HW2 STA521 Fall18

[Mingjie Zhao, mz136,mingjiezhao]

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. *Exclude text from final*

```
#https://rweb.stat.umn.edu/R/site-library/alr3/html/UN3.html
data(UN3, package="alr3")
library(car)
```

1. Create a summary of the data. How many variables have missing data? Which are quantitative and which are qualtitative?
   Answer: 6 variables have missing data, I think all the variables are quantitative variables

```
#check first a few rows and summary
head(UN3)
```

```
##             ModernC Change PPgdp Frate   Pop Fertility Purban
## Afghanistan      NA   3.88    98    NA 23897      6.80     22
## Albania          NA   0.68  1317    NA  3167      2.28     43
## Algeria          49   1.67  1784     7 31800      2.80     58
## Am.Samoa         NA   2.37    NA    42    57        NA     53
## Andorra          NA   2.59 14234    NA    64        NA     92
## Angola           NA   3.20   739    NA 13625      7.20     35
```

```
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
```

```
#check characteristics of the variables
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 ...
```

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

```
table1=t(as.matrix(rbind(round(apply(UN3[ 1:7],2,mean,na.rm=TRUE),3),
                         round(apply(UN3[ 1:7],2,sd,na.rm=TRUE),3))))
colnames(table1)=c("mean","standard deviation")
kable(table1,caption = c("mean and standard deviation of the quantitative variables"))
```

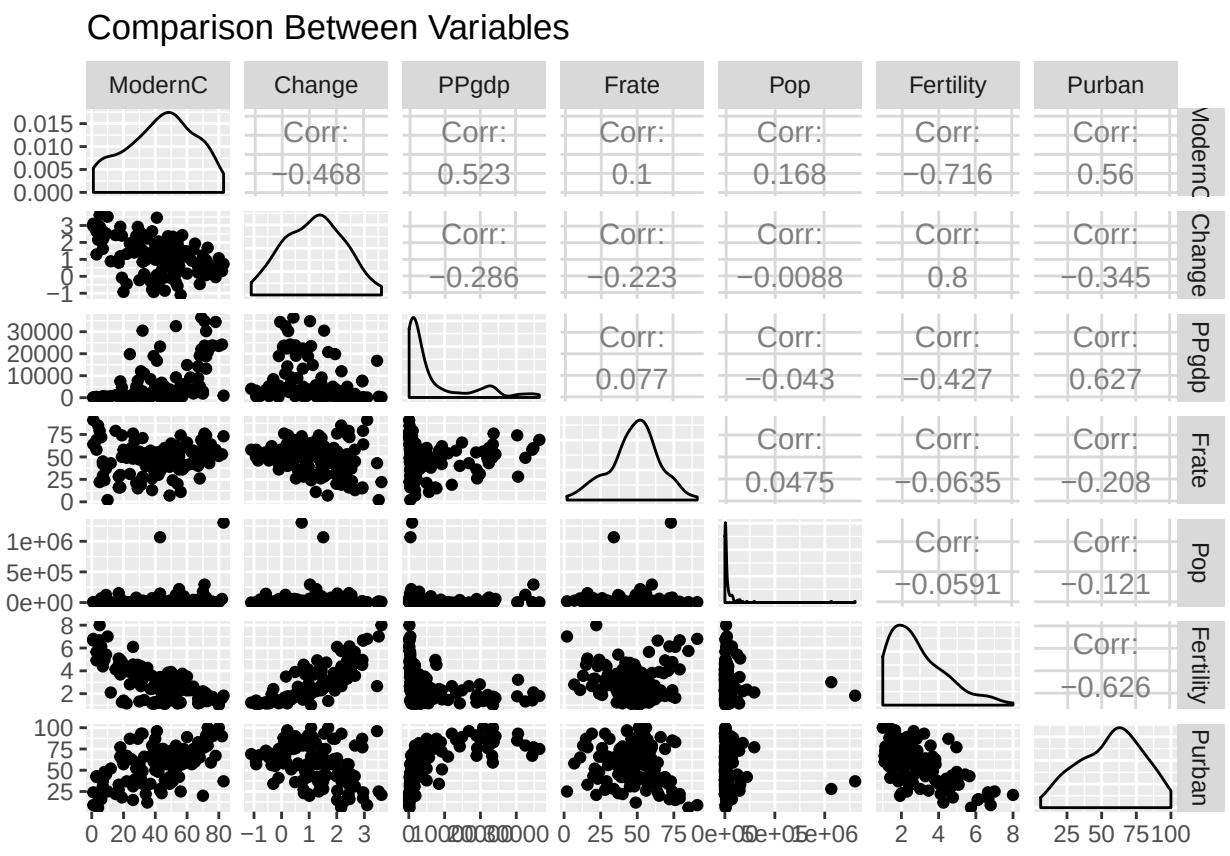
Table 1: mean and standard deviation of the quantitative variables

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

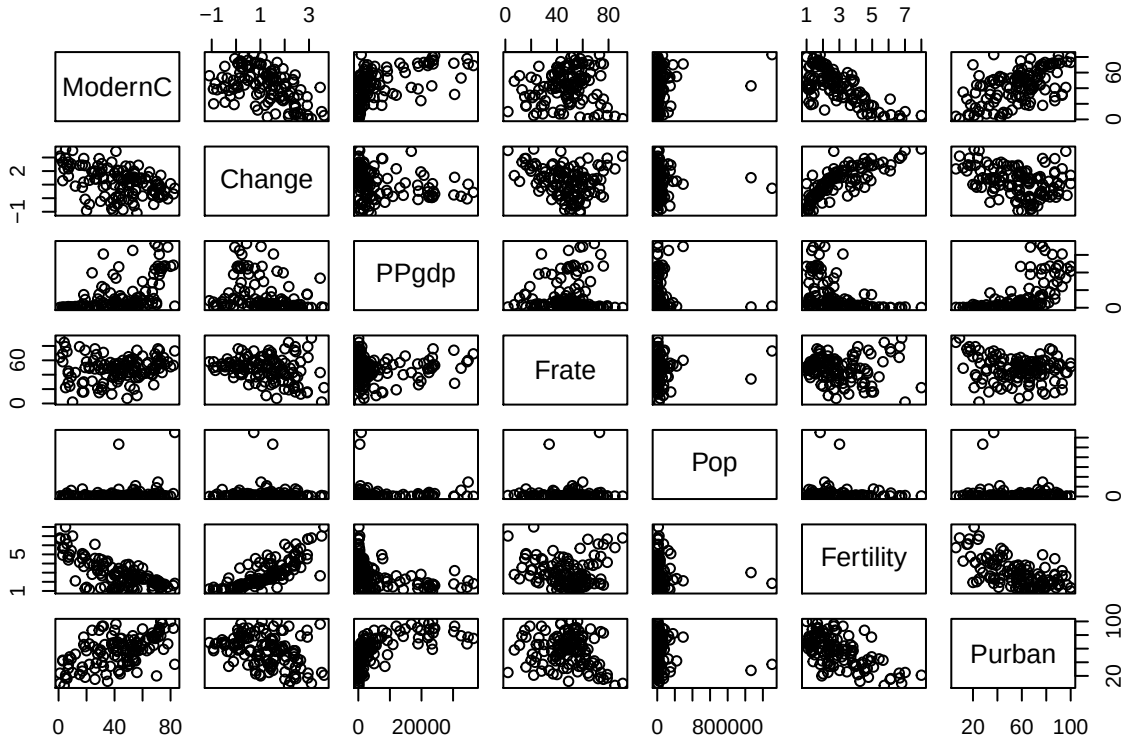
3. Investigate the predictors graphically, using scatterplots or other tools of your choice. Create some plots highlighting the relationships among the predictors. Comment on your findings regarding trying to predict `ModernC` from the other variables. Are there potential outliers, nonlinear relationships or transformations that appear to be needed based on your graphical EDA?

   Answer: from the ggpairs, we can see that the correlation values between ModernC and other variables range from 0.1 to -0.716. The PPdgp and Pop may need transformations because their distributions show bad normality. From plot generated by "pairs", it is hard to tell the relationships between Pop/ PPdgp and other variables, indicating that pop and PPdgp need to be transformed to reveal its relationship with other variables. There are potential outliers in the plot of ModernC with Change, PPdgp and Pop.

```
UN3_1=na.omit(UN3)
sp=ggpairs(UN3_1,  progress=F,title="Comparison Between Variables")
suppressWarnings(print(sp, progress=F))
```


Comparison Between Variables
ModernC
Change
PPgdp
Frate
Pop
Fertility
Purban
Corr: −0.468
Corr: 0.523
Corr: 0.1
Corr: 0.168
Corr: −0.716
Corr: 0.56
Corr: −0.286
Corr: −0.223
Corr: −0.0088
Corr: 0.8
Corr: −0.345
Corr: 0.077
Corr: −0.043
Corr: −0.427
Corr: 0.627
Corr: 0.0475
Corr: −0.0635
Corr: −0.208
Corr: −0.0591
Corr: −0.121
Corr: −0.626

```
pairs(UN3_1)
```


ModernC
Change
PPgdp
Frate
Pop
Fertility
Purban

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?

   Answer: there are 125 observations are used in my model fitting.
   From the residual plots:
   1) residuals vs fitted: the plot has relatively good constant variance, but there are some extreme data points, could be outliers
   2) Normal Q-Q: the plot has heavy tails close to 3 std dev. The observed quartiles are larger than expected under normality assumption. There might be some outliers at both left and right sides
   3) scale-location: indicates that the variance is relatively constant with the mean, but there are some potential outliers
   4) residuals vs leverage: Cook's distance show that there are a few influential points, like China and India

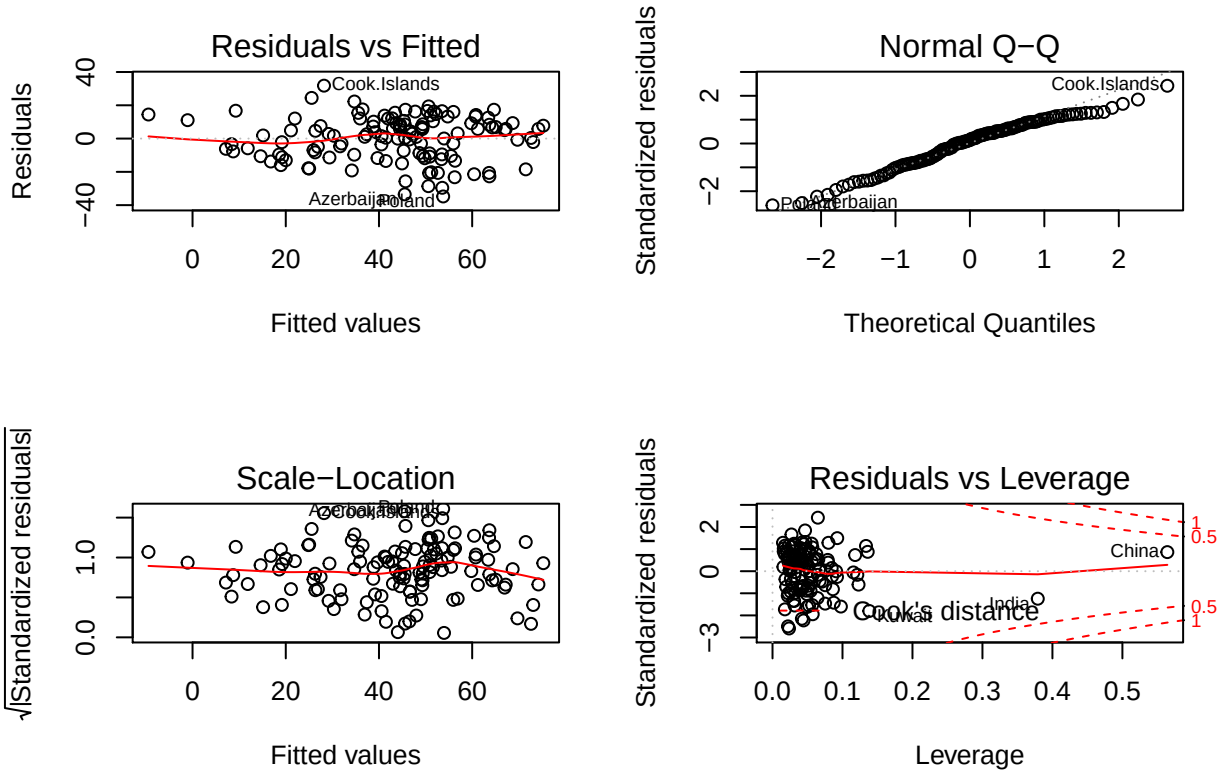
```
nrow(UN3_1)
```

```
## [1] 125
```

```
model_0=lm(ModernC~.,data=UN3_1)
summary(model_0)
```

```
##
## Call:
## lm(formula = ModernC ~ ., data = UN3_1)
##
## 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
```

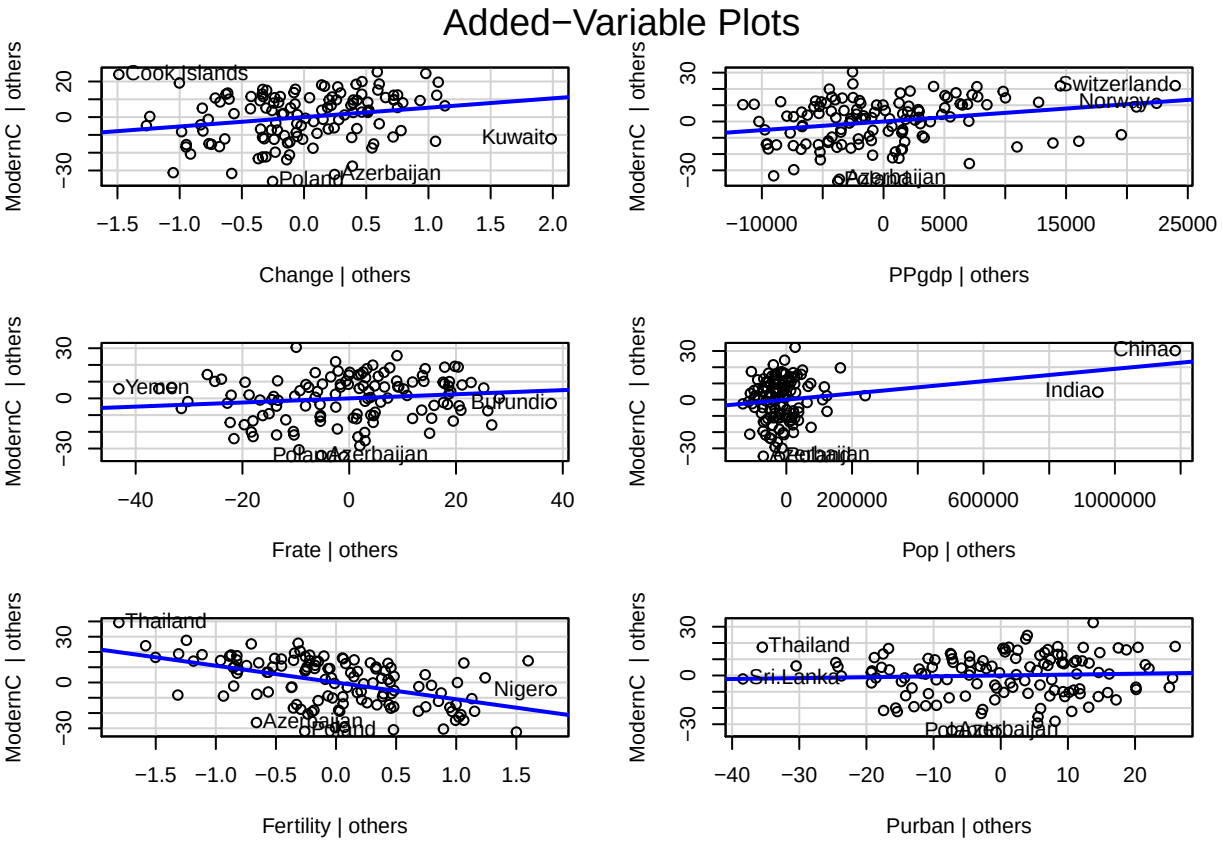
```
par(mfrow=c(2,2))
plot(model_0,ask=F)
```



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?

   -Answer: influential localities are China and India, Pop needs to be transformed
   -Explanation: In the added variable plots, the one for ModernC and Pop (2nd row, on the right side) caught my attention because the data is clustered together. There are two influential observations (China and India) are far away from the other observations, indicating that the variable Pop needs to be transformed before we use it as a prediction for ModernC.

```
par(mfrow=c(3,2))
avPlots(model_0)
```



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.

   - Answer: The range of Change is [-1.10 3.62], which include negative values. So I add 2.1 to all the values for Change to make the range become [1, 5.72]. There are 2 reasons: 1) I did not use abs() function, because the Change indicates the percentage of Annual population growth rate, we may interested in looking at the difference of Change for different counties, so we don't want to use abs(), 2) I want the minimum value of Change to be 1, so if there is a need for log transformation, all the values would be positive after the log transformation.
   - Based on the graphical EDA, I think the following variables need transformation: PPdgp and Pop
   - Reasons for transformations: 1) PPdgp has MLE of Lambda of -0.129, which is very close to 0, so I will do log transformation, 2) Pop as MLE of Lambda of 0.407, which is very close to 0.5, so I will do the square root transformation.
   - So at this point, my model 1 will be ModernC~Change+log(PPgdp)+Frate+sqrt(Pop)+Fertility+Purban

```
# 1 Change
range(UN3_1$Change)
```

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

```
UN3_2=UN3_1 #create a new dataframe
UN3_2$Change=UN3_2$Change+2.1
range(UN3_2$Change)
```

```
## [1] 1.00 5.72
```

```
# transform two variables
car::boxTidwell(ModernC~Pop+PPgdp,~Change+Frate+Fertility+Purban,data=UN3_1)
```

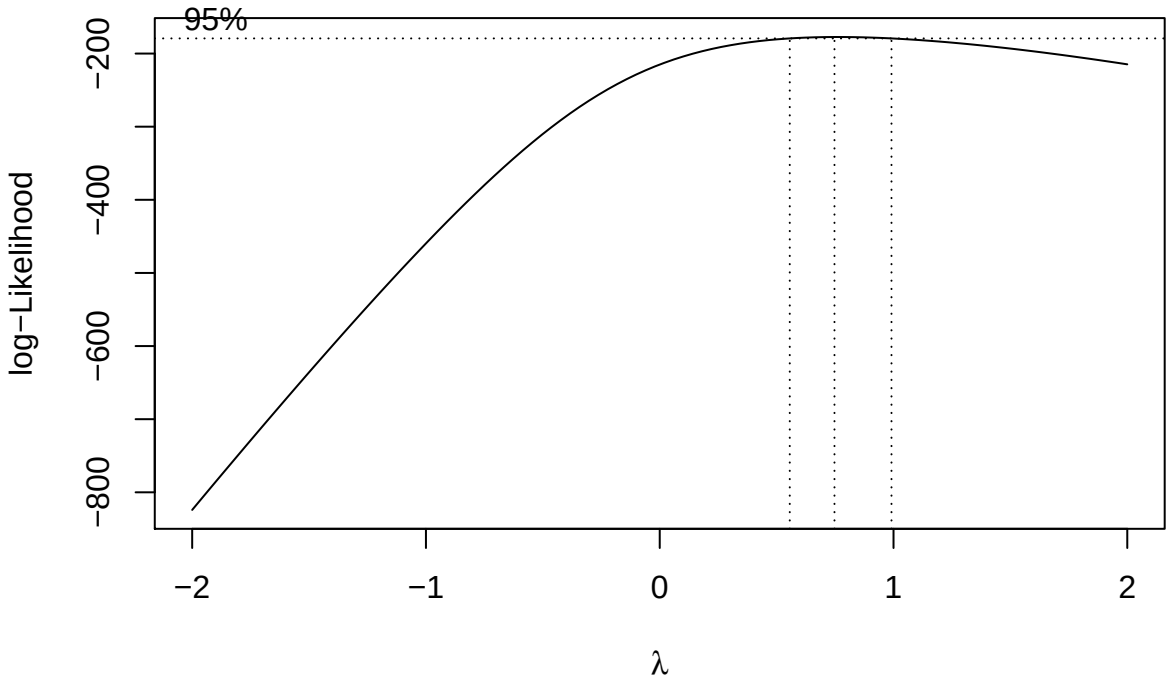
```
##       MLE of lambda Score Statistic (z) Pr(>|z|)
## Pop         0.40749             -0.7874   0.4310
## PPgdp      -0.12921             -1.1410   0.2539
##
## iterations =  4
```

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

   Answer: with the boxCox results, we can see that the 95% interval for lambda is about 0.6 to 1. So I think it is not necessary to transform y at this point because $\lambda = 1$ indicates no transformation is needed.

```
# regression after tranformation for x
model_1=lm(ModernC~Change+log(PPgdp)+Frate+sqrt(Pop)+Fertility+Purban,data=UN3_2)

MASS::boxcox(model_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.

   -Answer: From the regression results of Model 1, we can see that compared to the original model in Ex.4, Model 1 is better thanks to the transformation. Because the residual plots show better results, and R-squared values are improved.
   -Although the boxCox plot indicates that y does not need to be transformed, I am curious to see if transforming y can result in better regression results. So I tried two new models with transformed y: model 2 (use log transformation for y) and model 3 (use square root transformation for y)

   -Explanations: 1) log transformation is a common way to transform variable. Since I already have log transformation for one of my predictors PPgdp, I want to see if log transformation works for y too. 2)

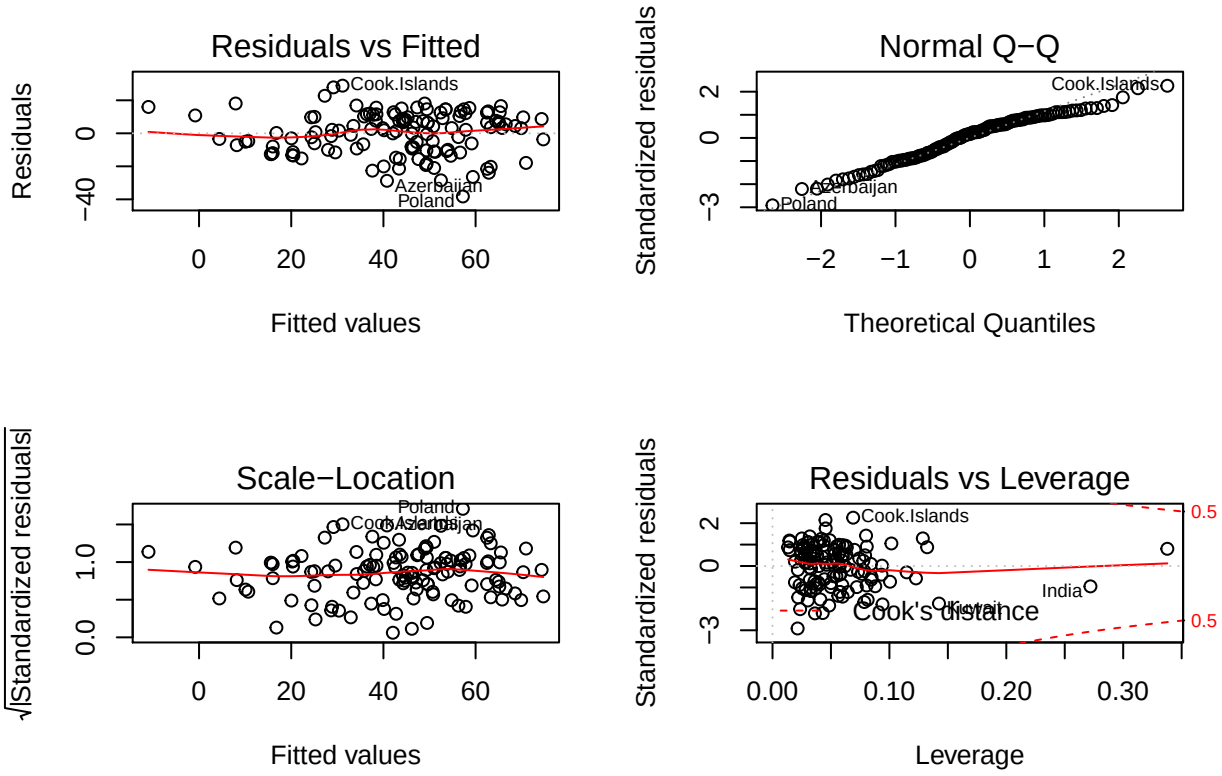
values of y might have a poisson distribution, and the squart root transformation is very common for poisson variables, so I want to try it.

-Results: with the regression results of model 2 and model 3, we can see that Model 3 fits slightly better than Model 2. Both models have slightly better R-squared values than model 1. But there is no obvious improvements in terms of graphical EDA and after y transformation, a few predictors (for example, Frate and Pop) are no longer significant in terms of their p-values. So I think Model 1 is still the best model I've got at this point.

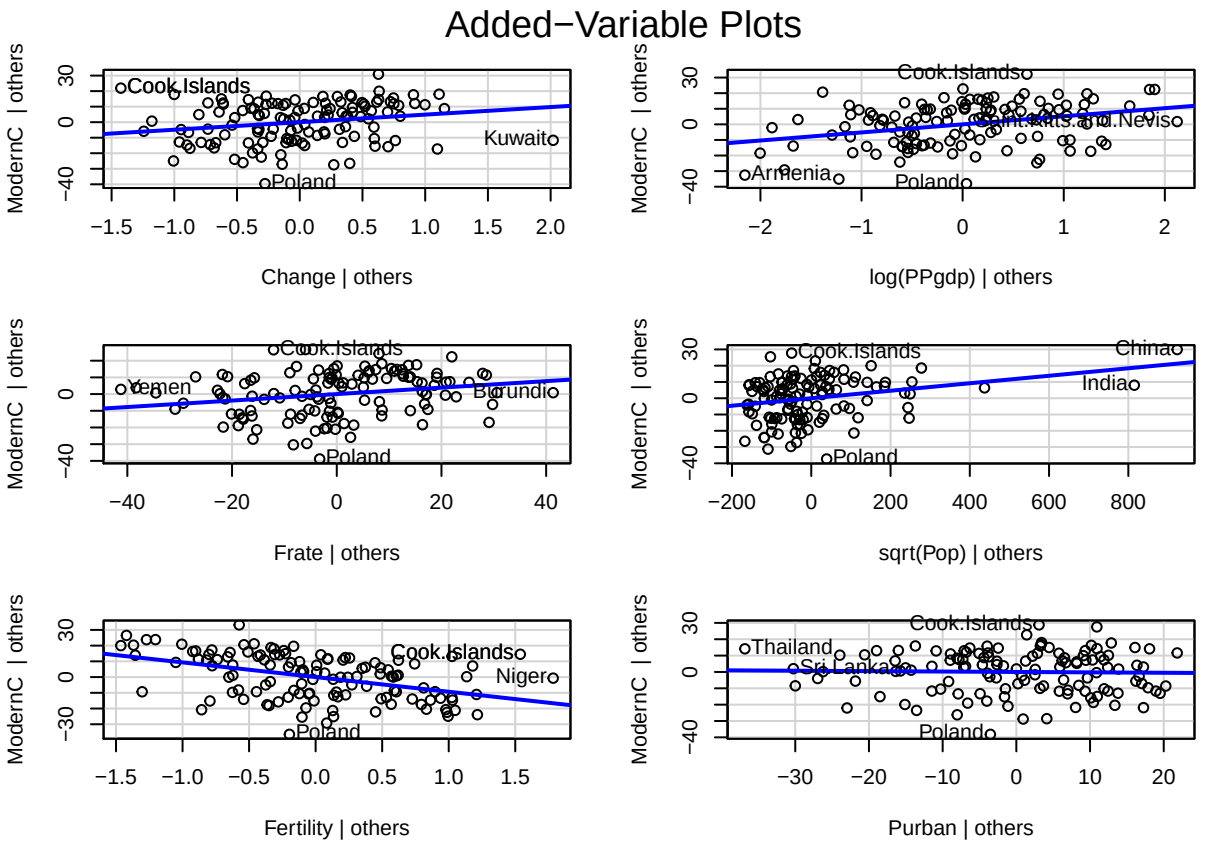
```
summary(model_1)
```

```
##
## Call:
## lm(formula = ModernC ~ Change + log(PPgdp) + Frate + sqrt(Pop) +
##     Fertility + Purban, data = UN3_2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -38.239  -9.995   2.133   9.961  28.861
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.742857  12.227659   0.224  0.82290
## Change       4.869147   2.042766   2.384  0.01874 *
## log(PPgdp)   5.182415   1.359076   3.813  0.00022 ***
## Frate        0.194010   0.076053   2.551  0.01202 *
## sqrt(Pop)    0.023017   0.007685   2.995  0.00335 **
## Fertility   -9.327572   1.749773  -5.331 4.77e-07 ***
## Purban      -0.025072   0.096557  -0.260  0.79558
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.26 on 118 degrees of freedom
## Multiple R-squared:  0.6362, Adjusted R-squared:  0.6177
## F-statistic:  34.4 on 6 and 118 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(model_1,ask=F)
```



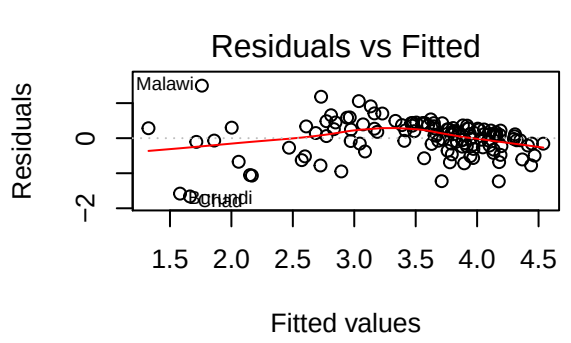

```
par(mfrow=c(2,2))
avPlots(model_1)
```

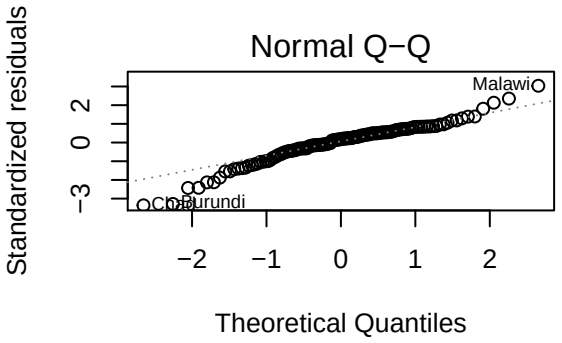


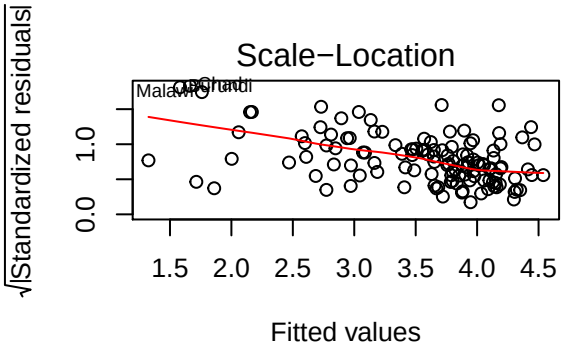
```r
# try log transformation for y
model_2=lm(log(ModernC)~Change+log(PPgdp)+Frate+sqrt(Pop)+Fertility+Purban,data=UN3_2)
summary(model_2)
```

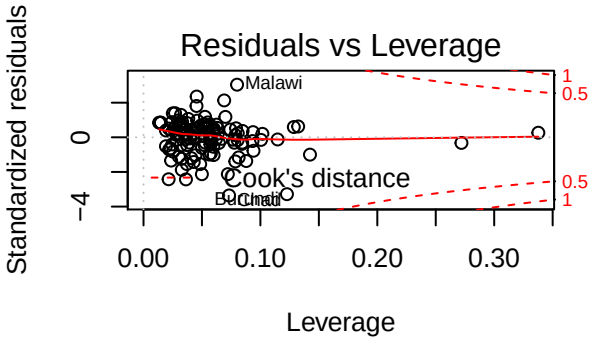
```
##
## Call:
## lm(formula = log(ModernC) ~ Change + log(PPgdp) + Frate + sqrt(Pop) +
##     Fertility + Purban, data = UN3_2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.66269 -0.22732  0.09175  0.28439  1.49837
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.5590286  0.4742487   7.505 1.27e-11 ***
## Change       0.2174183  0.0792285   2.744  0.00701 **
## log(PPgdp)   0.1491042  0.0527117   2.829  0.00549 **
## Frate       -0.0033012  0.0029497  -1.119  0.26535
## sqrt(Pop)    0.0004437  0.0002981   1.488  0.13929
## Fertility   -0.5132022  0.0678648  -7.562 9.45e-12 ***
## Purban      -0.0056600  0.0037450  -1.511  0.13337
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5142 on 118 degrees of freedom
## Multiple R-squared:  0.6628, Adjusted R-squared:  0.6457
## F-statistic: 38.67 on 6 and 118 DF,  p-value: < 2.2e-16
```

```r
par(mfrow=c(2,2))
plot(model_2,ask=F)
```

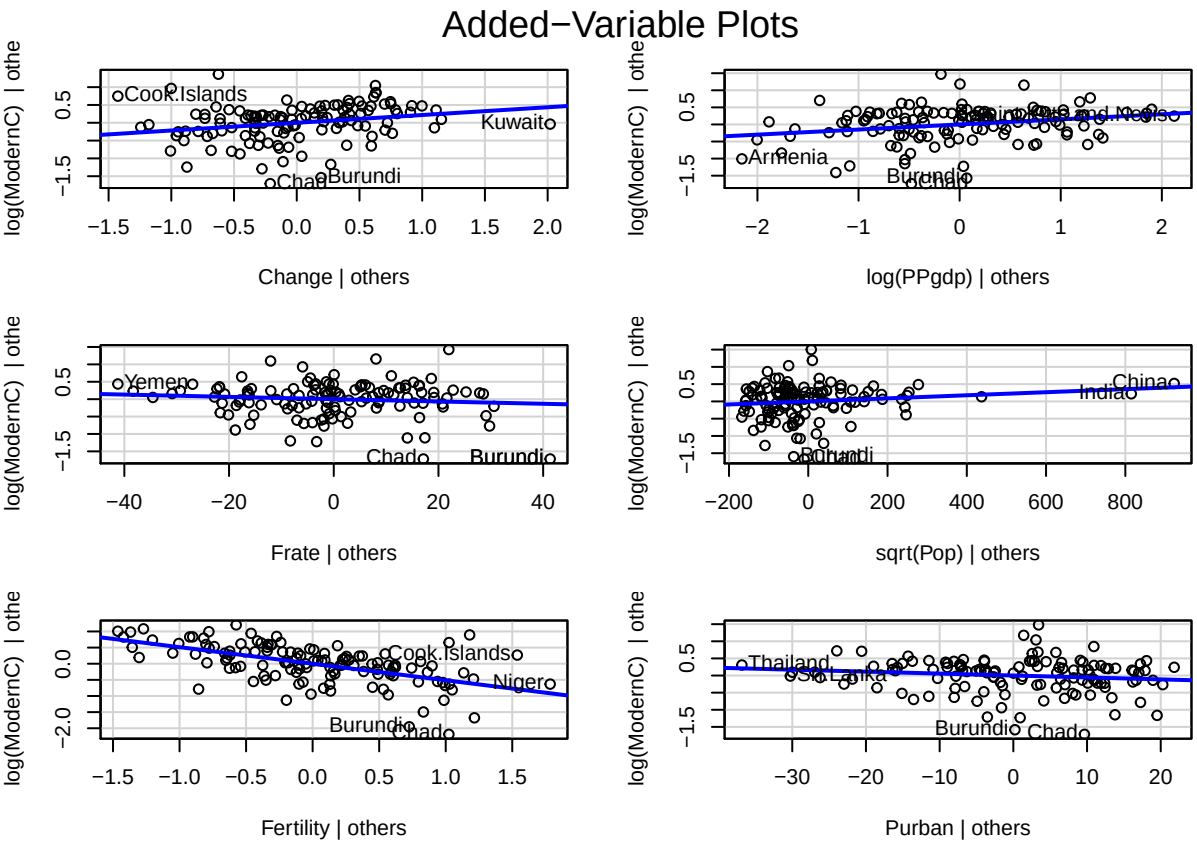








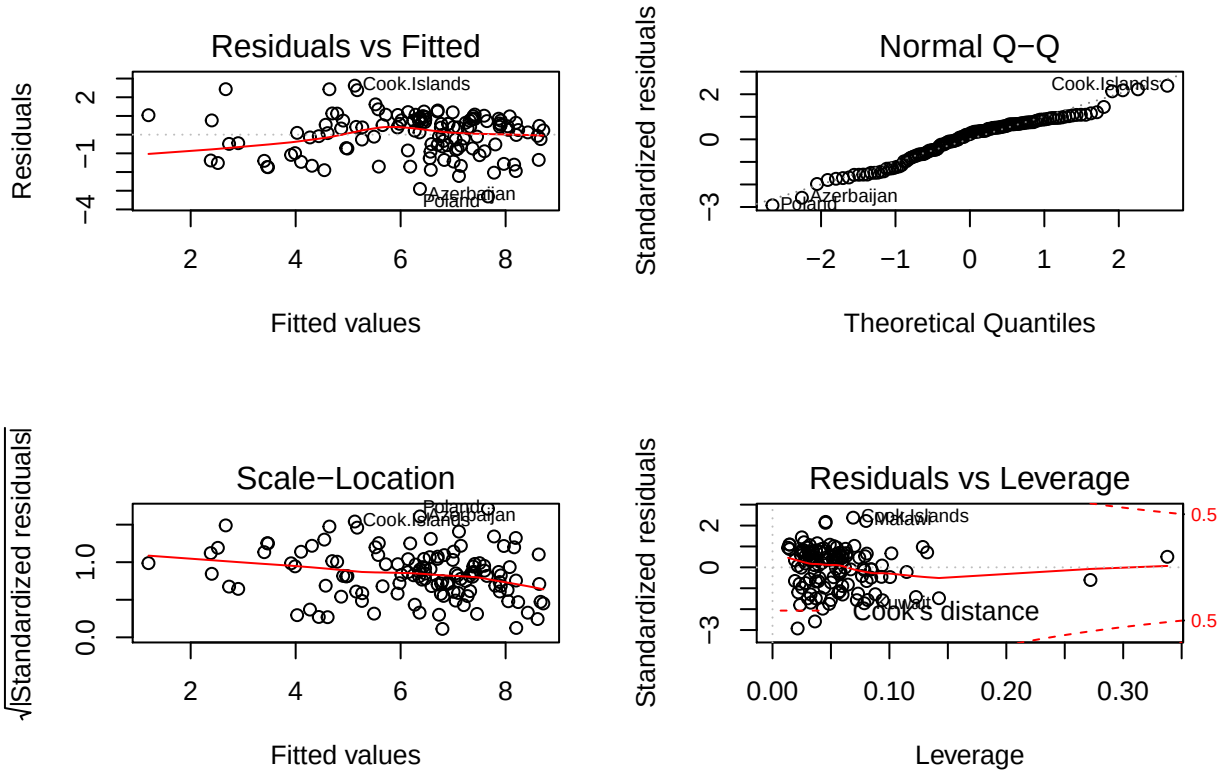

```
par(mfrow=c(3,2))
avPlots(model_2)
```



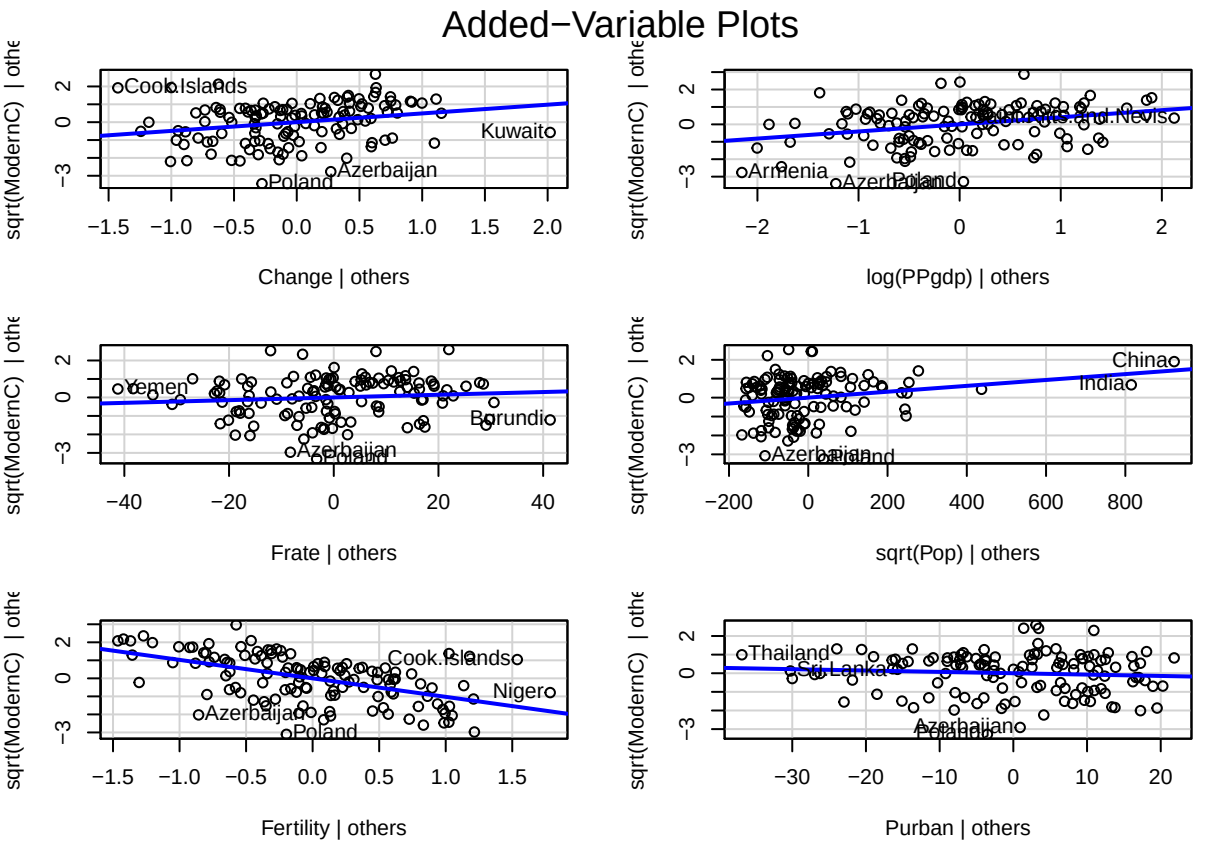
```r
# try sqrt transformation for y
model_3=lm(sqrt(ModernC)~Change+log(PPgdp)+Frate+sqrt(Pop)+Fertility+Purban,data=UN3_2)
summary(model_3)
```

```
##
## Call:
## lm(formula = sqrt(ModernC) ~ Change + log(PPgdp) + Frate + sqrt(Pop) +
##     Fertility + Purban, data = UN3_2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3061 -0.7433  0.2555  0.7604  2.6193
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.3096551  1.0538607   4.089 7.93e-05 ***
## Change       0.4893648  0.1760591   2.780 0.006337 **
## log(PPgdp)   0.4080004  0.1171342   3.483 0.000696 ***
## Frate        0.0073017  0.0065547   1.114 0.267562
## sqrt(Pop)    0.0015549  0.0006624   2.348 0.020562 *
## Fertility   -1.0250633  0.1508071  -6.797 4.63e-10 ***
## Purban      -0.0072729  0.0083219  -0.874 0.383922
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.143 on 118 degrees of freedom
## Multiple R-squared:  0.6678, Adjusted R-squared:  0.6509
## F-statistic: 39.53 on 6 and 118 DF,  p-value: < 2.2e-16
```

```r
par(mfrow=c(2,2))
plot(model_3,ask=F)
```



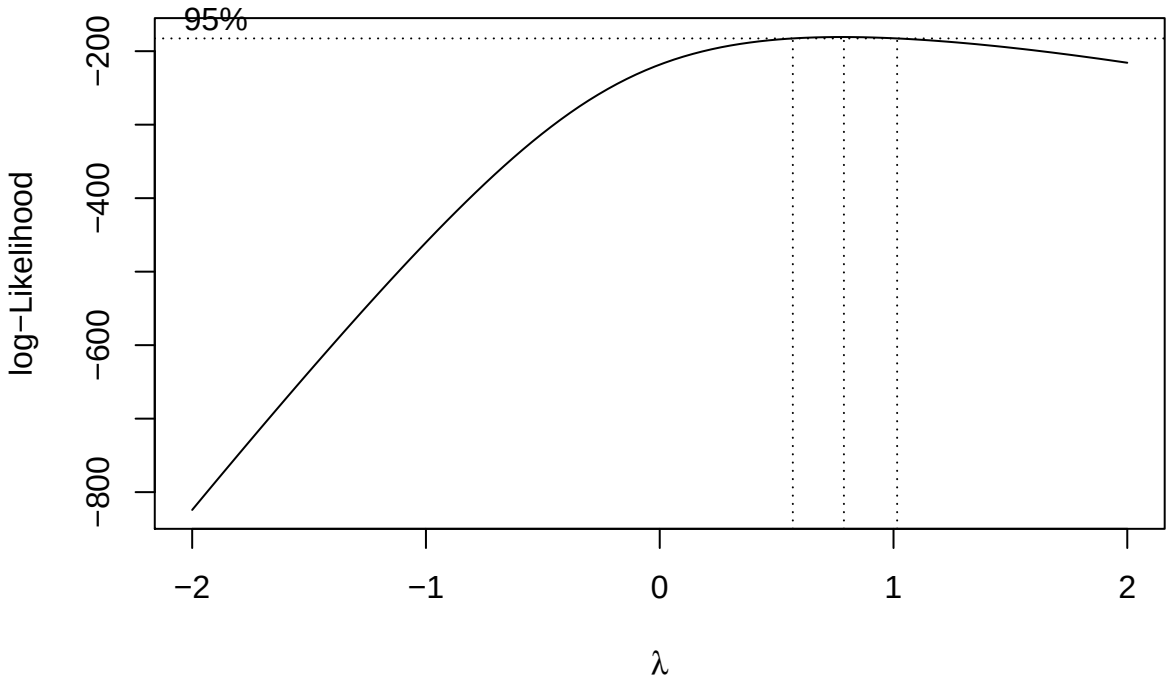

```
par(mfrow=c(3,2))
avPlots(model_3)
```



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?

   Answer: No I did not get a different model.
   Explanations: Step 1: to first find the best transformation of the response, I used Model 0 (original full model) for boxCox. And the resulted plot is similar to the one in Ex. 7, which indicates that y does not need to be transformed.
   Step 2: use the Box-Tidwell to find appropriate transformations of the predictor variables. The MLE of lambda are the same as that in Ex. 6. So Model 1 is still the best model so far.

```
# use Model 0 for boxCox
MASS::boxcox(model_0)
```




```
# transform two variables
car::boxTidwell(ModernC~Pop+PPgdp,~Change+Frate+Fertility+Purban,data=UN3_2)
```

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

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.

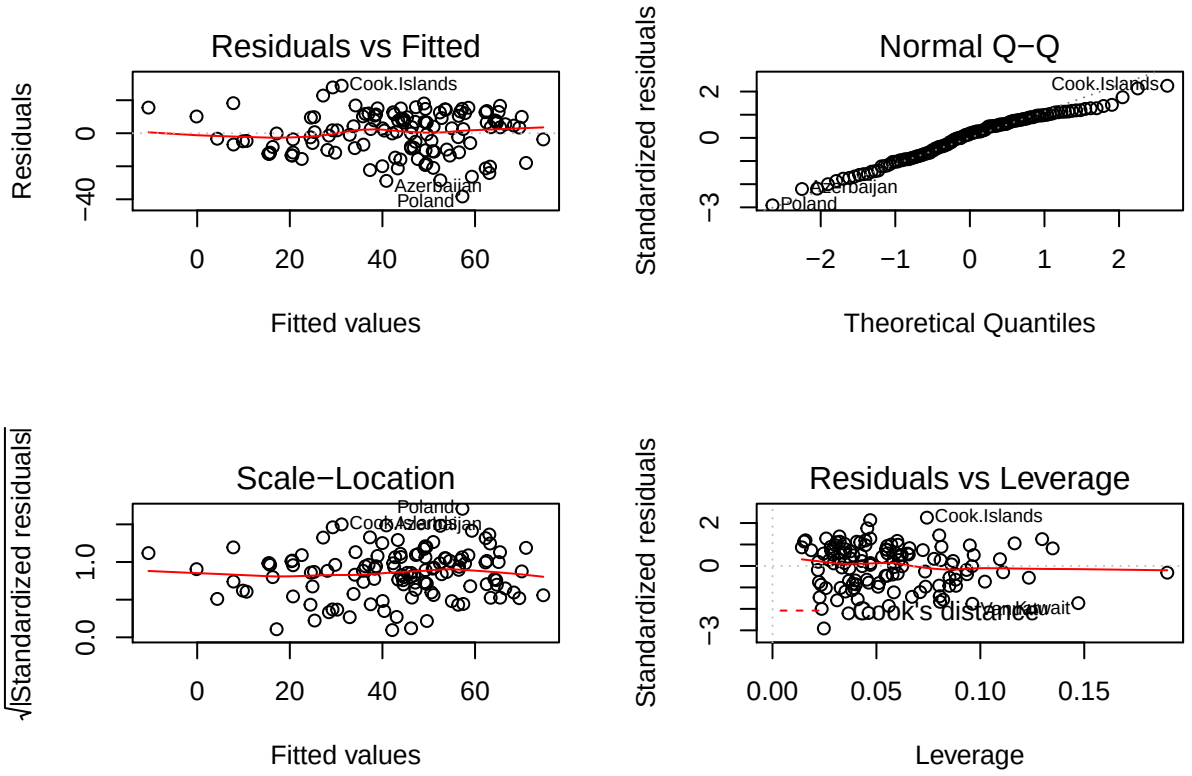
    Answer: Yes I think China and India are two influential points although I did transformation to Pop. So I removed them to check the results of Model 1. The regression results show that after removing the influential points, R-squared values does not change much and sqrt(Pop) becomes a less significant predictor in terms of p-value. But the residual plots and the added variable plots show better results, especially for Pop. So I think we should remove the points in the regression.

```
# remove outliers
ouliers=c("China","India")
UN3_3=UN3_2[which(!(rownames(UN3_2) %in% ouliers)),]
```

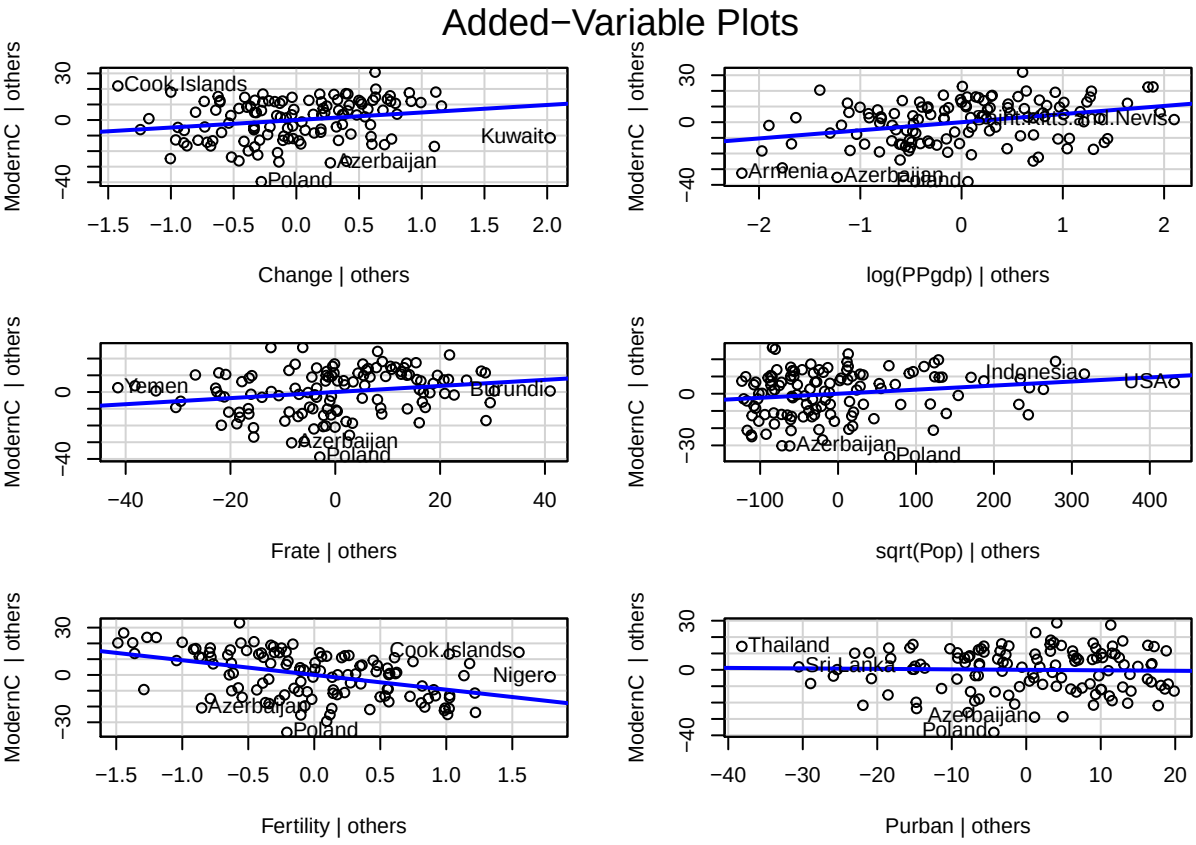
```r
model_1rm=lm(ModernC~Change+log(PPgdp)+Frate+sqrt(Pop)+Fertility+Purban,data=UN3_3)
summary(model_1rm)
```

```
##
## Call:
## lm(formula = ModernC ~ Change + log(PPgdp) + Frate + sqrt(Pop) +
##     Fertility + Purban, data = UN3_3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -38.278 -10.005   2.533  10.028  28.808
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.58599   12.32669   0.291 0.771638
## Change       4.83800    2.05078   2.359 0.019991 *
## log(PPgdp)   5.18311    1.36567   3.795 0.000236 ***
## Frate        0.18192    0.07724   2.355 0.020188 *
## sqrt(Pop)    0.02365    0.01151   2.055 0.042154 *
## Fertility   -9.31943    1.75699  -5.304 5.49e-07 ***
## Purban      -0.02949    0.09761  -0.302 0.763087
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.31 on 116 degrees of freedom
## Multiple R-squared:  0.6293, Adjusted R-squared:  0.6102
## F-statistic: 32.83 on 6 and 116 DF,  p-value: < 2.2e-16
```

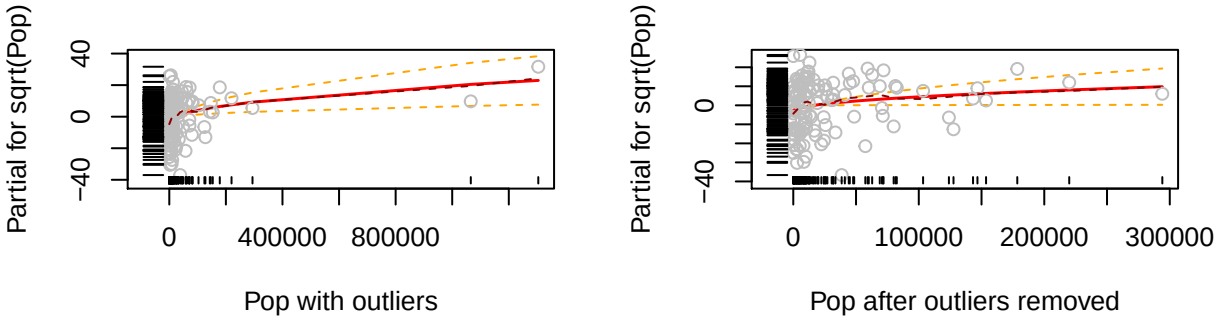
```r
par(mfrow=c(2,2))
plot(model_1rm,ask=F)
```




```
par(mfrow=c(2,2))
avPlots(model_1rm)
```



```
#check termplot
termplot(model_1,term="sqrt(Pop)",xlabs="Pop with outliers",partial.resid=T,
         se=T, rug=T,smooth=panel.smooth)
termplot(model_1rm,term="sqrt(Pop)",xlabs="Pop after outliers removed",partial.resid=T,
         se=T, rug=T,smooth=panel.smooth)
```



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!

    Answer: ModernC positively correlates to all predictors except Fertility and Purban.
    1) For predictors Change and Frate, when keep other predictors unchanged, one unit increase in the predictors would result in 4.838 and 0.182 unit increase in the response variable ModernC, respectively.
    2) For predictors Fertility and Purban, when keep other predictors unchanged, one unit increase in the predictors would result in 9.319 and 0.029 unit decrease in the response variable ModernC, respectively.
    3) For PPgdp, when keep other predictors unchanged, 10% increase in the predictor would result in $5.183 * log(1.1) = 0.494$ unit increase in the response variable ModernC
    4) For Pop, when keep other predictors unchanged, 10% unit increase in the predictor would result in $0.024 * sqrt(1.1) = 0.025$ unit increase in the response variable ModernC

```
re_table=cbind(round(confint(model_1rm),3),round(coef(model_1rm),3))
re_table=as.data.frame(re_table)
names(re_table)=c("CI 2.5%","CI 97.5%","Estimate")
kable(re_table,caption=c("95% confidence intervals and estimates for final model"))
```

Table 2: 95% confidence intervals and estimates for final model

	CI 2.5%	CI 97.5%	Estimate
(Intercept)	-20.829	28.001	3.586
Change	0.776	8.900	4.838
log(PPgdp)	2.478	7.888	5.183
Frate	0.029	0.335	0.182
sqrt(Pop)	0.001	0.046	0.024
Fertility	-12.799	-5.839	-9.319
Purban	-0.223	0.164	-0.029

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

    Answer: After removing two influential points (China and India), my best model is model 1: Mod-

ernC~Change+log(PPgdp)+Frate+sqrt(Pop)+Fertility+Purban
I chose the best model based on the following criteria:
1) Graphical EDA. I picked the model has relatively the best result from residual plots and added variable plots. That's why removing outliers are necessary because they improved the residual results.
2) Model complexity. I want an accurate model but I don't want over-fitting. And the model should be easy to explain and understand. That's why although transforming y may result in slightly better accuracy, I did not use y transformation because it would be harder to explain the model, and it's more complex, too.
3) Number of significant predictors, Since I am not dropping any predictors, I hope all the variables play a role in the model. So I prefer models in which most predictors are 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.*

$$e_y = \hat{\beta}_0 + \hat{\beta}_1 e_{x_i}$$

$$(I-H)Y = \hat{\beta}_0 + \hat{\beta}_1 (I-H)x_i$$

since OLS forβ is $(X'X)^{-1}X'Y$, here X equals to $(I-H)x_i$,and Y equals to $(I-H)Y$, so

$$(I-H)Y = 1_n'\hat{\beta}_0 + [((I-H)x_i)'((I-H)x_i)]'((I-H)x_i)'(I-H)Y(I-H)x_i$$

$$= 1_n'\hat{\beta}_0 + [x_i'(I-H)'(I-H)x_i]'x_i'(I-H)'(I-H)Y(I-H)x_i$$

since $(I-H)' = (I-H)$, and $(I-H)(I-H) = (I-H)$

$$(I-H)Y = 1_n'\hat{\beta}_0 + [x_i'(I-H)x_i]'x_i'(I-H)Y(I-H)x_i$$

since the terms $[x_i'(I-H)x_i]'$ and $x_i'(I-H)Y$ are both scalars, we can move them around in the equation. And we can multiply x_i' to both sides of the equation and get:

$$x_i'(I-H)Y = x_i'\hat{\beta}_0 + x_i'(I-H)x_i[x_i'(I-H)x_i]'x_i'(I-H)Y$$

$$x_i'(I-H)Y = x_i'\hat{\beta}_0 + x_i'(I-H)Y$$

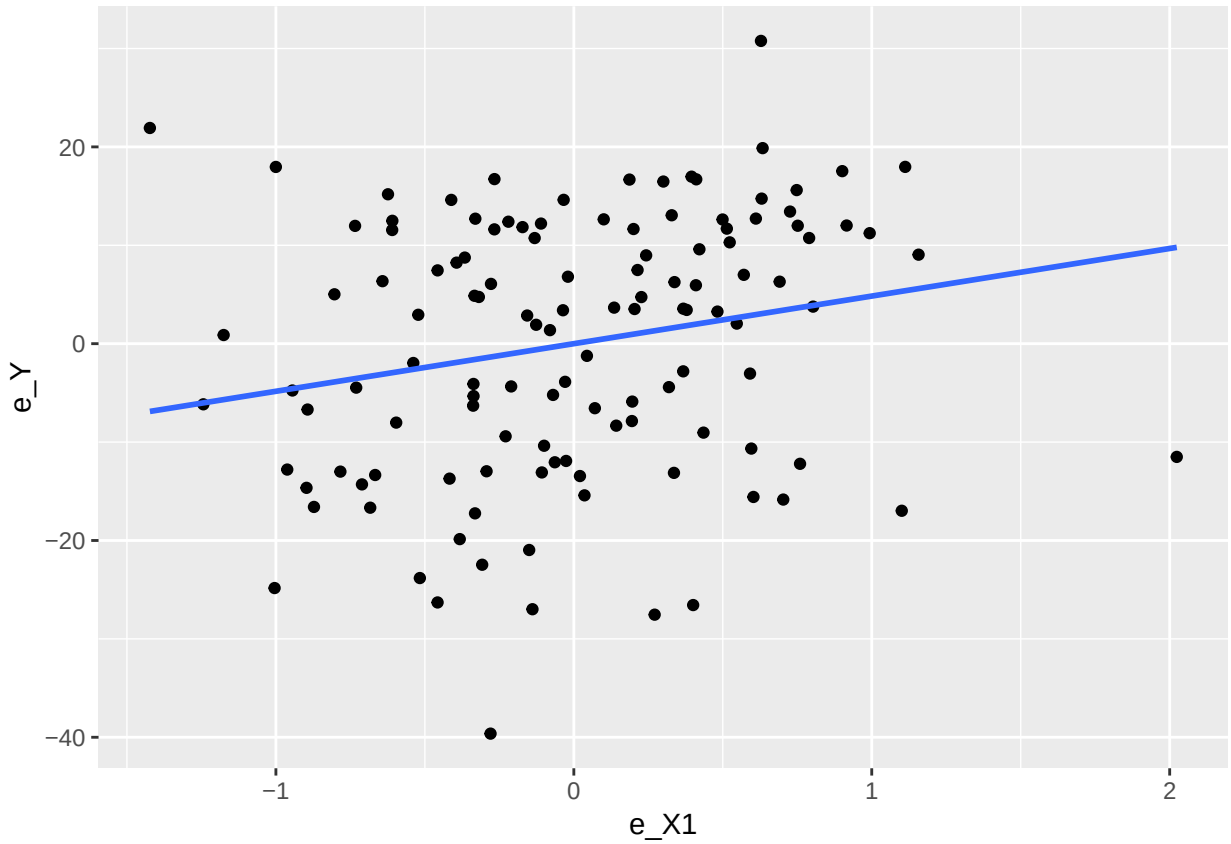
$$0 = x_i'\hat{\beta}_0$$

The above equation is valid for any random x_i, so $\hat{\beta}_0$ has to be 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.

Answer: Regress "ModernC" on all predictors except "Change" and obtain the residuals; call them e_Y. Regress "Change" on the other predictors and obtain the residuals (call e_{X1}). Then I regress e_Y on e_X. The results show that the slope for e_Y on e_X is the same as the estimated slope for "Change" in the full model with all the predictors.

```
e_Y = residuals(lm(ModernC~log(PPgdp)+Frate+sqrt(Pop)+Fertility+Purban, data=UN3_3))
e_X1 = residuals(lm(Change ~ log(PPgdp)+Frate+sqrt(Pop)+Fertility+Purban, data=UN3_3))
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)
```




```
summary(model_1rm)$coef
```

```
##                Estimate  Std. Error    t value     Pr(>|t|)
## (Intercept)  3.58598682 12.32669326  0.2909123 7.716379e-01
## Change       4.83799970  2.05078207  2.3590999 1.999099e-02
## log(PPgdp)   5.18310553  1.36567230  3.7952776 2.360648e-04
## Frate        0.18191944  0.07723895  2.3552812 2.018788e-02
## sqrt(Pop)    0.02365330  0.01151180  2.0547000 4.215428e-02
## Fertility   -9.31942717  1.75698682 -5.3042101 5.488885e-07
## Purban      -0.02949308  0.09761402 -0.3021398 7.630865e-01
```

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

```
##                  Estimate Std. Error       t value   Pr(>|t|)
## (Intercept) -6.406748e-16   1.174859 -5.453205e-16 1.00000000
## e_X1         4.838000e+00   2.007964  2.409406e+00 0.01748538
```