HW2 STA521 Fall18

Prabhakar Nanduri // Netid: pnn2 // github username: nanduriprabhakar

September 23, 2018

Exploratory Data Analysis

```
## Loading required package: car
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

Summary of the Data

```
colnames(UN3)
```

```
## [1] "ModernC"   "Change"    "PPgdp"     "Frate"     "Pop"       "Fertility"
## [7] "Purban"
```

```
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
```

```
glimpse(UN3)
```

```
## Observations: 210
## Variables: 7
## $ ModernC   <int> NA, NA, 49, NA, NA, NA, 51, NA, 22, NA, 72, 43, 12, ...
## $ Change    <dbl> 3.88, 0.68, 1.67, 2.37, 2.59, 3.20, 0.53, 1.17, -0.4...
## $ PPgdp     <int> 98, 1317, 1784, NA, 14234, 739, 8461, 7163, 687, NA,...
## $ Frate     <int> NA, NA, 7, 42, NA, NA, 63, 44, 51, 53, 55, 49, 43, 6...
## $ Pop       <dbl> 23897.000, 3167.000, 31800.000, 57.000, 64.000, 1362...
## $ Fertility <dbl> 6.80, 2.28, 2.80, NA, NA, 7.20, NA, 2.44, 1.15, NA, ...
```

```
## $ Purban    <int> 22, 43, 58, 53, 92, 35, 37, 88, 67, 51, 91, 67, 52, ...
```

1)From the summary analysis of the dataset, all the variables are quantitative and all the variables expect for Purban [ie. ModernC, Change, PPgdp, Frate, Pop, Fertility] have missing values.

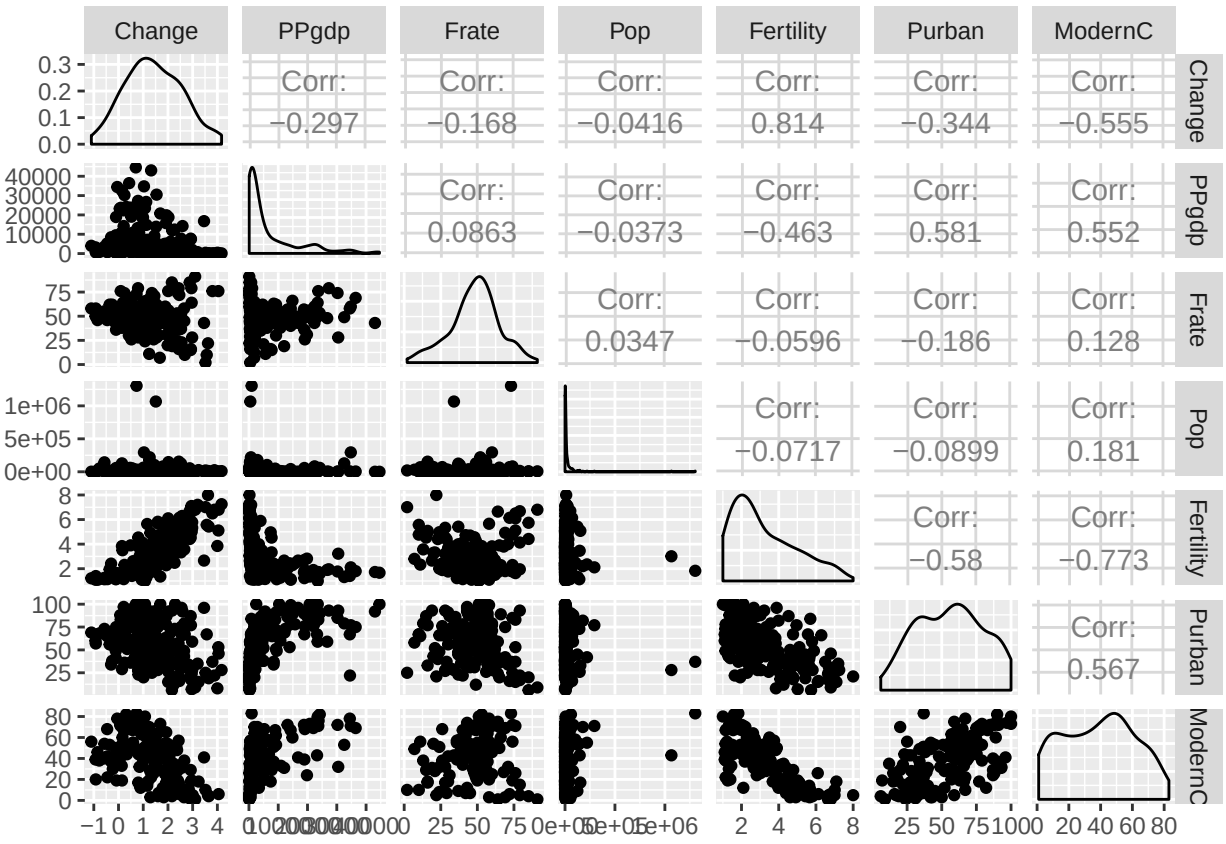
2)While calculating the mean and standard deviation of the quantitative predictors, we remove the NA values so that it does not affect measures of central tendency calculations.

```
ModernC_mean = mean(UN3$ModernC, na.rm = TRUE)
ModernC_std = sd(UN3$ModernC, na.rm = TRUE)
Change_mean = mean(UN3$Change, na.rm = TRUE)
Change_std = sd(UN3$Change, na.rm = TRUE)
PPgdp_mean = mean(UN3$PPgdp, na.rm = TRUE)
PPgdp_std = sd(UN3$PPgdp, na.rm = TRUE)
Frate_mean = mean(UN3$Frate, na.rm = TRUE)
Frate_std = sd(UN3$Frate, na.rm = TRUE)
Pop_mean = mean(UN3$Pop, na.rm = TRUE)
Pop_std = sd(UN3$Pop, na.rm = TRUE)
Fertility_mean = mean(UN3$Fertility, na.rm = TRUE)
Fertility_std = sd(UN3$Fertility, na.rm = TRUE)
Purban_mean = mean(UN3$Purban, na.rm = TRUE)
Purban_std = sd(UN3$Purban, na.rm = TRUE)

df=data.frame(matrix(nrow=0,ncol=3))
ModernC_df=c("ModernC",ModernC_mean,ModernC_std)
Change_df=c("Change",Change_mean,Change_std)
PPgdp_df=c("PPgdp",PPgdp_mean,PPgdp_std)
Frate_df=c("Frate",Frate_mean,Frate_std)
Pop_df=c("Pop",Pop_mean,Pop_std)
Fertility_df=c("Fertility",Fertility_mean,Fertility_std)
Purban_df=c("Purban",Purban_mean,Purban_std)

df=rbind(df,ModernC_df,Change_df,PPgdp_df,Frate_df,Pop_df,Fertility_df,Purban_df)

x=c("variable name","mean", "standard_deviation")
colnames(df)=x
kable(df)
```

variable name	mean	standard_deviation
ModernC	38.7171052631579	22.6366103759673
Change	1.41837320574163	1.13313267030361
PPgdp	6527.38805970149	9325.18855244529
Frate	48.3053892215569	16.5324480416909
Pop	30281.8714278846	120676.694478229
Fertility	3.214	1.70691793716661
Purban	56.2	24.1097570036514

3)The pair-plots here all remove missing data individually for each predictor.

```
ggpairs(UN3, columns=c(2,3,4,5,6,7,1))
```



From the pair-plots, we see that there is a high correlation between 'Fertility' and 'Change' (0.814), which can be intuitively explained (since high fertility leads to high population growth rate). There seem to be outlier points in 'Pop' and 'Purban'. The general distribution and scales of the predictors 'PPgdp' and 'Pop' seem to point towards taking their transformations. Plotting 'ModernC' as a dependent variable across the other predictors, the above mentioned transformations seem more pragmatic. The correlation values of 'ModernC' with 'Pop' and 'Frate' are relatively low and thus might not give much information about 'ModernC'.

Model Fitting

4)

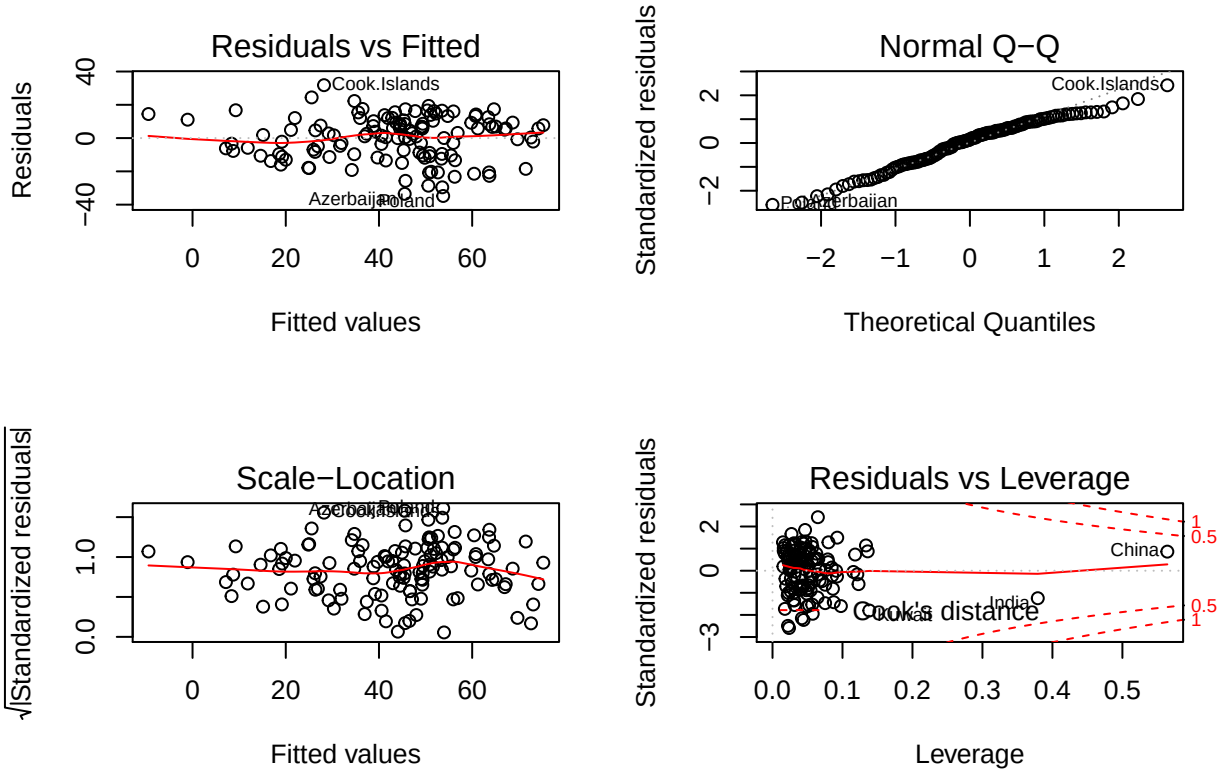
```
modernc.lm=lm(ModernC ~ ., data = UN3)
summary(modernc.lm)
```

```
##
## 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
```

From the summary of the linear model displayed above; we understand that 85 observations were deleted due to missing values. Thus of the 210 data points only (210-85 = 125) 125 data points were used to fit the linear model.

```
par(mfrow=c(2,2))
plot(modernc.lm, ask=FALSE)
```



For the linear model, from the residual plots we (more or less) see that the residuals are having a mean of zero with some constant variance. While we do not observe a perfect straight line, the overall trend is more or less centered around 0.

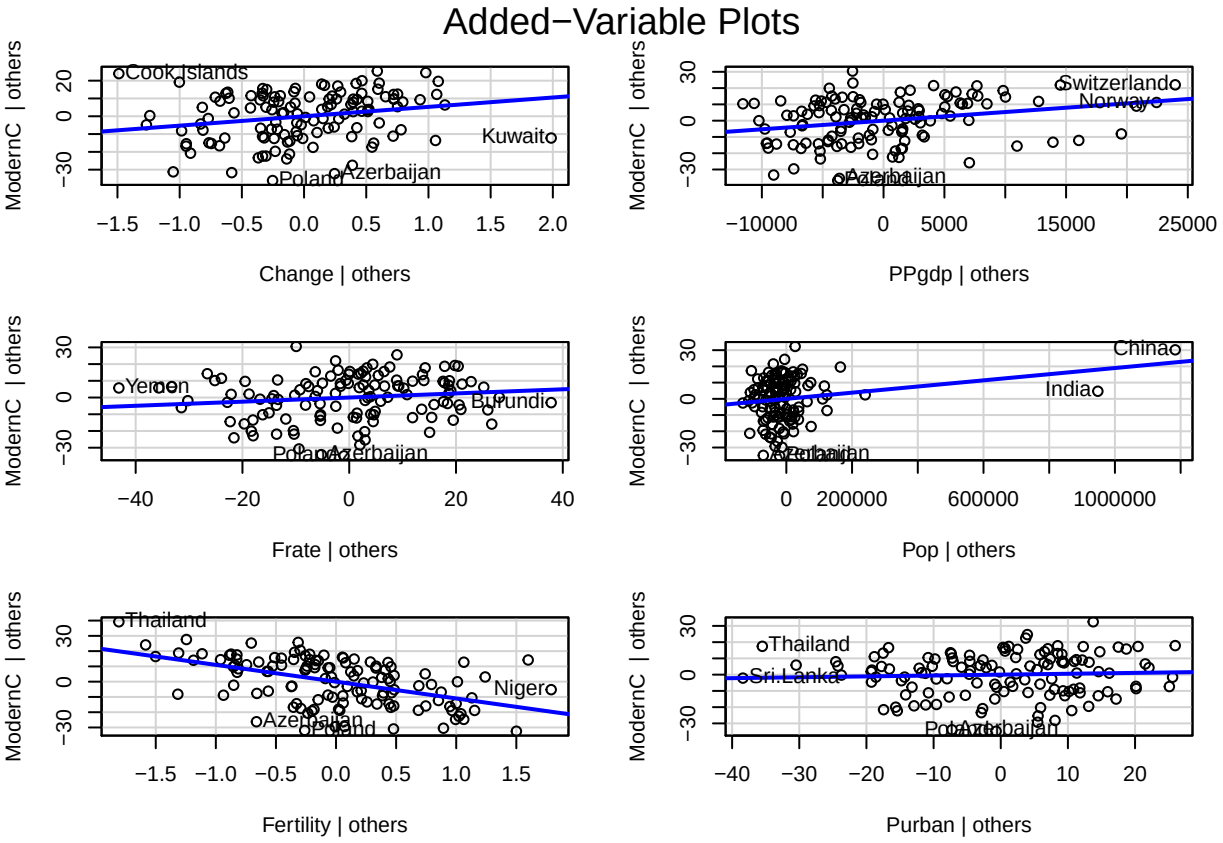
From the Q-Q plot, the assumption for ModernC being normally distributed doesnot completely hold true, since there seem to be a longer left tail(or a heavier right tail).

There do not seem to be influential outliers since none of the points have a cook's distance >[0.5 or 1], but the points of 'India' and 'China' do seem to be outliers nonetheless.

Added Variable Plots

5)

```
car::avPlots(lm(ModernC ~ ., data = UN3))
```



From the added variables plots above, it is evident that PPgdp and Pop might be needing some kind of transformations. [This can be deciphered from the scale of the axis and the slope steepness of the ModernC vs the 'particular variable' regression line]. While the plots for the Change and the Fertility variables do show steepness, the scale of the variables (in their units) do not intent for any transformations.

Transformations

6)

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

```
##  MLE of lambda Score Statistic (z) Pr(>|z|)
##        0.63309             -0.5543   0.5794
##
## iterations =  3
```

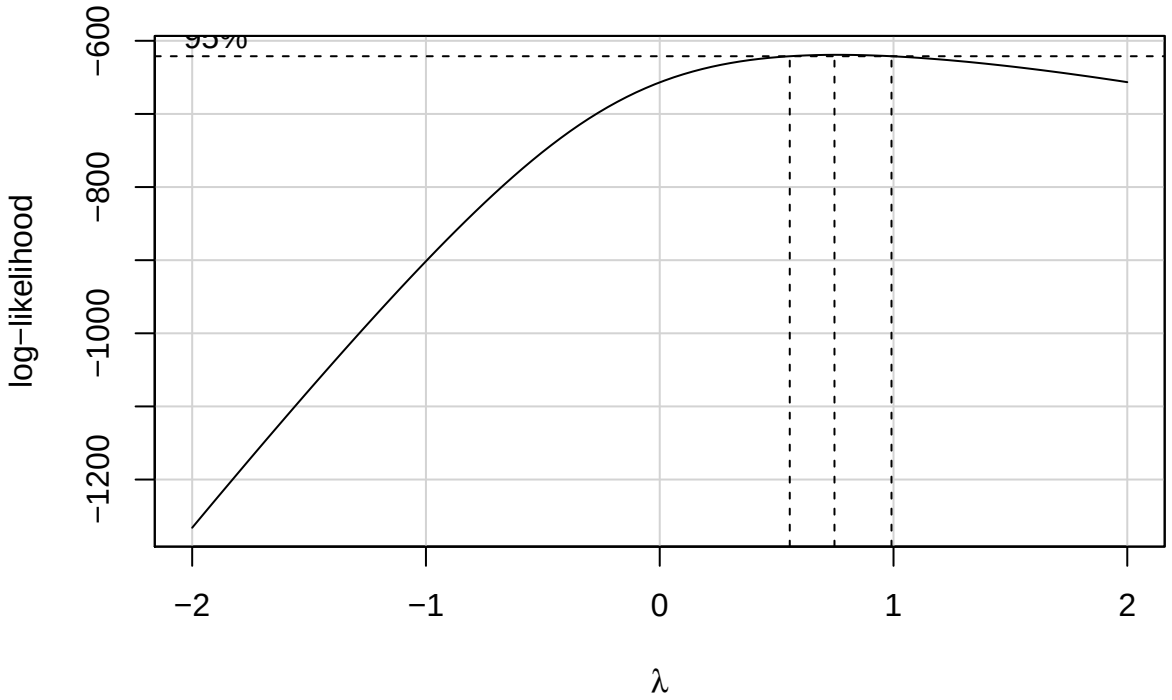
```
car::boxTidwell(ModernC ~ Pop+PPgdp ,  ~ Change +  Frate + Fertility + Purban  , data = UN3 )
```

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

Running the boxTidwell plots for two cases (first with transformation only in the population variable and second with transformations for both the population and PPgdp variables), the second case seems to be more pragmatic in the approach (since in both the cases we would be taking the square root of the Pop variable) to variable transformation. {Taking the log of PPgdp variable can be understood by looking at the scale of the variable's units.}

7)

```
# test1 = lm(ModernC ~ sqrt(Pop) + PPgdp + Change + Frate + Fertility + Purban, data = UN3)
test2 = lm(ModernC ~ sqrt(Pop) + log(PPgdp) + Change + Frate + Fertility + Purban, data = UN3)
car::boxCox(test2)
```



Plotting the boxCox plot for the response variable, with a reasonable confidence (of the 95% interval) it can be said that the response variable does not need any transformation as such. The 95% confidence interval leans to lambda = 1, which is easier to interpret rather than taking the absolute value of lambda. Taking the trade-off for interpretability over model accuracy, no transformation for the response variable would be needed.

Model Fitting and Additional Transformations

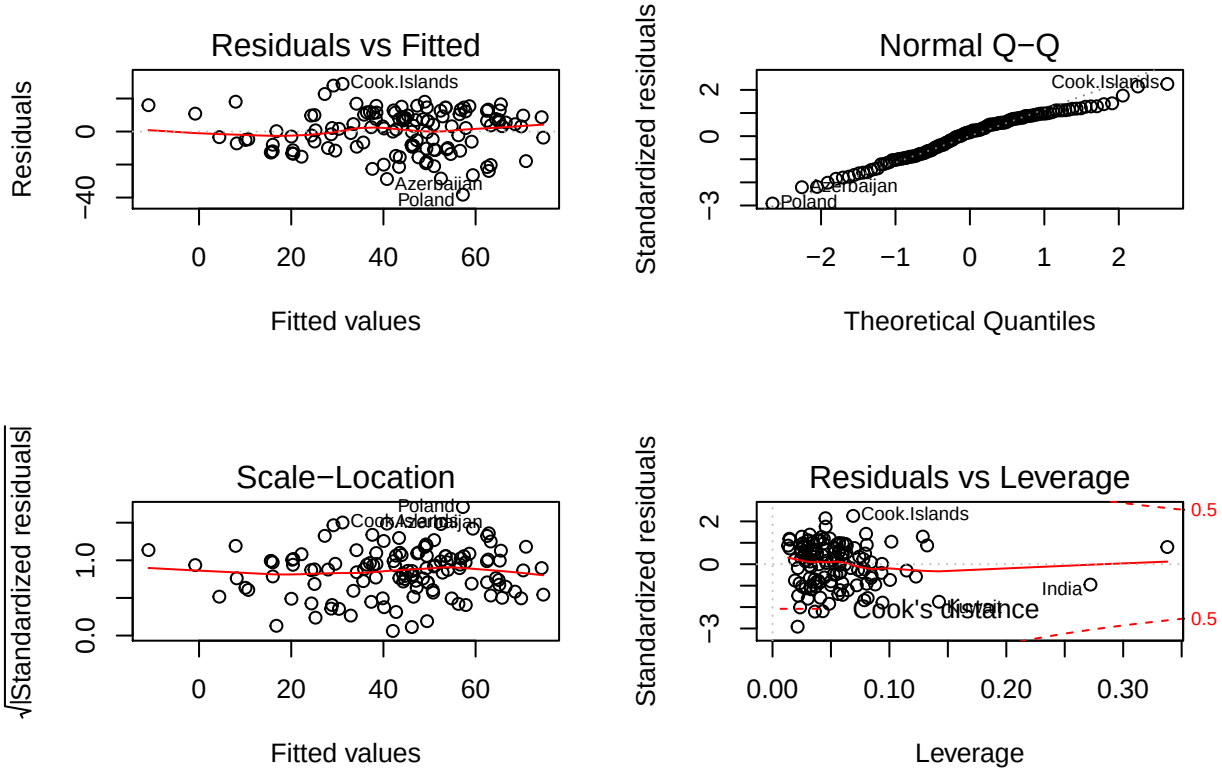
8)

```
test2.lm = lm(ModernC ~ sqrt(Pop) + log(PPgdp) + Change + Frate + Fertility + Purban, data = UN3)
summary(test2.lm)
```

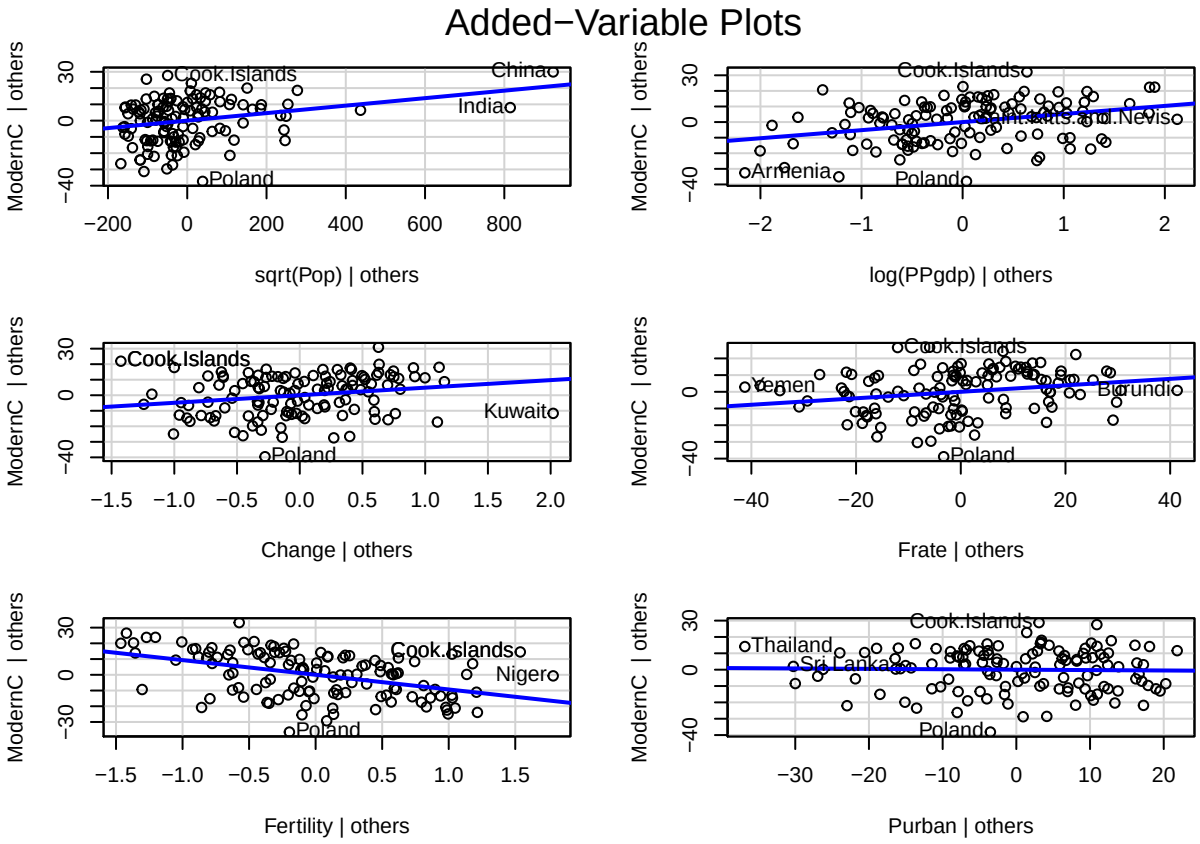
```
##
## Call:
## lm(formula = ModernC ~ sqrt(Pop) + log(PPgdp) + Change + Frate +
##     Fertility + Purban, data = UN3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -38.239  -9.995   2.133   9.961  28.861
```

```
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.968065  12.247992   1.059  0.29186
## sqrt(Pop)    0.023017   0.007685   2.995  0.00335 **
## log(PPgdp)   5.182415   1.359076   3.813  0.00022 ***
## Change       4.869147   2.042766   2.384  0.01874 *
## Frate        0.194010   0.076053   2.551  0.01202 *
## Fertility   -9.327572   1.749773  -5.331 4.77e-07 ***
## Purban      -0.025072   0.096557  -0.260  0.79558
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.26 on 118 degrees of freedom
##   (85 observations deleted due to missingness)
## Multiple R-squared:  0.6362, Adjusted R-squared:  0.6177
## F-statistic:  34.4 on 6 and 118 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(test2.lm, ask=FALSE)
```




```
car::avPlots(test2.lm)
```



With the new linear model fitted against the two transformed variables; the residual plots now are more aligned towards the general assumptions of a linear regression model. The residuals mean is now more centered towards zero and the variance in the residuals is relatively more equally distributed. The Normal distributions Q-Q plots still show some deviation from the normal distribution of the residuals { being heavier on the right tail}, but much better off than the un-transformed variable residuals. From the AVplots there still seems to be some transformation possibility for sqrt(Pop) but additional transformations might lead to a loss of interpretability of the model and thus the additional transformations are being avoided.

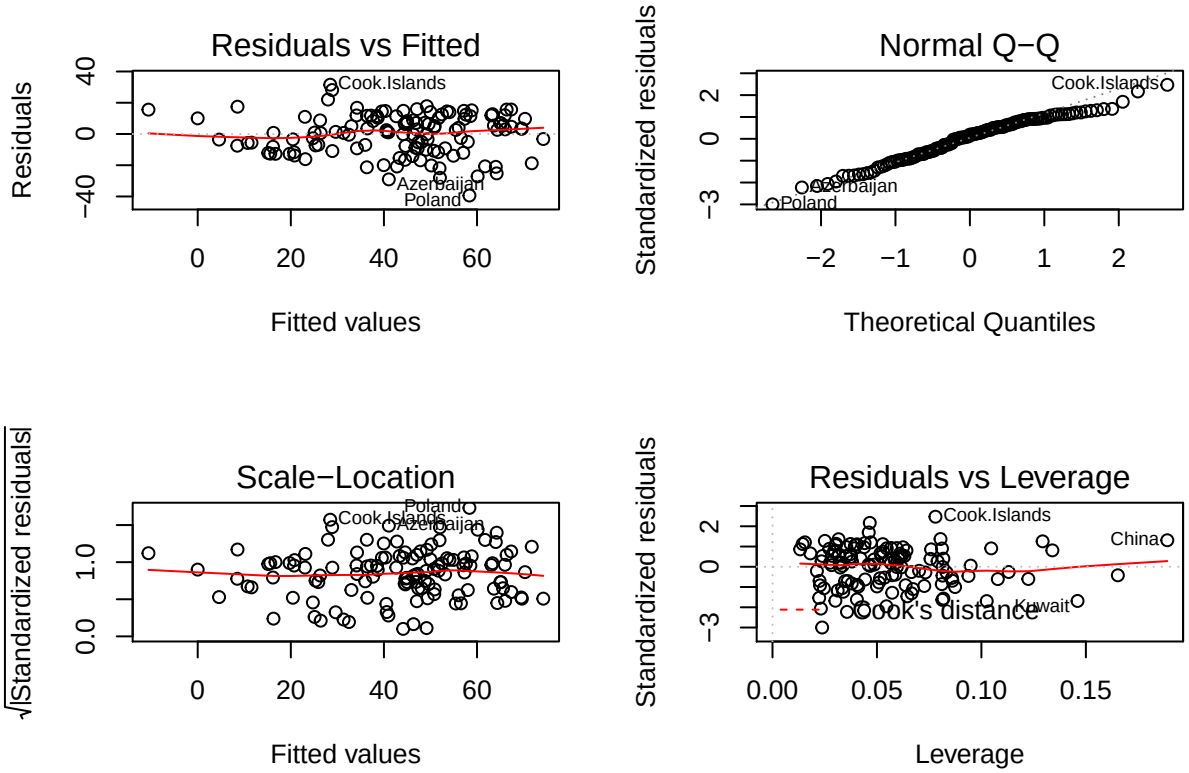
9)

```
test_.lm = lm(ModernC ~ sqrt(sqrt(Pop)) + log(PPgdp) + Change + Frate + Fertility + Purban, data = UN3)
summary(test_.lm)
```

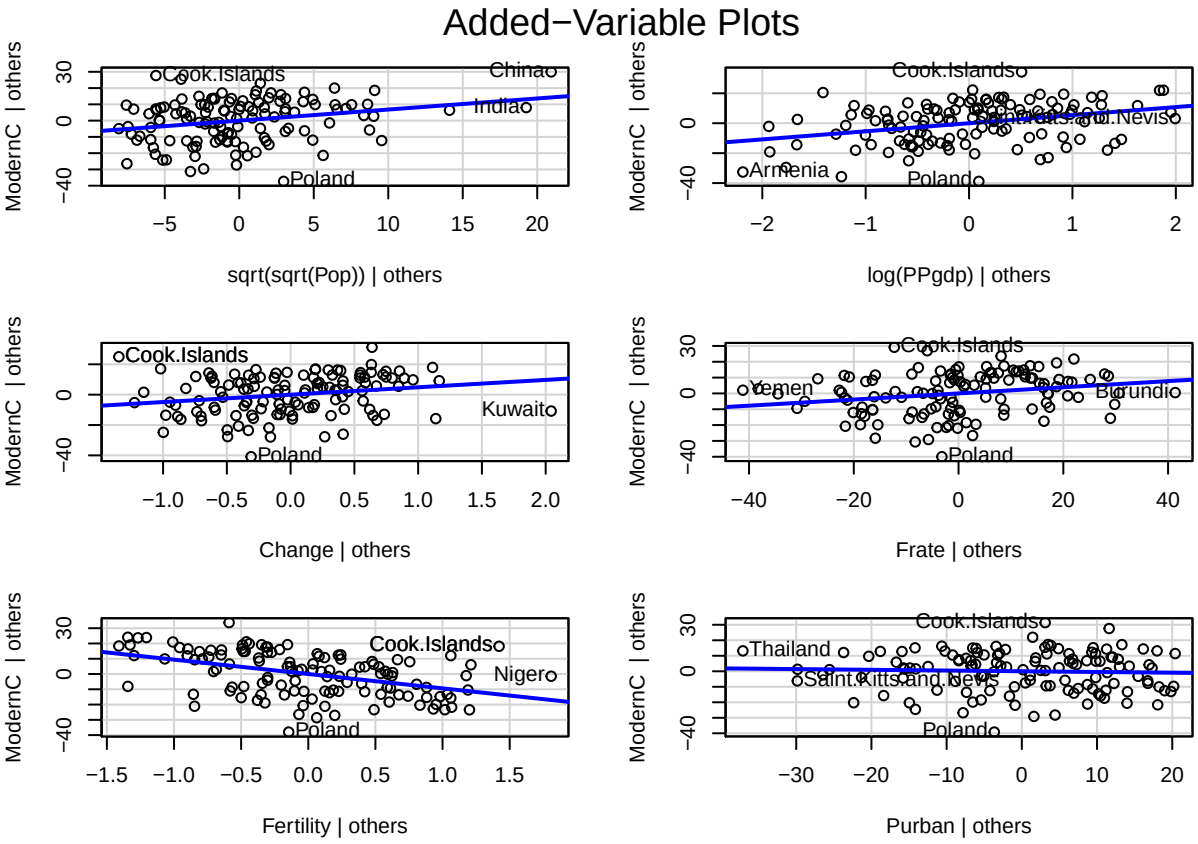
```
##
## Call:
## lm(formula = ModernC ~ sqrt(sqrt(Pop)) + log(PPgdp) + Change +
##     Frate + Fertility + Purban, data = UN3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -39.373  -9.209   2.468  10.238  31.530
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       8.72968   12.82655   0.681 0.497461
## sqrt(sqrt(Pop))   0.68095    0.23554   2.891 0.004573 **
## log(PPgdp)        5.39797    1.37207   3.934 0.000141 ***
## Change            4.86766    2.04957   2.375 0.019163 *
```

```
## Frate             0.19393    0.07624   2.544 0.012257 *
## Fertility        -9.40594    1.75092  -5.372 3.97e-07 ***
## Purban           -0.04680    0.09630  -0.486 0.627919
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.29 on 118 degrees of freedom
##   (85 observations deleted due to missingness)
## Multiple R-squared:  0.6345, Adjusted R-squared:  0.6159
## F-statistic: 34.14 on 6 and 118 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(test_.lm, ask=FALSE)
```




```
car::avPlots(test_.lm)
```



An attempt to take an additional sqrt of the (sqrt(Pop)) variable was done and the model was tried to fit. With interpretability of the model being a questionable issue, the residual plots of the fit model showed much deviation from a normal-distribution behaviour(along with a minor loss in the R squared value). This transformation is then ruled out as a worse model than the one deduced in question 8

Thus, after the various transformations of the predictors, the most pragmatic and interpretive friendly model is test2.lm = lm(ModernC ~ sqrt(Pop) + log(PPgdp) + Change + Frate + Fertility + Purban, data = UN3)

Removal of Outliers

10)From the previous model analysis, the points pertaining to countries “India” and "China seem to be outliers. {They are not necessarily influential since their cook’s distance is well below the threshold}

Re-analyzing the model by taking out the two rows pertaining to China (row 39) and India (row 86)

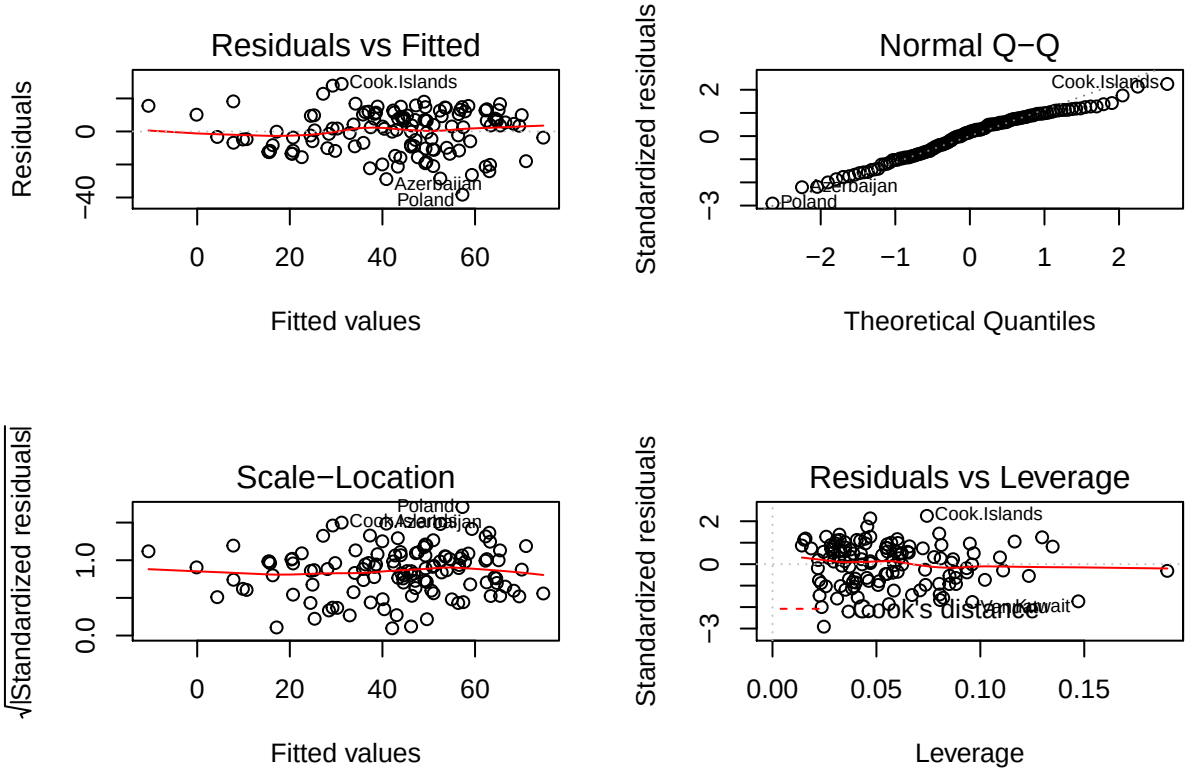
```
UN3_outliers_removed = UN3[-c(39,86),]
```

```
test3.lm = lm(ModernC ~ sqrt(Pop) + log(PPgdp) + Change + Frate + Fertility + Purban, data = UN3_outlie
summary(test3.lm)
```

```
##
## Call:
## lm(formula = ModernC ~ sqrt(Pop) + log(PPgdp) + Change + Frate +
##     Fertility + Purban, data = UN3_outliers_removed)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -38.278 -10.005   2.533  10.028  28.808
```

```
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.74579   12.34494   1.113 0.267806
## sqrt(Pop)    0.02365    0.01151   2.055 0.042154 *
## log(PPgdp)   5.18311    1.36567   3.795 0.000236 ***
## Change       4.83800    2.05078   2.359 0.019991 *
## Frate        0.18192    0.07724   2.355 0.020188 *
## Fertility   -9.31943    1.75699  -5.304 5.49e-07 ***
## Purban      -0.02949    0.09761  -0.302 0.763087
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.31 on 116 degrees of freedom
##   (85 observations deleted due to missingness)
## Multiple R-squared:  0.6293, Adjusted R-squared:  0.6102
## F-statistic: 32.83 on 6 and 116 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(test3.lm, ask=FALSE)
```



Refitting the model after removing the outliers, the residual plots now are more evenly distributed along the mean (0) value. While there is a minor decrease in the 'R-square' value (decrease in model's ability to explain variance), the trade-off towards developing a model that is closer towards the assumptions is acceptable. The residuals vs. Leverage plots also show even distribution around the mean (0) line with no evident outliers/influential points.

Summary of Results

11)

```
kable(summary(test3.lm)$coeff, digits = c(3,3,3,3))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	13.746	12.345	1.113	0.268
sqrt(Pop)	0.024	0.012	2.055	0.042
log(PPgdp)	5.183	1.366	3.795	0.000
Change	4.838	2.051	2.359	0.020
Frate	0.182	0.077	2.355	0.020
Fertility	-9.319	1.757	-5.304	0.000
Purban	-0.029	0.098	-0.302	0.763

```
kable(confint(test3.lm), digits = c(3,3))
```

	2.5 %	97.5 %
(Intercept)	-10.705	38.197
sqrt(Pop)	0.001	0.046
log(PPgdp)	2.478	7.888
Change	0.776	8.900
Frate	0.029	0.335
Fertility	-12.799	-5.839
Purban	-0.223	0.164

From the regression model fit (post transformations and outlier removals), the interpretation of the coefficients can be summarized as below:-

For the predictor variable (y) in its originial units (% of unmarried women using a modern method of contraception), -> An increase of population by 1000 members increase the % of unmarried women using modern methods of contraception{from now on referred to as just %} by (0.024)*1 = 0.024 % (approx. considering the base population levels are much, more higher than 1000 members) -> An increase of per capita GDP (2001, USD values) by x units, increase the % by 5.183* log(base-value + x / base-value) -> An increase in annual population growth rate by 1% increases the % by 4.838% -> An increase in females overage 15 that are economically active by 1% increases the % by 0.182%. -> An increase in expected number of liver births per female by 1 unit decrease the % by 9.319%. -> An increase in urban population by 1% is expected to decrease the % by 0.029%.

Model Summary Text

12)From the data collect by the UN containing the national health, welfare and education statistics of 210 places, only 125 data points (data from only 125 places) have complete entries that could be used to build a linear model. The model tries to explain the relationship between The % of unmarried women using modern methods of contraception to all the other macro economic factors viz. Annual population growth rate, Per capita GDP, place population, fertility of females and % of urban population. Two of the 125 data points(pertaining to India and China) have been removed from the model building exercise since they have extremely high population numbers which could result in wrong analysis. The model developed assumes that all the macro economic factors (with certain transformations for the factors of population and ppgdp) have a linear relationship with the predictor variable Y (% of women using modern methods of contraception). With this assumption, the model developed indicates that increase in population growth, per capita GDP, annual

population growth rate, fertility rate; and decrease in Feritlity and urban population % increases the % of unmarries women using modern methods of contraception. {For exact changes and value estimates please see the interpretations of coeffients above}.

The limitations of the model developed pertains to lack of causality. The model can not explain whether one factor(predictor variable) leads to a change in the ModernC variable or vice-versa. The model only explains correlations observed and does not explain any cause and effect relationships.

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.*

A) We know the premise that

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1.X$$

Thus applying for the errors in Y and X, we get

$$e_y = \hat{\beta}_0 + \hat{\beta}_1.e_x$$

we know,

$$e_y = (I-H)Y$$

and

$$e_x = (I-H)X$$

Thus, we get,

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

we know,

$$\hat{\beta}_1 = (X'.X)^{-1}X'.Y$$

thus,

$$(I-H)Y = \hat{\beta}_0 + (X'X)^{-1}X'Y(I-H)X$$

we know,

$$X = (I-H).X_j$$

simplifying further and rearranging the terms,

$$(I-H)Y = \hat{\beta}_0 + (((I-H)X_j)'(I-H)X_j)^{-1}((I-H)X_j)'Y(I-H)(I-H)X_j$$
$$(I-H)Y = \hat{\beta}_0 + (X_j'(I-H)'(I-H)X_j)^{-1}X_j'(I-H)'Y(I-H)(I-H)X_j$$
$$(I-H)Y = \hat{\beta}_0 + (X_j'(I-H)X_j)^{-1}X_j'(I-H)'Y(I-H)X_j$$

Note: We are using the property that $(I - H)$ is a diagonal and idempotent matrix.

Now, multiplying both sides of the equation with $(X_j)'$, we get,

$$X_j'(I-H)Y = X_j'\hat{\beta}_0 + X_j'(X_j'(I-H)X_j)^{-1}X_j'(I-H)'Y(I-H)X_j$$

$$X_j'(I-H)Y = X_j'\hat{\beta}_0 + X_j'(I-H)X_j(X_j'(I-H)X_j)^{-1}X_j'(I-H)Y$$

$$X_j'(I-H)Y = X_j'\hat{\beta}_0 + X_j'(I-H)Y$$

Thus,

$$X_j'\hat{\beta}_0 = 0$$

considering non-trivial cases, we conclude $\hat{\beta}_0 = 0$ thus, proving that the intercept value 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.

A) We will use the models developed in Q)10 to confirm on the slopes.

Let us take X_j to be Change.

```
UN3_outliers_removed = UN3_outliers_removed[complete.cases(UN3_outliers_removed),]
e_Y = residuals(lm(ModernC ~ sqrt(Pop) + log(PPgdp) + Frate + Fertility + Purban, data = UN3_outliers_r
e_X = residuals(lm(Change ~ sqrt(Pop) + log(PPgdp) + Frate + Fertility + Purban, data = UN3_outliers_re

linearmodel.lm = lm(e_Y ~ e_X)

coefficient_final_model = summary(test3.lm)$coefficients['Change',c('Estimate')]
coefficient_temp_model = summary(linearmodel.lm)$coefficients['e_X',c('Estimate')]


df_ = rbind(coefficient_final_model, coefficient_temp_model)

kable(df_)
```

coefficient_final_model	4.838
coefficient_temp_model	4.838

We see that the slope for the manually constructed AV plot (with the added variable being 'Change') is same as the slope of the final model generated in Q.10, Thus emperically proving the ask in the question.