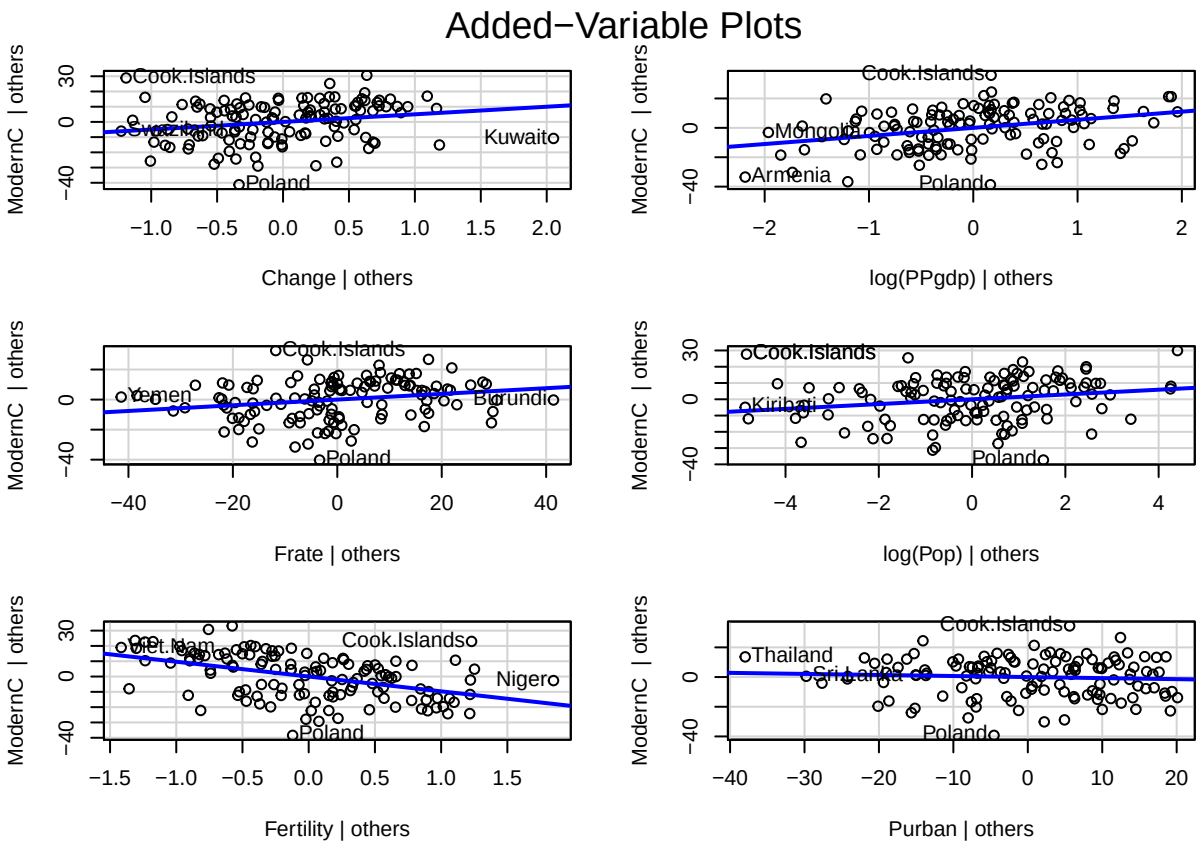
```
##
## Call:
## lm(formula = ModernC ~ Change + log(PPgdp) + Frate + log(Pop) +
##     Fertility + Purban, data = final)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -39.597  -9.540   2.238  10.024  34.840
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.11547   14.50854   0.284 0.777169
```

```
## Change       4.99296    2.07709   2.404 0.017781 *
## log(PPgdp)   5.50728    1.40505   3.920 0.000149 ***
## Frate        0.18939    0.07711   2.456 0.015500 *
## log(Pop)     1.47207    0.62875   2.341 0.020897 *
## 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
## 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(lmtransform)
```



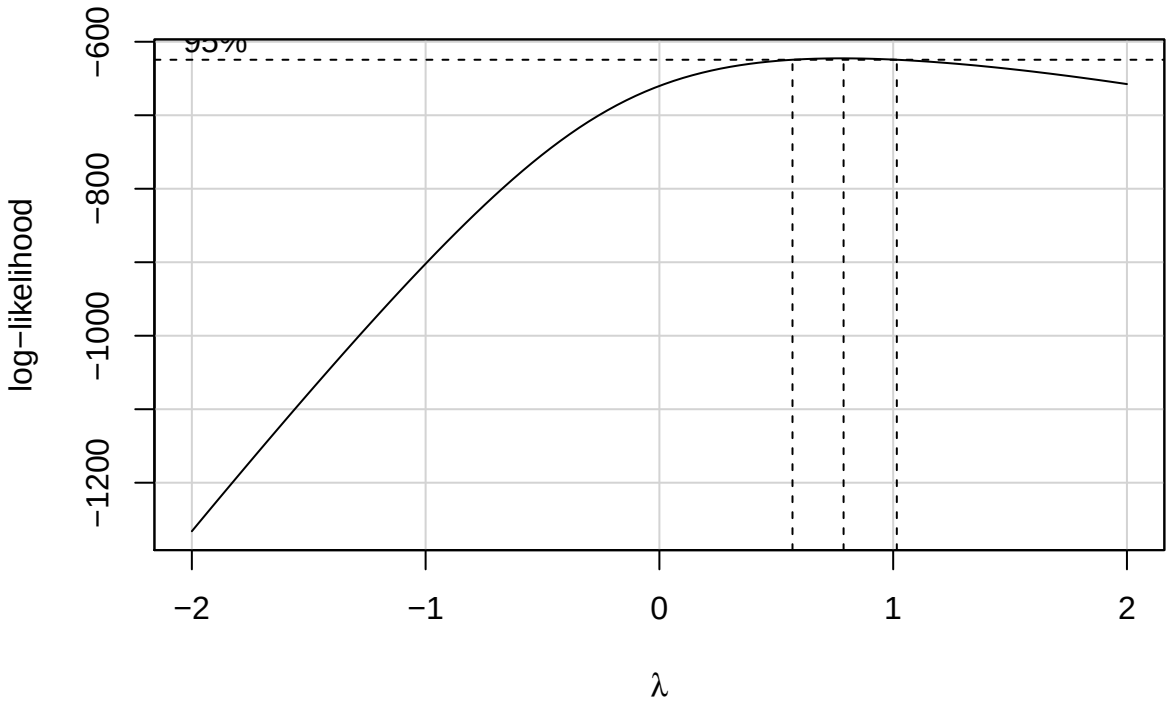

```
car::avPlots(lm(formula = ModernC~Change + log(PPgdp) + Frate + log(Pop) + Fertility + Purban, data = U
```



Ans: One of the major difference is we see that the significance value for log(PPgdp) went up compare to PPgdp, suggesting that log transformation of PPgdp fits better with ModernC. Also, we notice that the intercept is no longer significant, probably due to our transformed PPgdp. As for residual plot, the variance seems to be more consistent. Normal Q-Q suggest that points are closer to our theoretical line. Less influential points are also suggested by Leverage plot. The transformation gives results closer to our assumptions for OLS.

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?

```
car::boxCox(lm(formula = ModernC ~. , data = final))
```



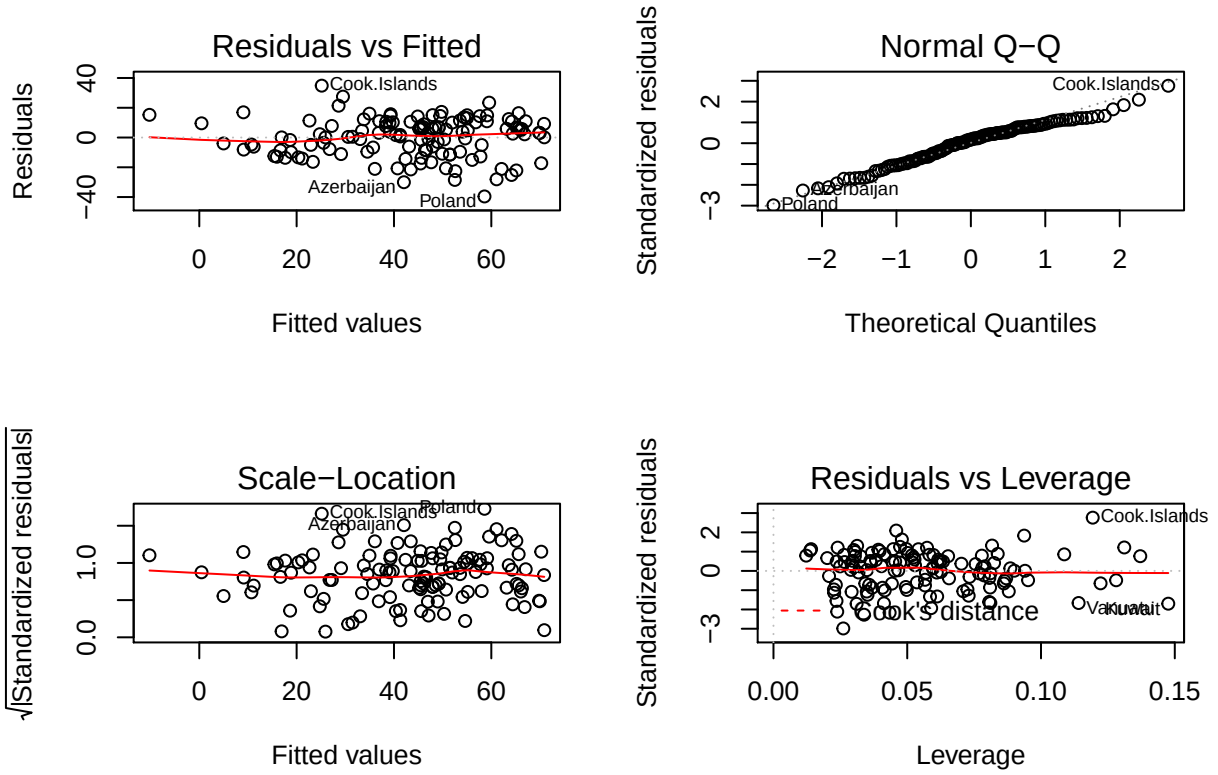

```
car::boxTidwell(ModernC ~ PPgdp + Pop, other.x = ~Change + Fertility + Purban + Frate, data = final)
```

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

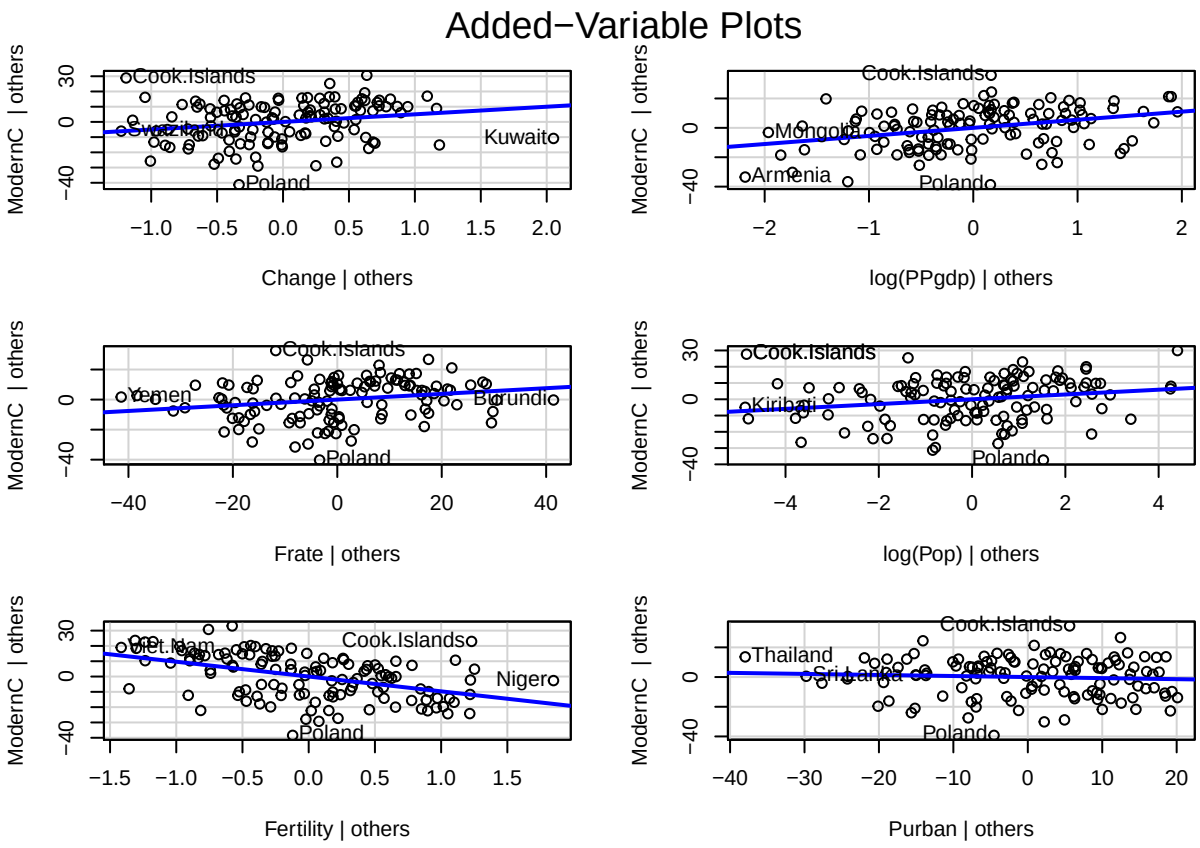
Ans: we see that the model is roughly the same in as in problem 8. It doesn't seem that a transformation of response is needed. Of course, if we choose not to transform the response variable, boxTidwell on predictors will give the same results.

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.

```
lmtransform = lm(formula = ModernC ~ Change + log(PPgdp) + Frate + log(Pop) + Fertility + Purban, UN3)
par(mfrow=c(2,2))
plot(lmtransform)
```




```
car::avPlots(lmtransform)
```



Ans: Yeah, China and India seem to look like outliers at first; however, they don't seem as much after log transformation. From a qualitative perspective, China and India are also too important to remove from the data set.

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!

```
fit = lm(formula = ModernC ~ log(PPgdp) +  Frate + log(Pop) + Fertility, data = final)
anova(lmtransform,fit)
```

```
## Analysis of Variance Table
##
## Model 1: ModernC ~ Change + log(PPgdp) + Frate + log(Pop) + Fertility +
##     Purban
## Model 2: ModernC ~ log(PPgdp) + Frate + log(Pop) + Fertility
##   Res.Df   RSS Df Sum of Sq      F  Pr(>F)
## 1    118 21325
## 2    120 22381 -2   -1056.4 2.9227 0.05769 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
t = data.frame(confint(fit))
coef = coefficients(fit)
t = cbind(t,coef)
```

```
row.names(t) = c("Intercept","PPgdp","Frate","Pop","Fertility")
kable(t)
```

	X2.5..	X97.5..	coef
Intercept	-31.6949920	25.2558395	-3.2195762
PPgdp	3.2244994	7.4977349	5.3611171
Frate	0.0088474	0.3046069	0.1567272
Pop	0.4829207	2.9588851	1.7209029
Fertility	-8.4937966	-4.1429402	-6.3183684

Ans: We choose to drop Purban and change as they are not significant in our previous models (confirmed by anova). PPgdp and Pop are log transformed; 95 CI suggests that if repeated samples are taken and 95% confidence inteval was computed for each sample, 95% of them would contain the true population mean. In this circumstance, we see that the intercept is between -31 and 25, so it gives virtually no information. The PPgdp,Frate and Pop all have positive coef and CI, suggesting an increase in these measures will increase Modern index by respective amount. Fertility has negaitive coefficients, suggesting it goes into the opposite direction of modernC. One way to explain log transformed predictor is that 10% increase in predictor will result in coef * log(1.1) increase in the response (given they move in the same direction).

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

Ans: In my final model, I used log(PPgdp), logPop, Fertility and Frate to predict ModernC. Log transformations were employed to conform our data with the normal assumption and stay current with the existing literature. Response variable ModernC was not transformed, supported by boxCox test. Almost half of the countries are omitted due to missing variables in one or more categories; however, do note that China and India are two very influential points before the transformation, although they are too important to omit; Our model gives a pretty decent result, with Pop, PPgdp and Frate influence modernC positively. More population, high per capita GDP and more female economic participation are commonly associated with modernization, while people in poor countries which lack contraception methods have high fertility rate.

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.

$$
\begin{aligned}
H &= X_j(X_jX_j^T)^{-1}X_j \\
e_Y &= Y - HY \\
&= (I-H)Y \\
e_X &= x_j - Hx_j \\
&= (I-H)x_j \\
e_Y &= \beta_0 1_n + \beta_1 e_x \\
(I-H)Y &= \beta_0 1_n + (x_j^T(I-H)^T(I-H)x_j)^{-1}x_j^T(I-H)Y(I-H)x_j \\
x_j^T(I-H)Y &= x_j^T\beta_0 + x_j^T(x_j^T(I-H)^T(I-H)x_j)^{-1}x_j^T(I-H)Y(I-H)x_j \\
x_j^T(I-H)Y &= x_j^T\beta_0 + x_j^T(I-H)x_j(x_j^T(I-H)x_j)^{-1}x_j^T(I-H)Y \\
x_j^T(I-H)Y &= x_j^T\beta_0 + x_j^T(I-H)Y \\
\beta_0 &= 0 \\
\frac{1}{n}\sum e_i = \frac{1}{n}1_n^T e_i &= \frac{1}{n}1_n^T(I-H)Y = 0
\end{aligned}
$$

X_j denotes feature matrix without jth feature and x_j is the jth feature.

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.

```
UN3final = UN3[complete.cases(UN3),]
e_Y = residuals(lm(ModernC ~  log(PPgdp) + Frate + log(Pop) + Fertility + Purban, UN3final)) #we omit c
lmtransformregressChange = lm(Change ~  log(PPgdp) + Frate + log(Pop) + Fertility + Purban, UN3final)
e_X = residuals(lmtransformregressChange)
lme = lm(e_Y ~ e_X)
lmtransform$coefficients["Change"]
```

```
##   Change
## 4.992957
```

```
lme$coefficients
```

```
##  (Intercept)          e_X
## 1.239519e-17 4.992957e+00
```

Ans: We see that the coefficent of the Change is equal to the coefficient of the slope in our manually constructed added variable plot.