HW2 STA521 Fall 17

[Your Name Here]

Due September 18, 2017

This exercise involves the UN data set from ALR. Download the `alr4` library and load the data to answer the following questions adding your code in the code chunks. Please add appropriate code to the chuncks to suppress messages and warnings as needed once you are sure the code is working properly and remove instructions if no longer needed. Please switch the output to pdf for your final version to upload to Sakai.

```
##
## Attaching package: 'alr3'

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

Exploratory Data Analysis

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 (i in 1:length(UN3))
{
  print(c(colnames(UN3[i]), is.numeric(UN3[,i])))

}
```

```
## [1] "ModernC" "TRUE"
## [1] "Change" "TRUE"
## [1] "PPgdp" "TRUE"
## [1] "Frate" "TRUE"
## [1] "Pop"  "TRUE"
## [1] "Fertility" "TRUE"
## [1] "Purban" "TRUE"
```

Answer: there are 6 variables that have missing data.
The variables with "TRUE" are quantitative – i.e., all vriables are quantitative

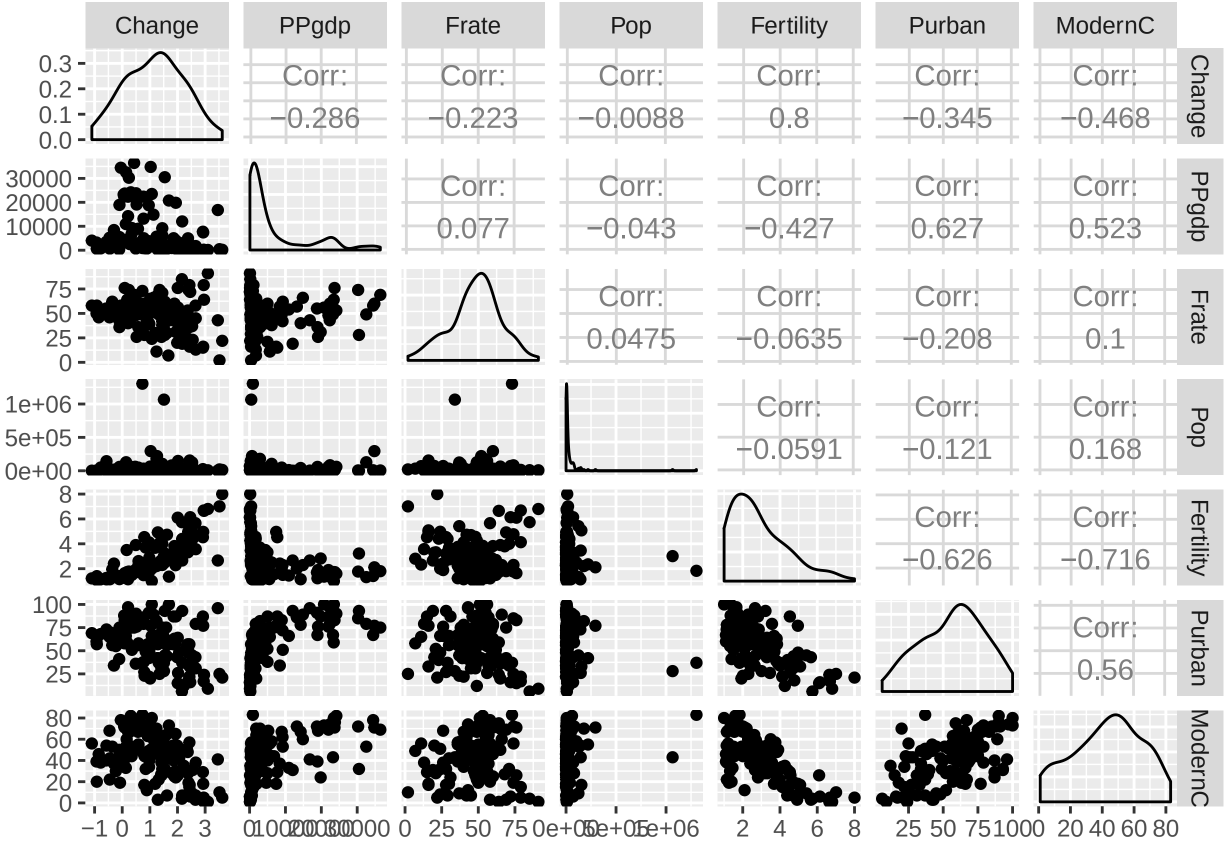
2. What is the mean and standard deviation of each quantitative predictor? Provide in a nicely formatted table.

```
mean_stan<-matrix(data=NA, nrow = 6, ncol = 3)
m<-1
for (i in 1:(length(UN3)-1)){
  mean_stan[m,]<-c(colnames(UN3[i]),mean(na.omit(UN3[,i])),sd(na.omit(UN3[,i])))
   m<-m+1
  }


stats.data <- data.frame(mean_stan)
colnames(stats.data)<-c("name","mean","sd")
knitr::kable(stats.data)
```

name	mean	sd
ModernC	38.7171052631579	22.6366103759673
Change	1.41837320574163	1.13313267030361
PPgdp	6527.38805970149	9325.18855244529
Frate	48.3053892215569	16.5324480416909
Pop	30281.8714278846	120676.694478229
Fertility	3.214	1.70691793716661

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?

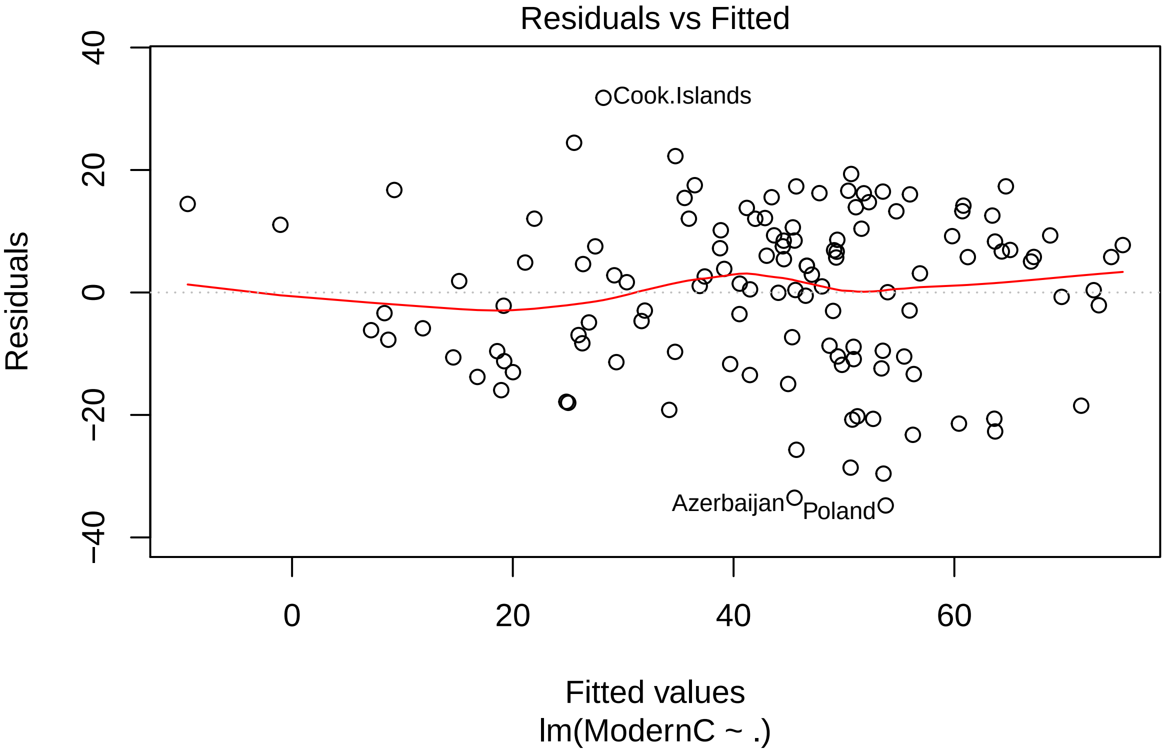
```
gg<-ggpairs(na.omit(UN3),columns=c(2:7,1))
gg
```

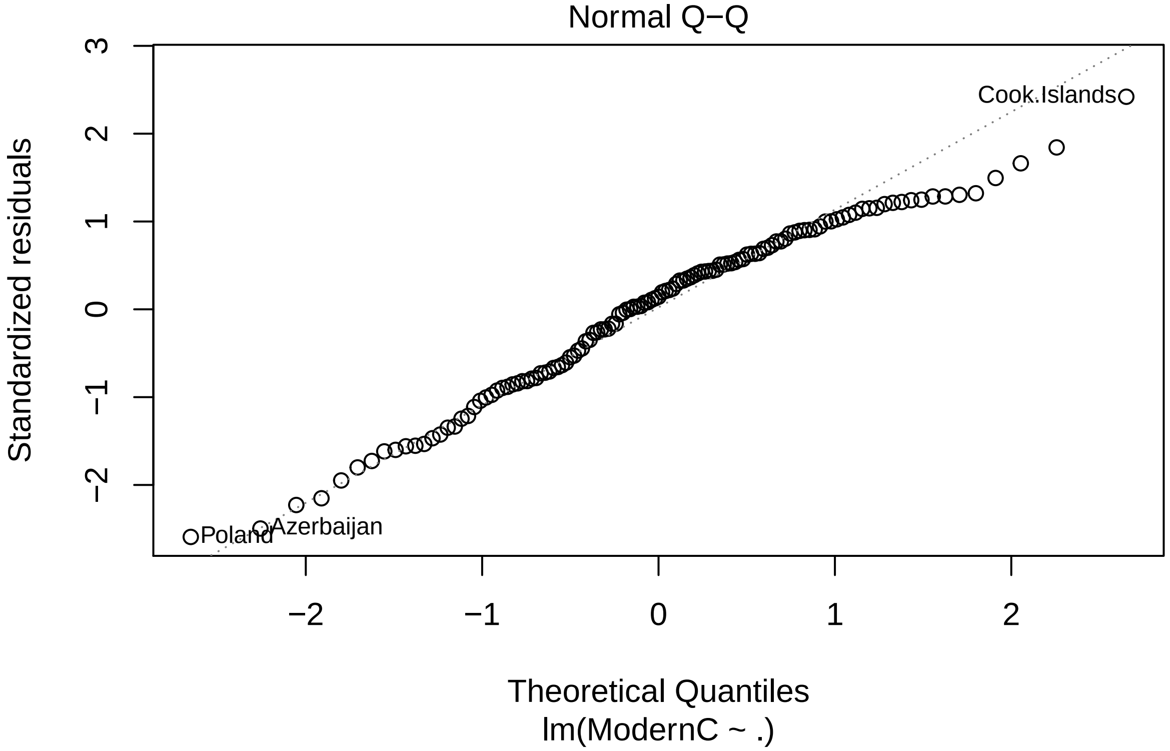


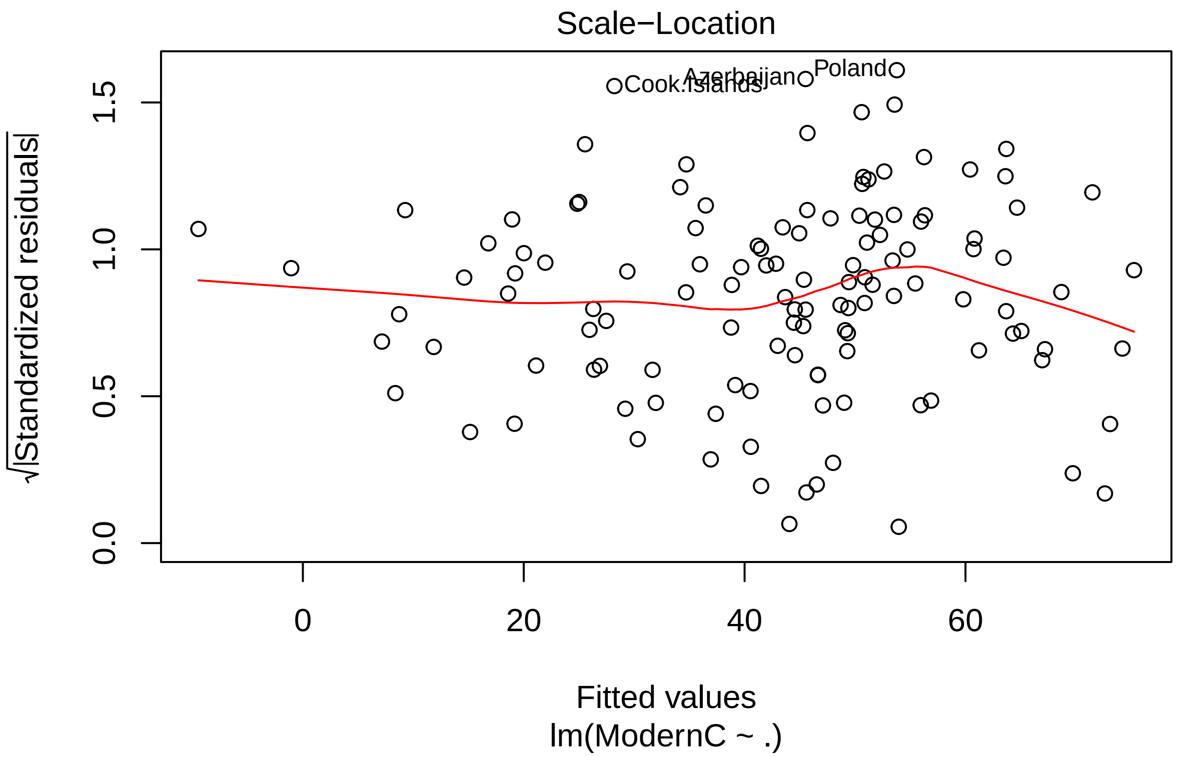
Answers: it seems that fertility, purban, ppdgp, and change are useful in predicting modernC. (corr coeff > 0.5) ## Model Fitting

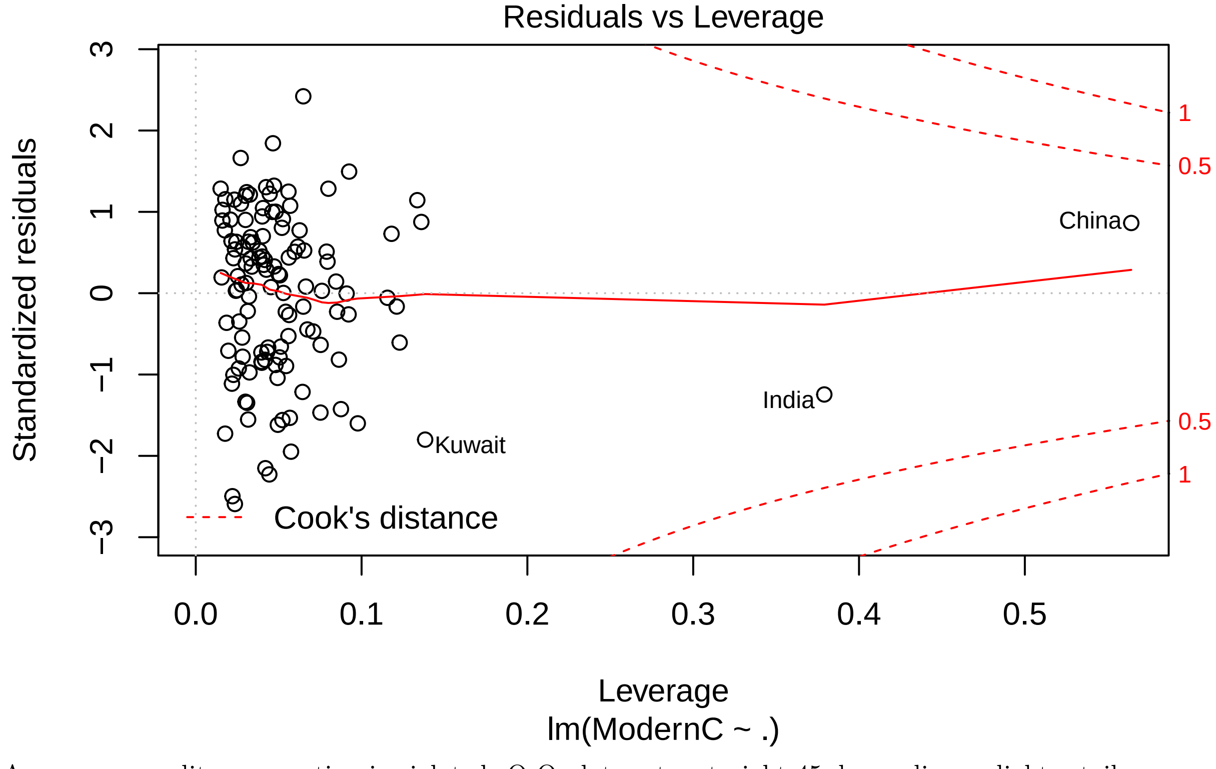
4. Use the `lm()` function to perform a multiple linear regression with `ModernC` as the response and all other variables as the predictors, using the formula `ModernC ~ .`, where the `.` includes all remaining variables in the dataframe. Create diagnostic residual plot from the linear model object and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

```r
g<-lm(ModernC~., data=UN3)
plot(g)
```


Residuals vs Fitted
Residuals
40
20
0
−20
−40
Cook.Islands
Azerbaijan
Poland
0
20
40
60
Fitted values
lm(ModernC ~ .)


Normal Q−Q
Standardized residuals
3
2
1
0
−1
−2
Cook.Islands
Poland
Azerbaijan
−2
−1
0
1
2
Theoretical Quantiles
lm(ModernC ~ .)





Answer: normality assumption is violated: Q-Q plot, not a straight 45-degree line: a lighter tail. others look good, except for figure 3: some non-random stuff in it

5. Using the Box-Tidwell `boxTidwell` from library `car` 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$Change_pos<-UN3$Change+5
#UN3_NA<-na.omit(UN3)
#m<-cbind(UN3_NA$ModernC,UN3_NA$Change_pos,UN3_NA$PPgdp,UN3_NA$Frate,UN3_NA$Pop,UN3_NA$Fertility,UN3_NA
#k<-powerTransform(m)
#k

UN3_NAA<-na.omit(UN3)
UN3_NAA$Change_pos<-NULL
k_bcn<-powerTransform(as.matrix(UN3_NAA[,-1])~.,family="bcnPower",data=UN3_NAA)
```

```
## Warning in sqrt(diag(res$invHess[1:d, 1:d, drop = FALSE])): NaNs produced
```

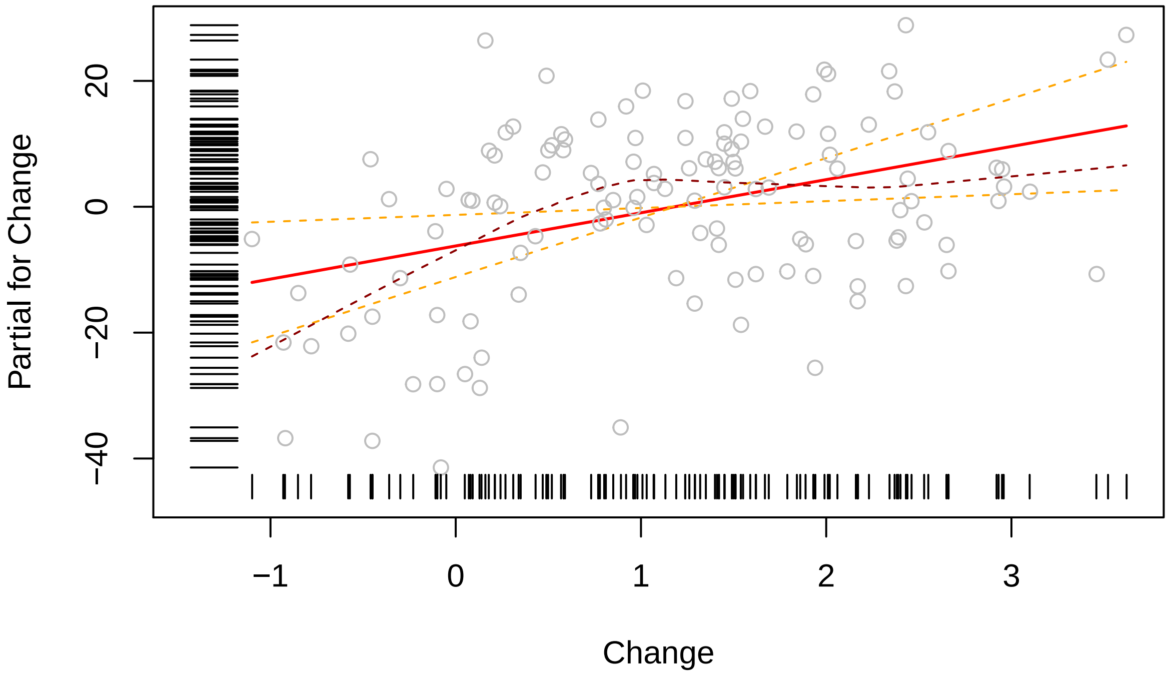
```
## Warning in sqrt(diag(res$invHess[(d + 1):(2 * d), (d + 1):(2 * d), drop =
## FALSE])): NaNs produced
```

```
k_bcn
```

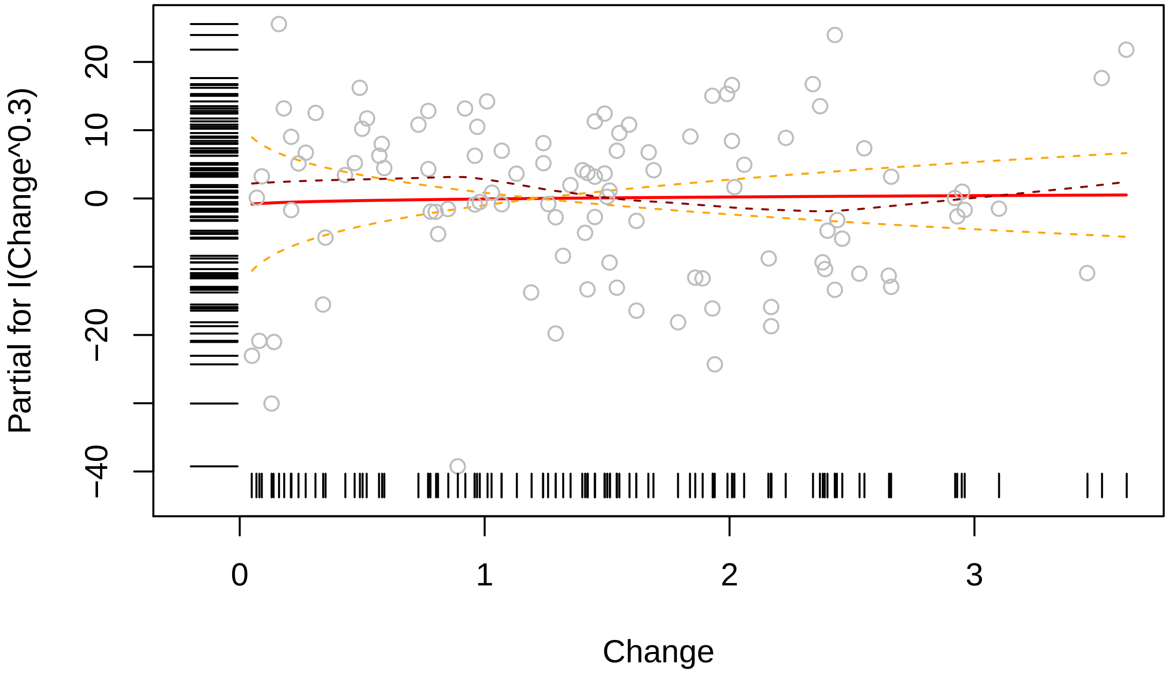
```
## Estimated transformation power, lambda
## [1] 0.2951946 0.9999996 0.9999975 0.3251072 0.9999648 0.9999841
## Estimated transformation location, gamma
## [1] 4.658143e+00 1.212076e+00 2.844007e-02 1.304196e+06 2.294808e-02
## [6] 1.428396e-01
```

In this problem, we can use powerTransform function to calculate the power of predictors. Rows with NA values are omitted. Using BCNpower family which accepts negative values will deal with the issues of "Change" (there are negative values inside)
Here we can see that Change and Pop have lambda values around 0.3, while all other 4 predictor variables have lambda values approximately 1. Therefore we transform Change and Pop according to their lambda values, while keeping the rest variables unchanged.

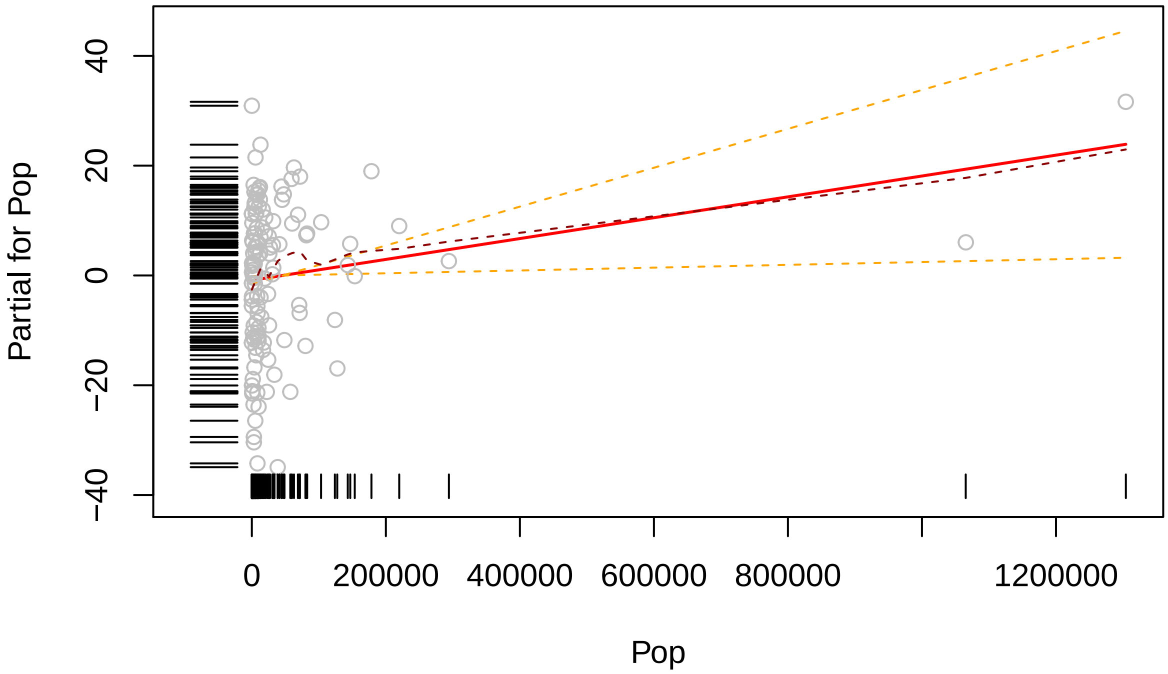
```
#z<-lm(ModernC~Change_pos+log(PPgdp)+Frate+log(Pop)+log(Fertility)+Purban,data=UN3_NA)
new_trans<-lm(ModernC~Purban+Frate+I(Change^0.3)+I(Pop^0.33)+Fertility+PPgdp,data=UN3_NAA)

termplot(g,terms="Change",partial.resid = T, se=T, rug=T,smooth = panel.smooth)
```



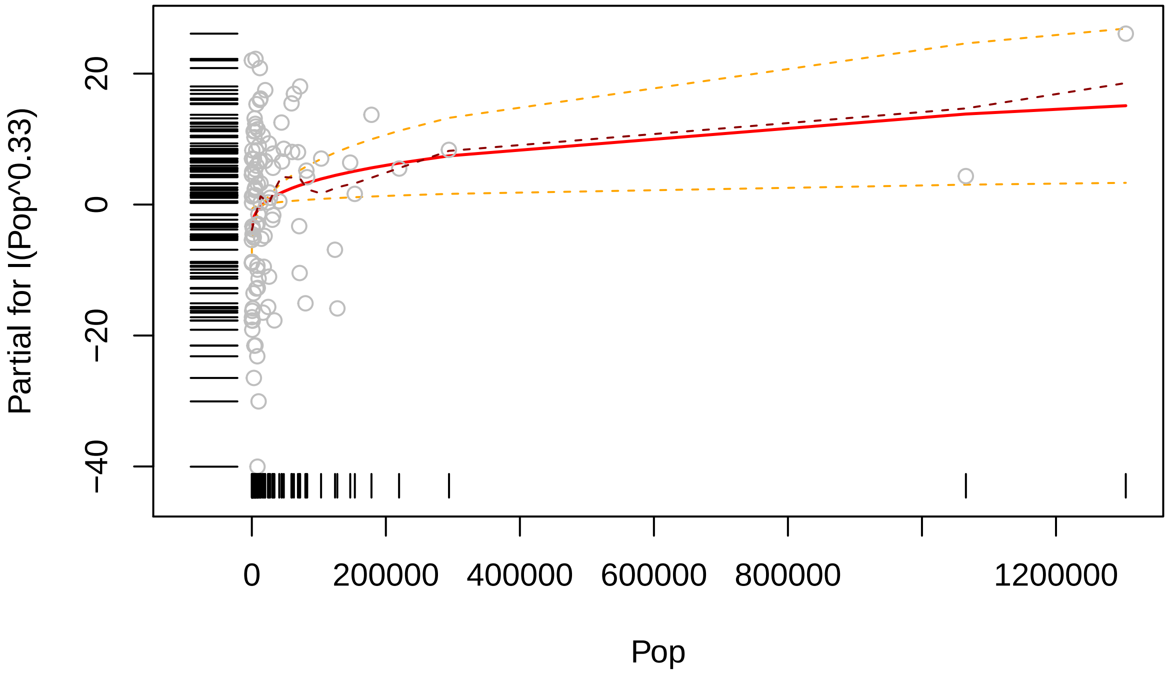
```r
#g is the original linear regression formula, in problem 3
termplot(new_trans,terms="I(Change^0.3)",partial.resid = T, se=T, rug=T,smooth = panel.smooth)
```



```r
termplot(g,terms="Pop",partial.resid = T, se=T, rug=T,smooth = panel.smooth)
```



```r
termplot(new_trans,terms="I(Pop^0.33)",partial.resid = T, se=T, rug=T,smooth = panel.smooth)
```



```r
#termplot(lm(ModernC ~.-Change,data=UN3_NA),terms="PPgdp",partial.resid = T, se=T, rug=T,smooth = panel
#termplot(lm(ModernC ~.-Change-PPgdp+log(PPgdp),data=UN3_NA),terms="log(PPgdp)",partial.resid = T, se=T
#termplot(lm(ModernC ~.-Change,data=UN3_NA),terms="Pop",partial.resid = T, se=T, rug=T,smooth = panel.s
#termplot(lm(ModernC ~.-Change-Pop+log(Pop),data=UN3_NA),terms="log(Pop)",partial.resid = T, se=T, rug=
#termplot(lm(ModernC ~.-Change,data=UN3_NA),terms="Fertility",partial.resid = T, se=T, rug=T,smooth = p
#termplot(lm(ModernC ~.-Change-Fertility+log(Fertility),data=UN3_NA),terms="log(Fertility)",partial.res
```

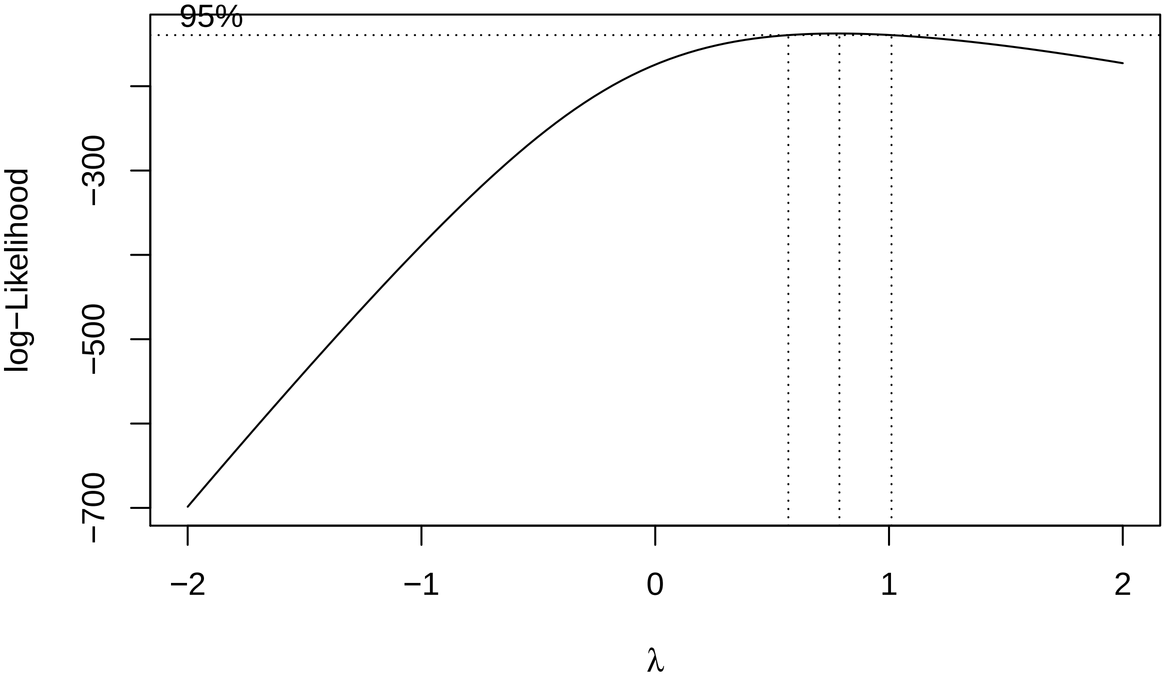
Comparing each pair of termplots, we can see that: 1. it seems that Change^0.3 is a little bit better than Change. 2. it seems that Pop^0.33 fits better.

We then look at addv plots to determine whether tranformations should take place.

```
#mod1 = lm(ModernC ~ .-Change, data=UN3_NA)
#avPlots(mod1)
#mod2 = lm(ModernC ~ .-Change-Pop+log(Pop), data=UN3_NA)
#avPlots(mod2)
#mod3 = lm(ModernC ~ .-Change-Pop-PPgdp+log(PPgdp)+log(Pop), data=UN3_NA)
#avPlots(mod3)
#mod4=lm(ModernC ~ .-Change-Pop-PPgdp-Fertility+log(PPgdp)+log(Pop)+log(Fertility), data=UN3_NA)
#avPlots(mod4)
```

By looking and comparing these addv plots, we can see that transferring PPgdp and Pop is necessary, while there is no need to transfer Fertility.
6. Given the selected transformations of the predictors, select a transformation of the response and justify.

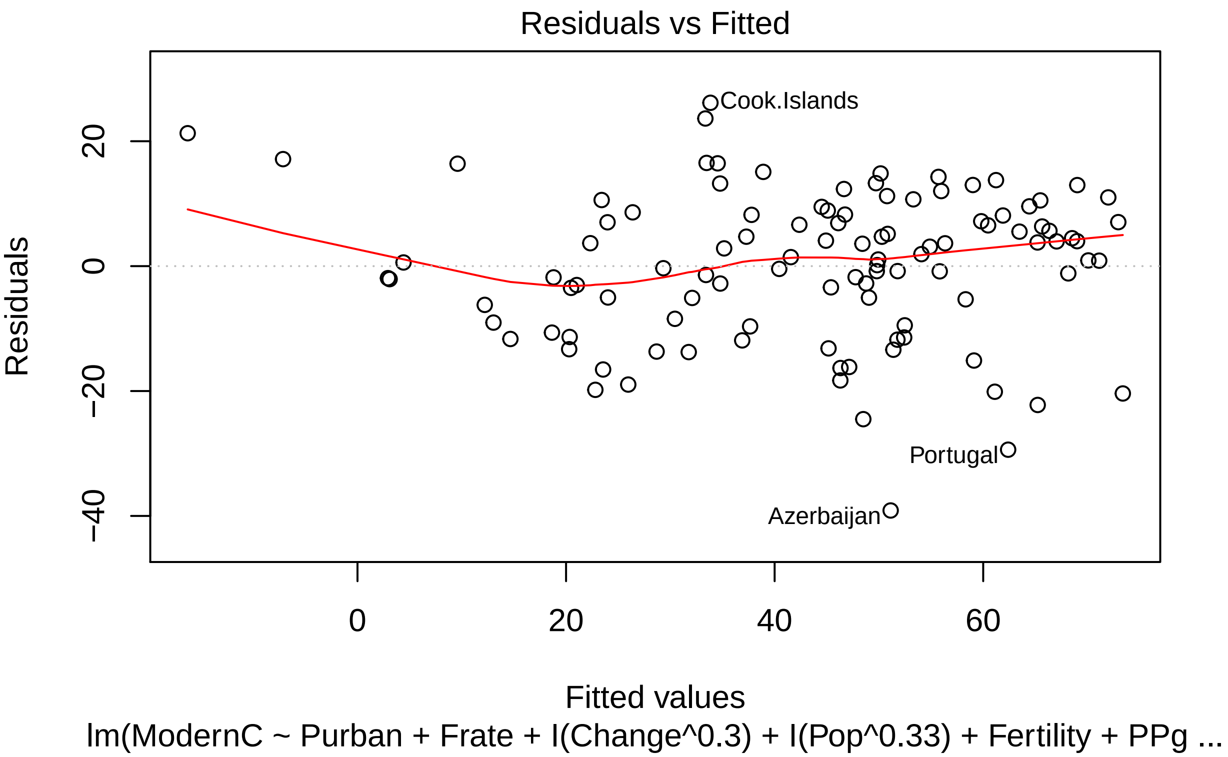
```
new_trans<-lm(ModernC~Purban+Frate+I(Change^0.3)+I(Pop^0.33)+Fertility+PPgdp,data=UN3_NAA)
boxcox(new_trans)
```

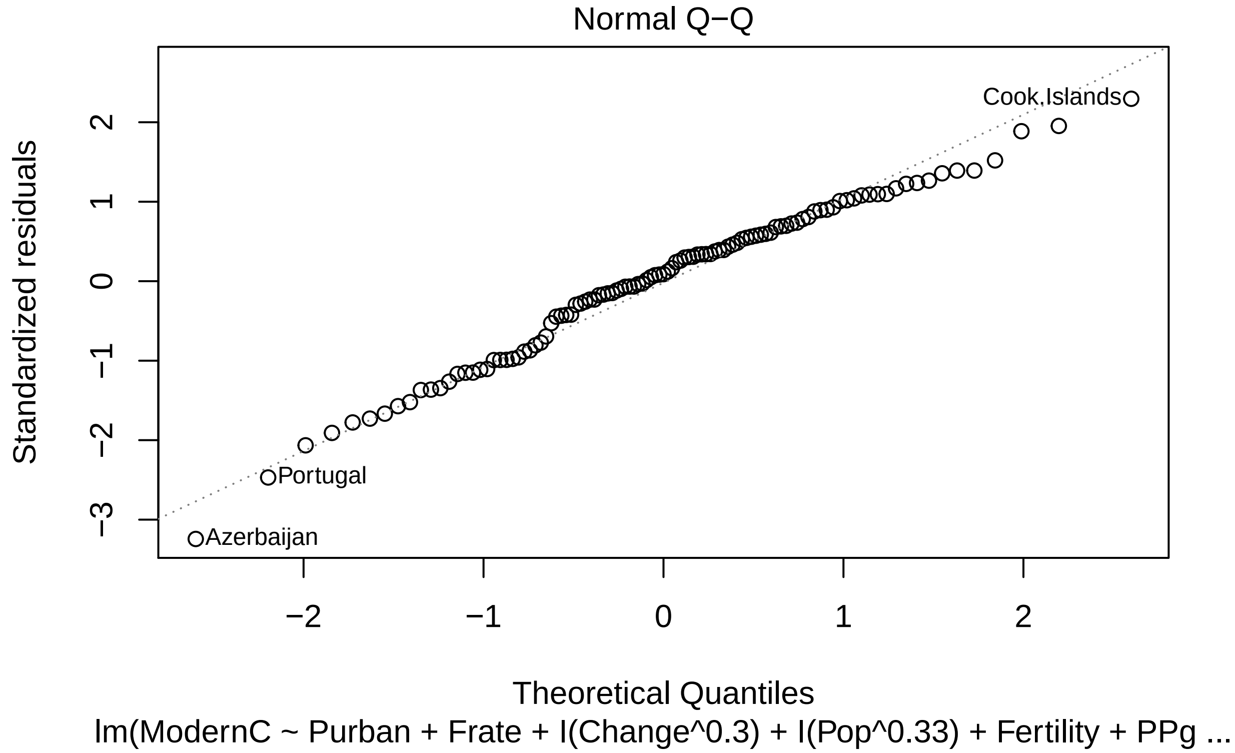


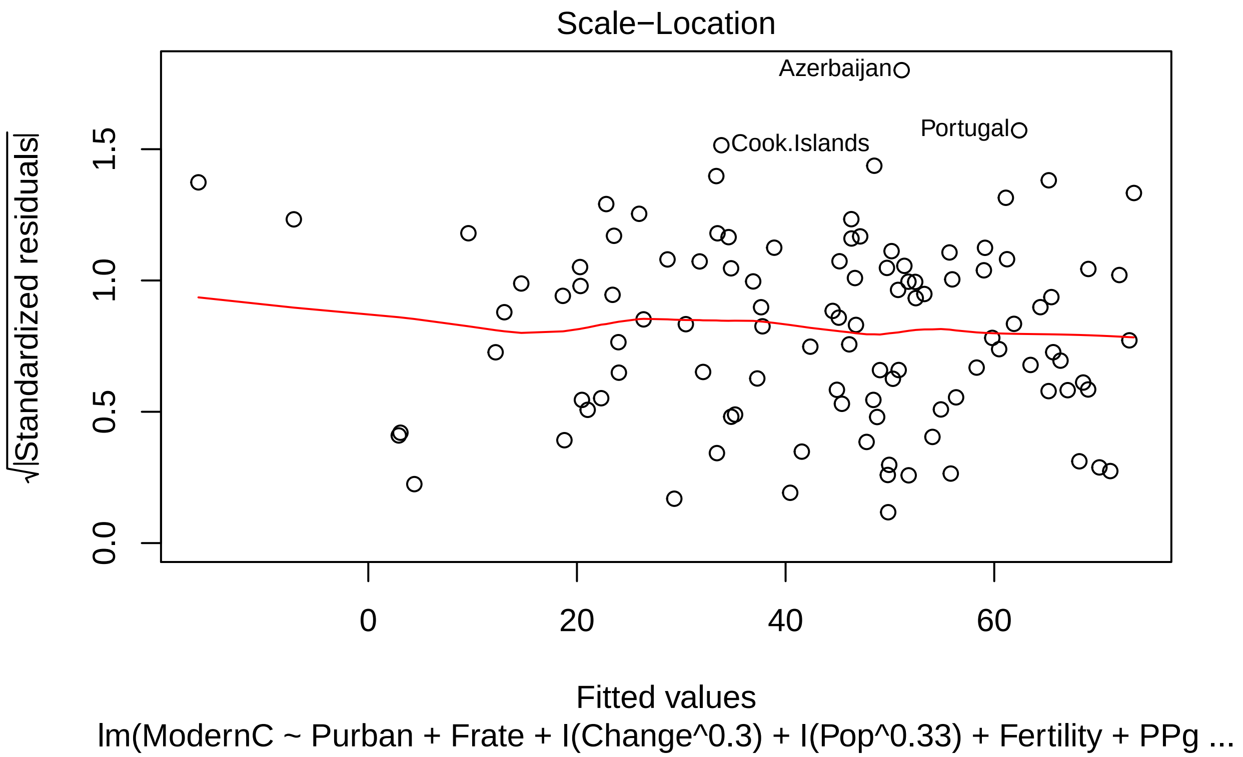
Answer: the graph indicates that lambda is around 1, so we don't need to transform the response.

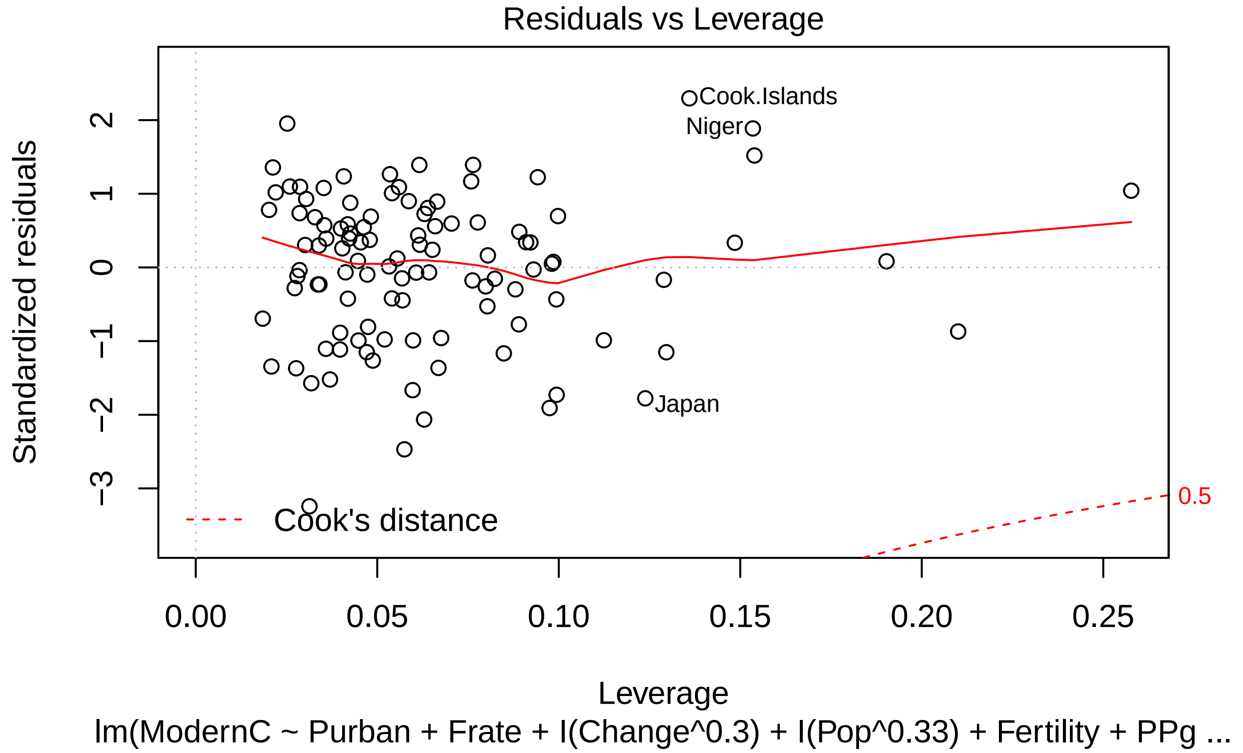
7. Fit the regression using the transformed variables. Provide residual plots and comment. Provide summaries of coefficients with 95% confidence intervals in a nice table with interpretations.

```
plot(new_trans)
```


Residuals vs Fitted
Residuals
20
0
−20
−40
Cook.Islands
Portugal
Azerbaijan
0
20
40
60
Fitted values
lm(ModernC ~ Purban + Frate + I(Change^0.3) + I(Pop^0.33) + Fertility + PPg ...


Normal Q−Q
Standardized residuals
2
1
0
−1
−2
−3
Cook.Islands
Portugal
Azerbaijan
−2
−1
0
1
2
Theoretical Quantiles
lm(ModernC ~ Purban + Frate + I(Change^0.3) + I(Pop^0.33) + Fertility + PPg ...





From the residual plots, we can see that the regression model of transformed variables looks better than the original one. Although there still exists a lighter tail in the normal Q-Q plot, it is much better than the original model. The residual vs leverage plot also becomes better.

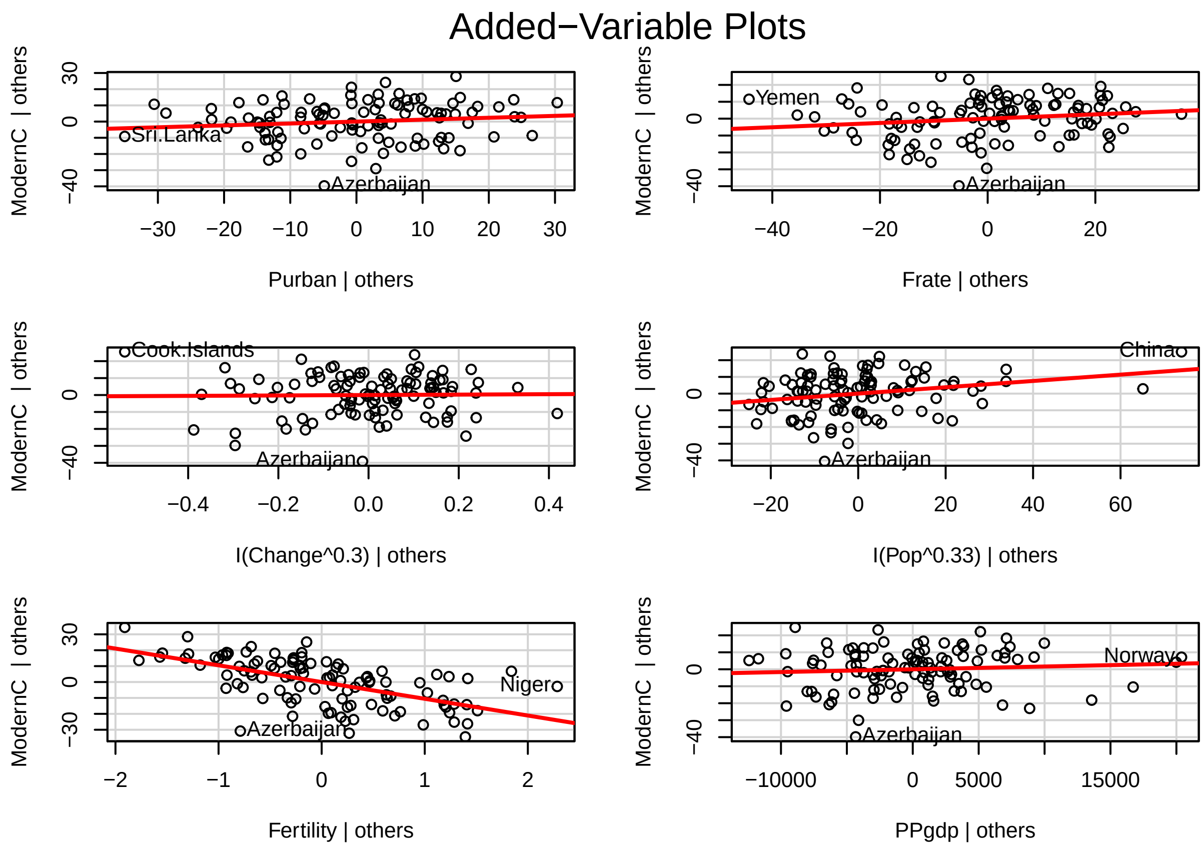
```
test<-matrix(data=NA, nrow = 6, ncol = 3)
cc<-summary(new_trans)
for (i in 2:length(coefficients(new_trans)))
```

```
{

  test[i-1,]<-c(rownames(cc$coefficients)[i],confint(new_trans, rownames(cc$coefficients)[i], level=0.9
}
ci_data<-data.frame(test)
colnames(ci_data)<-c("Var name","2.5%","97.5%")
knitr::kable(ci_data)
```

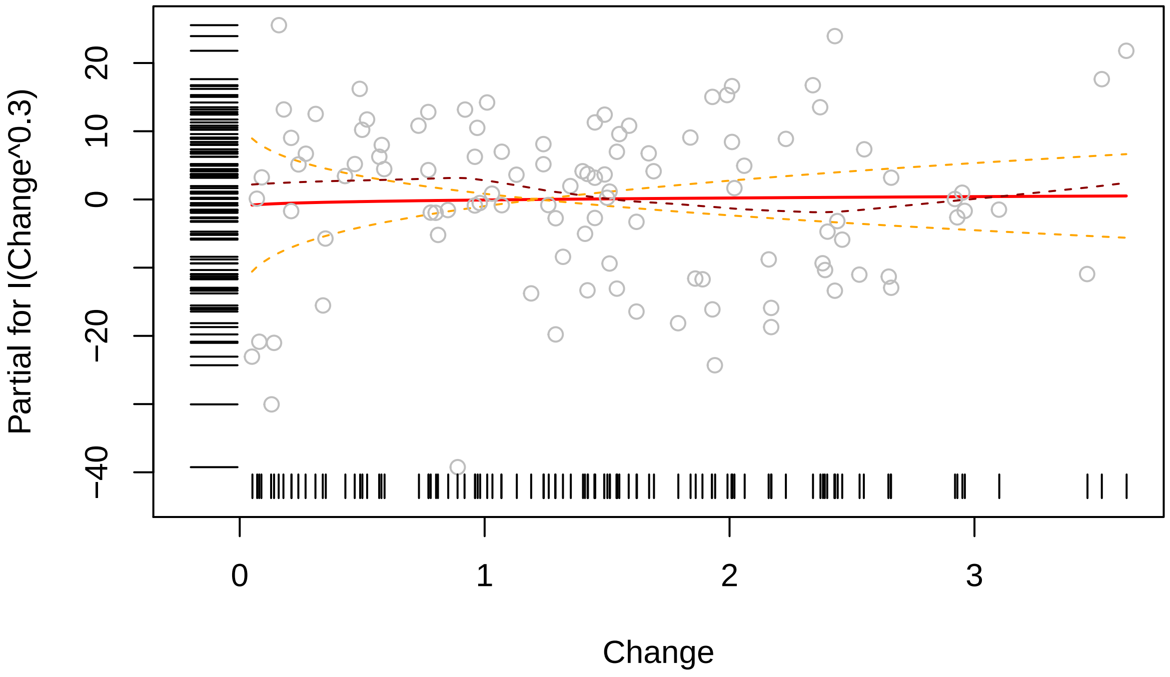
Var name	2.5%	97.5%
Purban	-0.0611365904590002	0.296582479567577
Frate	-0.0193813327722561	0.274195844975963
I(Change^0.3)	-13.5677697371068	16.0740167177822
I(Pop^0.33)	0.0428053120616708	0.335821186157936
Fertility	-13.3720702738129	-7.67866225503625
PPgdp	-0.000233343548163657	0.000559403607289659

8. Examine added variable plots and term plots for you model above. Is it likely that any of the localities are influential for any of the terms? Which localities? Which terms?

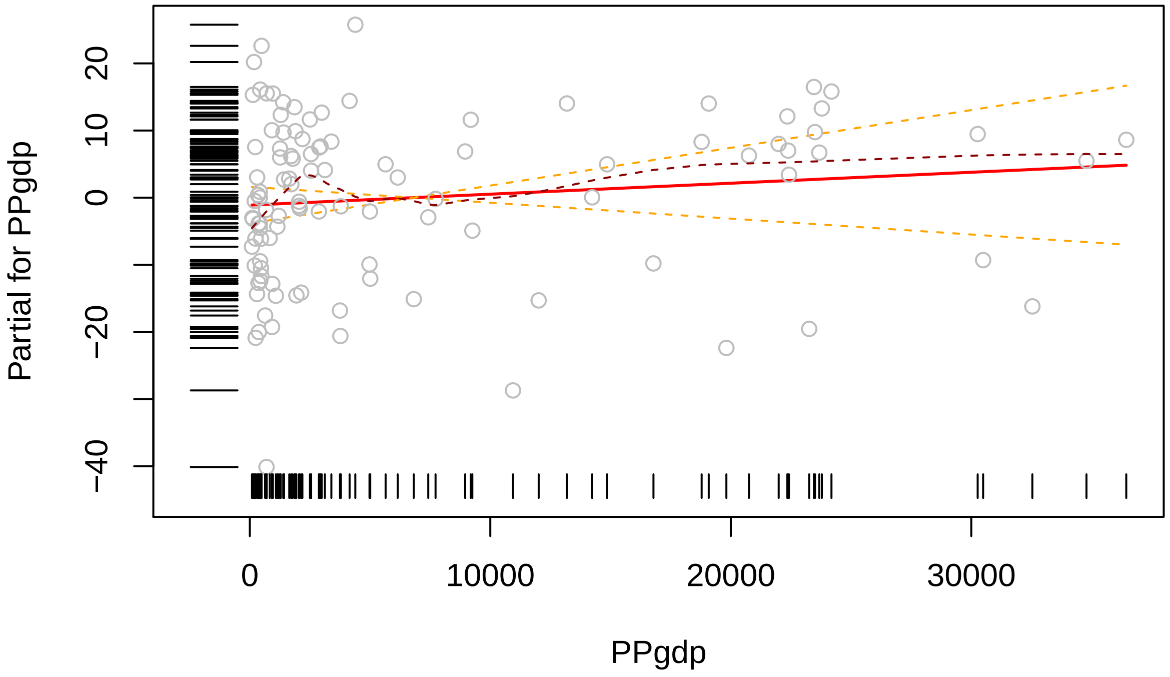
```
avPlots(new_trans,id.n=1)
```



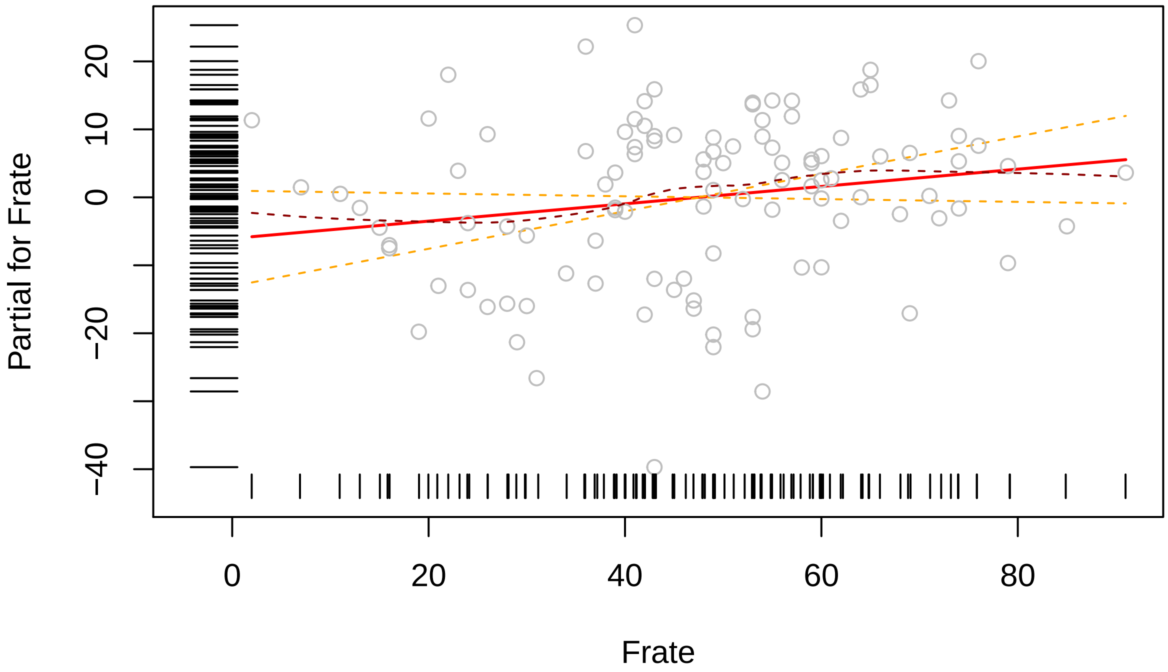
```
termplot(new_trans,terms="I(Change^0.3)",partial.resid = T, se=T, rug=T,smooth = panel.smooth)
```



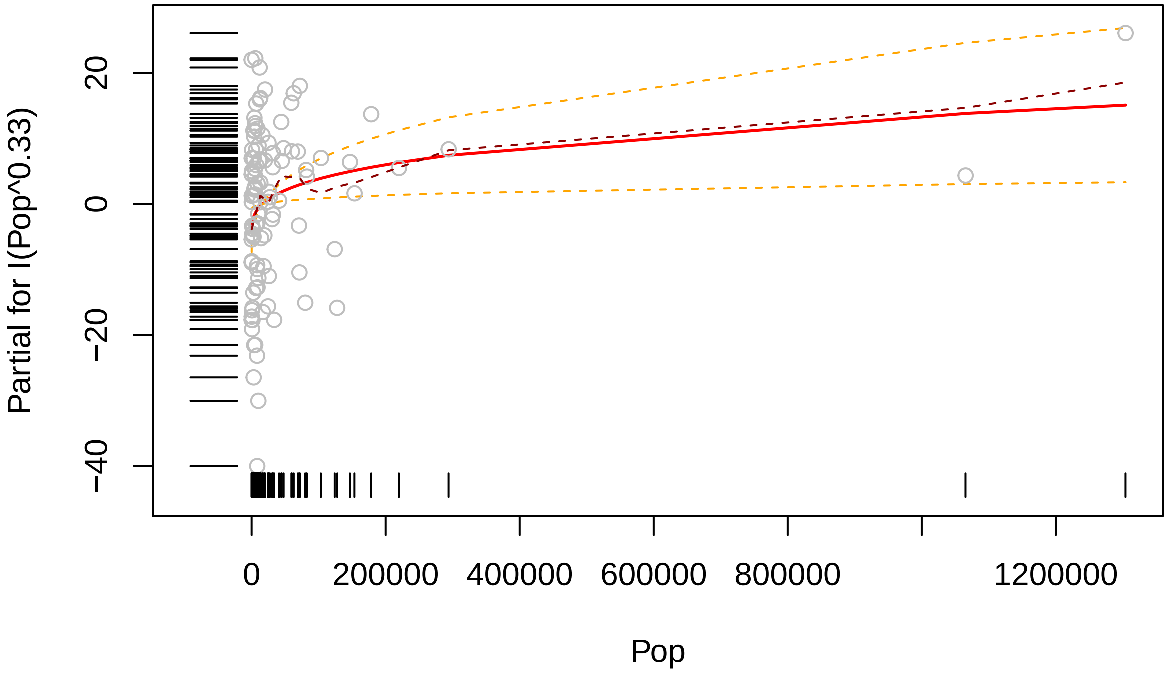
```
termplot(new_trans,terms="PPgdp",partial.resid = T, se=T, rug=T,smooth = panel.smooth)
```



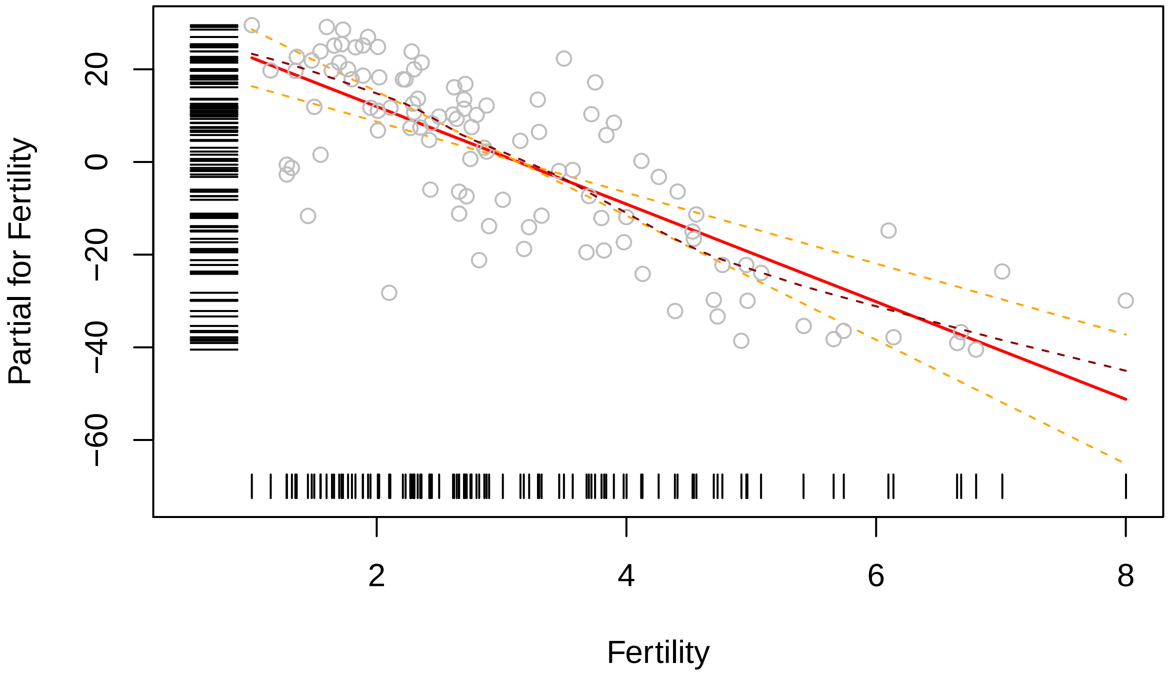
```
termplot(new_trans,terms="Frate",partial.resid = T, se=T, rug=T,smooth = panel.smooth)
```



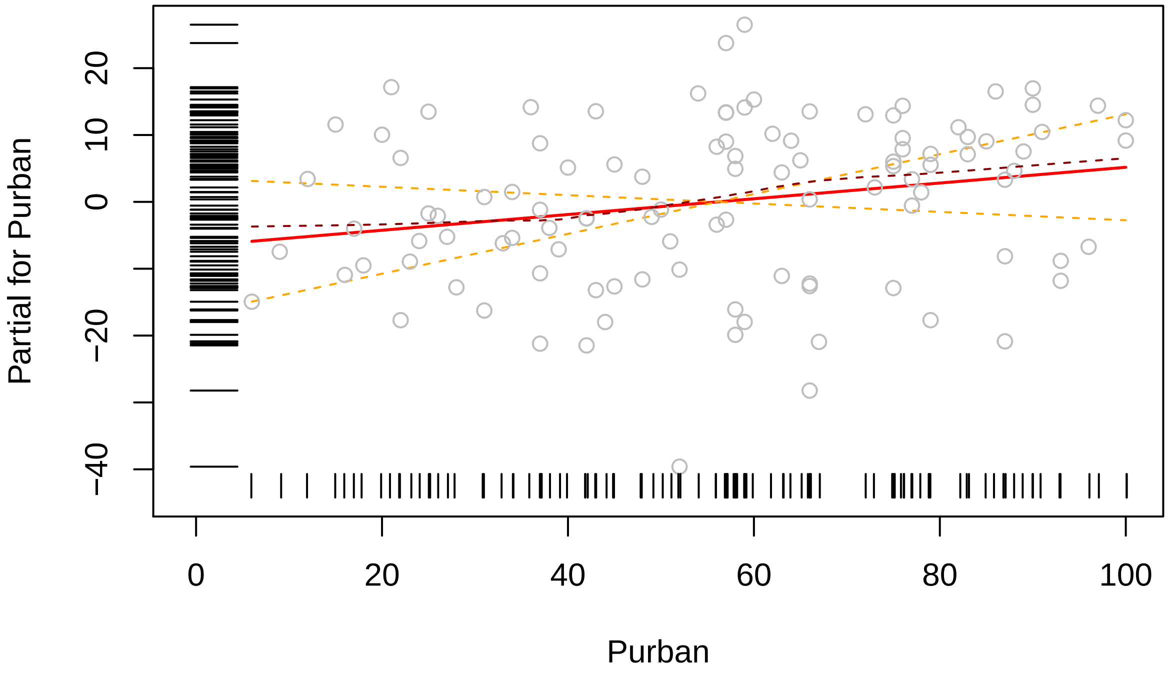
```
termplot(new_trans,terms="I(Pop^0.33)",partial.resid = T, se=T, rug=T,smooth = panel.smooth)
```



```
termplot(new_trans,terms="Fertility",partial.resid = T, se=T, rug=T,smooth = panel.smooth)
```



```r
termplot(new_trans,terms="Purban",partial.resid = T, se=T, rug=T,smooth = panel.smooth)
```



```r
#termplot(lm(ModernC ~.-Change_pos-Pop-PPgdp+log(Pop)+log(PPgdp),data=UN3_NA),terms="log(Pop)",partial.
#termplot(lm(ModernC ~.-Change_pos-Pop-PPgdp+log(Pop)+log(PPgdp),data=UN3_NA),terms="Fertility",partial
#termplot(lm(ModernC ~.-Change_pos-Pop-PPgdp+log(Pop)+log(PPgdp),data=UN3_NA),terms="Frate",partial.res
#termplot(lm(ModernC ~.-Change_pos-Pop-PPgdp+log(Pop)+log(PPgdp),data=UN3_NA),terms="Change",partial.re
#termplot(lm(ModernC ~.-Change_pos-Pop-PPgdp+log(Pop)+log(PPgdp),data=UN3_NA),terms="log(PPgdp)",partia
#termplot(lm(ModernC ~.-Change_pos-Pop-PPgdp+log(Pop)+log(PPgdp),data=UN3_NA),terms="Purban",partial.re
```

Answer:
1. from addv plot, we can see that for I(Pop^0.33), the locality seems to be China.

2. from termplot for I(Pop^0.33), it seems that there are 2 localities: China and India.
There seems to be no obvious localities in other plots.

9. Are there any outliers in the data? Explain. If so refit the model after removing any outliers.

Answer: I don't think so. According to the 4th residual plots (the one with cook's distance), we can see that there does not exist points with high cook's distance that will influence the result.

Summary of Results

10. Provide a brief paragraph summarizing your final model and findings suitable for the US envoy to the UN after adjusting for outlierd or influential points.

Theory

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

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

where h_{ii} is the ith diagonal element of $H = X(X^TX)^{-1}X^T$.
Start with the equation that we want to show, (1):

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

Multiply

$$(X^TX)(1 - h_{ii})$$

to each side of (1), WTS:

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

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

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

Multiply

$$X_{(i)}^T X_{(i)}$$

to each side again, then the equation becomes:

$$x_i x_i^T (1 - h_{ii}) = x_i x_i^T (X^TX)^{-1}(X^TX - x_i x_i^T)$$

$$x_i x_i^T (1 - h_{ii}) = x_i x_i^T (I - (X^TX)^{-1} x_i x_i^T)$$
$$x_i x_i^T h_{ii} = x_i x_i^T (X^TX)^{-1} x_i x_i^T$$

Notice that

$$h_{ii} = x_i(X^TX)^{-1}x_i)$$

, and it is a scalar So the equation we want to show turns out to be

$$x_i x_i^T h_{ii} = x_i h_{ii} x_i^T$$

Which is obvious to be true.

Therefore, starting with this equation and going back, we can prove

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

12. Use 11 to show that

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

where $\hat{\beta}_{(i)} = (X_{(i)}^T X_{(i)})^{-1} X_{(i)}^T Y_{(i)}$ and $e_i = y_i - x_i^T \hat{\beta}$. *Hint write* $X_{(i)}^T Y_{(i)} = X^T Y - x_i y_i$.

$$\hat{\beta}_{(i)} = (X_{(i)}^T X_{(i)})^{-1} X_{(i)}^T Y_{(i)}$$

$$= (X_{(i)}^T X_{(i)})^{-1} = [(X^T X)^{-1} + \frac{(X^T X)^{-1} x_i x_i^T (X^T X)^{-1}}{1 - h_{ii}}][X^T Y - x_i y_i]$$

$$= (X^T X)^{-1} X^T Y + (X^T X)^{-1} [\frac{x_i x_i^T (X^T X)^{-1} X^T Y - x_i y_i (1 - h_{ii}) - x_i x_i^T (X^T X)^{-1} x_i y_i}{1 - h_{ii}}]$$

$$= \hat{\beta} + \frac{(X^T X)^{-1}}{1 - h_{ii}} [x_i x_i^T (X^T X)^{-1} X^T X \hat{\beta} - x_i y_i + x_i y_i h_{ii} - x_i h_{ii} y_i]$$

$$= \hat{\beta} + \frac{(X^T X)^{-1}}{1 - h_{ii}} [x_i x_i^T \hat{\beta} - x_i y_i]$$

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