HW2 STA521 Fall18

Jingyi Zhang, jz139, rebeccazjy425

Due September 23, 2018 5pm

Backgound 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 qualtitative?

```r
UN3 = na.omit(UN3)
summary(UN3)
```

```
##     ModernC          Change           PPgdp           Frate      
## Min.   : 1.00   Min.   :-1.100   Min.   :   90   Min.   : 2.00  
## 1st Qu.:28.00   1st Qu.: 0.340   1st Qu.:  687   1st Qu.:39.00  
## Median :45.00   Median : 1.260   Median : 2077   Median :49.00  
## Mean   :43.27   Mean   : 1.182   Mean   : 6613   Mean   :48.11  
## 3rd Qu.:58.00   3rd Qu.: 1.940   3rd Qu.: 7724   3rd Qu.:58.00  
## Max.   :83.00   Max.   : 3.620   Max.   :36445   Max.   :91.00  
##       Pop             Fertility        Purban      
## Min.   :     19   Min.   :1.000   Min.   :  6.00  
## 1st Qu.:   3443   1st Qu.:1.700   1st Qu.: 40.00  
## Median :   8877   Median :2.500   Median : 58.00  
## Mean   :  46060   Mean   :2.876   Mean   : 56.98  
## 3rd Qu.:  31510   3rd Qu.:3.750   3rd Qu.: 75.00  
## Max.   :1304196   Max.   :8.000   Max.   :100.00  
```

```r
uni.val = apply(UN3, 2, unique)
func = function(x){length(uni.val[[x]])}
num.var = rbind(names(uni.val),sapply(1:length(uni.val), func))
kable(num.var, caption="Number of Unique Values of Each Variable")
```

From the summary and the table of numbers of unique values, 6 variables contain missing data. All of the 7 variables are quantitative, so none is qualitative.

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

```r
means = apply(UN3, 2, mean, na.rm=TRUE)
sds = apply(UN3, 2, sd, na.rm = TRUE)
summarys = rbind(means, sds)
kable(summarys, caption="Mean and Sd of Each Variable")
```

Table 1: Number of Unique Values of Each Variable

ModernC	Change	PPgdp	Frate	Pop	Fertility	Purban
69	105	123	57	125	109	70

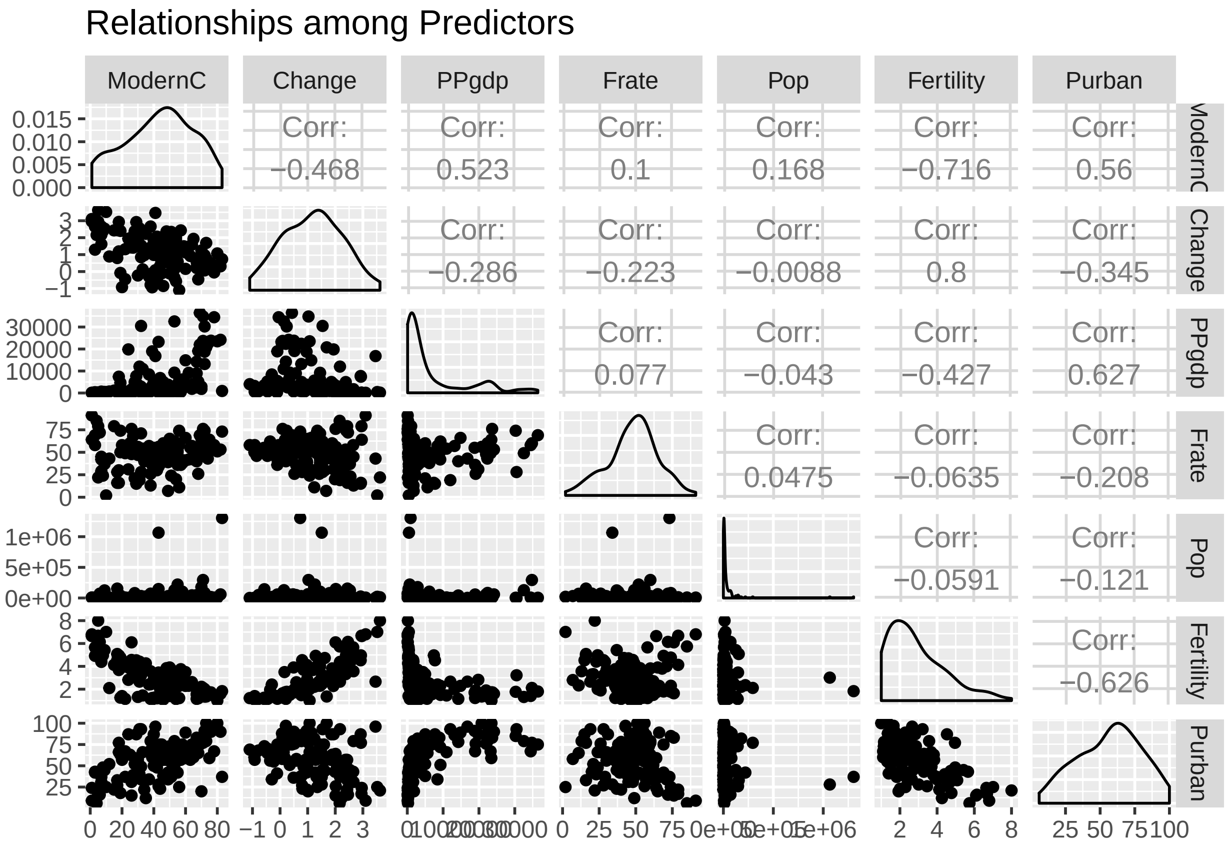
Table 2: Mean and Sd of Each Variable

	ModernC	Change	PPgdp	Frate	Pop	Fertility	Purban
means	43.27200	1.181680	6612.608	48.11200	46060.5	2.876240	56.9760
sds	21.44249	1.064931	9214.771	16.90448	153631.2	1.516948	22.5912

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?

From the ggpairs plot, we notice that, setting `ModernC` as the response variable, we notice that there are stronger linear relationship between the response variable and `Purban`, `Fertility`, `Change` and `PPgdp`. Linear relationships are less apparent between `ModernC` and `Pop` and `Frate`. We also notice some potentially influential points/outliers in `Pop` and `PPgdp` from the scatterplots. Specific potential outliers and influential points are discussed in question 5 with the avplots.

```r
ggpairs(UN3, na.rm = TRUE, title = "Relationships among Predictors")
```



Model Fitting

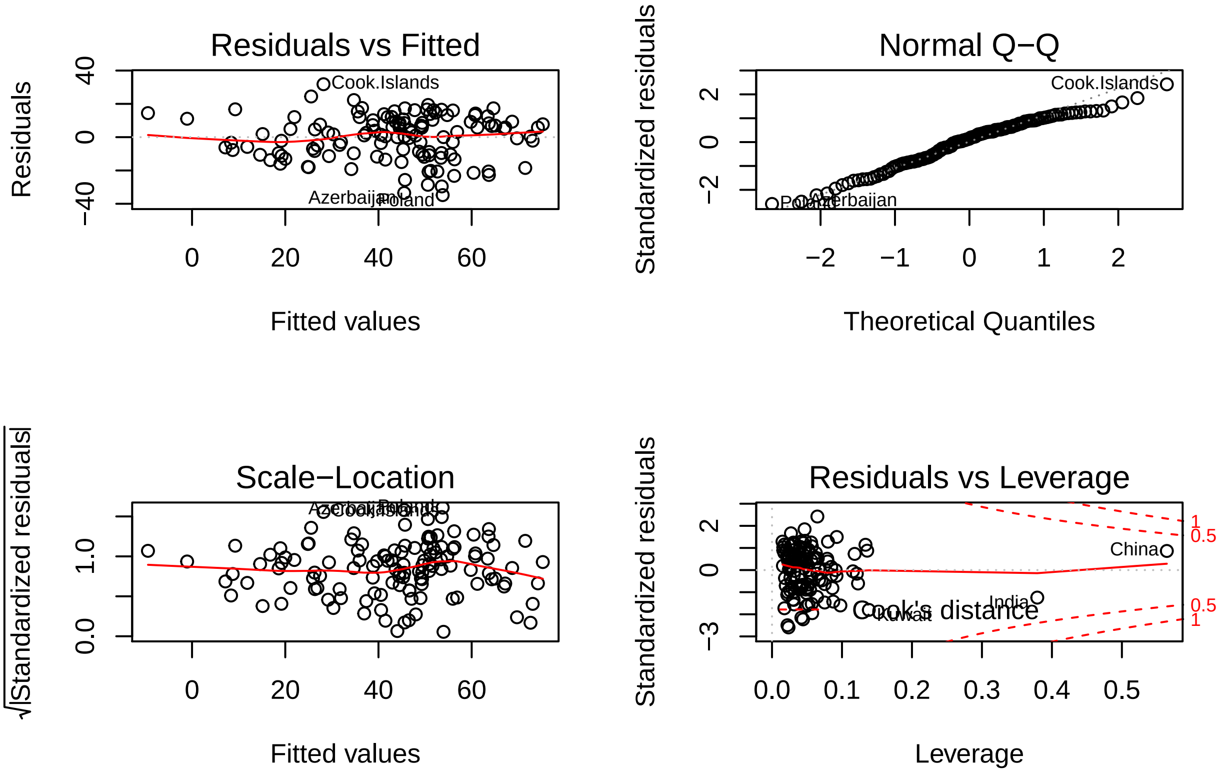
4. Use the `lm()` function to perform a multiple linear regression with `ModernC` as the response and all other variables as the predictors, using the formula `ModernC ~ .`, where the `.` includes all remaining variables in the dataframe. Create diagnostic residual plot from the linear model object and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

From the model summary, we know that 85 out of the 210 observations are deleted because of missingness. Thus only 125 observations are used in the model fitting. The four residual plots: in the residuals vs. fitted plot, the residuals are scattered around 0. There is no apparent pattern. The normal Q-Q plot show the normality assumption is plausible, except as moving on the the upper quantiles: there might be some potential outliers. The Scale-Location plot show some clustering at around fitted values 40-60. Last but not least, no point is outside of cook's distance in the leverage plot, but there are points with extremely high leverage. (eg. China, India). Different potentially influential cases are pointed out by different diagnostic plots. Further outlier tests need to be conducted to make decisions on whether some can be considered outlier cases.

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

```
## 
## 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
## Multiple R-squared:  0.6183, Adjusted R-squared:  0.5989 
## F-statistic: 31.85 on 6 and 118 DF,  p-value: < 2.2e-16
```

```
par(mfrow = c(2,2))
plot(lm_all)
```

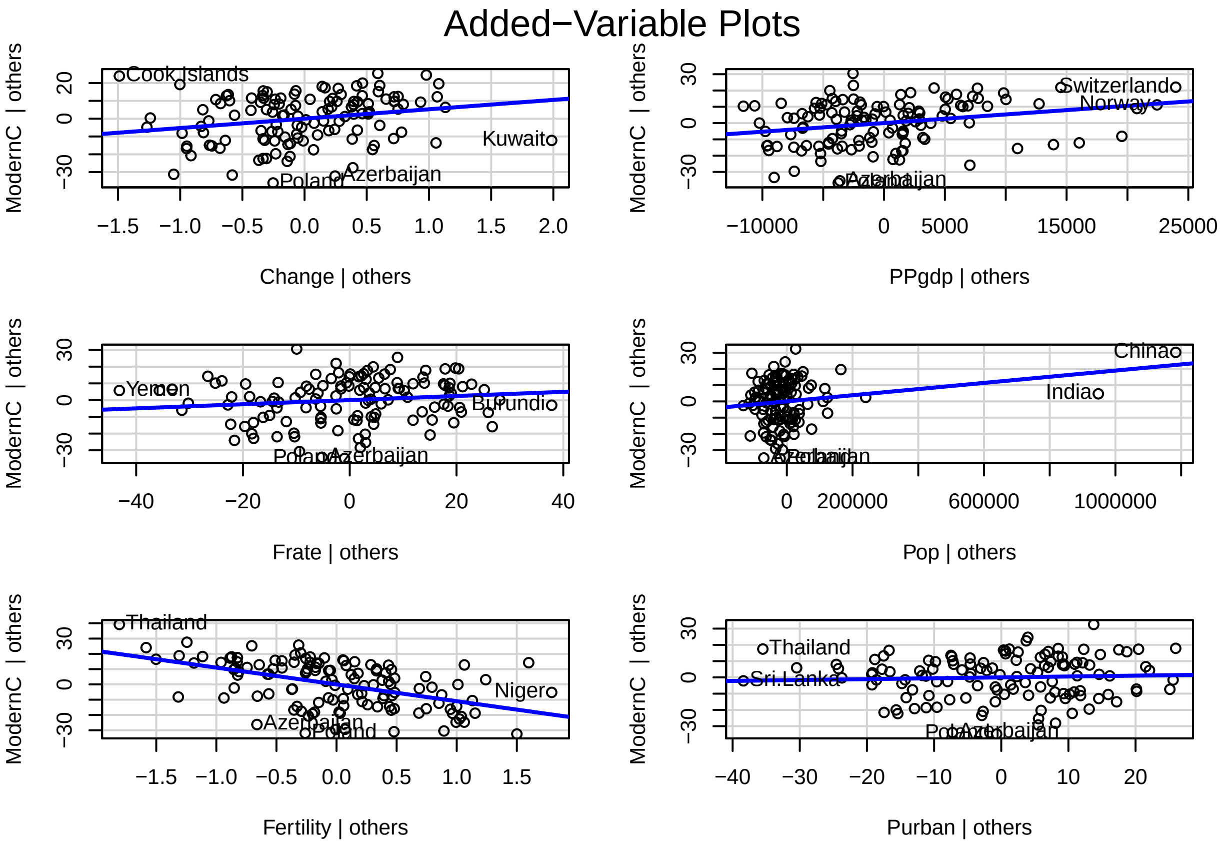


```r
par(mfrow = c(1,1))
```

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?

From the plots, it seems like `Pop` and `PPgdp` needs a log transformation, because the `Pop|others` and `PPgdp` plots show a concentration around lower values and sparsely spread out through out the rest of the plot. Influential localities for each variable include: `Change`: Poland, Azerbajian, Kuwaito, Cook Islands `PPgdp`: India, Poland, Norway, Switzerland, Azerbajian `Frate`: Poland, Azerbajian, Burundio, Yemen `Pop`: India, China, Azerbajian, Poland `Fertility`: Thailand, Azerbajian, Poland, Niger `Purban`: Sri Lanka, Thailand, Azerbajian, Poland In all the variables' potential influential localities, Poland and Azerbajian appear in all of them.

```r
avPlots(lm_all)
```



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.

First of all, from the summaries of the variables, only `Change` contains negative values, so we first transform `Change` all into non-negative values by subtracting the minimum value of `Change`. From question 5, we know that `Pop` and `PPgdp` might need a log transformation, so we use Box-Tidwell to examine if transformation are necessary. However, the Box-Tidwell suggest that transformations aren't necessary. However, the avplots in the previous questions did seem suspicious, so I still performed a log transformation and re-run the linear model with log transformations. It turns out that the log-transformed predictors show a much stronger linear relationship with the dependent variable and improved residual plots. Thus, log transformations are still kept for later analyses.

```
UN3_new = UN3  %>% mutate(Change_non = Change - min(na.omit(Change))) %>% select(-Change)
without_log = boxTidwell(ModernC ~  Pop + PPgdp, ~  Fertility + Change_non + Frate + Purban, data=UN3_n
without_log
```

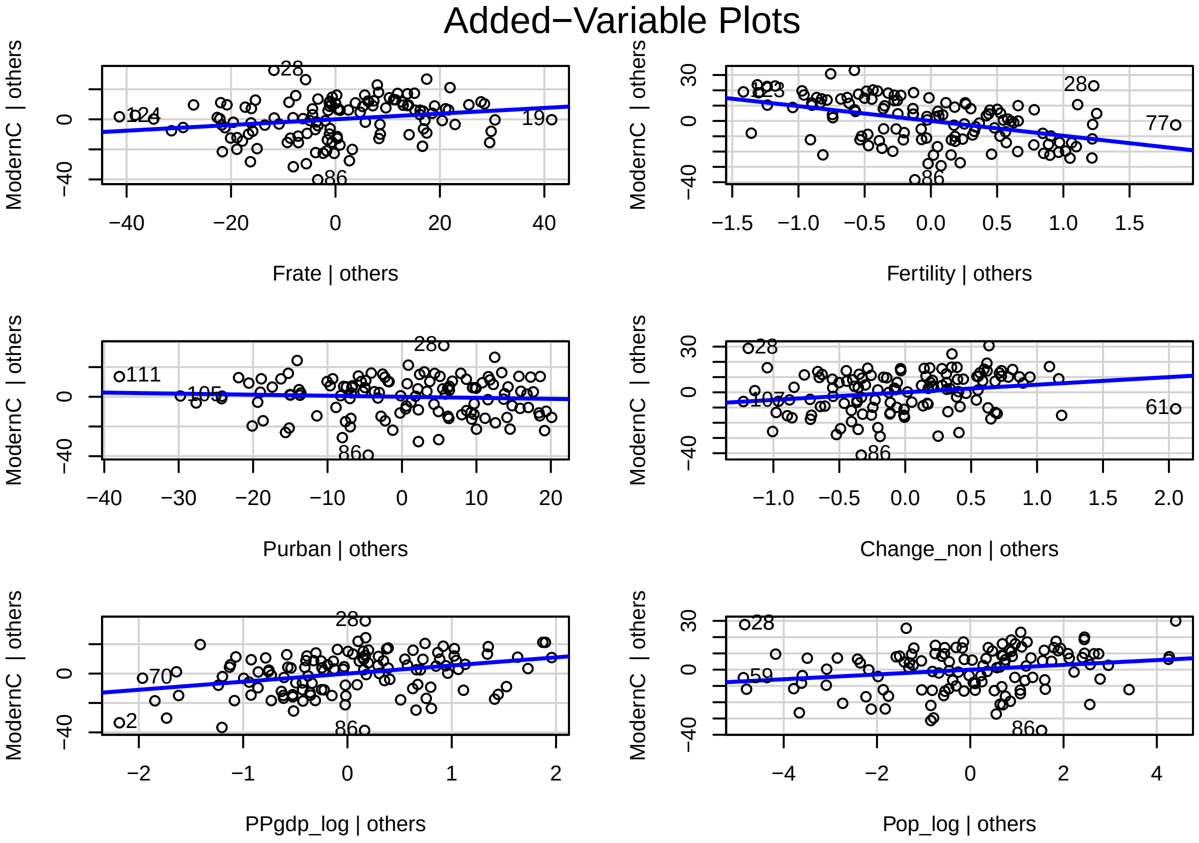
```
##       MLE of lambda Score Statistic (z) Pr(>|z|)
## Pop         0.40749             -0.7874   0.4310
## PPgdp      -0.12921             -1.1410   0.2539
##
## iterations =  4
```

```
UN3_log = UN3_new %>% mutate(PPgdp_log = log(PPgdp), Pop_log = log(Pop)) %>% select(-PPgdp,-Pop)
lm_log = lm(ModernC ~ ., data = UN3_log)
summary(lm_log)
```

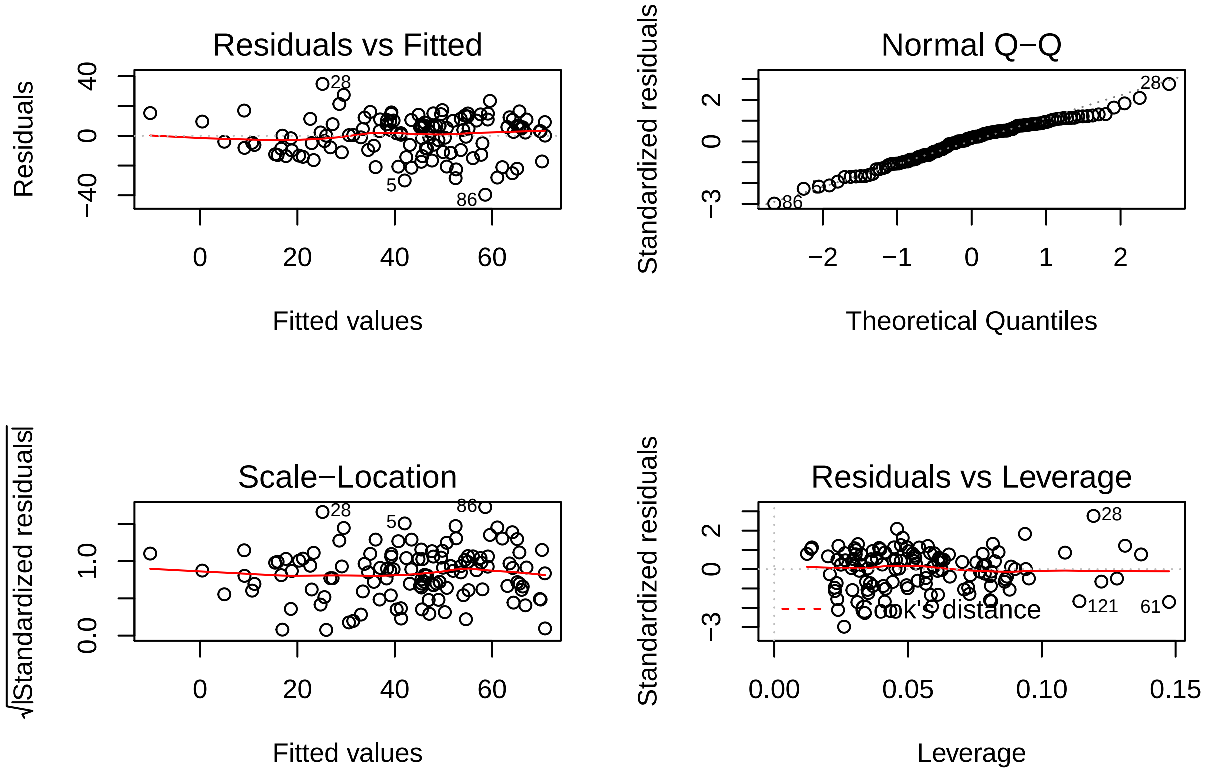
```
##
```

```
## Call:
## lm(formula = ModernC ~ ., data = UN3_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) -1.37678   14.19426  -0.097 0.922894
## 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
## Change_non   4.99296    2.07709   2.404 0.017781 *
## PPgdp_log    5.50728    1.40505   3.920 0.000149 ***
## Pop_log      1.47207    0.62875   2.341 0.020897 *
## ---
## 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
```

```
avPlots(lm_log)
```



```
par(mfrow = c(2,2))
plot(lm_log)
```

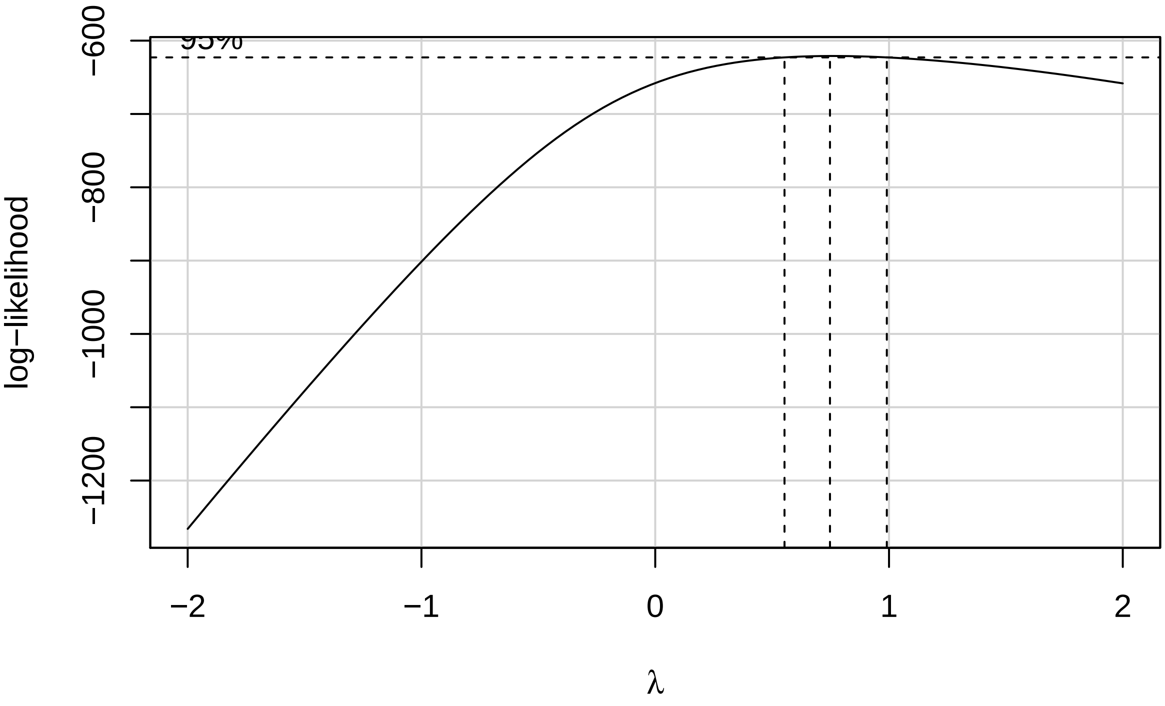


```
par(mfrow = c(1,1))
```

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

The boxCox method suggest a power transformation of the response variable to a power of 0.7585897.

```
car::boxCox(lm_log)
```



```r
powerTransform(lm_log)
```

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

```r
UN3_log_power = UN3_log %>% mutate(trans_modernC = ModernC ^ 0.7585897) %>% select(-ModernC)
lm_log_power = lm(trans_modernC ~ ., data = UN3_log_power)
```

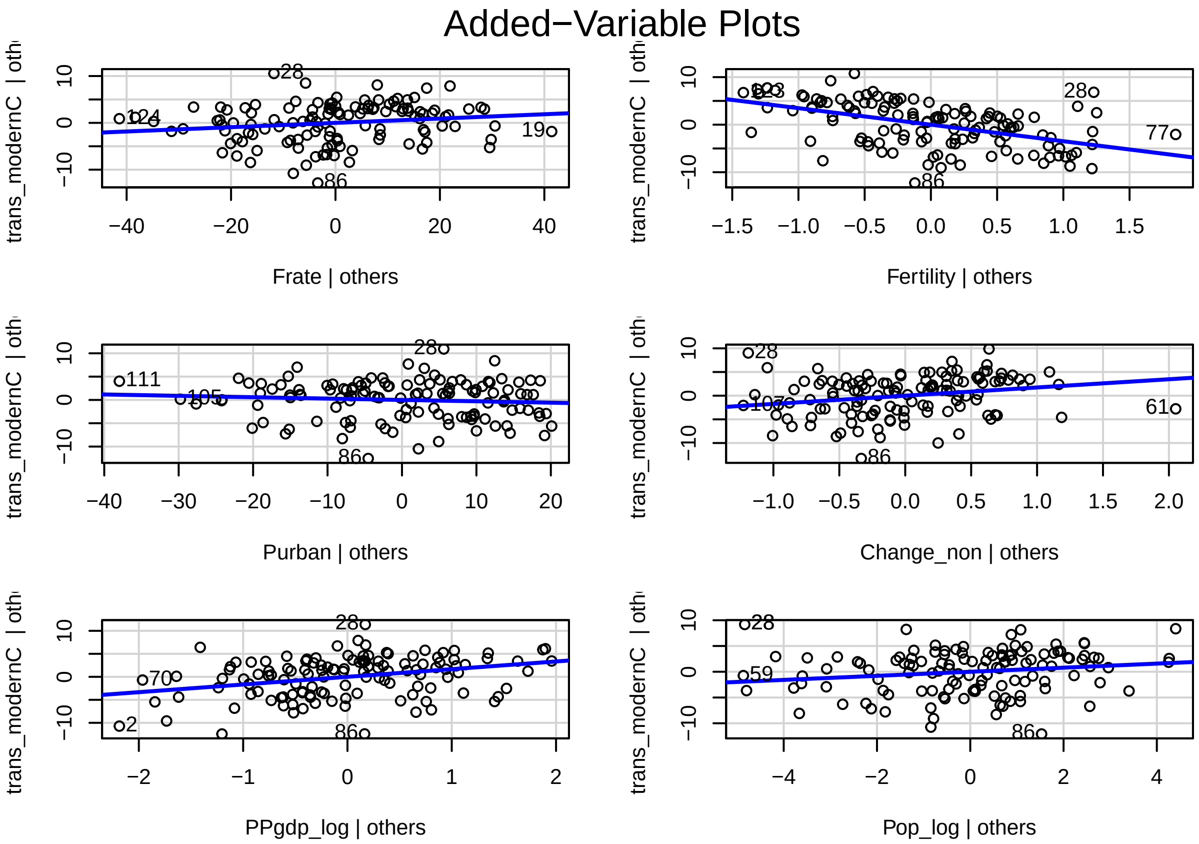
8. Fit the regression using the transformed variables. Provide residual plots and added variables plots and comment. If you feel that you need additional transformations of either the response or predictors, repeat any steps until you feel satisfied.

After applying the transformation onto the response variable and re-run the linear model, we look at the sumamry, avplots and diagnostic plots again. Unfortunately, there doesn't seem to be a significant improvement with the power transformation. Since we need to both consider the model fit and the interpretability of the model, I don't think the power transformation is necessary here. Only the log transformations of `Pop` and `PPgdp` are kept in the model.

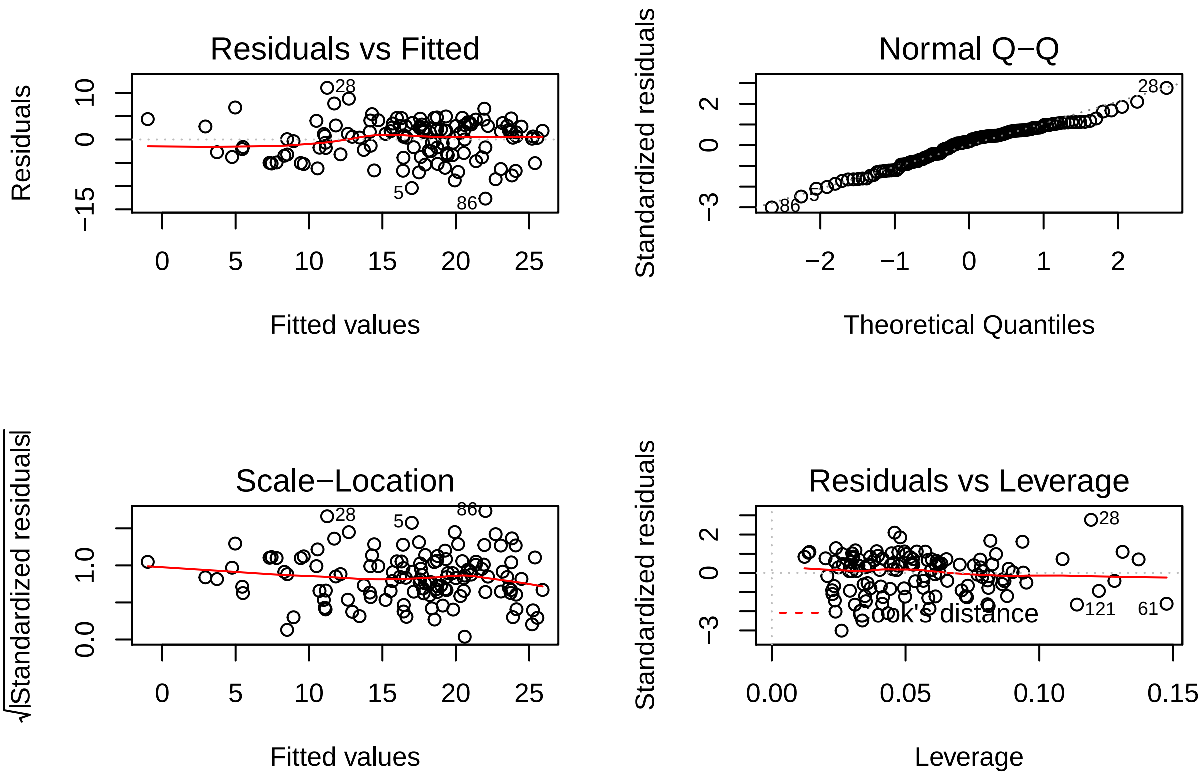
```r
summary(lm_log_power)
```

```
##
## Call:
## lm(formula = trans_modernC ~ ., data = UN3_log_power)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.6887  -2.9682   0.7644   2.9067  11.0914
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.91017    4.51170   1.310  0.19275
## Frate        0.04603    0.02451   1.878  0.06287 .
## Fertility   -3.47873    0.56121  -6.199 8.64e-09 ***
## Purban      -0.02928    0.03102  -0.944  0.34717
## Change_non   1.73626    0.66021   2.630  0.00968 **
## PPgdp_log    1.66841    0.44660   3.736  0.00029 ***
## Pop_log      0.39544    0.19985   1.979  0.05018 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.273 on 118 degrees of freedom
## Multiple R-squared:  0.643,  Adjusted R-squared:  0.6249
## F-statistic: 35.43 on 6 and 118 DF,  p-value: < 2.2e-16
```

```r
avPlots(lm_log_power)
```



```r
par(mfrow = c(2,2))
plot(lm_log_power)
```

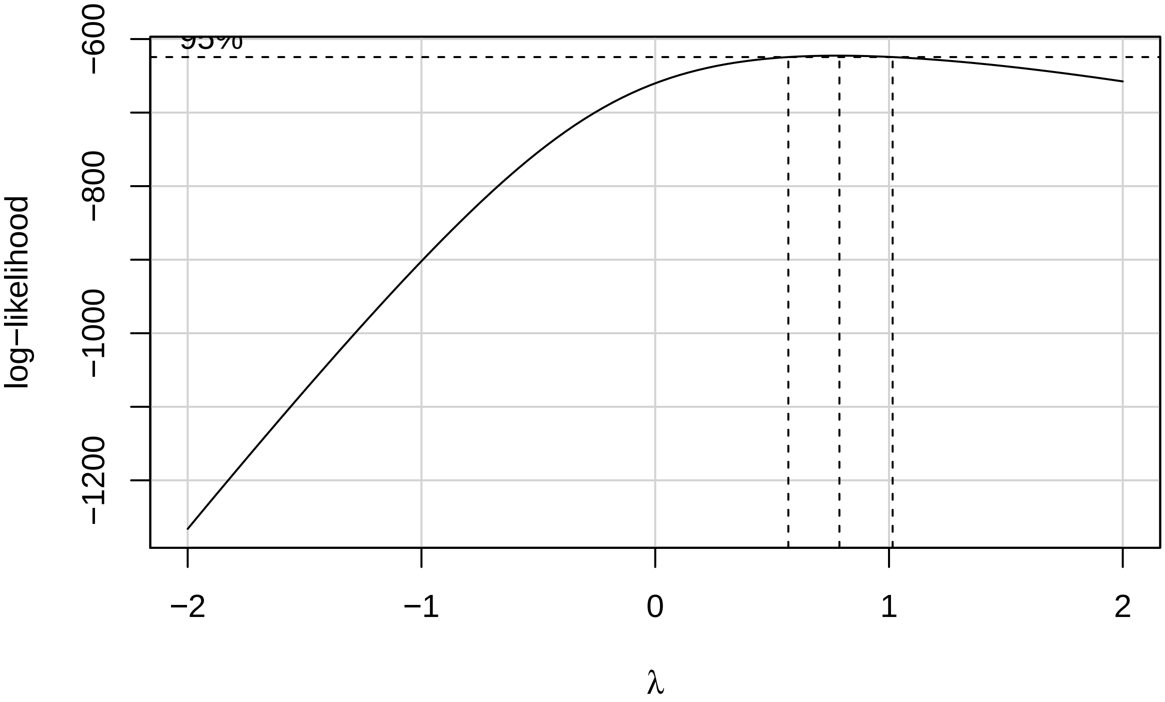


```r
par(mfrow = c(1,1))
```

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?

By switching the transformation order, I first use boxCox to determine if a tranformation for the response variable is necessary. Since this time, the 95% confidence interval of the power transformation includes the value 1, which is essentially indicating no transformation, I decided that no transformation for the response is included in the model. Although boxTidwell still suggests no transformation for predictors either, because of the reasons and improvements in selecting the model shown above with log transformation on `Pop` and `PPgdp`, I still keep these two transformations. Thus the final model is the same as question 8, with `Pop` and `PPgdp` log transformed, `Change` transformed to non-negative and no transformation on any other variable.

```r
car::boxCox(lm_all)
```



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

```
##       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.

First of all, the diagnostic plot from question 6 show that, there are influential points including Cook Islands, Vanuatu and Kuwait, but no point is outside the Coo's distance. The Bonferroni correction process also suggests that there is no outliers in the data, thus no point needs to be considered outliers and be removed.

```r
summary(lm_log)
```

```
##
## Call:
## lm(formula = ModernC ~ ., data = UN3_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) -1.37678   14.19426  -0.097 0.922894
## 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
## Change_non   4.99296    2.07709   2.404 0.017781 *
## PPgdp_log    5.50728    1.40505   3.920 0.000149 ***
## Pop_log      1.47207    0.62875   2.341 0.020897 *
## ---
## 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
```

```
p = 2*(1 - pt(abs(rstudent(lm_log)), lm_log$df - 1))
rownames(UN3)[p<.05/nrow(UN3_log)]
```

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

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!

I finally decided to not include Purban as a predictor from the conclusion of the ANOVA test. The coefficients and CI of the coefficients of the final model is as following and transformed back to the normal scale.

Interpretations of the coefficients: (Since it is impossible for all predictors to be 0 in practice, the interpretation of the intercept is essentially useless) Frate: each 1% increase in Frate results in a 0.2% increase in ModernC. Feritility: each unit increase in Fertility results in a 9.278% decrease in ModernC. Change: each 1% increase in Change results in a 4.698% increase in ModernC. PPgdp_log: each 1% increase in PPgdp results in a 0.048% increase in ModernC. Pop_log: each 1% increase in Pop results in a 0.014% increase in ModernC.

```
lm_noP = lm(ModernC ~ Frate + Fertility + Change_non + PPgdp_log + Pop_log, data = UN3_log)
anova(lm_log,lm_noP)
```

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

```
summary(lm_noP)
```

```
##
## Call:
```

Table 3: Estimates and CI of slope and intercept

	95% CI Lower bound	95% CI upper bound	Estimates
(Intercept)	-29.1025994	26.9716970	-1.0654512
Frate	0.0496977	0.3493943	0.1995460
Fertility	-12.5950756	-5.9617672	-9.2784214
Change	0.6727287	8.7227854	4.6977570
1% Increase in PPgdp	0.0270313	0.0696732	0.0483523
1% Increase in Pop	0.0020056	0.0266757	0.0143407

```
## lm(formula = ModernC ~ Frate + Fertility + Change_non + PPgdp_log +
##     Pop_log, data = UN3_log)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -39.276  -9.928   2.572  10.253  34.442
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.06545   14.15946  -0.075  0.94014
## Frate        0.19955    0.07568   2.637  0.00949 **
## Fertility   -9.27842    1.67499  -5.539 1.85e-07 ***
## Change_non   4.69776    2.03274   2.311  0.02255 *
## PPgdp_log    4.85936    1.08214   4.491 1.65e-05 ***
## Pop_log      1.44122    0.62606   2.302  0.02307 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.42 on 119 degrees of freedom
## Multiple R-squared:  0.6243, Adjusted R-squared:  0.6085
## F-statistic: 39.55 on 5 and 119 DF,  p-value: < 2.2e-16
```

```
estimates = as.data.frame(summary(lm_noP)$coef)
sum_sta_df = data.frame(confint(lm_noP),estimates$Estimate)
sum_sta_df["Pop_log",] = sum_sta_df["Pop_log", ]*log(1.01)
sum_sta_df["PPgdp_log", ] = sum_sta_df["PPgdp_log", ]*log(1.01)
rownames(sum_sta_df) = c("(Intercept)", "Frate", "Fertility",  "Change",
colnames(sum_sta_df) = c("95% CI Lower bound", "95% CI upper bound", "Estimates")
kable(sum_sta_df,caption="Estimates and CI of slope and intercept")
```

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.

First of all, there are a lot of influential points in terms of different predictors, with the most noticable being Poland and Azerbajian because they are potential outliers for almost all predictors. However, we decided that there is no case significantly enough to be considered an outlier, thus no case is removed. Second of all, in order to predict `ModernC`, we decided that whether the woman is from an urban location is not an influential predictor, thus we removed it from the group of predictors. The GDP per capital and population predictors do not have a clear linear relationship to `ModernC`, so we performed a log transformation on them. Other predictors remained on the natural scale. Last but not least, this dataset includes a large portion of cases with missing data. In the study above, missing data are simply omitted, but this could affect the precision of the analyses.

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.*

For addictive variable regression there is: $\hat{e}(y) = \hat{\beta}_0 + \hat{\beta}_1 \hat{e_{x_i}}$. This formula can be written into the following form: $(I-H)Y = \hat{\beta}_0 1_{n\times 1} + \hat{\beta}_1 (I-H)x_i$. From the OLS estimate we have: $\hat{\beta}_1 = [\big((I-H)x_i\big)^T((I-H)x_i)]^{-1}\big((I-H)x_i\big)^T(I-H)Y$, we can plug in the estimate into the previous equation and get:

$$(I-H)Y = \hat{\beta}_0 1_{n\times 1} + [\big((I-H)x_i\big)^T((I-H)x_i)]^{-1}\big((I-H)x_i\big)^T(I-H)Y(I-H)x_i$$

$$(I-H)Y = \hat{\beta}_0 1_{n\times 1} + \big(x_i^T(I-H)^T(I-H)x_i\big)^{-1}x_i^T(I-H)^T(I-H)Y(I-H)x_i$$

Since we have $(I-H)^T(I-H) = (I-H)$ and $(I-H)(I-H) = (I-H)$,

$$(I-H)Y = \hat{\beta}_0 1_{n\times 1} + \big(x_i^T(I-H)x_i\big)^{-1}x_i^T(I-H)Y(I-H)x_i$$

By multiplying x_i^T to both sides of the equation, we get:

$$x_i^T(I-H)Y = x_i^T\hat{\beta}_0 1_{n\times 1} + x_i^T\big(x_i^T(I-H)x_i\big)^{-1}x_i^T(I-H)Y(I-H)x_i$$

$$x_i^T(I-H)Y = \hat{\beta}_0 \sum x_i + x_i^T(I-H)Y$$

(The $x_i^T(I-H)x_i$ term is multiplied with its inverse to get 1) Thus we have: $0 = \hat{\beta}_0 \sum x_i$ If there is an intercept, the sample mean of the residuals will always be zero. However, in practice the sample mean of the residuals cannot be 0, the intercept, therefore, is always zero in the added variable, showing a zero intercept in the scatter plot.

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.

I selected `Fertility` as the predictor we're looking at.

```
lm_e_Y = lm(ModernC ~ Frate+ Fertility + Purban + PPgdp_log + Pop_log,data = UN3_log)
e_Y = residuals(lm_e_Y)
lm_e_X = lm(Change_non ~ Frate + Purban + Fertility + PPgdp_log + Pop_log, data = UN3_log)
e_X = residuals(lm_e_X)

q14_data = data.frame(e_Y,e_X)
lm_q14 = lm(e_Y ~ e_X, data = q14_data)
summary(lm_q14)
```

```
##
## Call:
## lm(formula = e_Y ~ e_X, data = q14_data)
##
## 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) -6.198e-17  1.178e+00   0.000   1.0000
## e_X          4.993e+00  2.034e+00   2.454   0.0155 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.17 on 123 degrees of freedom
## Multiple R-squared:  0.04668,    Adjusted R-squared:  0.03893
## F-statistic: 6.023 on 1 and 123 DF,  p-value: 0.01552
```

The coefficient matches with the coefficient of `Change` in Q10.