HW2 STA521 Fall18

[Yangfan Ren netID: yr47 git: renyangfan960902]

September 22, 2018

Backgound Reading

Readings: Chapters 3-4 in Weisberg Applied Linear Regression

Exploratory Data Analysis

```
options(repos="https://cran.rstudio.com" )
library(alr3)
data(UN3, package="alr3")
help(UN3)
library(car)
library(knitr)
library(dplyr)
install.packages("kableExtra")
```

```
##
## The downloaded binary packages are in
##  /var/folders/8n/4z61k9j50qzc4sf3wjs52y3w0000gn/T//RtmpEqc5IN/downloaded_packages
```

```
library(kableExtra)
library(GGally)
library(tibble)
```

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                      
```

```
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 ...
```

There are 6 variables have missing values. All of the variables are quantitative.

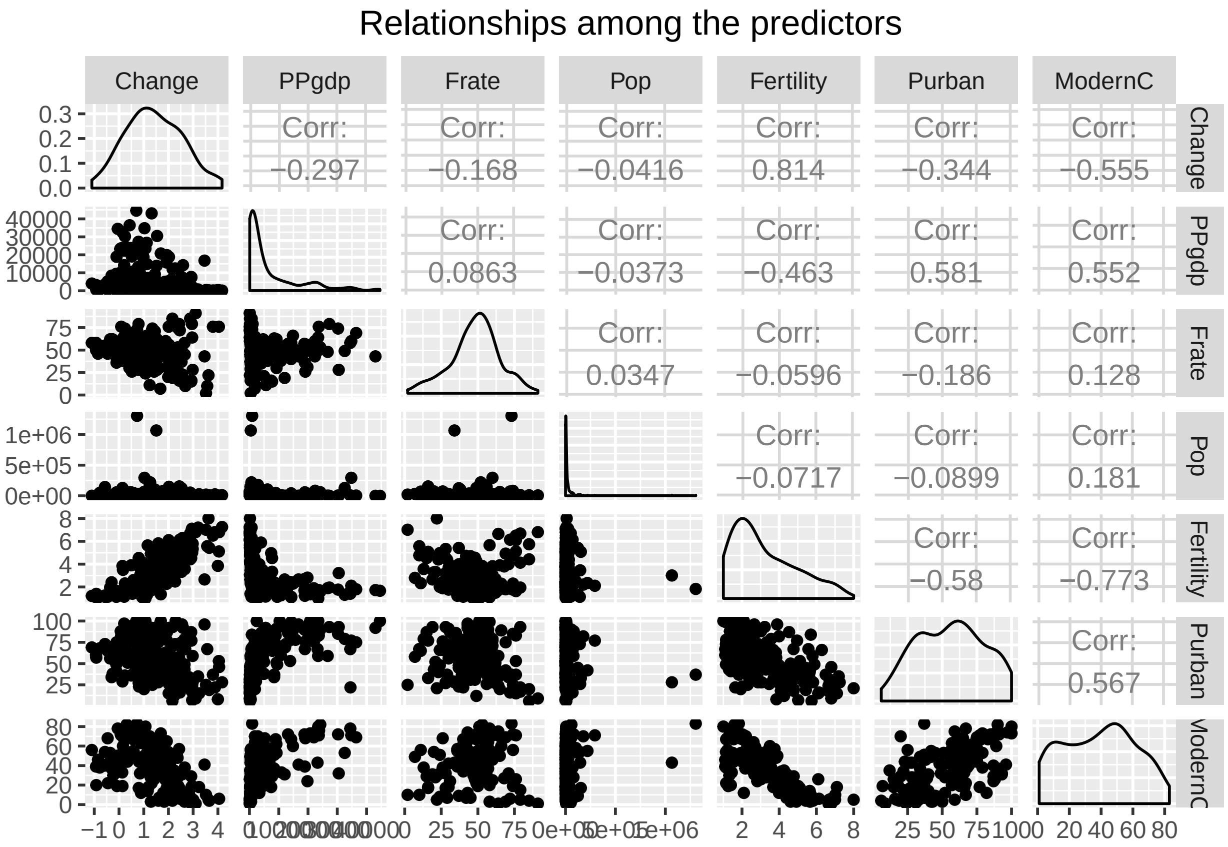
2. What is the mean and standard deviation of each quantitative predictor? Provide in a nicely formatted table.

```
mean<-lapply(na.omit(UN3[,1:7]),mean)
sd<-lapply(na.omit(UN3[,1:7]),sd)
m_sd<-t(as.data.frame(rbind(mean=mean,sd=sd)))
m_sd%>%
  kable() %>%
  kable_styling(bootstrap_options = "striped", full_width = F,position = "left")
```

	mean	sd
ModernC	43.272	21.4424872173103
Change	1.18168	1.06493126841671
PPgdp	6612.608	9214.77121685437
Frate	48.112	16.9044830854746
Pop	46060.504	153631.211314189
Fertility	2.87624	1.51694762705434
Purban	56.976	22.5912001137754

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?

```
ggpairs(UN3,columns=c(2:7,1),title="Relationships among the predictors")+theme(plot.title = element_tex
```

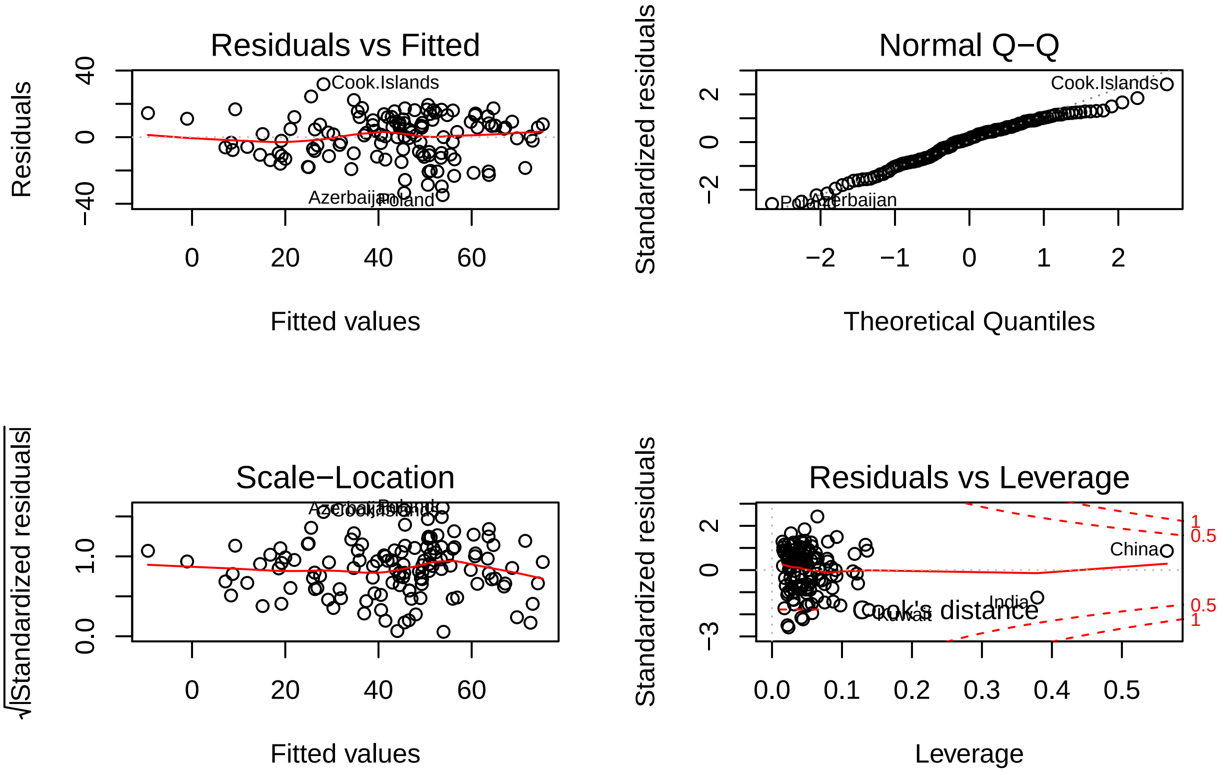


From the plots above, we could see that there are relatively clear linear relationships between ModernC and Change, Fertility, Purban respectively. The relationship between ModernC and PPgdp is kind of non-linear. And there are two obvious outliers in the plots of ModernC and Pop. And there is no specific pattern for ModernC~Frate. Therefore, some transformations seem to be needed for these three predictors.

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?

```
UN3.na.rm<-na.omit(UN3)
model1<-lm(ModernC~.,UN3.na.rm)
par(mfrow=c(2,2))
plot(model1)
```



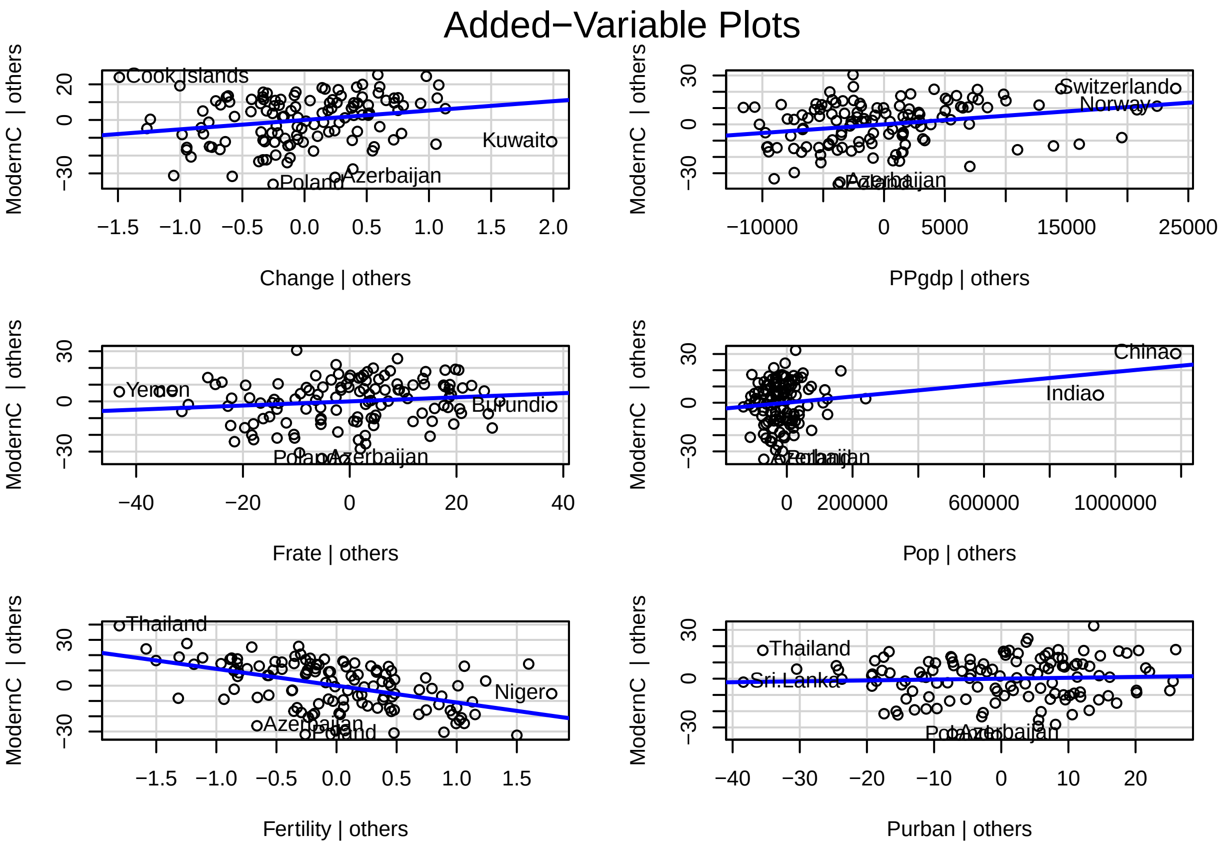
```r
summary(model1)
```

```
##
## Call:
## lm(formula = ModernC ~ ., data = UN3.na.rm)
##
## 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
```

The residual-fitted plot shows an almost horizontal line which indicates that we can not reject a linear relationship. And the scale-location plot shows that the residuals have constant variance. But the QQ plot tells us that the residuals are not strictlly normal distributed. Although there are some influential points like China, India and Kuwait, but they are not siginificant enough which indicates that there are no outlier that exceeds three standard deviations and no high leverage point. There are 125 observations used in the model.

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?

```
car::avPlots(model1)
```



From the avplots, we can clearly see that “Pop” needs transformation as there are two influential localities, India and China, which are far from other localities. All localities have population around or less than 200000 except that India and China have more than 800000. So transformation as log would be helpful.

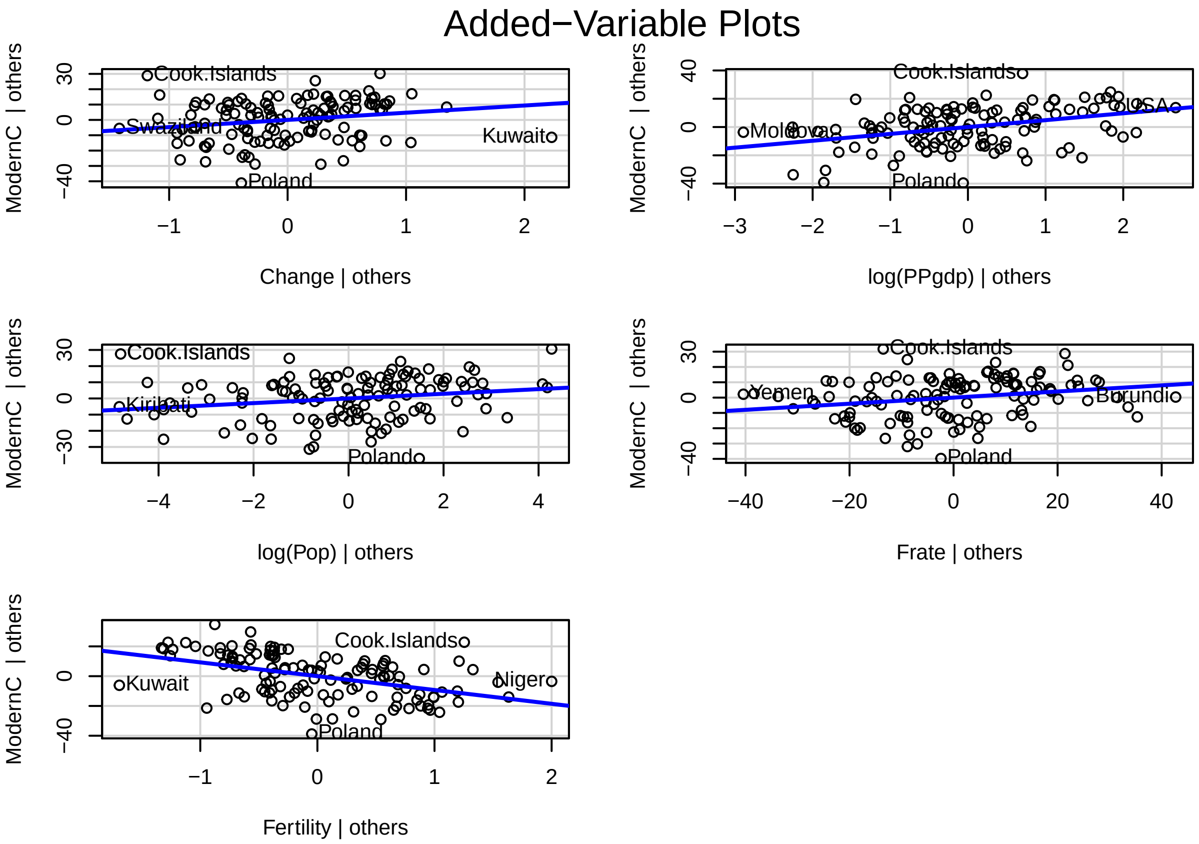
Besides, kuwait could be influential for “Change” since without it ,the slope of linear model might be a little larger.

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.na.rm["Change"]<-UN3.na.rm["Change"]+1.2
car::boxTidwell(ModernC~PPgdp+Pop,~Change+Fertility+Frate,data=UN3.na.rm)
```

```
##       MLE of lambda Score Statistic (z) Pr(>|z|)
## PPgdp      -0.13620             -1.2872   0.1980
## Pop         0.42227             -0.7941   0.4271
##
## iterations =  4
```

```
model2<-lm(ModernC~Change+log(PPgdp)+log(Pop)+Frate+Fertility,UN3.na.rm)
car::avPlots(model2)
```



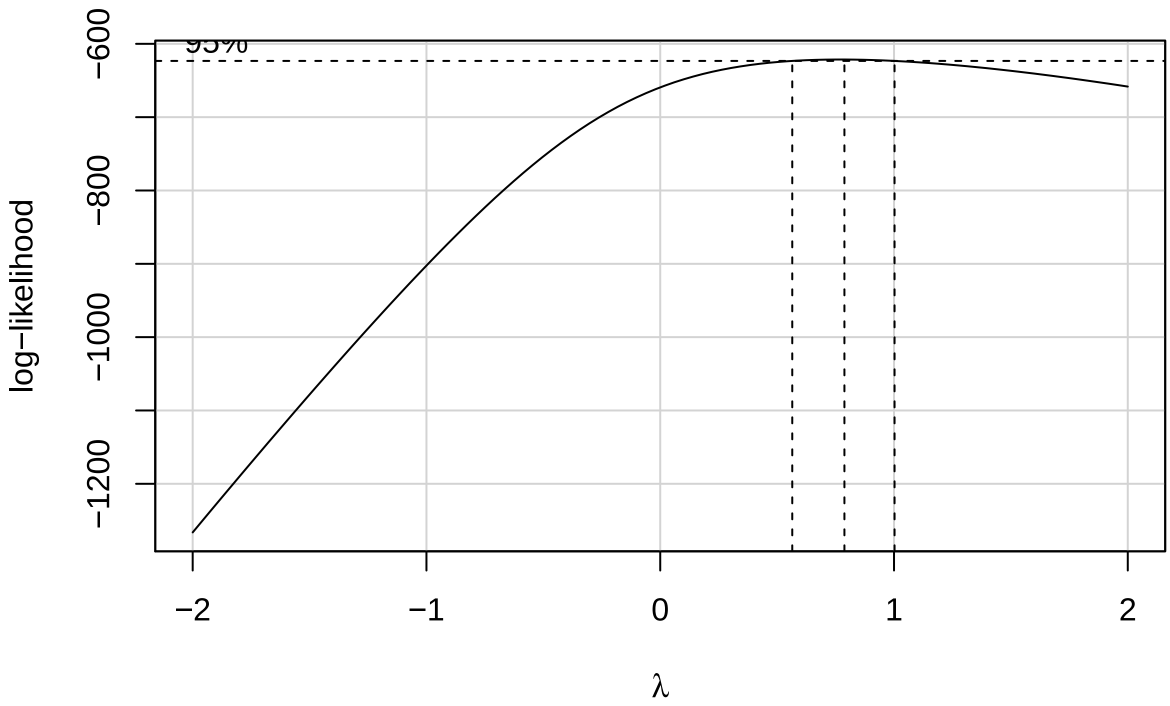
From the avplots, we can see that "Purban" has little influence on "ModernC". So I decide to exclude "Purban" in the next model. And I also exclude China and India due to their influences of "Pop".
Since the regression of "PPgdp" to "Change" is not optimal, I include "PPgdp" in transformation.

From the results of boxTidwell, the MLE of λ of "PPgdp" and "Pop" are -0.14 and 0.42 respectively. The λ of "PPgdp" and "Pop" are small that I choose to regard it as 0 which is more interpretable. Therefore we use transformations of log(PPgdp) and log(Pop).

Although the boxTidwell shows that the transformations of both "PPgdp" and "Pop" are not significant, from the avplots of model after transformation we can see better linear relationships of "PPgdP", "Pop" and "ModernC". So I decide to keep the transformations.

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

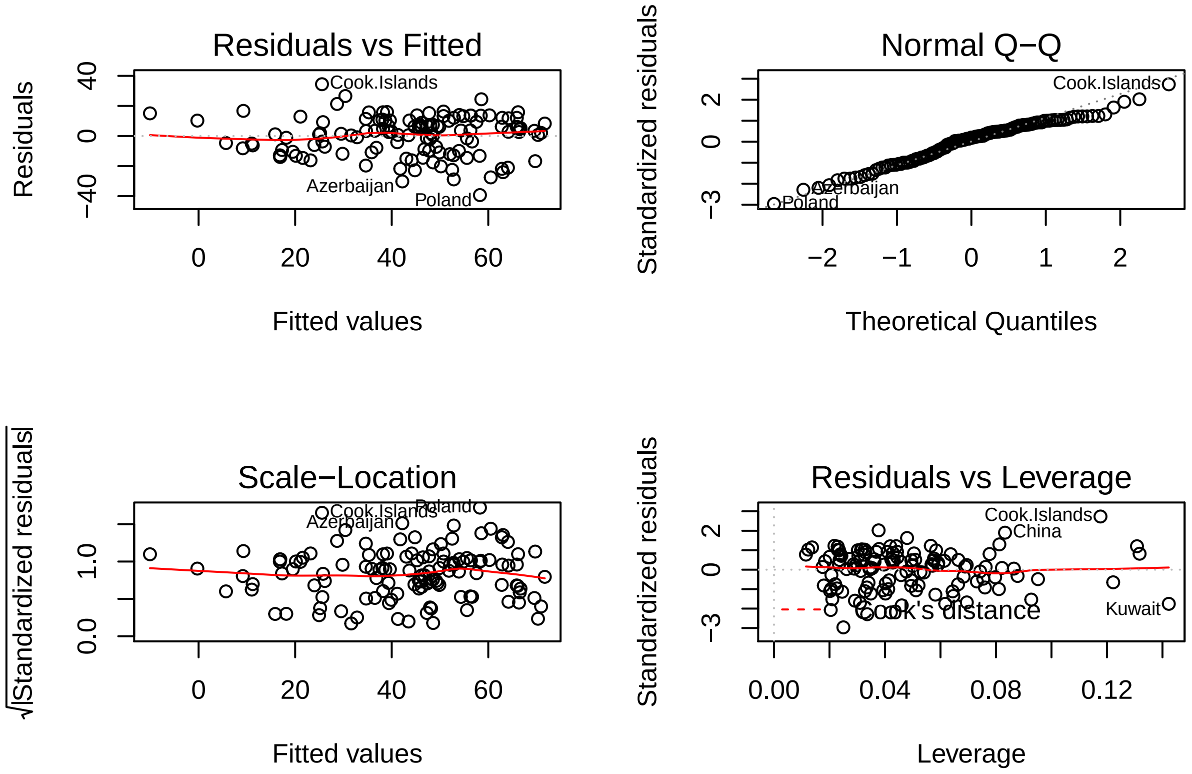
```
car::boxCox(model2)
```



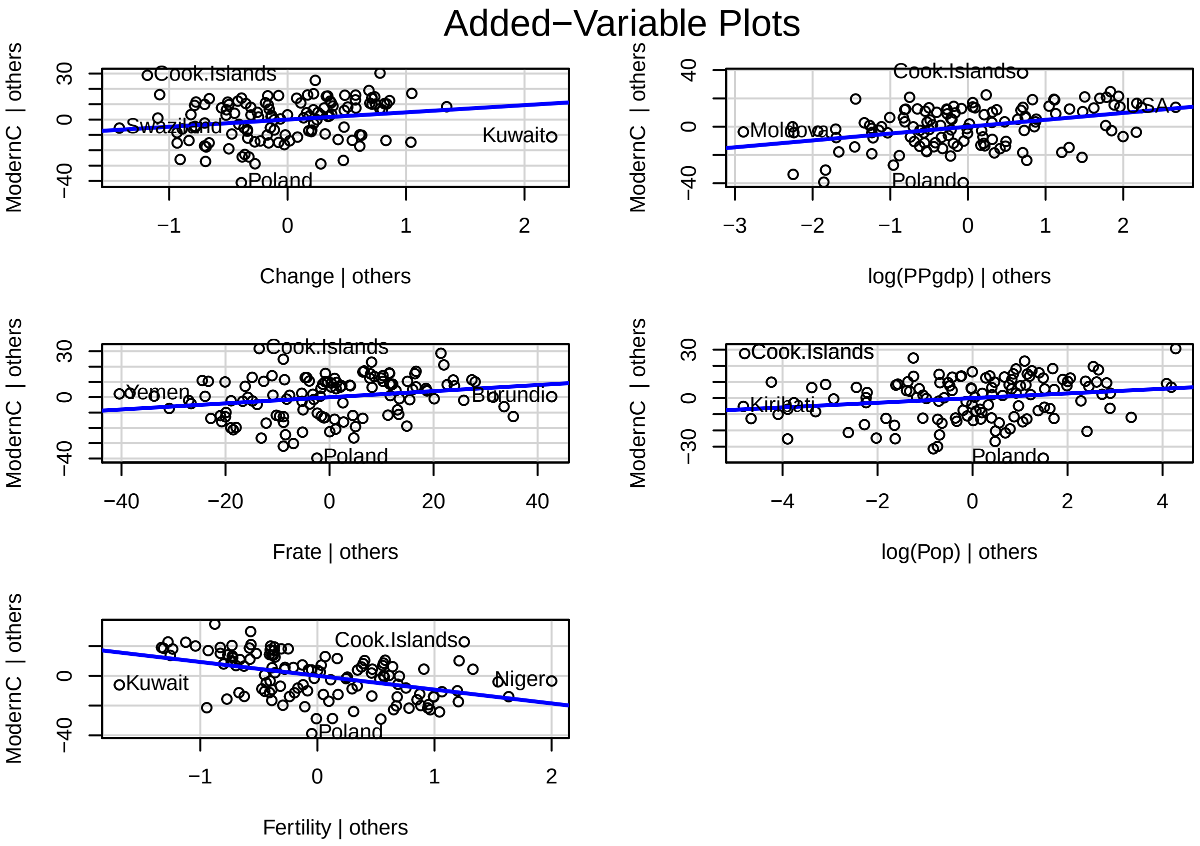
The λ is close to 0.8, I choose not to do any transformation for the reason of interpretability.

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.

```
model2<-lm(ModernC~Change+log(PPgdp)+Frate+log(Pop)+Fertility,UN3.na.rm)
par(mfrow=c(2,2))
plot(model2)
```



```
car::avPlots(model2)
```

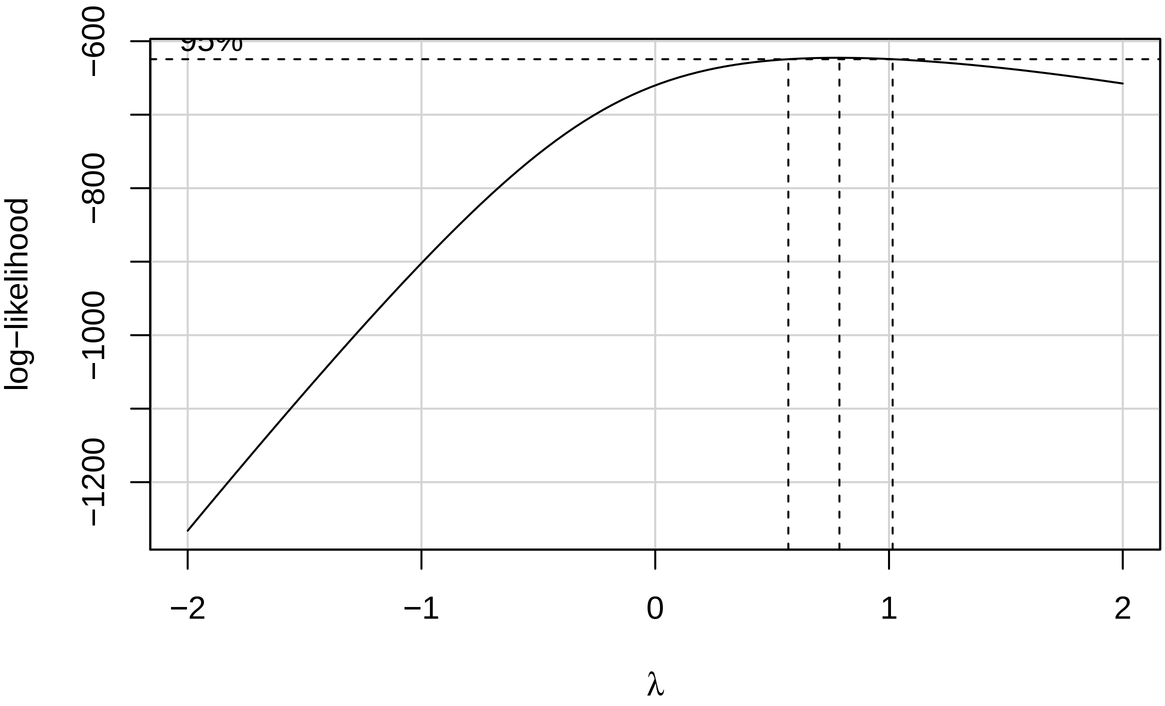


From the diagnostic plots, we can see that the cook's distance of China and India is smaller which indicates less influence. The QQ plot shows that it is still a little skewed, but it does have some improvement comparing to model1.

The added variable plots are all fit well. The avplot of "Pop" has significant improvement.

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?

```
car::boxCox(model1)
```



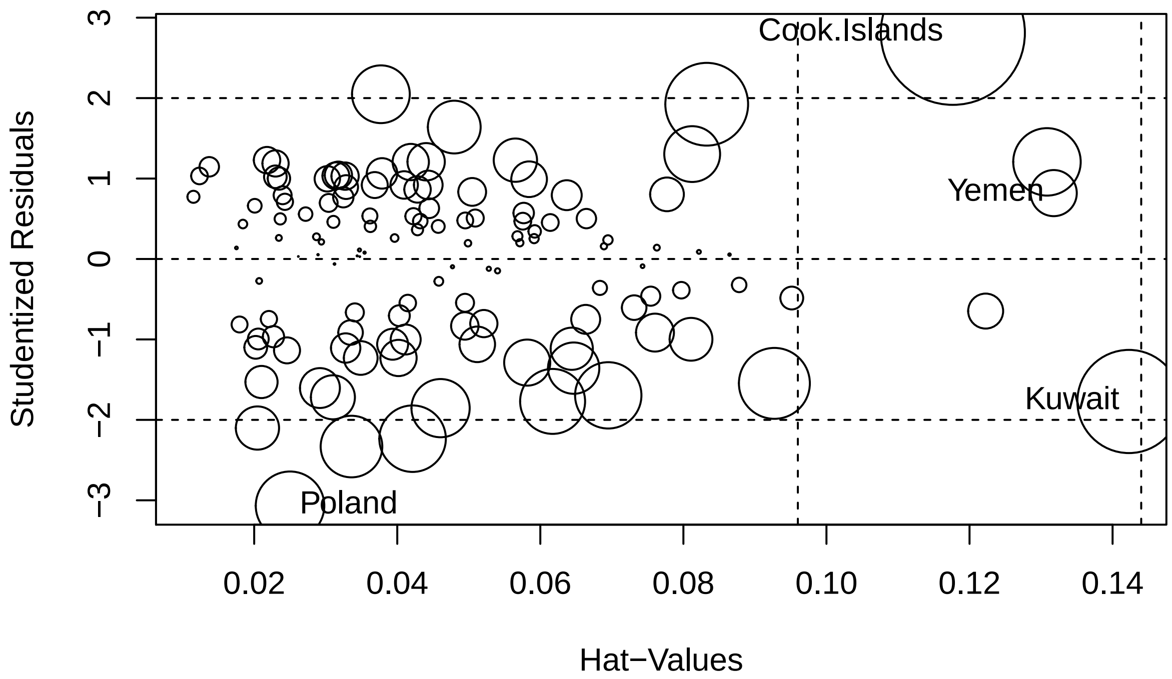
Since 1 is in the 95% confidence interval of λ which indicates no transformation of response, I end up with the same model as in 8.

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.

```
pval<- 2*(1- pt(abs(rstudent(model2)), model2$df - 1))
p<-as.data.frame(pval<0.05/nrow(UN3.na.rm))
UN3.na.rm$p<-p
rownames(UN3.na.rm[p==TRUE])
```

```
## NULL
```

```
influencePlot(model2)
```



```
##                  StudRes        Hat      CookD
## Cook.Islands  2.8111289 0.11766863 0.16601679
## Kuwait       -1.7715300 0.14229068 0.08524080
## Poland       -3.0677469 0.02501099 0.03758018
## Yemen         0.8190358 0.13177110 0.01701545
```

Using Bonferroni Correction, we can see that there is no outlier in the model.

From "influencePlot" function, we can know that although Cook.Islands, Kuwait, Poland and Yeman have somewhat high influence on the regression model, the cook's distance of them are either larger than 1 nor larger than 4/n. Therefore, there is no influential observation.

As a result, we don't need to refit the 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!

```
UN3.na.rm["Change"]=UN3.na.rm["Change"]-1.2
model3<-lm(ModernC~Change+log(PPgdp)+Frate+log(Pop)+Fertility,UN3.na.rm)
coef<-as.data.frame(confint(model3))
coef["log(PPgdp)",]<-exp(coef["log(PPgdp)",])
coef["log(Pop)",]<-exp(coef["log(Pop)",])
rownames(coef)<-c("Intercept",colnames(UN3[c(-1,-7)]))
coef%>%
  kable() %>%
  kable_styling(bootstrap_options = "striped", full_width = F,position = "left")
```

	2.5 %	97.5 %
Intercept	-24.5689501	32.7731131
Change	0.6727287	8.7227854
PPgdp	15.1291534	1098.9398923
Frate	0.0496977	0.3493943
Pop	1.2233091	14.5980723
Fertility	-12.5950756	-5.9617672

As “Change” had a transformation as adding 1.2, we need to transform “Change” into the row data. So we get model3.
From the table of 95% confidence interval of coefficients for model3, we can know that we have 95% confident that “ModernC would be between -24.56 and 32.77 when other predictors are 0.
Keeping others fixed, with a one unit change in”Change“, we have 95% confident that”Modernc" would change by 0.67 to 8.72.
Keeping others fixed, with a one unit change in “PPgdp”, we have 95% confident that “Modernc” would change by 15.12 to 1098.94.
Keeping others fixed, with a one unit change in “Frate”, we have 95% confident that “Modernc” would change by 0.05 to 0.35.
Keeping others fixed, with a one unit change in “Pop”, we have 95% confident that “Modernc” would change by 1.22 to 14.60.
Keeping others fixed, with a one unit change in “Fertility”, we have 95% confident that “Modernc” would change by -12.60 to -5.96.

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

```r
summary(model3)
```

```
##
## Call:
## lm(formula = ModernC ~ Change + log(PPgdp) + Frate + log(Pop) +
##     Fertility, data = UN3.na.rm)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -39.276  -9.928   2.572  10.253  34.442
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.10208   14.47959   0.283  0.77744
## Change       4.69776    2.03274   2.311  0.02255 *
## log(PPgdp)   4.85936    1.08214   4.491 1.65e-05 ***
## Frate        0.19955    0.07568   2.637  0.00949 **
## log(Pop)     1.44122    0.62606   2.302  0.02307 *
## Fertility   -9.27842    1.67499  -5.539 1.85e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.42 on 119 degrees of freedom
## Multiple R-squared:  0.6243, Adjusted R-squared:  0.6085
## F-statistic: 39.55 on 5 and 119 DF,  p-value: < 2.2e-16
```

I end up with a model as:

$$ModernC = 4.1021+4.6978(Change)+4.8594(log(PPgdp))+0.1996(Frate)+1.4412(log(Pop))-9.2784(Fertility)$$

As the model shows, we could say that the faster the population grows and the larger the population is, the more likely for unmariied women in a country to use modern method of contraception.
At the same time, the larger the per capital GDP is and the larger the percent of females over age 15 economically active is, the more likely for unmariied women in a country to use modern method of contraception.
However, the larger the expected number of live births per female is, the less likely for unmariied women in a country to use modern method of contraception.
We excluded 85 observations due to missing values. But we did not delete any outlier or influential observation.

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.

Assuming that our model is $y=\beta_0+\beta_1x_1$.
For added variable plots, we regress the residuals of model without certain variable on residuals of regressing the certain variable on others. Here we use x_1 as an example.
Since $e_i=Y_i-\hat{Y}_i$, we can know that $e_{(Y)}=(I-H)Y$ and $e_{(X_3)}=(I-H)X_3$ as $H=X(X^TX)^{-1}X^T$ and $\beta_1=(X^TX)^{-1}X^TY$.
Then we can get that

$$
\begin{aligned}
(1-H)Y &= \hat{\beta}_0I+\hat{\beta}_1(I-H)X_1 \\
(1-H)Y &= \hat{\beta}_0I+[X_1^T(I-H)^T(I-H)X_1]^{-1}((I-H)X_1)^T(I-H)Y(I-H)X_1 \\
(1-H)Y &= \hat{\beta}_0I+(X_1^T(I-H)X_1)^{-1}X_1^T(I-H)Y(I-H)X_1
\end{aligned}
$$

Since $(X_1^T(I-H)X_1)^{-1}$ and $X_1^T(I-H)Y$ are scalars, we can change the positions of them. We multiply X_1^T on both sides, then

$$X_1^T(1-H)Y=X_1^T\hat{\beta}_0I+X_1(I-H)X_1^T(X_1^T(I-H)X_1)^{-1}X_1^T(I-H)YX_1^T\hat{\beta}_0I=0$$

So $\sum_{i=1}^n x_1^{(i)}\hat{\beta}_0=0$, $\hat{\beta}_0=0$ as $x_3^{(i)}$ can not be all 0. So the intercept in the added variable scatter plot will always be zero.

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.

```
e_Y <- residuals(lm(ModernC ~ Change + log(PPgdp) + log(Pop) + Frate, data=UN3.na.rm))
UN3.na.rm["Change"]<-UN3.na.rm["Change"]+1.2
e_X <- residuals(lm(Fertility ~ Change + log(PPgdp) + log(Pop) + Frate, data=UN3.na.rm))
res<- data.frame(e_Y, e_X)
av <- lm(e_Y ~ e_X, data=res)
av$coef
```

```
##   (Intercept)           e_X
##  1.229701e-15 -9.278421e+00
```

```
model3$coef
```

```
## (Intercept)      Change  log(PPgdp)       Frate    log(Pop)   Fertility
##    4.102082    4.697757    4.859362    0.199546    1.441225   -9.278421
```

As we can see, the coefficient of e_X is exactly the same as coefficient of “Fertility” in the model in Ex10. Both of them are -9.2784.