HW2 STA521 Fall18

[(Henry) Yuren Zhou, yz482, HenryYurenZhou]

Due September 23, 2018 5pm

Backgound 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*

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                      
```

Therefore, 6 variables out of 7, "ModernC", "Change", "PPgdp", "Frate", "Pop" and "Fertility", have missing data.

Based on variable description of UN3 by `help(UN3)`, we see that all variables are quantitative.

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

```
library(knitr)
```

```
df <- data.frame(sapply(UN3, mean, na.rm=TRUE), sapply(UN3, sd, na.rm=TRUE))
kable(df, col.names = c("mean", "standard deviation"),
      caption = "mean and standard deviation of each quantitative predictor")
```

Table 1: mean and standard deviation of each quantitative predictor

	mean	standard deviation
ModernC	38.717105	2.263661e+01
Change	1.418373	1.133133e+00
PPgdp	6527.388060	9.325189e+03
Frate	48.305389	1.653245e+01
Pop	30281.871428	1.206767e+05
Fertility	3.214000	1.706918e+00
Purban	56.200000	2.410976e+01

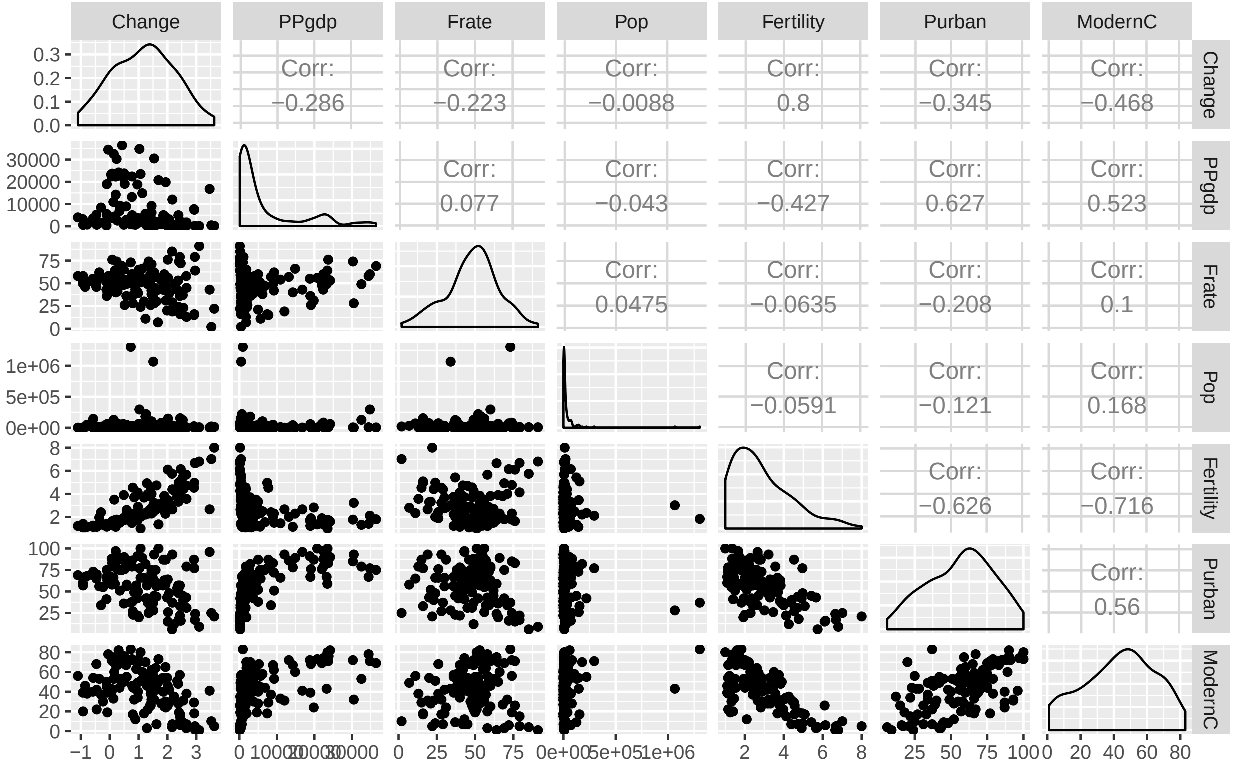

Figure 1: scatterplots for UN3 dataset

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?

```r
library(ggplot2)
library(GGally)
```

```r
ggpairs(na.omit(UN3)[, c(2:7, 1)])   # Fig.1
```

From Figure 1, we can observe that Change, Fertility and Purban seem to have approximately linear relationships with ModernC, while for other variables, the relationships appear to be non-linear. An approach to resolve non-linear relationship is to perform transformation. For PPgdp, log transformation could help distribute the data more evenly; Frate doesn't seem to have a clear relationship with ModernC, which is also shown in its correlation 0.1; Pop has two potential outliers (or high leverage points at least) China and India, and using log transformation could somehow reduce their leverages.

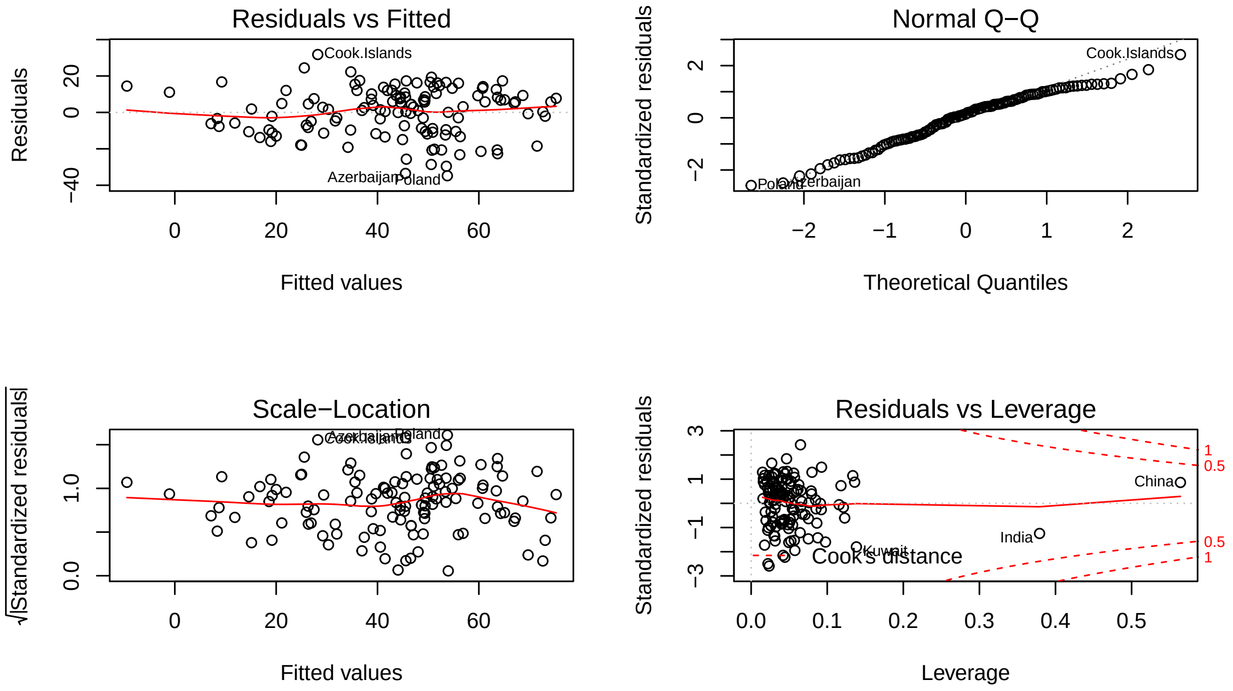

Figure 2: diagnostic residual plot from the linear model

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?

```
dim(na.omit(UN3))
```

```
## [1] 125   7
```

Therefore, after omitting all `NA` values, there are 125 observations used in the model.

```
lm_model <- lm(ModernC ~., UN3)
par(mfrow = c(2, 2))
plot(lm_model)   # Fig.2
```

From Residuals vs Fitted, Scale - Location plot, we can see that the distribution of residuals seems to be basically i.i.d., with only minor heteroscedasticity.

From Normal Q-Q plot, we can notice that the distribution of our sampled data is somehow skewed from normal distribution.

From Residuals vs Leverage plot, we see that China and India are high leverage points, but they don't have large Cook's distance and need further outlier tests for determination.

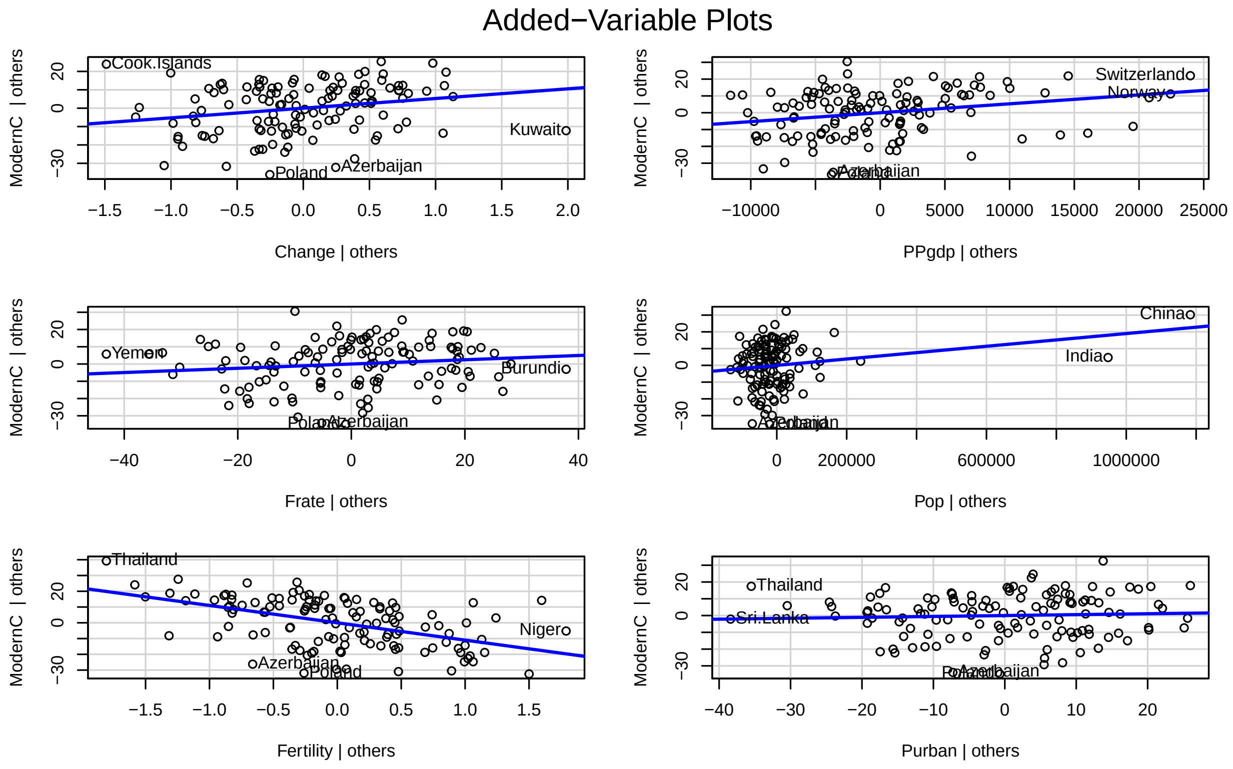

Figure 3: added variable plots

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_model)   # Fig.3
```

From the aded variable plots, we can see that using `log` transformation for Pop could potentially be a good idea, because this will reduce the high leverage of China and India while making a closer-to-linear relationship. Similarly, PPgdp might also need `log` transformation to distribute more evenly. Apart from Pop and PPgdp, other variables seem fine with the current linear relationship.

We can also notice that certain localities could be highly influential for certain terms, while not so influential for the rest. For example, China and India for Pop, Norway and Switzerland for PPgdp, Kuwait and Cook Islands for Change, Burundi and Yemen for Frate, etc.

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.

From previous analysis in Question 5, the most likely candidates for transformation are PPgdp and Pop, both of which are non-negative. Other variables are not examined for transformation in order to reduce computational complexity for `boxTidwell`*.*

```
boxTidwell(ModernC ~ PPgdp + Pop, ~ Change + Fertility + Purban + Frate, data = na.omit(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
```

From the results above, we see that `boTidwell` *suggests a* $\sqrt{\text{Pop}}$ *transformation and a* $\log(\text{PPgdp})$ *transformation, however the p-value for both MLEs of lambda is insignificant, meaning that there isn't enough evidence for the need of transformation.*

Nevertheless, taking `log` *appears to be a good idea for Pop and PPgdp, because of their high leverage points and crowded majorities. We will do so in the following models, where we will soon see that the assumptions of linear models are satisfied well after these* `log` *transformations.*

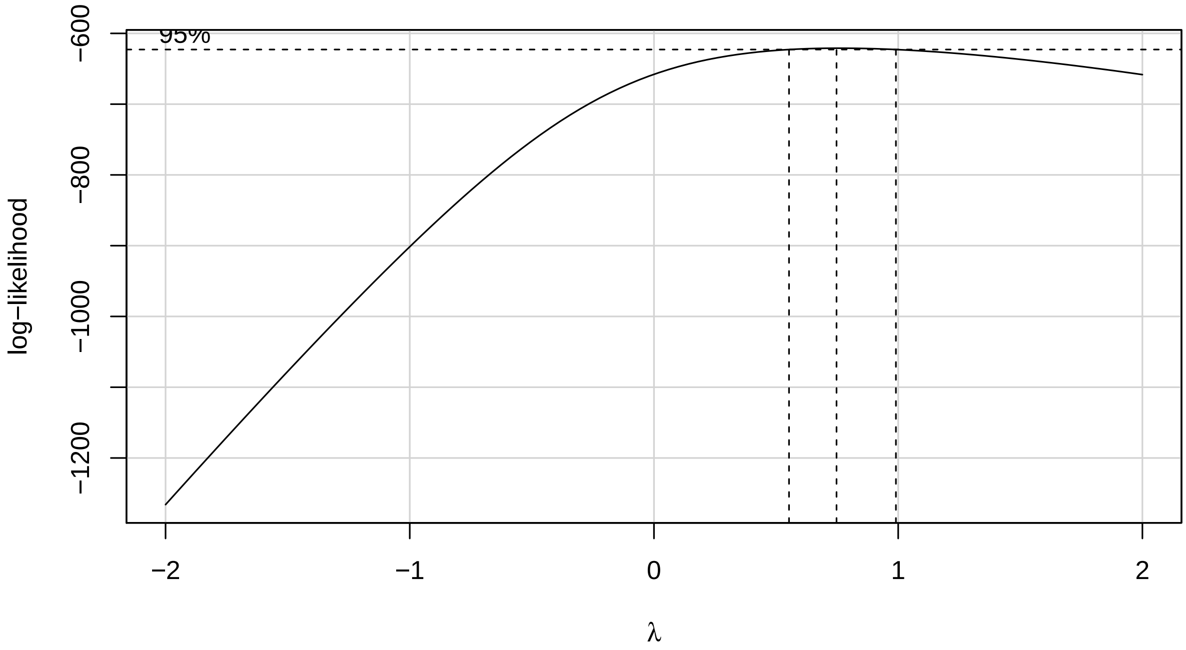

Figure 4: boxCox for ModernC

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

```
UN3$log_Pop <- log(UN3$Pop)
UN3$log_PPgdp <- log(UN3$PPgdp)
lm_model <- lm(ModernC ~ log_Pop + log_PPgdp + Change + Frate + Fertility + Purban, data = UN3)
boxCox(lm_model)  # Fig.4
```

From Figure 4, we can see that the optimal power for ModernC is around 0.8, then there is no need for transformation for sake of interpretation.

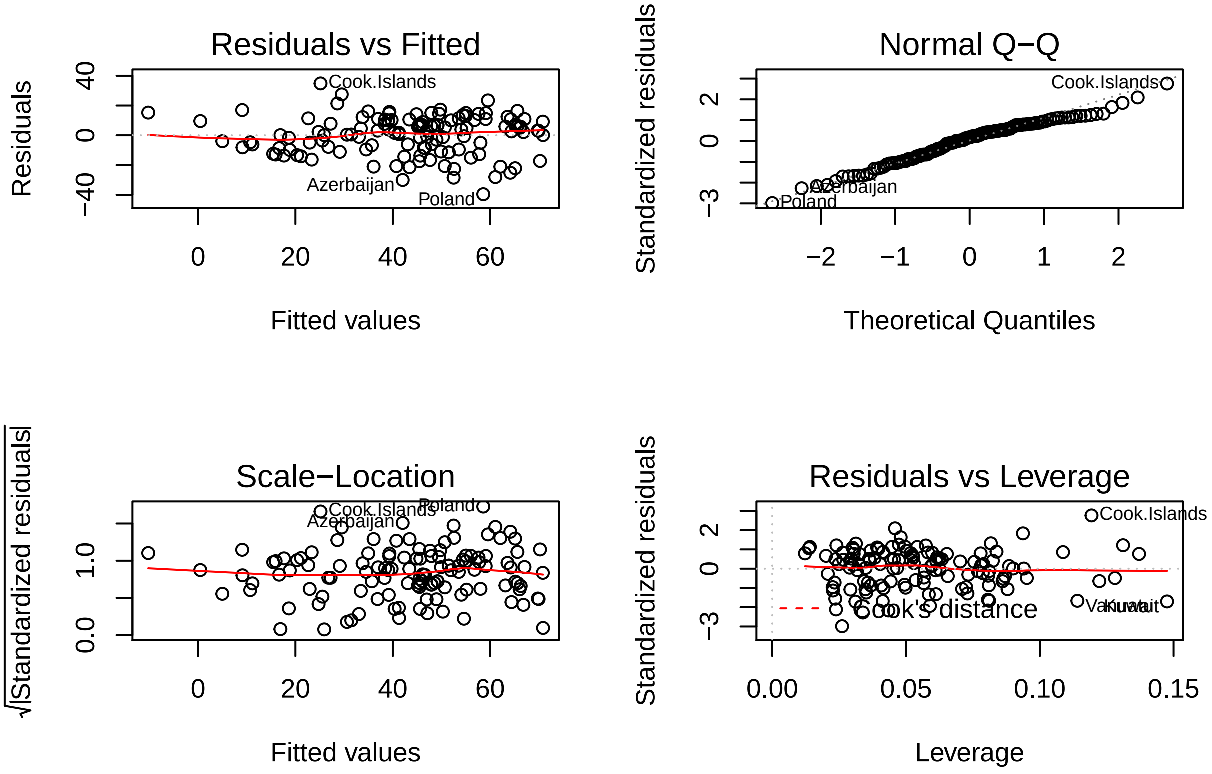

Figure 5: Residual plot for transformed linear model

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.

```r
par(mfrow = c(2, 2))
plot(lm_model)  # Fig.5
```

From Figure 5, we can notice that the problems mentioned in Question 4 still somehow exists, but less obvious. Overall speaking, almost everything looks good.

```r
avPlots(lm_model)  # Fig.6
```

In Figure 6 (on next page), we can see that the `log` *transformation for Pop and PPgdp does help improve the model by eliminating high leverage points and distributing the data evenly, and all six individual avplots seem to follow approximately linear relationships, with no need for further transformation.*

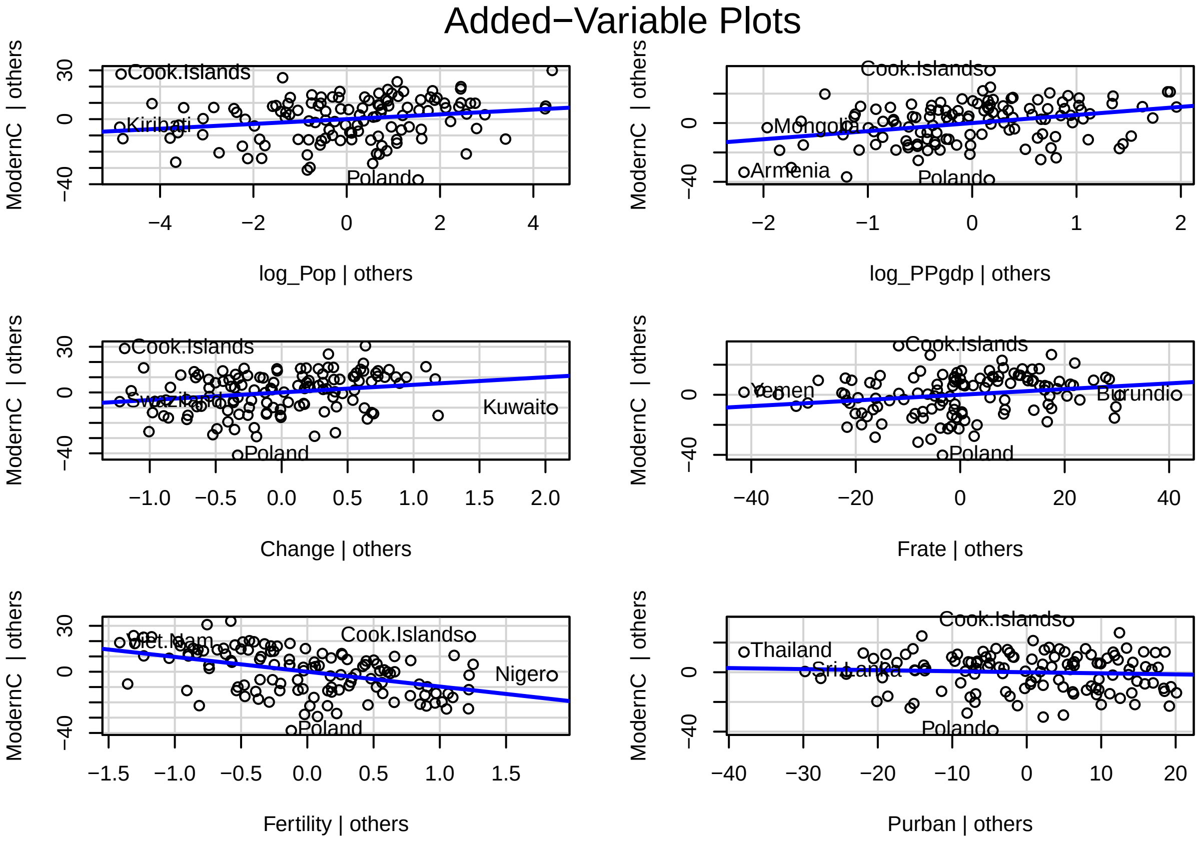

Figure 6: Added Variable plot for transformed linear model

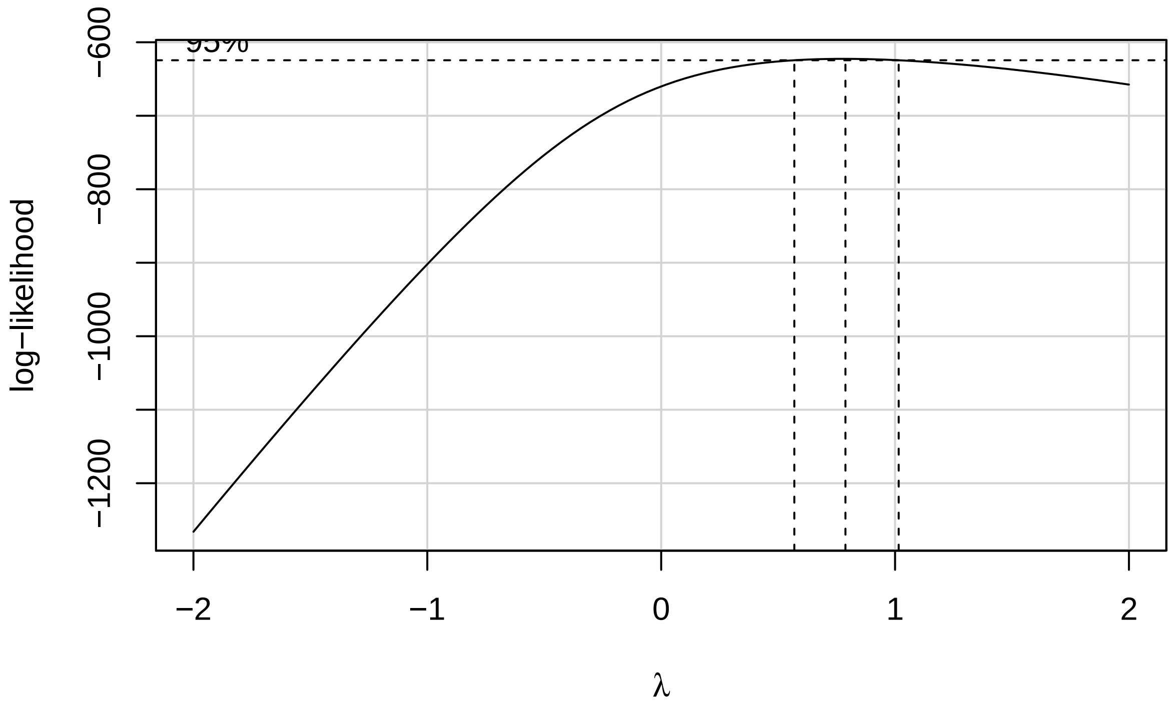

Figure 7: boxCox for the untransformed linear model

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?

```
boxCox(lm(ModernC ~ Pop + PPgdp + Change + Frate + Fertility + Purban, data = UN3))  # Fig.7
```

Similar to what we analyzed in Question 7, there is no need for transformation of reponse either even if we examine reponse first and predictors later. Therefore, the model will be the same as Question 8, with Pop and PPgdp being `log` transformed.

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.

```
any(cooks.distance(lm_model) >= 1)
```

```
## [1] FALSE
```

```
outlierTest(lm_model)
```

```
## No Studentized residuals with Bonferonni p < 0.05
## Largest |rstudent|:
##         rstudent unadjusted p-value Bonferonni p
## Poland -3.090987          0.0024937      0.31171
```

Based on Figure 5 and 6 in Question 8, as well as the Cook's distance and Bonferonni outlier test conducted above, we can see that there is no obvious outliers or influential points in the transformed data.

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!

The 95% confidence intervals are estimated by each coefficient's mean plus/minus 1.96 times its standard deviation, as follows.

```
kable_data1 <- Confint(lm_model)
kable_data2 <- exp(kable_data1[c("log_Pop", "log_PPgdp"), ])
row.names(kable_data2) <- c("Pop", "PPgdp")
kable_data <- rbind(kable_data1, kable_data2)
kable(kable_data, caption = "summaries of coefficients with 95% confidence intervals")
```

Table 2: summaries of coefficients with 95% confidence intervals

	Estimate	2.5 %	97.5 %
(Intercept)	4.1154711	-24.6153857	32.8463280
log_Pop	1.4720744	0.2269699	2.7171788
log_PPgdp	5.5072784	2.7249039	8.2896530
Change	4.9929573	0.8797496	9.1061651
Frate	0.1893936	0.0366943	0.3420929
Fertility	-9.6759414	-13.1723343	-6.1795485
Purban	-0.0707680	-0.2640391	0.1225031
Pop	4.3582664	1.2547921	15.1375564
PPgdp	246.4794010	15.2549478	3982.4518549

As interpretation of Table 2, the following statements hold on average:

- *For every 1 percent increase in annual population growth rate, there is a 4.993 percent increase of unmarried women using a modern method of contraception.*
- *For every 1 percent increase in females over age 15 economically active, there is a 0.189 percent increase of unmarried women using a modern method of contraception.*
- *For every 1 unit increase in expected numer of live births per female, there is a 9.676 percent decrease of unmarried women using a modern method of contraception.*
- *For every 1 percent increase in population that is urban, there is a 0.071 percent decrease of unmarried women using a modern method of contraception.*
- *For every 1 percent increase in population, there is a* $\log(1.01) * 1.472 = 0.0146$ *percent increase of unmarried women using a modern method of contraception.*
- *For every 1 percent increase in GDP per capita, there is a* $\log(1.01) * 5.507 = 0.0548$ *percent increase of unmarried women using a modern method of contraception.*

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

We have studied the percent of unmarried women using a modern method of contraception (ModernC) through six predictors: the impact of Annual population growth rate (Change), GDP per capita (PPgdp), Percent of females over age 15 economically active (Frate), population (Pop), expected numer of live births per female (Fertility) and percent of urban population (Purban).

210 pieces of data are collected in different countries and localities, while only 125 of them are actually used, due to the vast existence of missing entries.

Our model is constructed as

$$ModernC = 4.115 + 1.472\log(Pop) + 5.507\log(PPgdp) + 4.993Change + 0.189Frate - 9.676Fertility - 0.071Purban$$

From the model above, we can see that the proportion of unmarried women using a modern method of contraception could increase if

- *the population increases*
- *the GDP per capita increases*
- *the annual population grwoth rate increases*
- *the proportion of females over age 15 economically active increases*
- *the expected numer of liver births per female decreases*
- *the proportion of urban population decreases*

This is intuitively reasonable, because

- *if the population or annual population growth rate increase, families will have less desire for children due to the shortage of resources, and therefore more usage of modern contraception methods*
- *if GDP per capita grows or females become more economically active, modern contraception methods will be more affordable. and therefore more implementations*
- *a growth in expected number of live births per female represents families' desire for children, and therefore resulting less need for contraception*
- *finally, the rise in urban population percentage reflects a higher living standard, providing better opportunities to raise a child and therefore fewer families will consider contraception methods.*

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.*

When regressing Y against all variables except X_i, let $H_{(i)}$ denote its hat matrix, then the residual of Y is

$$Y-\hat{Y}=(I-H_{(i)})Y$$

And the residual's mean

$$\overline{Y-\hat{Y}}=\frac{1}{n}\mathbf{1}_n^\top(Y-\hat{Y})=\frac{1}{n}\mathbf{1}_n^\top(I-H_{(i)})Y=0$$

where the hint is used. Similarly, when regressing X_i against all other variables, the mean of its residuals is

$$\overline{X_i-\hat{X}_i}=\frac{1}{n}\mathbf{1}_n^\top(X_i-\hat{X}_i)=\frac{1}{n}\mathbf{1}_n^\top(I-H_{(i)})X_i=0$$

For a simple linear regression model $y\sim\beta_1x+\beta_0$, we know that the coefficient estimates are

$$\hat{\beta}_1=\frac{\overline{xy}-\overline{x}\;\overline{y}}{\overline{x^2}-\overline{x}^2},\qquad\hat{\beta}_0=\overline{y}-\hat{\beta}_1\overline{x}$$

where $^-$ represents the mean. This formula could be easily found in any regression textbook.

By substituting x and y with $X_i-\hat{X}_i$ and $Y-\hat{Y}$, we have the intercept estimator in added variable plot as

$$\hat{\beta}_0=\overline{Y-\hat{Y}}-\hat{\beta}_1\cdot\overline{X_i-\hat{X}_i}=0-\hat{\beta}_1\cdot 0=0$$

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 select Change as the leftout variable.

```
ModernC_residuals <- residuals(lm(ModernC ~ log_Pop + log_PPgdp + Frate + Fertility + Purban,
                                  data = na.omit(UN3)))
Change_residuals <- residuals(lm(Change ~ log_Pop + log_PPgdp + Frate + Fertility + Purban,
                                 data = na.omit(UN3)))
av_data <- data.frame(ModernC_residuals, Change_residuals)
av_model <- lm(ModernC_residuals ~ Change_residuals, data = av_data)
summary(av_model)$coefficients
```

```
##                      Estimate Std. Error       t value   Pr(>|t|)
## (Intercept)      -1.239519e-16   1.177707 -1.052485e-16 1.00000000
## Change_residuals  4.992957e+00   2.034437  2.454221e+00 0.01551866
```

From the summary above, we can see that the coefficient of Change in added variable model is 4.993

```
summary(lm_model)$coefficients
```

```
##                Estimate  Std. Error    t value     Pr(>|t|)
## (Intercept)  4.11547111 14.50853884  0.2836586 7.771692e-01
## log_Pop      1.47207436  0.62875419  2.3412557 2.089650e-02
## log_PPgdp    5.50727842  1.40504647  3.9196415 1.492131e-04
## Change       4.99295735  2.07709205  2.4038209 1.778126e-02
## Frate        0.18939357  0.07711025  2.4561402 1.550017e-02
## Fertility   -9.67594142  1.76561222 -5.4802189 2.444298e-07
## Purban      -0.07076799  0.09759825 -0.7250948 4.698293e-01
```

From the summary above, we can find that the coefficient of Change in our transformed model is also 4.993, which numerically verifies the question's proposition.