HW2 STA521 Fall18

Ziwei Zhu zz169 sophiazzw7

Due September 24, 2018 9am

Backgound Reading

Readings: Chapters 3-4 in Weisberg Applied Linear Regression

Exploratory Data Analysis

```
include = FALSE
suppressWarnings(library(car))
```

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

```
library(carData)
library(alr3)
data(UN3, package="alr3")
help(UN3)
library(car)
library(ggplot2)
library(knitr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## 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
```

```
library(knitr)
library(ggplot2)
library(GGally)
```

```
##
## Attaching package: 'GGally'

## The following object is masked from 'package:dplyr':
##
##     nasa
```

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

```r
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
str(UN3)
```

```
## 'data.frame':    210 obs. of  7 variables:
##  $ ModernC  : int  NA NA 49 NA NA NA 51 NA 22 NA ...
##  $ Change   : num  3.88 0.68 1.67 2.37 2.59 3.2 0.53 1.17 -0.45 2.02 ...
##  $ PPgdp    : int  98 1317 1784 NA 14234 739 8461 7163 687 NA ...
##  $ Frate    : int  NA NA 7 42 NA NA 63 44 51 53 ...
##  $ Pop      : num  23897 3167 31800 57 64 ...
##  $ Fertility: num  6.8 2.28 2.8 NA NA 7.2 NA 2.44 1.15 NA ...
##  $ Purban   : int  22 43 58 53 92 35 37 88 67 51 ...
```

All of the variables except Purban have missing values. All of the vairables are quantitative vaariables.

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

```r
sd = as.data.frame(round(apply(UN3[1:7],2,sd,na.rm=TRUE),3))
mean = as.data.frame(round(apply(UN3[1:7],2,mean,na.rm=TRUE),3))
sdMean_table = cbind(sd,mean)
rm(sd,mean)
colnames(sdMean_table) = c('SD','Mean')
kable(sdMean_table)
```

	SD	Mean
ModernC	22.637	38.717
Change	1.133	1.418
PPgdp	9325.189	6527.388
Frate	16.532	48.305
Pop	120676.694	30281.871
Fertility	1.707	3.214
Purban	24.110	56.200

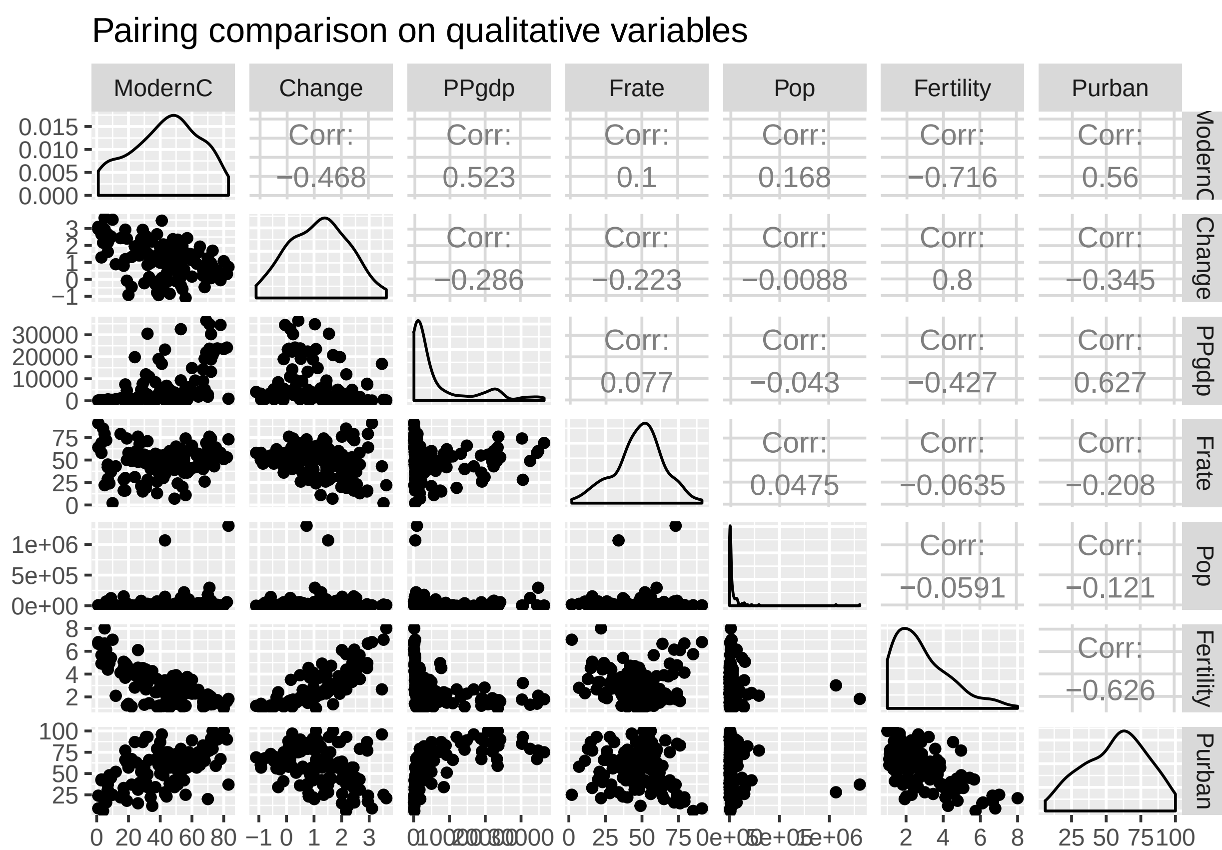
```r
sd = as.data.frame(round(apply(UN3[,1:7],2,sd,na.rm=TRUE),3))
mean = as.data.frame(round(apply(UN3[,1:7],2,mean,na.rm=TRUE),3))
sdMean_table = cbind(sd,mean)
rm(sd,mean)
```

```r
colnames(sdMean_table) = c('SD','Mean')
kable(sdMean_table)
```

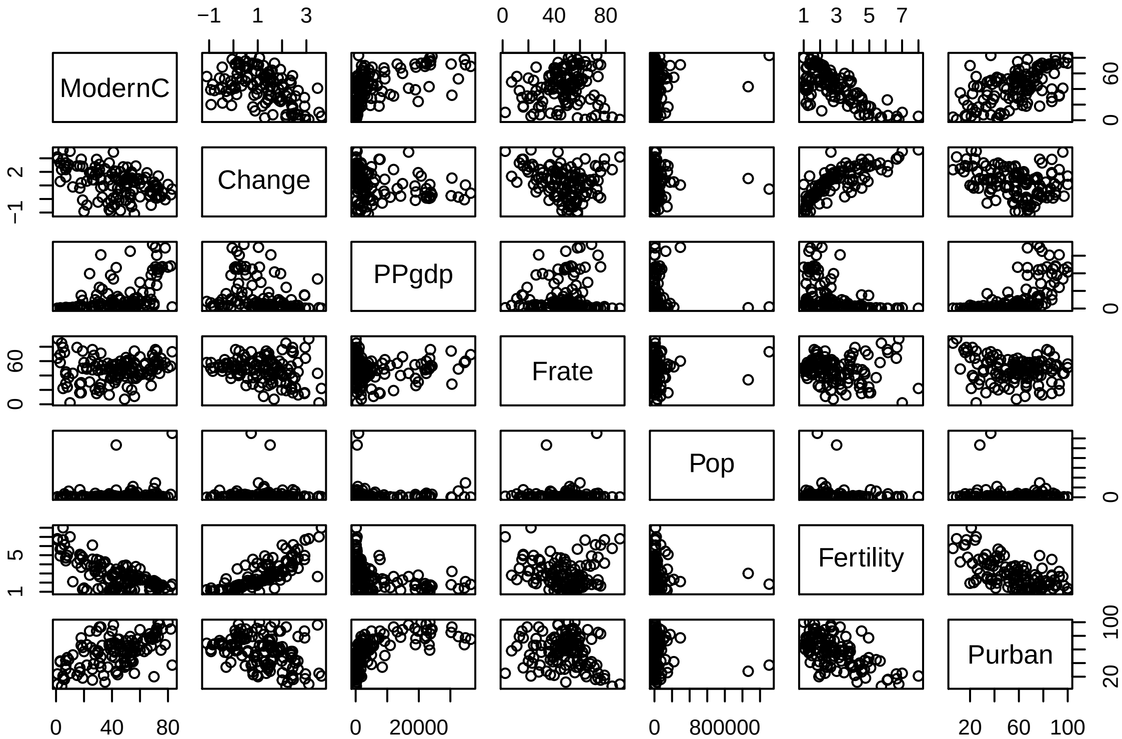
	SD	Mean
ModernC	22.637	38.717
Change	1.133	1.418
PPgdp	9325.189	6527.388
Frate	16.532	48.305
Pop	120676.694	30281.871
Fertility	1.707	3.214
Purban	24.110	56.200

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?

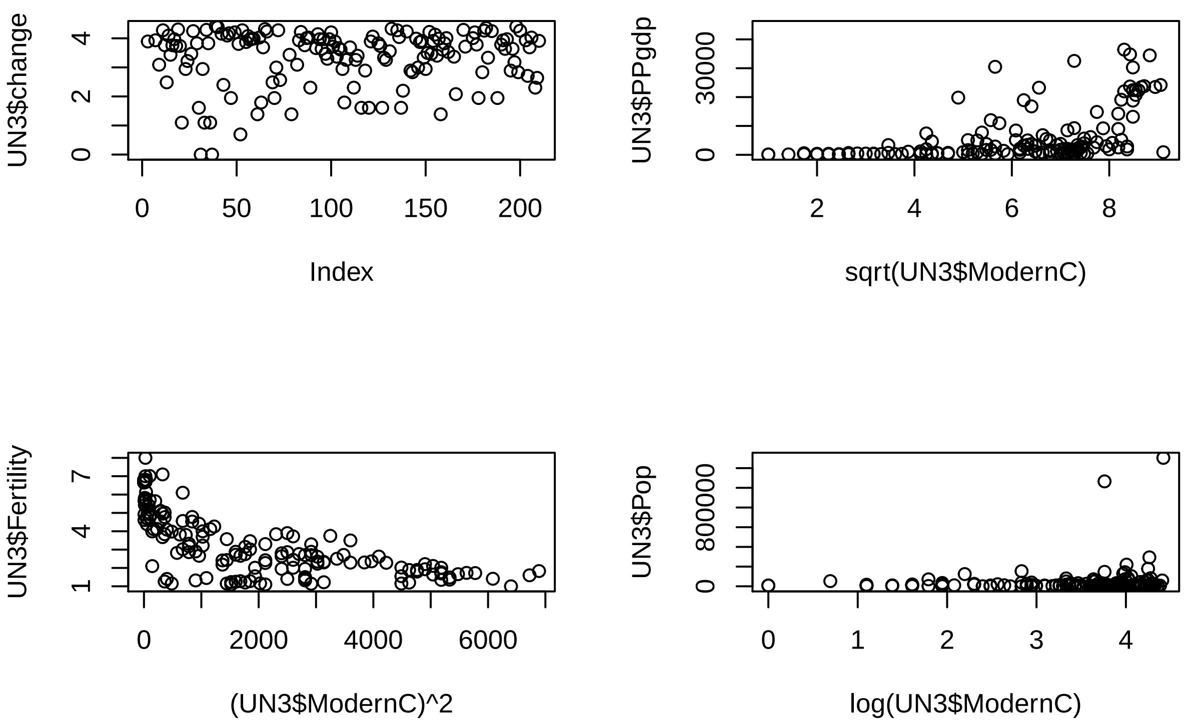
```r
UN3_0 = na.omit(UN3)
library(GGally)
ggpairs(UN3_0, progress = FALSE,title = "Pairing comparison on qualitative variables")
```




```r
pairs(UN3_0)
```




```r
par(mfrow = c(2, 2))
plot(log(UN3$ModernC), UN3$change)
plot(sqrt(UN3$ModernC), UN3$PPgdp)
plot((UN3$ModernC)^2, UN3$Fertility)
plot(log(UN3$ModernC), UN3$Pop)
```



From the graphs, i could identify that ModernC had nonlinear relationship with Pop and PPgdp, which may imply the need for further transformations Pop and ModernC has some potential outliers. Among all the precditors, Fertility would be the best variable to predict ModernC, since Fertility has the most linear relationship with "ModernC. And PPgdp may be of the most concern, since its relationship with ModernC seems most nonlinear.

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?

```r
coef(lm(ModernC ~ . , data= UN3_0))
```

```
##   (Intercept)        Change         PPgdp         Frate           Pop
##  5.529086e+01  5.268465e+00  5.300634e-04  1.232214e-01  1.899062e-05
##     Fertility        Purban
## -1.099843e+01  5.408230e-02
```

```r
anova(lm(ModernC ~ . , data= UN3_0))
```

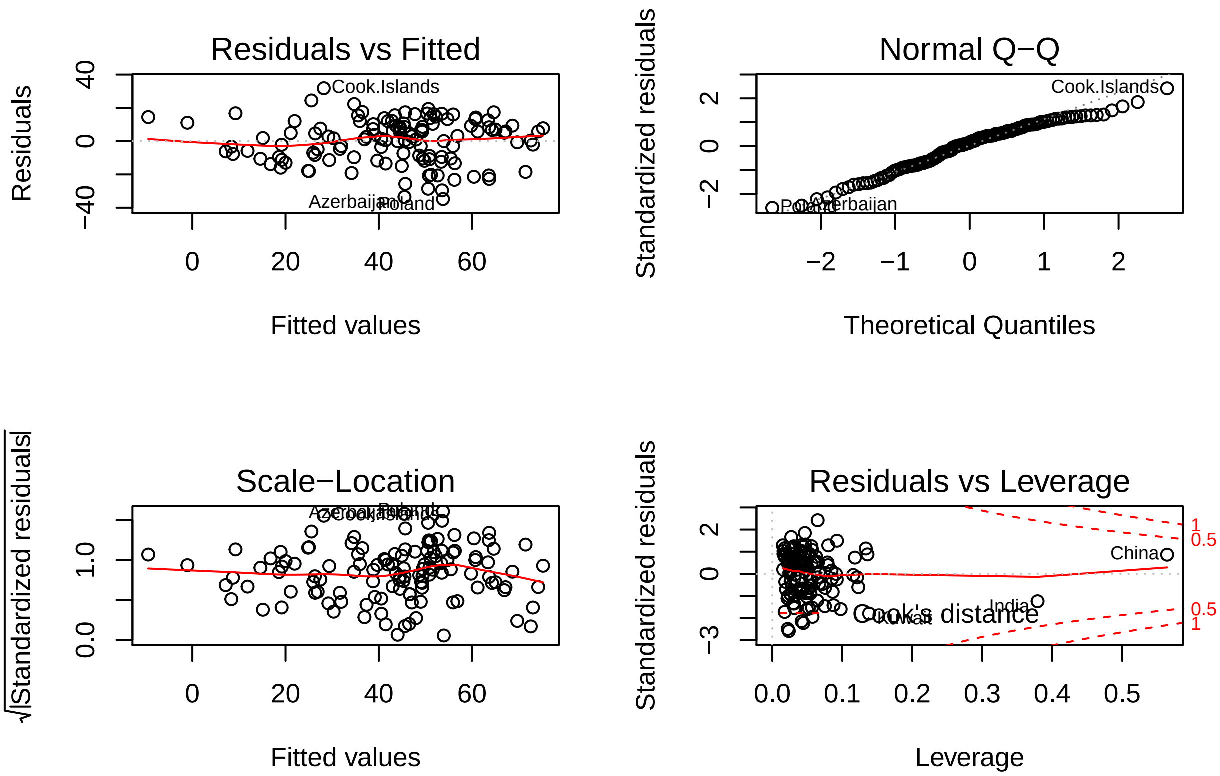
```
## Analysis of Variance Table
##
## Response: ModernC
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## Change      1 12493.0 12493.0 67.7356 2.846e-13 ***
## PPgdp       1  9407.7  9407.7 51.0076 8.162e-11 ***
## Frate       1     5.6     5.6  0.0303  0.862206
## Pop         1  1924.6  1924.6 10.4352  0.001602 **
## Fertility   1 11355.6 11355.6 61.5690 2.149e-12 ***
## Purban      1    62.6    62.6  0.3393  0.561344
## Residuals 118 21763.6   184.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
reg <- lm(ModernC ~ . , data= UN3_0)
summary(reg)
```

```
##
## Call:
## lm(formula = ModernC ~ ., data = UN3_0)
##
## 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
```

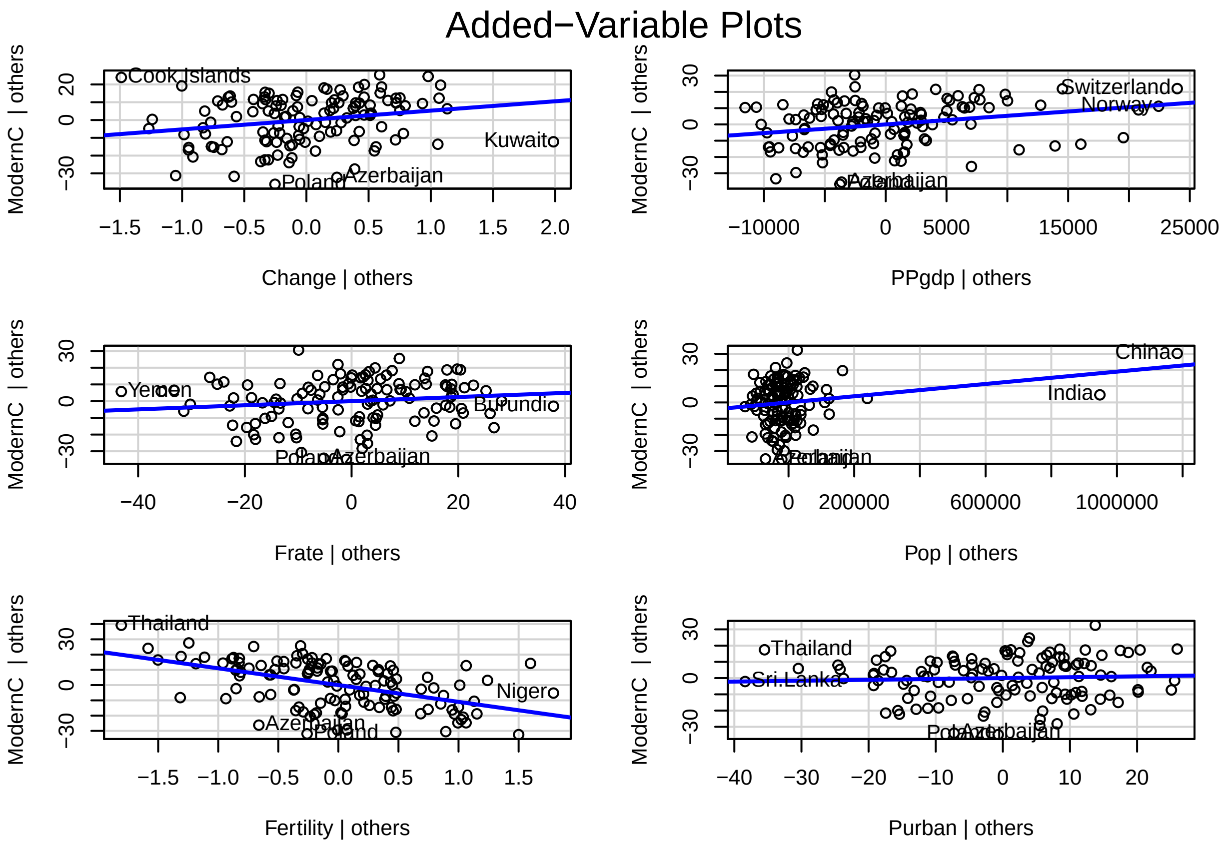
```r
par(mfrow = c(2, 2))
plot(reg)
```



We want residual randomly distributed around fitted line, and we saw the residual vs fitted graph looks fine. The normal QQ plot is showing a straight line trend rather than a curved shape,so we saw not necessiliy normality with a few outliers on the tails. heavy-tailed,quantile larger than normal value We wanted to see random pattern in the scale-location plot, and we kind of have it. For the residual vs.leverage plot, we could see India and China being marked out by R, but they do not appear to be influential. And 125 obeservations are used in my 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?

```r
car::avPlots(model = reg)
```



From the added-variable plots, i think the Pop is especially clustered and did not show linearlity, also the PPgdp shows a little clustering pattern, so i think transformation is needed for Pop and PPgdp.

From the graphs, Kuwaito and Cook's Islands are potential influential for Change. China and India are potential influential for Pop, since they may be responsible for the linear relationship seen on the graph.

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.

```
car::boxTidwell(ModernC~PPgdp+Pop,other.x=~Frate+Change+Fertility+Purban,data=UN3,max.iter=25, tol=0.00
```

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

```
powerTransform(as.matrix(UN3_0)~.,family="bcnPower",data=UN3_0)
```

```
## Estimated transformation power, lambda
## [1] 0.9999782 0.2951891 0.9999984 0.9999849 0.3251064 0.9994071 0.9999831
##
## Estimated location, gamma
## [1] 1.000000e-01 4.873502e+00 2.450958e+00 1.000000e-01 1.304196e+06
## [6] 1.000000e-01 1.000000e-01
```

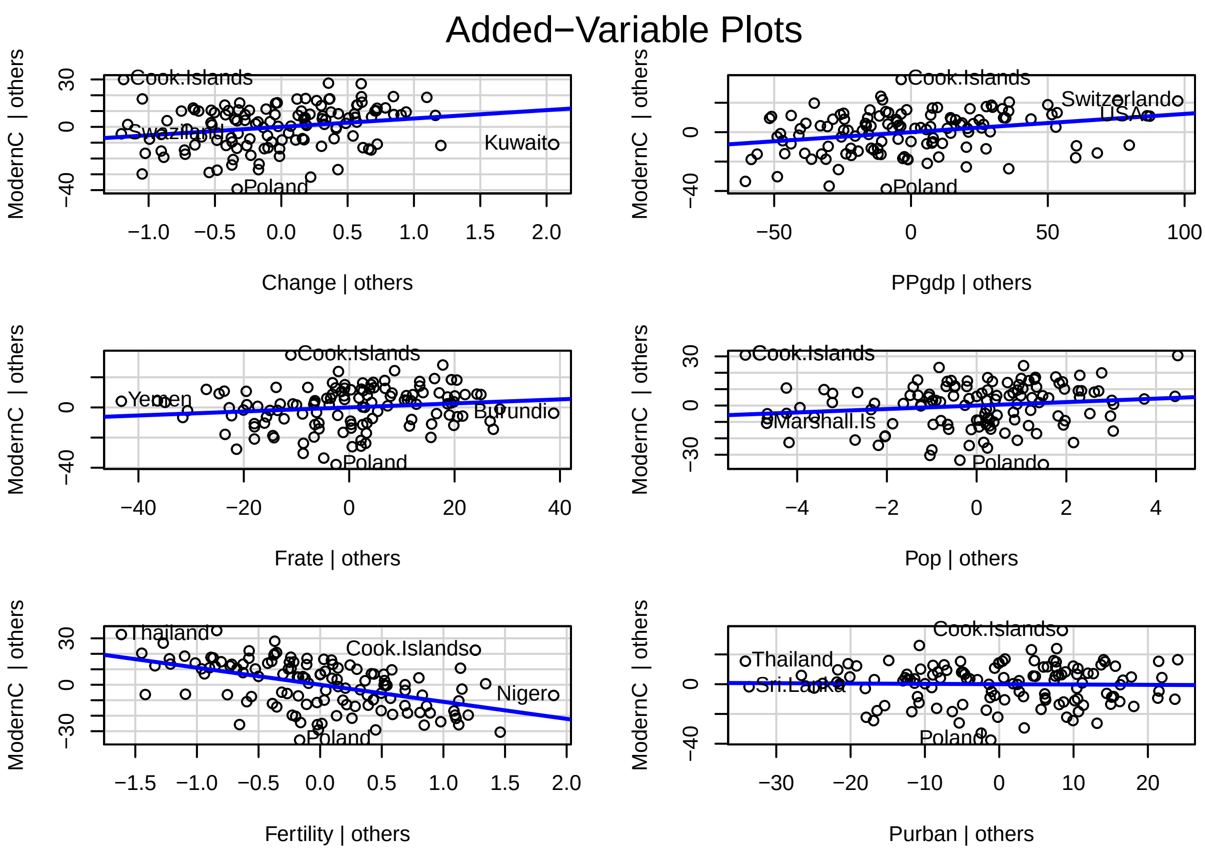
```
range(UN3_0['Change'])
```

```
## [1] -1.10  3.62
```

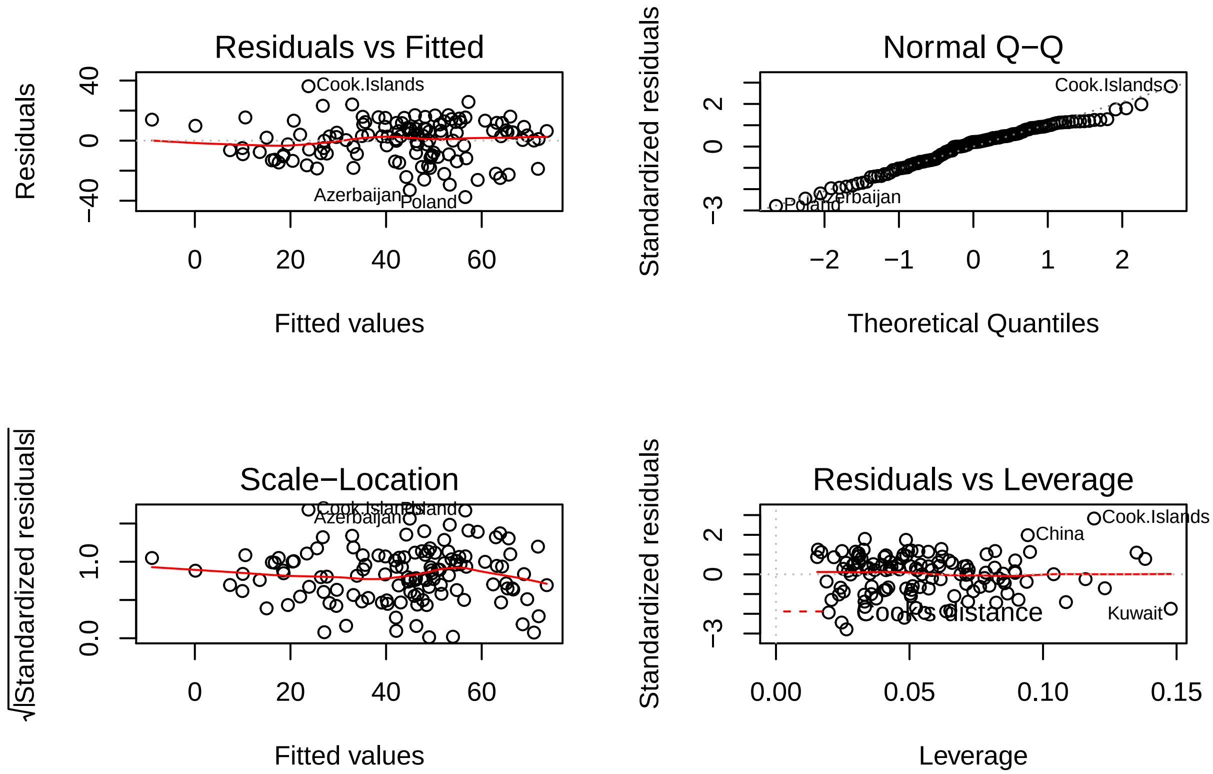
```r
UN3_1=UN3_0
UN3_1['Change']=UN3_0['Change']+2
powerTransform(as.matrix(UN3_1)~.,family="bcnPower",data=UN3_1)
```

```
## Estimated transformation power, lambda
## [1] 0.9999756 0.9991936 0.9999997 0.9999871 0.3250975 0.9993770 0.9999894
##
## Estimated location, gamma
## [1] 1.000000e-01 1.000000e-01 1.151552e+00 1.000000e-01 1.304196e+06
## [6] 1.000000e-01 1.000000e-01
```

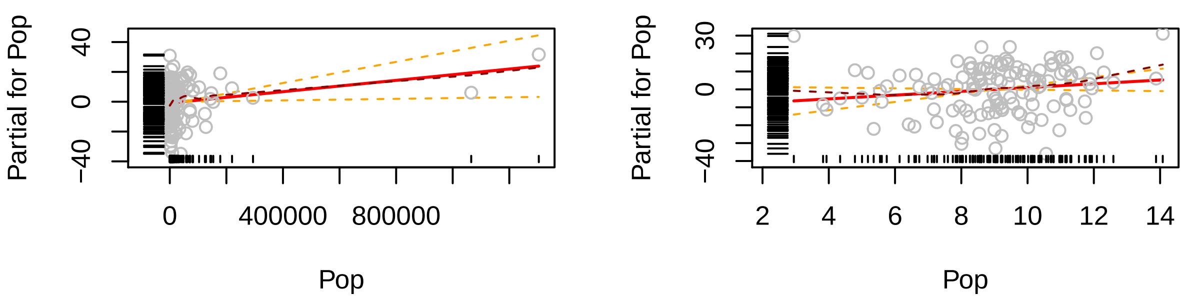
```r
UN3_trans=UN3_1
UN3_trans['Pop']=log(UN3_1['Pop'])
UN3_trans['PPgdp']=sqrt(UN3_1['PPgdp'])
reg_trans=lm(ModernC~.,data=UN3_trans)
par(mfrow = c(2, 2))
avPlots(reg_trans)
```




```r
plot(reg_trans)
```



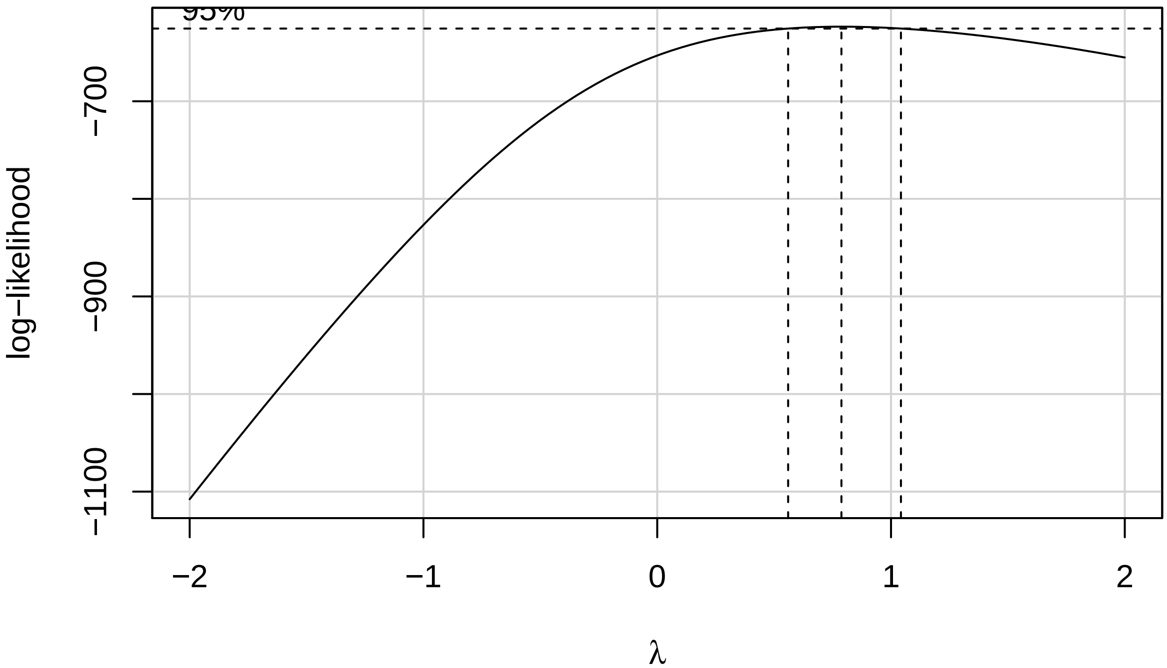

```r
termplot(reg , terms = "Pop",
         partial.resid = T, se=T, rug=T,
         smooth = panel.smooth)
termplot(reg_trans , terms = "Pop",
         partial.resid = T, se=T, rug=T,
         smooth = panel.smooth)
```



I initially looked at the added variable plots and saw PPgdp and Pop may be two variable needing transformation(since they are both clustered). I first tried the boxTidwell method and found that both variables give insigificance, Then i tried the powerTransform function, found that Change should be transformed to its 0.3 power, however, we would want to elimate the negative values in the variable Change. After bringing all values positive in Change, i apply powerTranform again to find that only variable needing transformation is Pop. Since 0.33 is relatively close to 0.5, a square root transformation will be approriate. Also see the disired transformation from the added variable graphs in question 5, since both graph for PPgdp and Pop seem clustered, we wanted a way to make them more spread. So i impose a log transformation on PPgdp, and from the termplot before&after transformation, my transformation did improve the graph.

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

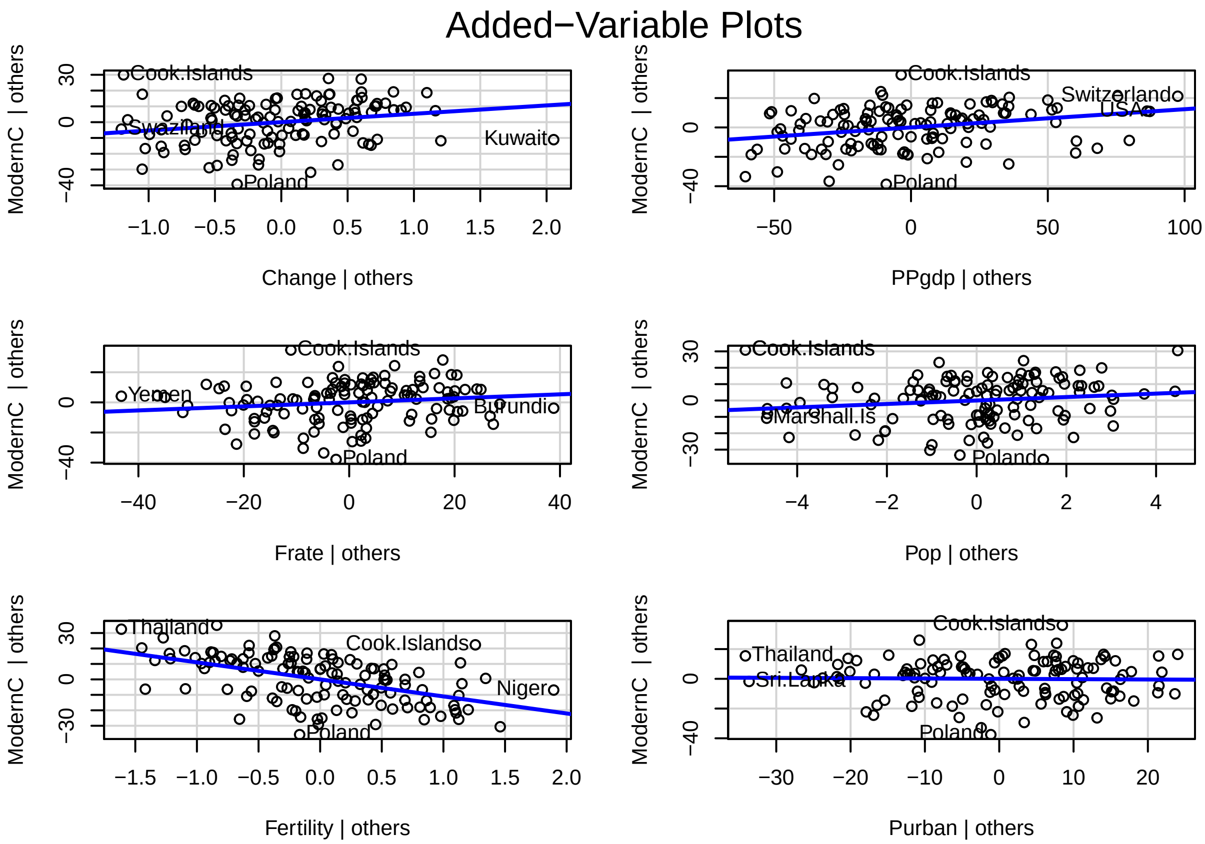
```r
car::boxCox(reg_trans,family="yjPower",plotit=TRUE)
```



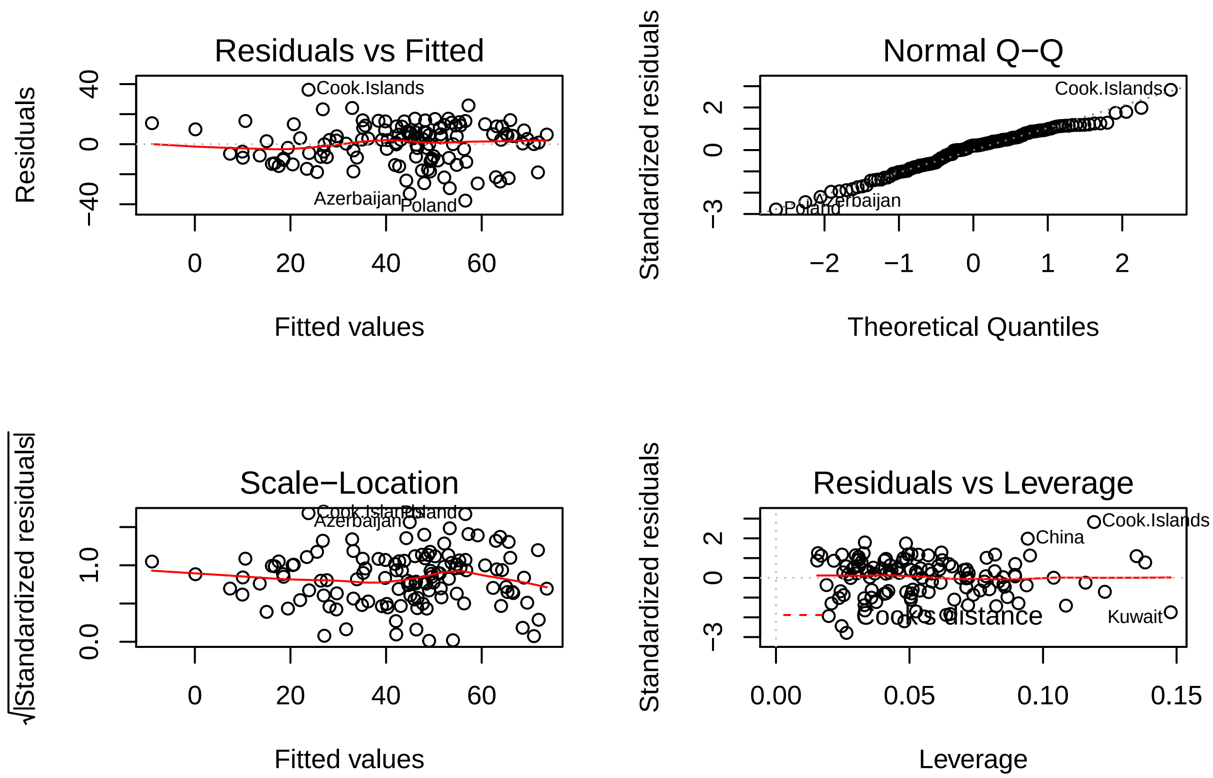
I decided not to impose any transformation on ModernC the response variable since lamda interval includes 1.

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.

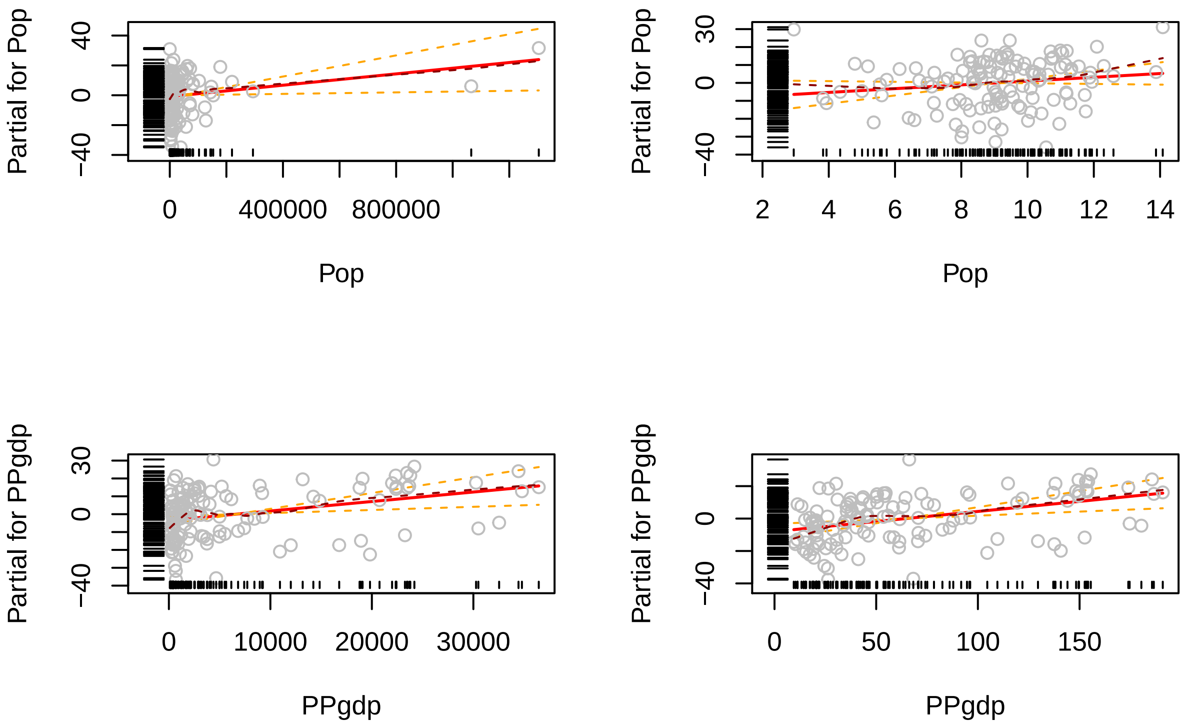
```r
reg_trans=lm(ModernC~.,data=UN3_trans)
par(mfrow = c(2, 2))
avPlots(reg_trans)
```




```r
plot(reg_trans)
```



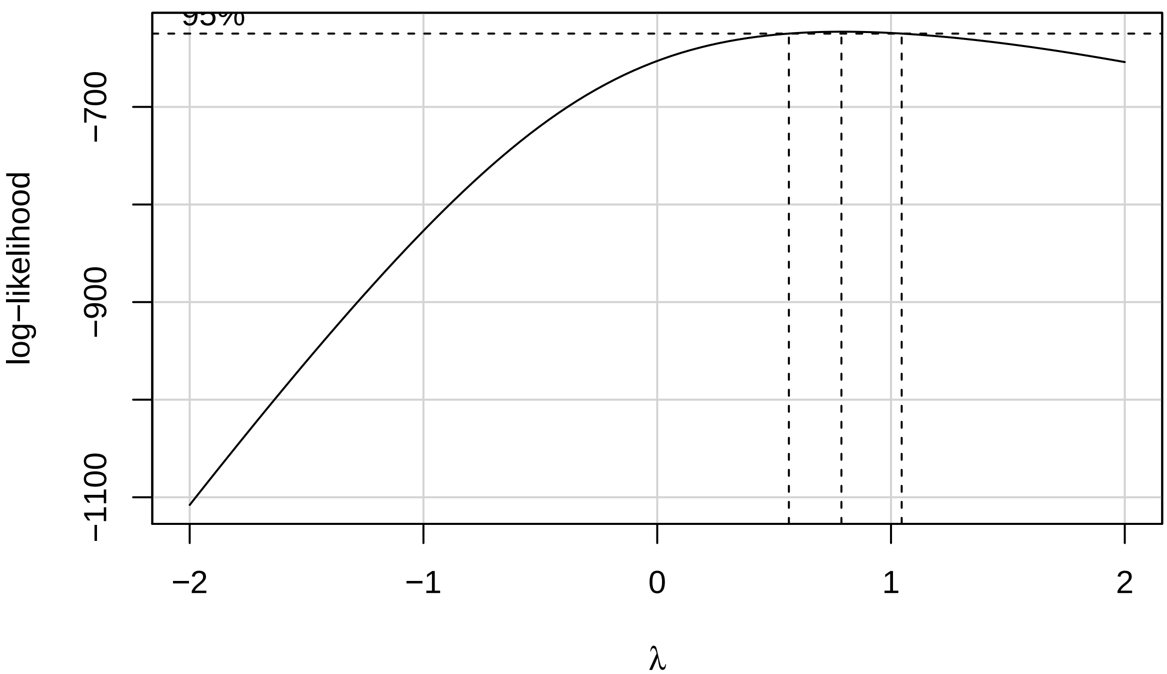
```r
termplot(reg , terms = "Pop",
         partial.resid = T, se=T, rug=T,
         smooth = panel.smooth)
termplot(reg_trans , terms = "Pop",
         partial.resid = T, se=T, rug=T,
         smooth = panel.smooth)
termplot(reg , terms = "PPgdp",
         partial.resid = T, se=T, rug=T,
         smooth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



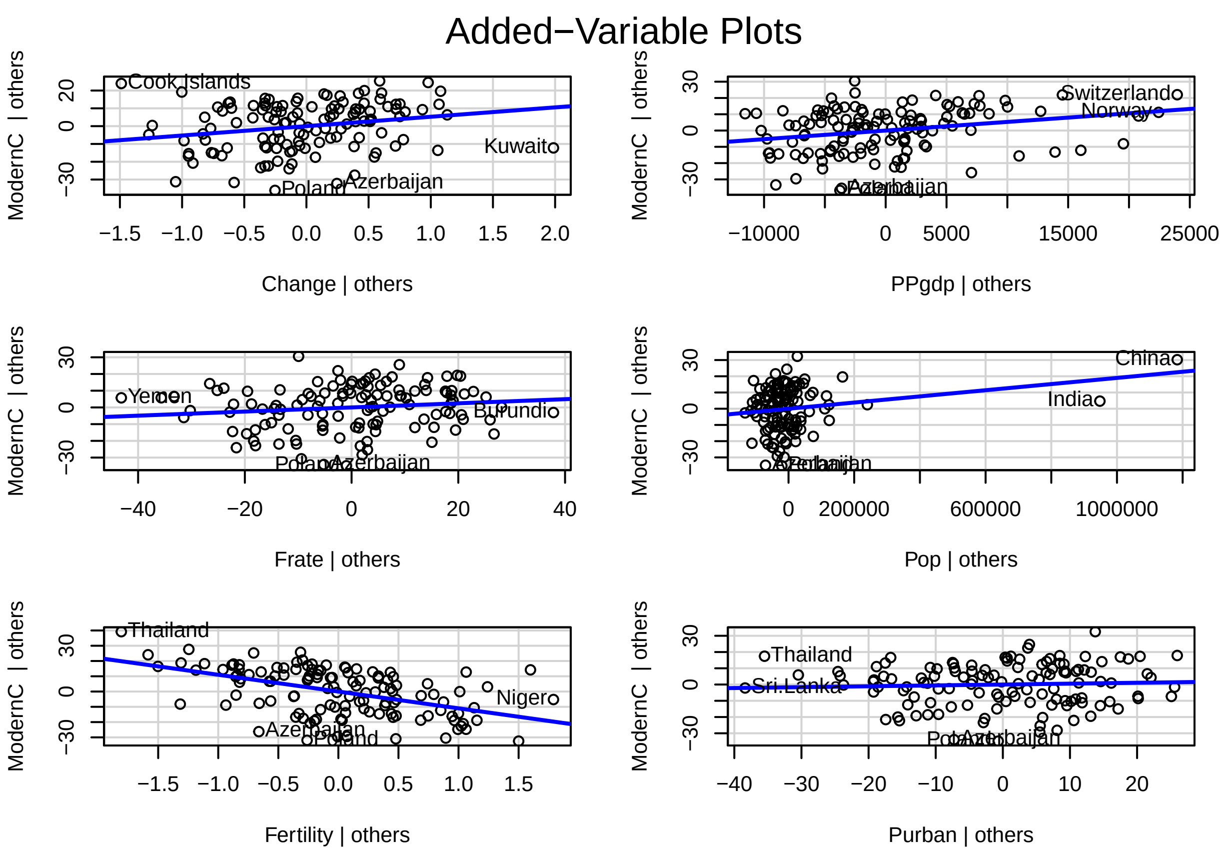
After the transformation, the termplot shows that both PPgdp and Pop are less clustered. We could also see this pattern from the added variable plot. Also, I observed improvements in residuals plots.Shape of the tail on normal QQ plot improved. The line on "Residuals vs. Leverage" became flatter.

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?

```r
boxCox(reg,family="yjPower",plotit=TRUE)
```




```
reg_test <- lm(ModernC~.,data=UN3_0)
car::avPlots(reg_test)
```




```
powerTransform(as.matrix(UN3_0)~.,family="bcnPower",data=UN3_0)
```

```
## Estimated transformation power, lambda
```

```
## [1] 0.9999782 0.2951891 0.9999984 0.9999849 0.3251064 0.9994071 0.9999831
##
## Estimated location, gamma
## [1] 1.000000e-01 4.873502e+00 2.450958e+00 1.000000e-01 1.304196e+06
## [6] 1.000000e-01 1.000000e-01
```

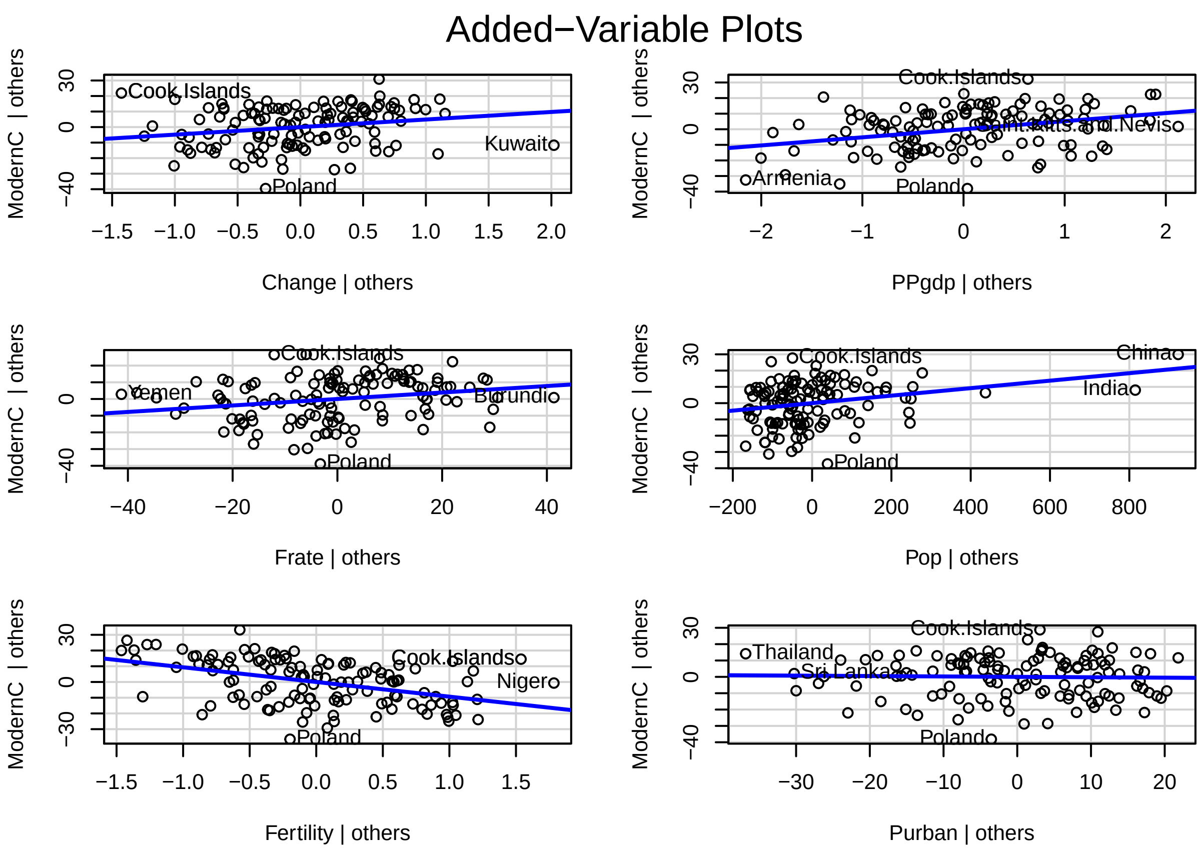
```r
range(UN3_0['Change'])
```

```
## [1] -1.10  3.62
```

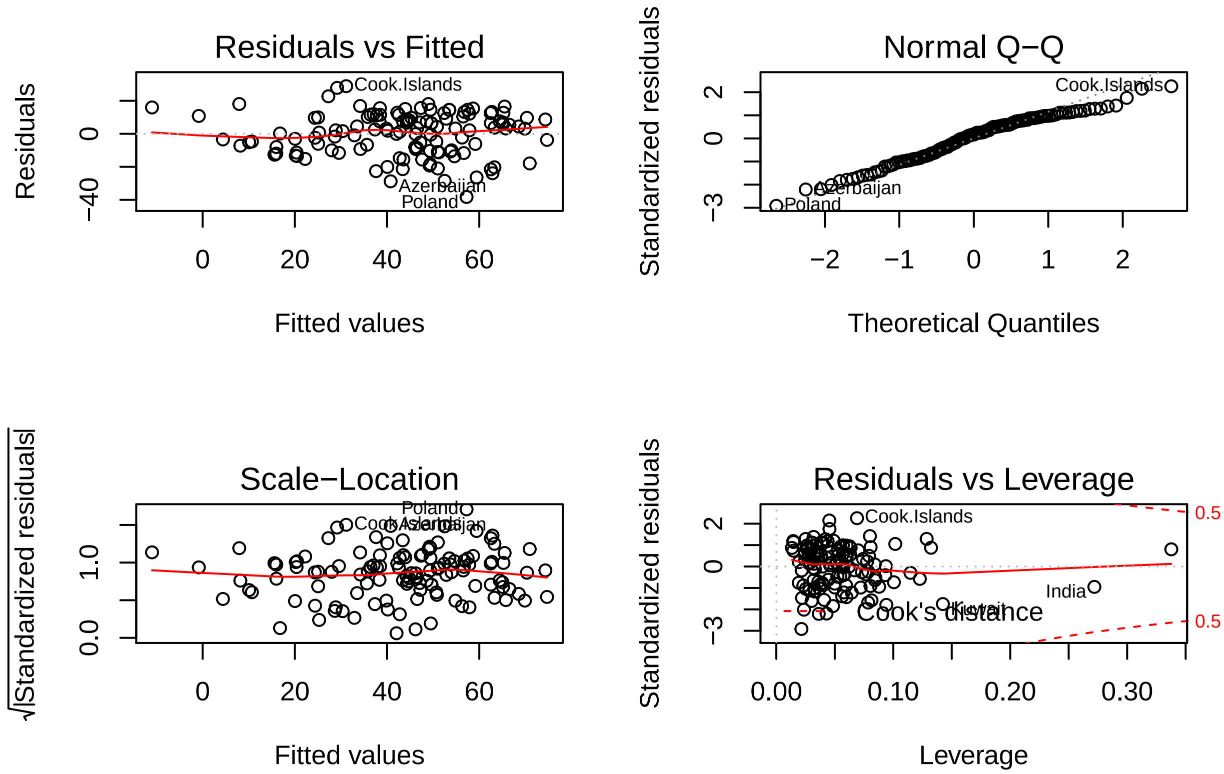
```r
UN3_1['Change']=UN3_0['Change']+2
powerTransform(as.matrix(UN3_1)~.,family="bcnPower",data=UN3_1)
```

```
## Estimated transformation power, lambda
## [1] 0.9999756 0.9991936 0.9999997 0.9999871 0.3250975 0.9993770 0.9999894
##
## Estimated location, gamma
## [1] 1.000000e-01 1.000000e-01 1.151552e+00 1.000000e-01 1.304196e+06
## [6] 1.000000e-01 1.000000e-01
```

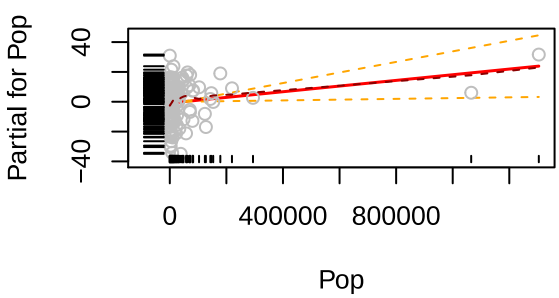
```r
UN3_trans1=UN3_1
UN3_trans1['Pop']=UN3_1['Pop']^0.5
UN3_trans1['PPgdp']=log(UN3_1['PPgdp'])
reg_trans1=lm(ModernC~.,data=UN3_trans1)
par(mfrow = c(2, 2))
avPlots(reg_trans1)
```

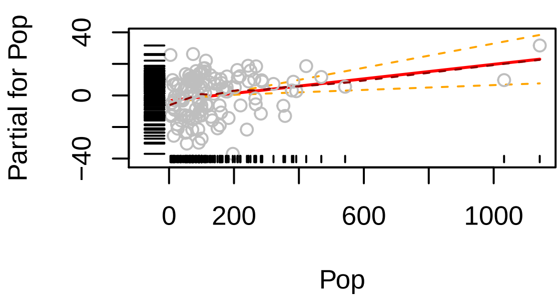



```r
plot(reg_trans1)
```




```r
termplot(reg , terms = "Pop",
         partial.resid = T, se=T, rug=T,
         smooth = panel.smooth)
termplot(reg_trans1 , terms = "Pop",
         partial.resid = T, se=T, rug=T,
         smooth = panel.smooth)
```





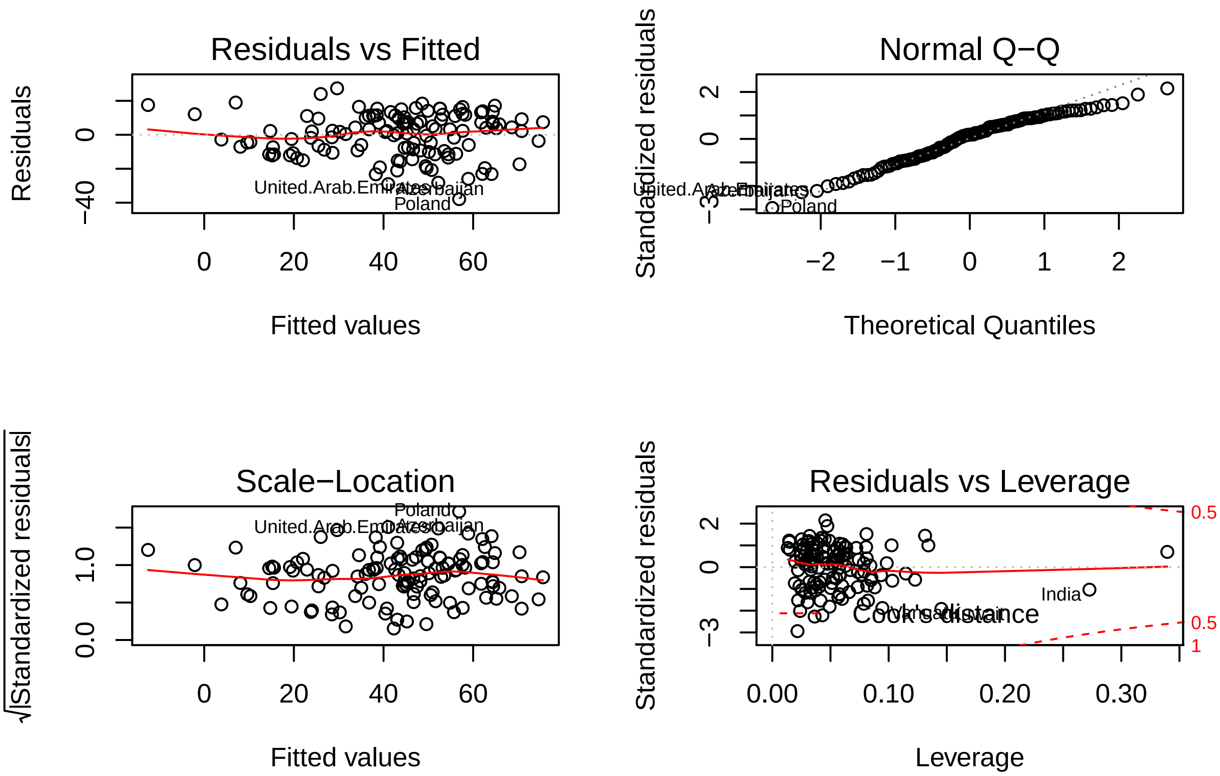
No, because no transformation has been imposed on the response variable, i end up with the same model as in question 8. For the predictors, i imposed same transformation after doing a powerTransform, changing the values in Change to be positive, and impose the powerTransform again to see a sqrt on PPgdp and see a log transform needed from the added variable plots.

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.

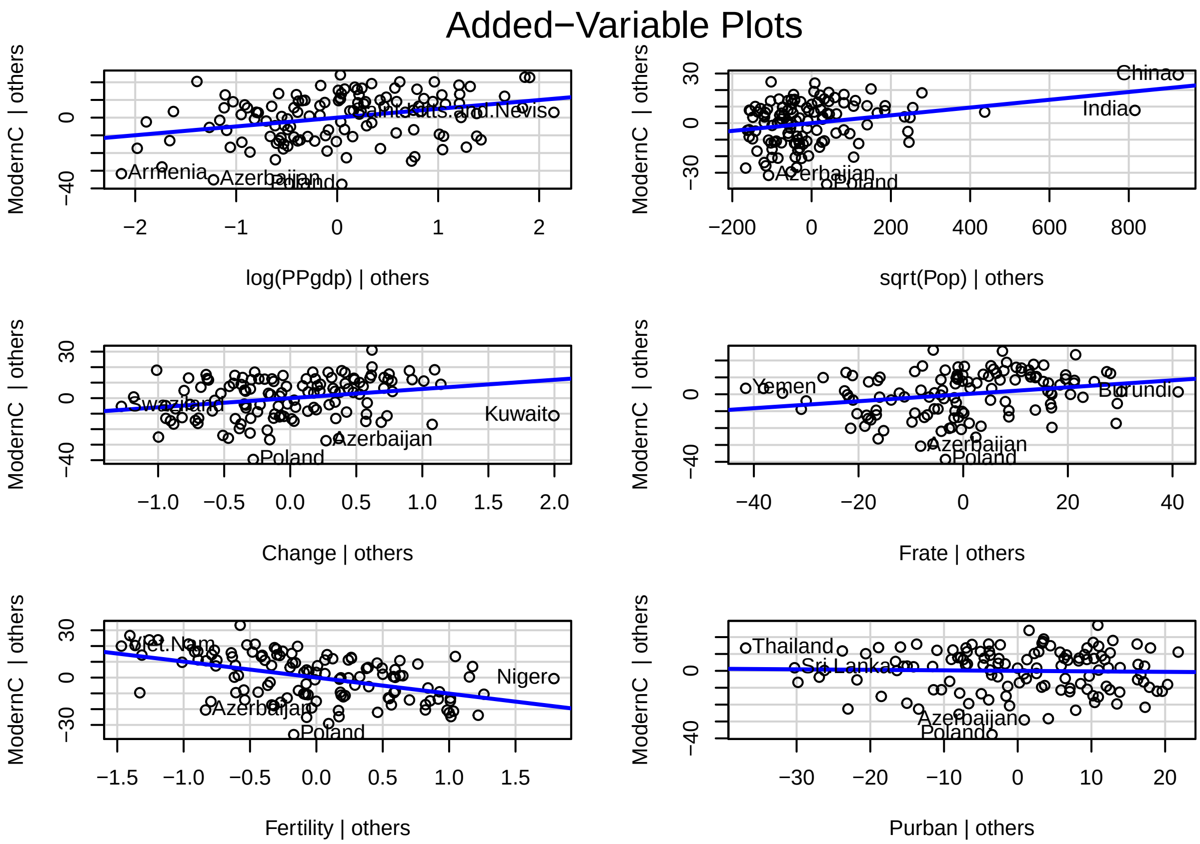
```r
outlierTest(reg_trans)
```

```
## No Studentized residuals with Bonferonni p < 0.05
## Largest |rstudent|:
##                rstudent unadjusted p-value Bonferonni p
## Cook.Islands 2.915207          0.0042608       0.5326
```

```r
UN3_out=UN3_1[-c(28),]
reg_rm<-lm(ModernC ~ log(PPgdp)+sqrt(Pop)+Change+Frate+Fertility+Purban,data=UN3_out)
par(mfrow = c(2, 2))
plot(reg_rm)
```




```r
car::avPlots(reg_rm)
```




```r
summary(reg_rm)
```

```
##
## Call:
## lm(formula = ModernC ~ log(PPgdp) + sqrt(Pop) + Change + Frate +
##     Fertility + Purban, data = UN3_out)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -37.886  -9.315   2.247  10.067  27.355
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.260328  11.979396   0.272  0.78598
## log(PPgdp)     4.975095   1.338155   3.718  0.00031 ***
## sqrt(Pop)      0.023532   0.007553   3.116  0.00231 **
## Change         5.918676   2.058110   2.876  0.00479 **
## Frate          0.206322   0.074903   2.755  0.00682 **
## Fertility    -10.158068   1.756536  -5.783 6.22e-08 ***
## Purban        -0.030130   0.094880  -0.318  0.75138
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.02 on 117 degrees of freedom
## Multiple R-squared:  0.6502, Adjusted R-squared:  0.6323
## F-statistic: 36.25 on 6 and 117 DF,  p-value: < 2.2e-16
```

China and India are points with high leverage and they are potential outliers, but not necessilary influential points. I tried to remove these two counrties. After removing these two points, another new point came to our eyes, Poland, marked by R, which is not that high leverage in comparison to China and India. However the residual plots did not change a lot,suggesting those two points may not be influential. After a outlierTest, cook's island seems to be one outlier, so we removed it. The normal QQ's tail seem to look better due to the removal.

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!

```r
summary(reg_rm)$coefficient
```

```
##                  Estimate    Std. Error    t value     Pr(>|t|)
## (Intercept)    3.26032826 11.979396234  0.2721613 7.859777e-01
## log(PPgdp)     4.97509455  1.338155045  3.7178760 3.098328e-04
## sqrt(Pop)      0.02353216  0.007553021  3.1155961 2.309446e-03
## Change         5.91867630  2.058110294  2.8757819 4.790079e-03
## Frate          0.20632203  0.074903428  2.7545072 6.818172e-03
## Fertility    -10.15806840  1.756535822 -5.7830124 6.223977e-08
## Purban        -0.03013016  0.094879567 -0.3175621 7.513832e-01
```

```r
a=as.matrix(summary(reg_rm)$coefficient)
b=data.frame("Estimate"=a[,1],"Lower Confidence Interval"=(a[,1]-a[,2]),"Upper Confidence Interval"=(a[
kable(b)
```

	Estimate	Lower.Confidence.Interval	
(Intercept)	3.2603283	-8.7190680	
log(PPgdp)	4.9750946	3.6369395	
sqrt(Pop)	0.0235322	0.0159791	
Change	5.9186763	3.8605660	
Frate	0.2063220	0.1314186	
Fertility	-10.1580684	-11.9146042	
Purban	-0.0301302	-0.1250097	
10% increse in	per capita 2	001 GDP will result in 14.93	% increase in percent of unmarried women using a modern m

10% increse in population(in thousands) will result in 0.024 unit increase in percent of unmarried women using a modern method of contraception. And the 95% confidence interval is [0.016,0.0031]

One unit increse in annual population growth rate percent will result in 5.91 unit increase in percent of unmarried women using a modern method of contraception.And the 95% confidence interval is [3.86,7.97]

One unit increse in percent of females over 15 economically active will result in 0.206 unit increase in percent of unmarried women using a modern method of contraception.And the 95% confidence interval is [0.131,0.281]

One unit increse in expected number of life births per female 2000 will result in -10.15 unit increase in percent of unmarried women using a modern method of contraception.And the 95% confidence interval is [-11.9,-8.4]

One unit increse in Percent of population that is urban, 2001 will result in -0.03 unit increase in percent of unmarried women using a modern method of contraception.And the 95% confidence interval is [-0.12,0.06]

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

I use na.omit to remove all rows containing NA's also, i decide not to removed India and China since they are not influential. After all these case deletions, i applied log transformation to Per capital GDP and square root transformation to Population. And the final model ModernC~Change+log(PPgdp)+Frate+sqrt(Pop)+Fertility+Purban.

$$Modern = 3.26 + 4.98log(PPgdp) + 0.02sqrt(Pop) - 10.15Fertility + 5.91Change - 0.03Purban + 0.21Frate$$

And my finding is after applying these transformations, the added-variable plots shows that the Population is not so clustered.

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.

$$e_{(y)} = \hat{\beta}_0 + \hat{\beta}_1 e_{(x)}$$
$$(I-H)y = \hat{\beta}_0 + \hat{\beta}_1 (I-H)X_i$$

We know that

$$be\hat{t}a_1 = (X^TX)^{-1}X^Ty, \text{and} X = (I-H)x_i, y = (I-H)y$$

Thus,

$$(I-H)y = \hat{\beta}_0 * \nVdash (I-H)x_i$$

and

$$\hat{\beta}_1 = (x^Tx)^{-1}x^Ty$$

, where

$$x = (I-H)x_i, y = (I-H)y$$

so, we have

$$(I-H)y = \hat{\beta}_0 \nVdash + [x_i^T(I-H)(I-H)X_i]^{-1}((I-H)X_i)^T(I-H)y(I-H)X_i$$
$$(I-H)y = \hat{\beta}_0 \nVdash + [x_i^T(I-H)X_i]^{-1}x_i^T(I-H)y(I-H)x_i$$
$$x_i^T(I-H)y = x_i^T\hat{\beta}_0 \nVdash + x_i^T[x_i^T(I-H)X_i]^{-1}x_i^T(I-H)y(I-H)x_i$$

$$x_i^T(I-H)y = x_i^T \nVdash \hat{\beta}_0 + x_i^T(I-H)x_i[x_i^T(I-H)X_i]^{-1}x_i^T(I-H)y$$
$$x_i^T(I-H)y = \sum_{j=1}^{n} x_{ij}\hat{\beta}_0 + x_i^T(I-H)y$$

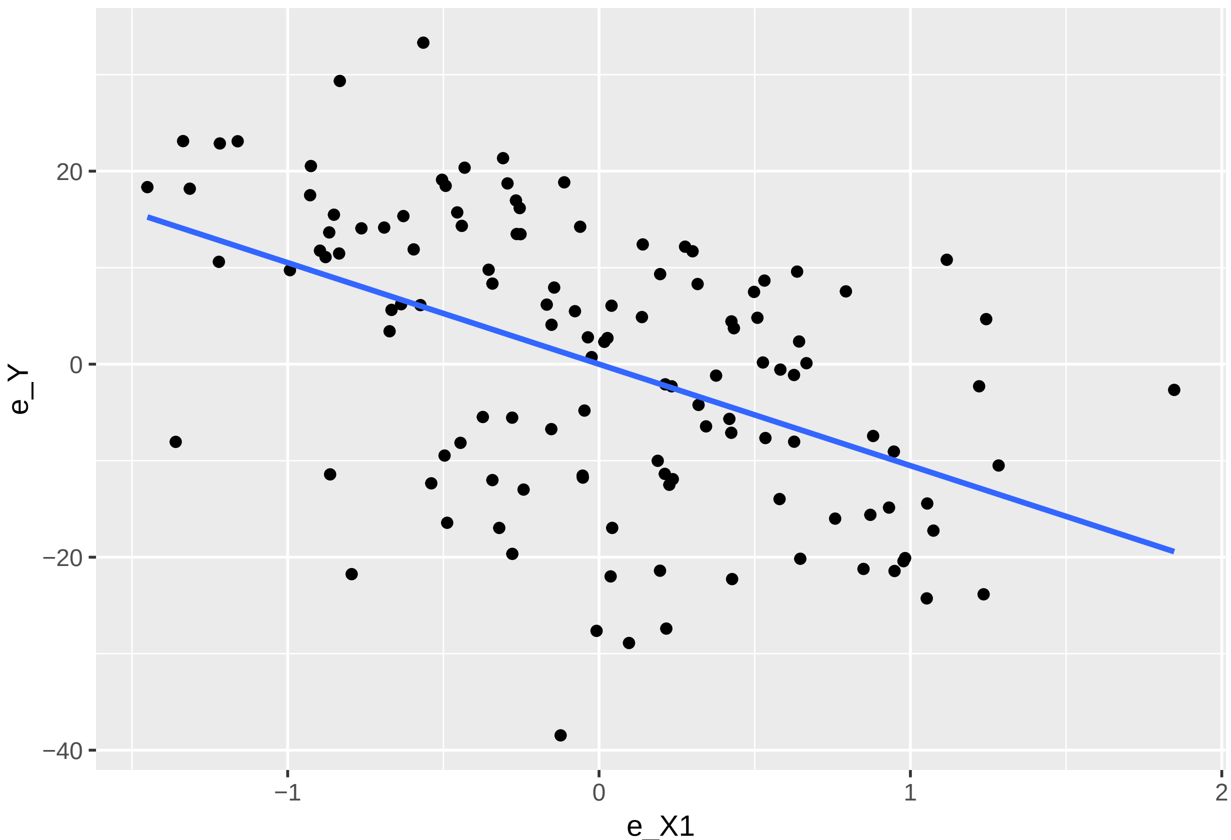
Thus, we have

$$\sum_{j=1}^{n} x_{ij}\hat{\beta}_0 = 0$$

And since$\sum_{j=1}^{n} x_{ij}$ 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~log(PPgdp)+Frate+log(Pop)+Change+Purban, data=UN3_out))
e_X1 = residuals(lm(Fertility ~ log(PPgdp)+Frate+log(Pop)+Change+Purban, data=UN3_out))
df = data.frame(e_Y=e_Y, e_X1=e_X1)
ggplot(data=df, aes(x = e_X1, y = e_Y)) +
geom_point() +
geom_smooth(method = "lm", se = FALSE)
```




```r
summary(reg_rm)$coef
```

```
##                   Estimate   Std. Error    t value     Pr(>|t|)
## (Intercept)   3.26032826 11.979396234  0.2721613 7.859777e-01
## log(PPgdp)    4.97509455  1.338155045  3.7178760 3.098328e-04
## sqrt(Pop)     0.02353216  0.007553021  3.1155961 2.309446e-03
## Change        5.91867630  2.058110294  2.8757819 4.790079e-03
## Frate         0.20632203  0.074903428  2.7545072 6.818172e-03
## Fertility   -10.15806840  1.756535822 -5.7830124 6.223977e-08
## Purban       -0.03013016  0.094879567 -0.3175621 7.513832e-01
```

```r
summary(lm(e_Y ~ e_X1, data=df))$coef
```

```
##                  Estimate Std. Error       t value     Pr(>|t|)
## (Intercept)  4.193481e-16   1.148273  3.651989e-16 1.000000e+00
## e_X1        -1.051515e+01   1.704067 -6.170618e+00 9.185067e-09
```

According to the result, the slope of our manually constructed added variable plot for predictor Fertility is -9.3, which is the same as the estimate from our model.