

HW2 STA521 Fall18

Min Chul Kim (NetID mk408, Github ID minchel93)

Due September 23, 2018 5pm

Background Reading

Readings: Chapters 3-4 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
##
## Attaching package: 'car'
## The following object is masked from 'package:dplyr':
##
##      recode
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 qualitative?

```
#Summary of Data
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
```

```
#Additional Missing Data
help(UN3)
```

```
#Data Type of Predictors
sapply(UN3, class)
```

```
## ModernC      Change      PPgdp      Frate      Pop Fertility      Purban
## "integer" "numeric" "integer" "integer" "numeric" "numeric" "integer"
```

According to the summary function, there are 58 data missing from ModernC, 1 missing from Change, 9 missing from PPgdp, 43 from Frate, 2 missing from Pop, 10 missing from Fertility, and 0 missing from Purban.

All the predictors are quantitative, according to the description of each predictor from R.

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

```
#Calculating the Mean and SD, w/ NA's Removed from Predictors
kable(sapply(UN3, mean, na.rm = TRUE), format = "markdown", col.names = "Mean")
```

	Mean
ModernC	38.717105
Change	1.418373
PPgdp	6527.388060
Frate	48.305389
Pop	30281.871428
Fertility	3.214000
Purban	56.200000

```
kable(sapply(UN3, sd, na.rm = TRUE), format = "markdown", col.names = "Standard Deviation")
```

	Standard Deviation
ModernC	2.263661e+01
Change	1.133133e+00
PPgdp	9.325189e+03
Frate	1.653245e+01
Pop	1.206767e+05
Fertility	1.706918e+00
Purban	2.410976e+01

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?

```
#Removing NA's from Data
UN3_No_NA = na.omit(UN3)
```

```
#Predictors
ModernC = UN3_No_NA$ModernC
```

```

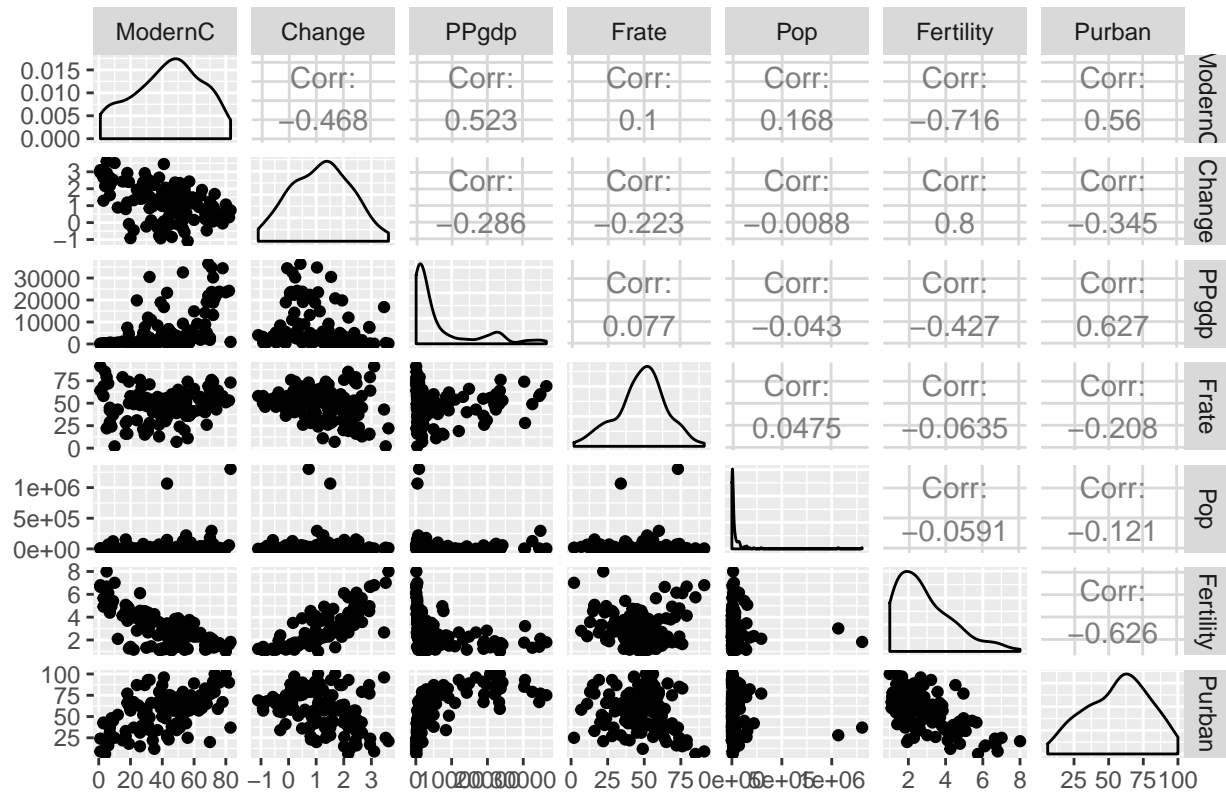
Change = UN3_No_NA$Change
PPgdp = UN3_No_NA$PPgdp
Frate = UN3_No_NA$Frate
Pop = UN3_No_NA$Pop
Fertility = UN3_No_NA$Fertility
Purban = UN3_No_NA$Purban

```

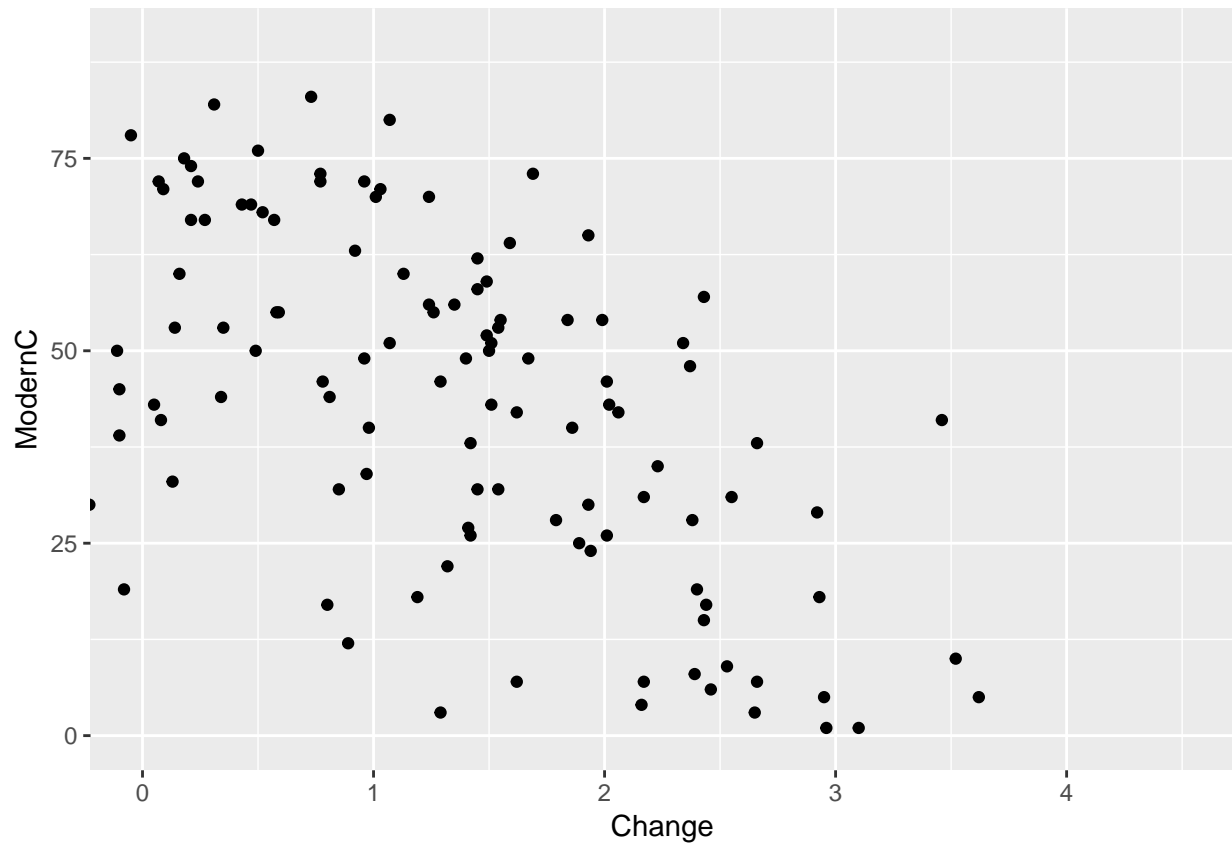
```
#Plots
```

```
ggpairs(UN3_No_NA, progress = FALSE, title = "Scatterplot Matrix for Dataset UN3")
```

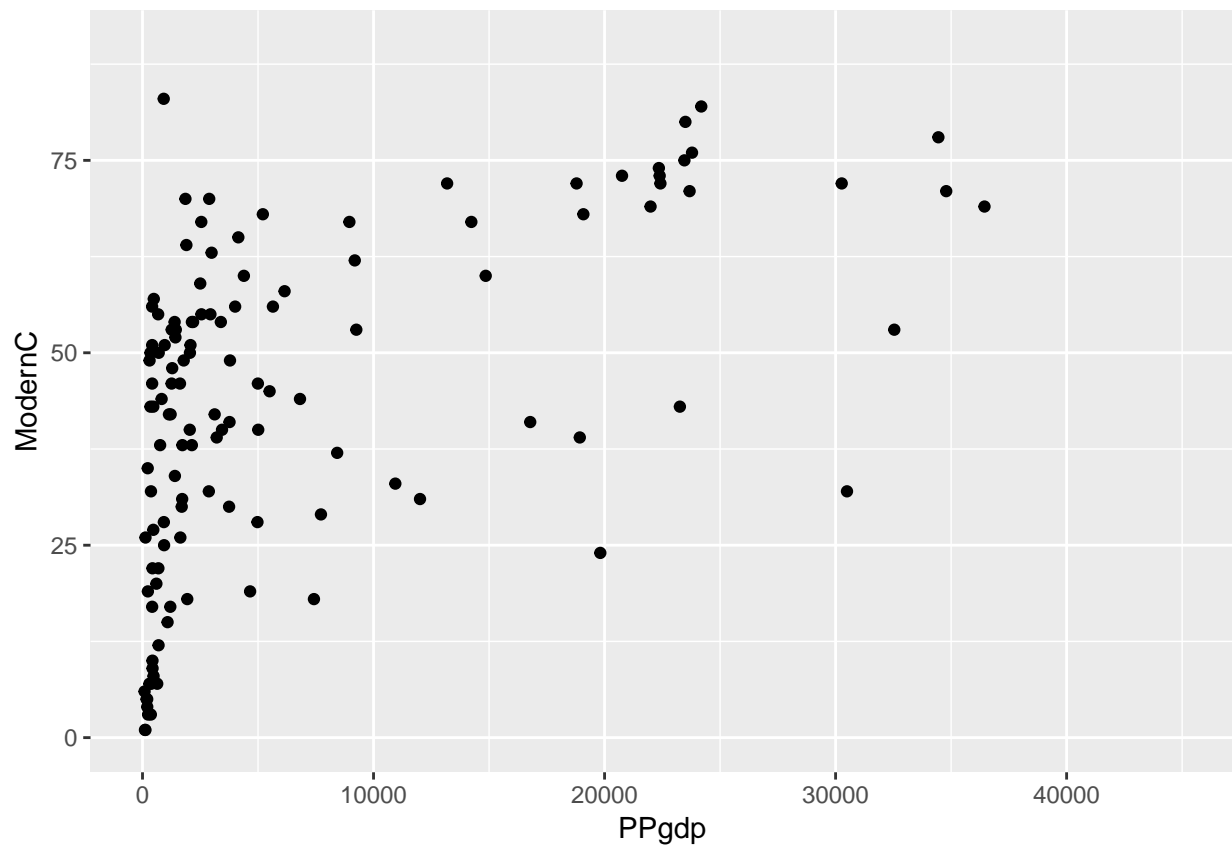
Scatterplot Matrix for Dataset UN3



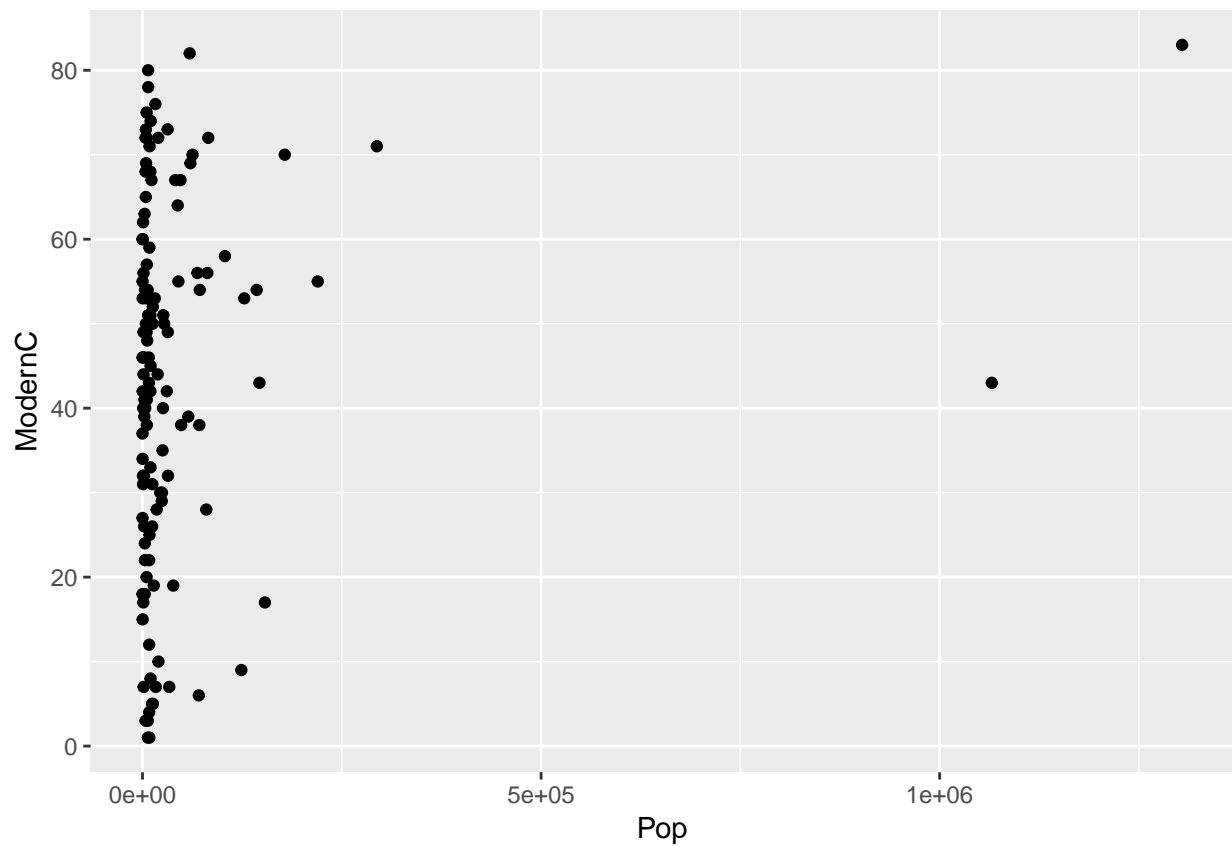
```
ggplot(UN3_No_NA, aes(Change, ModernC)) + geom_point() + coord_cartesian(xlim = c(0,4.5), ylim = c(0,90))
```



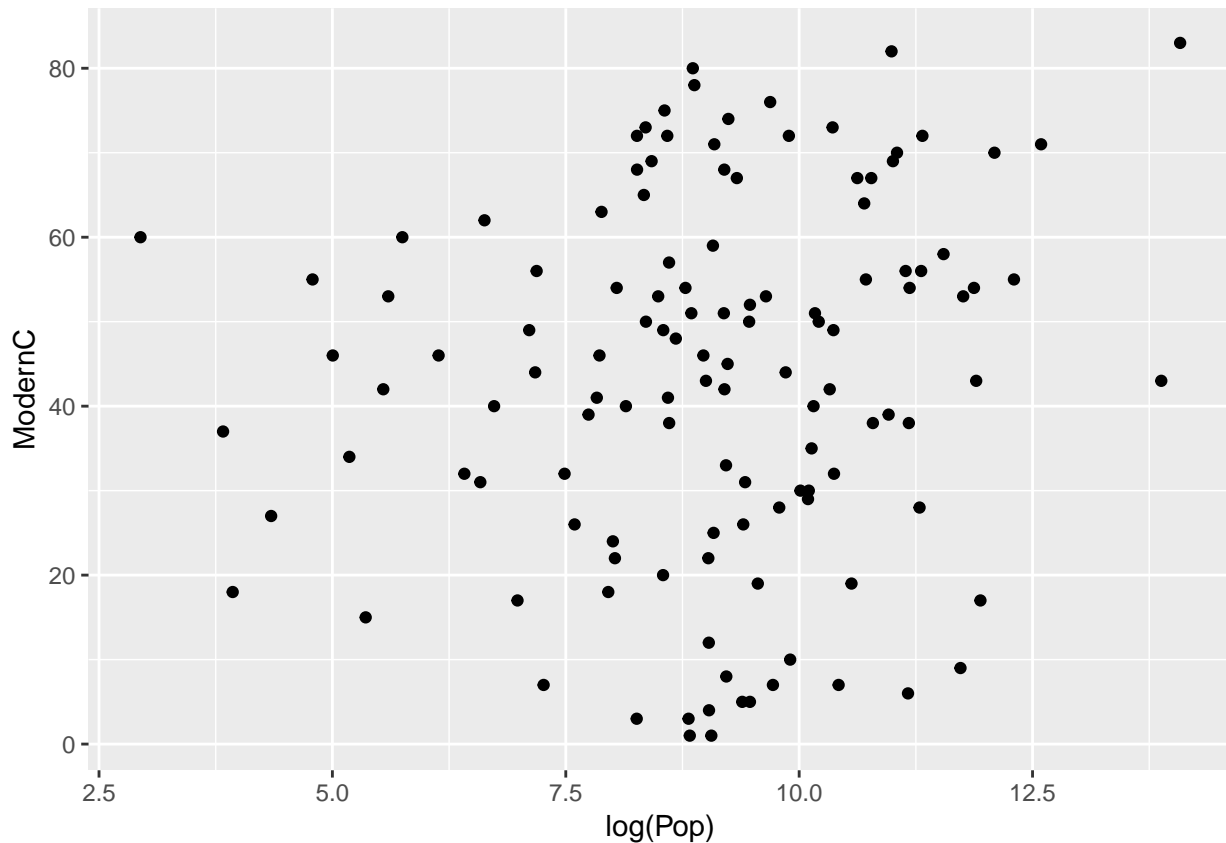
```
ggplot(UN3_No_NA, aes(PPgdp, ModernC)) + geom_point() + coord_cartesian(xlim = c(0,45000), ylim = c(0,9
```



```
ggplot(UN3_No_NA, aes(Pop, ModernC)) + geom_point()
```



```
ggplot(UN3_No_NA, aes(log(Pop), ModernC)) + geom_point()
```



By visual inspection, I was not able to identify potential outliers. However, I was able to observe that ModernC and PPgdp, PPgdp and Fertility, and PPgdp and Purban all have noticeable non-linear relationships, which may demand some transformations.

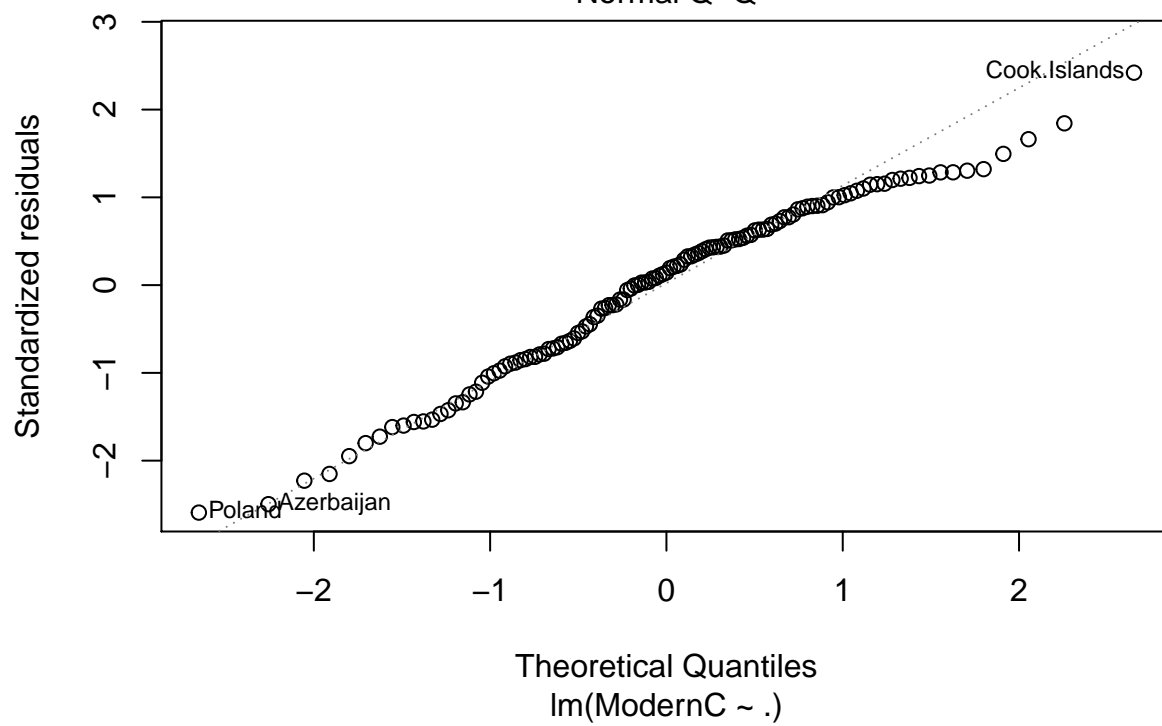
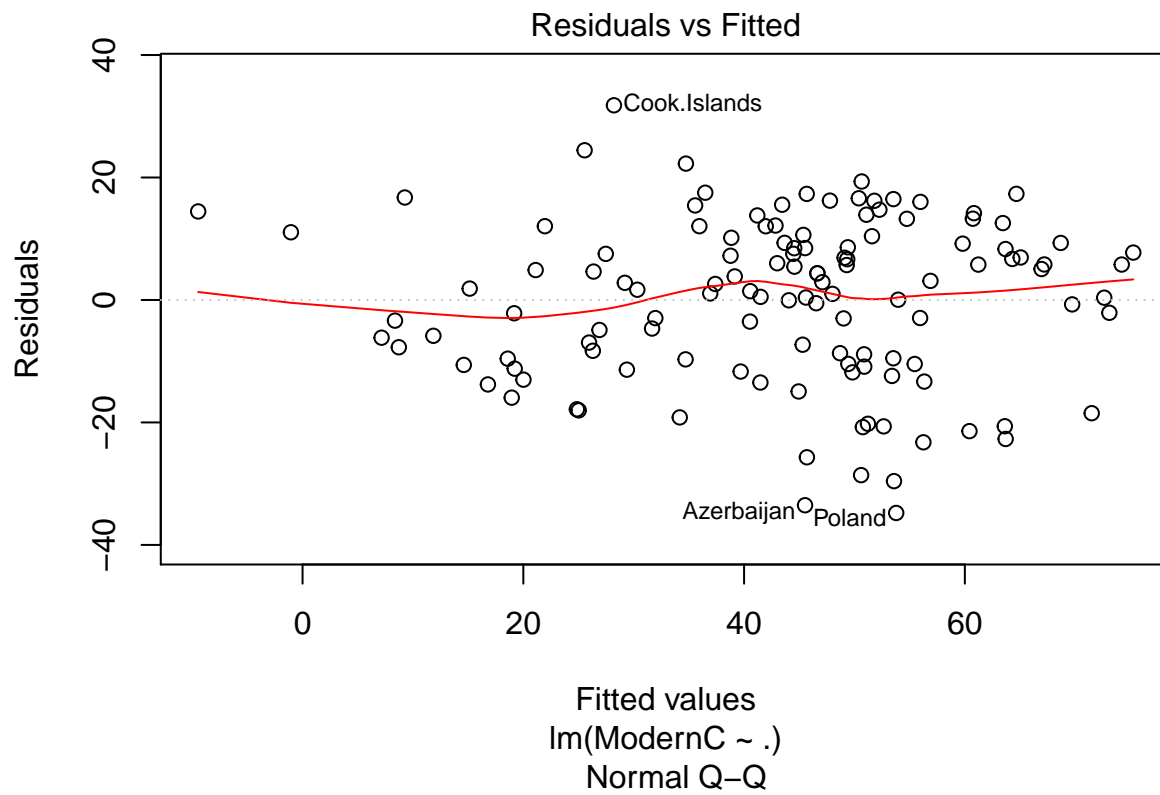
As for the findings regarding predicting “ModernC” using other variables, I concluded that the predictor “Fertility” would be the best pre-transformed variable to predict the response, since “Fertility” has the most linear relationship, among the predictors, with “ModernC.” I also concluded that “PPgdp” would be the worst pre-transformed variable to predict the response, since “PPgdp” has the most noticeable non-linear relationship, among the predictors, with “ModernC.”

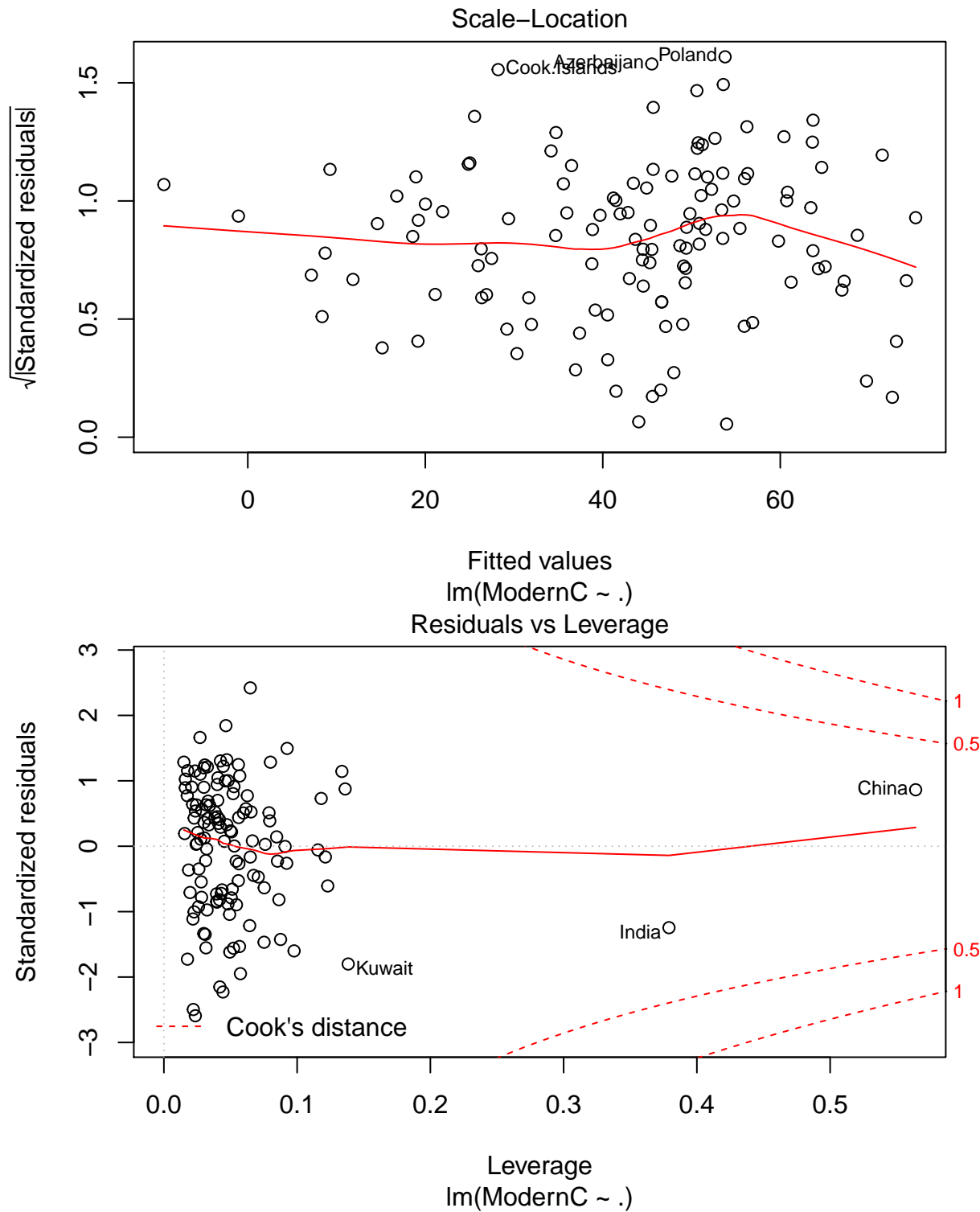
Model Fitting

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?

```
#Multiple Regression
LM_ModernC = lm(ModernC ~ ., data = UN3_No_NA)

#Multiple Regression Plot
plot(LM_ModernC)
```





```
#Observations Used in Multiple Regression
nobs(lm_ModernC)
```

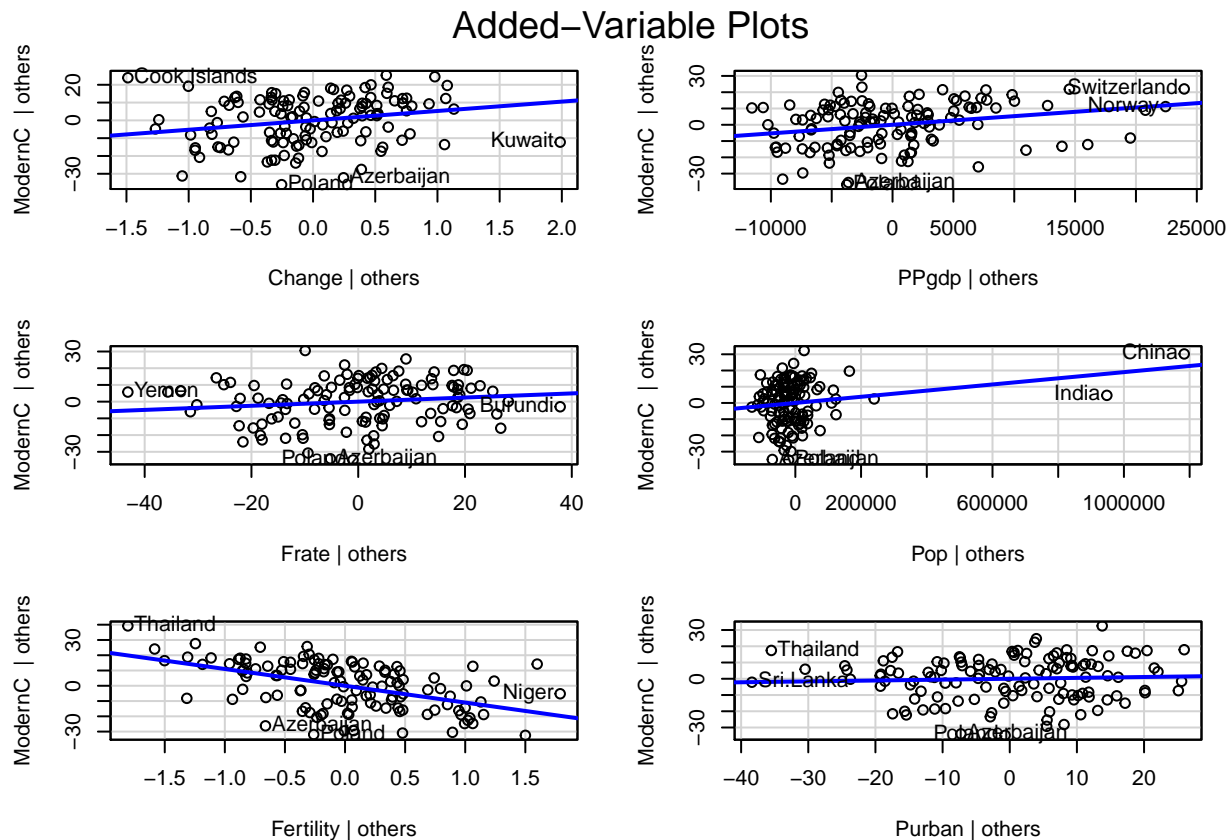
```
## [1] 125
```

- 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?

```
#Added Variable Plot
```

```
avPlots(LM_ModernC)
```



After visual inspection, I would like to suggest transforming the predictors Pop and PPgdp. The predictor “Pop” is too clustered around one section of the line, which I believe would become more spread out after a transformation. The predictor “PPgdp,” on the other hand, has a large x-scale, which hints that transformation would significantly change the plot.

As for the question about any of localities being influential for any of the terms, I would say China and India look like influential localities in ModernC|Others ~ Pop|Others graph. The presence of the two points is judged - subjectively - to be creating a positive relationship between the response and the predictor.

- 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.

```
#Transformed Variables
```

```
log_PPgdp = log(PPgdp)
```

```
log_Pop = log(Pop)
```

```
#New Data Frame
```

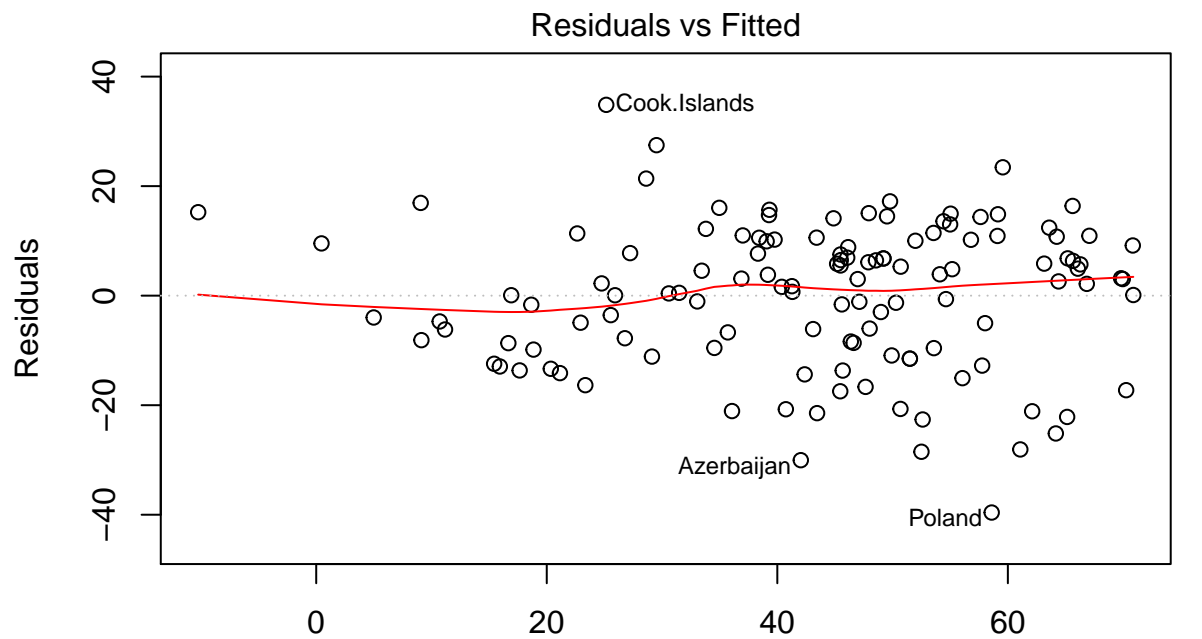
```
UN3_No_NA_Log = UN3_No_NA %>%
```

```
  rownames_to_column("Country") %>%
```

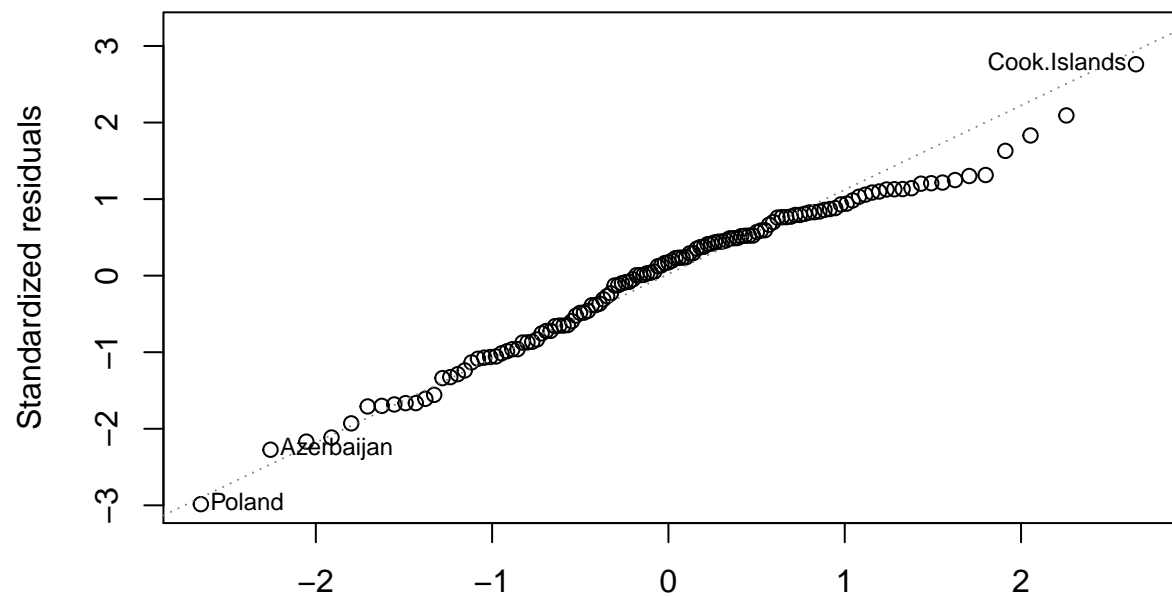
```
  mutate("log_PPgdp" = log(PPgdp), "log_Pop" = log(Pop)) %>%
```

```
  column_to_rownames("Country")
```

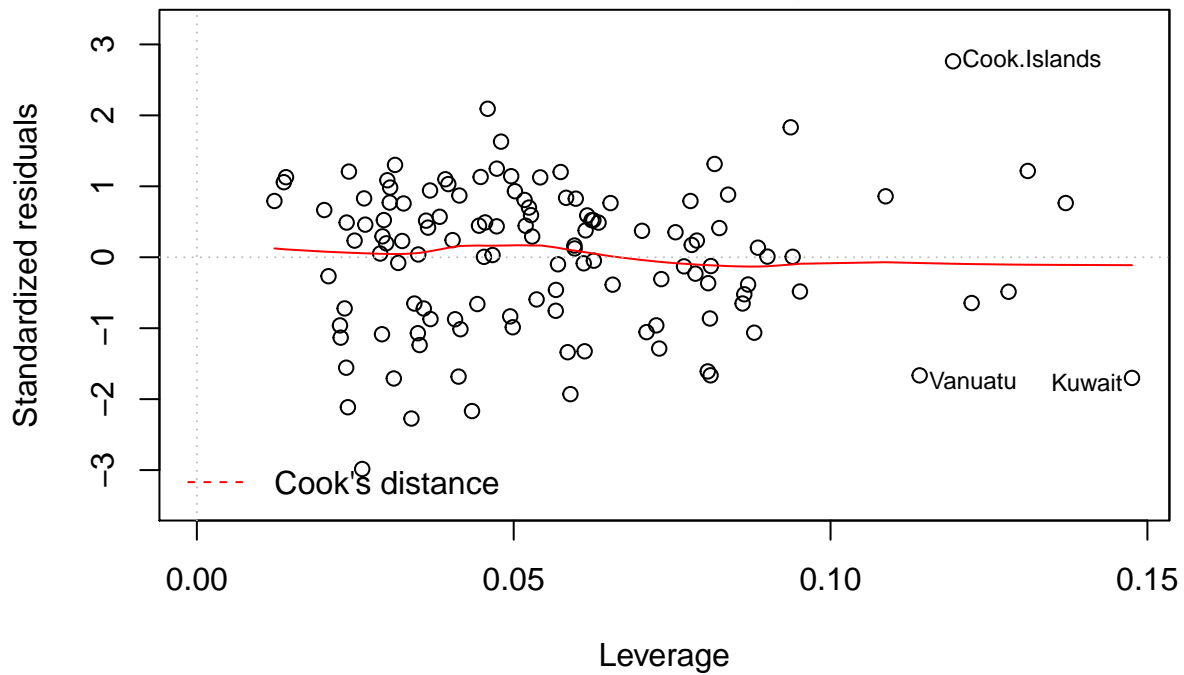
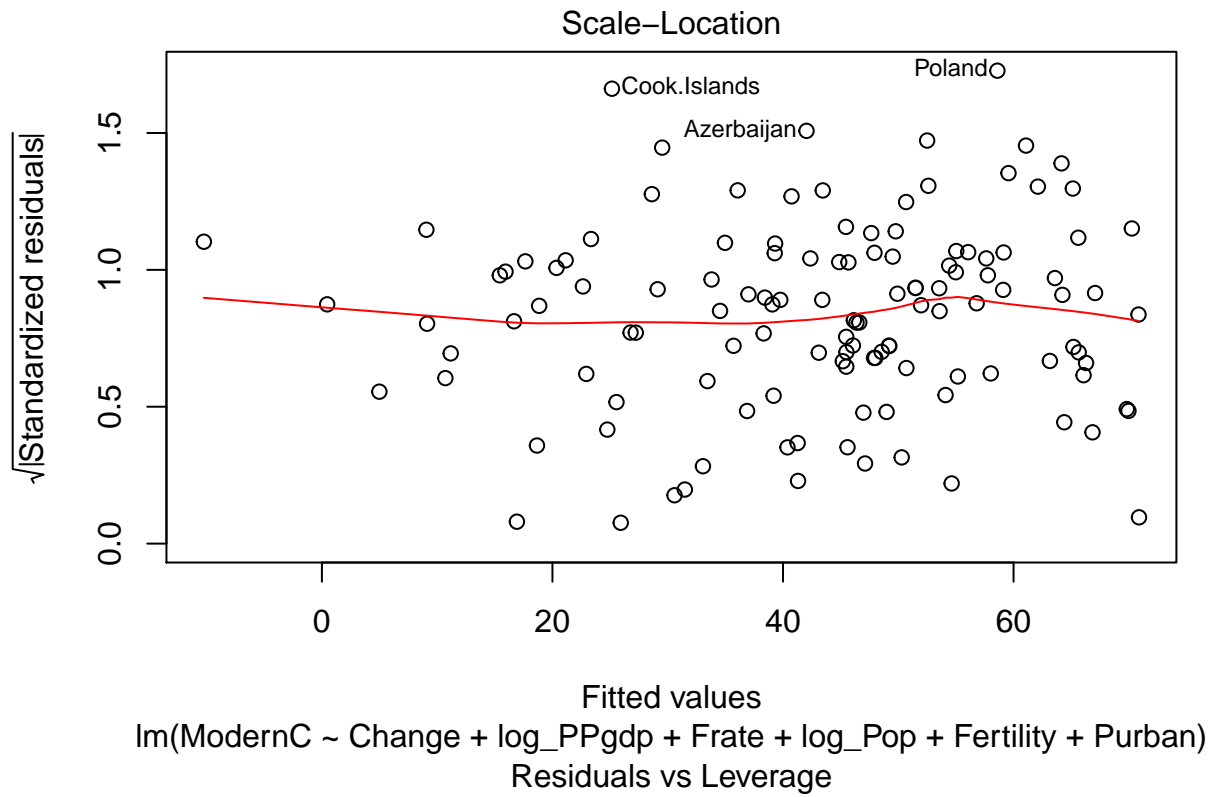
```
LM_ModernC_Transformed = lm(ModernC ~ Change + log_PPgdp + Frate + log_Pop + Fertility + Purban, data =  
plot(LM_ModernC_Transformed)
```



Fitted values
lm(ModernC ~ Change + log_PPgdp + Frate + log_Pop + Fertility + Purban)
Normal Q-Q

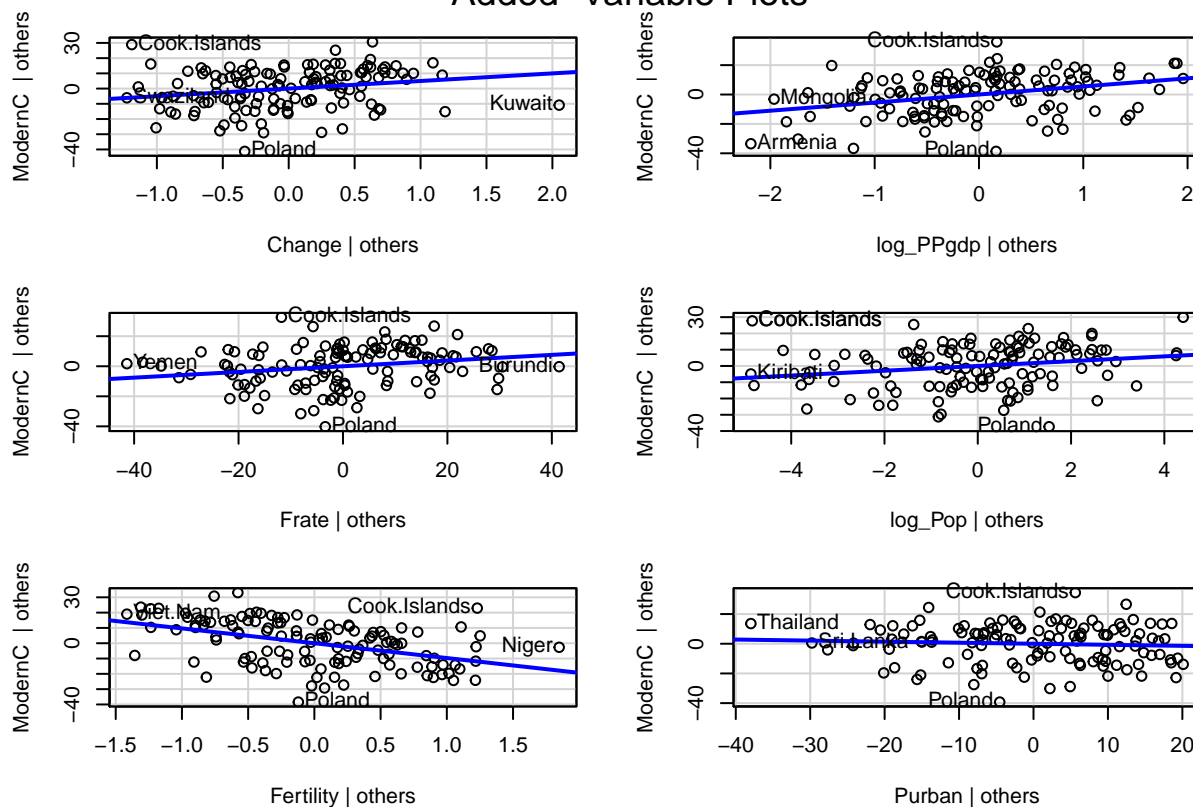


Theoretical Quantiles
lm(ModernC ~ Change + log_PPgdp + Frate + log_Pop + Fertility + Purban)

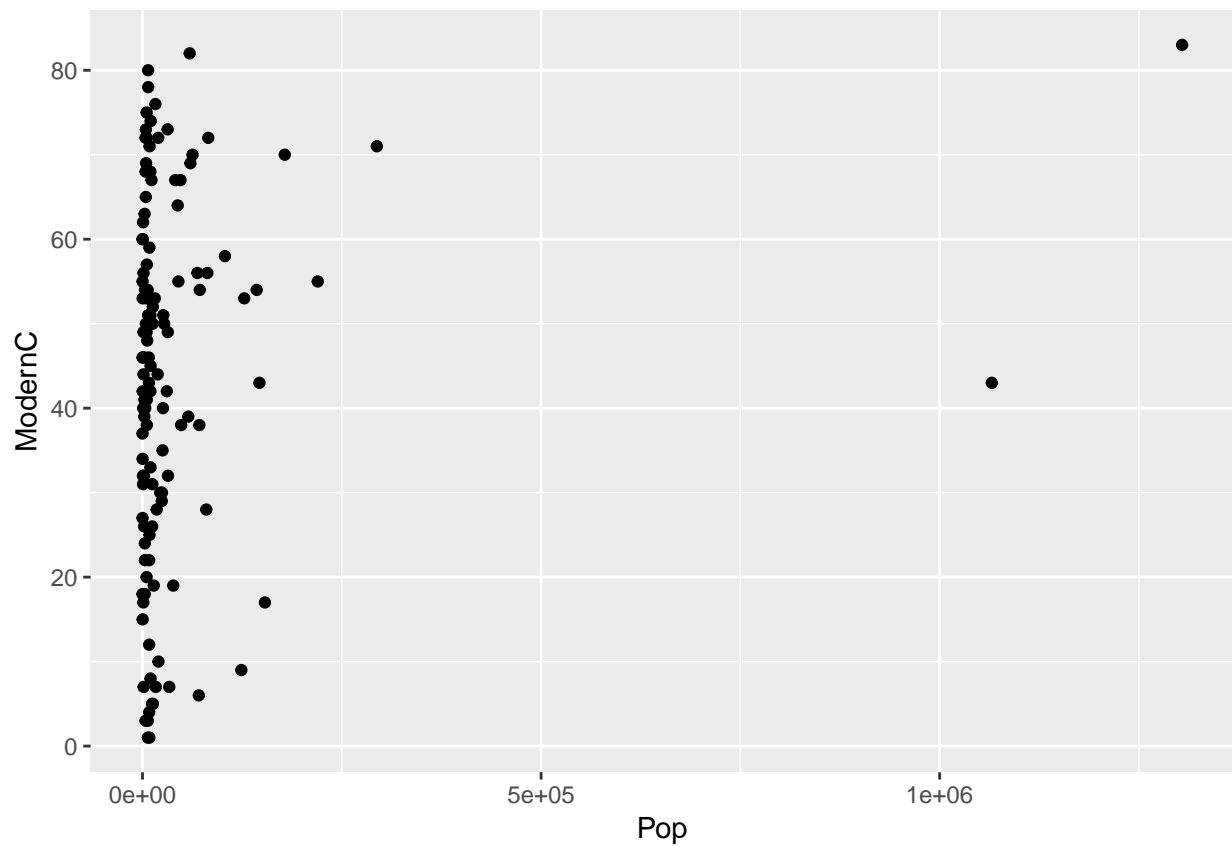


```
avPlots(LM_ModernC_Transformed)
```

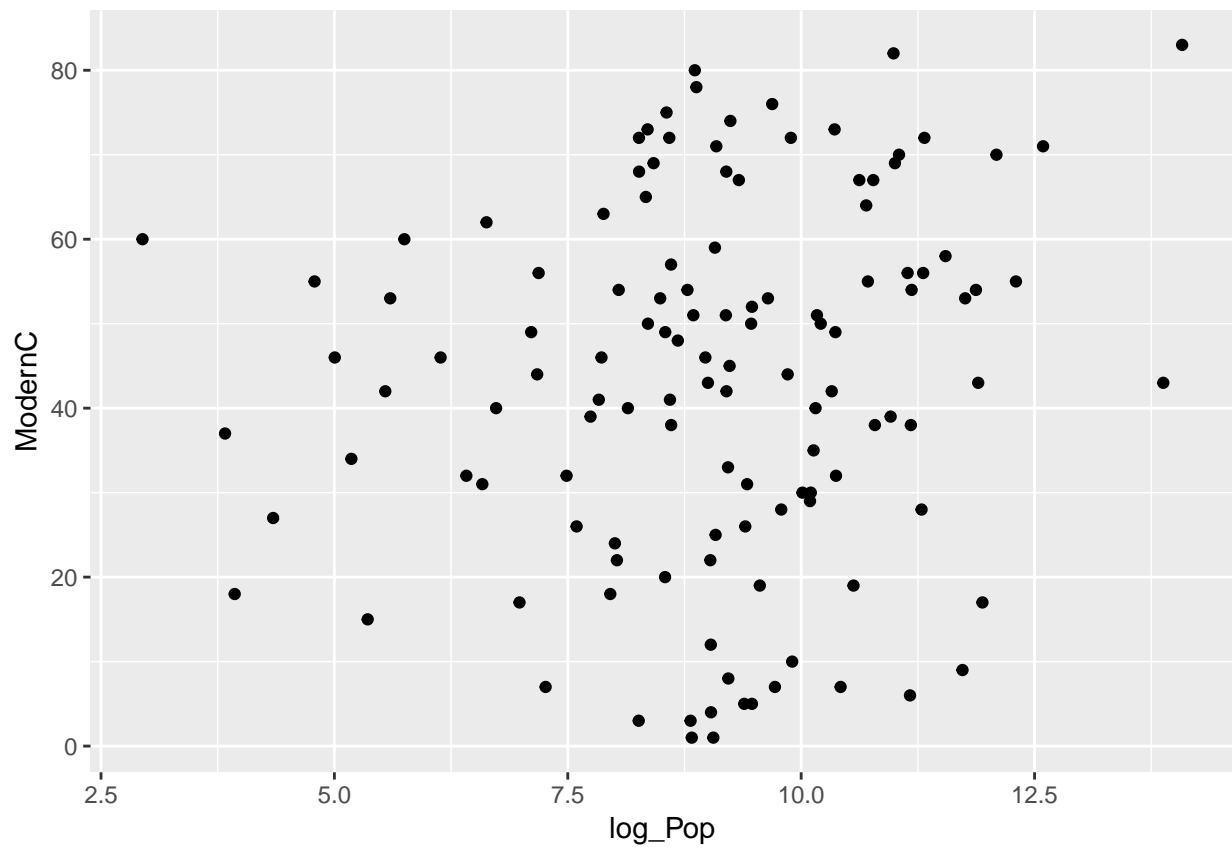
Added-Variable Plots



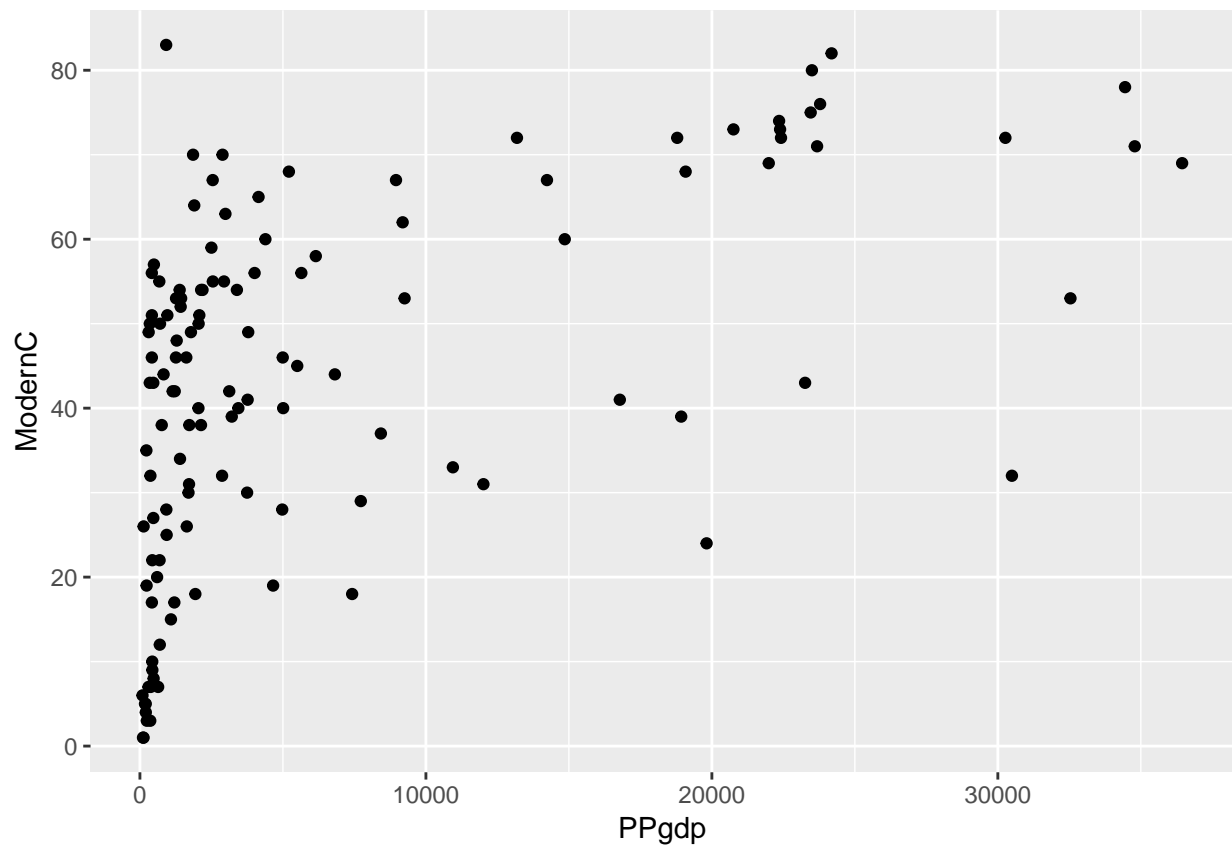
```
ggplot(UN3_No_NA, aes(Pop, ModernC)) + geom_point()
```



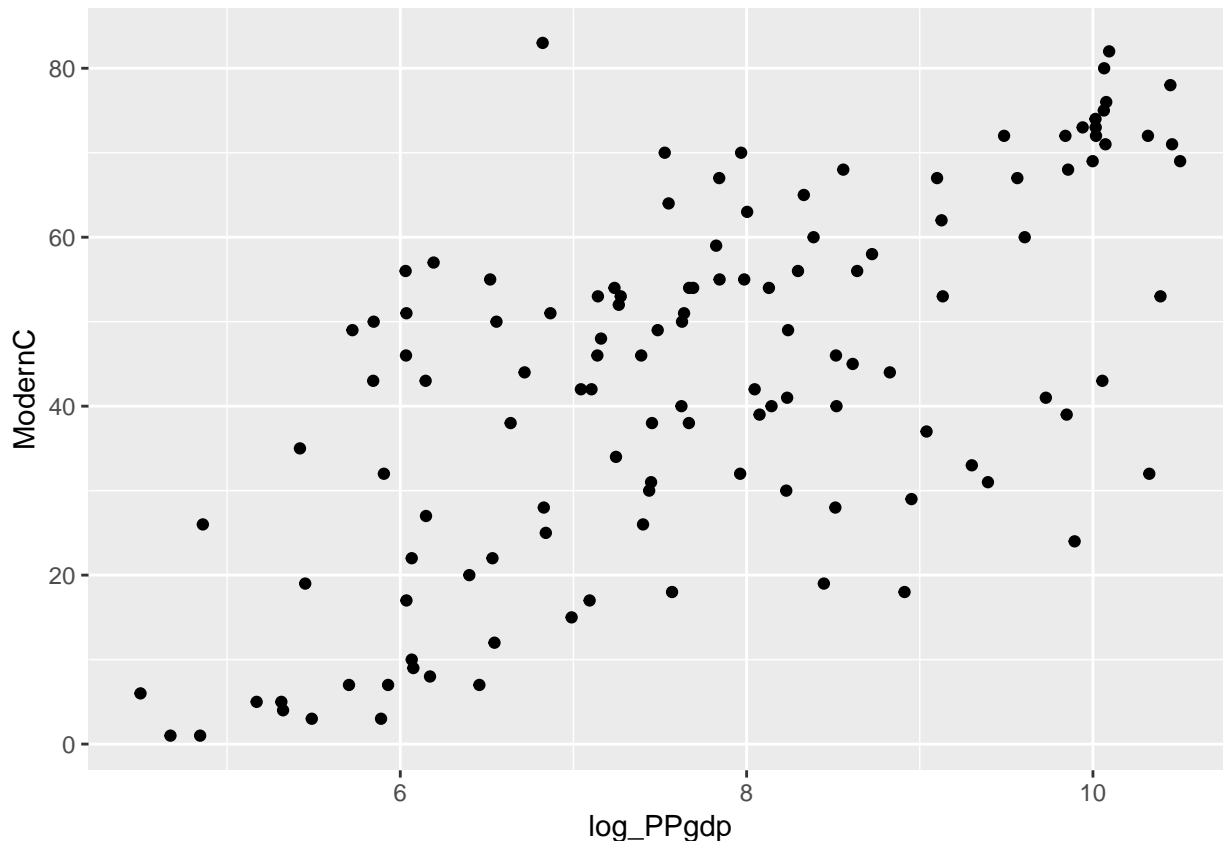
```
ggplot(UN3_No_NA, aes(log_Pop, ModernC)) + geom_point()
```



```
ggplot(UN3_No_NA, aes(PPgdp, ModernC)) + geom_point()
```



```
ggplot(UN3_No_NA, aes(log_PPgdp, ModernC)) + geom_point()
```

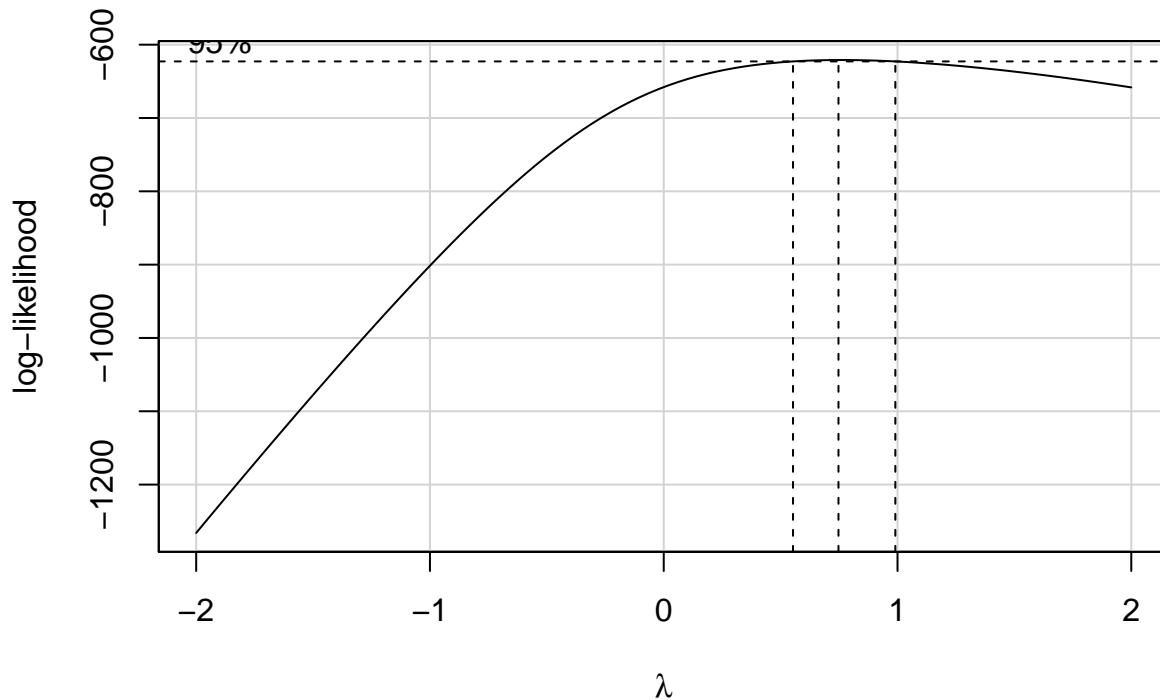
```
#BoxTidwell
boxTidwell(ModernC ~ Pop + PPgdp, other.x = ~ Change + Frate + Fertility + Purban, data = UN3_No_NA)

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

After running boxTidwell for predictors Pop and PPgdp, I obtained outputs that suggested statistical insignificance of doing predictors' transformations. However, to not to overlook suggestions about transformations I had made in #3 and #5, I log-transform the predictors, to see if I really end up with the outputs provided by boxTidwell. The results were different from what boxTidwell provided. Log transforming the predictors PPgdp and Pop actually altered the plots, providing avPlots with more linear trend, less pronounced residuals vs. fitted plot, and less pronounced residuals vs. leverage plot. Since log-transformations ended up improving many aspects of the linear model, despite the insignificance message from boxTidwell, I concluded to log-transform PPgdp and Pop.

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

```
#Plot for BoxCox
boxCox(LM_ModernC_Transformed)
```



```
powerTransform(LM_ModernC_Transformed)
```

```
## Estimated transformation parameter
##      Y1
## 0.7585897
```

BoxCox transformation suggests that there does not need to be a transformation done on the response variable, because the interval for lambda includes 1. This indicates that the power on the response is not significantly different from 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.

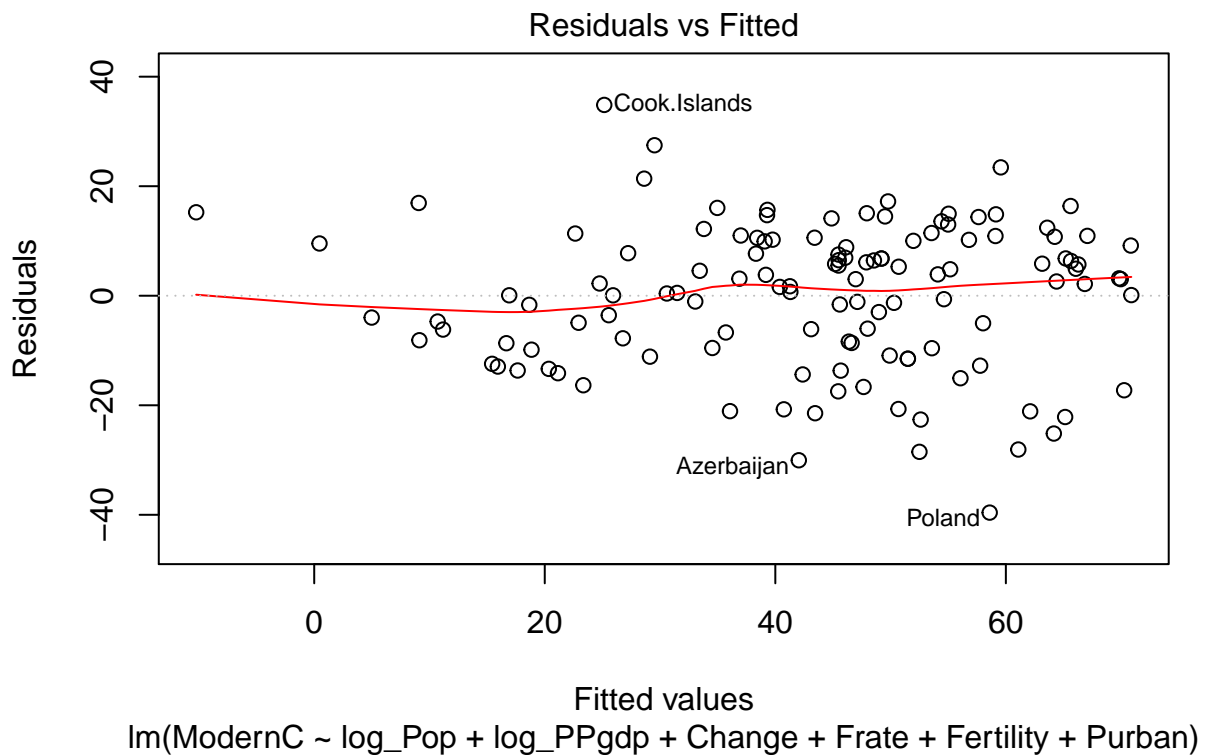
```
#Regression Using the Transformed Variables
```

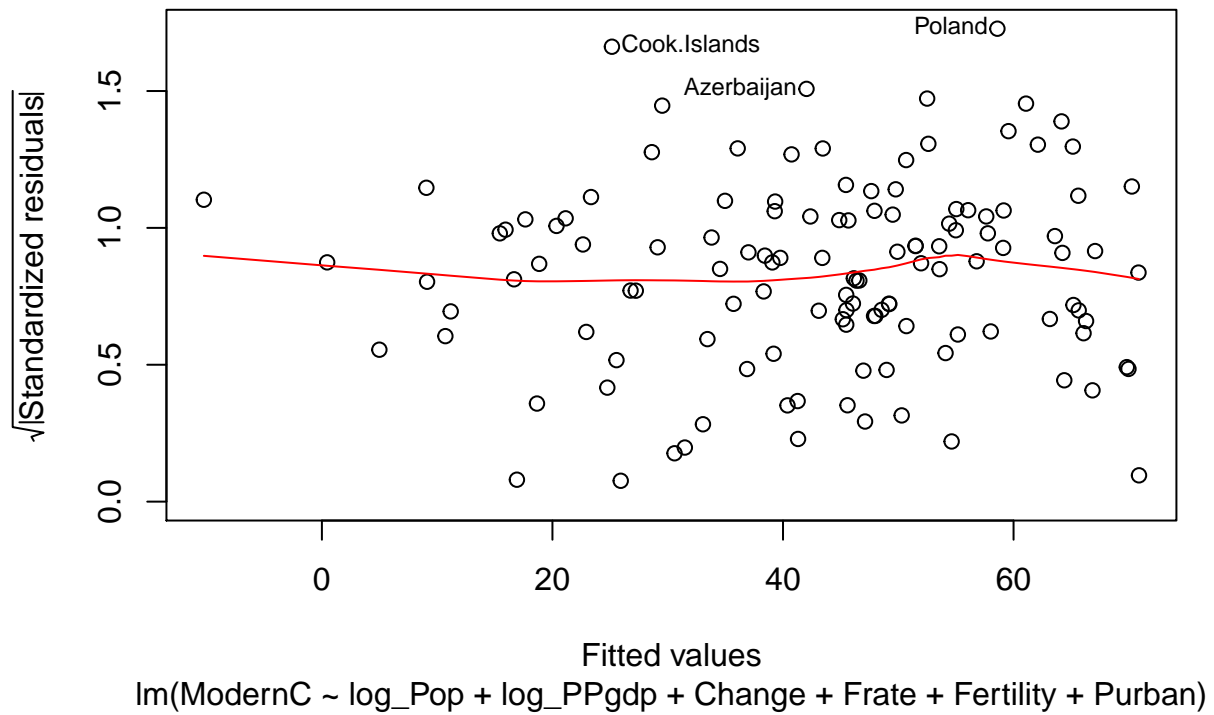
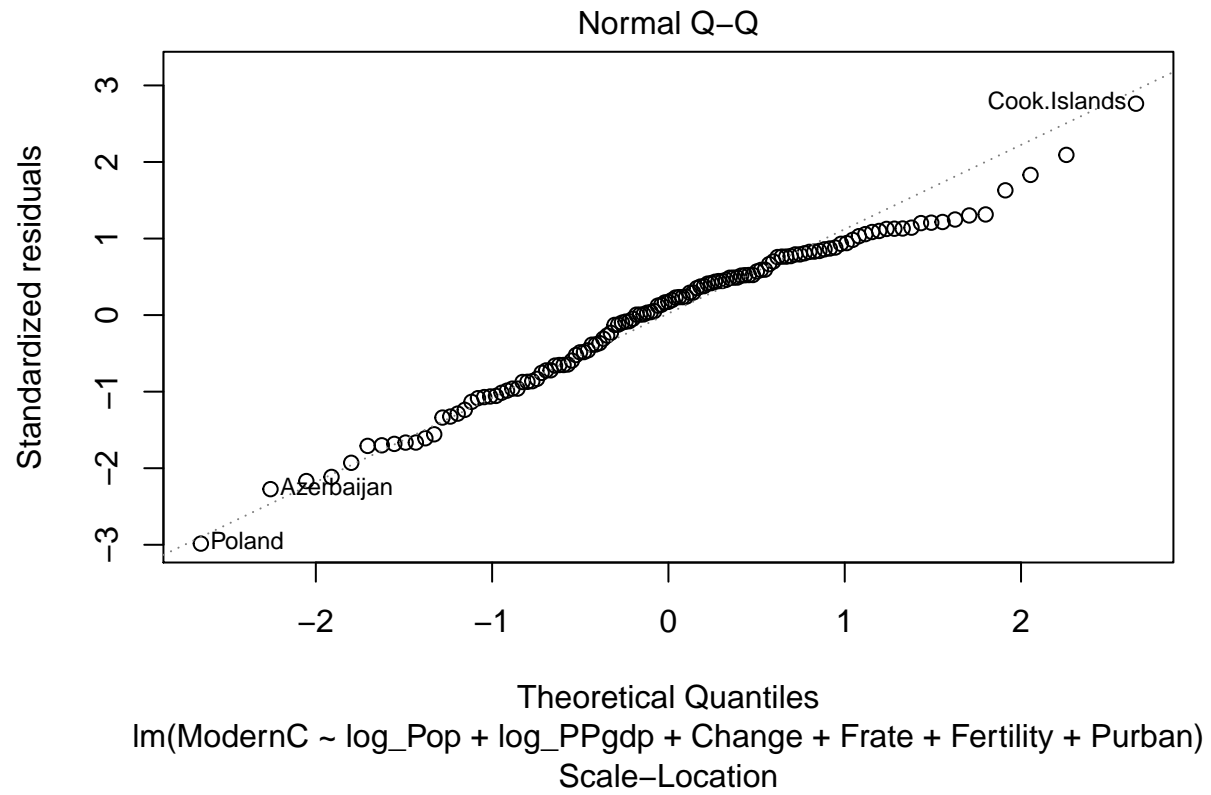
```
LM_ModernC_Transformed2 = lm(ModernC ~ log_Pop + log_PPgdp + Change + Frate + Fertility + Purban, data = UN3_No_NA_Log)
```

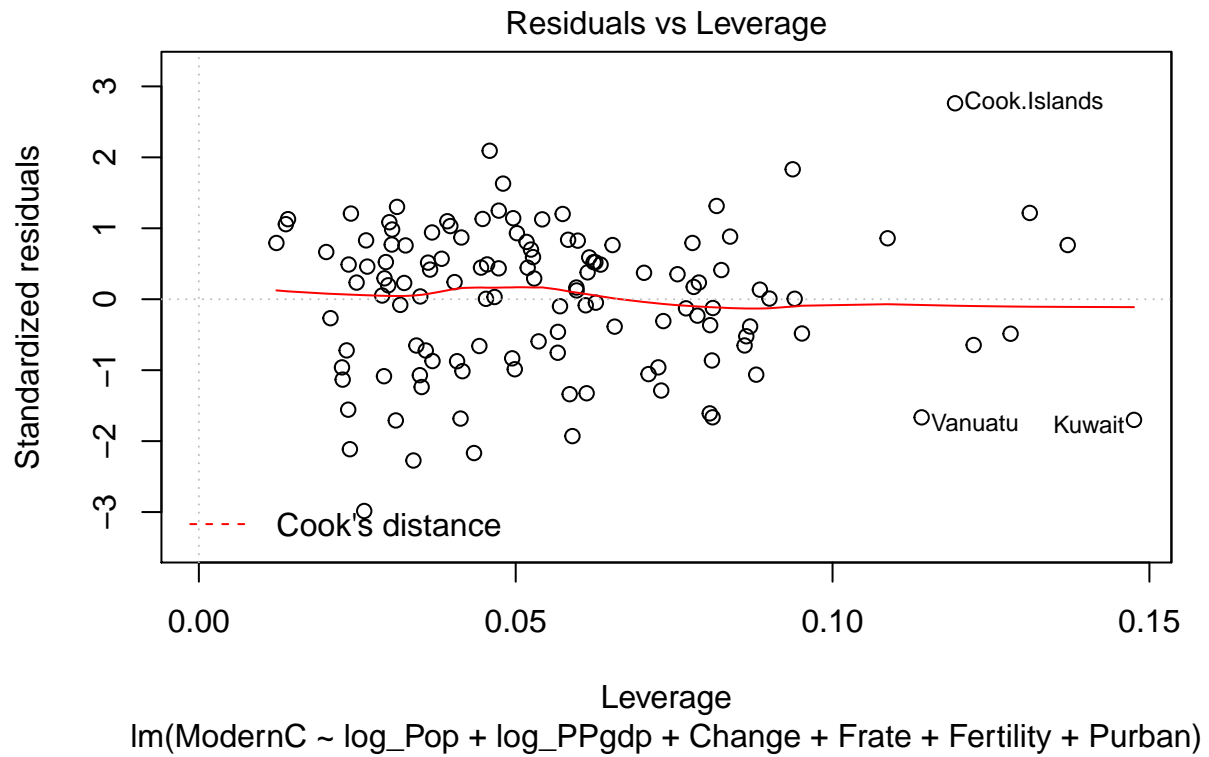
```
summary(LM_ModernC_Transformed2)
```

```
##
## Call:
## lm(formula = ModernC ~ log_Pop + log_PPgdp + Change + Frate +
##      Fertility + Purban, data = UN3_No_NA_Log)
##
## 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
## log_Pop       1.47207     0.62875   2.341 0.020897 *
## log_PPgdp     5.50728     1.40505   3.920 0.000149 ***
## Change        4.99296     2.07709   2.404 0.017781 *
```

```
## Frate      0.18939    0.07711    2.456 0.015500 *
## Fertility  -9.67594    1.76561   -5.480 2.44e-07 ***
## Purban     -0.07077    0.09760   -0.725 0.469829
## ---
## 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
#Residual Plot
plot(LM_ModernC_Transformed2)
```

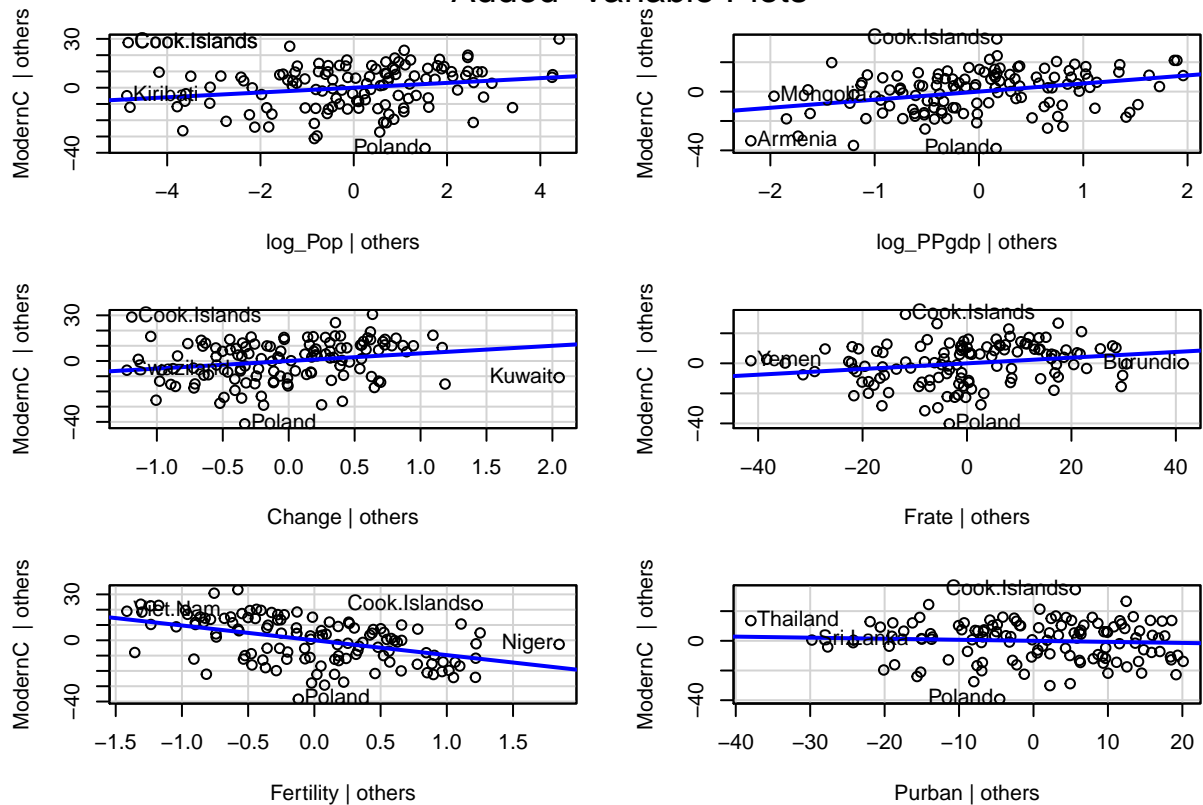






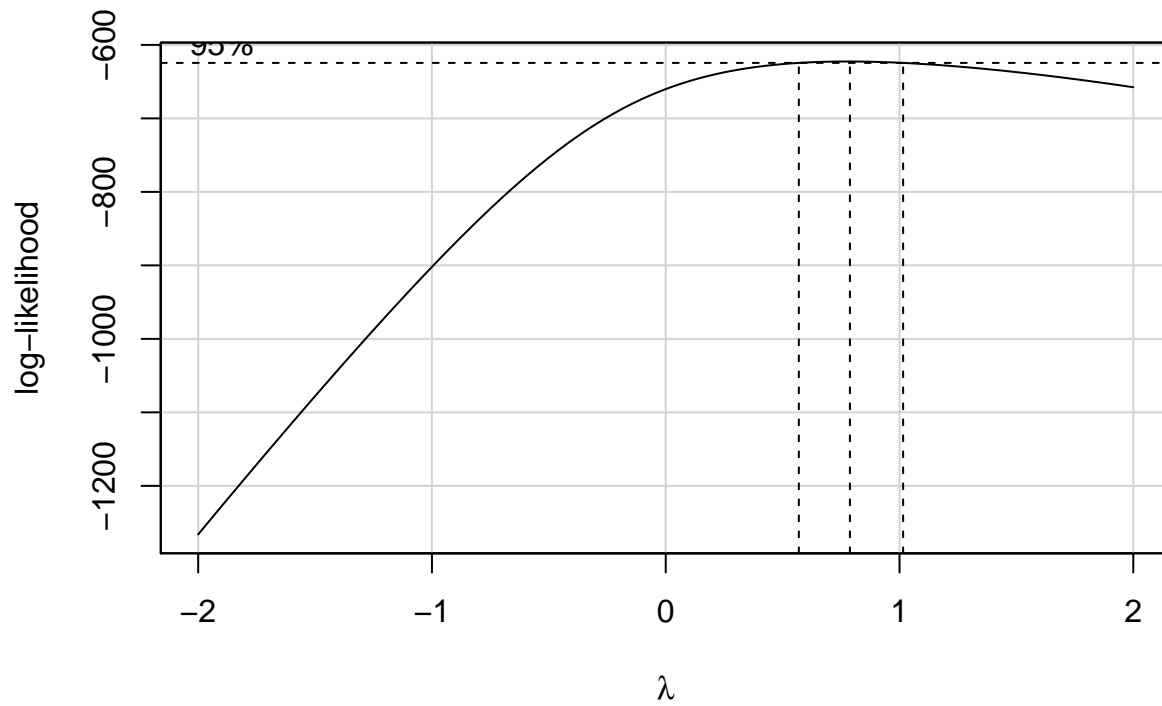
```
#Added Variables Plot
avPlots(LM_ModernC_Transformed2)
```

Added-Variable Plots



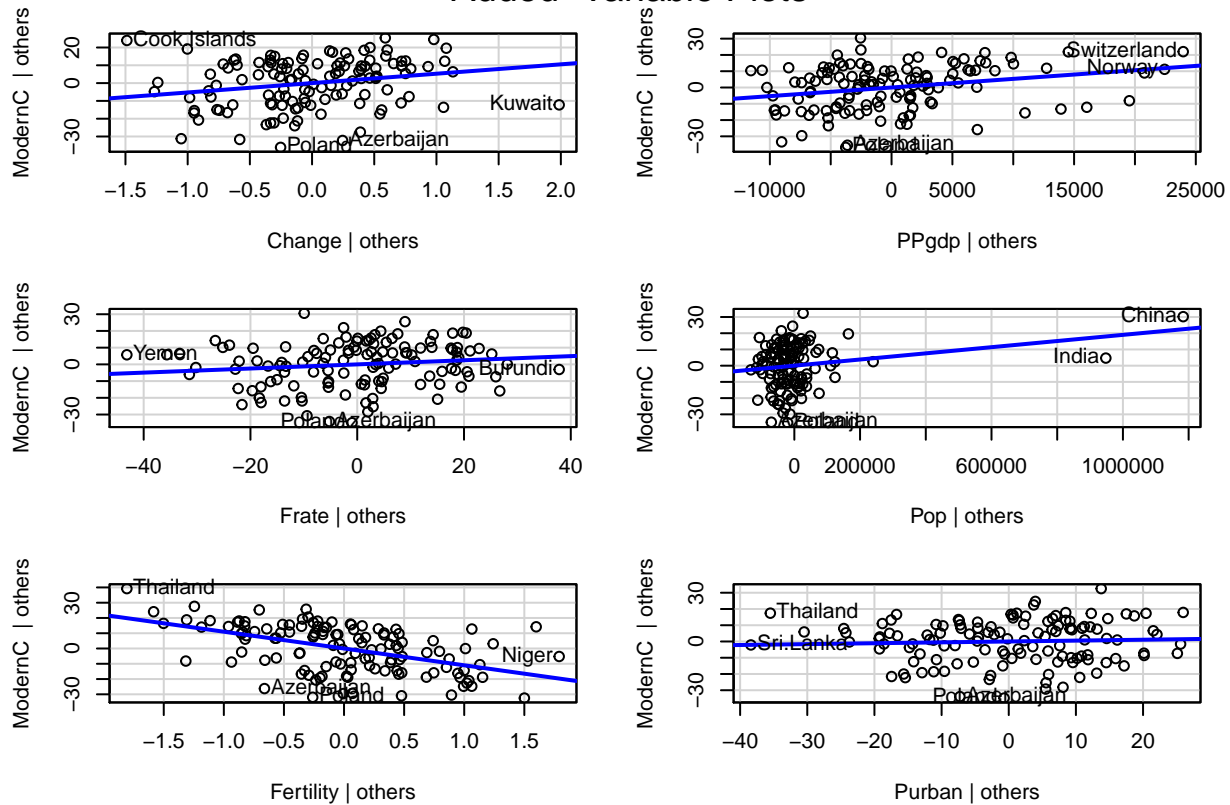
9. Start by finding the best transformation of the response and then find transformations of the predictors. Do you end up with a different model than in 8?

```
boxCox(LM_ModernC)
```



```
avPlots(LM_ModernC)
```

Added-Variable Plots



```
boxTidwell(ModernC ~ Pop + PPgdp, other.x = ~ Change + Frate + Purban + Fertility, data = UN3_No_NA)
```

```
##      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.

```
#Bonferroni Correction
pval = 2*(1 - pt(abs(rstudent(LM_ModernC_Transformed)), LM_ModernC_Transformed$df - 1))
rownames(UN3_No_NA_Log)[pval < .05/nrow(UN3_No_NA_Log)]
```

```
## character(0)
```

According to Bonferroni correction, there are no outliers in the data and therefore no data need to be removed.

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!

```
#Multiple Regression with Purban Removed
lm3 = lm(ModernC ~ Change + log_Pop + log_PPgdp + Frate + Fertility, data = UN3_No_NA_Log)
```

```
#Justification of Removing Purban
anova(lm3, LM_ModernC_Transformed2)
```

```
## Analysis of Variance Table
##
## Model 1: ModernC ~ Change + log_Pop + log_PPgdp + Frate + Fertility
## Model 2: ModernC ~ log_Pop + log_PPgdp + Change + Frate + Fertility +
##      Purban
##      Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1      119 21420
## 2      118 21325   1    95.016 0.5258 0.4698

confint(lm3)

##              2.5 %      97.5 %
## (Intercept) -24.56895012 32.7731131
## Change      0.67272868  8.7227854
## log_Pop     0.20155957  2.6808895
## log_PPgdp   2.71662357  7.0021013
## Frate       0.04969767  0.3493943
## Fertility   -12.59507558 -5.9617672

(newconfint.lm3 = data.frame(confint(lm3)) %>%
  rownames_to_column("Predictors") %>%
  filter(!row.names(confint(lm3)) %in% c("log_Pop", "log_PPgdp")) %>%
  column_to_rownames("Predictors"))

##              X2.5..      X97.5..
## (Intercept) -24.56895012 32.7731131
## Change      0.67272868  8.7227854
## Frate       0.04969767  0.3493943
## Fertility   -12.59507558 -5.9617672

Confind.Table = rbind(newconfint.lm3, Pop = c(0, 0), PPgdp = c(0, 0))

kable(Confind.Table, format = "markdown")
```

	X2.5..	X97.5..
(Intercept)	-24.5689501	32.7731131
Change	0.6727287	8.7227854
Frate	0.0496977	0.3493943
Fertility	-12.5950756	-5.9617672
Pop	0.0000000	0.0000000
PPgdp	0.0000000	0.0000000

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.

```
#e_Y on e_X
e_Y = residuals(lm3)
e_X = residuals(lm(Purban ~ Change + log_Pop + log_PPgdp + Frate + Fertility, data = UN3_No_NA_Log))

#Slope Comparison
lm_e_Y_on_e_X = lm(e_Y ~ e_X)
LM_ModernC_Transformed2$coefficients["Purban"]

##      Purban
## -0.07076799

x = data.frame(e_X = lm(e_Y ~ e_X)$coefficients["e_X"], LM_ModernC_Transformed2$coefficients["Purban"],

kable(x, col.names = c("e_X", "Model"), format = "markdown")
```

	e_X	Model
Coefficients	-0.070768	-0.070768

```
#AVPlots
avPlots(lm_e_Y_on_e_X)
```

