HW2 STA521 Fall18

Shuai Yuan, sy144, Kolin96

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?

```
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                      
```

Six variables have missing data: `ModernC`, `Change`, `PPgdp`, `Frate`, `Pop` and `Fertility`.

Quantitative: `ModernC`, `Change`, `PPgdp`, `Frate`, `Pop`, `Fertility`, `Purban`

Qualitative: (none)

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

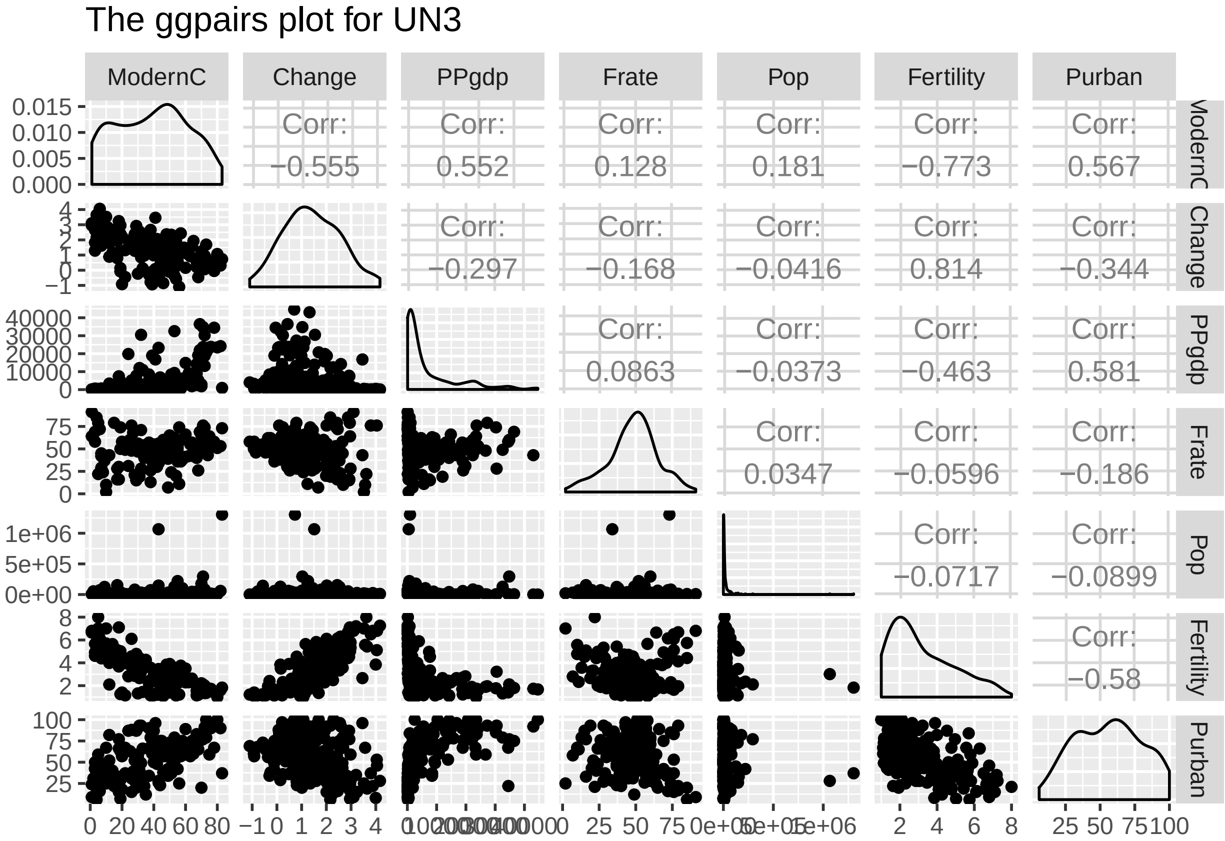
```
mean_result <- sapply(UN3, mean, na.rm=TRUE)
std_result <- sapply(UN3, sd, na.rm=TRUE)
kable(cbind(mean_result, std_result), col.names = c("mean", "std."), digits = 4)
```

	mean	std.
ModernC	38.7171	22.6366
Change	1.4184	1.1331
PPgdp	6527.3881	9325.1886
Frate	48.3054	16.5324
Pop	30281.8714	120676.6945
Fertility	3.2140	1.7069

	mean	std.
Purban	56.2000	24.1098

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
gp = ggpairs(UN3, titl="The ggpairs plot for UN3")
print(gp, progress=FALSE)
```



If we try to predict ModernC, it may be better to use the predictors Change, Frate, Fertility and Purban, since they are linear correlated.

Potential outliers: the Pop variable has two extremely large data point.

Nonlinear relationships: PPgdp seems to grow exponentially as ModernC increases. Frate also has some noised nonlinear relationship.

Transformation needed: PPgdp should be transformed to log scale.

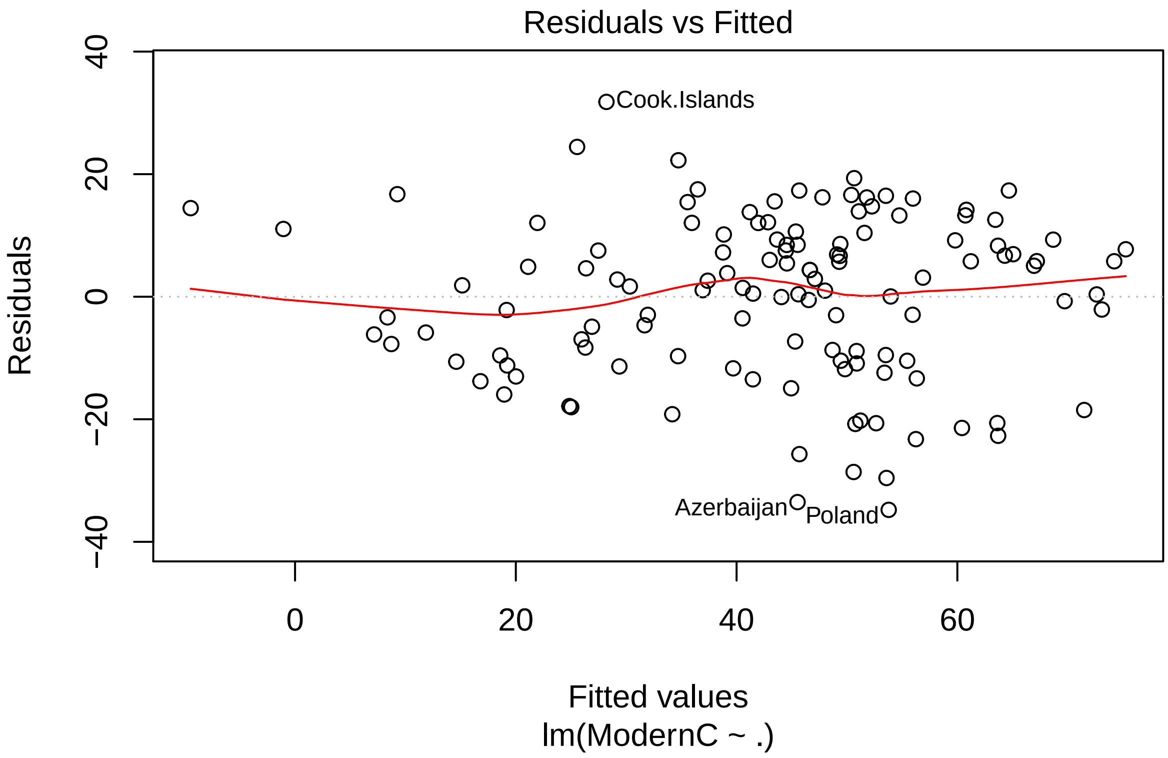
Model Fitting

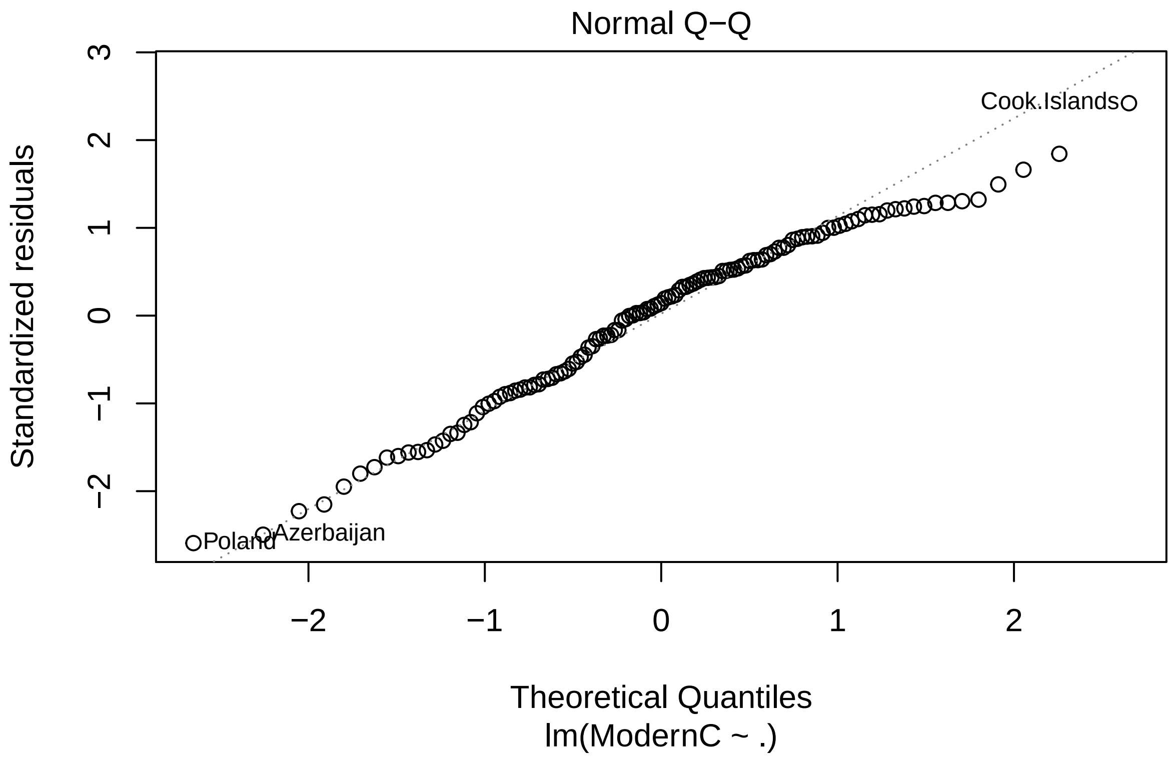
4. Use the lm() function to perform a multiple linear regression with ModernC as the response and all other variables as the predictors, using the formula ModernC ~ ., where the . includes all remaining variables in the dataframe. Create diagnostic residual plot from the linear model object and comment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

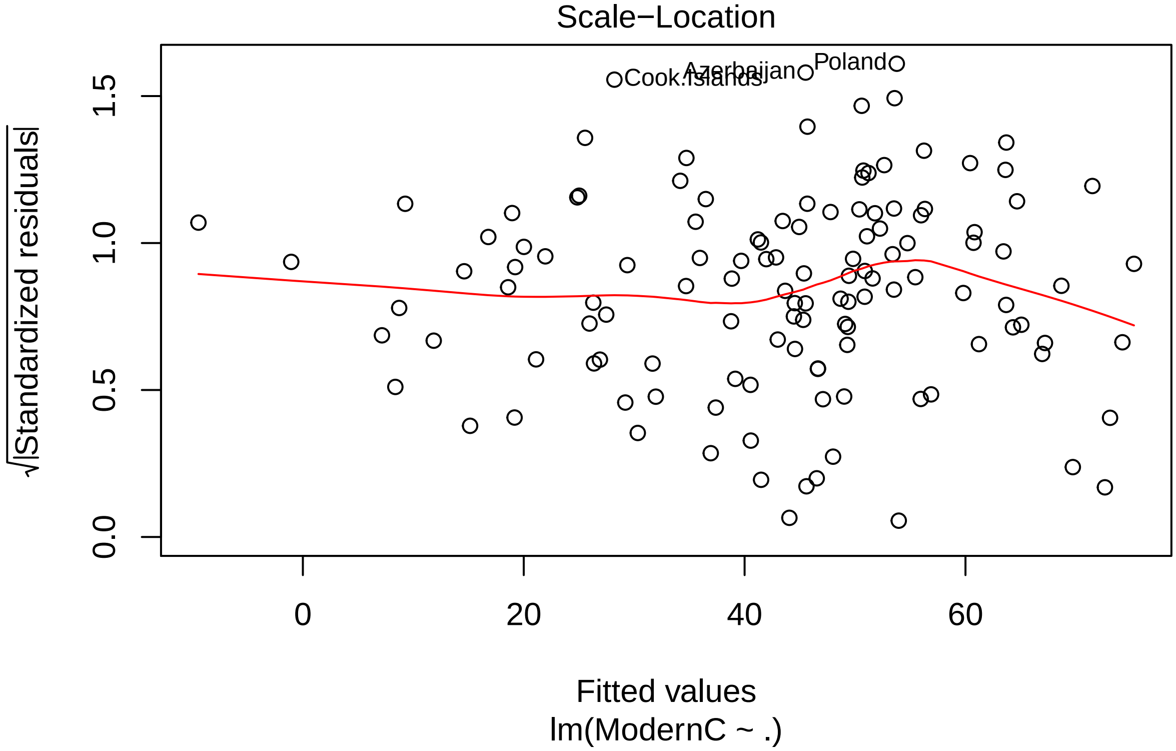
```
mymodel = lm(ModernC ~ ., data=UN3)
summary(mymodel)
```

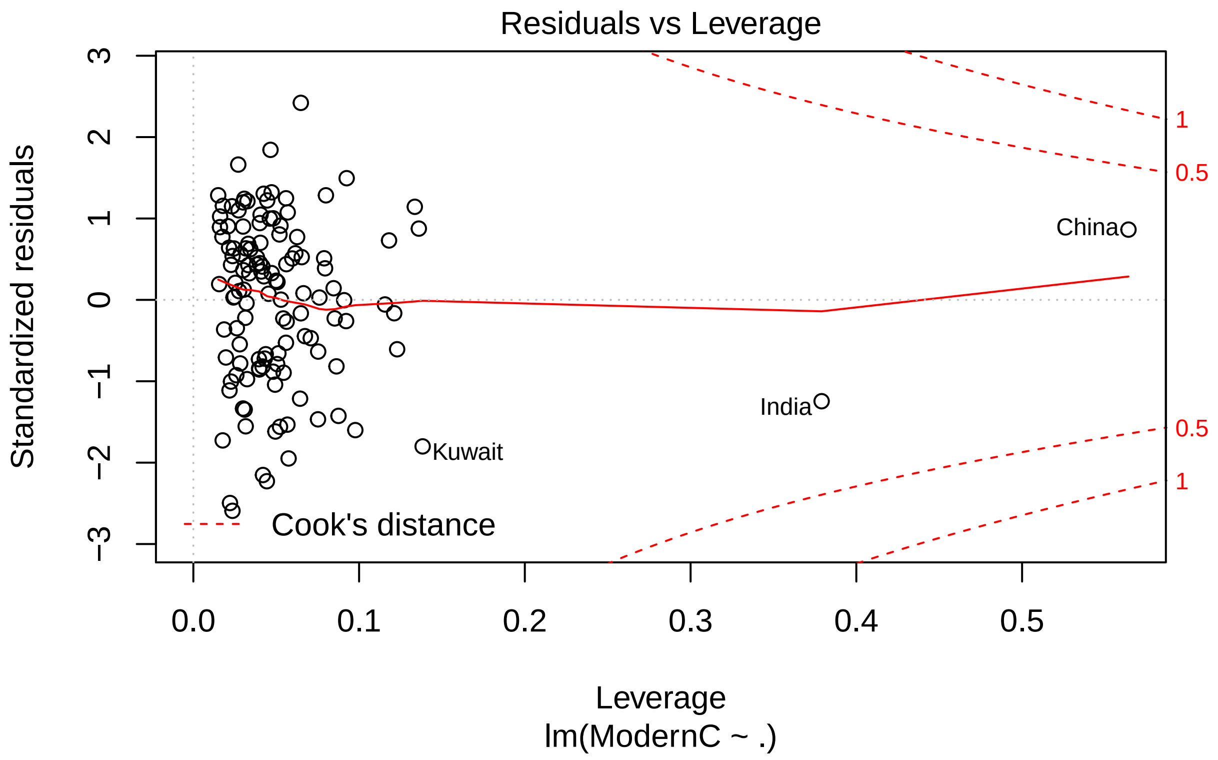
```
## 
## Call:
## lm(formula = ModernC ~ ., data = UN3)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -34.781  -9.698   1.858   9.327  31.791 
## 
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  5.529e+01  9.467e+00   5.841 4.69e-08 ***
## 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
```

```
# diagnostic residual plot
plot(mymodel)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(ModernC ~ .)
3
2
1
0
−1
−2
−2
−1
0
1
2
Cook.Islands
Poland
Azerbaijan


Scale−Location
√|Standardized residuals|
Fitted values
lm(ModernC ~ .)
1.5
1.0
0.5
0.0
0
20
40
60
Cook.Islands
Azerbaijan
Poland

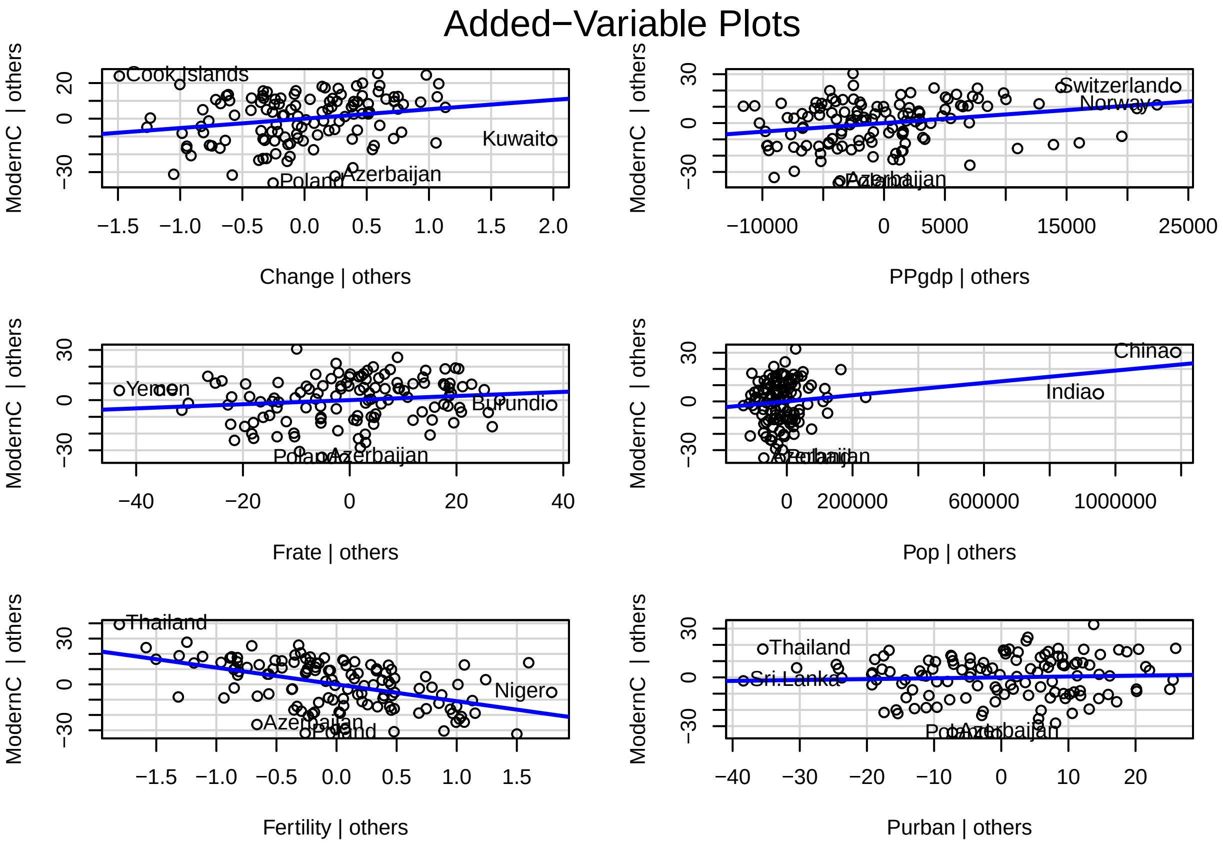


Comments: The linear model works fairly well, since the R-squared values are high (around 0.6) and the p-value is low (<2.2e-16). There are noises, but the residuals are generally independent of fitted values and have mean 0. For the Normal Q-Q plot, the higher quantiles can't match perfectly, but it is still not bad. Cook's distances are all below 0.5; China and Indea are potential outliers.

How many observations are used: 125, since the other 85 observations are deleted due to missingness.

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?

```
car::avPlots(mymodel)
```



Transformations that are needed: The `Pop | others` plot shows that we need to transform `pop` to log scale, because all data points are now concentrated near 0. `PPgdp` should also be transformed since the current datapoints are also not evenly distributed.

Influential localities for each term are all shown in the figures:

`Change`: Cook islands, Kuwaito, Poland, Azerbaijan;

`PPgdp`: Switzerland, Norway, Azerbaijan;

`Frate`: Yeman, Burundi, Poland, Azerbaijan;

`Pop`: China, Indea, Azerbaijan;

`Fertility`: Thailand, Niger, Poland, Azerbaijan;

`Purban`: Thailand, Sri.Lanka, Poland, Azerbaijan.

Note that Azerbaijan appears in all the lists.

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.

```
Change_min = min(UN3["Change"], na.rm = TRUE) - 1e-4 # to avoid numerical issues about 0
UN3_all = UN3 %>%
  mutate(name = row.names(UN3)) %>%
  na.omit() %>%
  mutate(Change_trans = Change - Change_min)
```

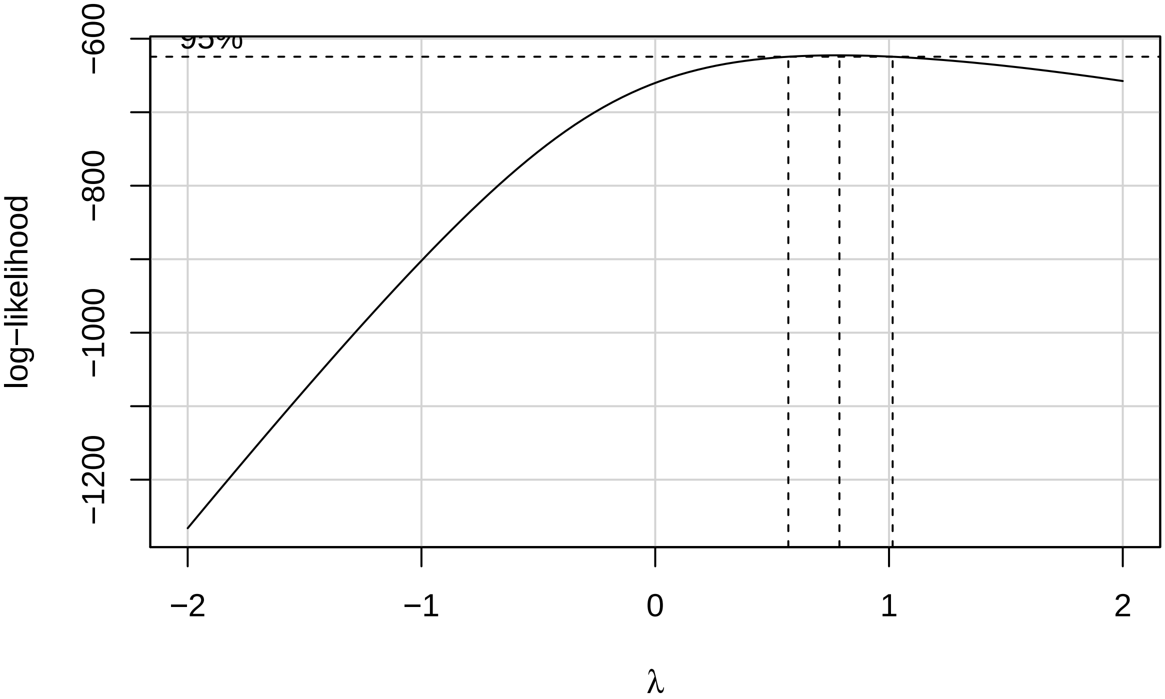
```
UN3_trans = UN3_all %>% select(c(ModernC, Change_trans, PPgdp, Frate, Pop, Fertility, Purban))
car::boxTidwell(ModernC ~  Pop + PPgdp, ~ Fertility + Change_trans + Frate + Purban, data=UN3_trans)
```

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

We first transform `Change` by subtracting its minimum value to make it nonnegative, since subtractions would not harm our linear assumption. Then, we transform `Pop` and `PPgdp`, the MLE of lambda computed by `car::boxTidwell` are 0.40749 and -0.12921.

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

```
mymodel = lm(ModernC ~ ., data=UN3_trans)
trials = car::boxCox(mymodel)
```



```
lambda = trials$x[trials$y == max(trials$y)]
```

By outputing car::boxCox(mymodel), we may find the best lambda is around 0.7879.

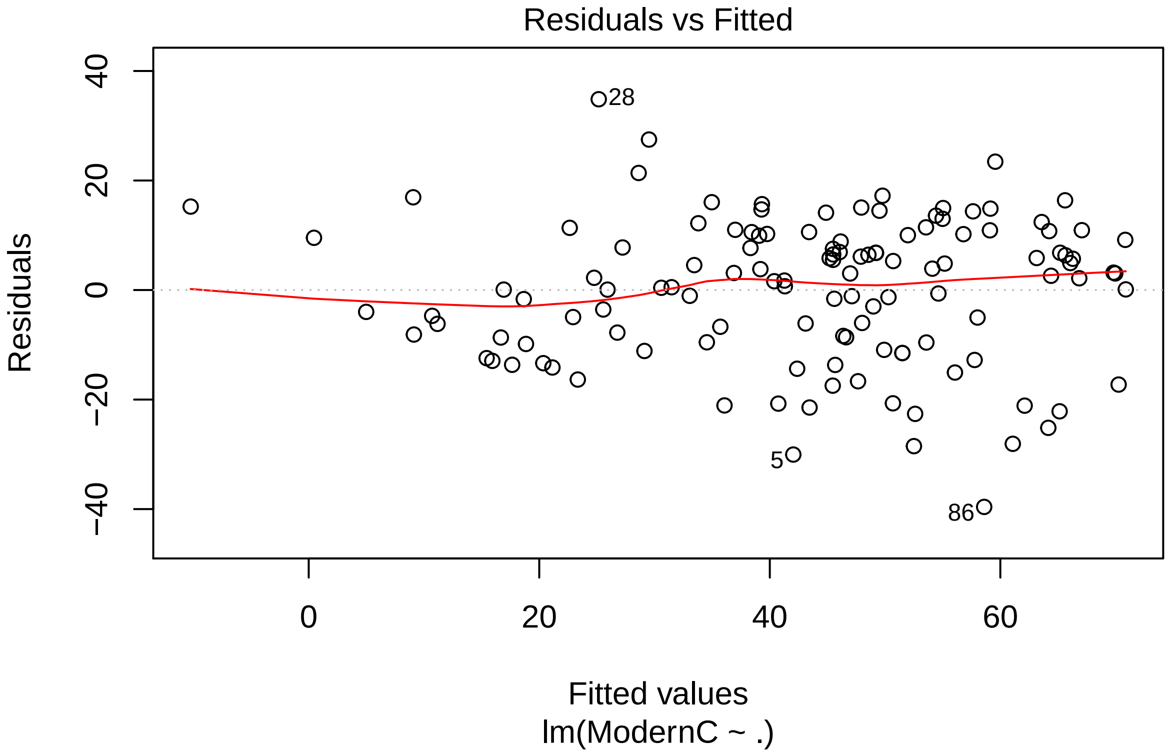
8. Fit the regression using the transformed variables. Provide residual plots and added variables plots and comment. If you feel that you need additional transformations of either the response or predictors, repeat any steps until you feel satisfied.

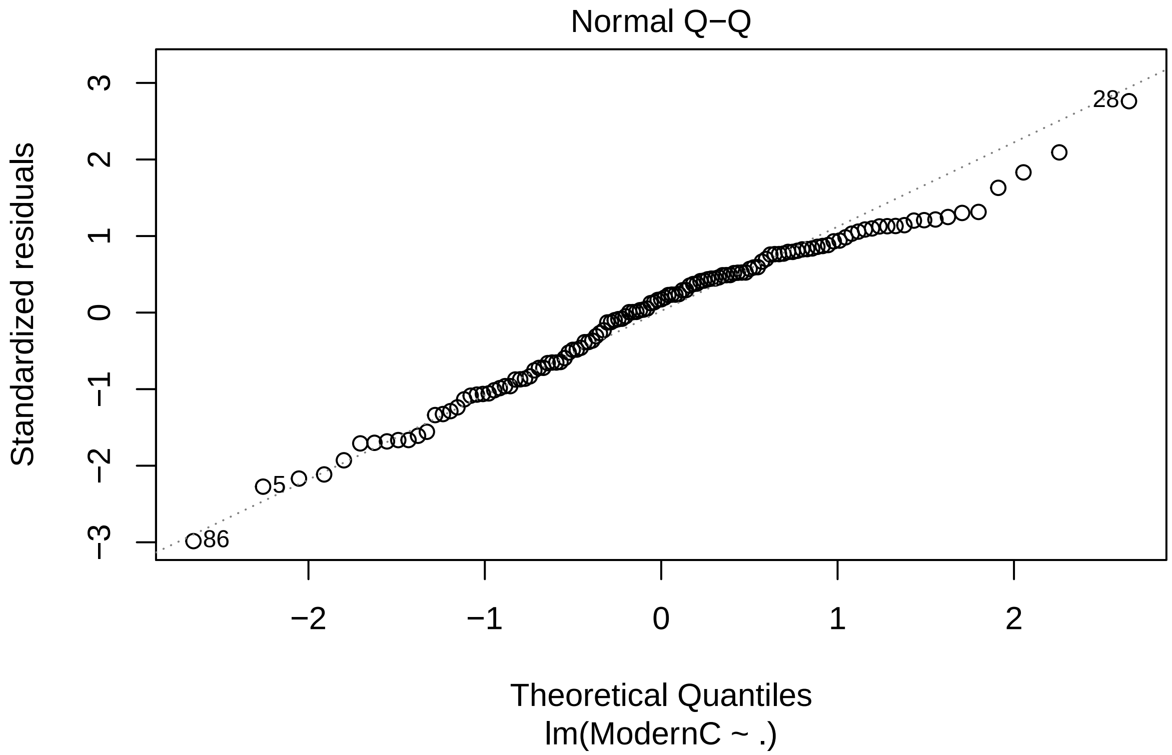
```
UN3_all = UN3_all %>%
  mutate(Pop_log = log(Pop)) %>%
  mutate(PPgdp_log = log(PPgdp))

UN3_trans = UN3_all %>% select(c(ModernC, Change_trans, PPgdp_log, Frate, Pop_log, Fertility, Purban))

mymodel = lm(ModernC ~ ., data=UN3_trans)
```

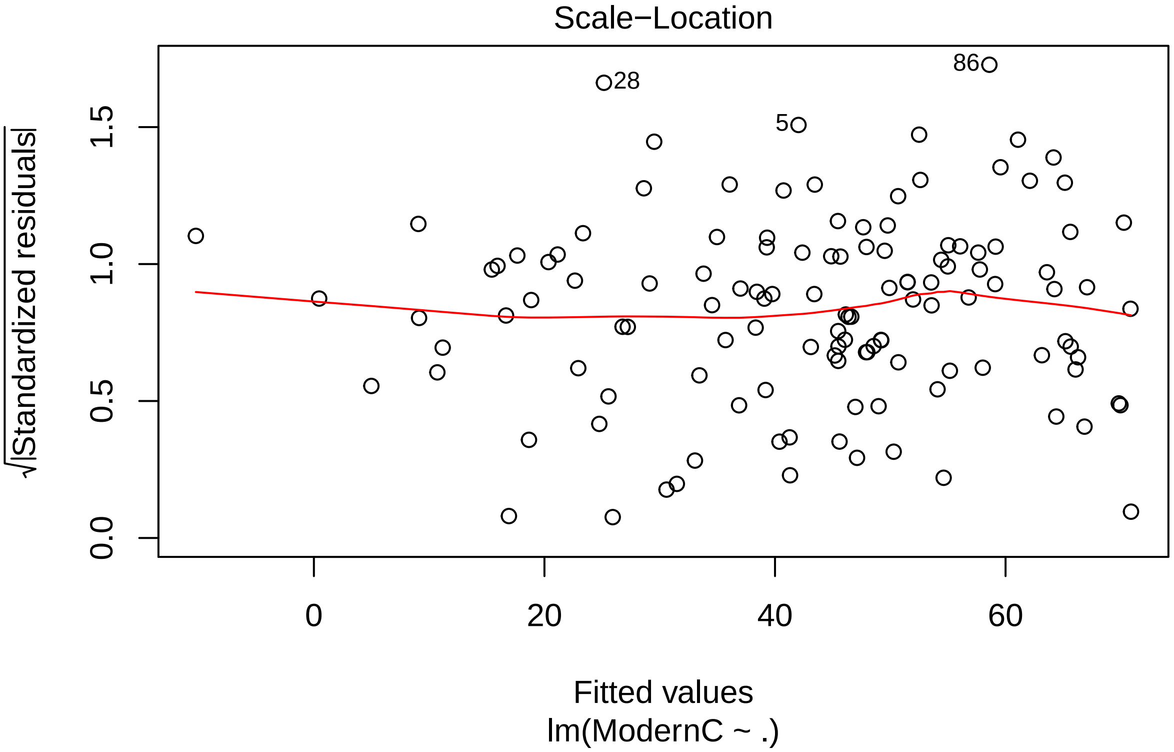
```r
summary(mymodel)
```

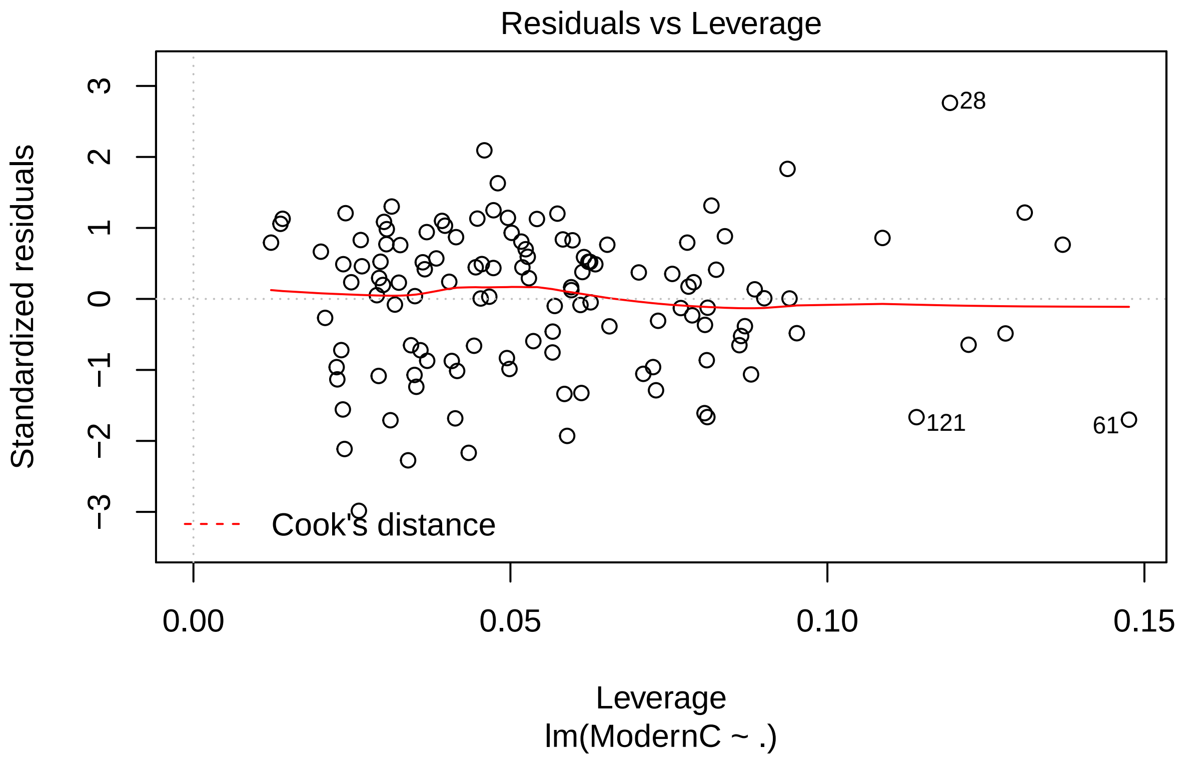
```
##
## Call:
## lm(formula = ModernC ~ ., data = UN3_trans)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -39.597  -9.540   2.238  10.024  34.840
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.37728   14.19424  -0.097 0.922866
## Change_trans  4.99296    2.07709   2.404 0.017781 *
## PPgdp_log     5.50728    1.40505   3.920 0.000149 ***
## Frate         0.18939    0.07711   2.456 0.015500 *
## Pop_log       1.47207    0.62875   2.341 0.020897 *
## 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
```

```r
plot(mymodel)
```

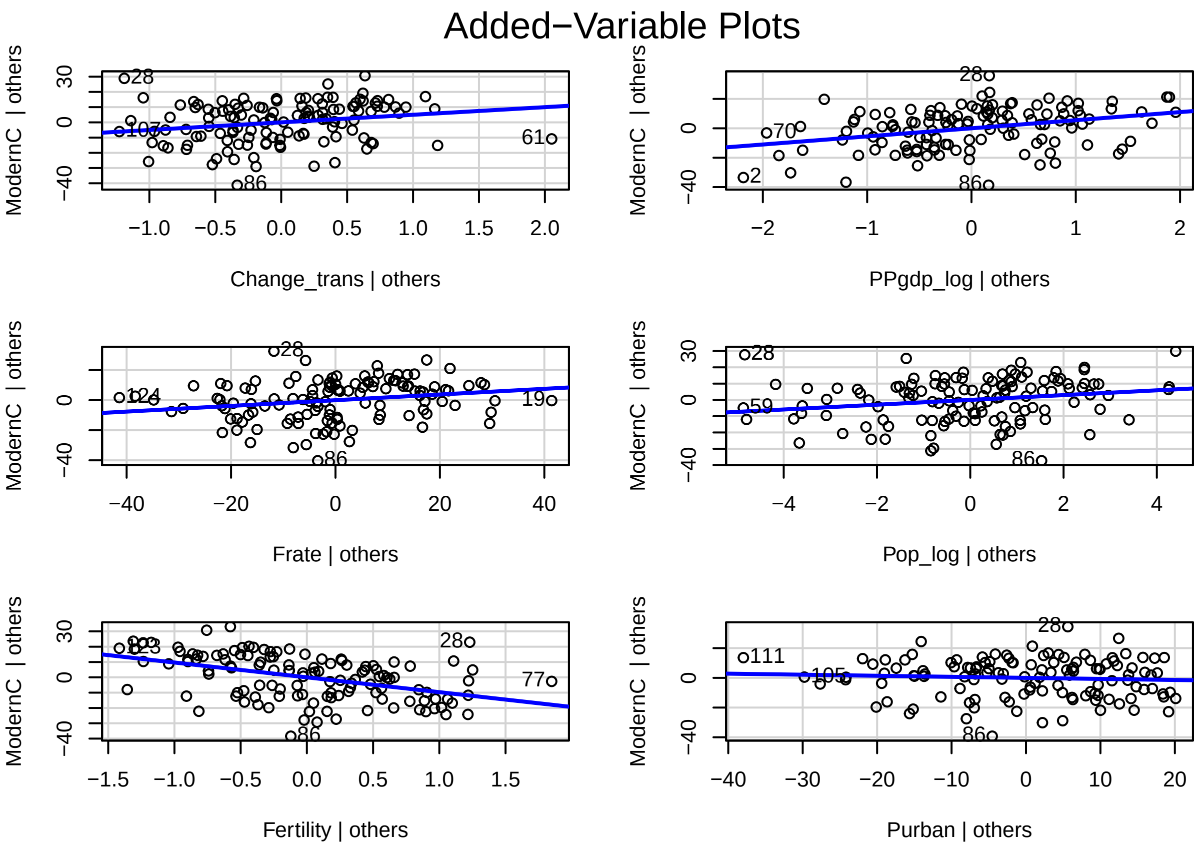



Normal Q−Q
Standardized residuals
3
2
1
0
−1
−2
−3
−2
−1
0
1
2
28
5
86
Theoretical Quantiles
lm(ModernC ~ .)


Scale−Location
√|Standardized residuals|
1.5
1.0
0.5
0.0
0
20
40
60
28
86
5
Fitted values
lm(ModernC ~ .)



```r
car::avPlots(mymodel)
```



We apply log transformation to Pop and PPgdp. The residual plots show that residuals are very small (close to mean 0) and independent to fitted values. The Cook's distance are all very small. In the added variable plot, the `logPop` plot is much more balanced than the original `Pop` plot. The new model works better than our previous 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?

Yes, we would get different models, but these models are not necessarily much better than our previous model, because the original response variable actually has strong linear correalations with many predictors. If we first transform the response variable by log, poly, etc., we still need to transform many predictors accordingly, making the model very messy. Actually, our previous model is good enough.

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.

```
abs.ti = abs(rstudent(mymodel))
pval= 2*(1- pt(max(abs.ti), mymodel$df - 1))
```

We can use student t test with Bonferroni Correction to determine whether there are outliers. The p-value for the observation that has the largest studentized residual is >0.0024, so all p-values should be much greater than 0.05/125, so we claim that there is no outlier. There should be also no influential points, since in the former plots, all Cook's distances are low.

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!

```
df = data.frame(cbind(coefficients(mymodel), confint(mymodel)))
kable(df, col.names = c("coefficients", "2.5%", "97.5%"), digits = 4)
```

	coefficients	2.5%	97.5%
(Intercept)	-1.3773	-29.4857	26.7312
Change_trans	4.9930	0.8797	9.1062
PPgdp_log	5.5073	2.7249	8.2897
Frate	0.1894	0.0367	0.3421
Pop_log	1.4721	0.2270	2.7172
Fertility	-9.6759	-13.1723	-6.1795
Purban	-0.0708	-0.2640	0.1225

Our model: `ModernC` = -1.3773 + 4.9930 `Change_trans` + 5.5073 log(`PPgdp`) + 0.1894 `Frate` + 1.4721 log(`Pop`) - 9.6759 `Fertility` - 0.0708 `Purban` + ϵ

Note: `Change_trans` is computed by subtracting `Change` by Change_min = -1.1001, which keeps the linearity of the equation.

Interpretations:

`(Intercept)`: The predicted value when all parameters approach 0. The confidence interval of the intercept is large, so a considerable noise may exist.

`Change`: Whenever `Change` increases, `ModernC` is expected to increase 4.9930 times the amount.

`PPgdp`: 10% increase in `PPgdp` implies a 5.5073 * log(1.1) = 0.5249 increase to `ModernC`.

`Frate`: Whenever `Frate` increases, `ModernC` is expected to increase 0.1894 times the amount.

`Pop`: 10% increase in `Pop` implies a 1.4721 * log(1.1) = 0.1403 increase to `ModernC`.

`Fertility`: Whenever `Fertility` increases, `ModernC` is expected to decrease 9.6759 times the amount.

`Purban`: Whenever `Purban` increases, `ModernC` is expected to decrease 0.0708 times the amount.

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

Our model: `ModernC` = -1.3773 + 4.9930 (`Change` + 1.1001) + 5.5073 log(`PPgdp`) + 0.1894 `Frate` + 1.4721 log(`Pop`) - 9.6759 `Fertility` - 0.0708 `Purban` + ϵ

In our study, we delete 85 observations with missing values, and our analysis are based on the remaining 125 UN member countries or regions.

When we talk about "the percent of unmarried women using a modern method of contraception" (cited from the R documentation for `UN3`), we find that many factors are correlated with it. Specificallly, greater annual population growth rate, higher GDP per capita, higher percent of females over 15 econnomically active and larger populations would all indicate more use of modern contraceptions among unmarried women, whereas increasing fertility or urban population percentage both imply a decline in the use of modern contraception. Morever, the population and GDP should be measured in log scale, since the their propotions between different countries are always exponential.

There are no significant outliers in the model, but we notice that the model is less accurate for China and Indea because of their huge population. Also, the model might work relatively poorly for countries like Thailand, Poland and Azerbaijan, for which the reasons are to be examined by sociologists.

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.*

Proof: For the added variable scatter plot, we are actually applying a simple linear regression between two regression residuals: $\hat{\mathbf{e}}_{(1)} = (\mathbf{I}_n - \mathbf{H})\mathbf{Y}$ and $\hat{\mathbf{e}}_{1|others} = (\mathbf{I}_n - \mathbf{H})\mathbf{X}_1$, where we suppose we are trying to add the variable x_1. Note that the hat matrices in the two expressions are the same, because both regressions are based on all predictors except x_1. The new regression model is

$$\hat{\mathbf{e}}_{(1)} = \beta_0 \mathbf{1}_n + \beta_1 \hat{\mathbf{e}}_{1|others} + \epsilon$$

A closed-form solution would be

$$\hat{\beta}_1 = (\hat{\mathbf{e}}_{1|others}^T \hat{\mathbf{e}}_{1|others})^{-1} \hat{\mathbf{e}}_{1|others}^T \hat{\mathbf{e}}_{(1)} = (\mathbf{X}_1^T(\mathbf{I}_n - \mathbf{H})\mathbf{X}_1)^{-1}\mathbf{X}_1^T(\mathbf{I}_n - \mathbf{H})\mathbf{Y},$$

in which we used $= (\mathbf{I}_n - \mathbf{H})^T(\mathbf{I}_n - \mathbf{H}) = \mathbf{I}_n - \mathbf{H}$. Therefore, we have

$$\hat{\beta}_0 \mathbf{1}_n = (\mathbf{X}_1^T(\mathbf{I}_n - \mathbf{H})\mathbf{X}_1)^{-1}\mathbf{X}_1^T(\mathbf{I}_n - \mathbf{H})\mathbf{Y}(\mathbf{I}_n - \mathbf{H})\mathbf{X}_1 - (\mathbf{I}_n - \mathbf{H})\mathbf{Y}$$

Multiplying (left) each side by $\mathbf{X}_1^T$, we obtain

$$\begin{aligned}
\hat{\beta}_0 \mathbf{X}_1^T \mathbf{1}_n &= \mathbf{X}_1^T \left(\mathbf{X}_1^T(\mathbf{I}_n - \mathbf{H})\mathbf{X}_1\right)^{-1} \left(\mathbf{X}_1^T(\mathbf{I}_n - \mathbf{H})\mathbf{Y}\right)(\mathbf{I}_n - \mathbf{H})\mathbf{X}_1 - \mathbf{X}_1^T(\mathbf{I}_n - \mathbf{H})\mathbf{Y} \\
&= \left(\mathbf{X}_1^T(\mathbf{I}_n - \mathbf{H})\mathbf{X}_1\right)^{-1} \left(\mathbf{X}_1^T(\mathbf{I}_n - \mathbf{H})\mathbf{Y}\right)\mathbf{X}_1^T(\mathbf{I}_n - \mathbf{H})\mathbf{X}_1 - \mathbf{X}_1^T(\mathbf{I}_n - \mathbf{H})\mathbf{Y} \\
&= \mathbf{X}_1^T(\mathbf{I}_n - \mathbf{H})\mathbf{Y} - \mathbf{X}_1^T(\mathbf{I}_n - \mathbf{H})\mathbf{Y} \\
&= 0,
\end{aligned}$$

in which we note that $\mathbf{X}_1^T(\mathbf{I}_n - \mathbf{H})\mathbf{X}_1$ and $\mathbf{X}_1^T(\mathbf{I}_n - \mathbf{H})\mathbf{Y}$ are scalars. Consequently, since $\mathbf{X}_1^T \mathbf{1}_n = \sum_{i=1}^n X_{i,1} \neq 0$, we could state that $\hat{\beta}_0 = 0$.

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.

```
lm_x = lm(Fertility ~ Change_trans + PPgdp_log + Frate + Pop_log + Purban, data = UN3_trans)
lm_y = lm(ModernC ~ Change_trans + PPgdp_log + Frate + Pop_log + Purban, data = UN3_trans)
av_plot_data = data.frame(cbind(residuals(lm_x), residuals(lm_y)))
colnames(av_plot_data) = c("e_x", "e_y")

lm_avplot = lm(e_y ~ e_x, data = av_plot_data)
summary(lm_avplot)
```

```
##
## Call:
## lm(formula = e_y ~ e_x, data = av_plot_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) -4.914e-16  1.178e+00   0.000        1
## e_x         -9.676e+00  1.729e+00  -5.595 1.36e-07 ***
## ---
## 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:  0.2029, Adjusted R-squared:  0.1964
## F-statistic: 31.31 on 1 and 123 DF,  p-value: 1.36e-07
```

We manually constrcted an avplot for `Fertility`. The estimated slope is -9.676, which is the same as its coefficient.