HW2 STA521 Fall18

[Lingxi Song]

Due September 23, 2018 5pm

Backgound Reading

Readings: Chapters 3-4 in Weisberg Applied Linear Regression

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
```

```
apply(UN3, 2, anyNA)
```

```
##   ModernC    Change     PPgdp     Frate       Pop Fertility    Purban 
##      TRUE      TRUE      TRUE      TRUE      TRUE      TRUE     FALSE
```

```
apply(UN3,2,is.numeric)
```

```
##   ModernC    Change     PPgdp     Frate       Pop Fertility    Purban 
##      TRUE      TRUE      TRUE      TRUE      TRUE      TRUE      TRUE
```

As illustrated by the R result,6 variables (ModernC,Change,PPgdp,Frate,Pop,Fertility) have missing data. All the variables are quantitative.

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

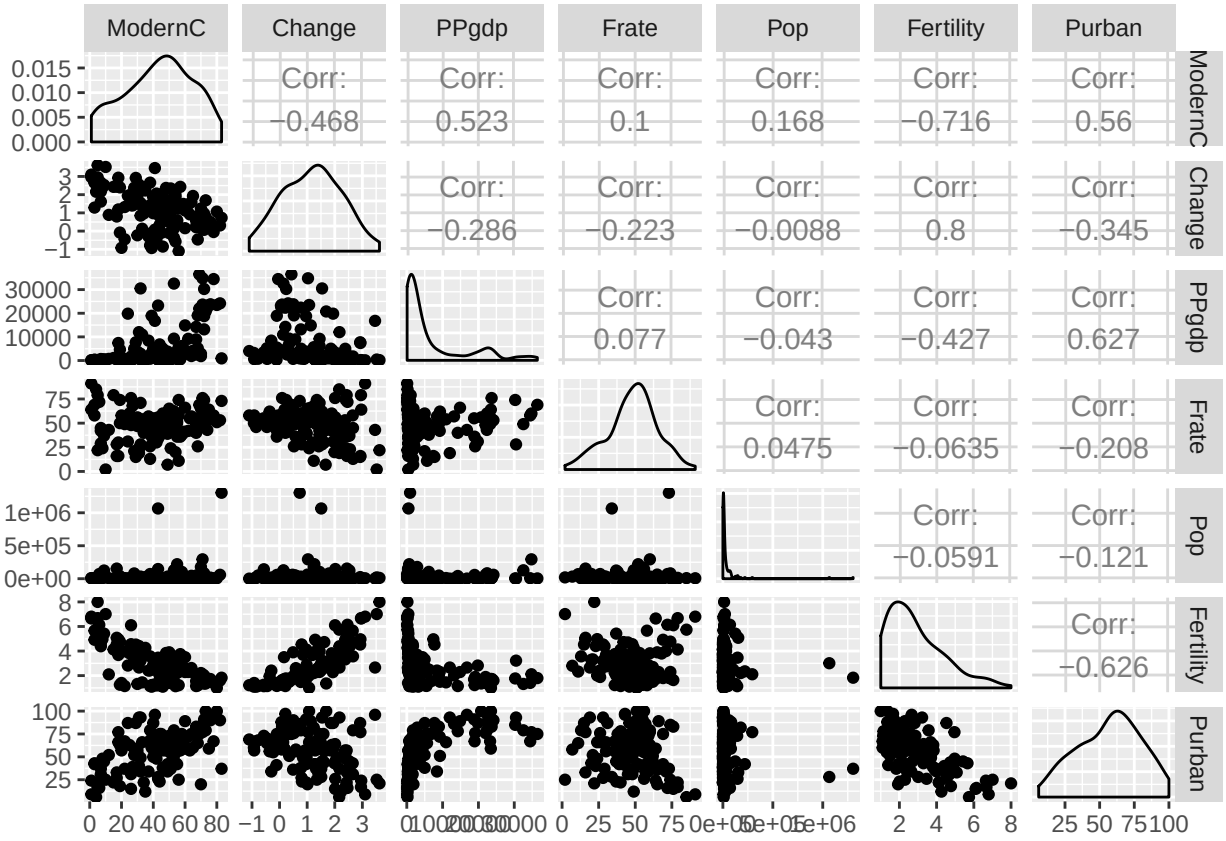
```
mstable<-matrix(nrow=ncol(UN3),ncol = 3)
colnames(mstable)<-c("variable","mean","stanard deviation")
mstable[,1]<-colnames(UN3)
mstable[,2]<-apply(UN3,2,function(x){mean(x,na.rm=TRUE)})
```

```
mstable[,3]<-apply(UN3,2,function(x){sd(x,na.rm=TRUE)})
knitr::kable(mstable)
```

variable	mean	stanard deviation
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 based on your graphical EDA?

```
ggp<-ggpairs(na.omit(UN3))
ggp
```



From correlation coefficient we can guess purban,fertility, ppdgp, and change are useful in predicting modernC.Also the scatter plot of Frate and PPgdp doesn't seem so linear, so transformation may needed. The scatter plots of Fertility and Purban show there may be high leverage points and we can only see potential outliers from Pop.

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?

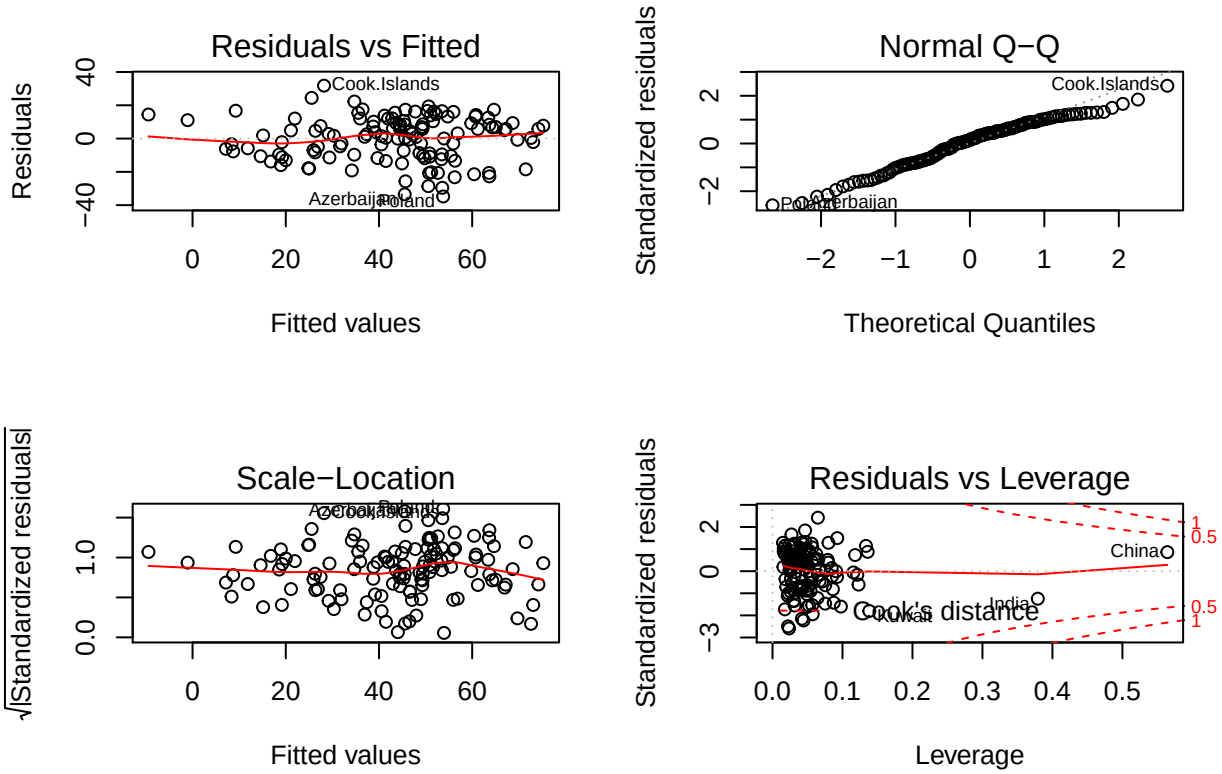
```
modernc.lm<-lm(ModernC~.,data=na.omit(UN3))
summary(modernc.lm)
```

```
##
## Call:
## lm(formula = ModernC ~ ., data = na.omit(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
## Multiple R-squared:  0.6183, Adjusted R-squared:  0.5989
## F-statistic: 31.85 on 6 and 118 DF,  p-value: < 2.2e-16
```

```
#studentized Breusch-Pagan test
lmtest::bptest(modernc.lm)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  modernc.lm
## BP = 4.697, df = 6, p-value = 0.5832
```

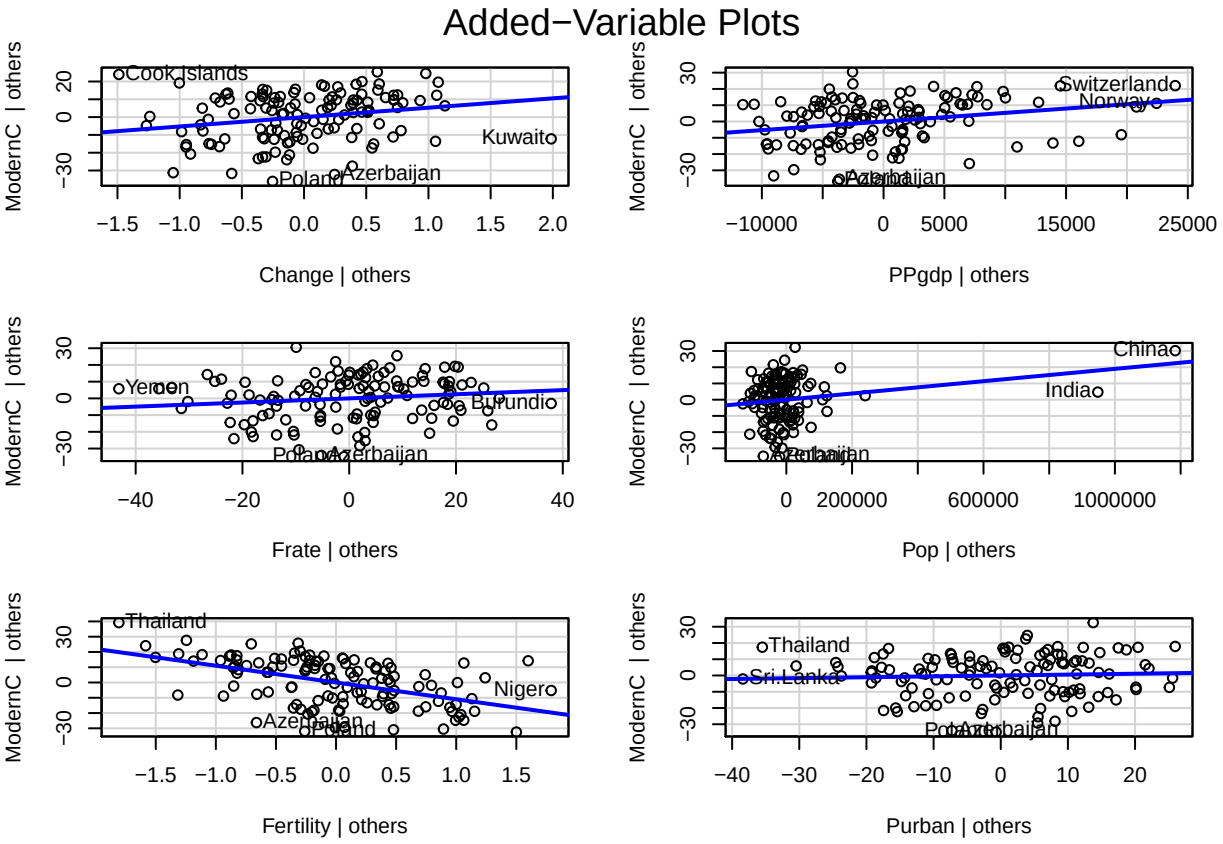
```
par(mfrow=c(2,2))
plot(modernc.lm,ask=FALSE)
```



I used 125 observations because na.omit function deleted some. As the first and third plots suggest, residual is not random. Also, the Normal Q-Q plot is not a straight 45-degree line, indicating a right tail. The last graph shows China and Indias have high leverage, so they have the potential to be influencial points. However, no points have cook's distance bigger than 1. We need to do further tests and transformations.

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)
```



The avplot for Pop shows clearly that a transformation is needed and the locality seems to be China and India.
From avplot for Change, it seems that there are 4 localities: Cook's Island,Kuwaito,Azerbaijian and Poland.
From avplot for PPgdp, it seems that there are 2 localities: Switzerland and Norway.
From avplot for Fertility, it seems that there are 2 localities: Thailand and Nigero.
From avplot for Purban, it seems that there are 2 localities: Thailand and Sri.Lanka.

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.

```
UN3_nao<-na.omit(UN3)
summary(powerTransform(cbind(PPgdp,Pop,Fertility,Purban,Change,Frate)~.,
                       family="yjPower",data = UN3_nao))
```

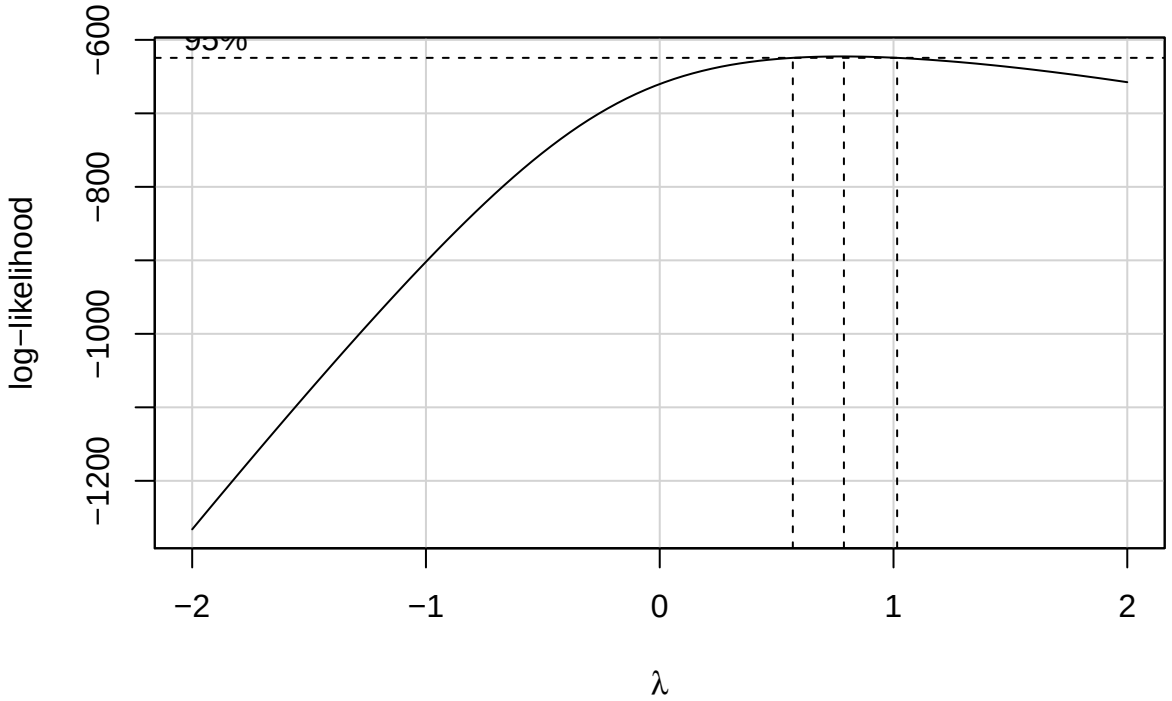
```
## yjPower Transformations to Multinormality
##            Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## PPgdp       -0.1523       -0.15      -0.2477      -0.0570
## Pop          0.0624        0.00      -0.0050       0.1298
## Fertility   -0.0991        0.00      -0.4482       0.2499
## Purban       0.9336        1.00       0.6212       1.2461
## Change       0.9654        1.00       0.7251       1.2057
## Frate        1.1020        1.00       0.7607       1.4433
##
##  Likelihood ratio test that all transformation parameters are equal to 0
##                                    LRT df       pval
## LR test, lambda = (0 0 0 0 0 0) 205.9254  6 < 2.22e-16
```

Instead of BoxTidewell, we can use powerTransform function to figure out the power of predictors. By adding

yjPower we can deal with the negative values in "Change".Accorcing to the output, Fertility,PPgdp and Pop have lambda values other than 1. But Fertility is a "good" variable so far. So I will only do log() to Pop and PPgdp. Our model is now: $ModernC \sim Change + log(PPgdp) + Frate + log(Pop) + Fertility + Purban$

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

```
boxCox(modernc.lm, lambda = seq(-2, 2, 1/10))
```



To reach the max likelihood, $\lambda \in [0.8, 1]$. But for interpretation, we can choose $\lambda = 1$(no transformation of ModernC).

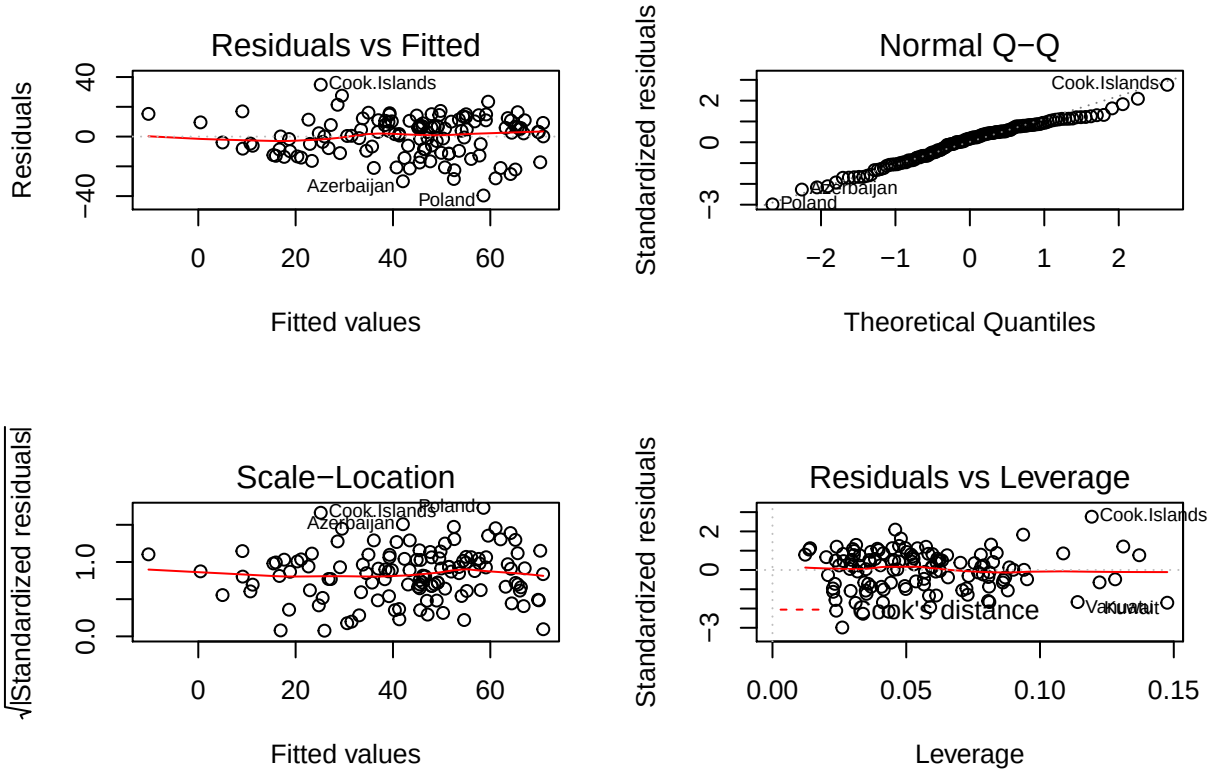
8. Fit the regression using the transformed variables. Provide residual plots and added variables plots and comment. If you feel that you need additional transformations of either the response or predictors, repeat any steps until you feel satisfied.

```
modernc.lm.2<-lm(ModernC~
                 Purban+Frate+Change+I(log(Pop))+Fertility+I(log(PPgdp)),
                 data=UN3_nao)
summary(modernc.lm.2)
```

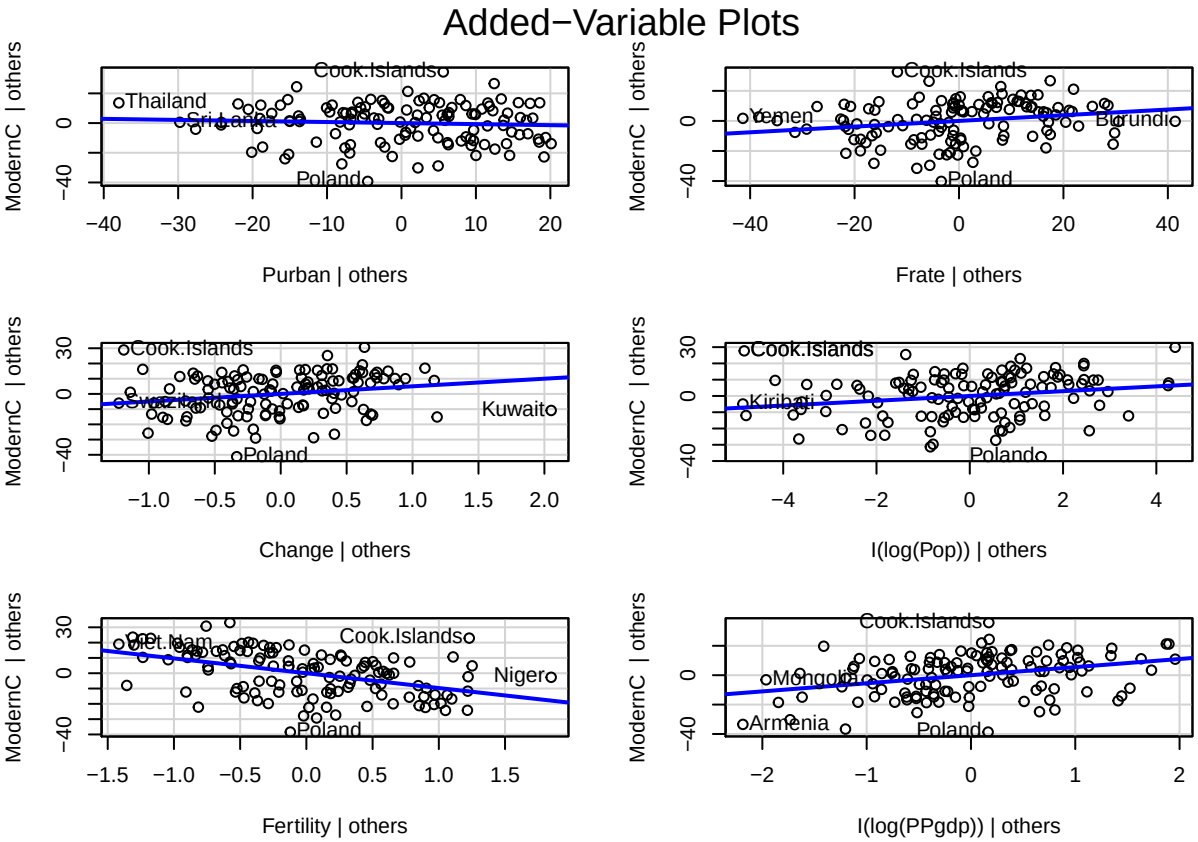
```
##
## Call:
## lm(formula = ModernC ~ Purban + Frate + Change + I(log(Pop)) +
##     Fertility + I(log(PPgdp)), data = UN3_nao)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -39.597  -9.540   2.238  10.024  34.840
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4.11547   14.50854   0.284 0.777169
## Purban        -0.07077    0.09760  -0.725 0.469829
## Frate          0.18939    0.07711   2.456 0.015500 *
```

```
## Change          4.99296    2.07709   2.404 0.017781 *
## I(log(Pop))     1.47207    0.62875   2.341 0.020897 *
## Fertility      -9.67594    1.76561  -5.480 2.44e-07 ***
## I(log(PPgdp))   5.50728    1.40505   3.920 0.000149 ***
## ---
## 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(modernc.lm.2,ask=FALSE)
```




```
avPlots(modernc.lm.2)
```



From the residual plots we see random distributrd variables and the Normal Q-Q plot is more likely a straight 45-degree line. The added variable plots show that log(Pop),log(PPgdp) are better than the original variables. After checking the plots we can say the model is satisfying.

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?

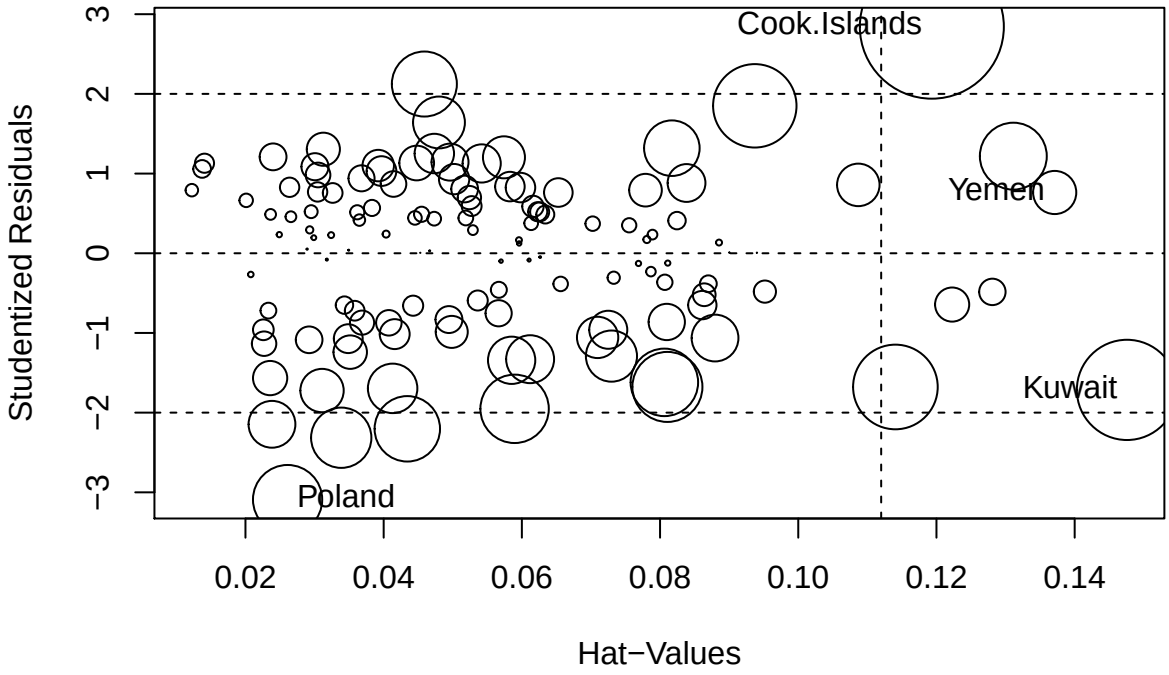
I end up with the same model in question 8.

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.

```
outlierTest(modernc.lm.2)
```

```
## No Studentized residuals with Bonferonni p < 0.05
## Largest |rstudent|:
##         rstudent unadjusted p-value Bonferonni p
## Poland -3.090987          0.0024937      0.31171
```

```
influencePlot(modernc.lm.2)
```




```
##                 StudRes        Hat      CookD
## Cook.Islands  2.8433168 0.11933235 0.14763090
## Kuwait       -1.7145962 0.14757081 0.07152969
## Poland       -3.0909870 0.02609592 0.03410030
## Yemen         0.7626724 0.13711004 0.01325056
```

The function outlierTest() and influencePlot() provide a quick way to do this. Even if the point Poland has an unadjusted p-value of 0.0036, the Bonferonni P (equals to the unadjusted P multiplies observation number) is larger than 0.05. Therefore, we can't reject the H_0: There are no outliers in the data. And it's obvious from the plot that Yeman, Poland, Kuwait and Cook.islands are influencial points. So we don't have to refit our final model.

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!

```
c<-confint(modernc.lm.2,level=0.95)
knitr::kable(c)
```

	2.5 %
(Intercept)	-24.6153857
Purban	-0.2640391
Frate	0.0366943
Change	0.8797496
I(log(Pop))	0.2269699
Fertility	-13.1723343
I(log(PPgdp))	2.7249039

95% confidence i nterval means the frequency of possible confidence intervals that contain the true value of the unknown

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.

```
summary(modernc.lm.2)
```

```
##
## Call:
## lm(formula = ModernC ~ Purban + Frate + Change + I(log(Pop)) +
##     Fertility + I(log(PPgdp)), data = UN3_nao)
##
## 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
## Purban        -0.07077    0.09760  -0.725 0.469829
## Frate          0.18939    0.07711   2.456 0.015500 *
## Change         4.99296    2.07709   2.404 0.017781 *
## I(log(Pop))    1.47207    0.62875   2.341 0.020897 *
## Fertility     -9.67594    1.76561  -5.480 2.44e-07 ***
## I(log(PPgdp))  5.50728    1.40505   3.920 0.000149 ***
## ---
## 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
```

The final model is ModernC ~ Frate+Fertility+Purban+Change+log(Pop)+log(PPgdp). From the adjusted R-squared and residual plots we can conclude the final model is better. As suggested in the result, only Fertility is negatively correlated with ModernC. Variable Purban is not significant.

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.

$e_Y = \beta_0 + \beta_1 e_x$

$\beta_0 = \bar{e_Y} - \beta_1 \bar{e_x}$(regression line passes through the center point)

since $\bar{e_Y} = \frac{1}{n}1_n^T(I-H)Y$,$\bar{e_x} = \frac{1}{n}1_n^T(I-H)X_i$

therefore,use the hint we can get

$\beta_0 = \frac{1}{n}1_n^T(I-H)Y - \frac{1}{n}\beta_1 1_n^T(I-H)X_i = 0$

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.

```
e1<-residuals(lm(ModernC~Purban+Frate+Change+I(log(Pop))+I(log(PPgdp)),
              data=UN3_nao))
e2<-residuals(lm(Fertility~Purban+Frate+Change+I(log(Pop))+I(log(PPgdp)),
              data=UN3_nao))
test<-lm(e1~e2)
summary(test)$coef
```

```
##                  Estimate Std. Error       t value     Pr(>|t|)
## (Intercept) -8.276956e-16   1.177707 -7.028028e-16 1.000000e+00
## e2          -9.675941e+00   1.729353 -5.595121e+00 1.359835e-07
```

```
summary(modernc.lm.2)$coef
```

```
##                  Estimate  Std. Error    t value     Pr(>|t|)
## (Intercept)    4.11547111 14.50853884  0.2836586 7.771692e-01
## Purban        -0.07076799  0.09759825 -0.7250948 4.698293e-01
## Frate          0.18939357  0.07711025  2.4561402 1.550017e-02
## Change         4.99295735  2.07709205  2.4038209 1.778126e-02
## I(log(Pop))    1.47207436  0.62875419  2.3412557 2.089650e-02
## Fertility     -9.67594142  1.76561222 -5.4802189 2.444298e-07
## I(log(PPgdp))  5.50727842  1.40504647  3.9196415 1.492131e-04
```

We can use the variable log(Fertility) as an example to confirm the statement. As suggested above: coefficients have the same value -9.676, which means the same slope.