HW2 STA521

[Weijie Yi, wy65]

Due September 14, 2019 10am

Background Reading

Readings: Chapters 3-4, 8-9 and Appendix in Weisberg Applied Linear Regression

This exercise involves the UN data set from `alr3` package. Install `alr3` and the `car` packages and load the data to answer the following questions adding your code in the code chunks. Please add appropriate code to the chunks to suppress messages and warnings as needed once you are sure the code is working properly and remove instructions if no longer needed. Figures should have informative captions. Please switch the output to pdf for your final version to upload to Sakai. **Remove these instructions for final submission**

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*

```
library(alr3)
```

```
## Loading required package: car
```

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

```
data(UN3, package="alr3")
help(UN3)
library(car)
```

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

Answer: Six variables have missing data, and all variables are quantitative.

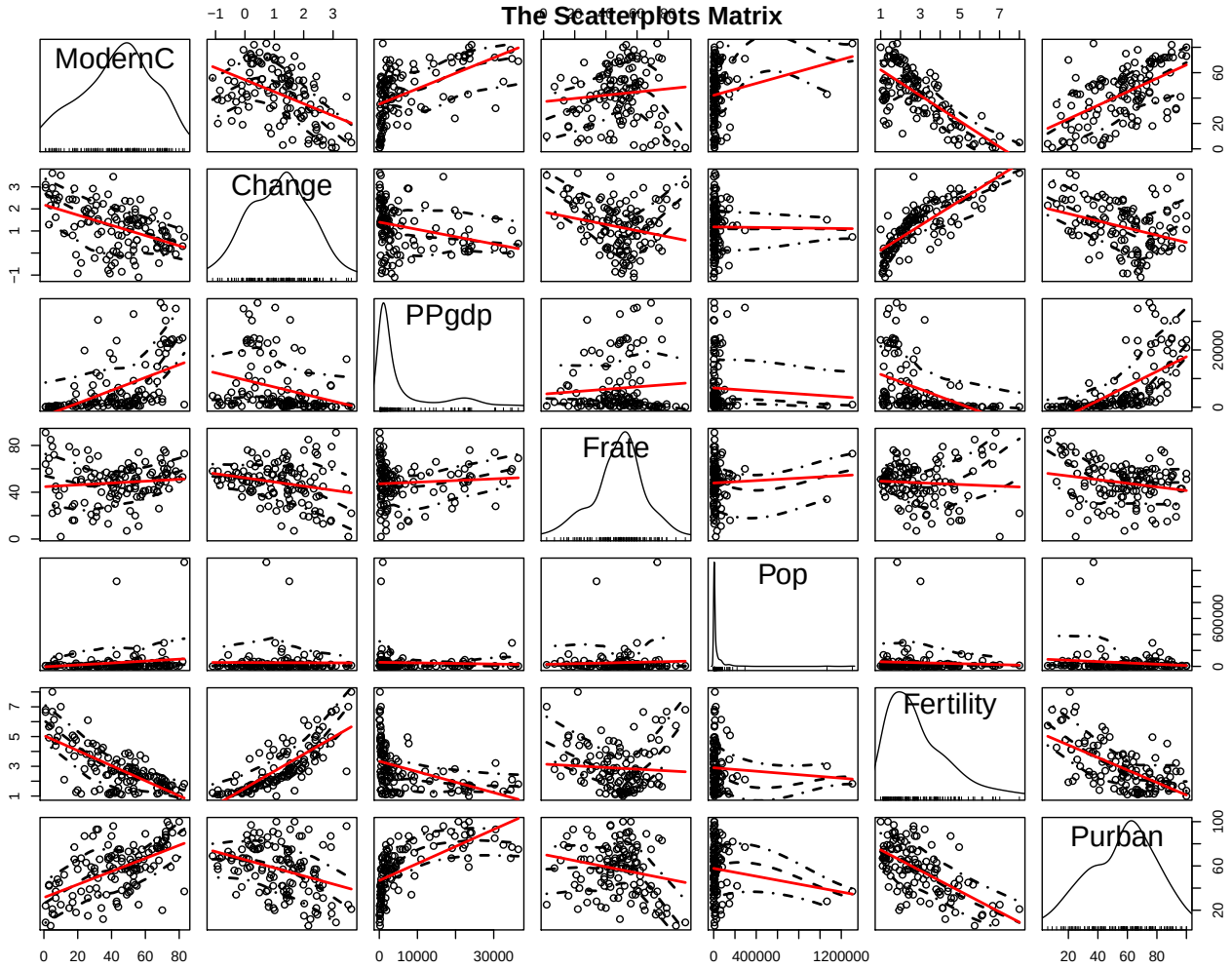
```
summary(UN3)
```

```
##     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
```

2. 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: Based on the scatterplots matrix, the predictors other than `Pop` are related to ModernC.The predictor `Pop` may need to transform. There are some outliers, especially in predictor `Pop`. There seems to be a nonlinear relationship between the predictors `PPgdp`, `Frate` and `ModernC`.

```
car::scatterplotMatrix(UN3, col = 1, regLine = list(method=lm, col = 2))
title(main = 'The Scatterplots Matrix')
```

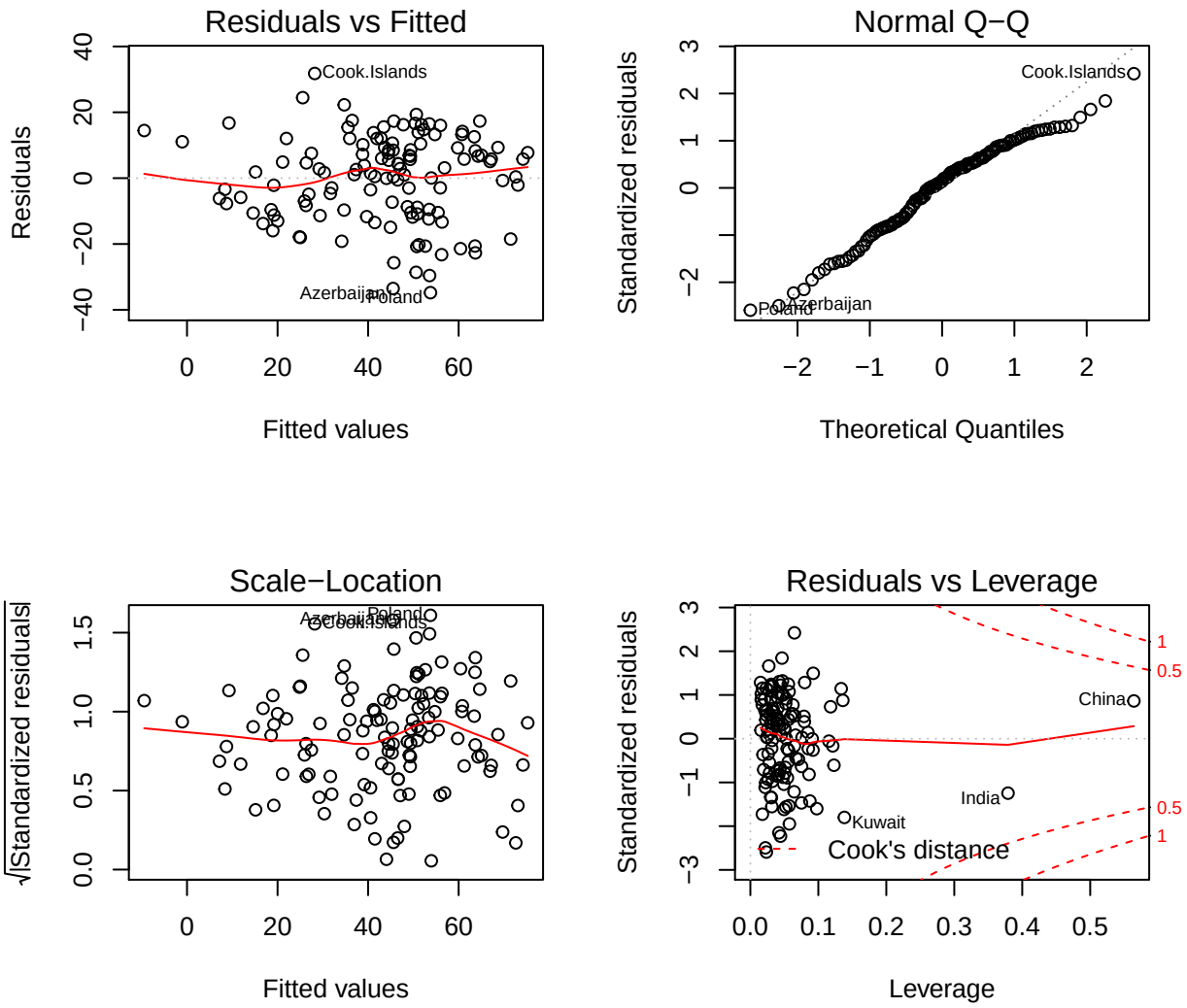


Model Fitting

3. 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: Based on the diagnostic residual plot, all assumptions of the linear model are satisfied. 125 observations are used in my model fitting.

```
fit0 = lm(ModernC ~ ., UN3)
par(mfrow = c(2, 2))
plot(fit0)
```



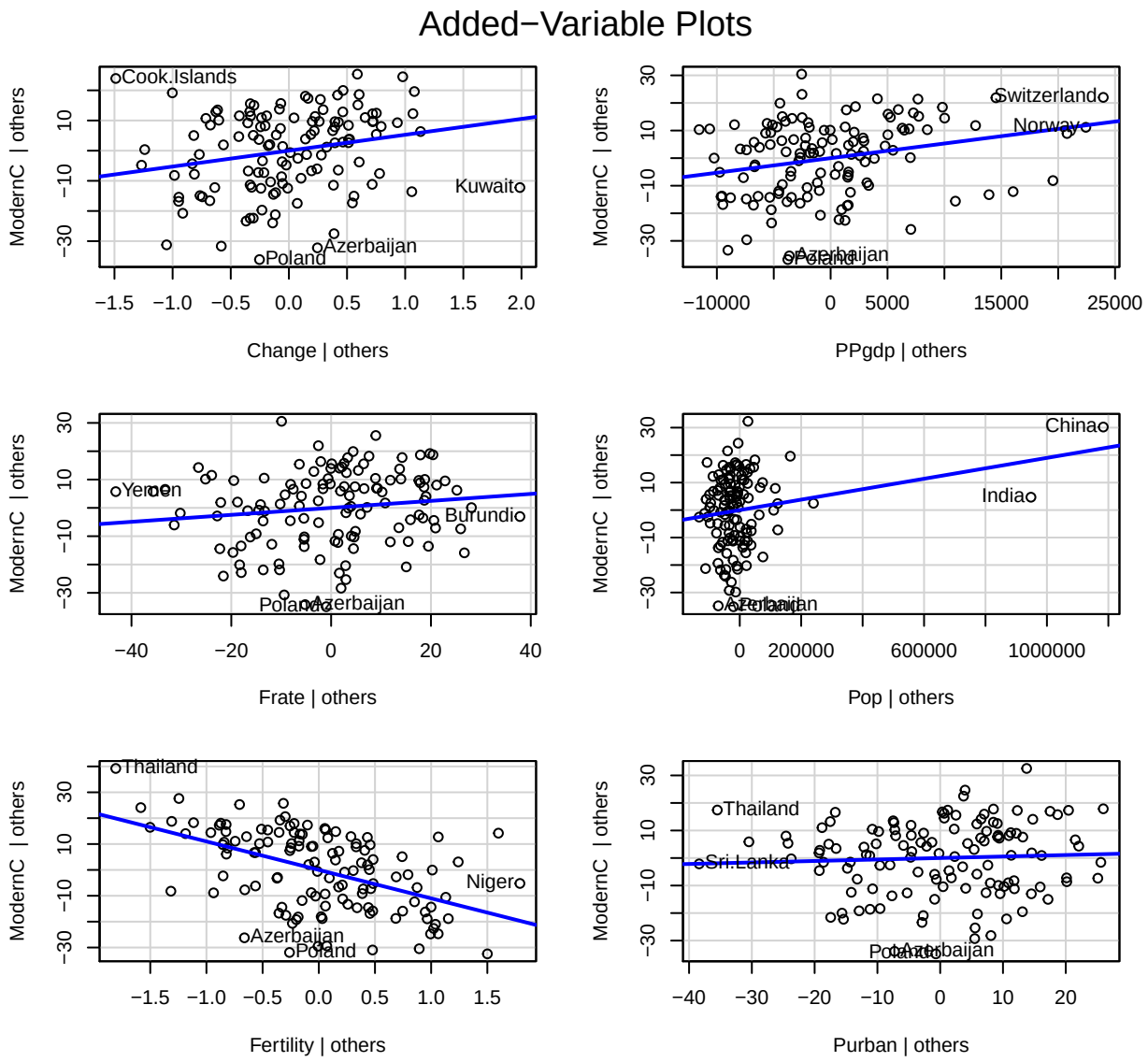

```
nobs(fit0)
```

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

4. 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: The added variable plots suggest that a transformations are needed for predictor `Pop` in the model. The localities `Cook.Islands` and `Kuwait` are influential for predictor `Change`. The localities `Switzerlando` and `Norway` are influential for predictor `GGgdp`. The localities `Yemen` and `Burundio` are influential for predictor `Frate`. The localities `China` and `India` are influential for predictor `Pop`. The localities `Thailand` and `Nigero` are influential for predictor `Fertility`. The localities `Thailand` and `SriLanka` are influential for predictor `Purban`.

```
car::avPlots(fit0)
```



5. Using the multivariate BoxCox `car::powerTransform` find appropriate transformations of the response and predictor variables for the linear model. If any predictors are negative, you may need to transform so that they are non-negative. Summarize the resulting transformations.

Answer: According to the table below, we can see that Pop, Fertility and PPgdp need to be transformed. We can use log transformation for them.

```
pt = car::powerTransform(UN3, family='bcnPower')$lambda
names(pt) = names(UN3)
knitr::kable(t(round(pt, 2)))
```

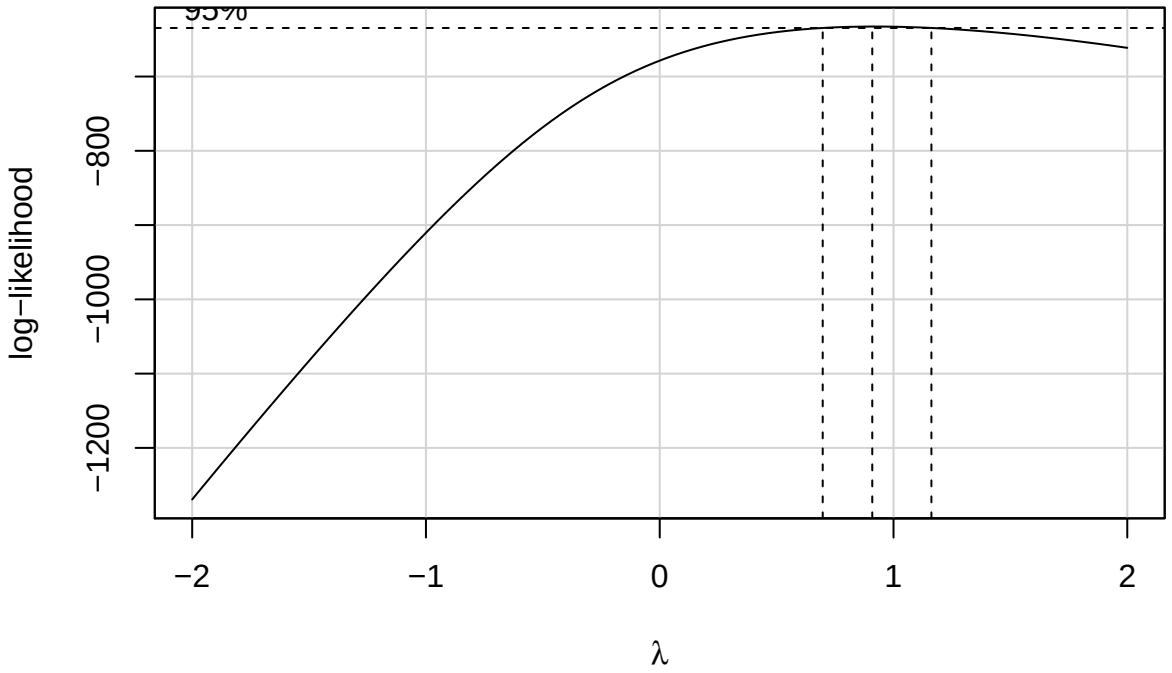
ModernC	Change	PPgdp	Frate	Pop	Fertility	Purban
1.63	0.23	-0.14	0.93	0.06	0.16	1.09

6. Given the selected transformations of the predictors, verify the transformation of the response using `MASS::boxcox` or `car::boxCox` and justify. Do you get the same transformation if you used

`car::powerTransform` above? Do you get the same transformation for the response if you do not transform any of the predictors? Discuss briefly the findings.

Answer: Given the selected transformations of the predictors, the two transformations for the response are roughly the same.

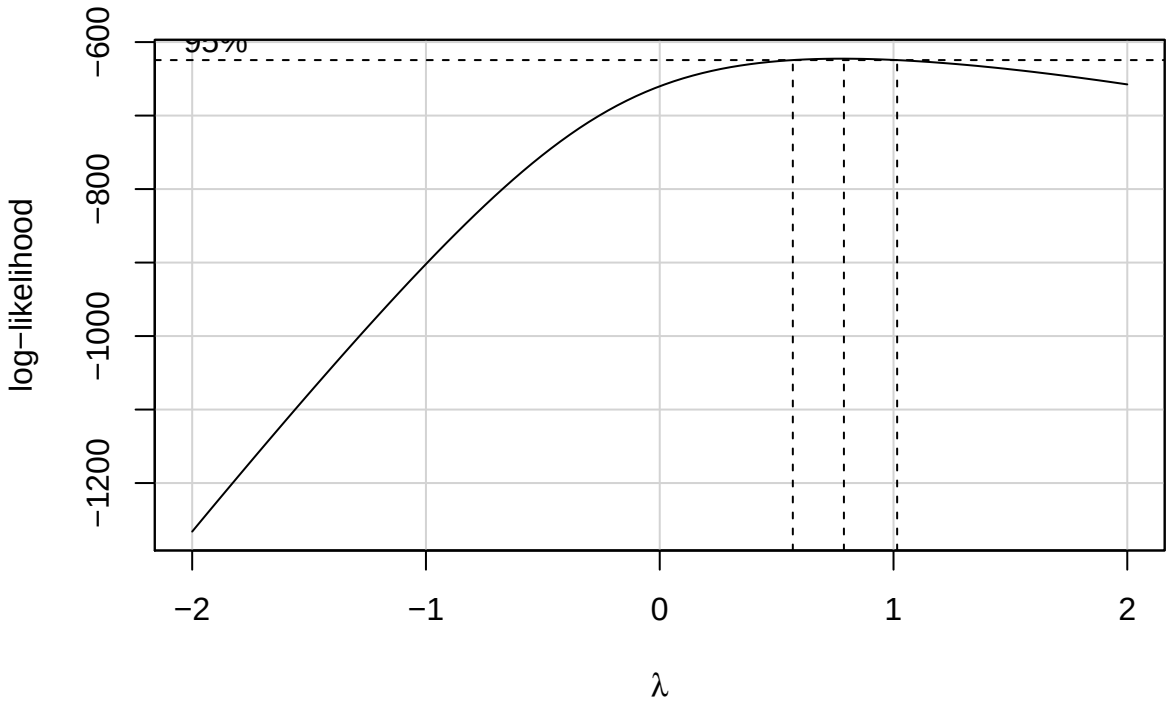
```
fit1 = lm(ModernC ~ Change + log(PPgdp) + Frate + log(Pop) + log(Fertility) + Purban, UN3)
bc1 = car::boxCox(fit1, grid = TRUE)
```




```
lam1 = c(boxcox=bc1$x[which.max(bc1$y)],
         powerTransform=car::powerTransform(fit1)$lambda[[1]])
knitr::kable(t(round(lam1, 4)))
```

boxcox	powerTransform
0.9091	0.9184

```
bc0 = car::boxCox(fit0, grid = TRUE)
```




```
lam0 = c(boxcox=bc0$x[which.max(bc0$y)],
         powerTransform=car::powerTransform(fit0)$lambda[[1]])
knitr::kable(t(round(lam0, 4)))
```

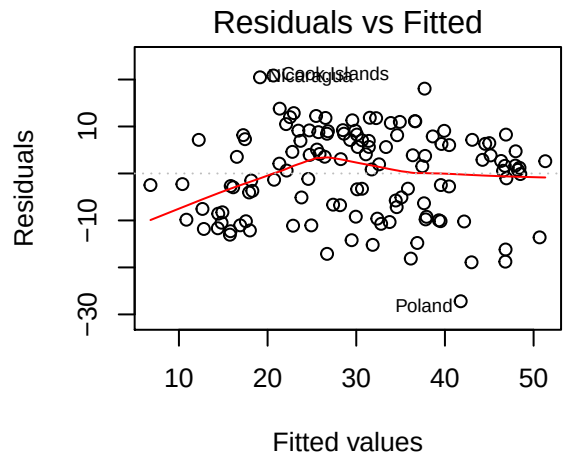
boxcox	powerTransform
0.7879	0.779

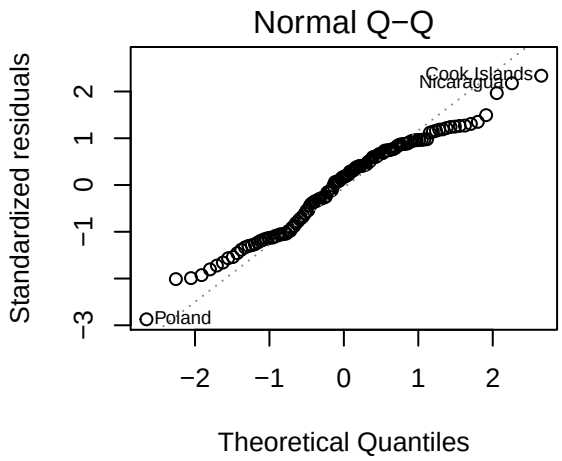
Answer: If we do not transform the predictors, the two transformations (boxcox and powerTransform) for the response are also roughly the same.

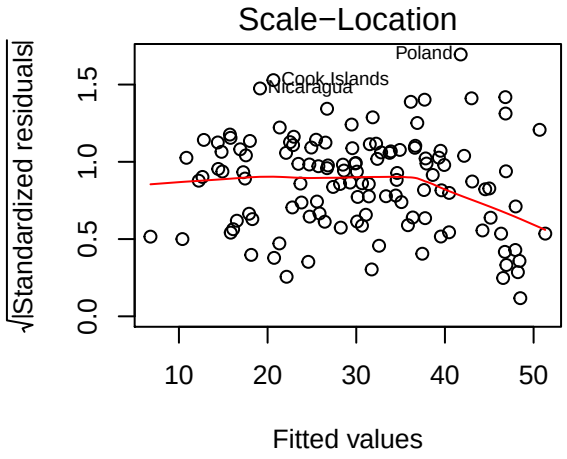
7. Fit the regression using the transformed variables. Provide residual plots and added variables plots and comment. If you feel that you need additional transformations of either the response or predictors, repeat any steps until you feel satisfied with the model and residuals.

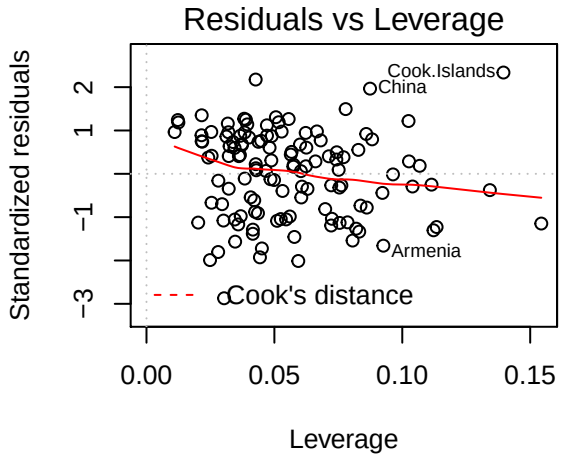
Answer: Based on the diagnostic residual plot, all assumptions of the linear model are satisfied.

```
fit = lm(ModernC^0.91 ~ Change + log(PPgdp) + Frate + log(Pop) + log(Fertility) + Purban, UN3)
par(mfrow = c(2, 2))
plot(fit)
```

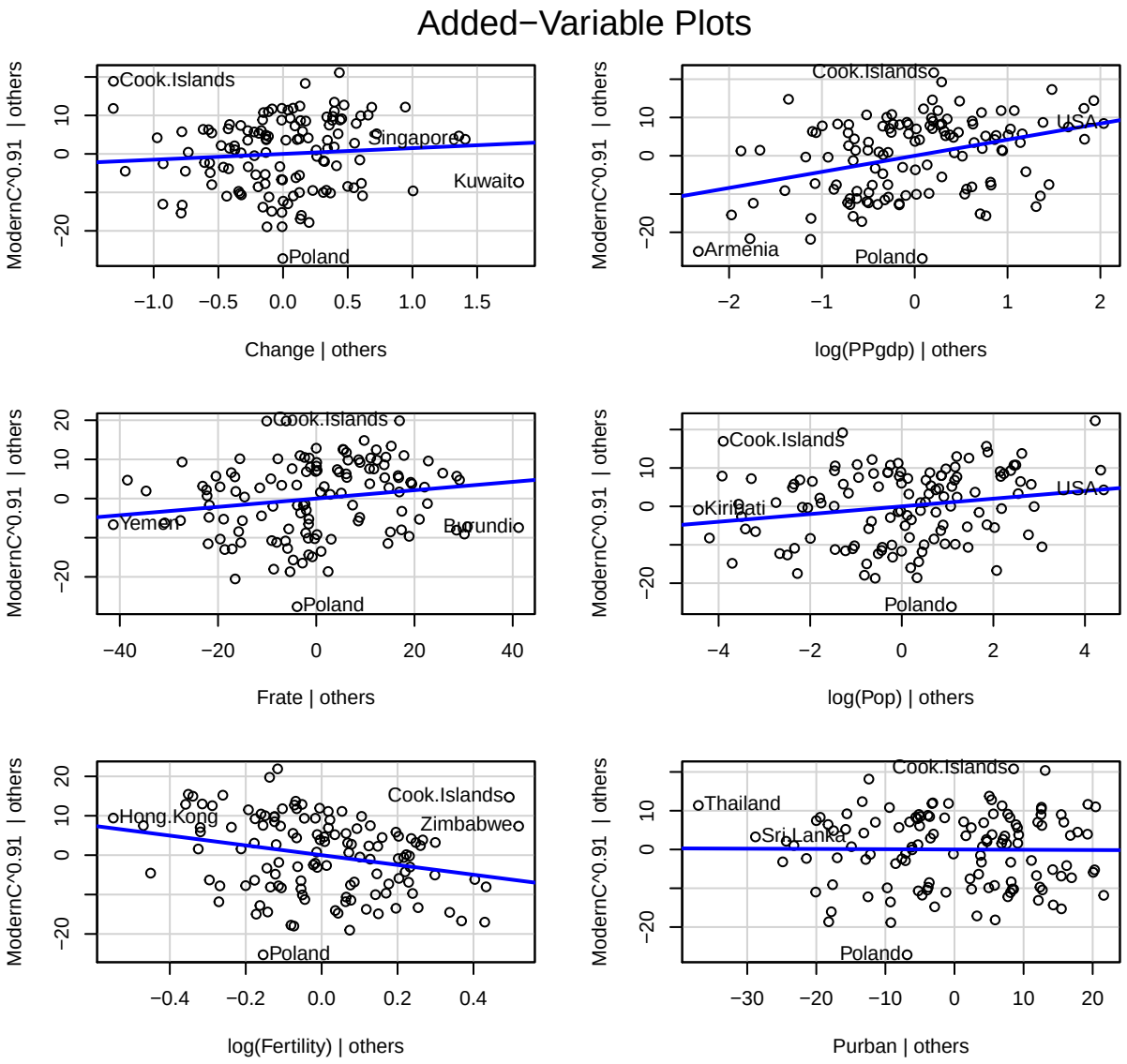









```
car::avPlots(fit)
```

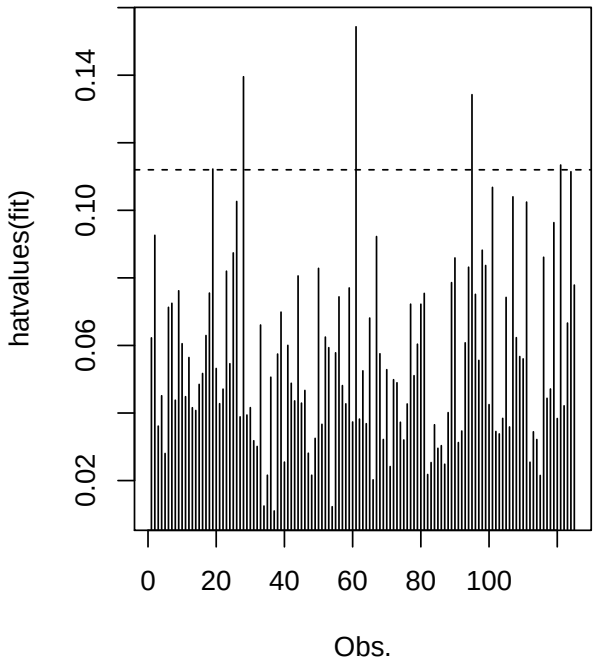


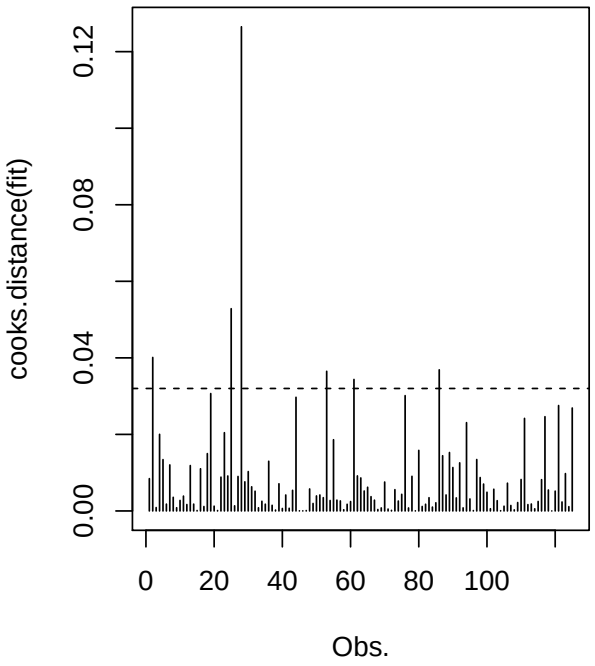
Answer: According to the results above, It seems to all predictors are satisfied.

8. Are there any outliers or influential points in the data? Explain. If so, refit the model after removing any outliers/influential points and comment on residual plots.

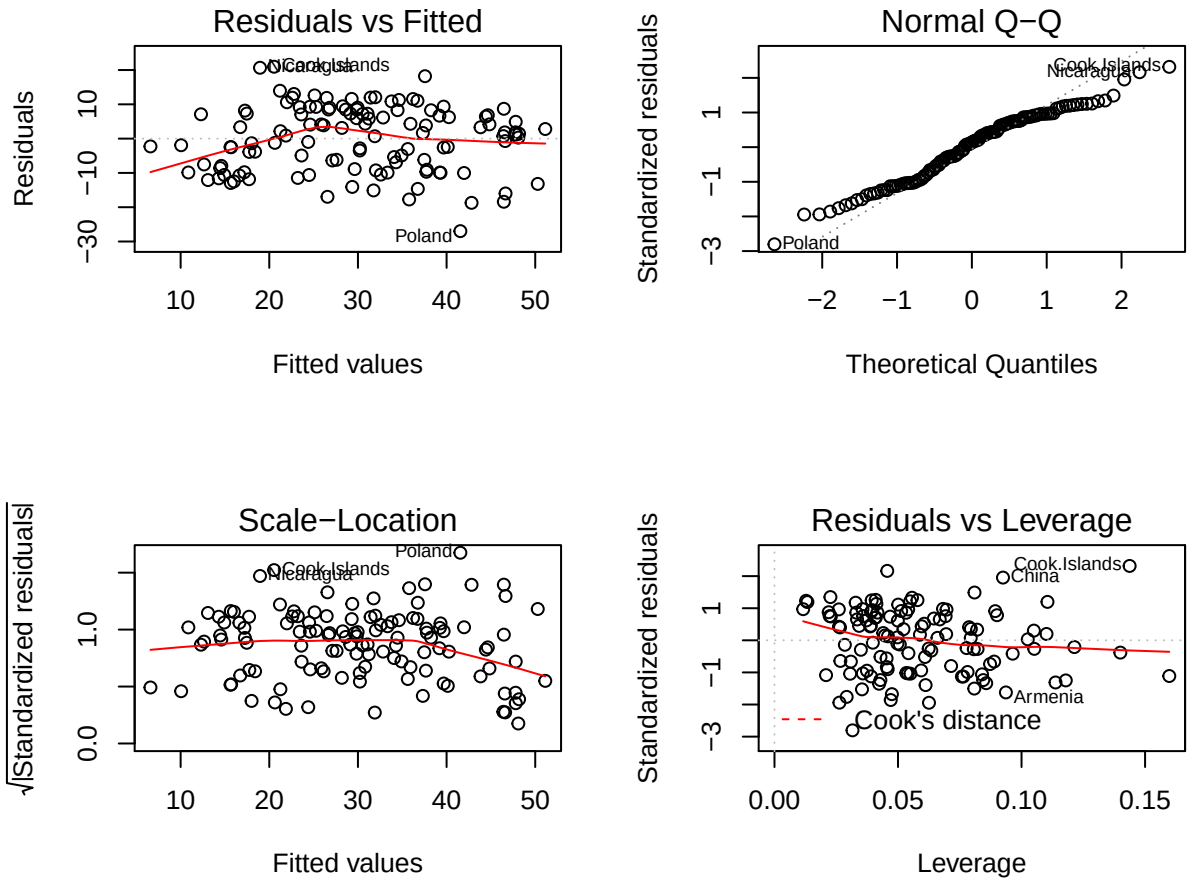
Answer: Based on the plot, there are some outliers/influential points (The points which over the dotted line I set).

```
par(mfrow = c(1, 2))
plot(hatvalues(fit), type = 'h', xlab = 'Obs.')
abline(h = 2*mean(hatvalues(fit)), lty = 2)
plot(cooks.distance(fit), type = 'h', xlab = 'Obs.')
abline(h = 4/nobs(fit), lty = 2)
```






```
ind = which(hatvalues(fit) > 2*mean(hatvalues(fit)))
ind = c(ind, which(cooks.distance(fit) > 4/nobs(fit)))
dat = UN3[-ind, ]
reg = lm(ModernC^0.91 ~ Change + log(PPgdp) + Frate + log(Pop) + log(Fertility) + Purban, dat)
par(mfrow = c(2, 2))
plot(reg)
```



Answer: I removed the outliers/influential points and All assumptions of the linear model are satisfied.

Summary of Results

9. 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!

```
#knitr::kable(round(cbind(summary(reg)$coef, confint(reg)), 3))
round(cbind(summary(reg)$coef, confint(reg)), 3)
```

```
##                Estimate Std. Error t value Pr(>|t|)   2.5 % 97.5 %
## (Intercept)      -5.622     11.055  -0.509    0.612 -27.523 16.280
## Change            1.595      1.748   0.912    0.364  -1.869  5.058
## log(PPgdp)        4.186      1.031   4.058    0.000   2.142  6.229
## Frate             0.112      0.058   1.929    0.056  -0.003  0.227
## log(Pop)          0.951      0.485   1.963    0.052  -0.009  1.912
## log(Fertility)  -12.666      4.352  -2.910    0.004 -21.288 -4.043
## Purban           -0.012      0.075  -0.164    0.870  -0.160  0.136
```

Answer: The coefficient $\beta_{\text{Change}} = 1.595$, means the ModernC$^{0.91}$ is expected to increase by 1.595% for 1% increase of `Change`.

The coefficient $\beta_{\log(\text{PPgdp})} = 4.186$, means the ModernC$^{0.91}$ is expected to increase by 4.186% for 1% increase of log(PPgdp).

The coefficient $\beta_{\text{Frate}} = 0.112$, means the ModernC$^{0.91}$ is expected to increase by 0.112% for 1% increase of `Frate`.

The coefficient $\beta_{\log(\text{Pop})} = 0.951$, means the ModernC$^{0.91}$ is expected to increase by 0.951% for 1% increase of log(Pop.

The coefficient $\beta_{\log(\text{Fertility})} = -12.666$, means the ModernC$^{0.9}$ is expected to decrease by 12.666% for 1% increase of log(Fertility.

The coefficient $\beta_{\text{Purban}} = -0.012$, means the ModernC$^{0.9}$ is expected to decrease by 0.012% for 1% increase of `Purban`.

10. 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(reg)
```

```
##
## Call:
## lm(formula = ModernC^0.91 ~ Change + log(PPgdp) + Frate + log(Pop) +
##     log(Fertility) + Purban, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -26.977  -8.664   1.318   7.513  20.965
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -5.62162   11.05491  -0.509  0.61208
## Change           1.59466    1.74826   0.912  0.36364
## log(PPgdp)       4.18555    1.03139   4.058 9.14e-05 ***
## Frate            0.11203    0.05809   1.929  0.05629 .
## log(Pop)         0.95148    0.48461   1.963  0.05206 .
## log(Fertility) -12.66571    4.35233  -2.910  0.00435 **
## Purban          -0.01223    0.07462  -0.164  0.87011
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.777 on 113 degrees of freedom
##   (81 observations deleted due to missingness)
## Multiple R-squared:  0.5529, Adjusted R-squared:  0.5292
## F-statistic: 23.29 on 6 and 113 DF,  p-value: < 2.2e-16
```

The final model is:

$$\begin{aligned}\text{ModernC}^{0.91} = &-5.622 + 1.595\text{Change} + 4.186\log(\text{PPgdp}) + 0.112\text{Frate} \\ &+ 0.951\log(\text{Pop}) - 12.666\log(\text{Fertility}) - 0.012\text{Purban}\end{aligned}$$

Answer: We remove some outliers/influential points in order to fit the final model. In this model, all assumptions are satisfied. And the model can explain 52.92% of the variability of ModernC$^{0.91}$, but the predictors `Change` and `Purban` are not significant at 10% level of significance.

Methodology

11. Exercise 9.12 from ALR

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

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

where h_{ii} is the ith diagonal element of $H = X(X^T X)^{-1} X^T$ using direct multiplication and simplify in terms of_ h_{ii}.

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

$$\begin{aligned}
&\Rightarrow \left(X_{(i)}^T X_{(i)}\right) \left((X^T X)^{-1} + \frac{(X^T X)^{-1} x_i x_i^T (X^T X)^{-1}}{1 - h_{ii}}\right) \\
&= \left(X^T X - x_i x_i^T\right) \left((X^T X)^{-1} + \frac{(X^T X)^{-1} x_i x_i^T (X^T X)^{-1}}{1 - h_{ii}}\right) \\
&= I + \frac{x_i x_i^T (X^T X)^{-1}}{1 - h_{ii}} - x_i x_i^T (X^T X)^{-1} - \frac{x_i x_i^T (X^T X)^{-1} x_i x_i^T (X^T X)^{-1}}{1 - h_{ii}} \\
&= I + \frac{x_i x_i^T (X^T X)^{-1}}{1 - h_{ii}} - x_i x_i^T (X^T X)^{-1} - \frac{x_i h_{ii} x_i^T (X^T X)^{-1}}{1 - h_{ii}} \\
&= I + \frac{1 - (1 - h_{ii}) - h_{ii}}{1 - h_{ii}} x_i x_i^T (X^T X)^{-1} \\
&= I
\end{aligned}$$

$$\Rightarrow (X_{(i)}^T X_{(i)})^{-1} = (X^T X)^{-1} + \frac{(X^T X)^{-1} x_i x_i^T (X^T X)^{-1}}{1 - h_{ii}}$$

12. Exercise 9.13 from ALR. Using the above, show

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

$$\begin{aligned}
\hat{\beta}_{(i)} &= (X_{(i)}^T X_{(i)})^{-1} X_{(i)}^T Y_{(i)} \\
&= \left((X^T X)^{-1} + \frac{(X^T X)^{-1} x_i x_i^T (X^T X)^{-1}}{1 - h_{ii}}\right) \left(X^T Y - x_i y_i\right) \\
&= \hat{\beta} - (X^T X)^{-1} x_i y_i + \frac{(X^T X)^{-1} x_i x_i^T \hat{\beta}}{1 - h_{ii}} - \frac{(X^T X)^{-1} x_i h_{ii} y_i}{1 - h_{ii}} \\
&= \hat{\beta} - \frac{(X^T X)^{-1} x_i y_i}{1 - h_{ii}} + \frac{(X^T X)^{-1} x_i \hat{y}_i}{1 - h_{ii}} \\
&= \hat{\beta} - \frac{(X^T X)^{-1} x_i (y_i - \hat{y}_i)}{1 - h_{ii}} \\
&= \hat{\beta} - \frac{(X^T X)^{-1} x_i e_i}{1 - h_{ii}}
\end{aligned}$$

13. (optional) Prove that the intercept i n the added variable scatter plot will always be zero. *Hint: use the fact that if H is the projection matrix for X which contains a column of ones, then $1_n^T (I - H) = 0$ or $(I - H) 1_n = 0$. Use this to show that the sample mean of residuals will always be zero if there is an intercept.*

If X has the intercept term, then

$$1_n^T (I - H) = 0$$

$$
\begin{aligned}
\Rightarrow 1_n^T e &= 1_n^T (Y - \hat{Y}) \\
&= 1_n^T (Y - HY) \\
&= 1_n^T (I - H) Y \\
&= 0Y \\
&= 0
\end{aligned}
$$

$$
\Rightarrow \frac{1}{n} 1_n^T e = \frac{1}{n} \sum_{i=1}^{n} e_i = 0
$$