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

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.**

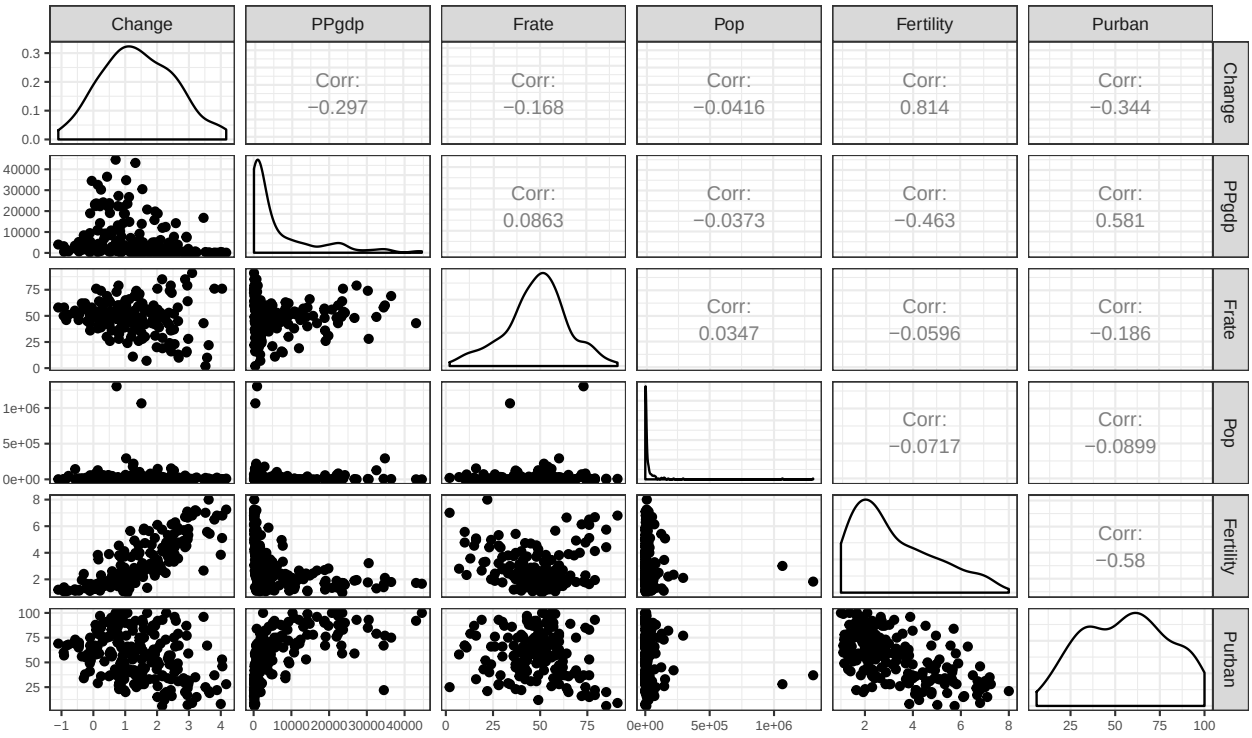
```
## 'data.frame':    210 obs. of  7 variables:
##  $ ModernC  : int  NA NA 49 NA NA NA 51 NA 22 NA ...
##  $ Change   : num  3.88 0.68 1.67 2.37 2.59 3.2 0.53 1.17 -0.45 2.02 ...
##  $ PPgdp    : int  98 1317 1784 NA 14234 739 8461 7163 687 NA ...
##  $ Frate    : int  NA NA 7 42 NA NA 63 44 51 53 ...
##  $ 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

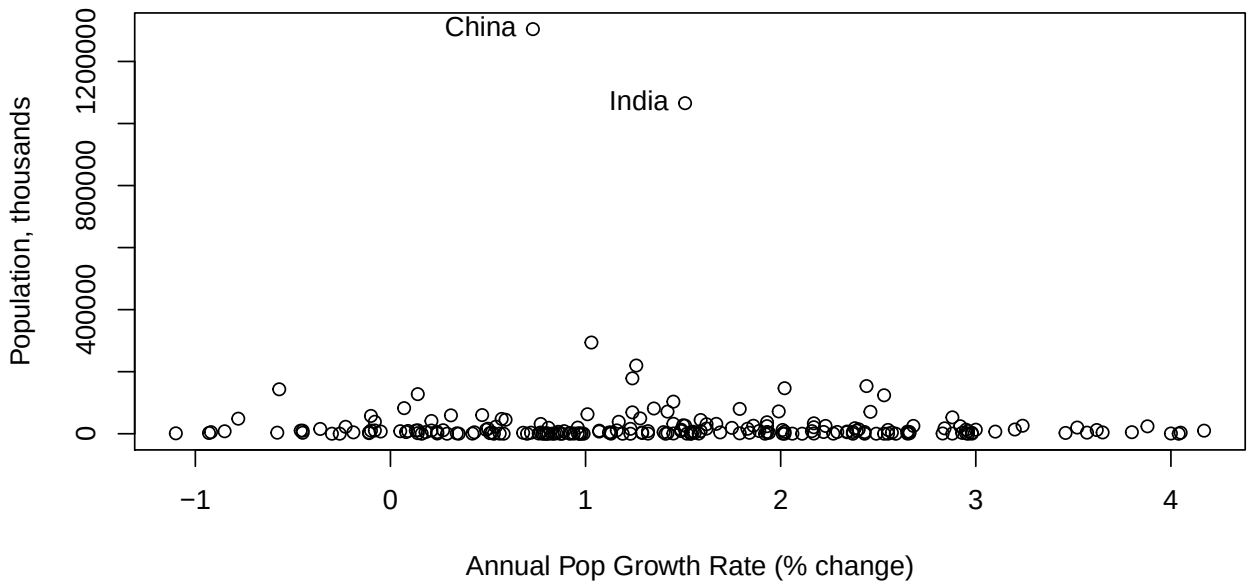
3. Investigate the predictors graphically, using scatterplots or other tools of your choice. Create some plots highlighting the relationships among the predictors. Comment on your findings regarding trying to predict 'ModernC' from the other variables. Are there potential outliers, nonlinear relationships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

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



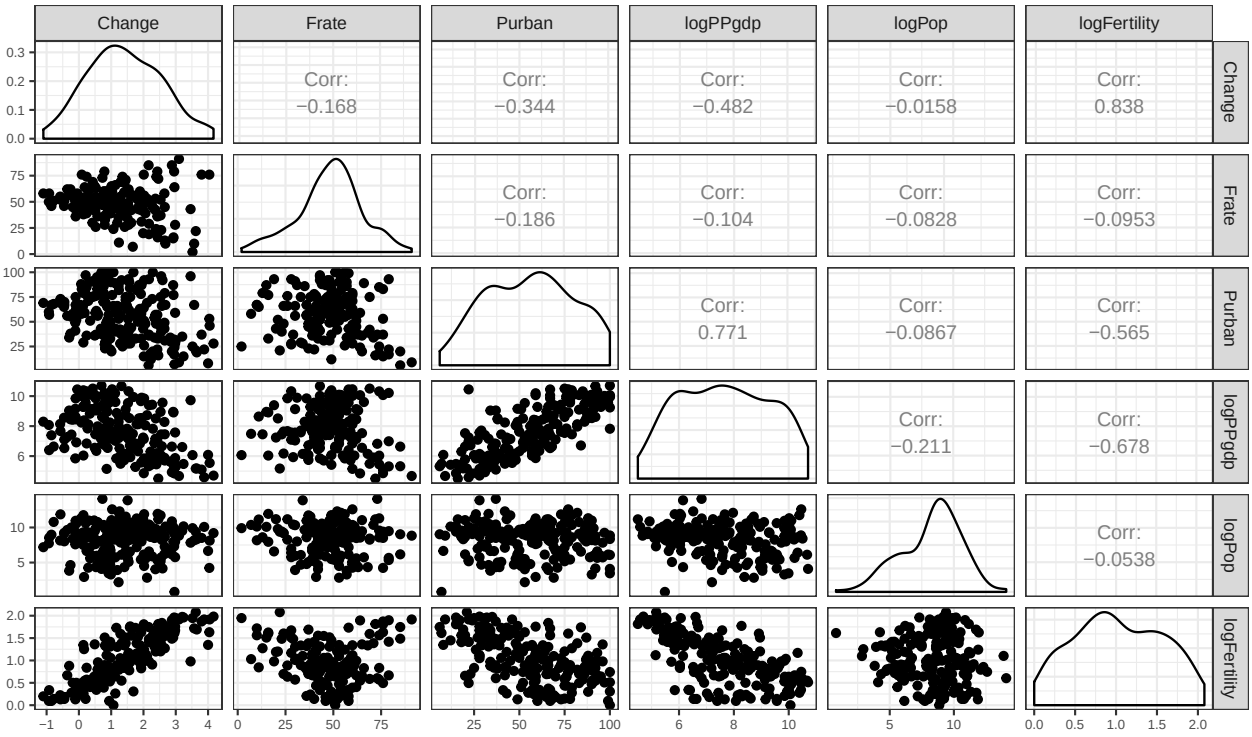
Using the `ggpairs` function we can do an initial assessment of the relationships among the predictor values. From **Fig 1** above it is easy to notice at first glance that many relationships among predictor variables seem non-linear. The relationships between the `Change`, `Purban`, `Fertility`, and `Frate` variables seem to have somewhat of a linear relationship. The `Frate` variable does seems to have a non-multicolinear relationship with some of the variables as well. However, two predictors stand out from this plot - `PPgdp` and `Pop`. The `PPgdp`'s relationships seem to follow an increasing or decreasing exponential, while in `Pop` we can nottice what could be two potential outliers (**Fig2**, below). In this plot we can see that China and India seem as potential outliers in terms of population vs other entities in the dataset.

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



Neverthless, it is important to note that the scale of `Pop` and `PPgdp` are several orders of magnitude higher compared to other predictors, which brings to mind possible linear transformations as a remedy. Thus, I explored whether a simple `log()` transformation would improve the linear relationship for some of the variables **Fig 1** (e.g. `logFertility`, `logPPgdp`, `logPop`). Even though this was a crude first transformation on the data, we can notice in **Fig 3** below that it does improve linear relationships among our predictors.

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



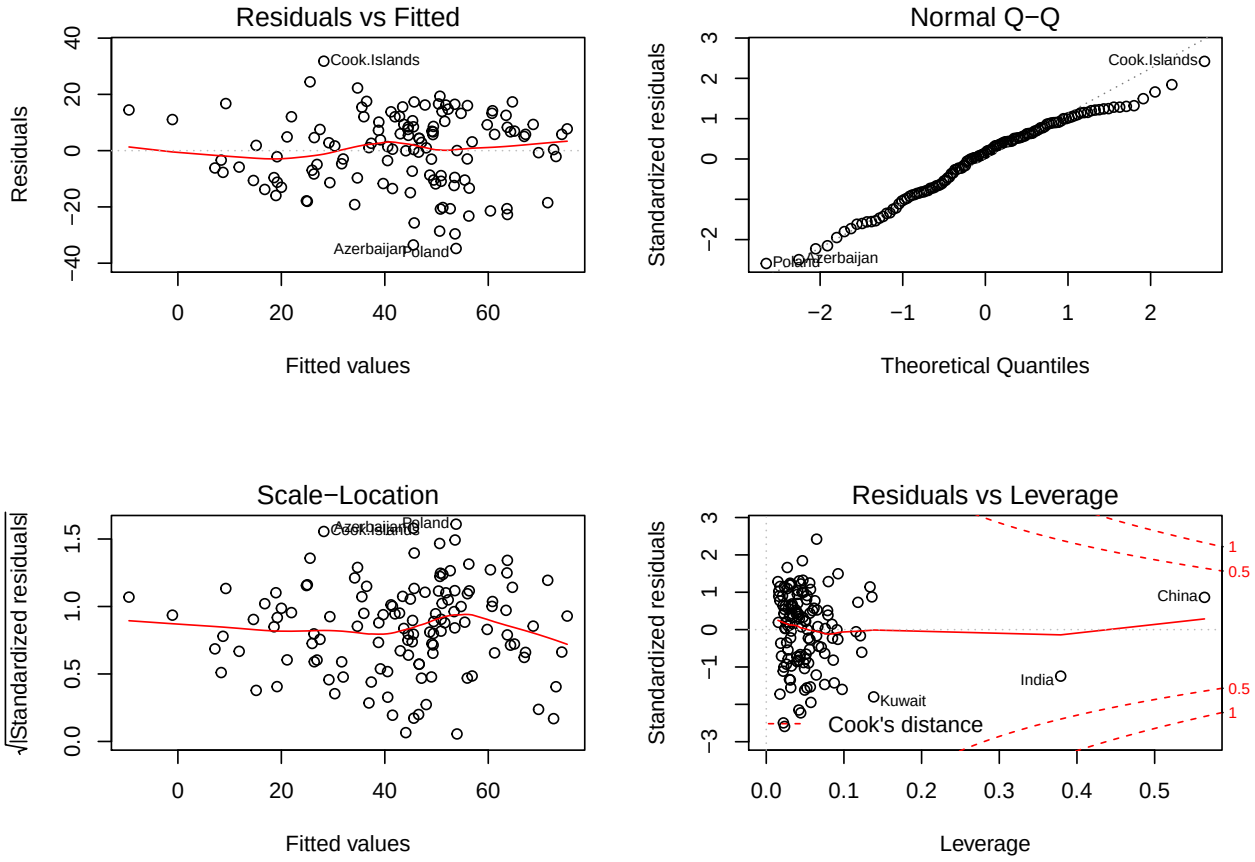
Finally, we can see that a linear combination of the these variables could be helpful to predict `ModernC`, although some would require transformations and others wouldn't. However, it is important to note that some predictors exhibit multicolinearily which means adding them to the linear model would be redundant as these would be contributing to explain the same variance and we should consider as we continue to build and assess the fit of the model.

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 from the initial comparisons in **Fig3** 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 mentioned model and following 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 **Fig 5** we can notice a minor heteroscedastic trend on the fitted vs residual plot, which shows that the variances is non-constant. However, this is a slight trend and doesn't have much effect on our assumptions of normality 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 assumptions still hold as the observed standard deviations seem to follow the theortical ones. Similarly, these plots show few standardize errors above 1 standard deviation and also few observations with high leverage that could be influencing the fit (i.e. China and India). However, these high leverage points aren't significant enough to discard our assumptions (i.e. Cook's Distance < 0.5). Overall, Cook Islands, Azerbaijian, Poland, China, and India are candidates for outlier testing given these observations tend to be farther away from the rest of the data or are potentially influencial.

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



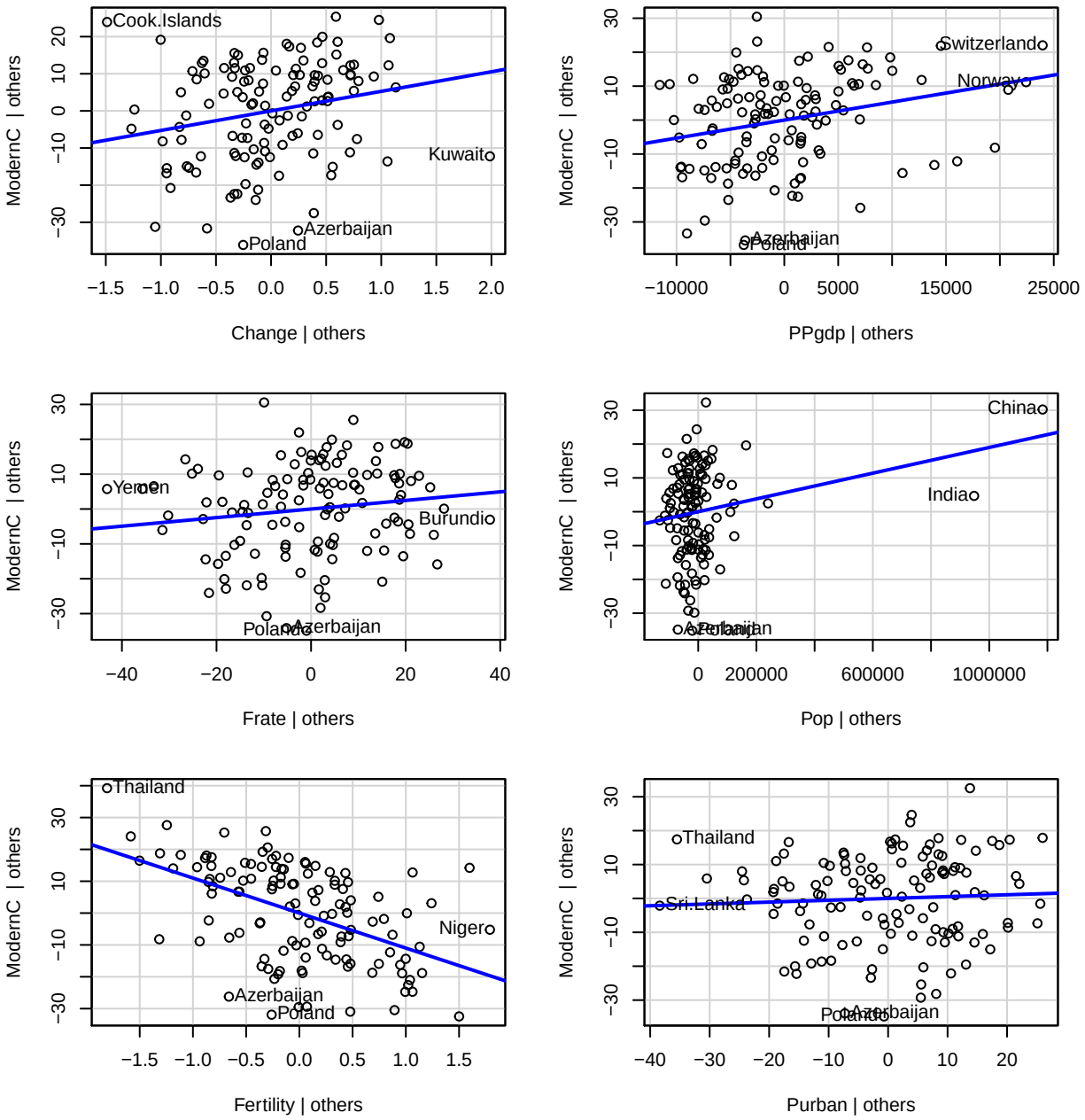
5. Examine added variable plots 'car::avPlot' or 'car::avPlots' for your model above. Are there any plots that suggest that transformations are needed for any of the terms in the model? Describe. Is it likely that any of the localities are influential for any of the terms? Which localities? Which terms?

From the plots in **Fig 6** we can observe that a transformation on `Pop` would be helpful. This is noticeble by the large amount of data points concentrated near zero while two points - China and India - are orders of magnitude apart. Adding a linear transformation, such as `log`, would help to reduce the skewedness of the plot by bringing the values near zero and far away from zero closer together. It could also help reduce the possible 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 they 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.

Fig 6. Added−Variable Plots for 6 Predictors of ‘modern_lm‘ Model



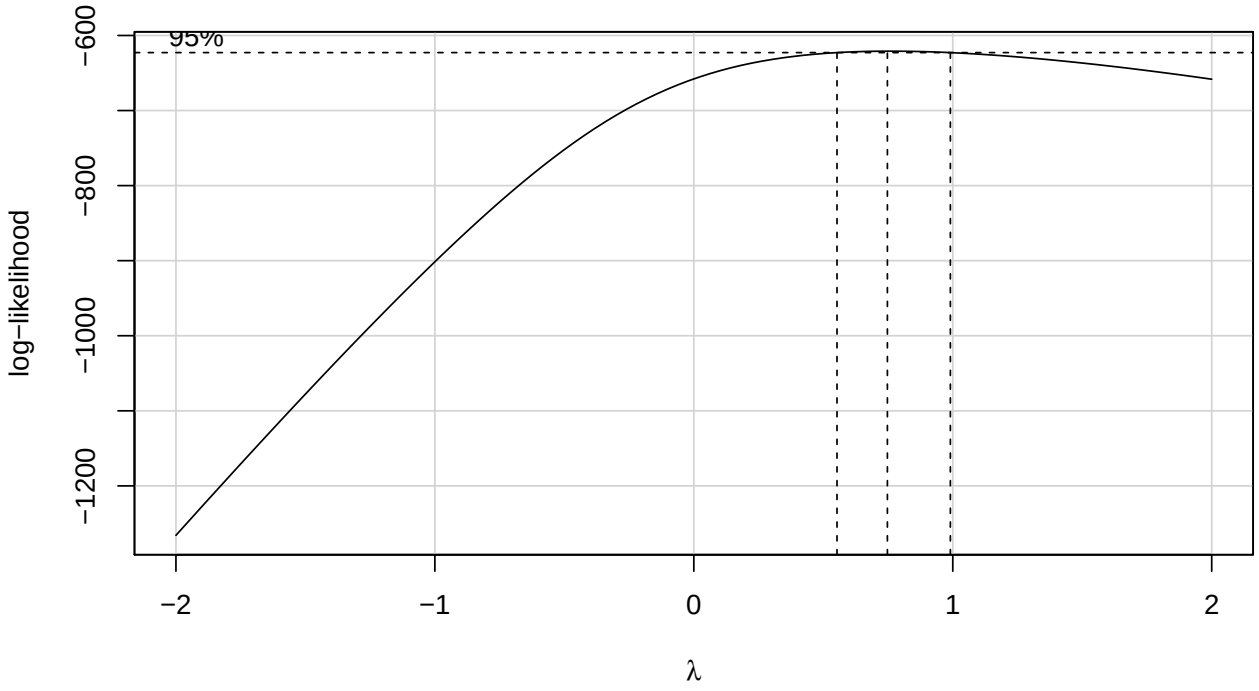
6. Using the Box-Tidwell ‘car::boxTidwell‘ or graphical methods find appropriate transformations of the predictor variables to be used as predictors in the linear model. If any predictors are negative, you may need to transform so that they are non-negative. Describe your method and the resulting transformations.

As mentioned in the previous question, `Pop` seemed as the most promising candidate that would benefit from a transformation given the data was clustered near zero with two points spread far away. Using the `boxTidwell` built-in function the optimal transformation is close to $\lambda = 0.5$ (i.e $\sqrt{Pop}$). However, from this method the transformation does not significantly improve the fit as shown by a non-significant p-value with $\alpha = 0.5$ (i.e. the transformation isn't significantly different from $\lambda = 1$ - our H_0. I similarly tested other predictors, but the suggested transformations didn't provide enough evidence to reject the H_0. Just to double check that highly influential points on `Pop` such as China and India weren't giving us a different transform

estimation, I removed them from a modified dataframe and re-test. Even when we remove these we still fail to reject H_0.
However, graphically from **Fig 3** we can see that a `log()` transformation does improve the linear relationship between the `Pop`, `PPgdp` and `ModernC` - thus I will use these transformations.

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

From the plot below we can observe the MLE for λ using `boxCox`. Since the 95% confidence interval doesn't include 1 and the `boxCox` $\lambda = 0.76$, I decided to use the closest λ that would still allow for some interpretation $\lambda = 0.5$. This is because we want a model that is both improved by a transformation, but at the same time it is interpretable.



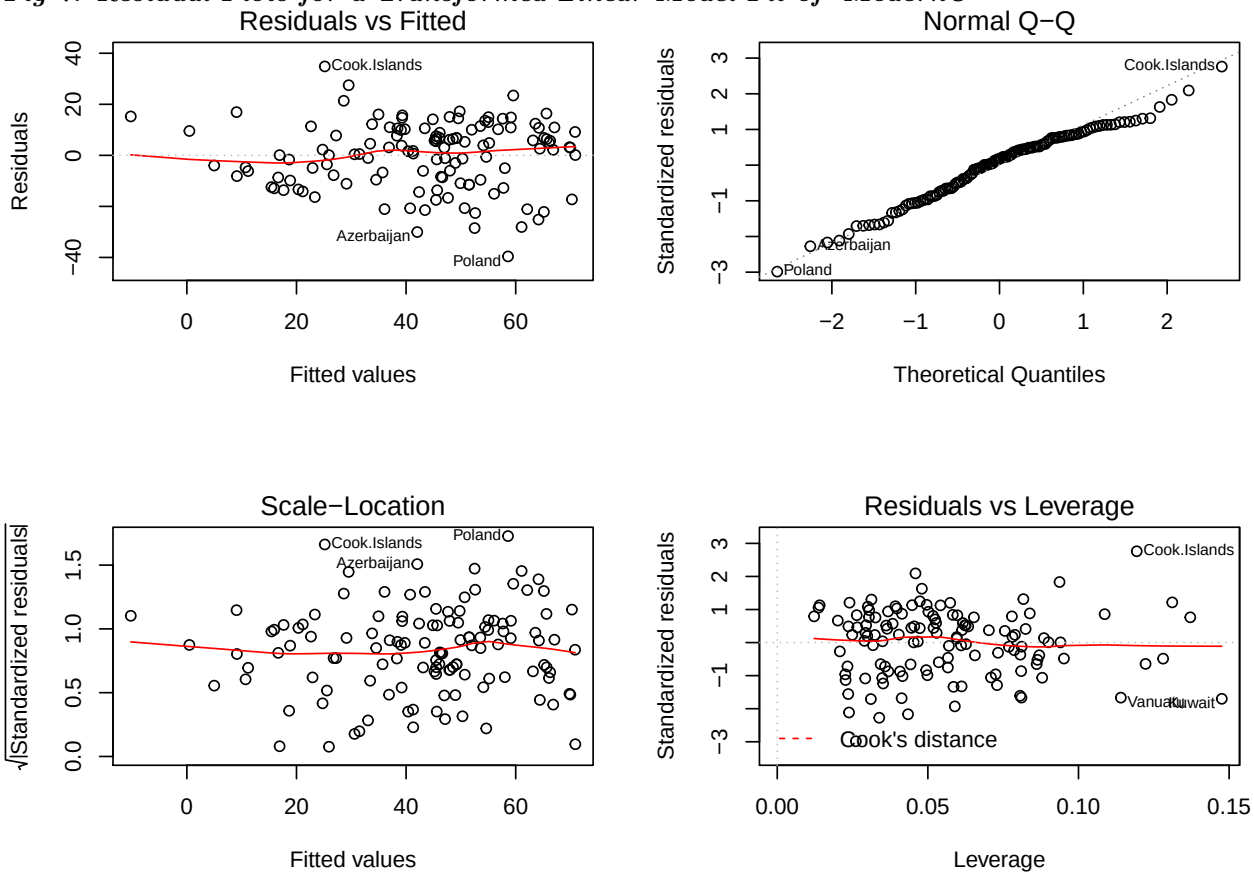
8. Fit the regression using the transformed variables. Provide residual plots and added variables plots and comment. If you feel that you need additional transformations of either the response or predictors, repeat any steps until you feel satisfied.

From the previous question I had decided to use $\lambda = 0.5$ for the response variable given it was close to the significant MLE for this parameter with $\alpha = 0.50$. However, after plotting the residual and added variable plots I noticed this transformation wasn't benefitial. After trying the λ's MLE and $\lambda = 1$ as well, I decided to proceed with the later given it improves the model and interpretability. Thus in **Fig 7** we observe the residual plots for,

$$ModernC = \beta_0 + \beta_1 Change + \beta_2 logPPgdp + \beta_3 Frate + \beta_4 logPop + \beta_5 Frate + \beta_6 Purban$$

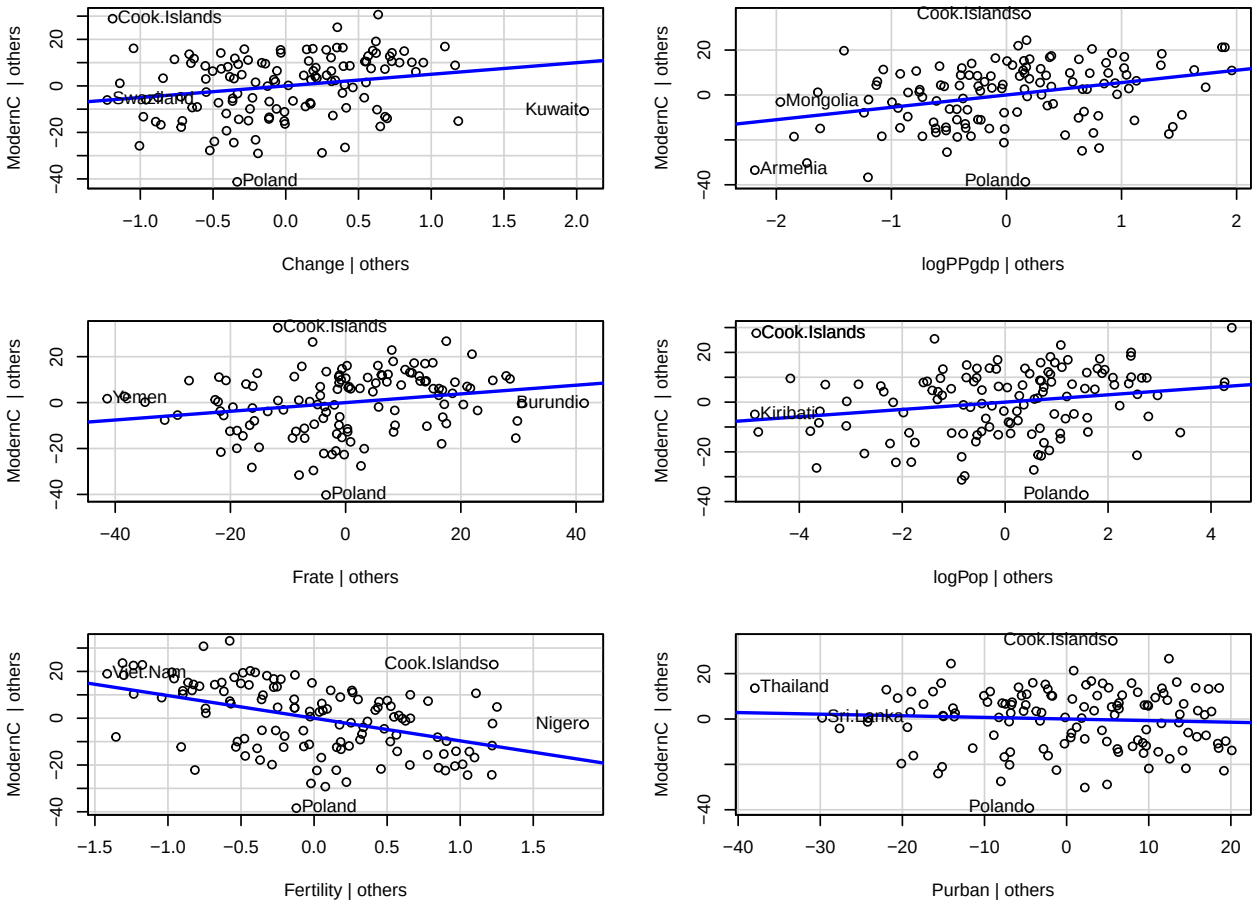
Compared to **Fig 5**, we can observe two improvements: 1) we don't see the previous high leverage points (China/India) near a Cook's Distance that would raise a red flag for influential points nor any other point, 2) there is a minor correction of the points in both tails toward the diagonal line in the Normal Q-Q. Even when the transformations did very slight corrections to the minor heteroscedastic trends, these are still within reasonable bounds.

Fig 7. Residual Plots for a Transformed Linear Model Fit of 'ModernC'



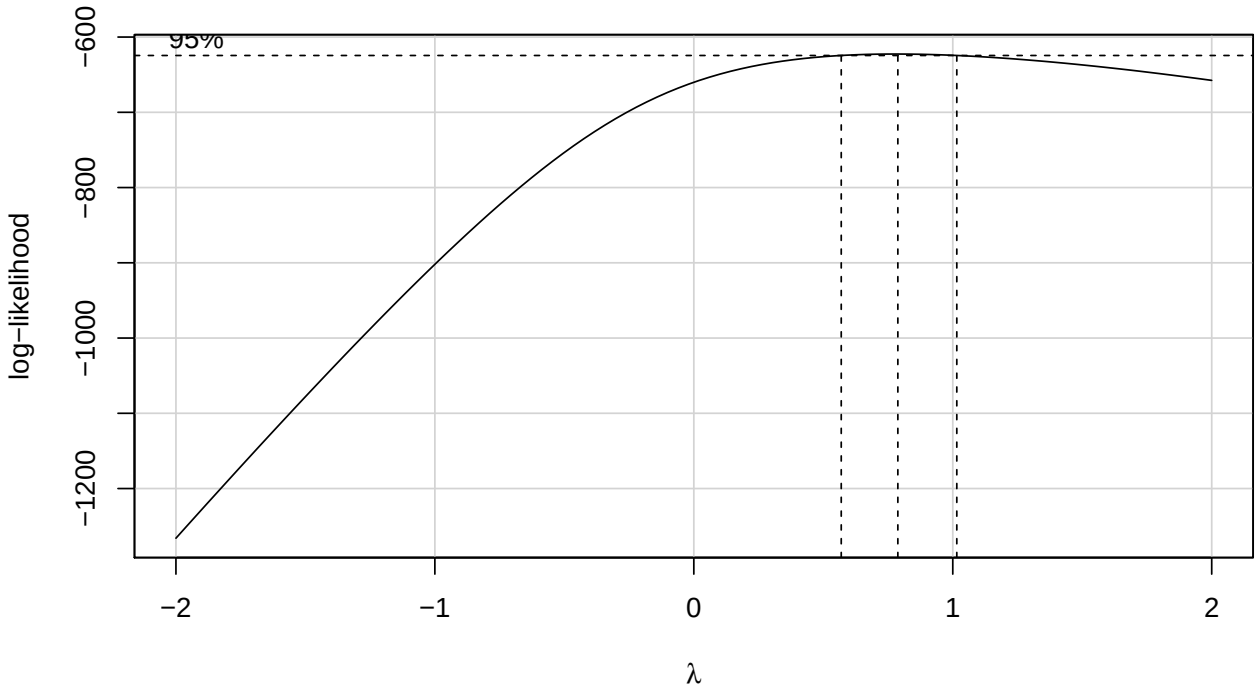
Comparing the added-variable plots to **Fig 6**, we can see that in **Fig 8** there are two changes on the transformed predictors. `logPop`'s slope is less prominent than without a transformation now that we have China or India closer to the other observations and not exerting an higher influence on the value of the coefficient. In `lopPPgdp` the data is closer to each other as expected with the transformation and we don't see Norway or Switzerland as potential outliers as we did before.

Fig 8. Added−Variable Plots for 6 Predictors of 'modern_lm2' 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 have a slightly different model if we start with a transformation for the response that is close to the `boxCox` MLE of $\lambda = 0.78$ and that is also interpretable (i.e. $\lambda = 0.5$, see plot below). Assuming we proceed with this value, we would find a `boxTidwell` λ for `Fertility` that is significantly different ($\alpha = 0.05$) from 1 - i.e. $\lambda = 1.5$. To make it more interpretable I would have chosen $\lambda = 2$ instead for `Fertility`. However, given the 95% confidence interval for the response λ includes 1 (as seen below), I wouldn't feel comfortable choosing any transformation in the first place. Thus, we would be back looking to find predictor λs with an untransformed response which was shown in Question 6.



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.

I used the Bonferroni Correction to test for outliers where $p_i < \frac{\alpha}{n}$ where $\alpha = 0.5$. From this test no observations have signficant p-value [results and code hidden], thus none can be discarded as outliers. Regarding influential points, we can observe from **Fig 7** that even though with the transformations there are points with high leverage such as `Cook Islands`, `Vanuatu` and `Kuwait`. However, none are influential points.

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!
    Before finalizing the model and deciding which predictors to include, I tested whether keeping `Purban` would improve the model and reduce the RSS. To do so I used an ANOVA to test H_0 - i.e. whether $\beta_6 = 0$ or not. From the resulting p-value (with $\alpha = 0.05$), we fail to reject H_0 and therefore I have excluded `Purban` from the final model:

$$Final\ Model \rightarrow \quad ModernC = \beta_0 + \beta_1 Change + \beta_2 logPPgdp + \beta_3 Frate + \beta_4 logPop + \beta_5 Fertility$$

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

Below is summary table of the coefficients with 95% confidence interval in terms of the original units.

	2.5%	97.5%
(Intercept)	-24.569	32.773
Change	0.673	8.723
PPgdp	15.129	1098.940
Frate	0.050	0.349
Pop	1.223	14.598
Fertility	-12.595	-5.962

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

The following model provides a linear relationship between the percent of unmarried women using a modern method of contraception (`ModernC`) and a combination of socio-economic indicators (`Change`, `PPgdp`, `Frate`, `Pop`, `Fertility`, `Purban`) from 125 entities.

$$ModernC = 4.10 + 4.70(Change) + 4.85(logPPgdp) + 0.20(Frate) + 1.44(logPop) - 9.27(Fertility)$$

This model can be interpreted in the following manner,
1. Every unit increase in annual population growth rate implies a 4.70 unit increase in percent of unmarried women using contraception (with all else held constant)
2. Similar claims can be done for `Frate` and `Fertility`. A 5 unit increase in `Frate` implies a 0.2 unit increase in `ModernC`, while a 1 unit increase in `Fertility` implies a -9.27 decrease in `ModernC`. For each case we must keep the others indicators constant.
3. For `Pop`, we can interpret it using percent changes in `Pop`. For example, every doubling of the population (i.e. 100% increase) implies an increase of 0.43 units in `ModernC` with every other indicator help constant.
4. Similarly, for `PPgdp` when the per capita GDP (in USD) is doubled (i.e. 100% increase) this implies an increase of 1.46 units in `ModernC` with all else held constant.

Although the original data contained data from 210 entities, 85 entities were excluded due to missing data in one or more socio-economic indicator.

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.
Before starting some clarifications on the notation: Let's assume our model is,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

- $\hat{Y}_{(1)}$ refers to the Y vector regressed on all the predictors minus the X_1 predictor
- $\hat{X}_{1\cdot 2}$ refers the X_1 predictor being regressed on the rest of the predictors (i.e. X_2)
- Therefore, $X_1 - \hat{X}_{1\cdot 2} = \hat{e}_{1\cdot 2}$ refers to the residuals from the regression of predictor X_1 on the X_2.
- Similarly, $Y - \hat{Y}_{(1)} = \hat{e}_{(1)}$ refers to the residuals from the regression of Y on all the predictors minus the X_1
- Althought the correct notation for the hat matrix without X_1 would be $H_{(1)}$, we will use H for simplicity given both regressions use the same hat matrix

Knowing the above we can prove that β_0 of an added variable plot will always be zero from,

$$\hat{e}_{(1)} = \vec{1}\hat{\beta}_0 + \hat{\beta}_1 \hat{e}_{1\cdot 2}$$

Thus we can find $\hat{\beta}_1$ using the $(X^TX)^{-1}X^TY$ notation, but now thinking of $X \equiv \hat{e}_{1\cdot 2}$ and $Y \equiv \hat{e}_{(1)}$.

$$\begin{aligned}
\hat{e}_{(1)} &= \vec{1}\hat{\beta}_0 + \overbrace{\Big[\big((I-H)X_1\big)^T(I-H)X_1\Big]^{-1}\Big[(I-H)X_1\Big]^T Y}^{\hat{\beta}_1}\,(I-H)X_1 \\
&= \vec{1}\hat{\beta}_0 + \Big[X_1^T \underbrace{(I-H)(I-H)}_{(I-H)} X_1\Big]^{-1} \underbrace{X_1^T(I-H)Y}_{1x1\ scalar}\underbrace{(I-H)X_1}_{1x1\ scalar}
\end{aligned}$$

If you multiply both sides by X_1^T and rearrange the scalars you get,

$$X_1^T(I-H)Y = X_1^T\vec{1}\hat{\beta}_0 + \underbrace{X_1^T(I-H)X_1\big[X_1^T(I-H)X_1\big]^{-1}}_{I} X_1^T(I-H)Y$$

$$\therefore \qquad X_1^T(I-H)Y = \sum_{i=1}^{n} X_{i,1}\hat{\beta}_0 + X_1^T(I-H)Y$$

$$\sum_{i=1}^{n} X_{i,1}\hat{\beta}_0 = X_1^T(I-H)Y - X_1^T(I-H)Y = 0$$

Thus we can see that the only way this relationship can only be zero is if $\hat{\beta}_0 = 0$ (i.e. the intercept is 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.

To confirm that the coefficients are the same, I obtained the residuals from regressing `ModernC` on all predictors except `Purban` (e_Y) and the residuals from regressing `Purban` on the rest of the predictors (e_Purban). Then, regressing these two linear models provides us with the coefficient for `Purban` when compared to that of the original model (i.e. Y regressed on all predictors). **Fig 9** below shows the plot e_Y vs. e_Purban and demonstrates a slope of -0.070768

	Original	A-V Plot
Coeffs	-0.070768	-0.070768

Fig 9. Added−Variable Plot for ‘ModernC‘ ~ X w/o ‘Purban‘ vs ‘Purban‘ ~ X w/o ‘Purban‘