HW2 STA521 Fall18

[Eric Su, es351, Eric-Su-2718]

Due September 23, 2018 5pm

Exploratory Data Analysis

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
```

The variables `Modernc`, `Change`, `PPgdp`, `Frate`, `Pop`, and `Fertility` have missing values. All variables are quantitative.

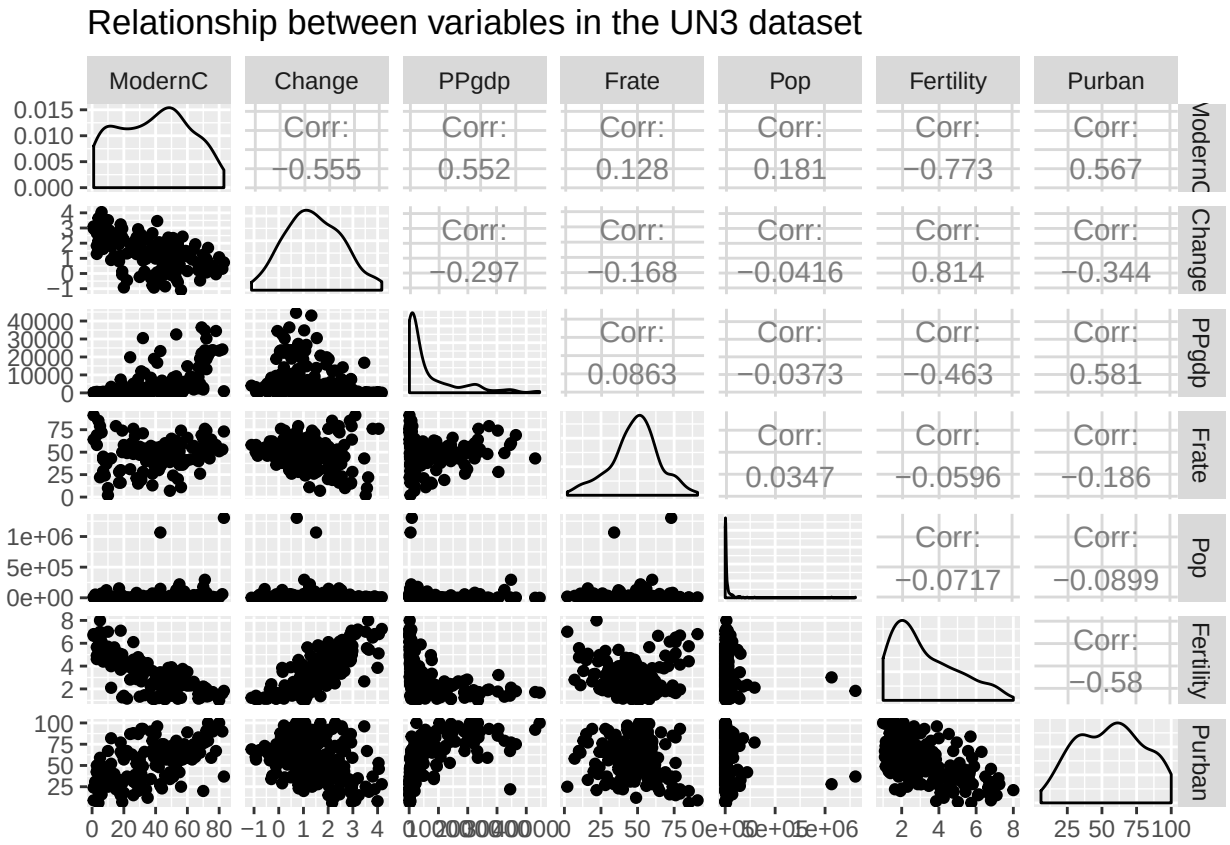
2. What is the mean and standard deviation of each quantitative predictor? Provide in a nicely formatted table.

```
mu_sd_data = rbind(apply(UN3[, 1:7], 2, mean, na.rm = TRUE), apply(UN3[, 1:7], 2, sd, na.rm = TRUE))
mu_sd_data = t(mu_sd_data)
colnames(mu_sd_data) = c("Mean", "Standard deviation")
library(knitr)
kable(mu_sd_data, digits = 2)
```

	Mean	Standard deviation
ModernC	38.72	22.64
Change	1.42	1.13
PPgdp	6527.39	9325.19
Frate	48.31	16.53
Pop	30281.87	120676.69
Fertility	3.21	1.71
Purban	56.20	24.11

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?

```
library(GGally)
ggpairs(UN3, title = "Relationship between variables in the UN3 dataset")
```

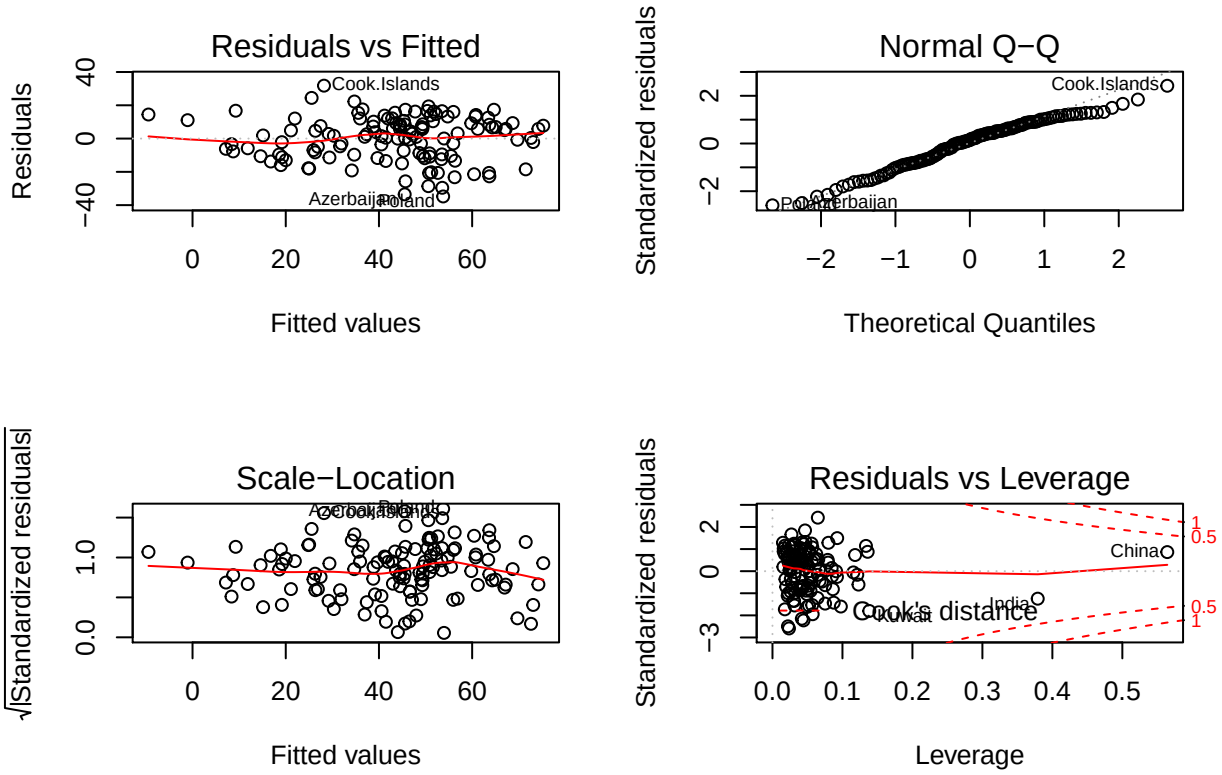


Based on the scatter plots, the variables `Change`, `Fertility` and `Purban` seem to have linear relationships with `ModernC`. On the other hand, `ModernC` has non-linear relationships with variables `PPgdp`, `Frate` and `Pop` and thus may need transformation. There also appears to be outliers as can be seen in the scatter plots involving `Pop`. Two observations (*China*‘ and *India*) are significantly away from others.

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?

```
lm_no_trans = lm(ModernC ~ ., data = UN3)
par(mfrow = c(2, 2))
plot(lm_no_trans, ask = FALSE)
```



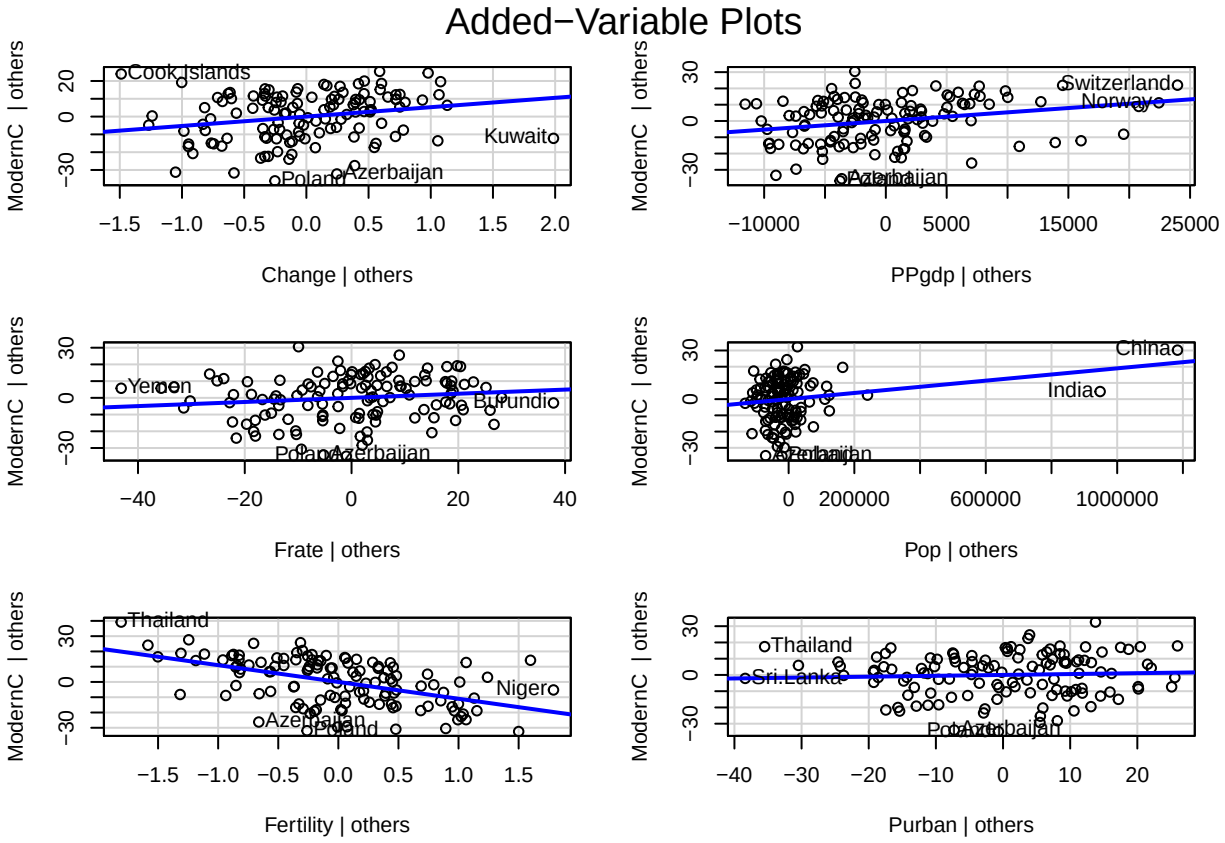

```
summary(lm_no_trans)
```

```
##
## 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
```

According to the diagnostic plots, we don't think there is a serious problem regarding unequal variance. However, observations `Cook Islands` and `Poland` could be outliers since both have large residuals and seem to be far from the theoretical normal quantiles. 85 observations were deleted due to having missing values, thus only 125 observations were used in this 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?

```
avPlots(lm_no_trans)
```



It seems that `PPgdp` and `Pop` need to be transfomed.

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.

```
boxTidwell(ModernC ~ PPgdp + Pop, ~ Change + Frate + Fertility + Purban, data = UN3)
```

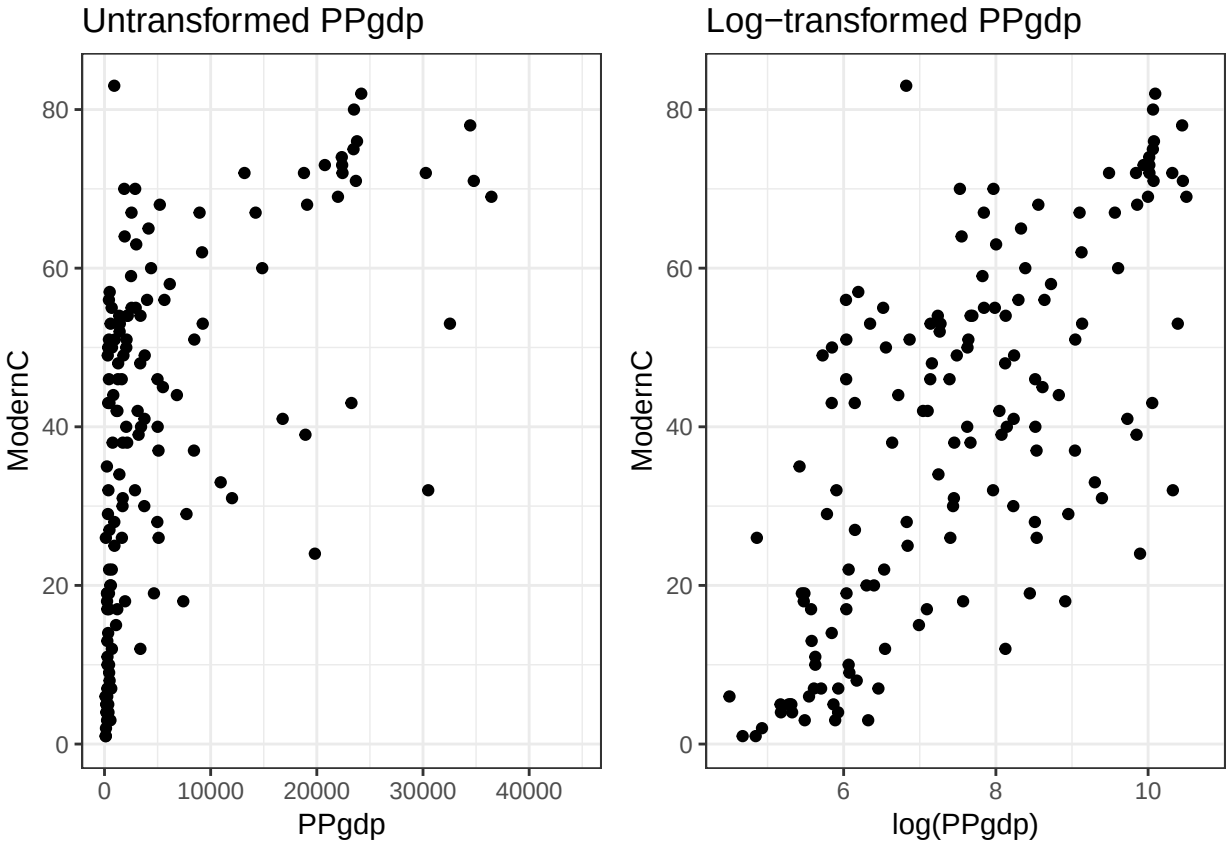
```
##       MLE of lambda Score Statistic (z) Pr(>|z|)
## PPgdp      -0.12921             -1.1410   0.2539
## Pop         0.40749             -0.7874   0.4310
##
## iterations =  4
```

According to the Box-Tidwell test, we cannot reject $H_0 : \lambda = 1$ and thus transformation is not needed for both `PPgdp` and `Pop`. However, we should also examine the scatter plots of `PPgdp` and `Pop` against `ModernC` to see if transformations make sense using graphical methods.

Below are scatter plots of `PPgdp` against `ModernC` with `PPgdp` untransformed and log-transformed.

```
library(ggplot2)
library(gridExtra)
PPgdp_p1 = ggplot(UN3, aes(x = PPgdp, y = ModernC))+
  geom_point()+
  theme_bw()+
  labs(title = "Untransformed PPgdp")

PPgdp_p2 = ggplot(UN3, aes(x = log(PPgdp), y = ModernC))+
  geom_point()+
  theme_bw()+
  labs(title = "Log-transformed PPgdp")
grid.arrange(PPgdp_p1, PPgdp_p2, ncol = 2)
```



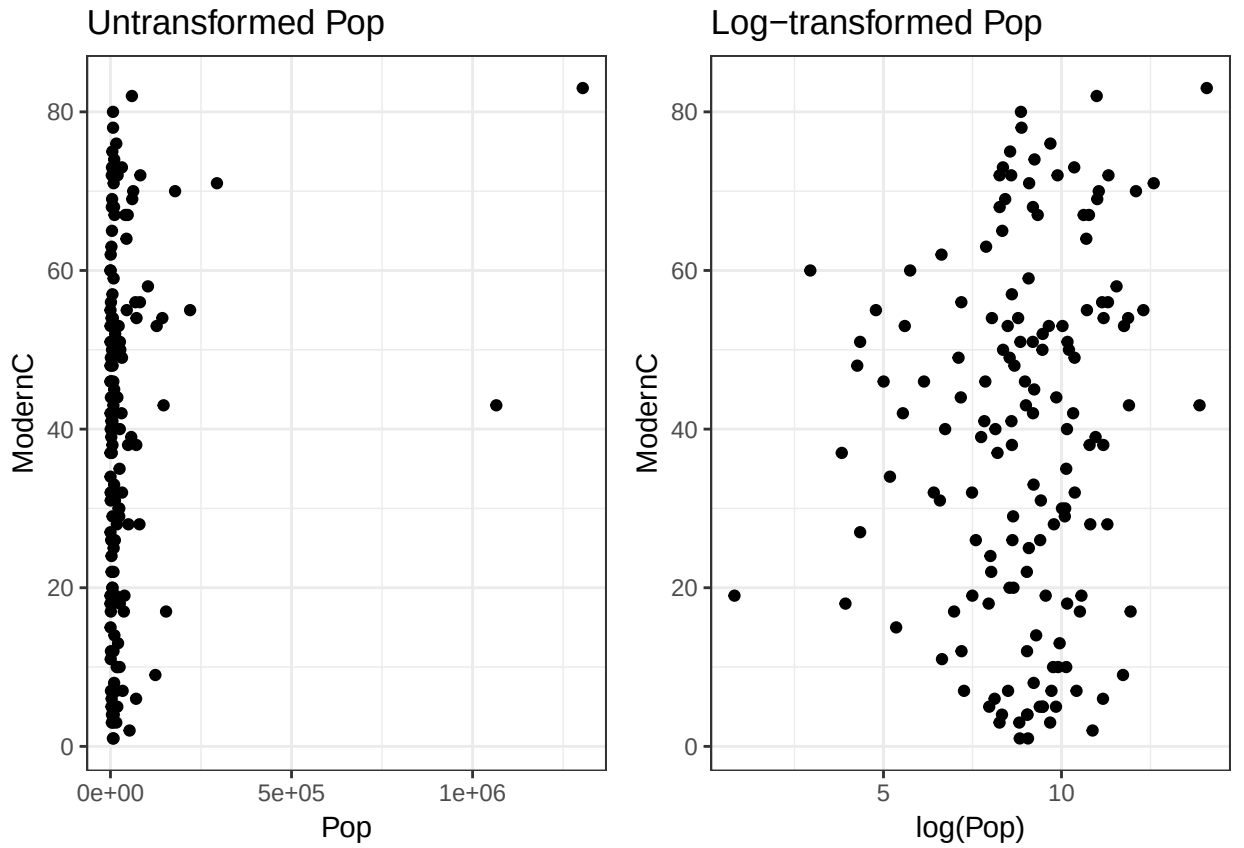
We can clearly see that the relationship between `log(PPgdp)` and `ModernC` is close to linear and thus will still apply the log transformation to `PPgdp` in our model.

Next we look at scatter plots of `Pop` against `ModernC` with `Pop` untransformed and log-transformed.

```
Pop_p1 = ggplot(UN3, aes(x = Pop, y = ModernC))+
  geom_point()+
  theme_bw()+
  labs(title = "Untransformed Pop")

Pop_p2 = ggplot(UN3, aes(x = log(Pop), y = ModernC))+
  geom_point()+
  theme_bw()+
```

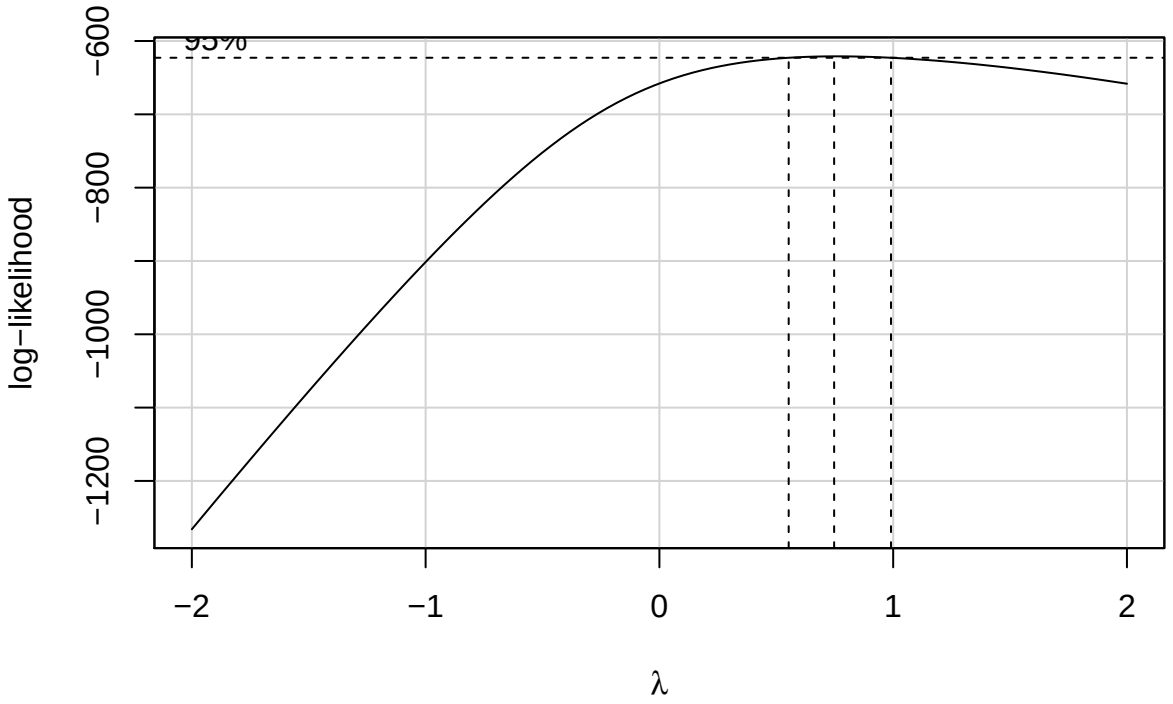
```
  labs(title = "Log-transformed Pop")
grid.arrange(Pop_p1, Pop_p2, ncol = 2)
```



Similarly, a log-transformation helps seperate countries with different population number and makes the relationship clearer. Therefore, we will also apply log-transformation to `Pop`.

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

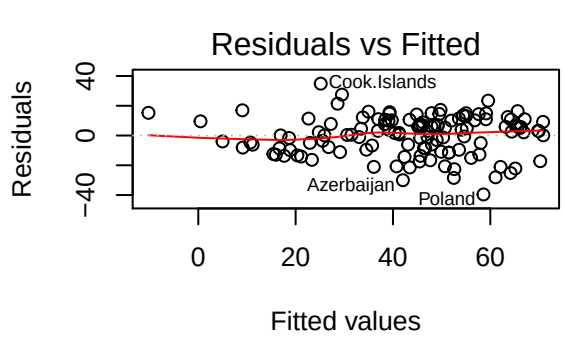
```
lm_pred_trans = lm(ModernC ~ Change + log(PPgdp) + Frate + log(Pop) + Fertility + Purban, data = UN3)
box_cox = boxCox(lm_pred_trans)
```

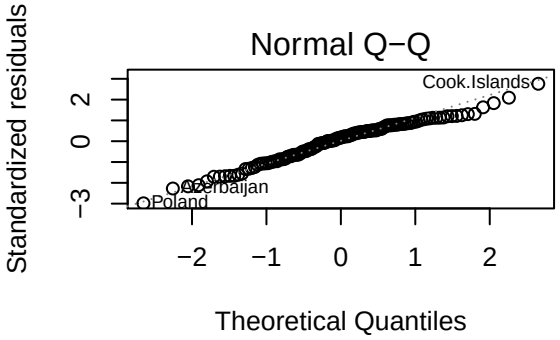


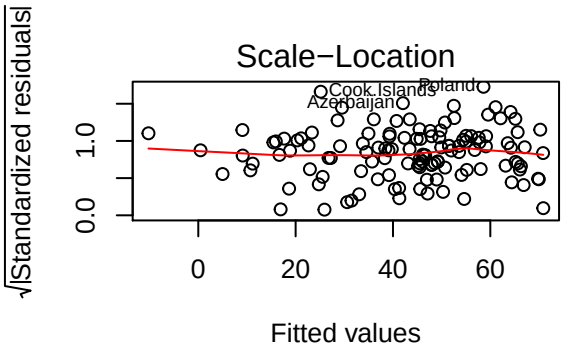
The optimal λ suggested by the Box-Cox method is very close to 1, therefore we conclude that no transformation is necessary for the response variable `ModernC`.

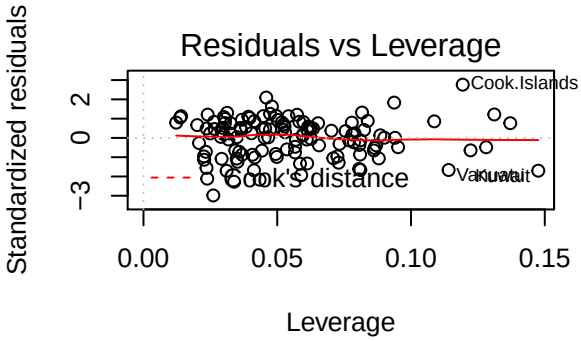
8. Fit the regression using the transformed variables. Provide residual plots and added variables plots and comment. If you feel that you need additional transformations of either the response or predictors, repeat any steps until you feel satisfied.

```
par(mfrow = c(2, 2))
plot(lm_pred_trans, ask = FALSE)
```

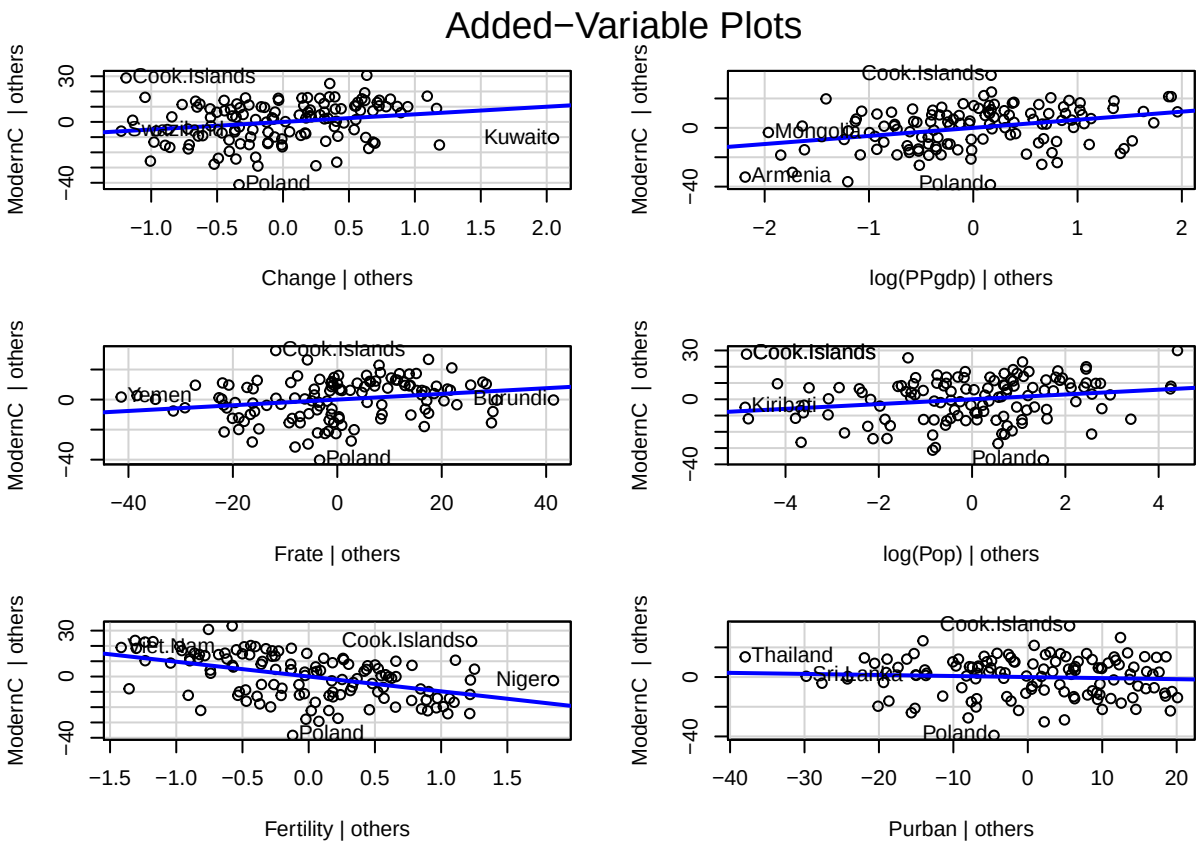








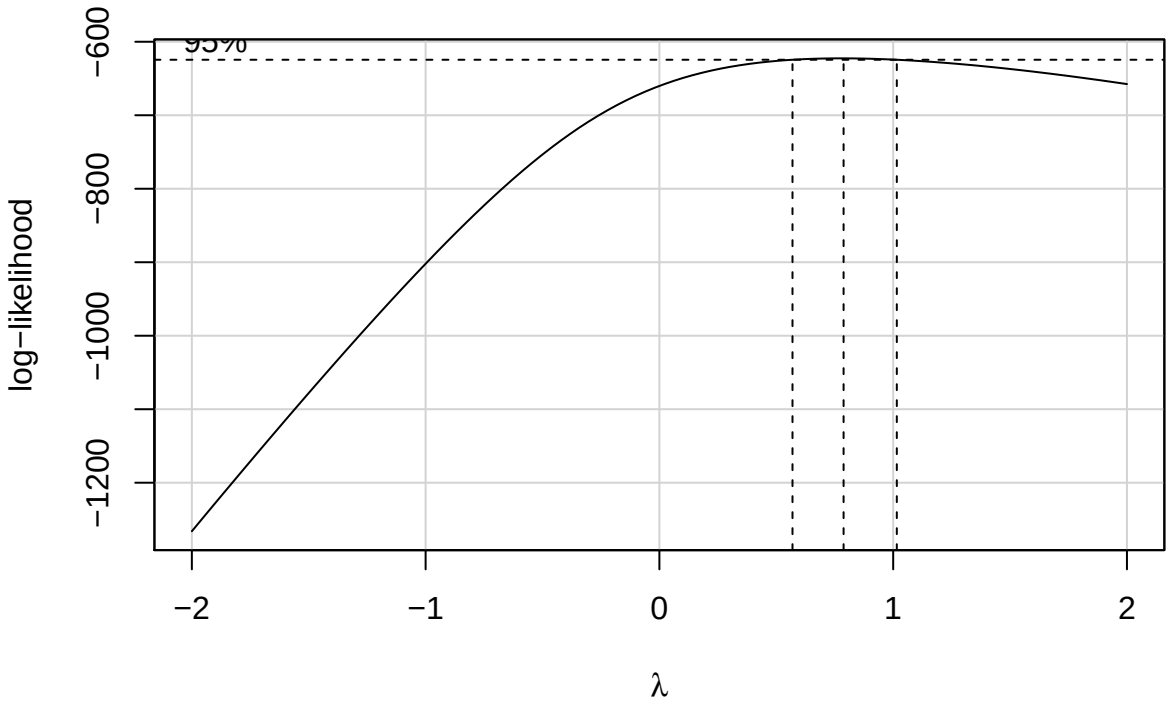

```
avPlots(lm_pred_trans)
```



Based on these plots, it looks like we have reduced the problem of non-normality and have also reduced the number of influential points in our model. However, some outliers still exist and we might want to remove them later.

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?

```
box_cox_untrans = boxCox(lm_no_trans)
```




```
MLE_lambda2 = box_cox_untrans$x[which(box_cox_untrans$y == max(box_cox_untrans$y))]
```

Based on the Box-Cox method, we conclude that the optimal λ is close to 1 and thus no tranforamtion is needed.

```
boxTidwell(ModernC ~ PPgdp + Pop, ~ Change + Frate + Fertility + Purban, data = UN3)
```

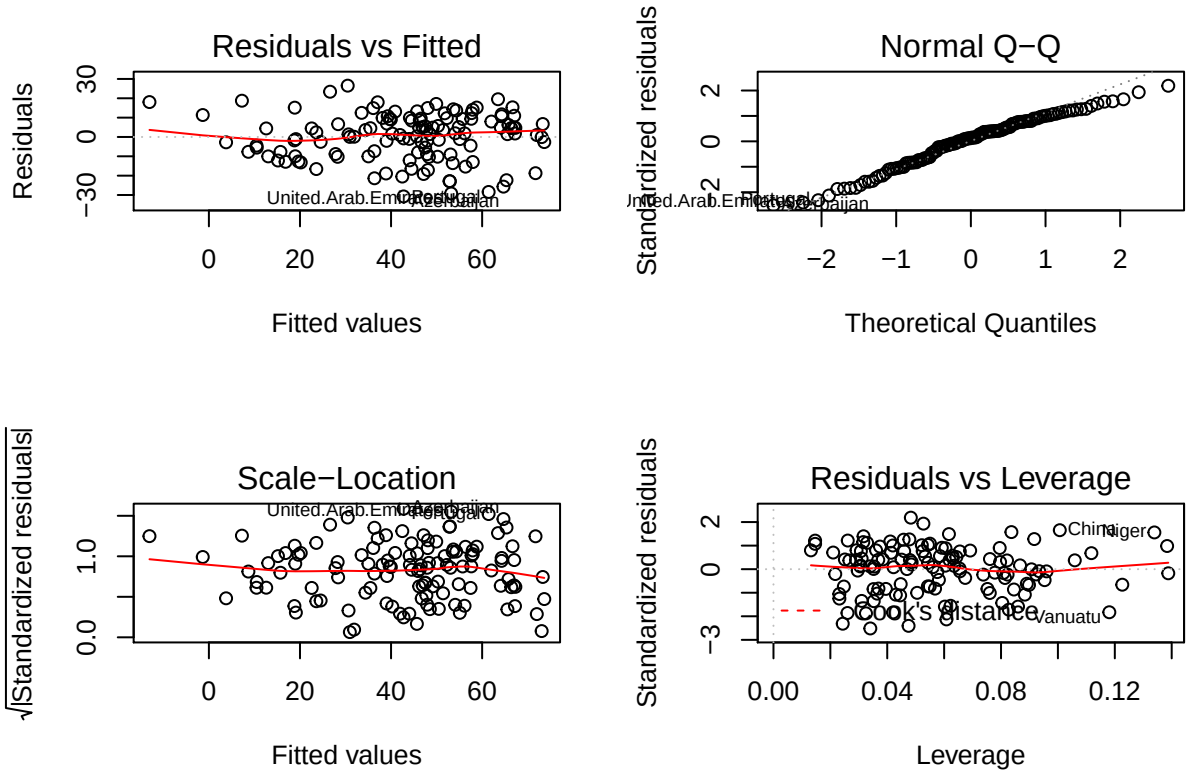
```
##       MLE of lambda Score Statistic (z) Pr(>|z|)
## PPgdp      -0.12921             -1.1410   0.2539
## Pop         0.40749             -0.7874   0.4310
##
## iterations =  4
```

The Box-Tidwell test suggest no transformation for the predictor variables. However, using similar arguments we used previously, we will still apply log-transformations to both `PPgdp` and `Pop`. As a result, we ended up to with the identical model we had before.

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.

Based on the regression diagnostic plots and the added-variable plots, observations `Cook.Islands`, `Poland` and `Kuwait` seem to be outliers and we will refit our model after removing them.

```
lm_trans_rmout = lm(ModernC ~ Change + log(PPgdp) + Frate + log(Pop) + Fertility + Purban, data = UN3[-c
par(mfrow = c(2, 2))
plot(lm_trans_rmout, ask = FALSE)
```



Although we removed some outliers and influential points in our original model, new ones seem to pop up. If we keep removing outliers in our model, the same issue is likely to arise repeatedly. As the new outliers do not seem to cause new issues, we will stop removing them.

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!

```
conf_coef = confint(lm_trans_rmout)
conf_coef["log(PPgdp)", ] = conf_coef["log(PPgdp)", ] * log(110 / 100)
conf_coef["log(Pop)", ] = conf_coef["log(Pop)", ] * log(110 / 100)
rownames(conf_coef)[3] = "PPgdp (10% increase)"
rownames(conf_coef)[5] = "Pop (10% increase)"
kable(conf_coef, digits = 2, caption = "95% confidence interval of coefficients")
```

Table 2: 95% confidence interval of coefficients

	2.5 %	97.5 %
(Intercept)	-25.05	28.22
Change	2.96	11.18
PPgdp (10% increase)	0.28	0.77
Frate	0.07	0.36
Pop (10% increase)	0.07	0.30
Fertility	-14.56	-7.87
Purban	-0.26	0.10

	2.5 %	97.5 %

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.

```
library(stargazer)
stargazer(lm_trans_rmout, title = "Summary of final regression model", header = FALSE, type = "latex", s
```

Table 3: Summary of final regression model

	Dependent variable:
	ModernC
Change	7.069*** (2.073)
log(PPgdp)	5.498*** (1.300)
Frate	0.215*** (0.072)
log(Pop)	1.919*** (0.603)
Fertility	−11.216*** (1.689)
Purban	−0.079 (0.091)
Constant	1.584 (13.445)
Observations	122
R^2	0.683
Adjusted R^2	0.667
Residual Std. Error	12.433 (df = 115)
F Statistic	41.362*** (df = 6; 115)
Note:	*p<0.1; **p<0.05; ***p<0.01

As one can see from the summary table, all predictors except `Purban` 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.*

We have

$$\begin{aligned}
\bar{\hat{e}} &= \frac{\sum_{i=1}^n \hat{e}}{n} \\
\sum_{i=1}^n \hat{e} &= 1_n^T \vec{\hat{e}} \\
&= 1_n^T(Y - X\hat{\beta}) \\
&= 1_n^T(Y - X(X^TX)^{-1}X^TY) \\
&= 1_n^T(I_n - X(X^TX)^{-1}X^T)Y \\
&= 1_n^T(I_n - H)Y \\
&= 0
\end{aligned}$$

Thus

$$\begin{aligned}\bar{\hat{e}} &= \frac{\sum_{i=1}^{n} \hat{e}}{n} \\ &= \frac{0}{n} \\ &= 0\end{aligned}$$

The intercept in a linear regression is $\bar{Y} - \hat{\beta}\bar{X}$, and therefore the intercept in the added variable plot is given by $\bar{\hat{e}_y} - \hat{\beta}\bar{\hat{e}_x} = 0 - 0 = 0$

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.

We will construct a added variable plot for the predictor `Fertility`. The summary table is shown below.

```
UN3_noNA = na.omit(UN3)
e_Y = residuals(lm(ModernC ~ Change + log(PPgdp) + Frate + log(Pop) + Purban, data = UN3_noNA))
e_X = residuals(lm(Fertility ~ Change + log(PPgdp) + Frate + log(Pop) + Purban, data = UN3_noNA))
lm_av_ModerFert = lm(e_Y ~ e_X)
stargazer(lm_av_ModerFert, title = "Summary of regression from Ex.14", header = FALSE, type = "latex", s
```

Table 4: Summary of regression from Ex.14

	Dependent variable:
	e_Y
e_X	-9.676^{***} (1.729)
Constant	0.000 (1.178)
Observations	125
R^2	0.203
Adjusted R^2	0.196
Residual Std. Error	13.167 (df = 123)
F Statistic	31.305^{***} (df = 1; 123)
Note:	*p<0.1; **p<0.05; ***p<0.01

The coefficient of e_X is -9.6759, which is the same as the coefficient for `Fertility` that we got from Ex.10. The summary table for Ex.10 is shown below.

```
stargazer(lm_pred_trans, title = "Summary of regression model from Ex.10", header = FALSE, type = "late
```

Reference

1. Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables. R package version 5.2.2. https://CRAN.R-project.org/package=stargazer

Table 5: Summary of regression model from Ex.10

	Dependent variable:
	ModernC
Change	4.993** (2.077)
log(PPgdp)	5.507*** (1.405)
Frate	0.189** (0.077)
log(Pop)	1.472** (0.629)
Fertility	−9.676*** (1.766)
Purban	−0.071 (0.098)
Constant	4.115 (14.509)
Observations	125
R^2	0.626
Adjusted R^2	0.607
Residual Std. Error	13.443 (df = 118)
F Statistic	32.912*** (df = 6; 118)
Note:	*p<0.1; **p<0.05; ***p<0.01