HW2 STA521 Fall 17

[Yunxuan Li]

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

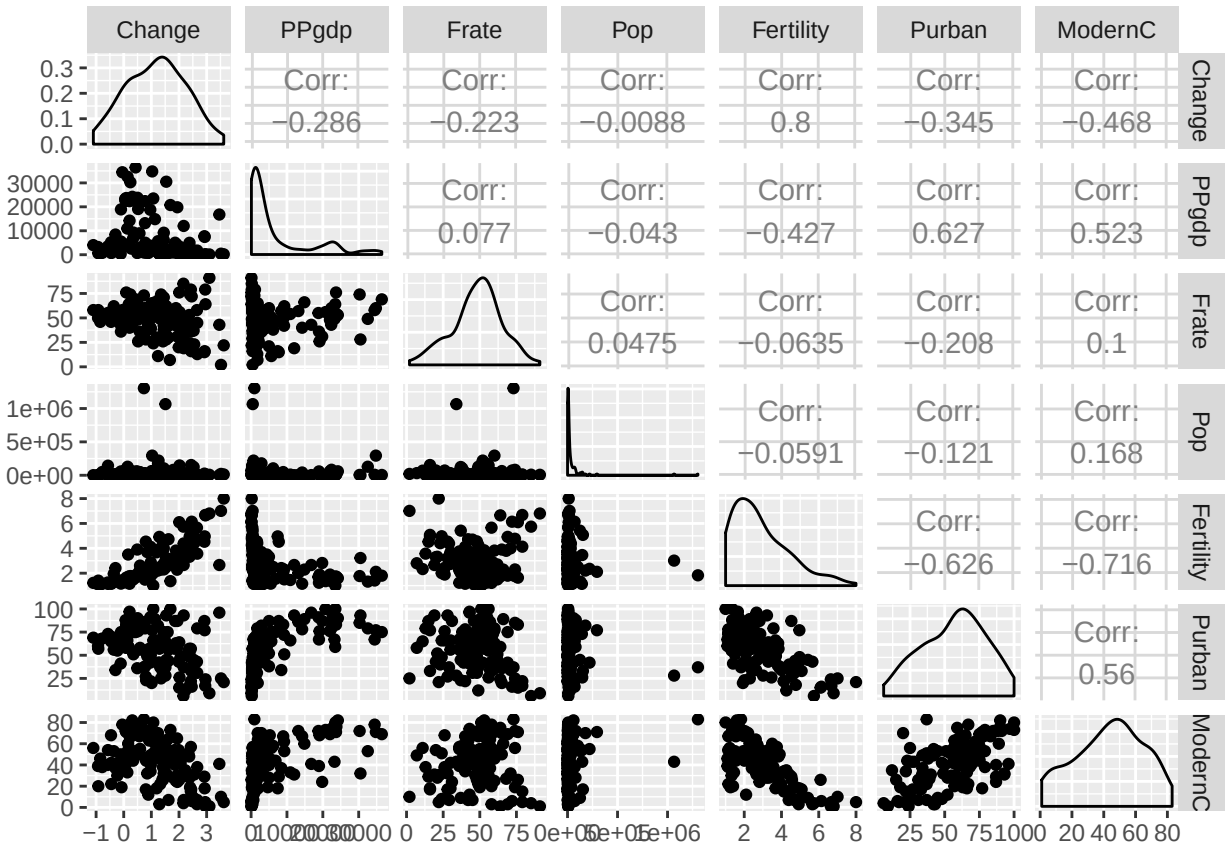
2. What is the mean and standard deviation of each quantitative predictor? Provide in a nicely formatted table.

```
mean_stan<-matrix(data=NA, nrow = 7, ncol = 3)
m<-1
for (i in 1:(length(UN3))){
  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
Purban	56.2	24.1097570036514

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
```

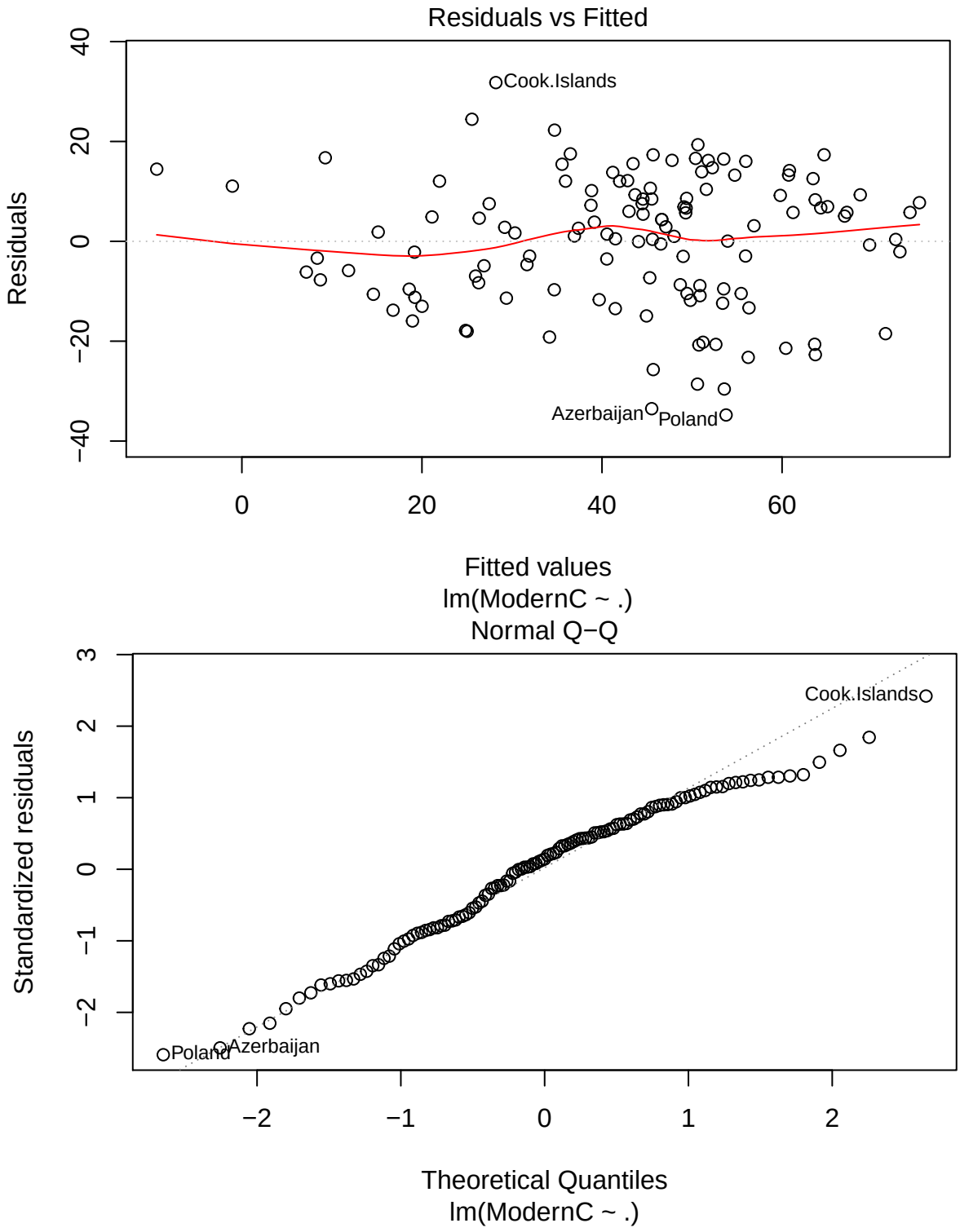


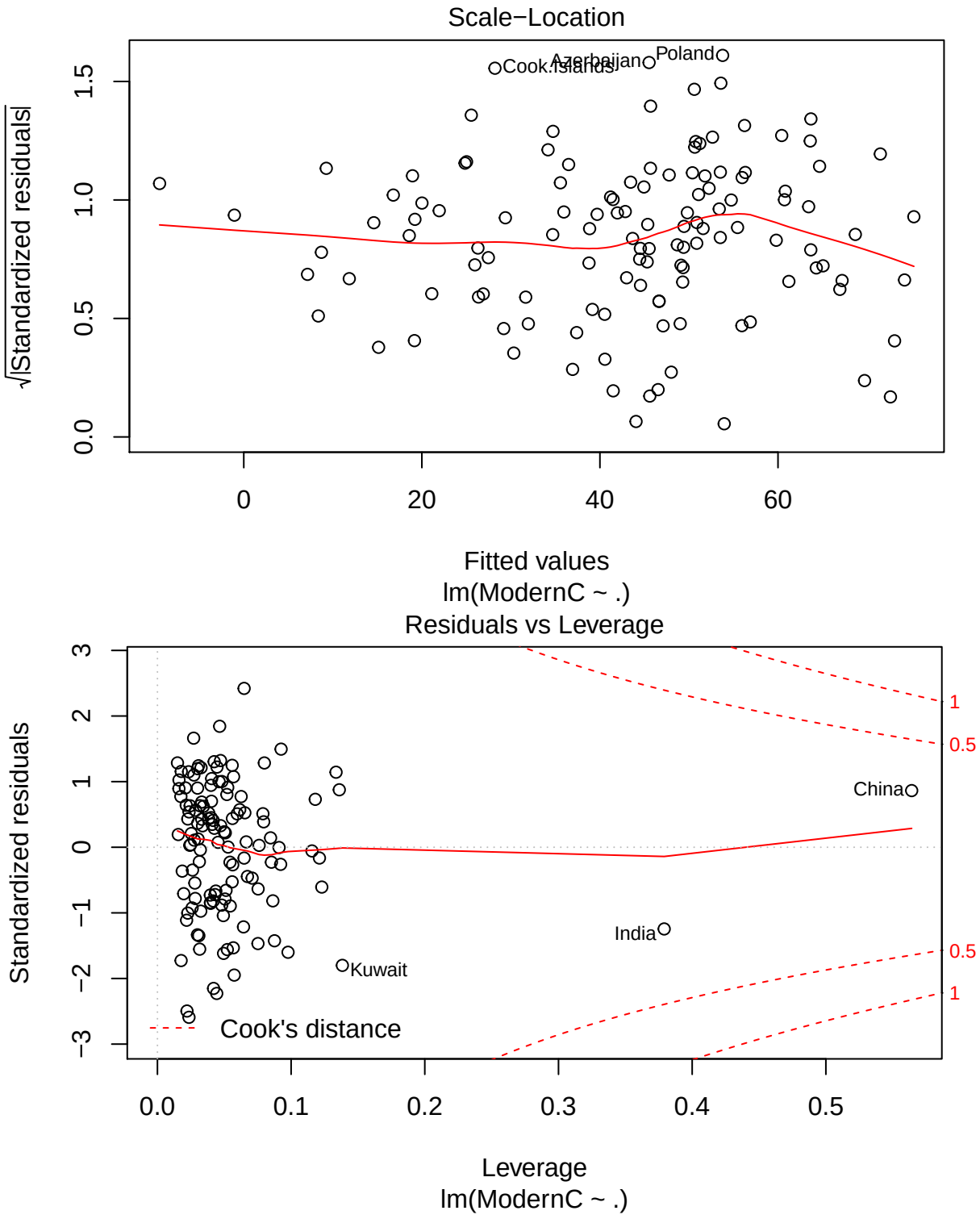
Answers: it seems that fertility, purban, ppdgp, and change are useful in predicting modernC. (corr coeff > 0.5)

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.

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


Residuals vs Fitted
Residuals
Fitted values
lm(ModernC ~ .)
Cook.Islands
Azerbaijan
Poland
Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(ModernC ~ .)



Answer:
It looks like the residual is not random.
Also, the Q-Q plot is not a straight 45-degree line: there is a lighter tail.
We need to do some transformations.

5. Using the Box-Tidwell `boxTidwell` from library `car` or graphical methods find appropriate transfor-

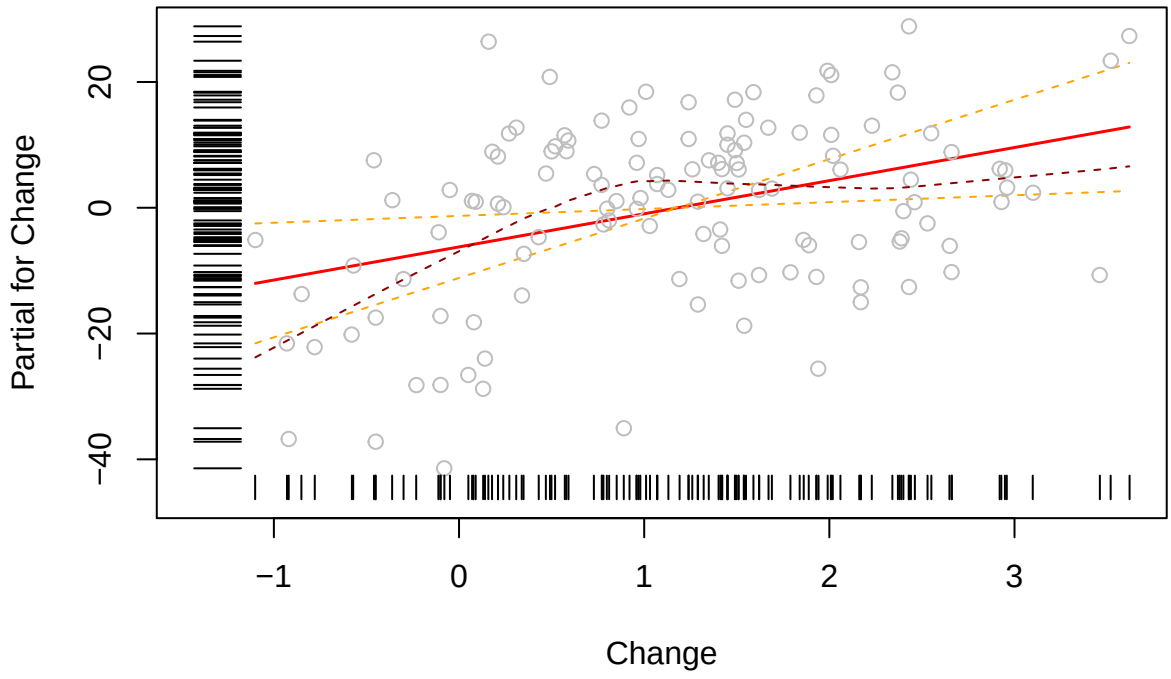
mations of the predictor variables to be used as predictors in the linear model. If any predictors are negative, you may need to transform so that they are non-negative. Describe your method and the resulting transformations.

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

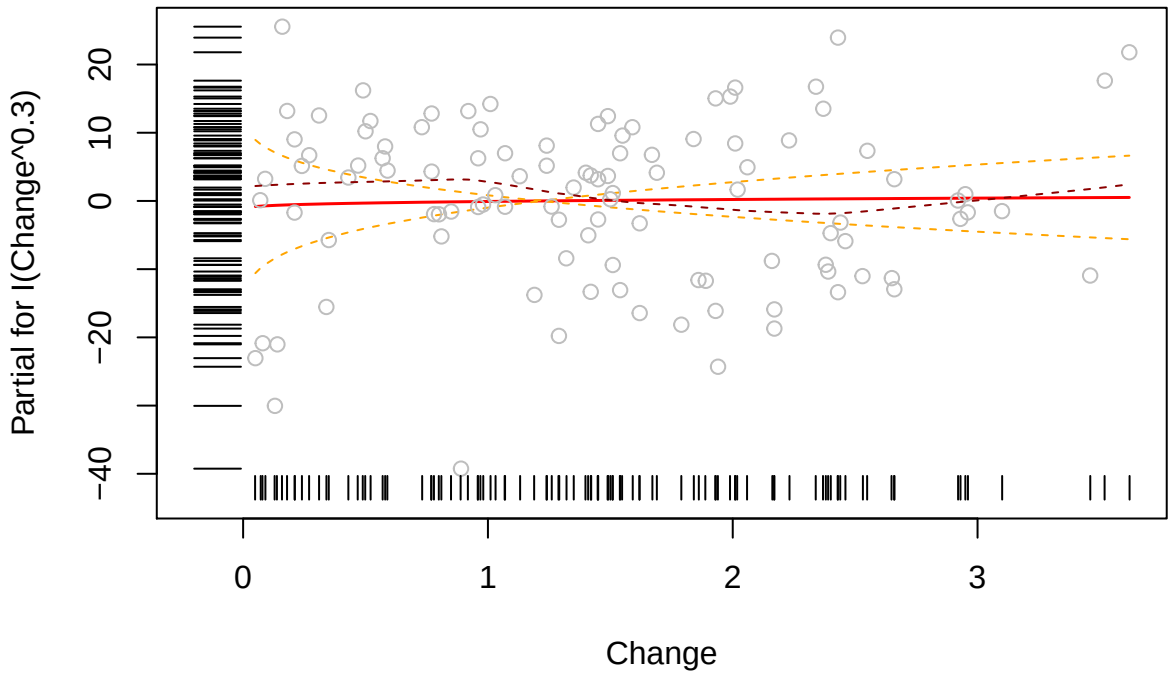
```
## Estimated transformation power, lambda
## [1] 0.2951946 0.9999996 0.9999975 0.3251072 0.9999648 0.9999841
## Estimated transformation location, gamma
## [1] 4.658143e+00 1.212076e+00 2.844007e-02 1.304196e+06 2.294808e-02
## [6] 1.428396e-01
```

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. We compare the termplots before transformation and afterwards to see if this really works.

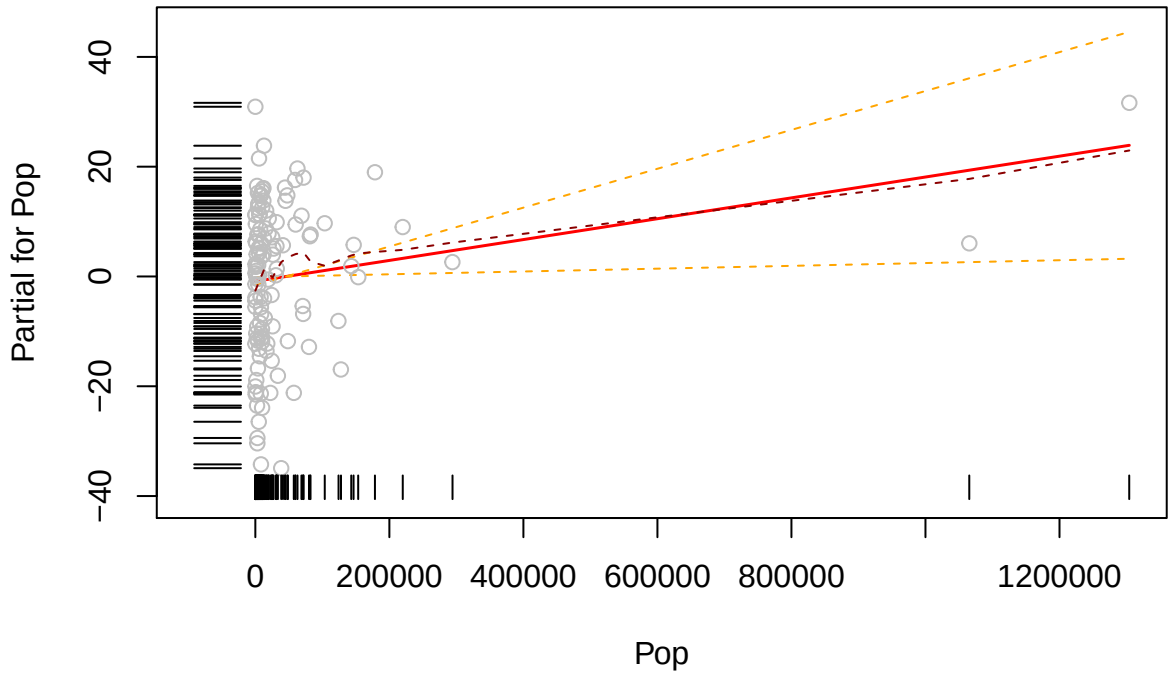
```
new_trans<-lm(ModernC~Purban+Frate+I(Change^0.3)+I(Pop^0.33)+Fertility+PPgdp,data=UN3_NAA)

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



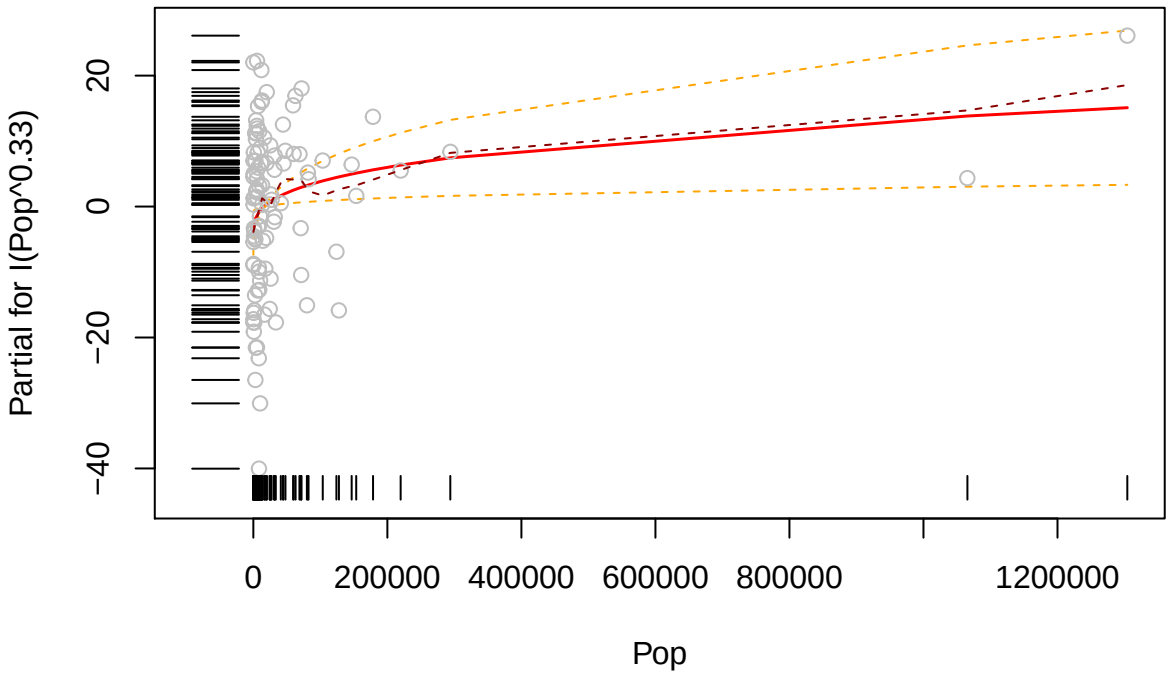

```
#g is the original linear regression formula, in problem 3
termplot(new_trans,terms="I(Change^0.3)",partial.resid = T, se=T, rug=T,smooth = panel.smooth)
```




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

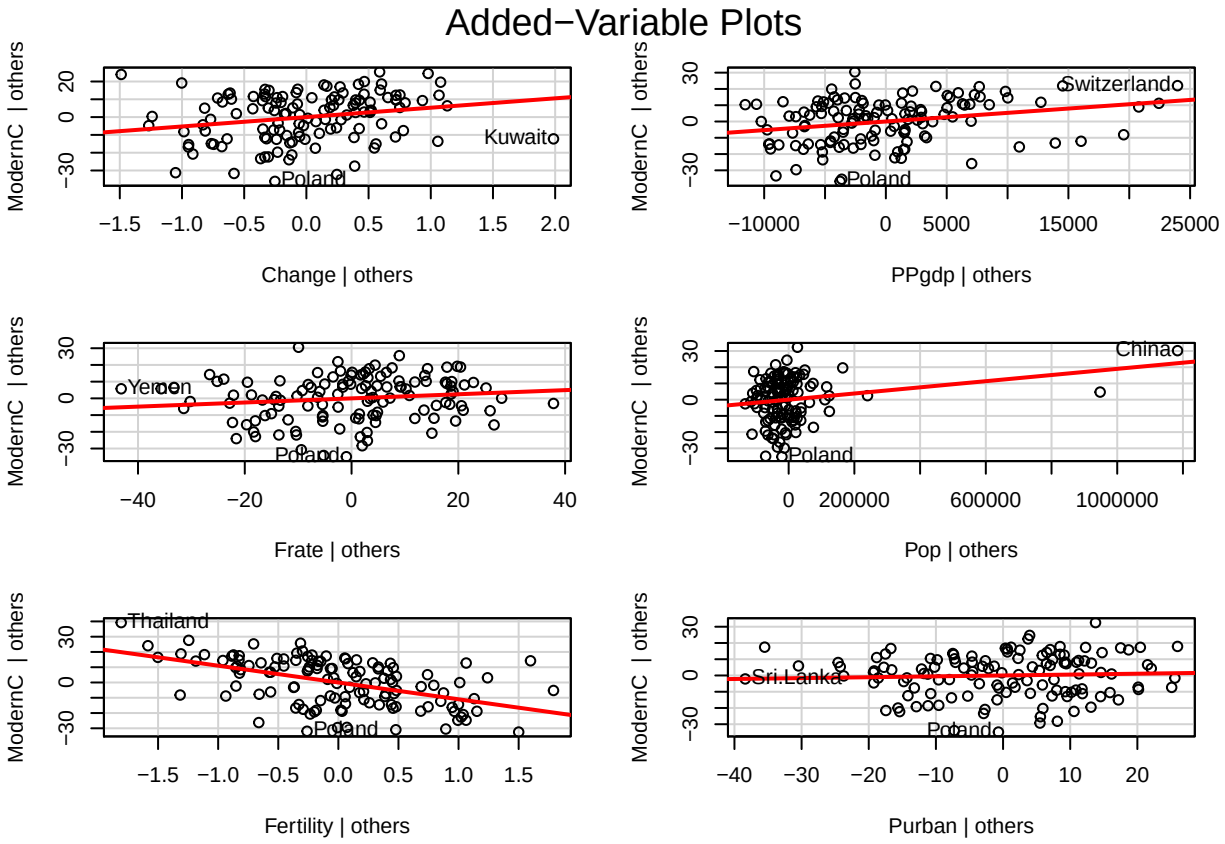



```
termplot(new_trans,terms="I(Pop^0.33)",partial.resid = T, se=T, rug=T,smooth = panel.smooth)
```

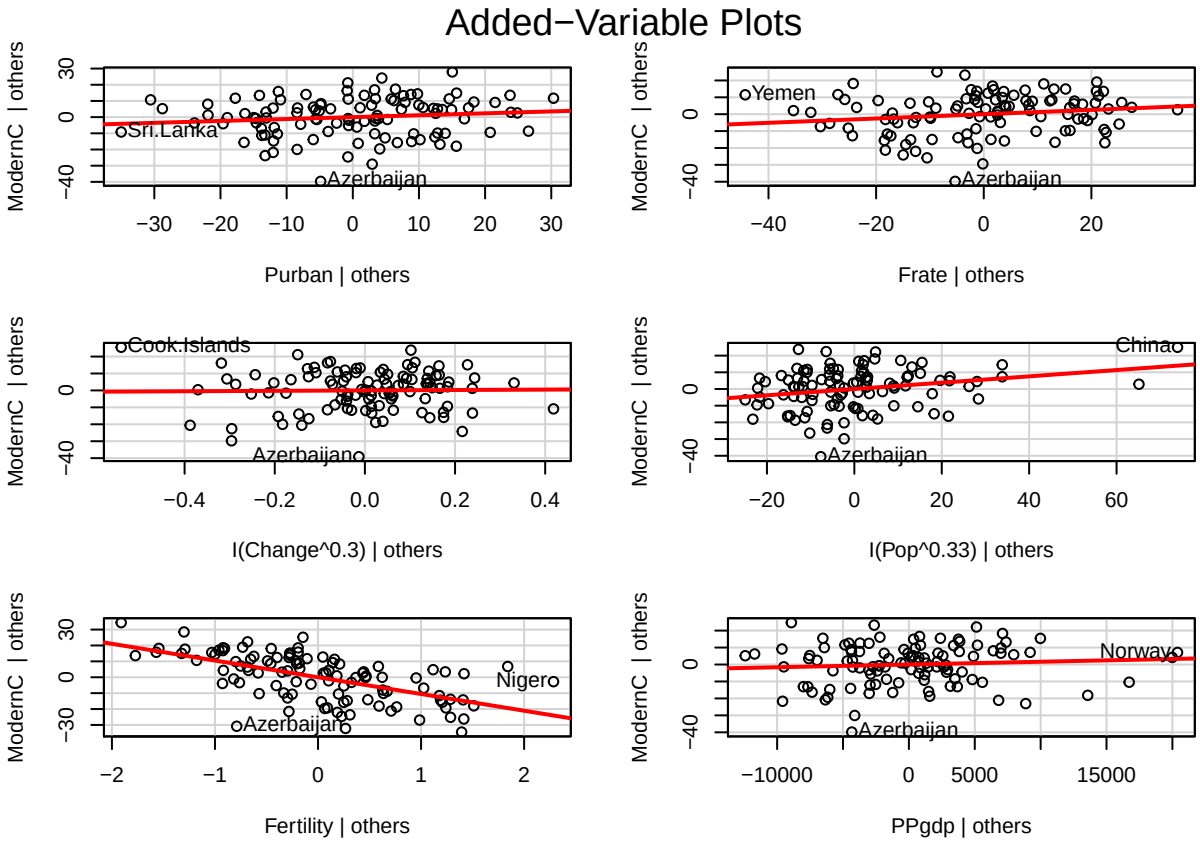


Comparing each pair of termplots, we can see that:
1. it seems that Changeˆ0.3 is a little bit better than Change.
2. it seems that Popˆ0.33 fits better.
We then look at addv plots to determine whether tranformations should take place.

```
mod1 = lm(ModernC ~ ., data=UN3_NAA)
avPlots(mod1,id.n=1)
```



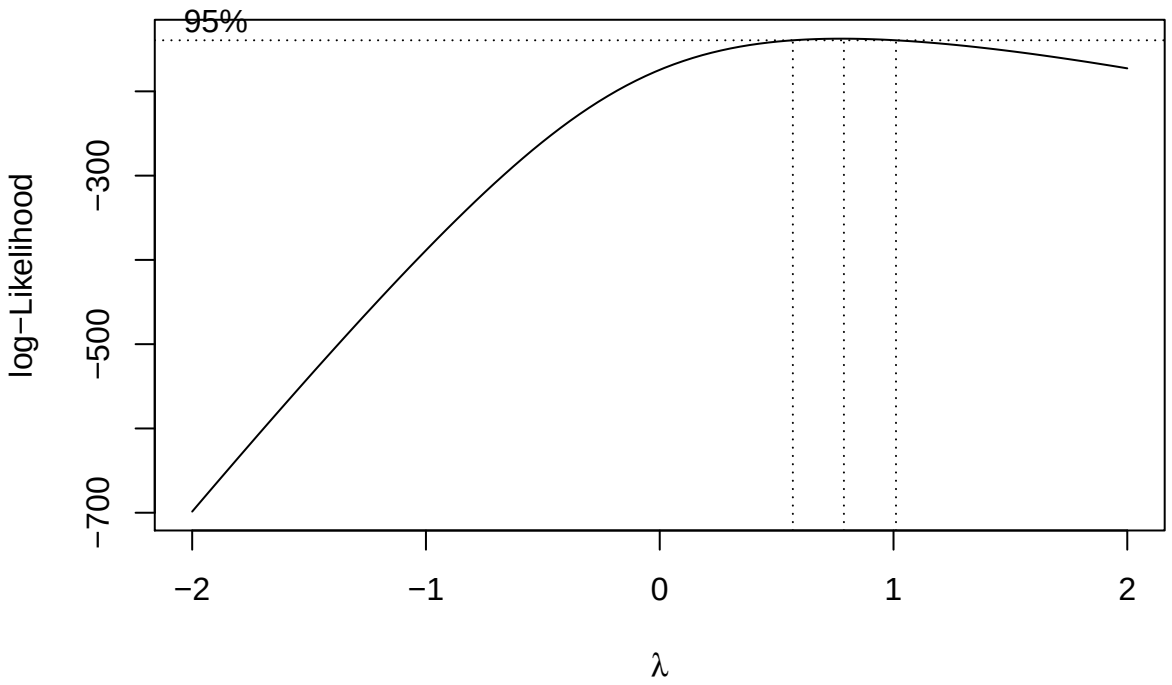

```
avPlots(new_trans,id.n=1)
```



The added variable plots show that Popˆ0.33 is much better then Pop.
After checking these plots, i finally decided to tranform "Change" and "Pop", with power 0.3 and 0.33 respectively. (These numbers are from the powerTransform results.

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)
boxx=boxcox(new_trans)
```

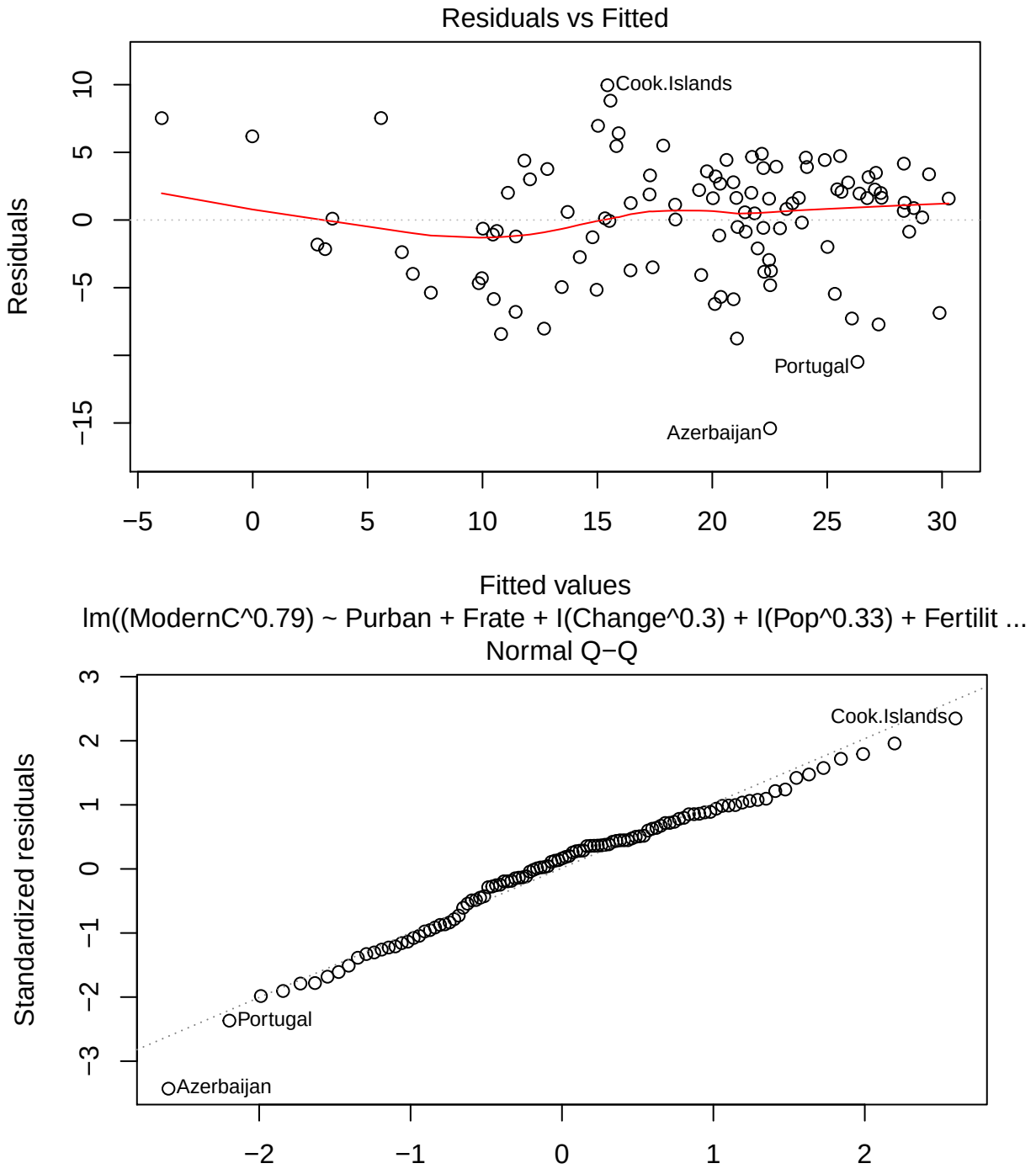


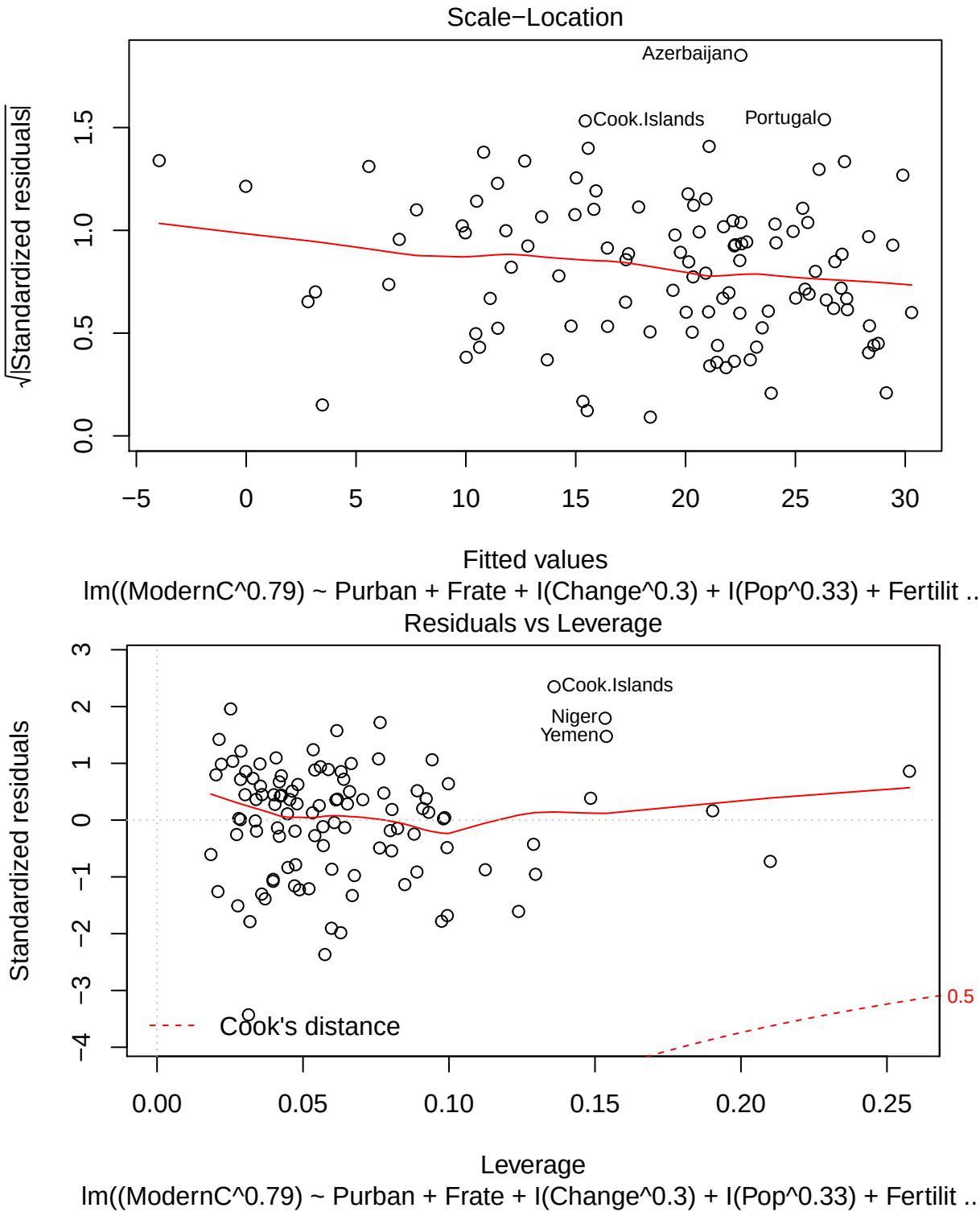

```
lambda = boxx$x
likeli = boxx$y
maxlambda<-cbind(lambda,likeli)[order(-likeli),][1]
maxlambda
```

```
## [1] 0.7878788
```

Answer: the graph indicates that lambda of ModernC is 0.7878788, so we transform it with power 0.79.
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.

```
newnew_trans<-lm((ModernC^0.79)~Purban+Frate+I(Change^0.3)+I(Pop^0.33)+Fertility+PPgdp,data=UN3_NAA)
plot(newnew_trans)
```


Residuals vs Fitted
Residuals
Fitted values
lm((ModernC^0.79) ~ Purban + Frate + I(Change^0.3) + I(Pop^0.33) + Fertilit ...
Cook.Islands
Portugal
Azerbaijan
Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm((ModernC^0.79) ~ Purban + Frate + I(Change^0.3) + I(Pop^0.33) + Fertilit ...
Cook.Islands
Portugal
Azerbaijan



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.

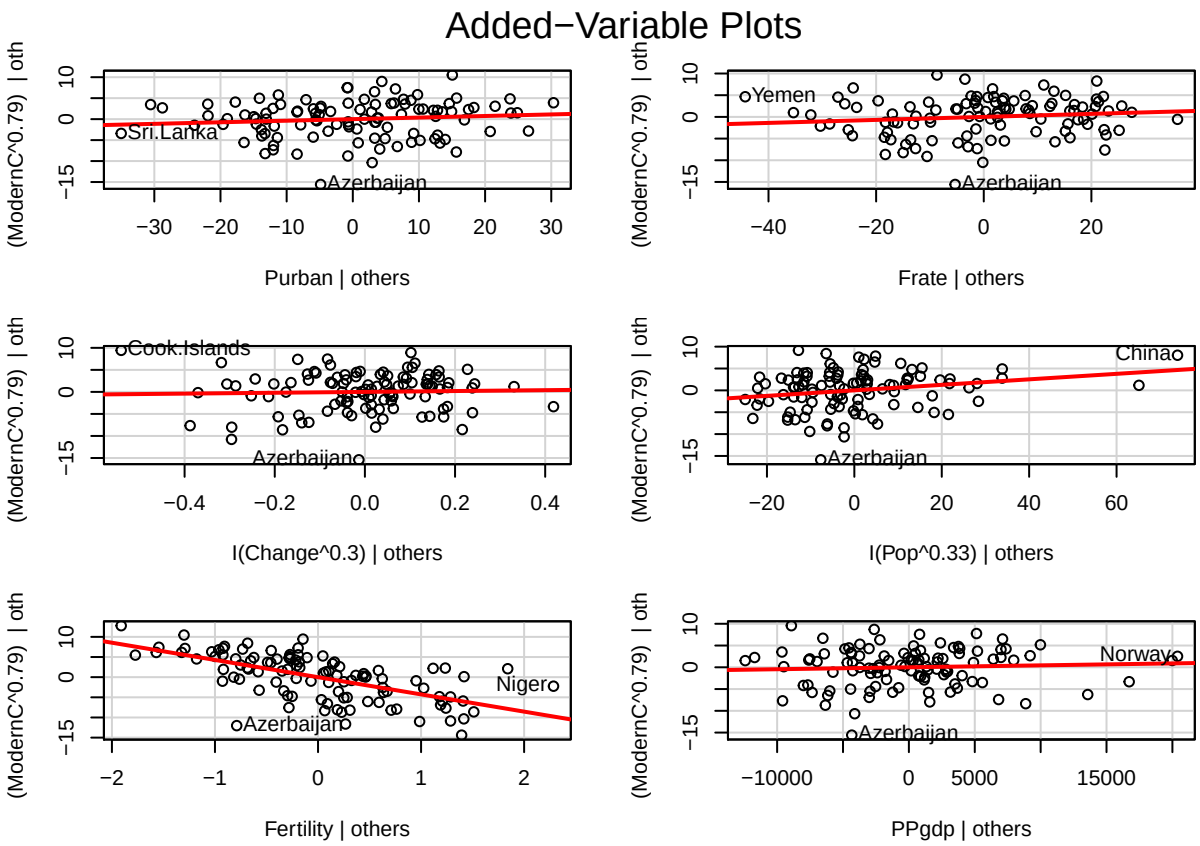
```
test<-matrix(data=NA, nrow = 6, ncol = 3)
cc<-summary(newnew_trans)
for (i in 2:length(coefficients(newnew_trans)))
```

```
{

  test[i-1,]<-c(rownames(cc$coefficients)[i],confint(newnew_trans, rownames(cc$coefficients)[i], level=0
}
ci_data<-data.frame(test)
colnames(ci_data)<-c("Var Name","2.5%","97.5%")
knitr::kable(ci_data)
```

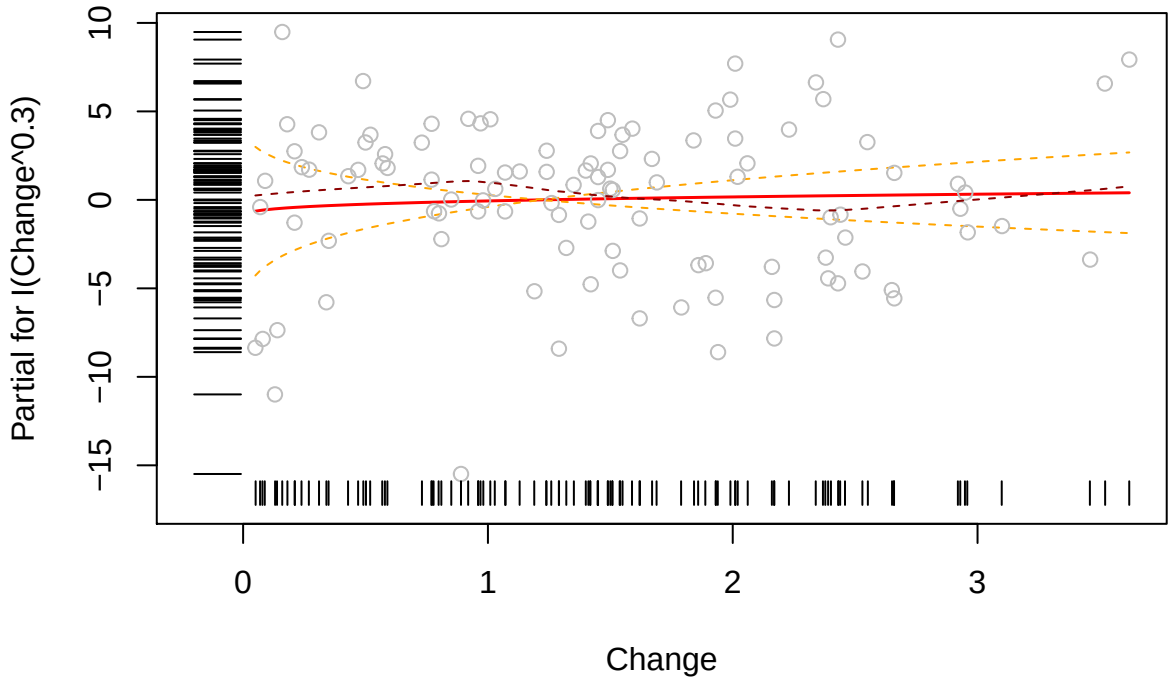
Var Name	2.5%	97.5%
Purban	-0.029170227293176	0.103939307729938
Frate	-0.0198301735677872	0.0894117564768152
I(Changeˆ0.3)	-4.53799423618096	6.4919025260229
I(Popˆ0.33)	0.00846017352609325	0.11749323892137
Fertility	-5.33282447055709	-3.21427123262118
PPgdp	-0.000102985800039471	0.000192000445019949

Answer: The confidence intervals mean that we are 95% confident that the coefficient of a specific predictor will fall in the corresponding 2.5%~97.5% range, which is listed above.
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?

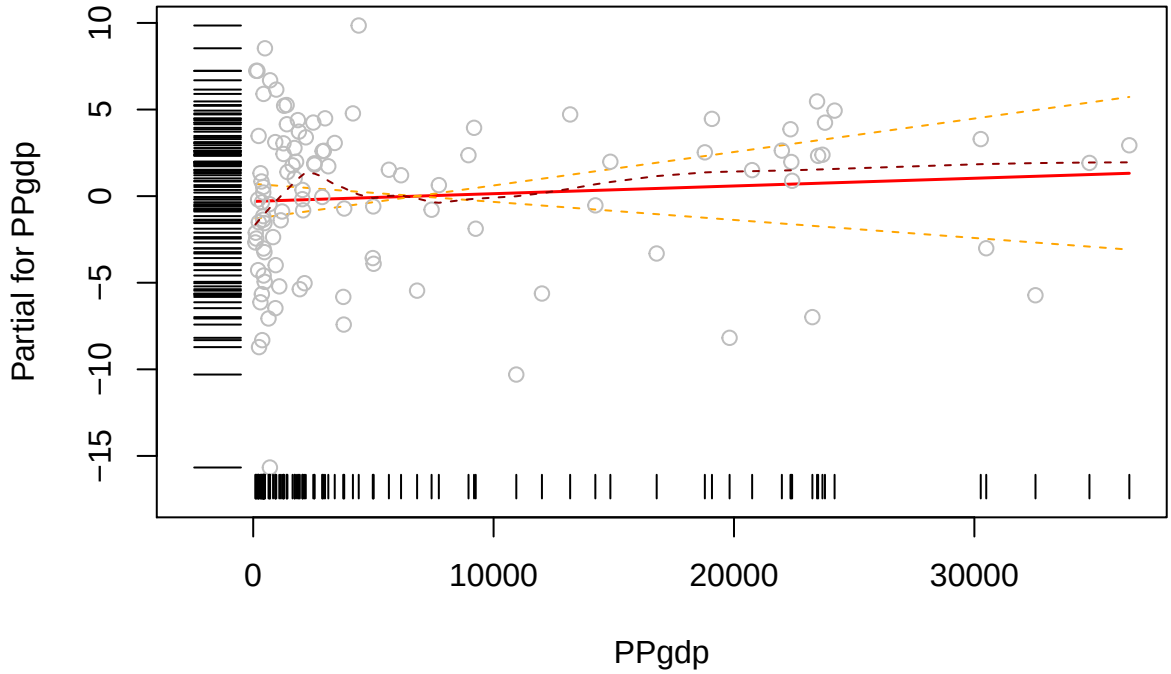
```
avPlots(newnew_trans,id.n=1)
```



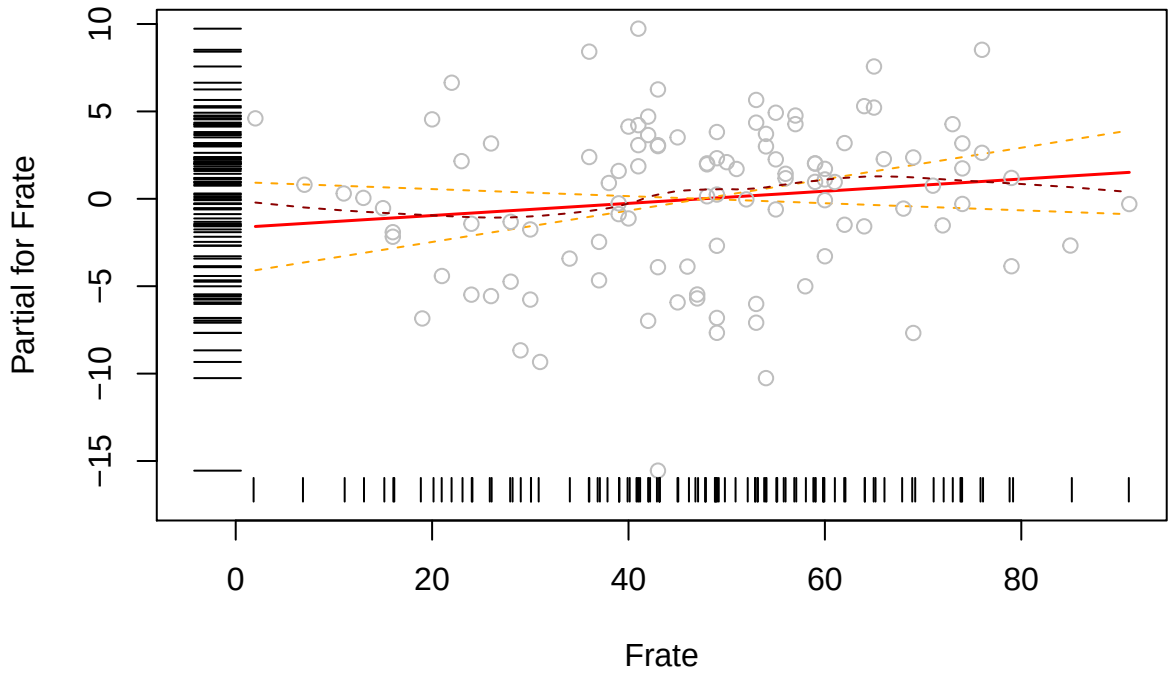

```
termplot(newnew_trans,terms="I(Change^0.3)",partial.resid = T, se=T, rug=T,smooth = panel.smooth)
```



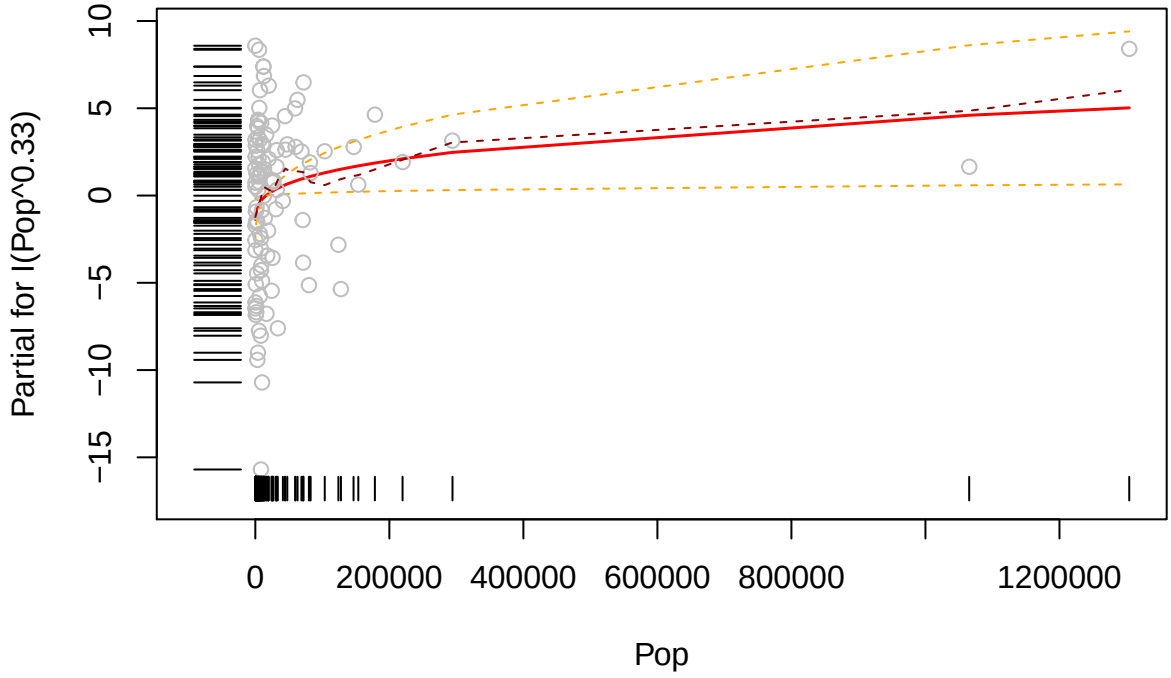

```
termplot(newnew_trans,terms="PPgdp",partial.resid = T, se=T, rug=T,smooth = panel.smooth)
```



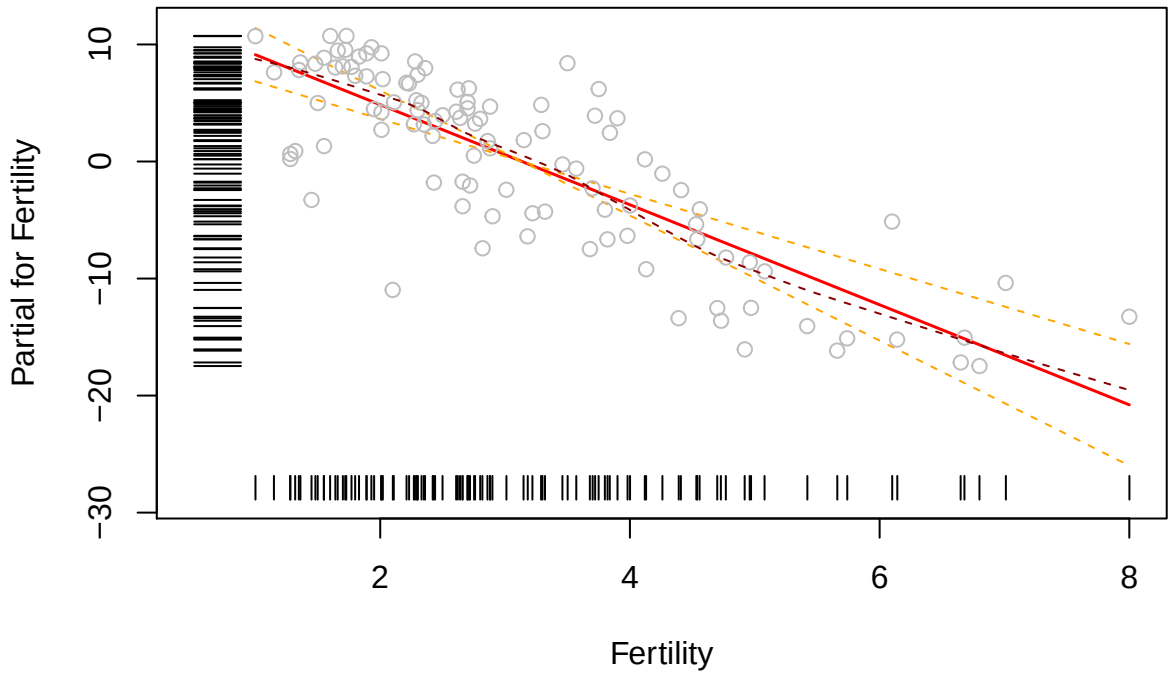

```
termplot(newnew_trans,terms="Frate",partial.resid = T, se=T, rug=T,smooth = panel.smooth)
```



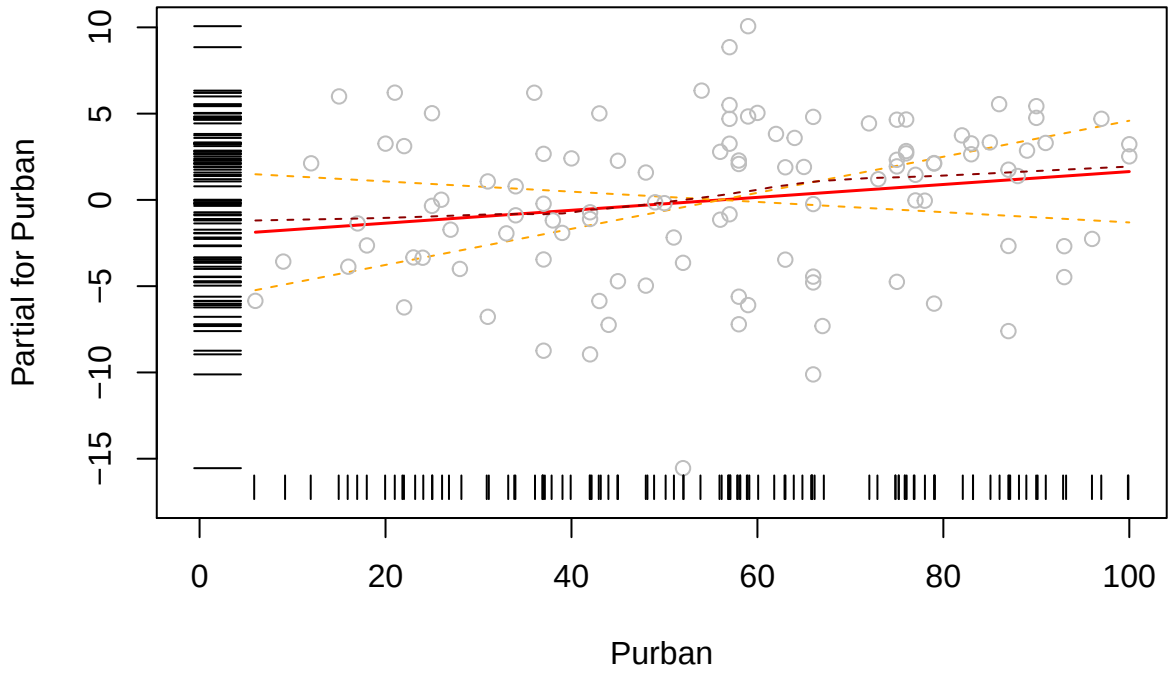

```
termplot(newnew_trans,terms="I(Pop^0.33)",partial.resid = T, se=T, rug=T,smooth = panel.smooth)
```




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



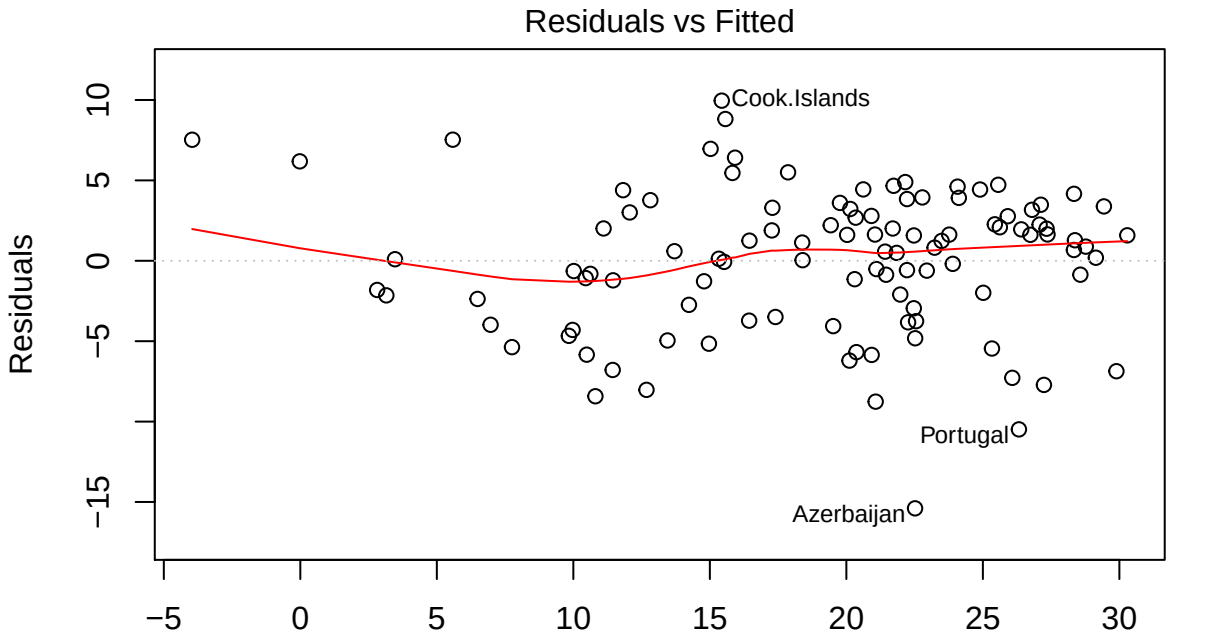

```
termplot(newnew_trans,terms="Purban",partial.resid = T, se=T, rug=T,smooth = panel.smooth)
```

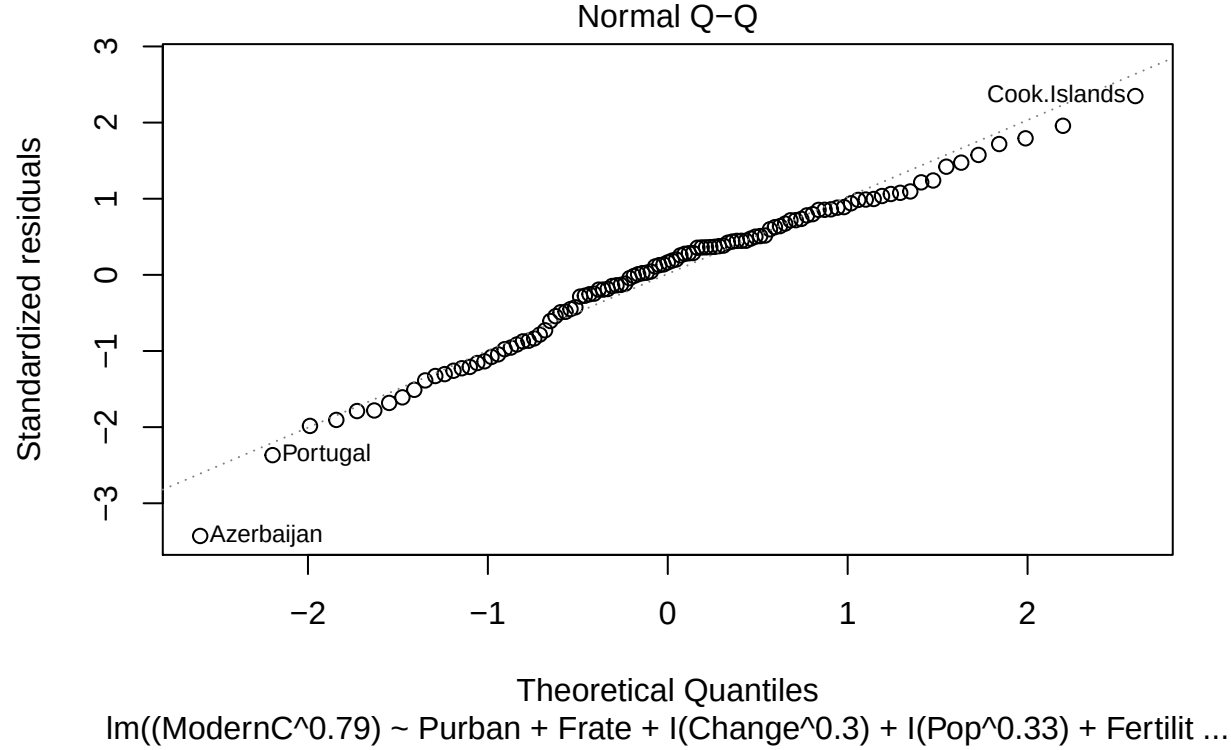


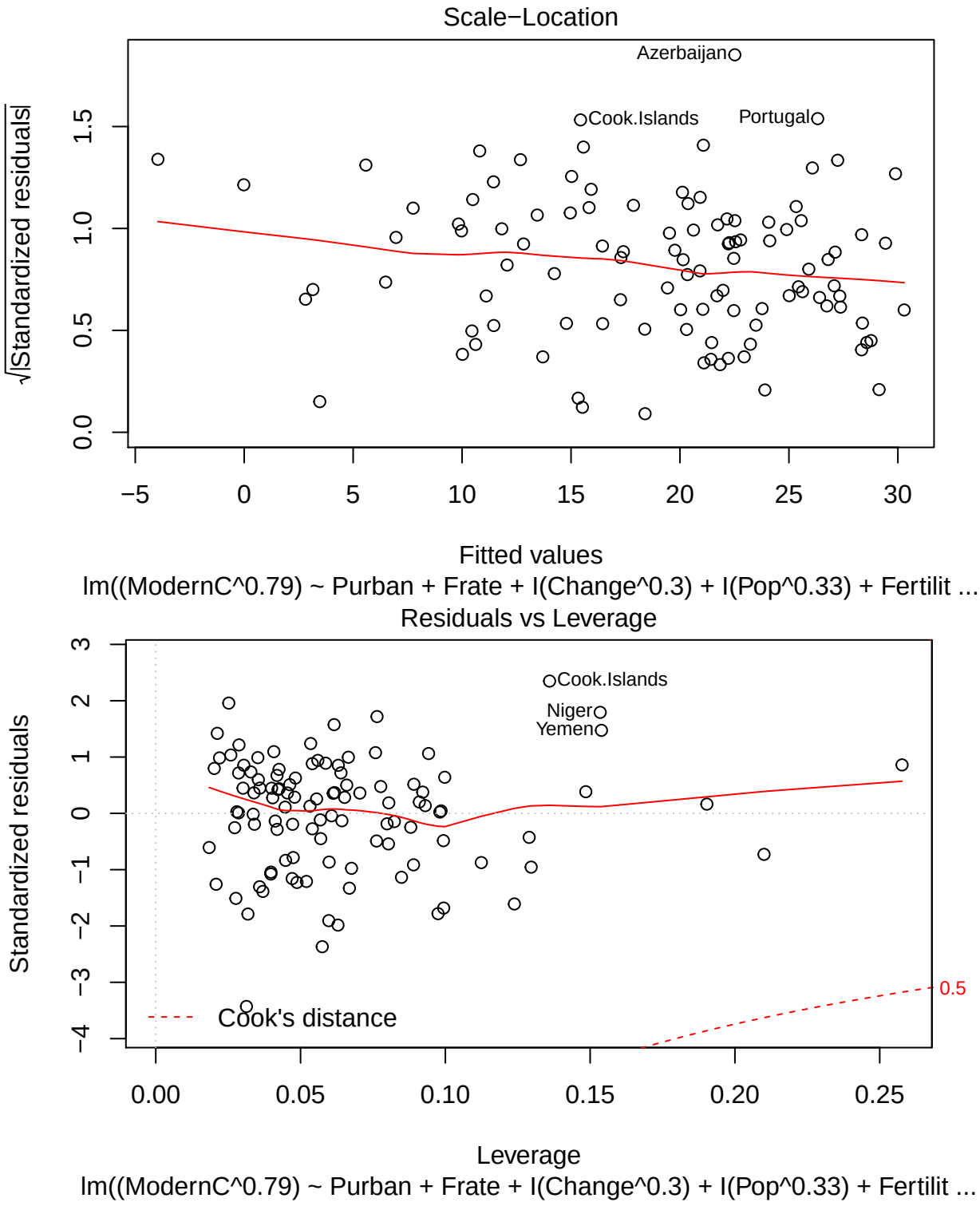
Answer:
1. from addv plot, we can see that for I(Popˆ0.33), the locality seems to be China.
2. from termplot for I(Popˆ0.33), it seems that there are 2 localities: China and India.
There seems to be no obvious localities in other plots.

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

```
plot(newnew_trans)
```







Answer: No, there is no outliers – no points have cook's distance larger than 0.5.

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 outliers or influential points.

```r
summary(g)
```

```
##
## 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
```

```r
summary(newnew_trans)
```

```
##
## Call:
## lm(formula = (ModernC^0.79) ~ Purban + Frate + I(Change^0.3) +
##     I(Pop^0.33) + Fertility + PPgdp, data = UN3_NAA)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.395  -2.846   0.672   3.088   9.958
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2.584e+01  3.828e+00   6.750 9.81e-10 ***
## Purban         3.738e-02  3.355e-02   1.114    0.268
## Frate          3.479e-02  2.753e-02   1.264    0.209
## I(Change^0.3)  9.770e-01  2.780e+00   0.351    0.726
## I(Pop^0.33)    6.298e-02  2.748e-02   2.292    0.024 *
## Fertility     -4.274e+00  5.339e-01  -8.004 2.23e-12 ***
## PPgdp          4.451e-05  7.434e-05   0.599    0.551
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.561 on 100 degrees of freedom
##   (18 observations deleted due to missingness)
## Multiple R-squared:  0.7261, Adjusted R-squared:  0.7097
## F-statistic: 44.19 on 6 and 100 DF,  p-value: < 2.2e-16
```

Answer: So my final model is ModernCˆ0.79 ~ Frate+Fertility+Purban+Changeˆ{0.3}+ Popˆ{0.3}+PPgdp. I think my model is better because the original lm model has R-squared value: 0.6183, while the new model has R-squared value: 0.7261, which means the new model fits much better to the dataset than the old one does.
Findings:
1. Percent of females over 15, per capita GDP, percent of urban ppl, popˆ0.3 –these four predictors have only a very very small influence on our final response variable, the percent of unmarried women using contraception. (Notice that such influence is insignificant).
2. The expected number of live births per female plays a critical role in predicting the percentage of unmarried woman using contraception. This correlation is very significant. Specifically, for each 1 more live birth, the percentage of unmarried women using contraception decreases by 4.27.
3. The annual ppl growth rate, when taking power of 0.3 (i.e, Changeˆ0.3) have some influence in predicting the percentage of unmarried women using contraception, although the correlation is not significant. For each unit increase in Changeˆ0.3, the precentage of unmarried women increases by approximately 1. 4. Populationˆ0.33 is positively correlated with percentage of unmarried woman using contraception, but the correlation is very small. Notice that such correlation is significant though.

Theory

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

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

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):

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

Multiply

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

to each side of (1), WTS:

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

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

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

Multiply

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

to each side again, then the equation becomes:

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

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

Notice that

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

, 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* $X_{(i)}^T Y_{(i)} = X^T Y - x_i y_i$.

(1)

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

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

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

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

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

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