

# HW2 STA521 Fall18

Shuai Yuan, sy144, Kolin96

Due September 23, 2018 5pm

## Background 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 qualitative?

```
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
```

Six variables have missing data: ModernC, Change, PPgdp, Frate, Pop and Fertility.

Quantitative: ModernC, Change, PPgdp, Frate, Pop, Fertility, Purban

Qualitative: (none)

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

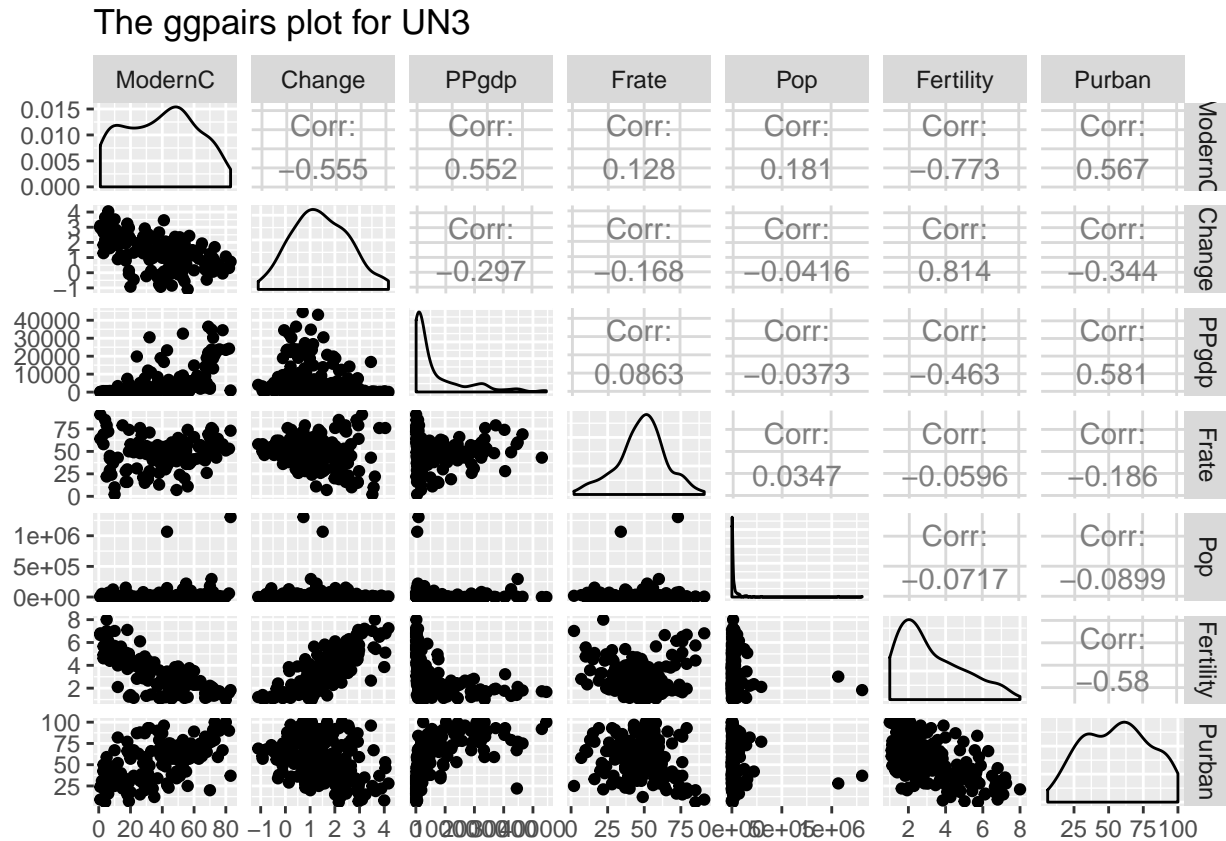
```
mean_result <- sapply(UN3, mean, na.rm=TRUE)
std_result <- sapply(UN3, sd, na.rm=TRUE)
kable(cbind(mean_result, std_result), col.names = c("mean", "std."), digits = 4)
```

	mean	std.
ModernC	38.7171	22.6366
Change	1.4184	1.1331
PPgdp	6527.3881	9325.1886
Frate	48.3054	16.5324
Pop	30281.8714	120676.6945
Fertility	3.2140	1.7069

	mean	std.
Purban	56.2000	24.1098

- 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?

```
gp = ggpairs(UN3, titl="The ggpairs plot for UN3")
print(gp, progress=FALSE)
```



If we try to predict `ModernC`, it may be better to use the predictors `Change`, `Frate`, `Fertility` and `Purban`, since they are linear correlated.

Potential outliers: the `Pop` variable has two extremely large data point.

Nonlinear relationships: `PPgdp` seems to grow exponentially as `ModernC` increases. `Frate` also has some noised nonlinear relationship.

Transformation needed: `PPgdp` should be transformed to log scale.

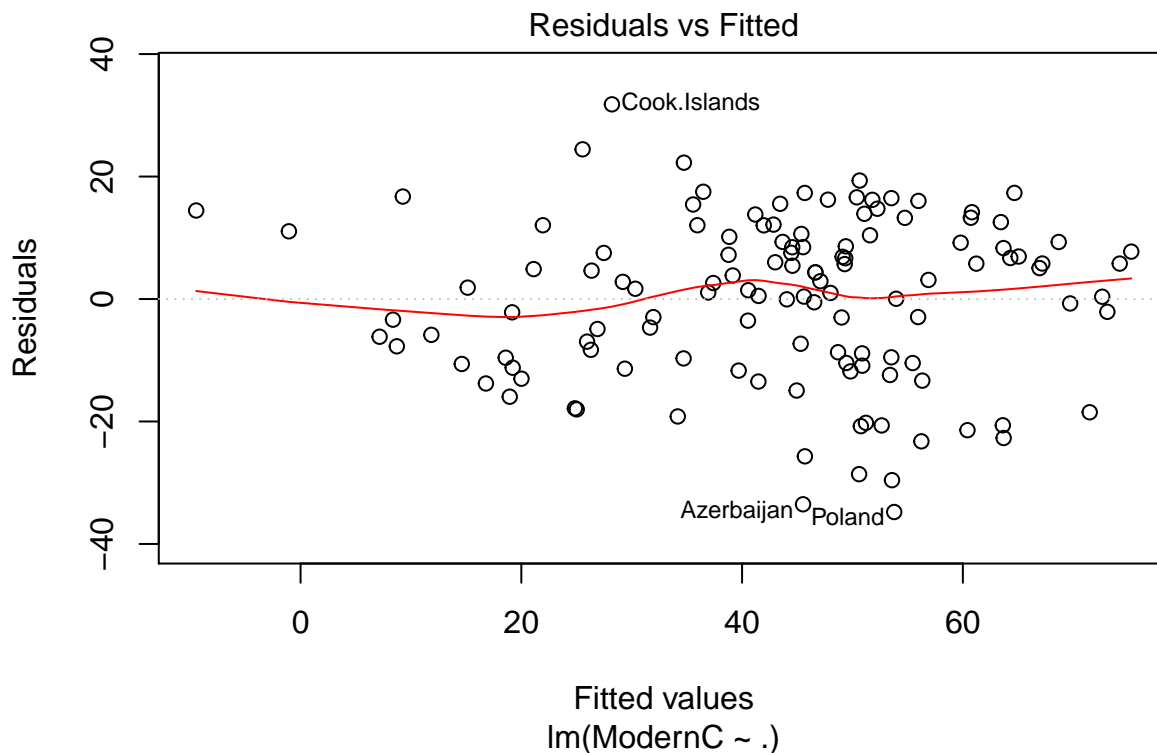
## Model Fitting

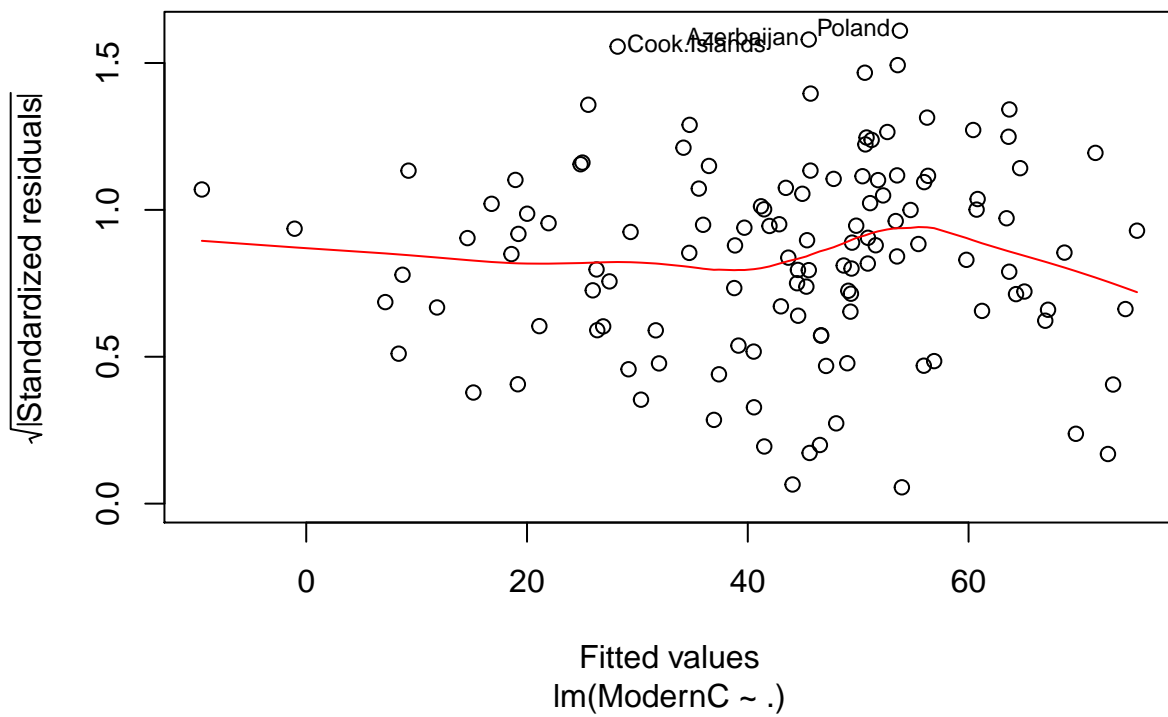
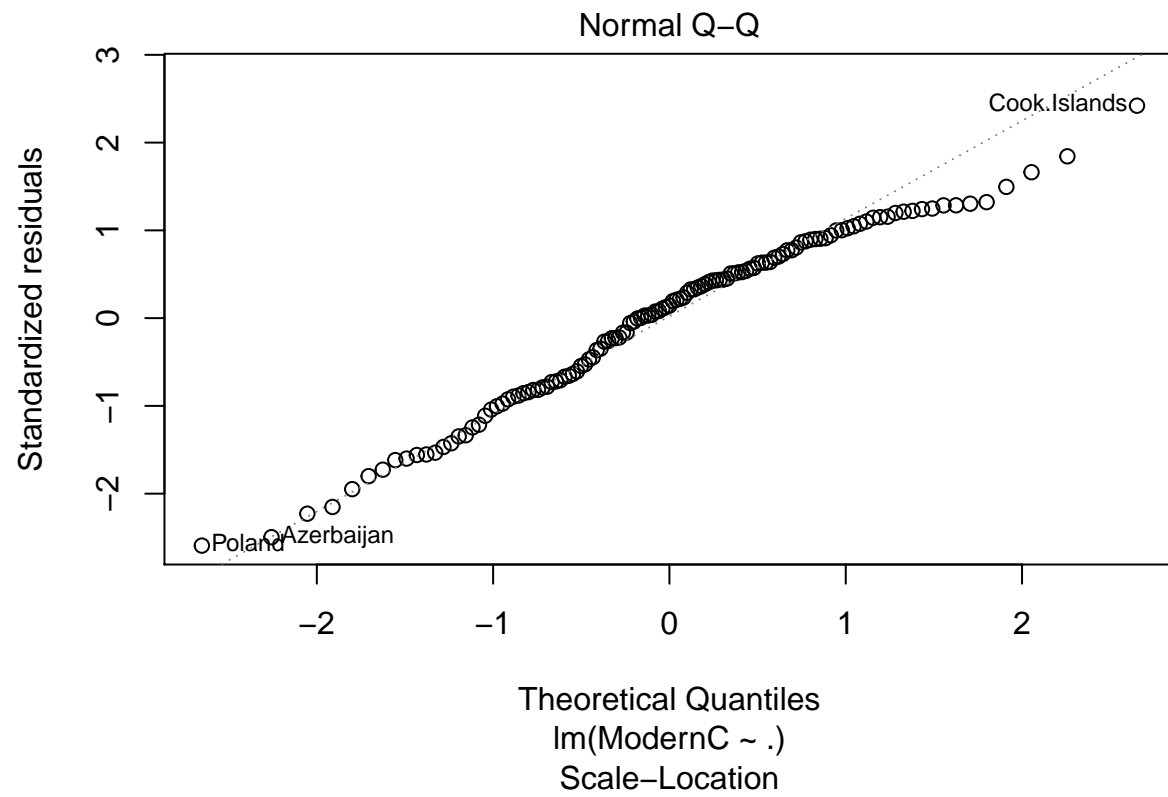
- Use the `lm()` function to perform a multiple linear regression with `ModernC` as the response and all other variables as the predictors, using the formula `ModernC ~ .`, where the `.` includes all remaining variables in the dataframe. Create diagnostic residual plot from the linear model object and comment on results regarding assumptions. How many observations are used in your model fitting?

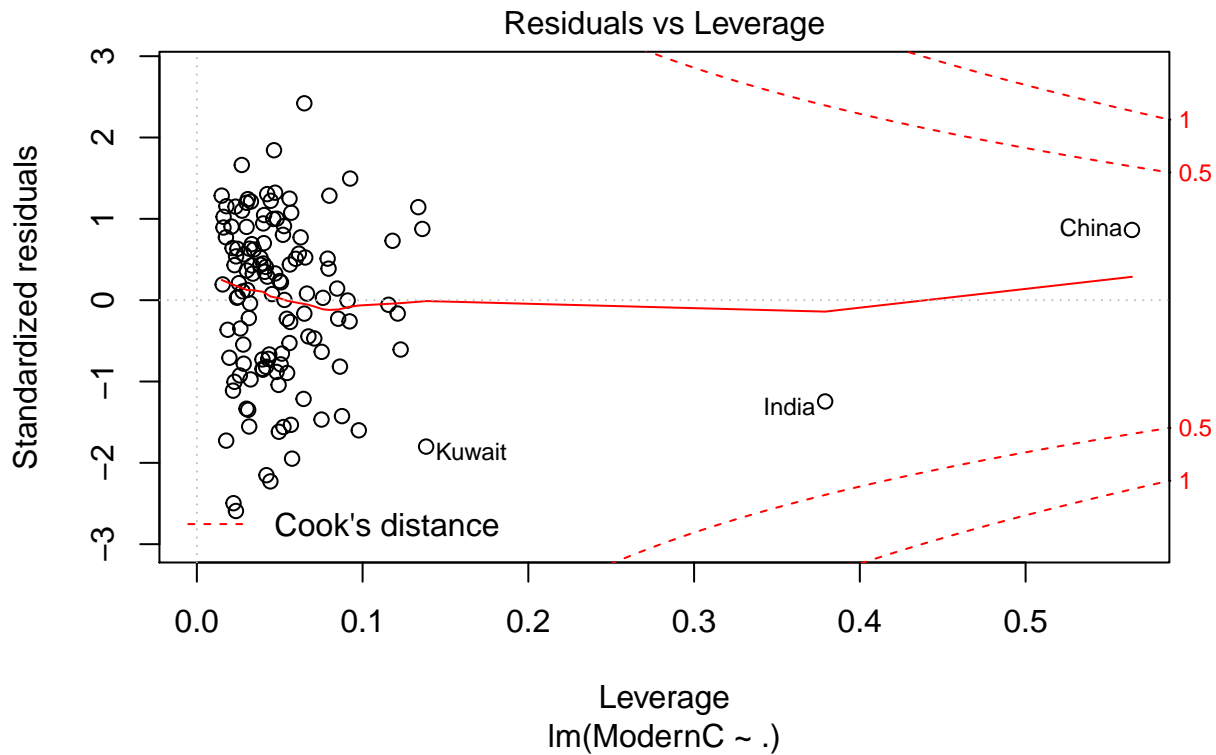
```
mymodel = lm(ModernC ~ ., data=UN3)
summary(mymodel)
```

```
##
## 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
```

```
# diagnostic residual plot
plot(mymodel)
```







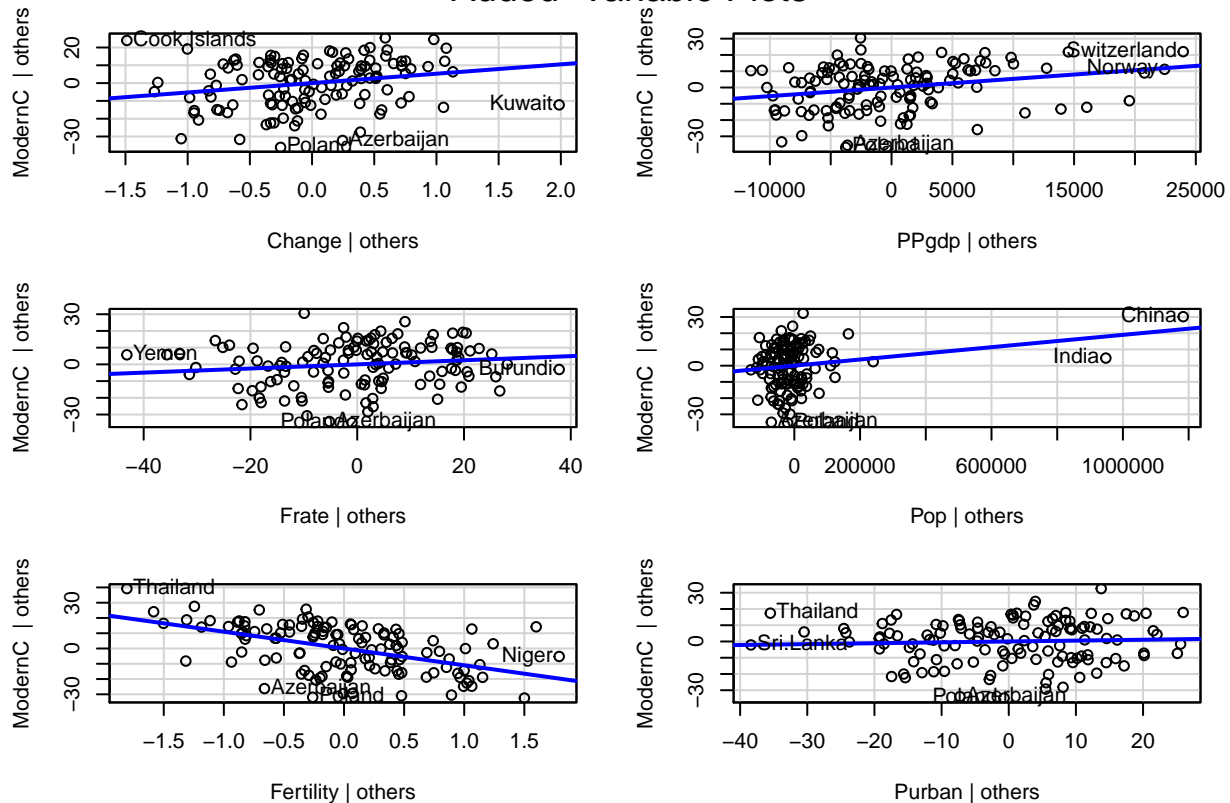
Comments: The linear model works fairly well, since the R-squared values are high (around 0.6) and the p-value is low ( $<2.2e-16$ ). There are noises, but the residuals are generally independent of fitted values and have mean 0. For the Normal Q-Q plot, the higher quantiles can't match perfectly, but it is still not bad. Cook's distances are all below 0.5; China and India are potential outliers.

How many observations are used: 125, since the other 85 observations are deleted due to missingness.

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(mymodel)
```

## Added-Variable Plots



Transformations that are needed: The `Pop | others` plot shows that we need to transform `pop` to log scale, because all data points are now concentrated near 0.

Influential localities for each term are all shown in the figures:

**Change:** Cook islands, Kuwaito, Poland, Azerbaijan;

**PPgdp:** Switzerland, Norway, Azerbaijan;

**Frate:** Yeman, Burundi, Poland, Azerbaijan;

**Pop:** China, Indea, Azerbaijan;

**Fertility:** Thailand, Niger, Poland, Azerbaijan;

**Purban:** Thailand, Sri.Lanka, Poland, Azerbaijan.

Note that Azerbaijan appears in all the lists.

- 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.

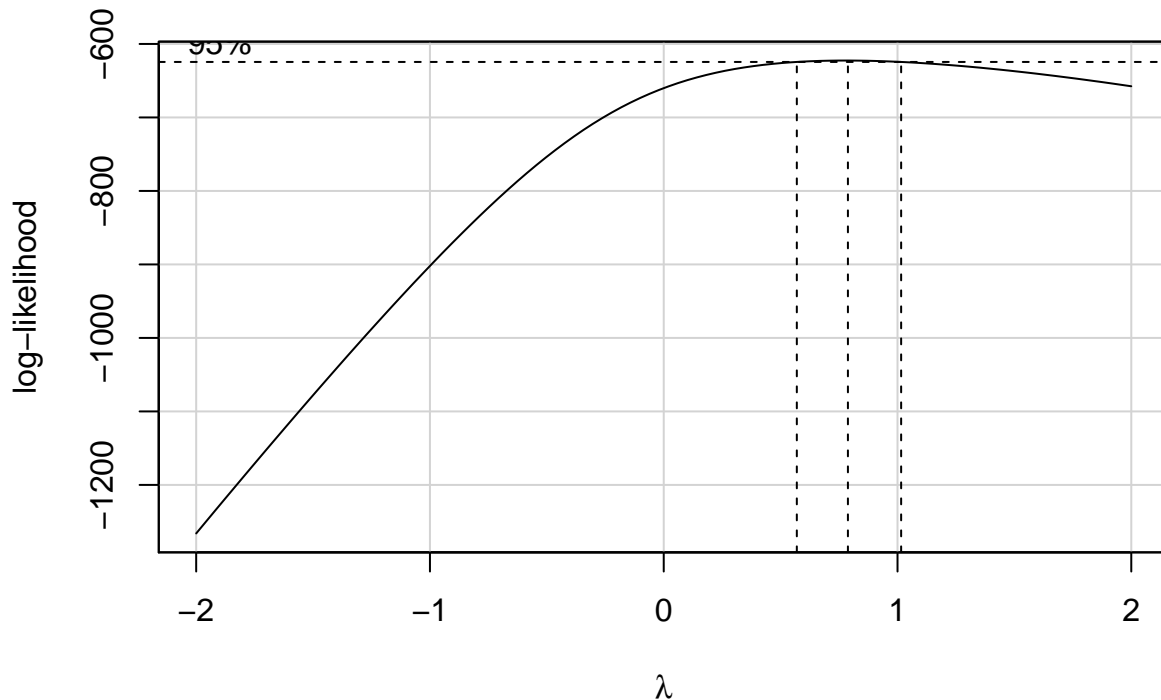
```
UN3_nonnega = UN3
UN3_nonnega['Change'] = UN3_nonnega['Change'] - min(UN3['Change'], na.rm=TRUE) + 1
car::boxTidwell(ModernC ~ Pop, ~ PPgdp + Fertility + Change + Frate + Purban, data=UN3_nonnega, na.act.

## MLE of lambda Score Statistic (z) Pr(>|z|)
##      0.63309      -0.5543      0.5794
##
## iterations = 3
```

We first transform `Change` by subtracting its minimum value to make it nonnegative. Then, we transform `Pop`, the MLE of  $\lambda$  computed by `car::boxTidwell` is 0.63309.

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

```
mymodel = lm(ModernC ~ ., data=UN3_nonnega)
lambda = car::boxCox(mymodel)
```



By outputting `car::boxCox(mymodel)`, we may find the best  $\lambda$  is around 0.7879.

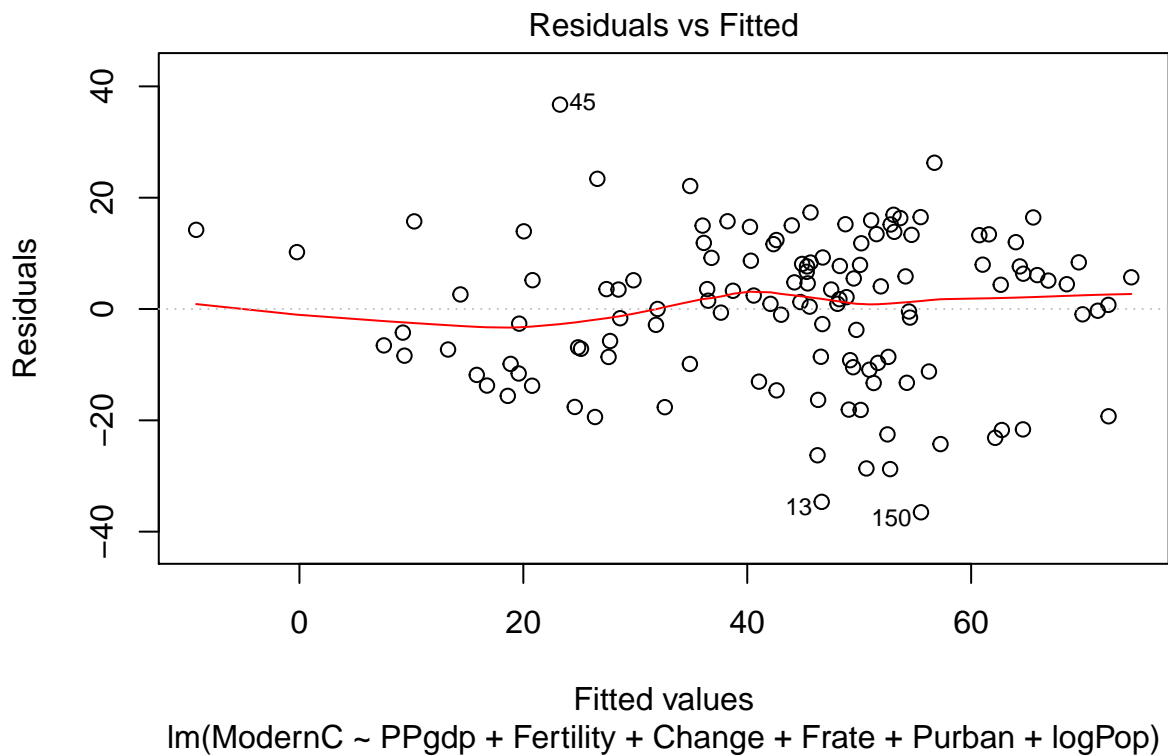
- Fit the regression using the transformed variables. Provide residual plots and added variables plots and comment. If you feel that you need additional transformations of either the response or predictors, repeat any steps until you feel satisfied.

```
UN3_nonnega = UN3_nonnega %>% mutate(logPop = log(Pop))
mymodel = lm(ModernC ~ PPgdp + Fertility + Change + Frate + Purban + logPop, data=UN3_nonnega)
summary(mymodel)
```

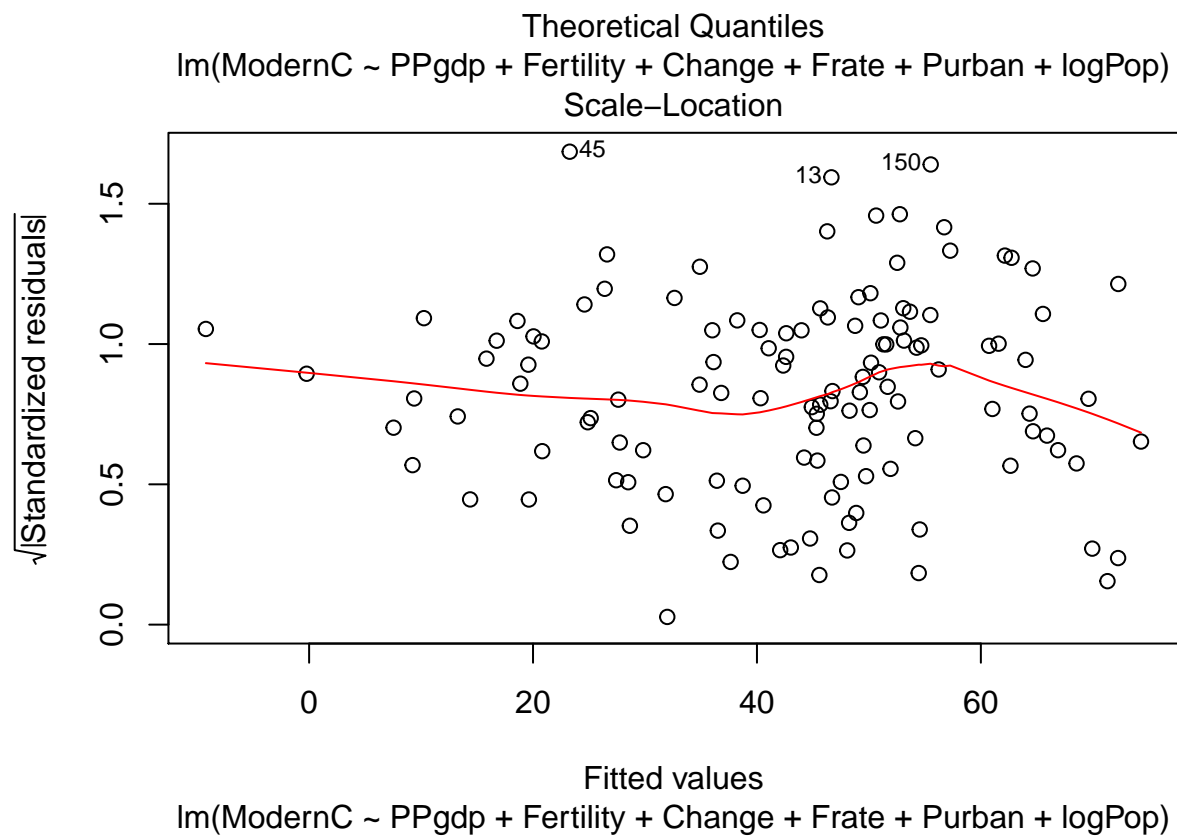
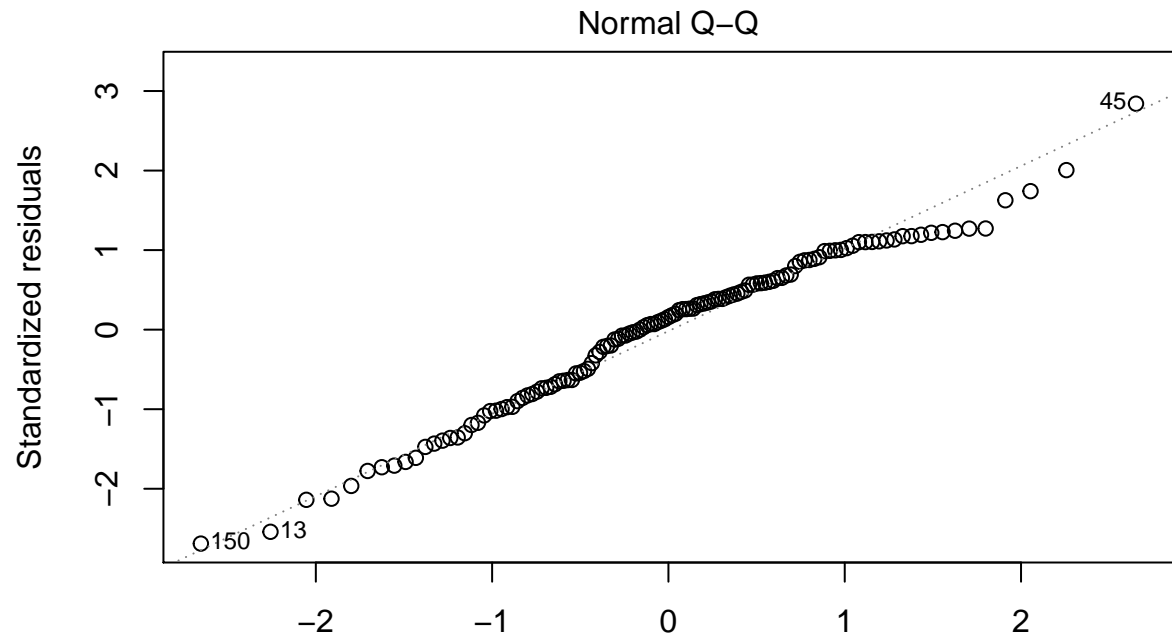
```
##
## Call:
## lm(formula = ModernC ~ PPgdp + Fertility + Change + Frate + Purban +
##     logPop, data = UN3_nonnega)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -36.520  -9.676   2.132   9.191  36.723
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.994e+01  1.114e+01   3.585 0.000492 ***
## PPgdp        5.384e-04  1.795e-04   3.000 0.003296 **
## Fertility    -1.145e+01  1.763e+00  -6.495 2.06e-09 ***
## Change       5.444e+00  2.129e+00   2.557 0.011827 *
```

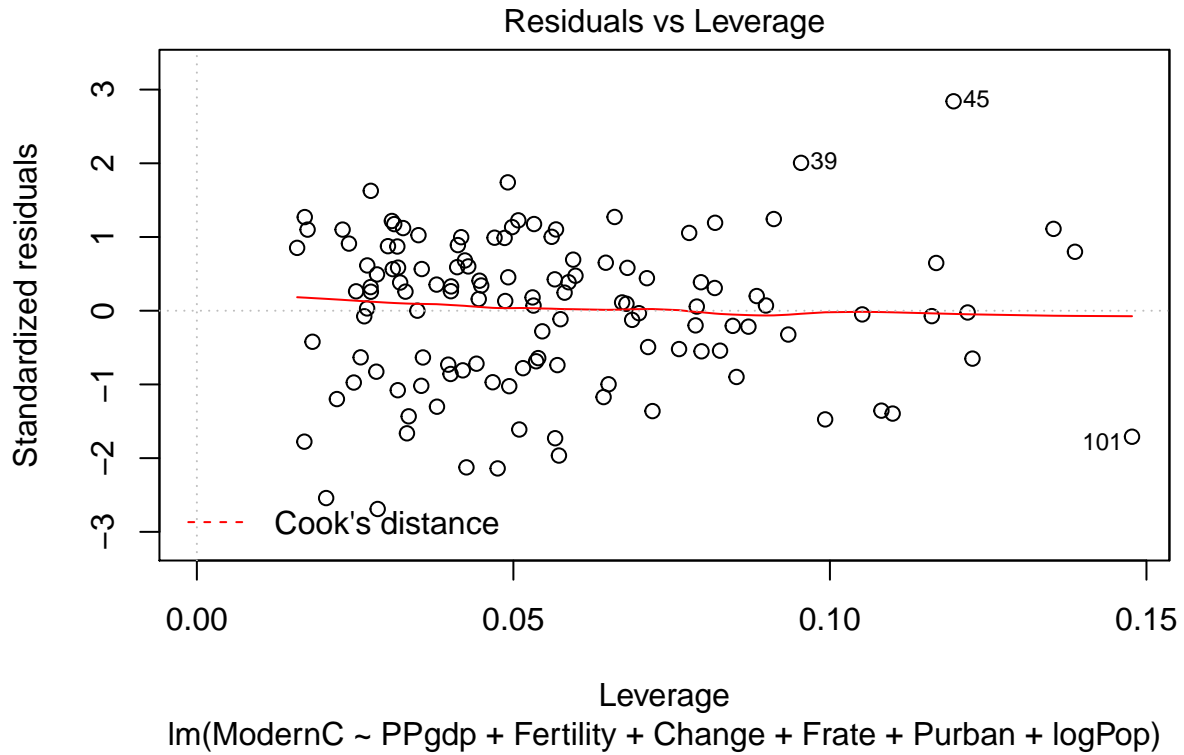
```
## Frate      1.226e-01  8.176e-02  1.499 0.136442
## Purban     2.014e-02  9.240e-02  0.218 0.827855
## logPop     8.623e-01  6.297e-01  1.369 0.173447
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.78 on 118 degrees of freedom
## (85 observations deleted due to missingness)
## Multiple R-squared:  0.6072, Adjusted R-squared:  0.5872
## F-statistic: 30.4 on 6 and 118 DF, p-value: < 2.2e-16
```

```
plot(mymodel)
```

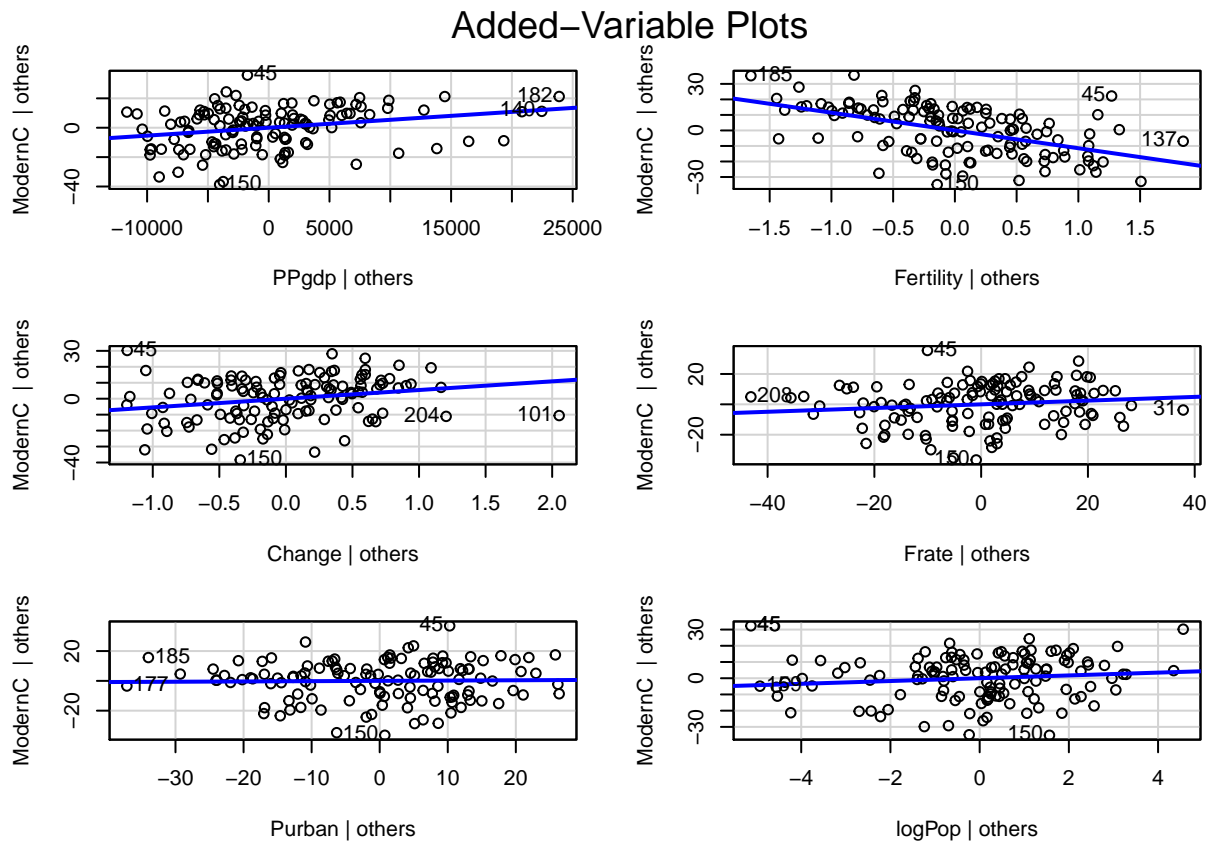








```
car::avPlots(mymodel)
```



We also applies log transformation to Pop. The residual plots show that residuals are very small (close to

mean 0) and independent to fitted values. The Cook's distance are all very small. In the added variable plot, the **logPop** plot is much more balanced than the original **Pop** plot. The new model works better than our previous model.

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?

Yes, we would get different models, but these models are not necessarily better than our previous model, because the original response variable actually has strong linear correlations with many predictors. If we first transform the response variable by log, poly, etc., we still need to transform many predictors accordingly, making the model very messy.

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.

## 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!
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

## 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.*
14. For multiple regression with more than 2 predictors, say a full model given by  $Y \sim X_1 + X_2 + \dots$ .  $X_p$  we create the added variable plot for variable  $j$  by regressing  $Y$  on all of the  $X$ 's except  $X_j$  to form  $e_Y$  and then regressing  $X_j$  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.