HW2 STA521 Fall18

Eduardo Coronado - ec243 - ecoronado92

Due September 23, 2018 5pm

Backgound Reading

Readings: Chapters 3-4 in Weisberg Applied Linear Regression

Exploratory Data Analysis

0. Preliminary read in the data. After testing, modify the code chunk so that output, messages and warnings are suppressed.

1. Create a summary of the data. How many variables have missing data? Which are quantitative and which are qualtitative?
   From the summary data below 6 out of the 7 variables have missing data, with `ModernC` and `Frate` being the ones with most NAs present. Also from the metadata we know that only 125 observations out of 210 have complete data for all variables.

```
##     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
```

Additionally, **all variables are quantitative.**

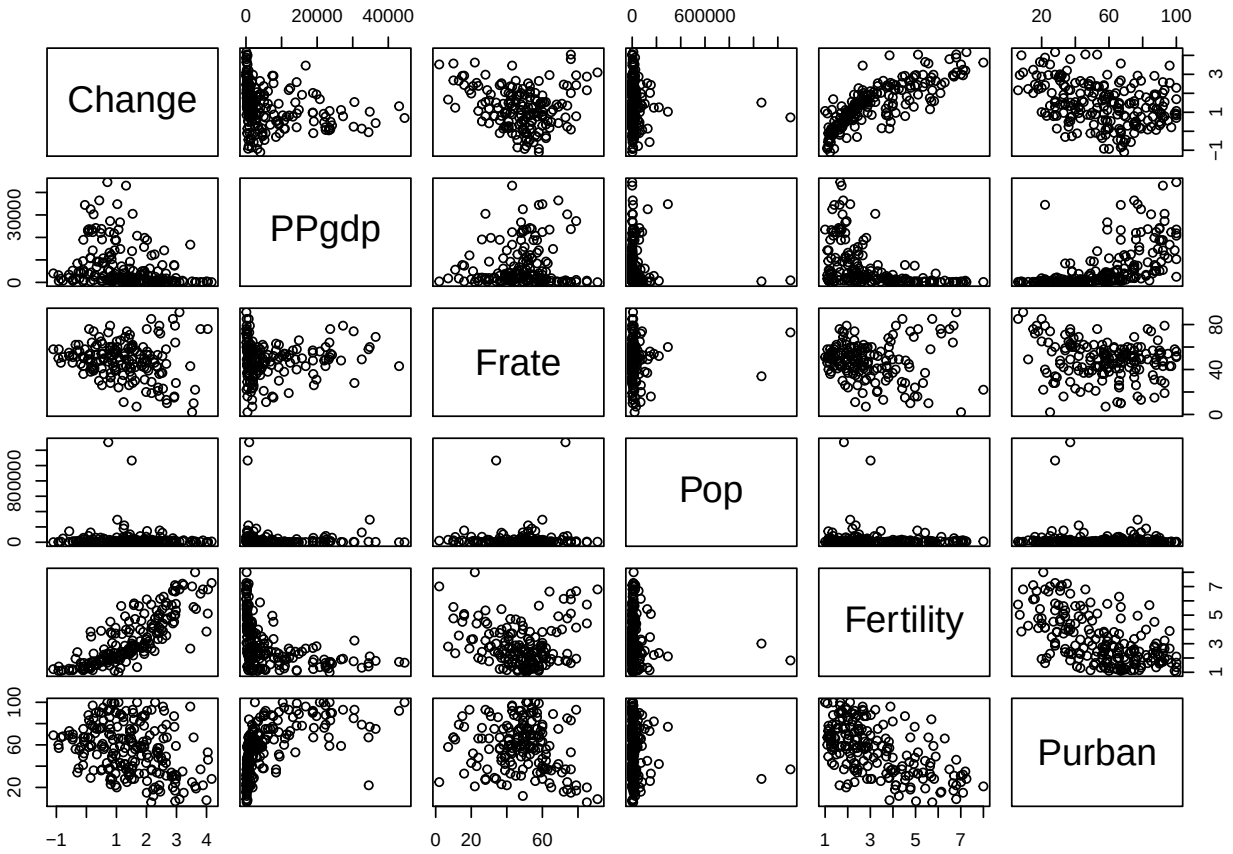
```
## 'data.frame':    210 obs. of  7 variables:
##  $ ModernC  : int  NA NA 49 NA NA NA 51 NA 22 NA ...
##  $ Change   : num  3.88 0.68 1.67 2.37 2.59 3.2 0.53 1.17 -0.45 2.02 ...
##  $ PPgdp    : int  98 1317 1784 NA 14234 739 8461 7163 687 NA ...
##  $ Frate    : int  NA NA 7 42 NA NA 63 44 51 53 ...
##  $ Pop      : num  23897 3167 31800 57 64 ...
##  $ Fertility: num  6.8 2.28 2.8 NA NA 7.2 NA 2.44 1.15 NA ...
##  $ Purban   : int  22 43 58 53 92 35 37 88 67 51 ...
```

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

	Mean	Standard Dev
ModernC	38.72	22.64
Change	1.42	1.13
PPgdp	6527.39	9325.19
Frate	48.31	16.53
Pop	30281.87	120676.69
Fertility	3.21	1.71
Purban	56.20	24.11

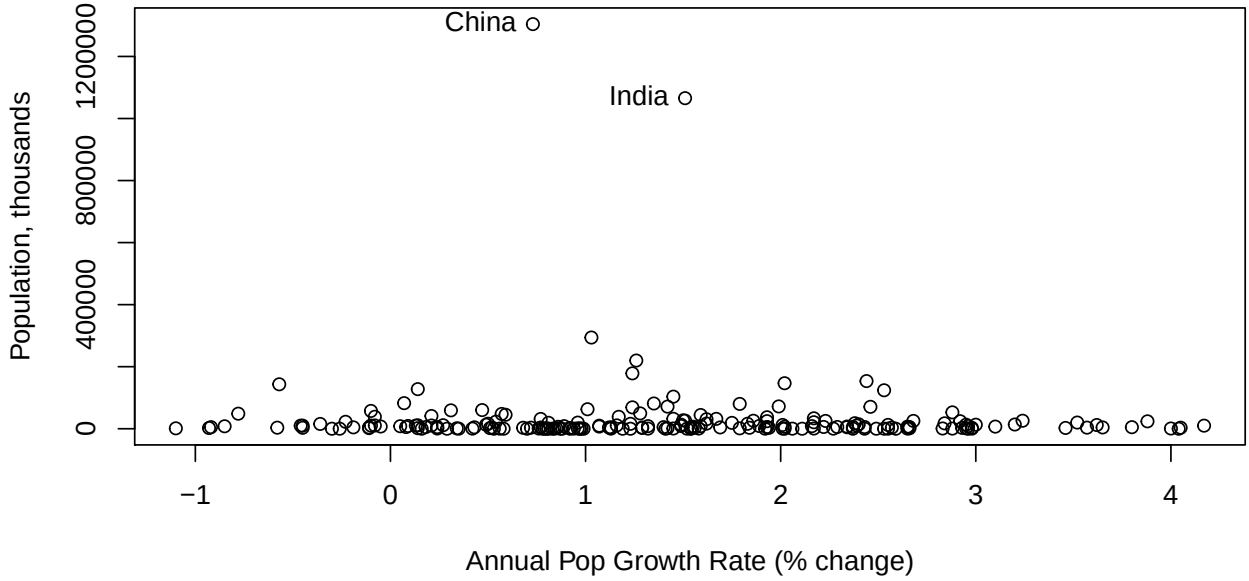
3. Investigate the predictors graphically, using scatterplots or other tools of your choice. Create some plots highlighting the relationships among the predictors. Comment on your findings regarding trying to predict `ModernC` from the other variables. Are there potential outliers, nonlinear relationships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

Fig 1. Pairwise Comparisons of Quanitative Predictor Variables from UN3 Dataset



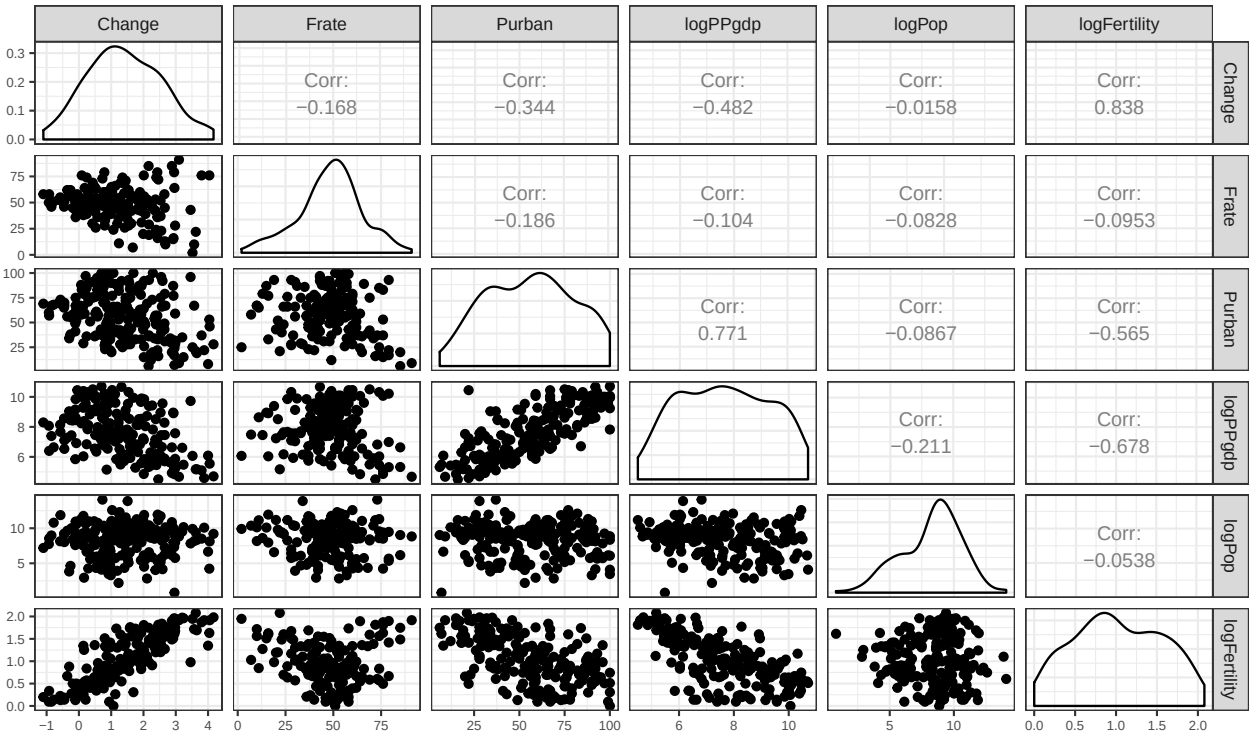
Using the `pairs` function we can do an initial assessment of the relationships among the predictor values. From **Fig 1** above it is easy to notice at first glance that many relationships among predictor variables seem non-linear. The relationships between the `Change`, `PUrban` and `Fertility` variables seem to be the ones that mostly resemble a somewhat linear relationship. The `Frate` variable does seems to have a non-multicolinear relationship with some of the variables. However, two plots standout from this plot - `PPgdp` and `Pop`. The `PPgdp` predictor's relationships seem to follow an increasing or decreasing exponential, while the `Pop` helps signal what seem to be two clear outliers (**Fig2**, below). From this plot we can see that China and India seem as outliers in terms of population vs other countries.

Fig 2. Annual Pop Growth Rate Change vs Population for 210 Countries



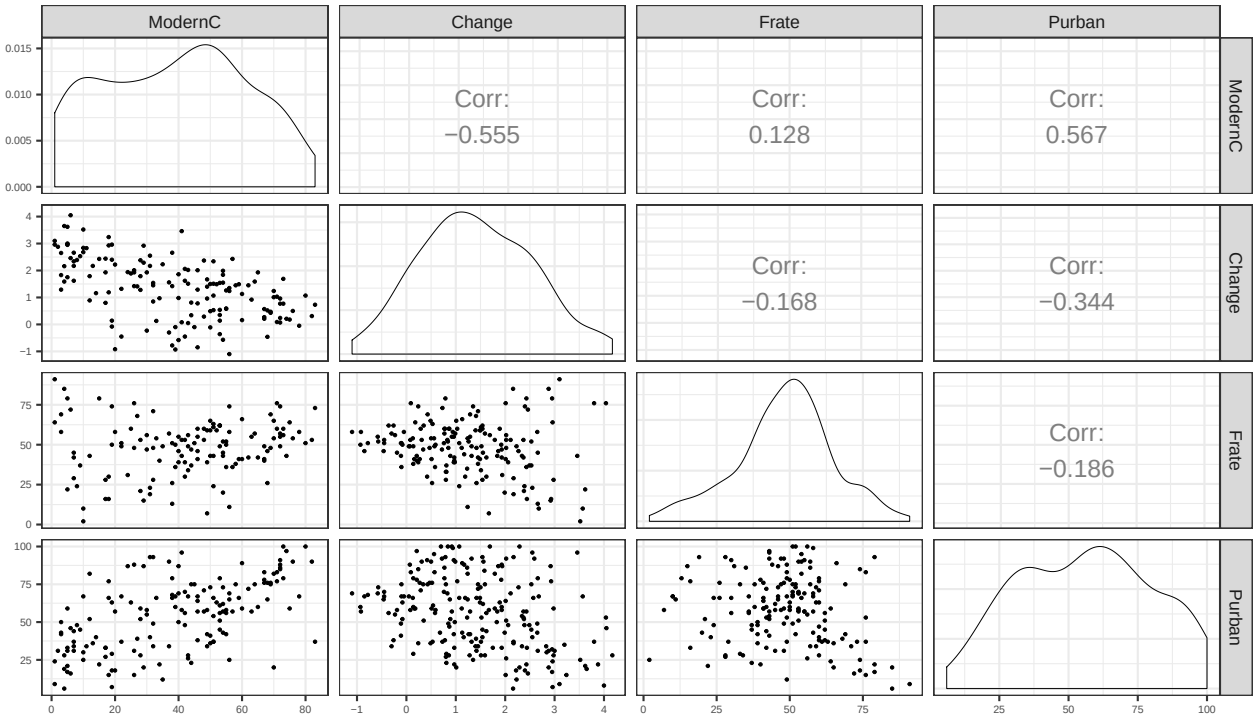
Neverthless, it is important to note that the scale of these two variables is several orders of magnitude higher than the predictor they are compared against, which brings to mind possible linear transformations for further evaluations. Thus, I explored whether a simple `log()` transformation would suffice to demonstrate a linear relationship among the non-linear relationships in **Fig 1** (i.e. `logFertility`, `logPPgdp`, `logPop`). Even though this was a crude first transformation on the date, in **Fig 3** below we can notice that transforming the data can help elucidate potential linear relationships among our predictors, as well as the presence or lack of multicolinearity.

Fig 3. Pairwise Comparisons of Transformed Predictor Variables from UN3 Dataset



Finally, from **Fig 4** below we can see that a linear combination of the predictor variables could be helpful to predict `ModernC` given these seem to follow a linear relationship. However, it is important to note that some predictors exhibit multicolinearily which means adding them to the linear model would be redundant as these would be contributing the same variance to the response variable. It is something that we should consider as we continue to build and assess the fit of the model.

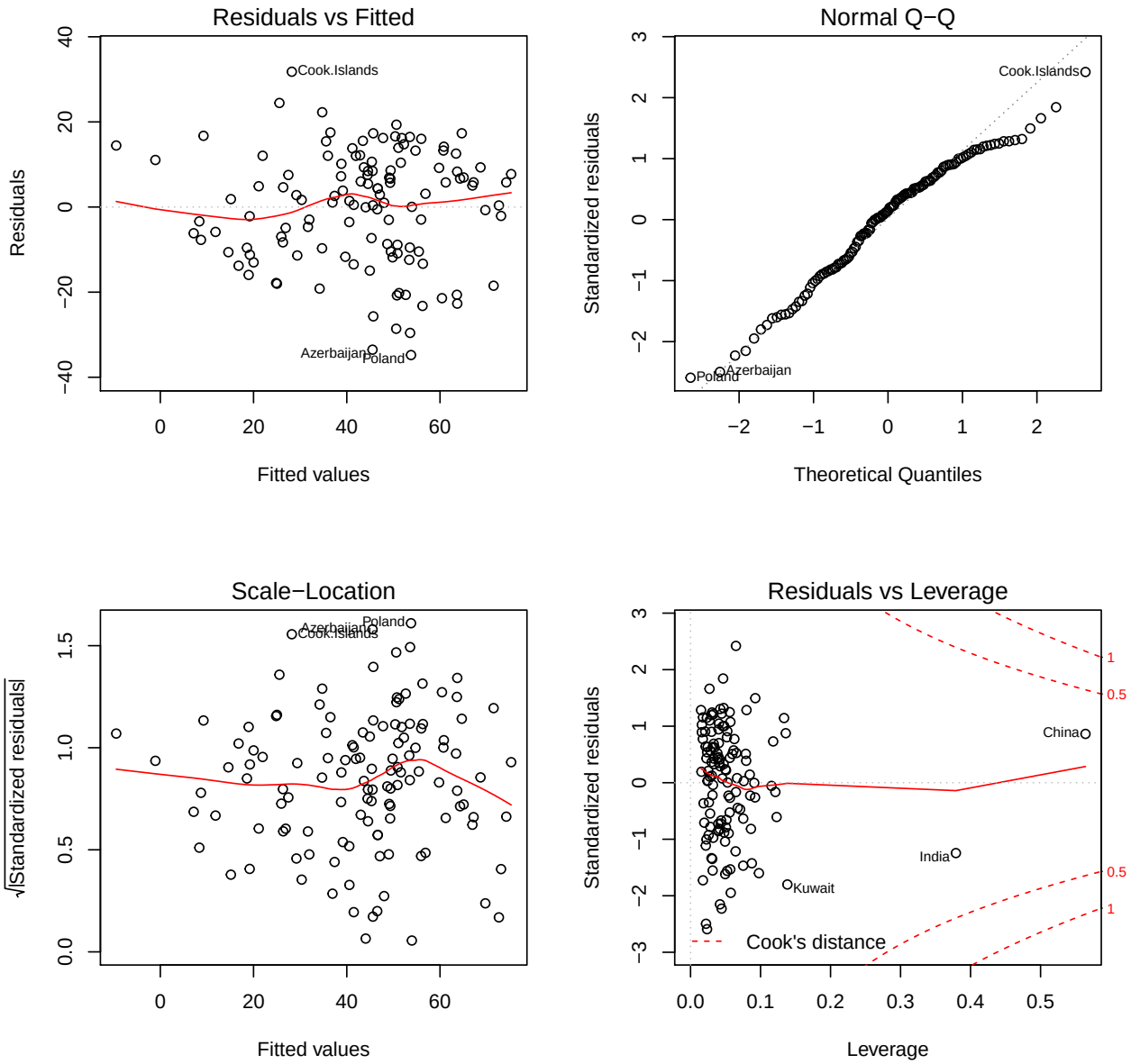
Fig 4. Pairwise Comparison of Six Predictor Variables and 'ModernC' Response Variable from the UN3 Dataset



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?

Fig 5. Diagnostic Residual Plots for a Linear Model Fit of 'ModernC'



From from the initial comparisons in **Fig4** we know that the predictor variables have a linear relationship which allows us to interpret these multiple regression diagnostic plots as those of a simple regression model. ***Note**: the above linear model and diagnostic plots were done using the original, non-transformed data*

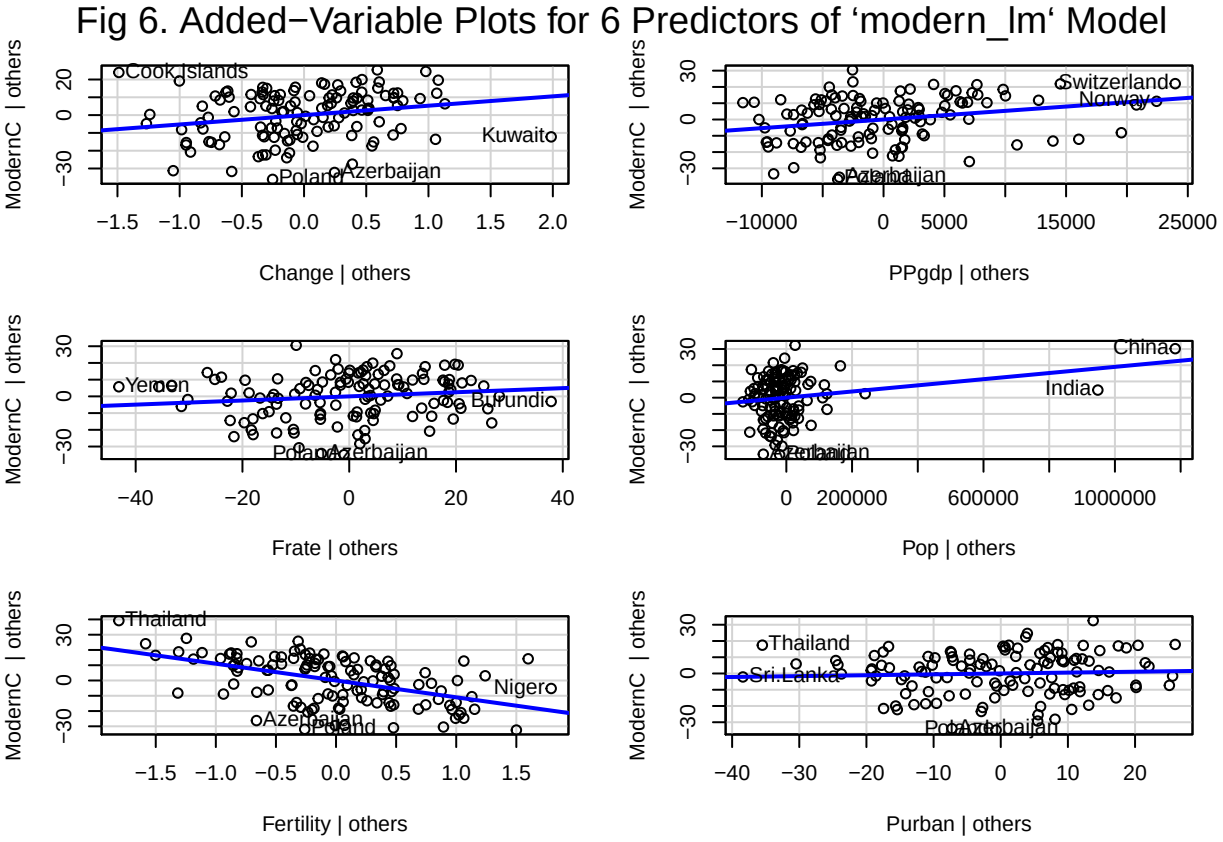
Using the `summary` function, we notice that the `lm` function automatically excluded 85 observations. Therefore, the multiple regression model was done 125 observations. Looking at these plots we can notice a minor heteroscedastic trend on the top-left plot, which shows that the variances is non-constant. However, this trend isn't that significant thus our assumptions of linearity still hold. On the Normal Q-Q plot we observe some points diverging from the normal line - especially on the top-right - which means our data follows a skewd normal distribution. Yet, our linerity assumption still holds as the observed standard deviations seem

to follow the theortical ones. Similarly, show that the errors for some observations are above 1 standard deviation and there are observations having a higher influence on the fit (i.e. China and India). However, these aren't significant enough to discard our assumptions of linearity. Overall, Cook's Island, Azerbaijian, Poland, China, and India are likely candidates for outlier testing given these observations tend to be farther away from the rest of the data or are highly influencial.

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 in **Fig 6** we can observe that a transformation on the `Pop` term would be helpful. This is noticeble by the large amount of data points concentrated around zero while only China and India seem to be spread out. Adding a linear transformation, such as `log`, would help to reduce the skewedness of the plot by expanding those values near zero and contracting those away from zero. It could also help reduce the current influence of the China/India observations on the fit of the model.
It is also noticeable how certain countries are influential for specific terms. For example, Kuwait and Cook's Island are an influential point for `Change`. India and China, as seen before, are influential on the `Pop` term. Noticeably, these plots brought up new influential localities for specific terms such as Norway and Switzerland for `PPgdp`, Niger and Thailand for `Fertility`, Yemen and Burundi for `Frate`, and Sri Lanka and Thailand for `Purban`. Again, we can notice previous localities such as Poland, Azerbaijian, and Cook's Island as being influential for certain terms such as `PPgdp` or `Purban`, among others.
Overall, from these plots we can notice the explanatory power of each predictor on the response variable after accounting for all the other predictors. From the slope we can notice almost all terms have either a positive or negative linear relationship with the response variable and thu contribute to explain the variability. `PUrban` and `Frate` seem to have the least explanatory power after accounting for all other predictors. A possible cause for this would be an existing multi-colinear relationship with another term already accounted in the linear model.

```
avPlots(model = modern_lm, main="Fig 6. Added-Variable Plots for 6 Predictors of `modern_lm` Model")
```

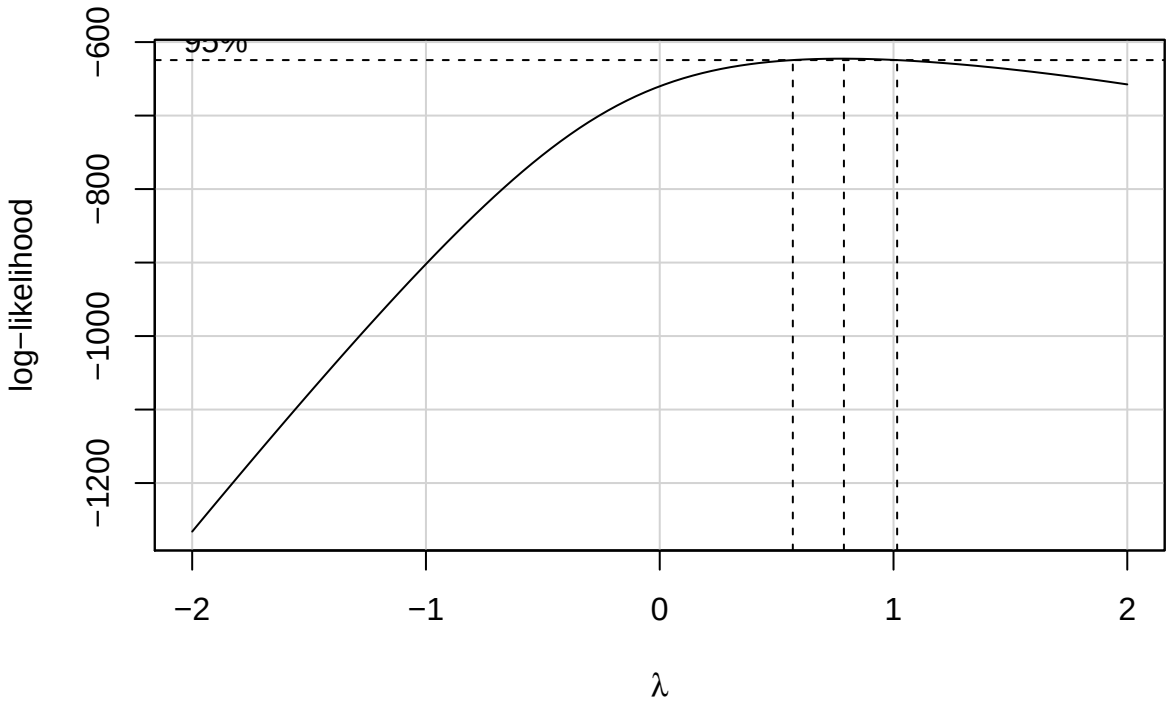


6. Using the Box-Tidwell `car::boxTidwell` or graphical methods find appropriate transformations of the predictor variables to be used as predictors in the linear model. If any predictors are negative, you may need to transform so that they are non-negative. Describe your method and the resulting transformations.

   As mentioned in the previous question, `Pop` seemed as the a candidate that would mostly benefit from a transformation given most of the data is all clustered near zero with two points spread far away. Using the `boxTidwell` built-in function the optimal transformation for $(Pop)^{\lambda}$ is $\lambda = 0.5$ which is close to the MLE of lambda given by `boxTidwell` $\lambda = 0.63$. In other words, $\sqrt{Pop}$.I tested for other predictors as well and although there were transformations that would help optimize the predictors for linearity, these didn't seem to aid in the interpretation. For example, the optimal transformation for `Fertility` was $\lambda = \frac{3}{2}$, however if we were to interpret what a change in $\sqrt{(Fertility^3)}$ means for `ModernC` it wouldn't be intuitive. Similar examples happened with other predictors such as `PPgdp` with $\lambda = .12$ and `Purban` with $\lambda = 2$ were optimal transformations, but these transformations wouldn't aid our interpretation which is our end goal.

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

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



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.

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?

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.

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!

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

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.*

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.