HW2 STA521 Fall18

[Zhaolin Ying zy70 github:sallyying]

Due September 24, 2018 9am

Exploratory Data Analysis

1. Create a summary of the data. How many variables have missing data? Which are quantitative and which are qualtitative?

```
##     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                      
```

```
##   ModernC    Change     PPgdp     Frate       Pop Fertility    Purban 
##      TRUE      TRUE      TRUE      TRUE      TRUE      TRUE     FALSE 
```

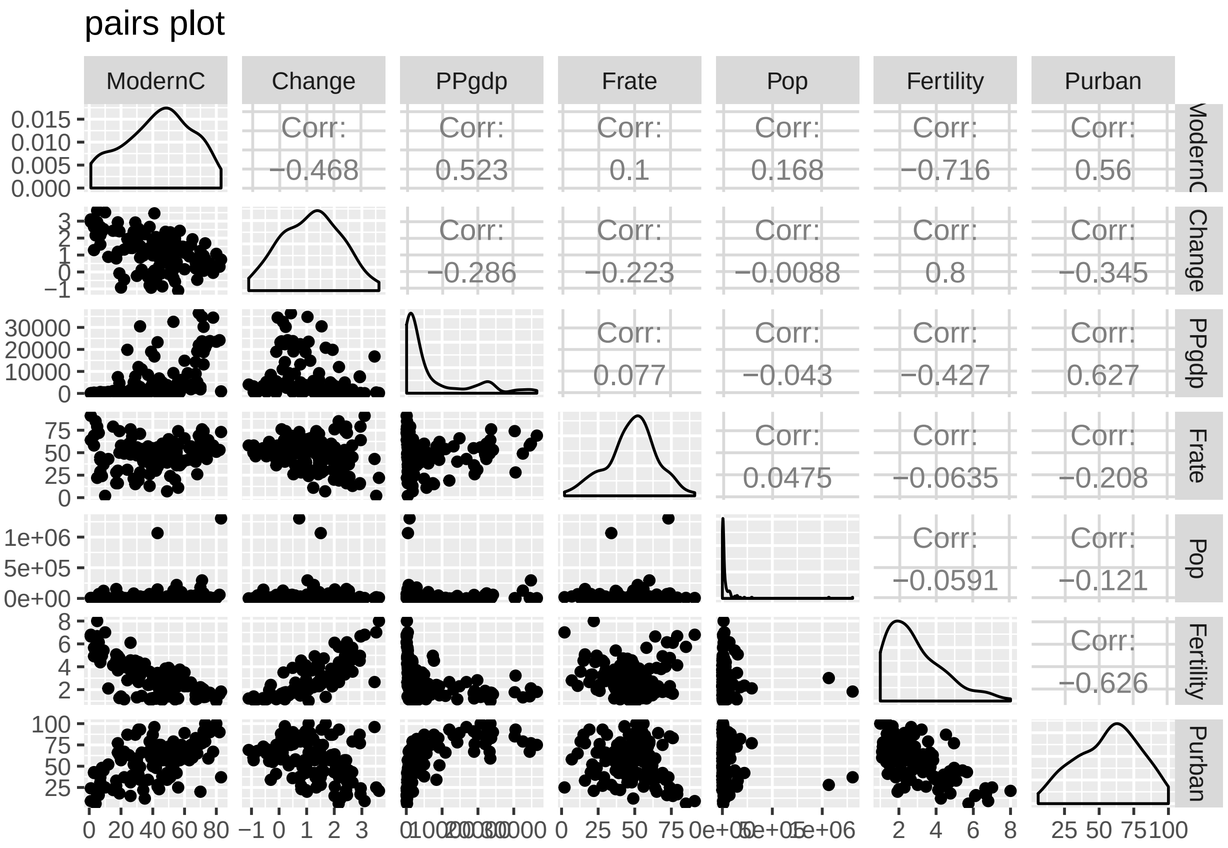
```
##   ModernC    Change     PPgdp     Frate       Pop Fertility    Purban 
##      TRUE      TRUE      TRUE      TRUE      TRUE      TRUE      TRUE 
```

The result shows that all variables, except for "Purban", have missing data. All variables are quantitative.

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

	mean	sd
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

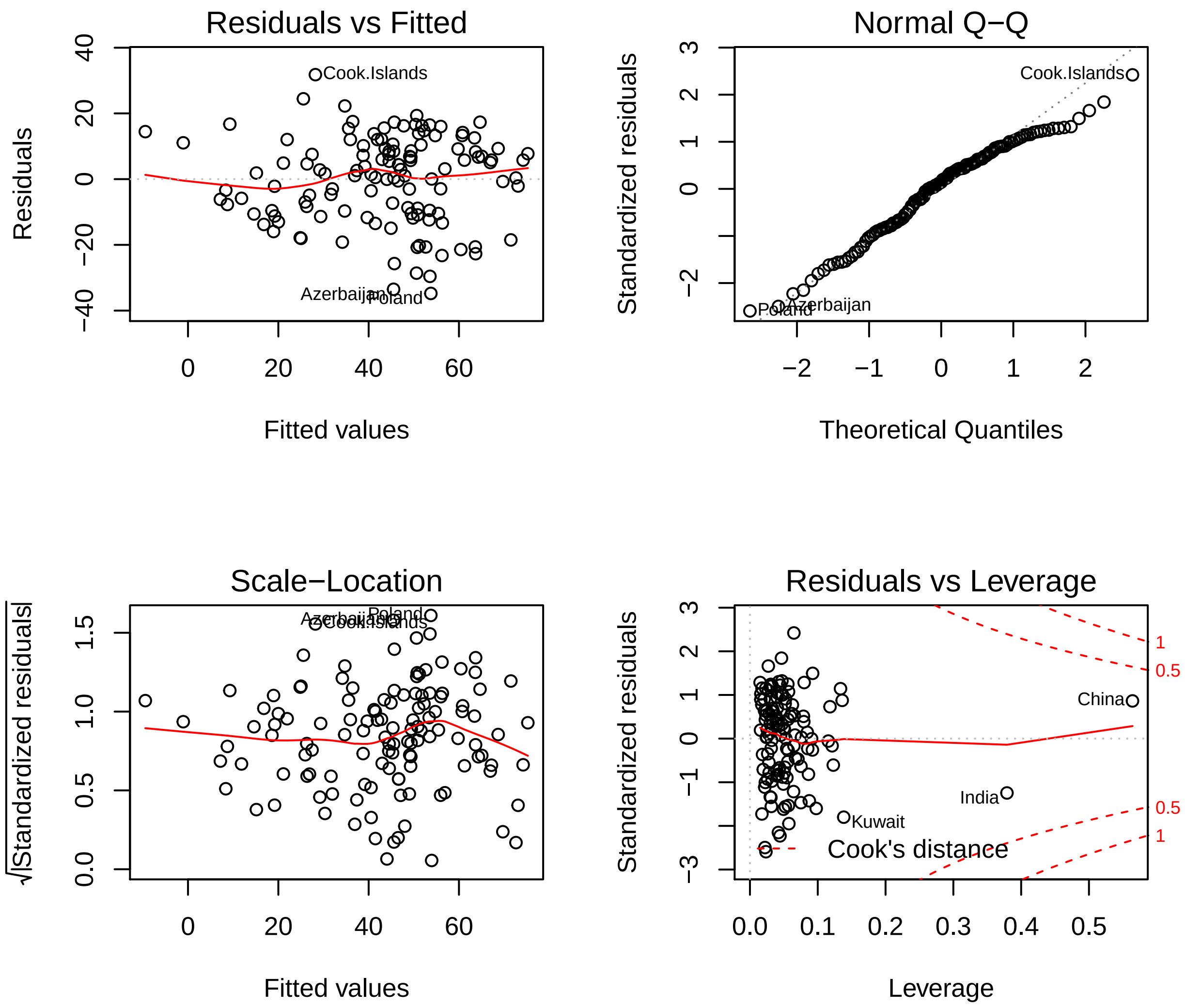
3. Investigate the predictors graphically, using scatterplots or other tools of your choice. Create some plots highlighting the relationships among the predictors. Comment on your findings regarding trying to predict `ModernC` from the other variables. Are there potential outliers, nonlinear relationships or transformations that appear to be needed based on your graphical EDA?



Variables including “Fertility”, “Purban”, “PPdgp”, and “Change” may be useful in predicting ModernC. There may be nonlinear relationships between “ModernC” and “Change”, “PPgdp” and “Pop”. The plots between “ModernC” and “PPgdp”, “Purban” shows there are potential outliers, given there are data points obviously far away from other data.

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?

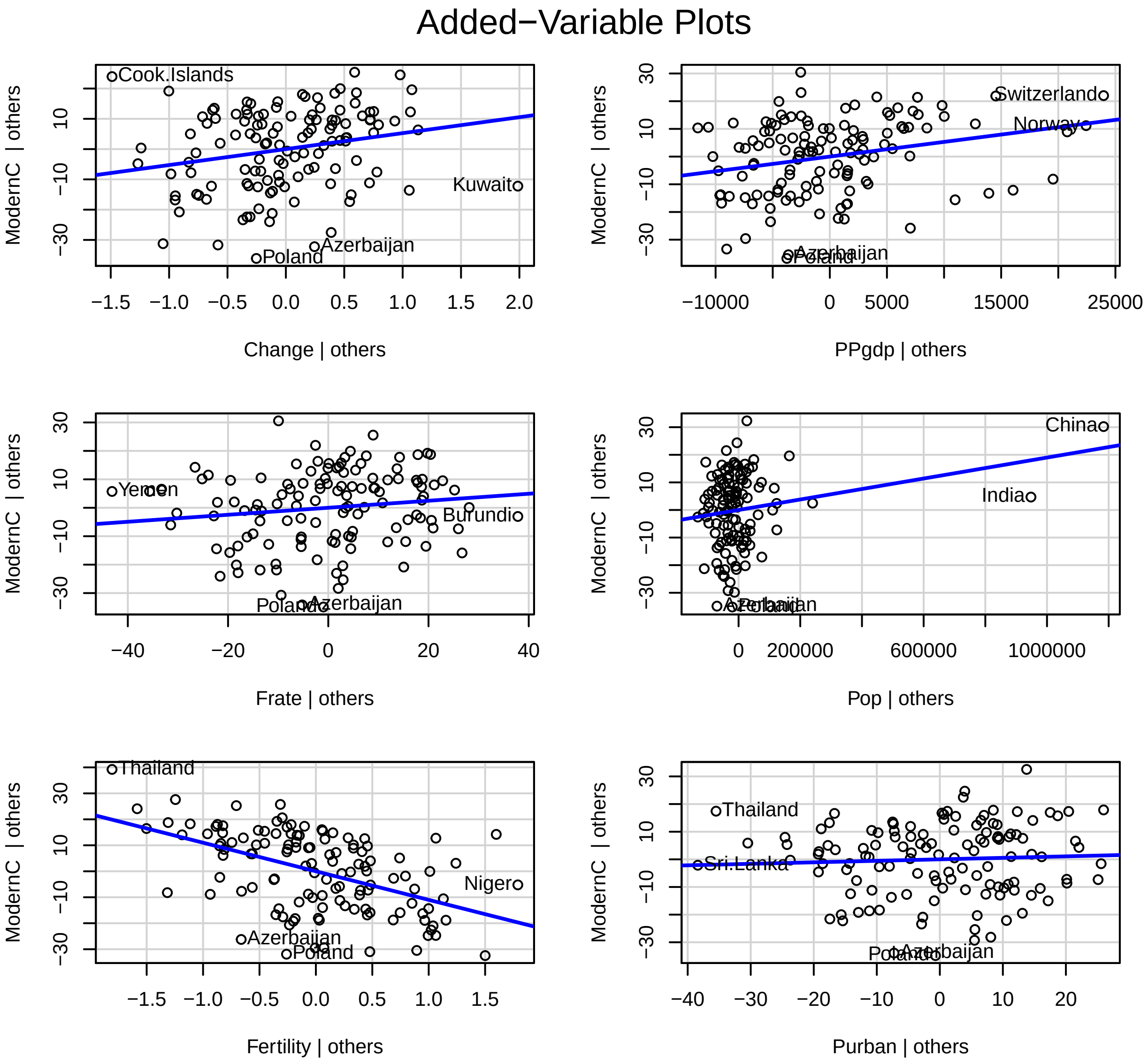


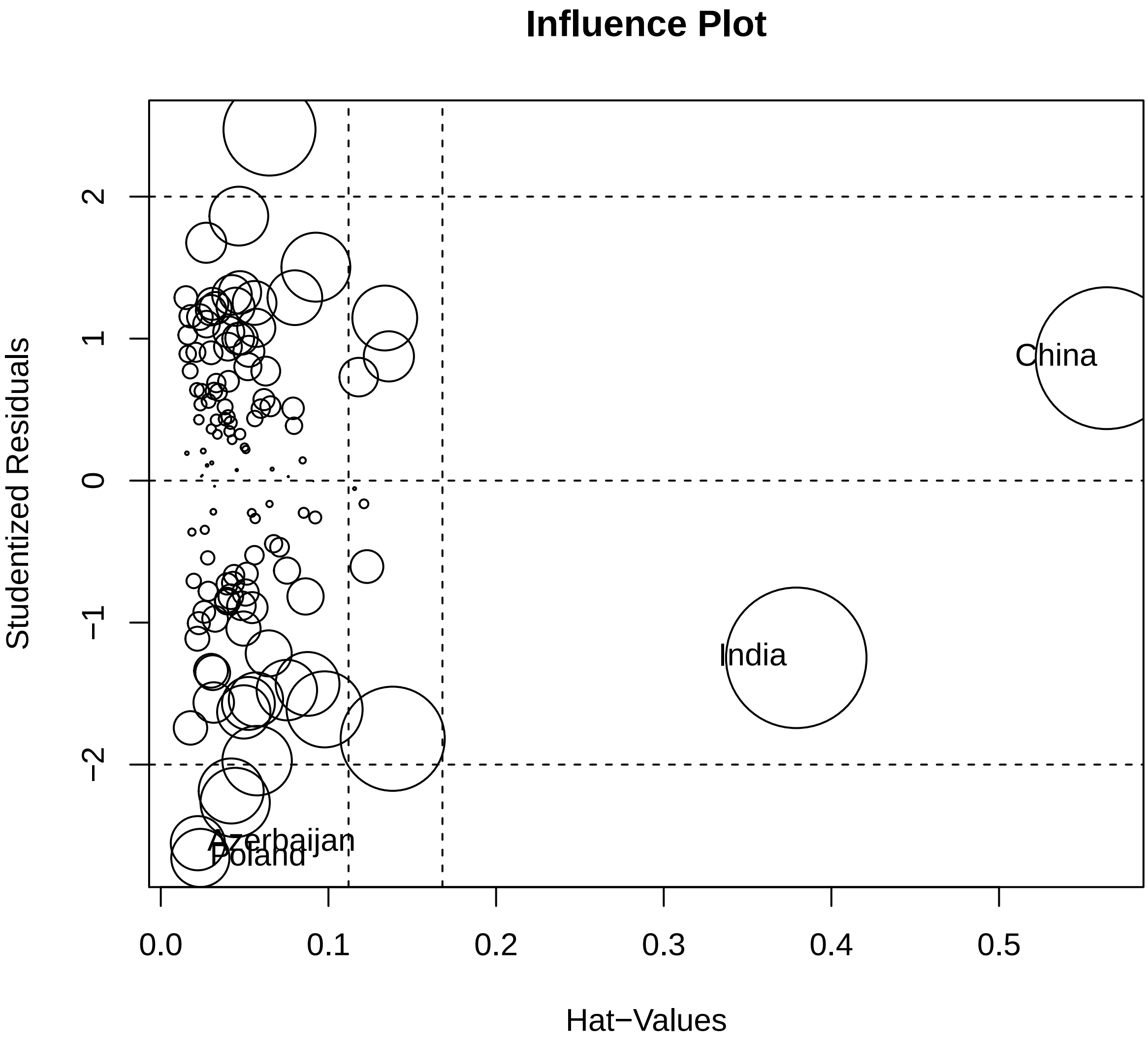
```
## [1] 125
```

```
## [1] 210
```

There is funnel shape in plot “Residual-Fitted values”. This shows variance of error term may not be constant. We might consider log transforms. Also, the standarized residual of “Poland”, “Cook.Islands”, and “Azerbaijian” are obviously bigger than other points. They may be outliers. “China” and “India” are points with high leverage. We can see that 125 observations are used in the model fitting, while originally we have 210 observations in total.

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?


Added−Variable Plots
ModernC | others
Change | others
−1.5 −1.0 −0.5 0.0 0.5 1.0 1.5 2.0
−30 −10 10
Cook.Islands
Kuwait
Azerbaijan
Poland
ModernC | others
PPgdp | others
−10000 0 5000 15000 25000
−30 −10 10 30
Switzerland
Norway
Azerbaijan
Poland
ModernC | others
Frate | others
−40 −20 0 20 40
−30 −10 10 30
Yemen
Burundi
Poland
Azerbaijan
ModernC | others
Pop | others
0 200000 600000 1000000
−30 −10 10 30
China
India
Azerbaijan
Poland
ModernC | others
Fertility | others
−1.5 −1.0 −0.5 0.0 0.5 1.0 1.5
−30 −10 10 30
Thailand
Niger
Azerbaijan
Poland
ModernC | others
Purban | others
−40 −30 −20 −10 0 10 20
−30 −10 10 30
Thailand
Sri.Lanka
Poland
Azerbaijan



```
##                StudRes        Hat      CookD
## Azerbaijan -2.5537604 0.02210682 0.02012037
## China       0.8623741 0.56418981 0.13783693
## India      -1.2481804 0.37905760 0.13522653
## Poland     -2.6575494 0.02359719 0.02319194
```

For the avplot between ModernC and Pop, the residual Pop values are concenterated in the left area. We may consider log transformation of the Pop variable.

Also, from the avplots, we can see that Kuwait and Cook Islands may be influential on Change. Switzerland and Norway may be influential on PPgdp. Yemen and Burundio may be influential on Frate. India and China may be influential on Pop. Thailand and Niger may be influential on Fertility. Sri Lanka and Thailand may be influential on Purban.

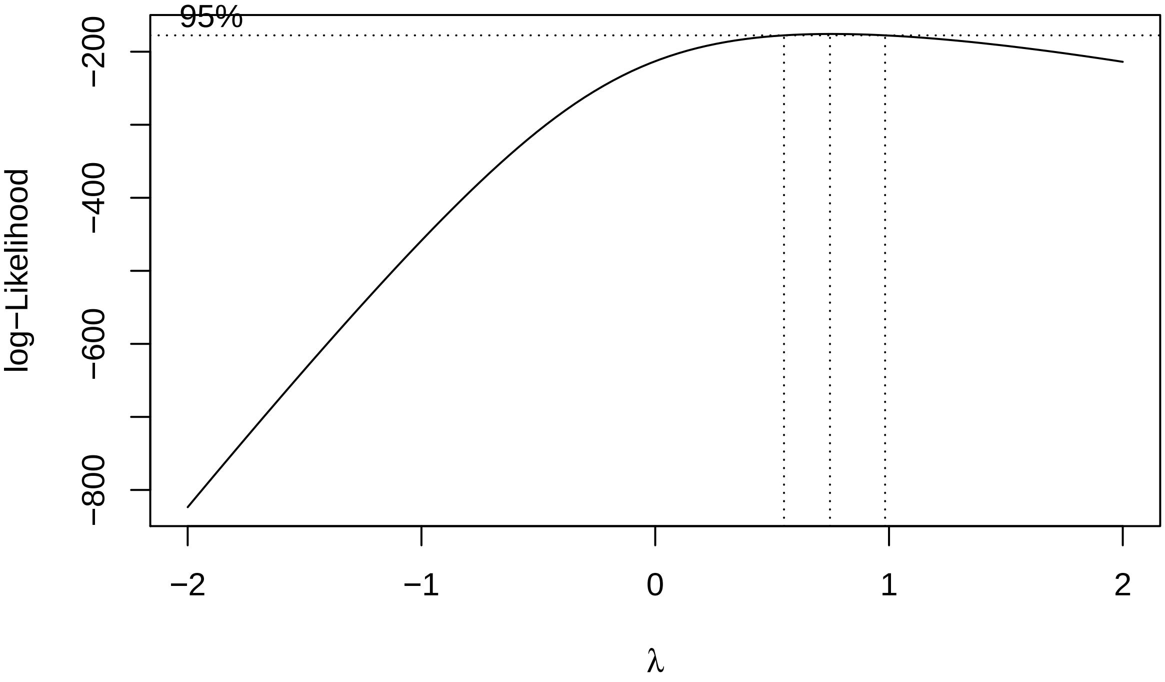
6. Using the Box-Tidwell `car::boxTidwell` or graphical methods find appropriate transformations of the predictor variables to be used as predictors in the linear model. If any predictors are negative, you may need to transform so that they are non-negative. Describe your method and the resulting transformations.

```
##        MLE of lambda Score Statistic (z) Pr(>|z|)
## Change      -1.21571             -2.9009 0.003721 **
## Pop          0.41276             -0.6563 0.511646
## PPgdp       -0.11773             -1.0037 0.315541
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## iterations =  26
```

We use the function boxTidwell in library car to find appropriate transformations. And we minus min(Change) and add 1 to variable "Change" to make it positive.

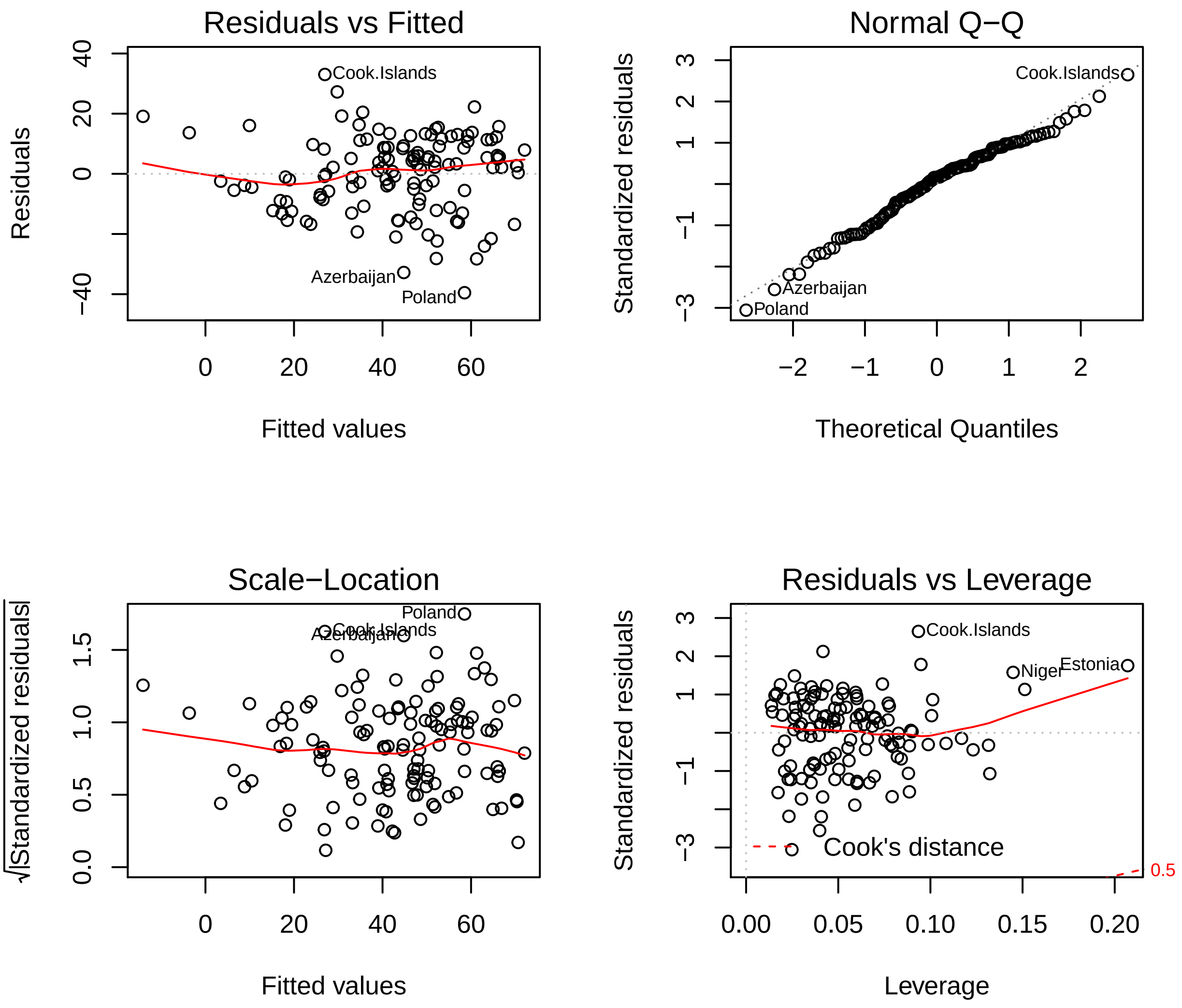
The result shows that we should take reciprocal of the variable ""Change", and take log transformation of "PPgdp" and "Pop".

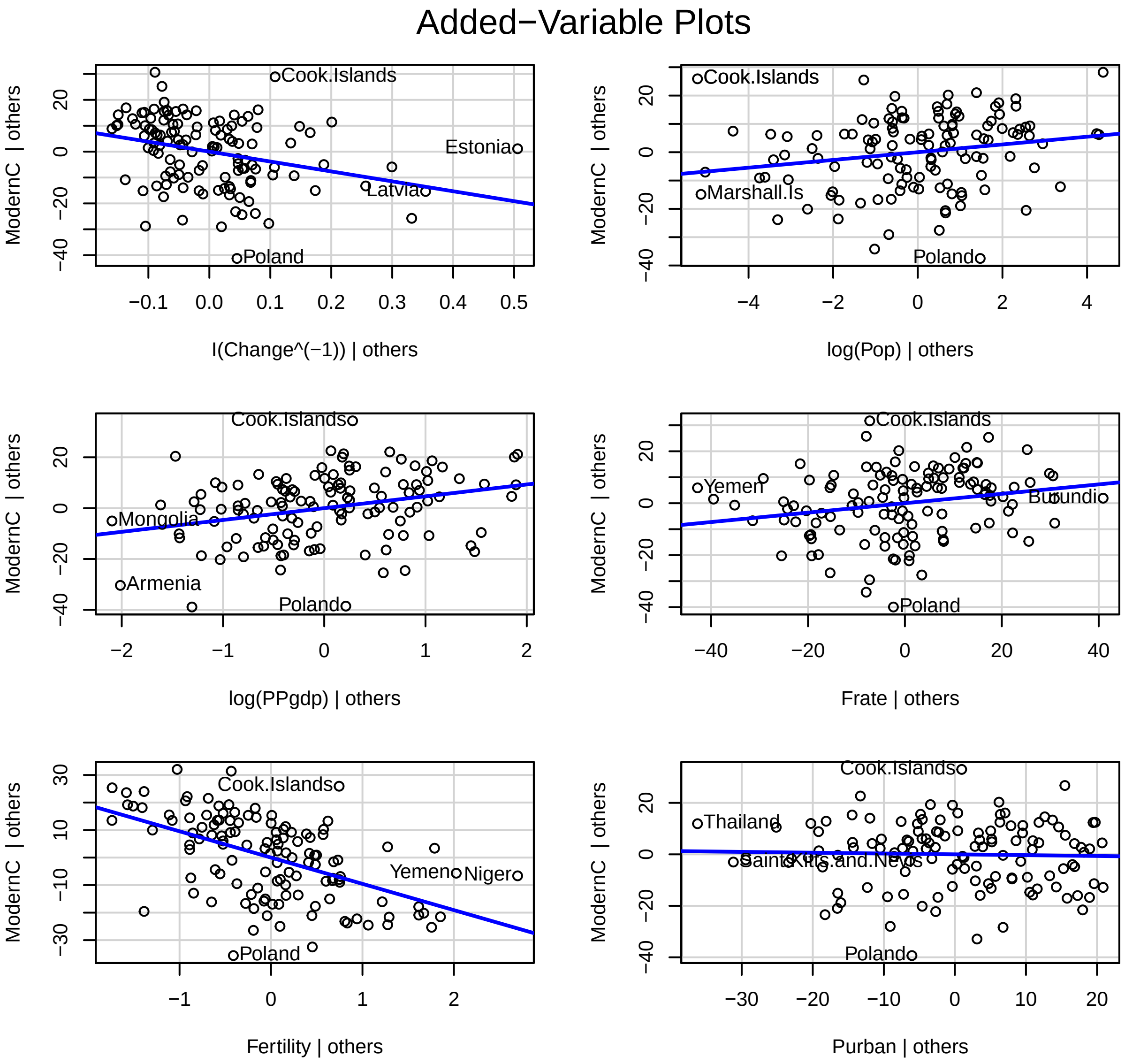
7. Given the selected transformations of the predictors, select a transformation of the response using `MASS::boxcox` or `car::boxCox` and justify.



Using the boxcox method, we see that $\lambda = 1$ is very close to right side of the 95% confidence interval, and is extremely close to the maximum log-likelihood, which suggests we can set $\lambda = 1$ and don't transform Y. We also tried to transfrom response with $\lambda = 0.75$ which realizes the maximum likelihood. But it didn't show much difference in their diagnostic plots or adjusted R-square. So we would not transform the response.

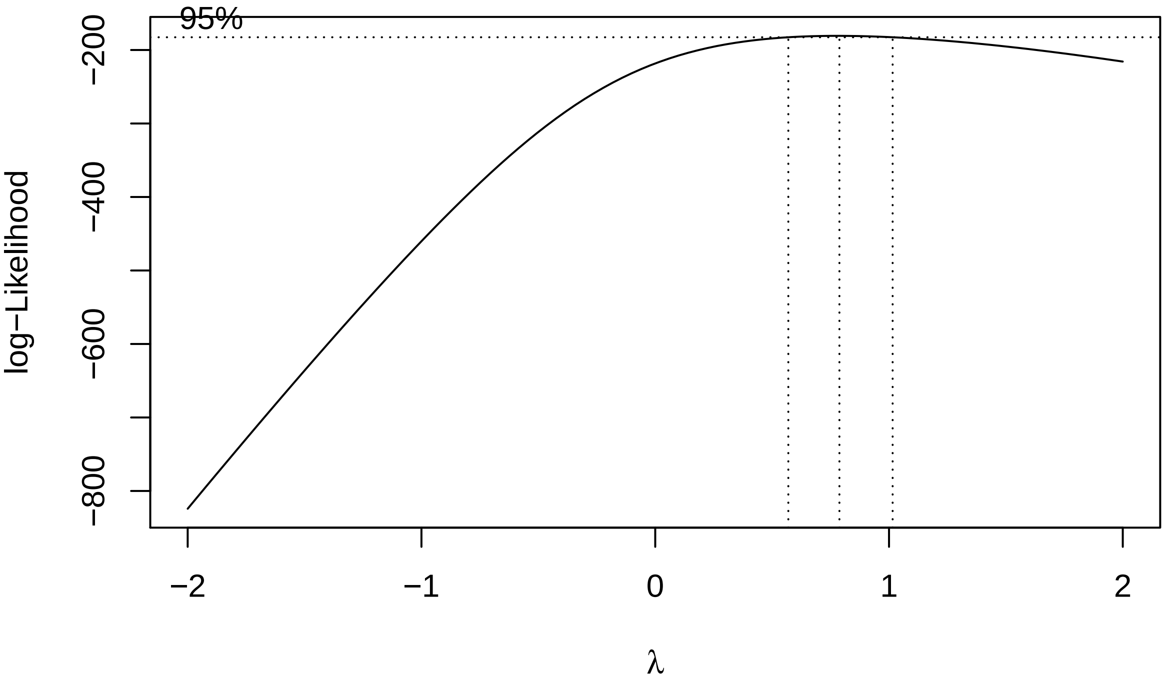
8. Fit the regression using the transformed variables. Provide residual plots and added variables plots and comment. If you feel that you need additional transformations of either the response or predictors, repeat any steps until you feel satisfied.


Residuals vs Fitted
Residuals
Fitted values
40
20
0
−40
0
20
40
60
Cook.Islands
Azerbaijan
Poland
Normal Q−Q
Standardized residuals
Theoretical Quantiles
3
2
1
−1
−3
−2
−1
0
1
2
Cook.Islands
Azerbaijan
Poland
Scale−Location
√|Standardized residuals|
Fitted values
1.5
1.0
0.5
0.0
0
20
40
60
Poland
Cook.Islands
Azerbaijan
Residuals vs Leverage
Standardized residuals
Leverage
3
2
1
−1
−3
0.00
0.05
0.10
0.15
0.20
Cook.Islands
Niger
Estonia
Cook's distance
0.5



We can see that after taking log of Pop, it shows much better linear relationship with ModernC in added-variable plot. Also, the residual plots showed that the new model fit the data better, with obvious improvements in the normal Q-Q plot. The high leverages also decrease a lot.

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?

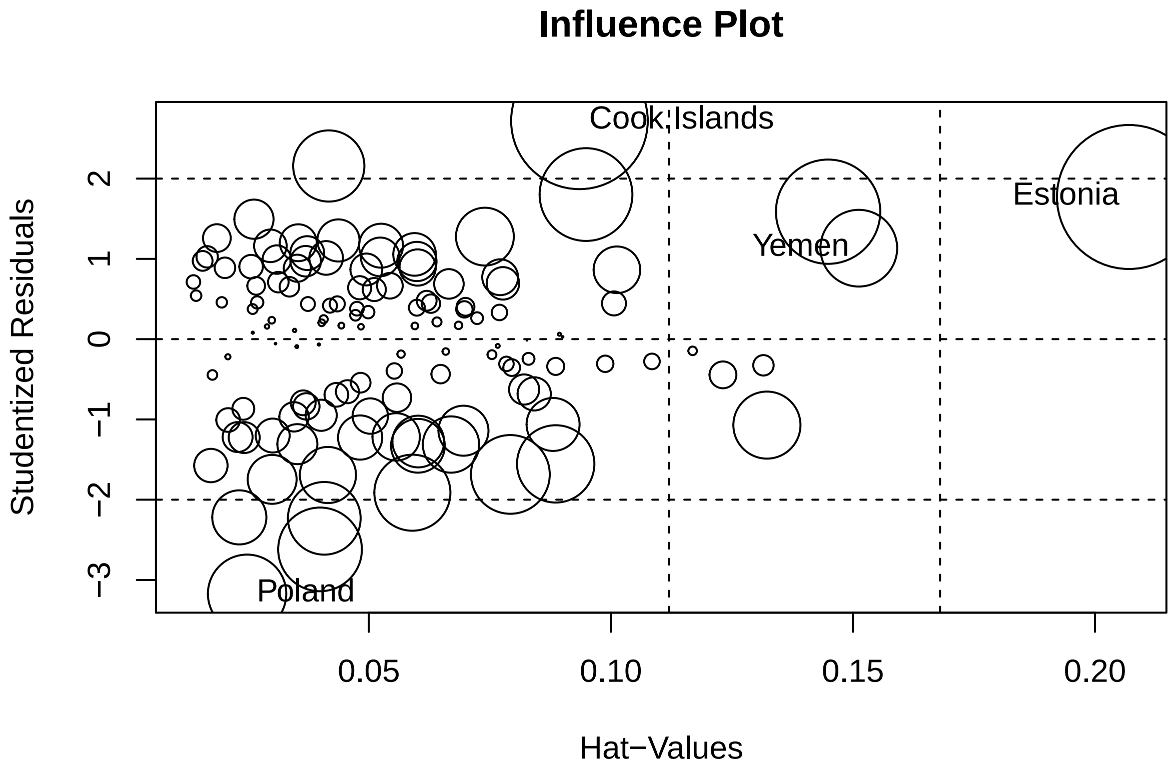


```
##        MLE of lambda Score Statistic (z) Pr(>|z|)
## Change      -1.21571             -2.9009 0.003721 **
## Pop          0.41276             -0.6563 0.511646
## PPgdp       -0.11773             -1.0037 0.315541
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## iterations =  26
```

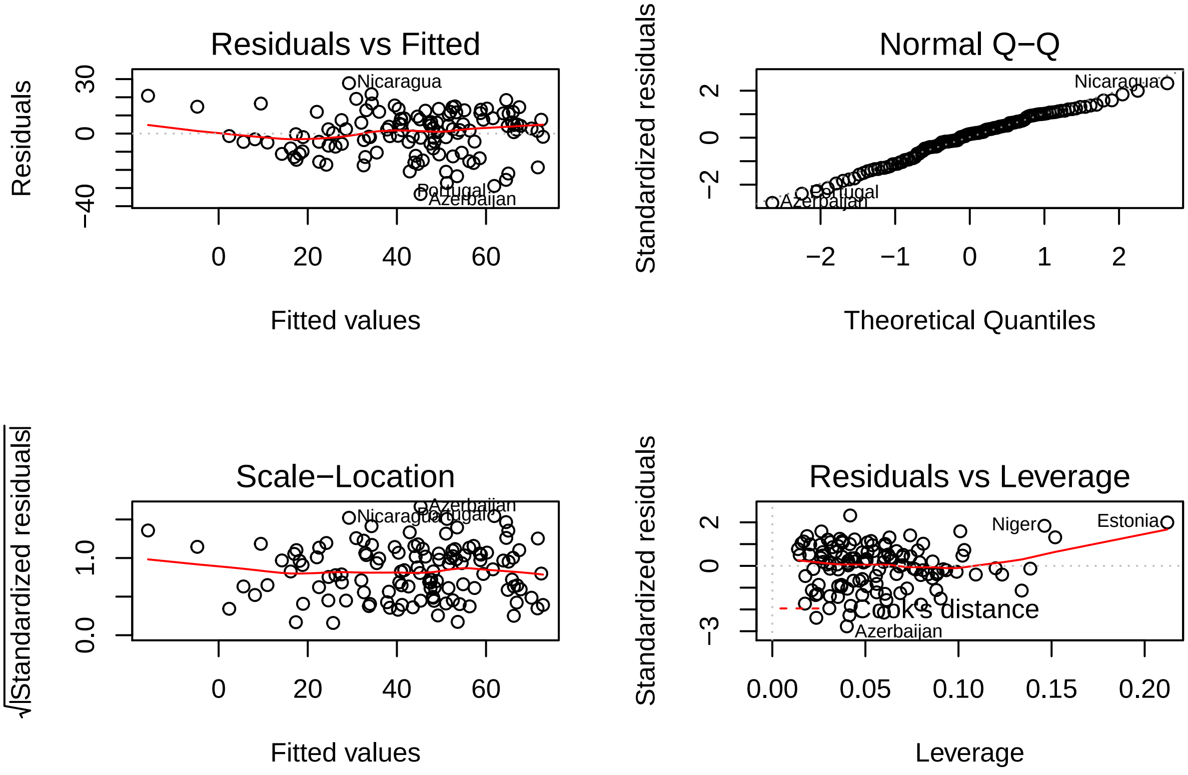
we see that $\lambda = 1$ is in the 95% confidence interval, and is extremely close to the maximum log-likelihood, which suggests we can set $\lambda = 1$ and don't transform Y. Then for the predictors, the result shows that we should take reciprocal of the variable "“Change”, and take log transformation of “PPgdp” and “Pop”.

So the model we get is totally the same as the model in 8.

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.



```
##                  StudRes        Hat      CookD
## Cook.Islands   2.719601 0.09350532 0.10338515
## Estonia        1.771267 0.20703149 0.11493549
## Poland        -3.171816 0.02484809 0.03401030
## Yemen          1.133305 0.15121377 0.03260942
```



From the result, we can see that for points “Cook.Islands” and “Poland”, their studentized residual is outside the ±2 range. So they are considered outliers, statistically significant at 95% level. All the points’ CookDistance is smaller than 1. So there are no influential points in the data.

So we delete outliers “Cook.Islands” and “Poland”. After removing the outliers, the new model is slightly better, although it didn’t make much difference. Because these two outliers are not influential.

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!

	2.5 %	97.5 %
(Intercept)	-4.2064631	57.0986737
I(Change^(-1))	-59.9628672	-19.5302624
log(Pop)	0.7453949	3.0915993
log(PPgdp)	2.0514967	7.2206136
Frate	0.0492671	0.3238701
Fertility	-12.6796773	-7.4342077
Purban	-0.2192580	0.1265653

We didn’t make transformations on the response “ModernC”. So it is already in its original units. So when reciprocal of “change” increase one unit, then there is 95% possibility that the response “ModernC” will decrease between 19.5302624 and 59.9628672. When log(Pop) increase one unit, then there is 95% possibility that the response will increase between 0.7453949 and 3.0915993. Other coefficients’ interpretations are similar.
Variables including “Change”, “Pop”,“PPgdp”, and “Frate” have positive relationship with “ModernC”, while “Fertility” and “Purban” have negative relationship with “ModernC”.

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

In my final model, I take reciprocal of the variable "“Change”, and take log transformation of “PPgdp” and “Pop”. And I delete two outliers, “Poland” and “Cook.Islands”. The regression result shows that variables including “Change”, “Pop”,“PPgdp”, and “Frate” have positive relationship with “ModernC”, while “Fertility” and “Purban” have negative relationship with “ModernC”.

The two outliers that I delete have studentized residual outside the ±2 range. So they are considered outliers, statistically significant at 95% level. After removing the two outliers, the new model is better as reflected in residual plots.

```
##
## Call:
## lm(formula = ModernC ~ I(Change^(-1)) + log(Pop) + log(PPgdp) +
##     Frate + Fertility + Purban, data = UN3R)
##
## Coefficients:
##    (Intercept)  I(Change^(-1))        log(Pop)      log(PPgdp)
##       26.44611       -39.74656         1.91850         4.63606
##          Frate       Fertility          Purban
##        0.18657       -10.05694        -0.04635
```

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.*

Suppose we regress y on x_1 and x_2, and x_3 is added variable. $e_{(y)}$ means residual of regression y on x_1 and x_2, $e_{(x_3)}$ means residual of regression x_3 on x_1 and x_2.

$$\overrightarrow{e_{(y)}} = \hat{\beta}_0 \overrightarrow{i} + \hat{\beta}_1 \overrightarrow{e_{(x_3)}}$$

$$(I-H)Y = \hat{\beta}_0 \overrightarrow{i} + \hat{\beta}_1 (I-H)x_3$$

where $\overrightarrow{i}$ is a $n*1$ vector of 1 and $H = X(X^TX)^{-1}X^T$ since

$$\hat{\beta}_1 = (((I-H)x_3)^T(I-H)x_3)^{-1}((I-H)x_3)^T(I-H)Y$$

and $I-H$ is symmetric and idempotent, thus we have

$$\begin{aligned}(I-H)Y &= \hat{\beta}_0 \overrightarrow{i} + (((I-H)x_3)^T(I-H)x_3)^{-1}((I-H)x_3)^T(I-H)Y(I-H)x_3 \\ &= \hat{\beta}_0 \overrightarrow{i} + [x_3^T(I-H)(I-H)x_3]^{-1}x_3^T(I-H)(I-H)Y(I-H)x_3 \\ &= \hat{\beta}_0 \overrightarrow{i} + [x_3^T(I-H)x_3]^{-1}x_3^T(I-H)Y(I-H)x_3\end{aligned}$$

we know that both $[x_3^T(I-H)x_3]^{-1}$ and $x_3^T(I-H)Y$ is a $1*1$ scalar, so we can move them anywhere we want. so

$$(I-H)Y = \hat{\beta}_0 \overrightarrow{i} + (I-H)x_3[x_3^T(I-H)x_3]^{-1}x_3^T(I-H)Y$$

let's left multiply x_3^T on both sides, then we get

$$\begin{aligned}x_3^T(I-H)Y &= x_3^T\hat{\beta}_0 \overrightarrow{i} + x_3^T(I-H)x_3[x_3^T(I-H)x_3]^{-1}x_3^T(I-H)Y \\ &= x_3^T\hat{\beta}_0 \overrightarrow{i} + x_3^T(I-H)Y\end{aligned}$$

so we have

$$x_3^T\hat{\beta}_0 \overrightarrow{i} = 0$$

i.e.,

$$\sum_{i=1}^{n} x_3^i \hat{\beta}_0 = 0$$

since the sum of elements in x_3 won't always be zero, so the intercept in the added variable scatter plot $\hat{\beta}_0$ will always be zero.

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 will take variable "Fertility" as example.

```
##    (Intercept) I(Change^(-1))       log(Pop)     log(PPgdp)          Frate
##    26.44610533   -39.74656482     1.91849709     4.63605514     0.18656860
##      Fertility         Purban
##   -10.05694249    -0.04634634
```

```
##   (Intercept) Fertility_e_x
##  4.027019e-16 -1.005694e+01
```

We can see that the coefficient of "Fertility" in the added variable regression is -10.05694, which is exactly the same as the Fertility's coefficient in the full model.