# HW2 STA521 Fall18

*Evan Poworoznek netID: elp28 github: Poworoznek*

*Due September 23, 2018 5pm*

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
```

```
?UN3
head(UN3)
```

```
##             ModernC Change PPgdp Frate   Pop Fertility Purban
## Afghanistan      NA   3.88    98    NA 23897      6.80     22
## Albania          NA   0.68  1317    NA  3167      2.28     43
## Algeria          49   1.67  1784     7 31800      2.80     58
## Am.Samoa         NA   2.37    NA    42    57        NA     53
## Andorra          NA   2.59 14234    NA    64        NA     92
## Angola           NA   3.20   739    NA 13625      7.20     35
```

```
sapply(UN3, function(x) {sum(is.na(x))})
```

```
##   ModernC    Change     PPgdp     Frate       Pop Fertility    Purban 
##        58         1         9        43         2        10         0
```

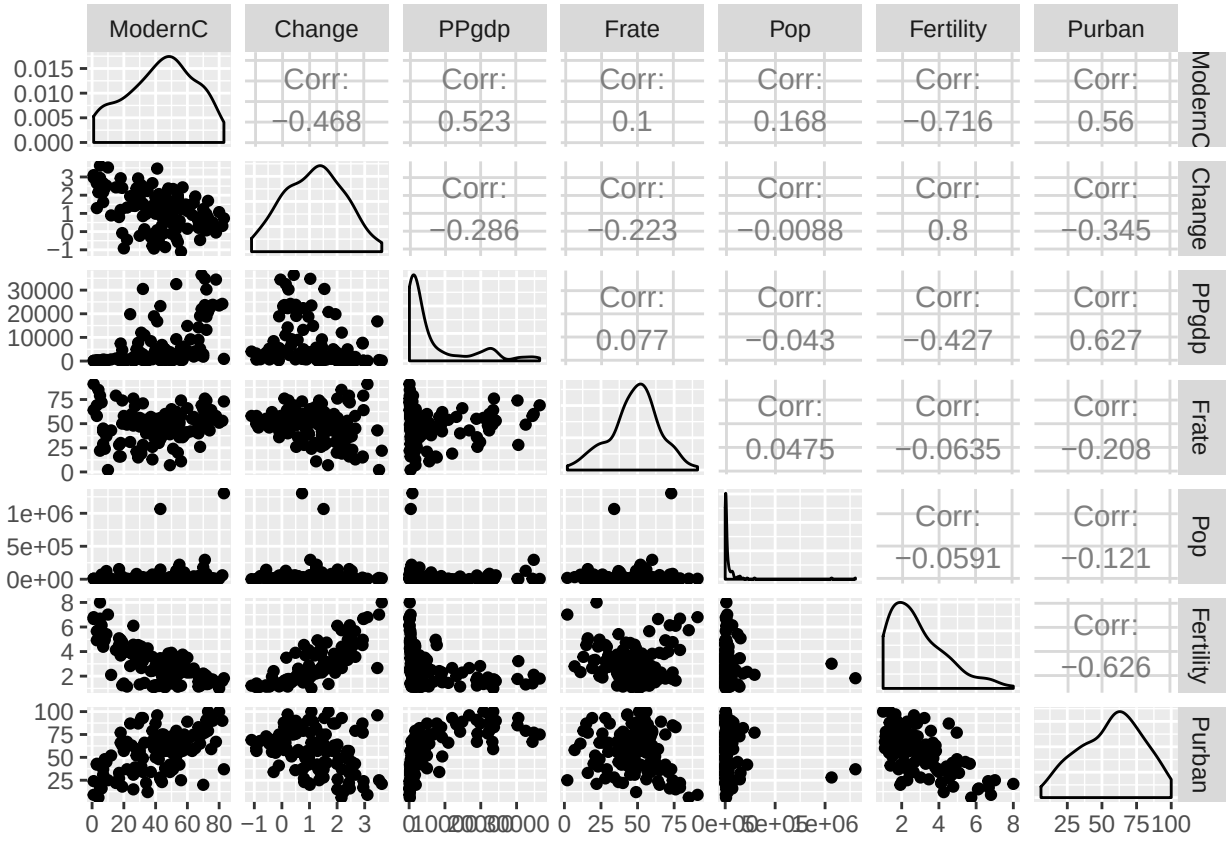
All of the variables in this dataset are quantitative, and the help file confirms that none of them are nominal or ordinal factors masquerading as numeric. Purban is the only variable without missing values, and the remaining variables have between 1 and 58 missing values.

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

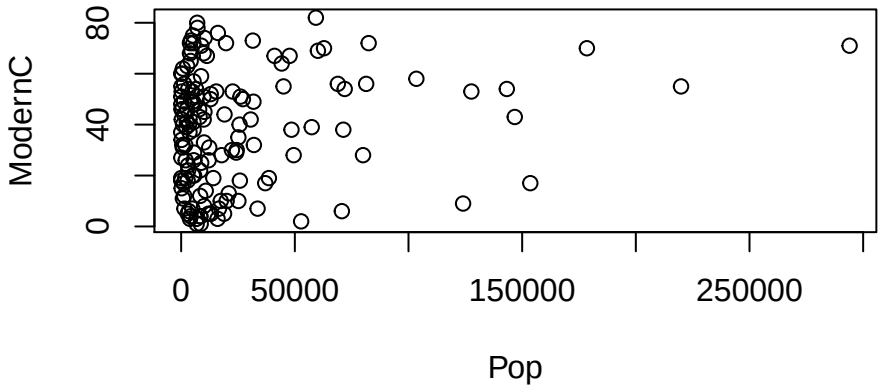
```
tabledata = cbind(sapply(UN3, function(x) {mean(na.omit(x))}),
                  sapply(UN3, function(x) {sd(na.omit(x))}))
kable(tabledata, col.names = c('mean', 'standard deviation'), booktabs = T)
```

| | mean | standard deviation |
|---|---|---|
| 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 |

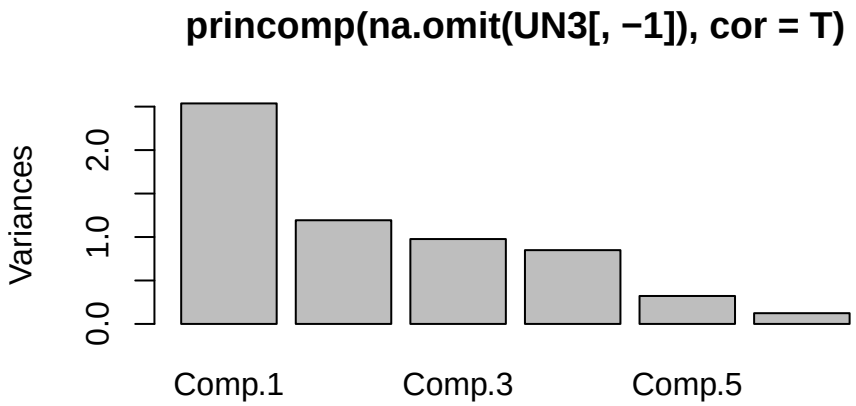
3. Investigate the predictors graphically, using scatterplots or other tools of your choice. Create some plots highlighting the relationships among the predictors. Comment on your findings regarding trying to predict `ModernC` from the other variables. Are there potential outliers, nonlinear relationships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




The scatterplot matrix shows that there are some issues in this regressor set. The predictors have pearson correlation coefficients with magnitudes between 0.1 and 0.7 for their relationship with ModernC, so they do appear to be associated with ModernC. Some of the relationships between ModernC and the predictors appear only slightly nonlinear (Change, Frate, Fertility, Purban), but the relationship between ModernC and PPgpd appears very non-linear and very heteroskedastic. The relationship between ModernC and Pop is difficult to determine because of the two population outliers, India and China.




The scatterplot for these variables with India and China excluded is shown below the scatterplot matrix, and does not show a visible association. Just among The regressor set there are issues. This does not qualify as a set of linear predictors as defined in the text, because there are clear non-linear relationships between the regressors (PPgpd shows a non-linear relationship with just about every other predictor, and the relationship between fertility and change is apparently non-linear). That indicates that our standard residual plots are no longer interpretatible in the same manner.



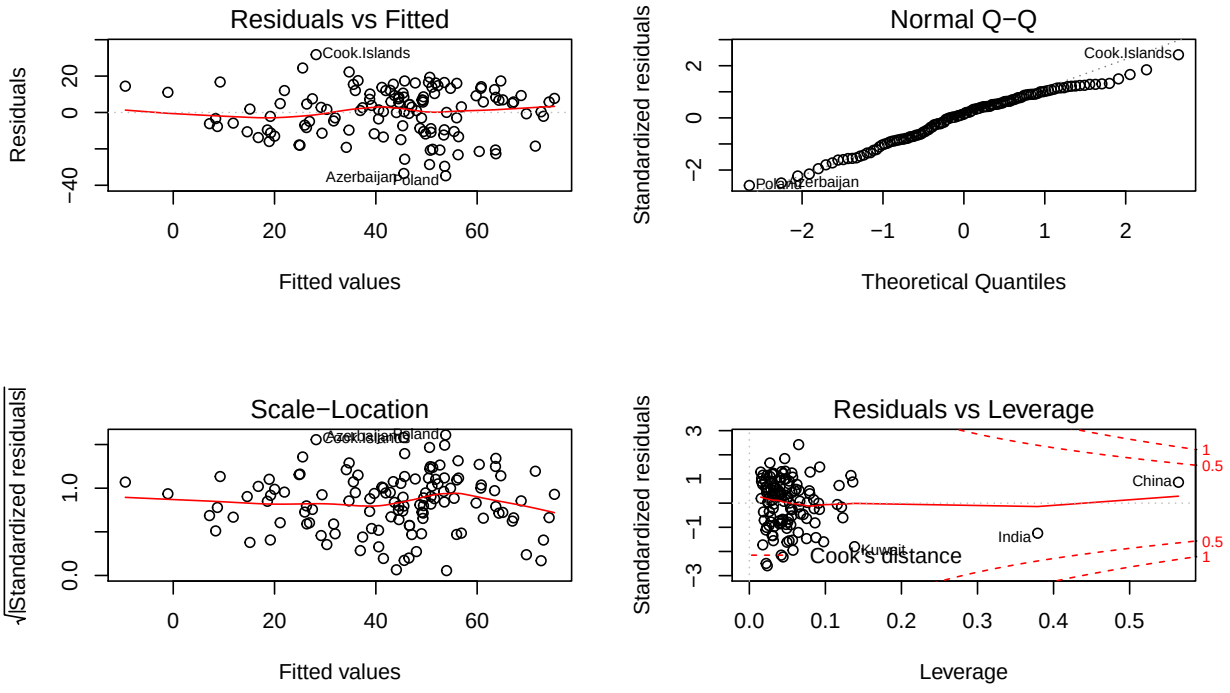

If we consider only the linear part of the relationships in the regressor set with a linear approximation in a principle component analysis we can see that there is a decent amount of variance split between the regressors in this set, so the set is not close to singular. The plot showing the comparative variance amounts in this principal components analysis is shown above.

## 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?

```
##
## 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
```



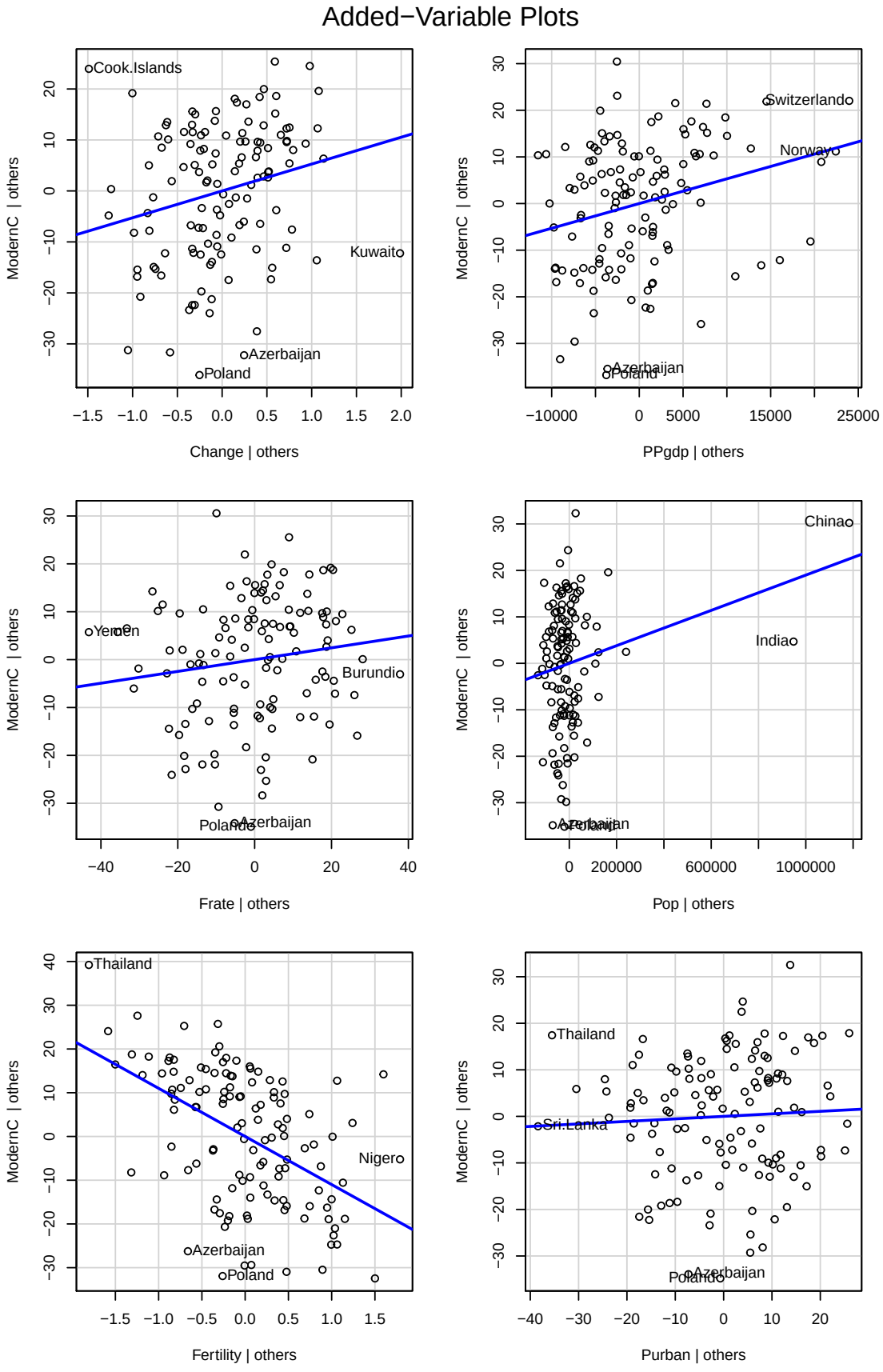
85 observations are excluded by the lm function because those 85 rows contain at least one missing value, and the default behavior of the lm function is to exclude completely rows without a complete set of values for the specified regressors. 125 observations remain with which to fit the linear model. The diagnostic plots for the standard least squares regression with all the covariates in the datasets appear to support the set of assumptions required to make conclusions. The residual by fitted plot does not appear to show a non-linear pattern or a major trend in the variance across the fitted axis. There appears to be a slight reduction in variance toward the lower end of the fitted values, but that appearance could just be a result of the fewer datapoints in that area. From this it appears valid to assume that the residuals are independent and identically distributed around mean 0 with constant variance.

The normal quantile-quantile plot of the residuals indicates near-normality, but with a slight left skew. Overall the normal assumption is relatively robust and the slight skew might only affect very sensitive inference. The Cook Islands look somewhat isolated in the normal q-q plot, but they are not very far off the 1:1 line, or the expected pattern. Likewise, the Scale-Location plot does not show any concerning features, and appears to be patternless noise. The residuals-leverage plot shows that india and china have very high leverage, but not very large standardized residuals.

Just because these plots appear to show that our assumptions hold, does not mean that they necessarily do. These plots are unreliable due to the pairwise non-linear relationships visible in the scatterplot matrix above, and particularly when we divide by the diagonal hat value to standardize, that does not appropriately capture the variance at the point without a set of linear predictors. for this reason, the individual relationships between the predictors and the model should be evaluated.

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?

```
par(mfrow = c(1,1))
avPlots(un3.lm)
```


Added−Variable Plots
ModernC | others
Change | others
Cook.Islands
Kuwait
Azerbaijan
Poland
PPgdp | others
Switzerland
Norway
Frate | others
Yemen
Burundi
Chinao
India
Pop | others
Thailand
Nigero
Fertility | others
Sri.Lanka
Purban | others

The a-v plots for Frate and Change indicate that the linearity assumption is upheld for those variables. The a-v plots for PPgpd and Purban have distinct non-linear trends, and indicate that a transformation in either the response or those predictors is necessary to restore linearity. The a-v plot for fertility is inconclusive, as there is a hint of a nonlinear U shaped pattern, but it is not strong enough to indicate that a transformation is required. The a-v plot for Pop shows that India and China are extremely influential in deciding the model's population effect. It is difficult to assess linearity for Pop, because India and China prevent any patterns in the rest of the data from showing on the plot. It is also hard to tell if the relationship between population and ModernC is similar for the datapoints not including India and China.

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.

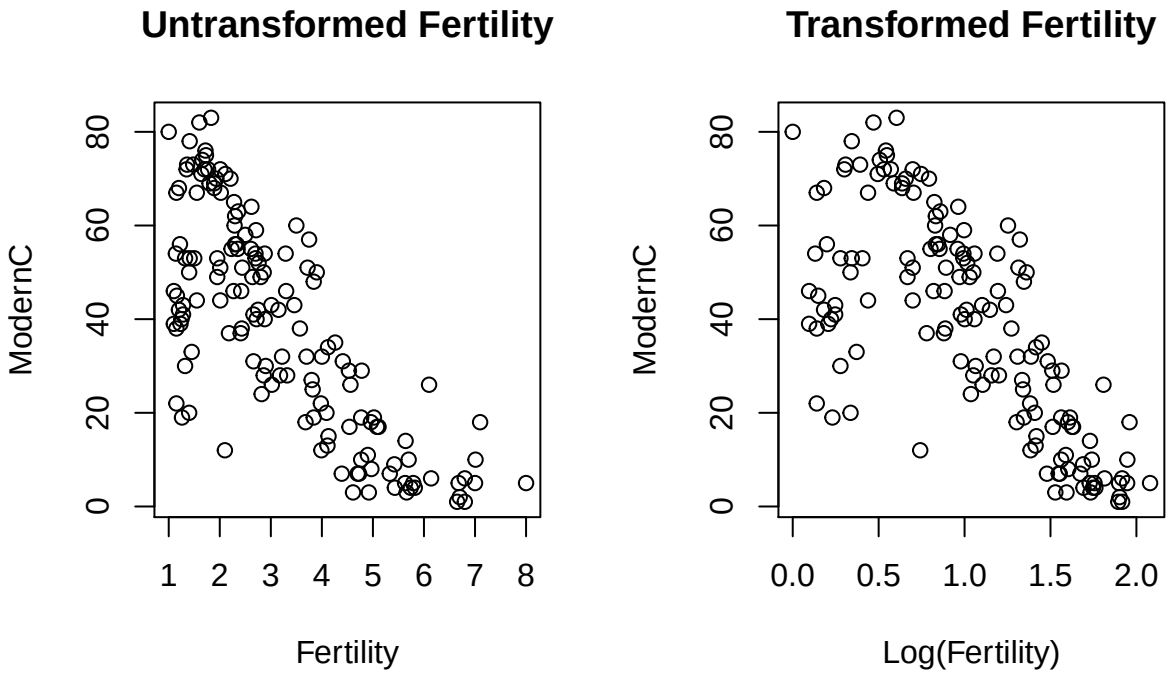
```
un3.na.o = na.omit(UN3)
boxTidwell(ModernC ~ PPgdp + Fertility + Purban,data = un3.na.o)
```

```
## Warning in boxTidwell.default(y, X1, X2, max.iter = max.iter, tol = tol, :
## maximum iterations exceeded
```

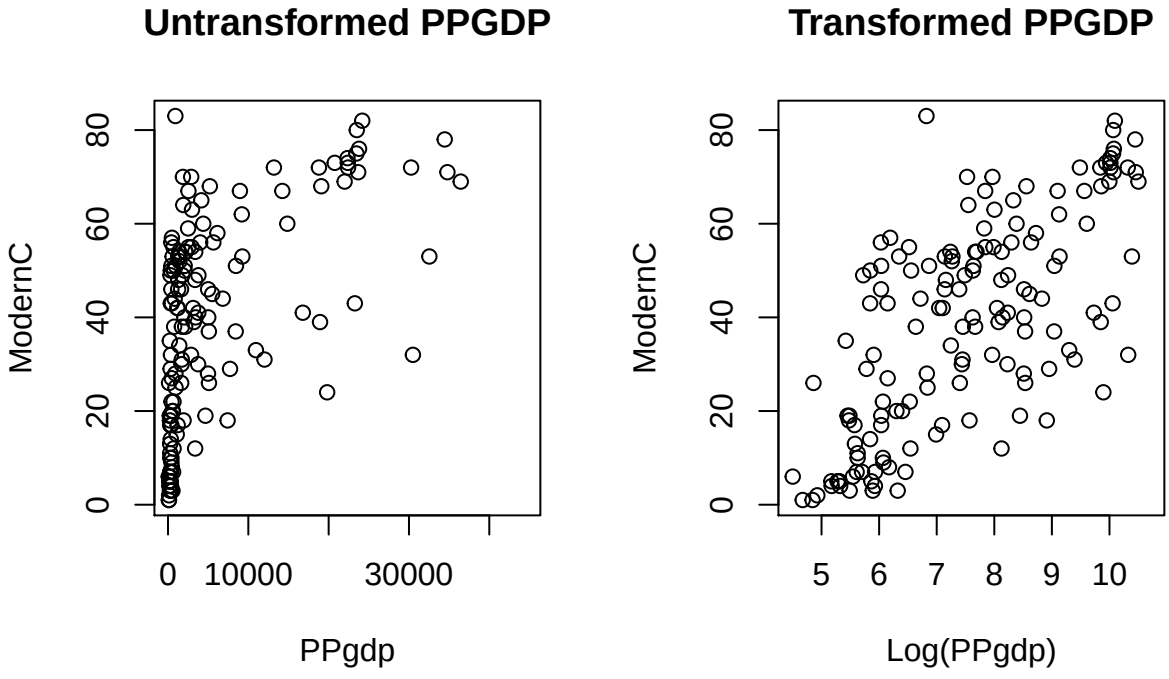
```
##           MLE of lambda Score Statistic (z) Pr(>|z|)
## PPgdp           0.66675             -0.3645 0.715509
## Fertility       1.61788             -2.8141 0.004891 **
## Purban          4.87644              1.4484 0.147519
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## iterations =  26
```

Change was not considered a candidate for transformation in boxTidwell because it contains negative values and its a-v plot appeared to indicate that it was linear in its effect on ModernC. In addition, including Pop and Frate in the formula of the boxTidwell function resulted in failure. For these reasons boxTidwell is used to test PPgdp, fertility, and Purban, and the other predictors are analysed graphically. It is also important to note that in the boxTidwell function, an optimization function did not converge nicely (indicated by the "maximum iterations exceeded" warning), and the values shown above are perhaps unreliable.

In the boxTidwell summary, fertility had a significant p value (0.005) indicating its relationship with the response is significantly non-linear. This is confirmed looking at the bivariate plot of Fertility and ModernC, shown below, where a clearly non-linear pattern is evident. The log transform results in the majority of the data following a linear pattern, shown in the second bivariate plot below. There is a clump of data at different levels of ModernC, but log(fertility) values around 0. That there are these two clumps of data for which separate linear relationships appear appropriate indicates that there may be clusters in our data that deserve separate analysis or special attention.

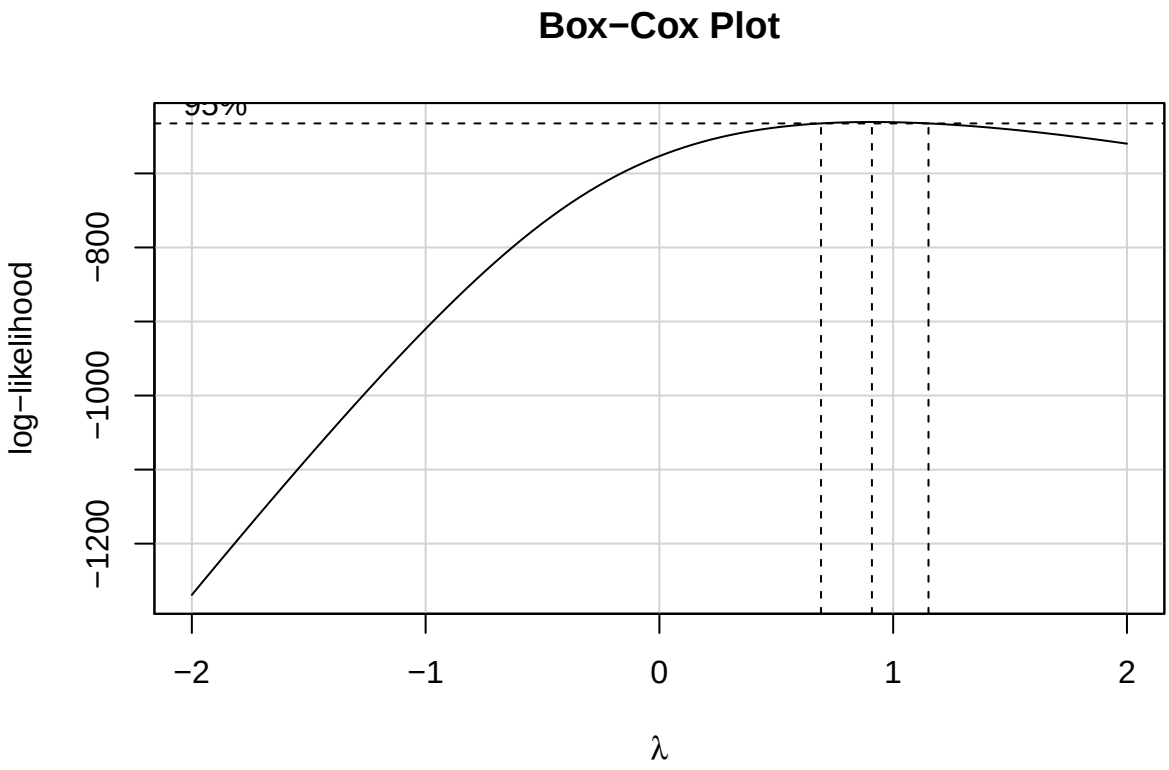



Fertility was the only predictor the boxTidwell function considered significant in the nonlinearity of its relationship with the response, but I have reservations about other variables. As shown below, the log transform of PPgdp results in a linear relationship between that transformed predictor and the response, when the original scale of PPgdp had a strongly non-linear relationship with ModernC.



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

```
boxCox(un3.na.o$ModernC ~ log(un3.na.o$Fertility) + un3.na.o$Purban +
         log(un3.na.o$PPgdp) + un3.na.o$Frate + un3.na.o$Pop + un3.na.o$Change)
title('Box-Cox Plot')
```




The Box-Cox plot for the model with log transformed fertility and PPgdp with the full set of other regressors shows that the lambda value 1 falls on the 95% confidence interval for the optimal lambda, so no transformation is recommended in this case.

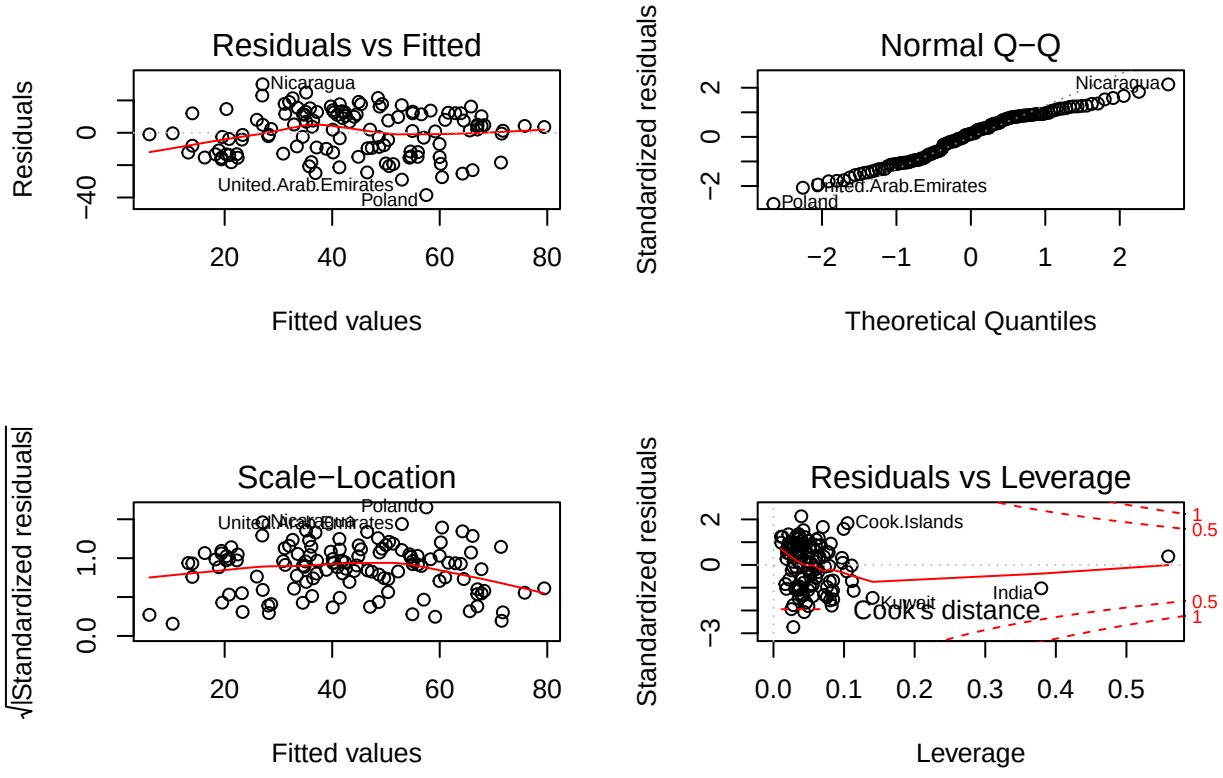
8. Fit the regression using the transformed variables. Provide residual plots and added variables plots and comment. If you feel that you need additional transformations of either the response or predictors, repeat any steps until you feel satisfied.

```
un3.lm2 = lm(ModernC ~ log(Fertility) + log(PPgdp) + Pop + Change +
               Frate + Purban, data = UN3)
summary(un3.lm2)
```

```
##
## Call:
## lm(formula = ModernC ~ log(Fertility) + log(PPgdp) + Pop + Change +
##     Frate + Purban, data = UN3)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -38.506 -12.176   1.907  12.010  29.949
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     6.688e-01  1.255e+01   0.053 0.957573
## log(Fertility) -1.951e+01  6.014e+00  -3.244 0.001531 **
## log(PPgdp)      5.796e+00  1.447e+00   4.007 0.000108 ***
## Pop             2.630e-05  8.574e-06   3.067 0.002680 **
## Change          2.988e+00  2.444e+00   1.222 0.224024
## Frate           1.770e-01  8.217e-02   2.154 0.033306 *
## Purban          4.388e-02  1.058e-01   0.415 0.679094
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.31 on 118 degrees of freedom
##   (85 observations deleted due to missingness)
## Multiple R-squared:  0.576,  Adjusted R-squared:  0.5544
## F-statistic: 26.72 on 6 and 118 DF,  p-value: < 2.2e-16
```

```
par(mfrow = c(2,2))
plot(un3.lm2)
```

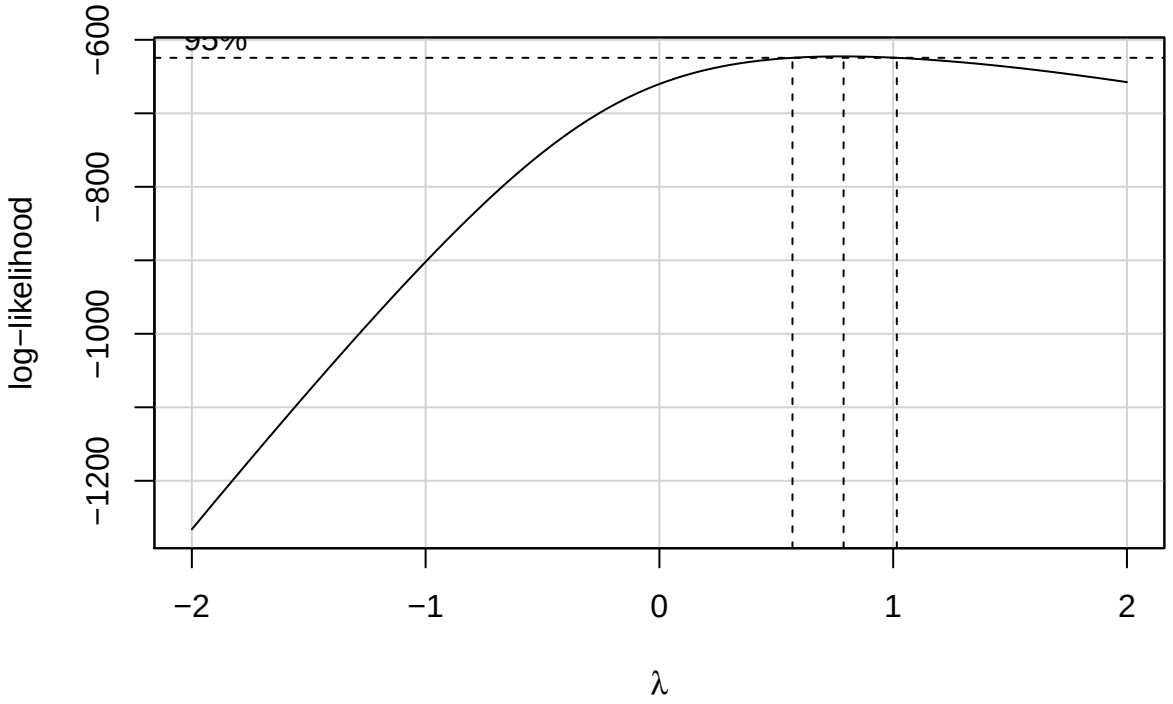



The regression of ModernC by the transformed regressor set better satisfies the assumptions, so the residual plots are more reliable. The standard residual by fitted plot shows residuals with no patterns present, supporting the linearity and constant variance assumptions. The Normal Q-Q plot of the residuals shows

that the residuals likely follow a distribuiton with lighter tails than the normal, but there is no apparent skew. None of the points show up as major outliers that do not follow the end of the general data. The Scale-Location plot shows another patternless cloud of data, supporting the assumptions of linearity and constant variance, and while the spline shows some curvature, it does not appear to be significant and centers on the edges of the point cloud, where the spline has the highest uncertainty. The Residuals vs Leverage plot shows that none of the data in this model have a Cook's distance higher than 0.5, and that while India and China continue to have high leverage, their standardized residuals are not large.

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?

```
boxCox(un3.na.o$ModernC ~ un3.na.o$Fertility + un3.na.o$Purban +
         un3.na.o$PPgdp + un3.na.o$Frate + un3.na.o$Pop + un3.na.o$Change)
```




The value 1 falls on the 95% confidence interval for the optimal value of lambda, so at the standard significance level of 0.05 we would not reject the null hypothesis that no transformation is necessary. Because there is then no transformation applied to the response the optimal transformations of the predictors are the same transformations as done in the previous sections, so the final model is the same as in part 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.

```
## Hat (hii) values

##     India     China
## 0.3793581 0.5592546
```

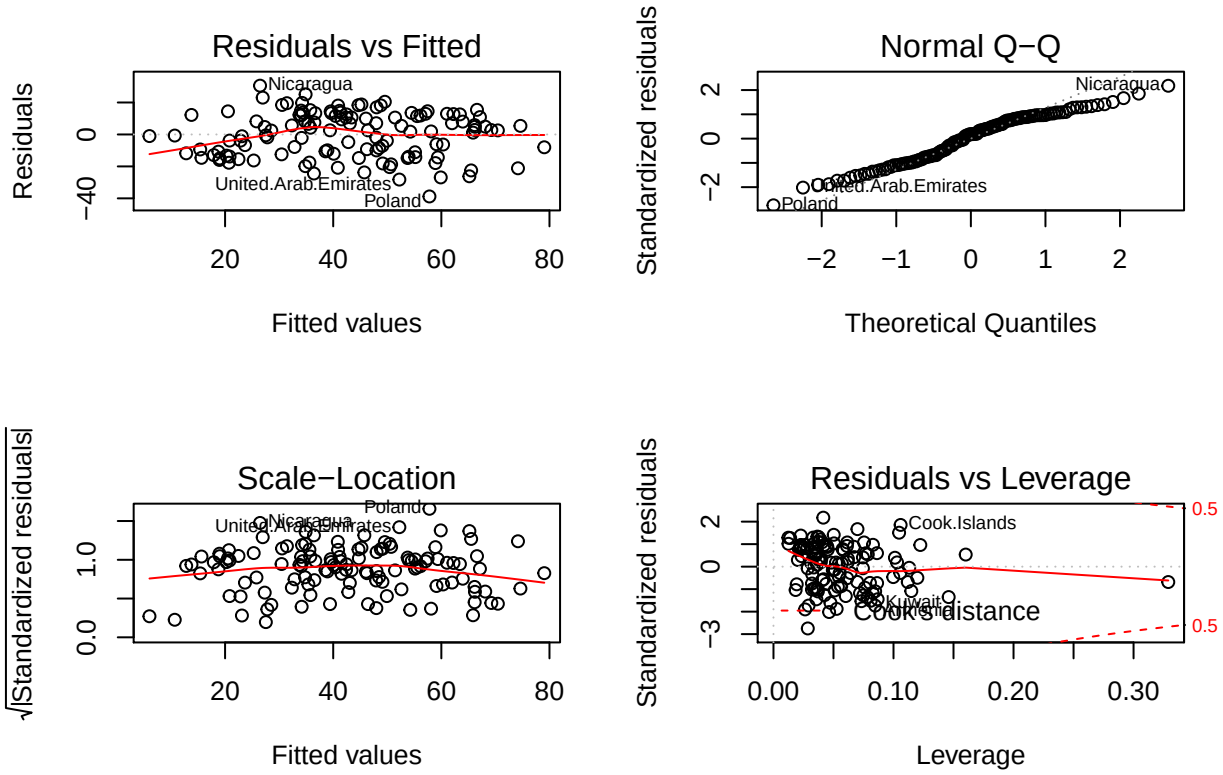
```
## Respective Pop values

## [1] 1065462 1304196
```

India and China have the largest two hat values in this models, both of which are large enough to qualify those points as very influential. They are influential because they are outliers in the Pop variable, with population values larger than one billion. In both of the previous Residuals vs Leverage plots they have stood out from the rest of the data, so performing the analysis with those values excluded is worthwhile to gauge their effect on the model. Below is the summary table for the model with the same formula as the chosen transformation based model, but with the rows corresponding to India and China removed from the dataset.

```
##
## Call:
## lm(formula = ModernC ~ log(Fertility) + log(PPgdp) + Pop + Change +
##     Frate + Purban, data = UN3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -38.506 -12.176   1.907  12.010  29.949
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     6.688e-01  1.255e+01   0.053 0.957573
## log(Fertility) -1.951e+01  6.014e+00  -3.244 0.001531 **
## log(PPgdp)      5.796e+00  1.447e+00   4.007 0.000108 ***
## Pop             2.630e-05  8.574e-06   3.067 0.002680 **
## Change          2.988e+00  2.444e+00   1.222 0.224024
## Frate           1.770e-01  8.217e-02   2.154 0.033306 *
## Purban          4.388e-02  1.058e-01   0.415 0.679094
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.31 on 118 degrees of freedom
##   (85 observations deleted due to missingness)
## Multiple R-squared:  0.576,  Adjusted R-squared:  0.5544
## F-statistic: 26.72 on 6 and 118 DF,  p-value: < 2.2e-16
```

We can see the effect of these outliers by comparing the model summaries, and in this case there are clear differences. The standard error for the coefficient for the Pop predictor has been reduced in magnitude from the previous model and Pop's corresponding p value (0.0527489) is no longer significant at the 0.05 significance level. The origin of the high influence values of the India and China data was primarily in their very large populations, so to see the coefficient for Pop change from a significant effect with those rows included to a non-significant effect when those rows, which were very few in number compared with the size of the full data, are removed implies that Pop may not contain much information about ModernC. The diagnostic plots for the model without india and china are shown below.




The residual plots, with the exception of the Residuals vs Leverage plot, look almost exactly the same as the residual plots generated for the model fit without excluding India and China. The Residuals vs Leverage plot now does not include the two highest leverage values that stood out in the previous iteration of the plot, but still looks similar because another point now appears to be high leverage in comparison to the rest of the data.

## 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!

| | Estimate | Std. Error | t value | Pr(>\|t\|) | CI Lower Bound | CI Upper Bound |
|---|---|---|---|---|---|---|
| (Intercept) | 0.6688427 | 12.5456906 | 0.0533125 | 0.9575730 | -24.1750401 | 25.5127256 |
| log(Fertility) | -19.5105909 | 6.0135031 | -3.2444634 | 0.0015311 | -31.4189642 | -7.6022176 |
| log(PPgdp) | 5.7964240 | 1.4465926 | 4.0069499 | 0.0001081 | 2.9317769 | 8.6610711 |
| Pop | 0.0000263 | 0.0000086 | 3.0672047 | 0.0026799 | 0.0000093 | 0.0000433 |
| Change | 2.9879320 | 2.4444814 | 1.2223173 | 0.2240236 | -1.8528067 | 7.8286708 |
| Frate | 0.1769695 | 0.0821741 | 2.1535918 | 0.0333064 | 0.0142424 | 0.3396967 |
| Purban | 0.0438777 | 0.1057985 | 0.4147287 | 0.6790935 | -0.1656321 | 0.2533874 |

Interpretations:

All interpretations are provided for the case when the variable of interest is changed but all others are held constant.

Fertility: For a doubling of the fertility value we expect ModernC to reduce in value by about 13.5 points

PPgpd: For a doubling of the PPgpd value we expect ModernC to increase in value by about 4 points

Pop: For an increase of 38000 people in population we expect ModernC to increase by about 1 point

Change: For a 1 point increase in Change we expect ModernC to increase by about 3 points

Frate: For a 5 point increase in Frate we expect ModernC to increase by about 1 point

Purban: For a 23 point increase in Purban we expect ModernC to increase by about 1 point

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

```
##
## Call:
## lm(formula = ModernC ~ log(Fertility) + log(PPgdp) + Pop + Change +
##     Frate + Purban, data = UN3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -38.506 -12.176   1.907  12.010  29.949
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     6.688e-01  1.255e+01   0.053 0.957573
## log(Fertility) -1.951e+01  6.014e+00  -3.244 0.001531 **
## log(PPgdp)      5.796e+00  1.447e+00   4.007 0.000108 ***
## Pop             2.630e-05  8.574e-06   3.067 0.002680 **
## Change          2.988e+00  2.444e+00   1.222 0.224024
## Frate           1.770e-01  8.217e-02   2.154 0.033306 *
## Purban          4.388e-02  1.058e-01   0.415 0.679094
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.31 on 118 degrees of freedom
##   (85 observations deleted due to missingness)
## Multiple R-squared:  0.576,  Adjusted R-squared:  0.5544
## F-statistic: 26.72 on 6 and 118 DF,  p-value: < 2.2e-16
```

The model was computed using log transformations on Fertility and PPgdp in order to satisfy the linearity assumption and no cases were excluded in the model fit. China and India were not excluded, despite being heavily influential points, because their standardized residuals were small and they belong to the population of interest, so excluding them does not seem valid. By considering the terms with significant effect sizes, this model tells us that places with increased fertility rates also tend to have lower rates of unmarried women using a modern contraceptive. In addition, this tells us that places with higher GDP per capita tend to have higer rates of modern contraceptive usage in unmarried women. Largely due to the influence of India and China, this model indicates that places with a larger population also tend to have higher rates of usage of modern contraceptives among unmarried women. Lastly, the Frate term is marginally significant, so with a somewhat loose significance level of 0.05 we can conclude that places with higher rates of economically active women above the age of 15 tend to have higher rates of unmarried women using modern conntraceptives. That last conclusion may not hold at more conservative significance levels such as 0.005.

## 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.*

In bivariate regression with quadratic loss we are guarenteed that the regression line will pass through the centriod of the data. The added variable plot relates the sets of residuals from two quadratic loss regressions. The simple least squares regression line on this plot must then pass though the centriod of the residuals. Therefore $\bar{e}_a = \hat{\beta}_0 + \hat{\beta}_1 \bar{e}_b$ where $e_a$ and $e_b$ are residuals from the regression of the other regressors on the response and the residuals from the regression of the other regressors on the regressor of interest respectively. For a generic set of residuals with response $\mathbf{Y}$ and hat matrix $\mathbf{H}$ in quadratic loss regression we know:

$$\bar{e} = \frac{1}{n}\mathbf{1}'\mathbf{e} = \frac{1}{n}\mathbf{1}'(\mathbf{Y}-\mathbf{H}\mathbf{Y}) = \frac{1}{n}\mathbf{1}'(\mathbf{I}-\mathbf{H})\mathbf{Y} = 0$$

Therefore

$$\bar{e}_a = \hat{\beta}_0 + \hat{\beta}_1 \bar{e}_b \implies 0 = \hat{\beta}_0 + \hat{\beta}_1 0 \implies 0 = \hat{\beta}_0$$

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.

```
e_Y = residuals(lm(ModernC ~ log(PPgdp) + Pop + Change + Frate + Purban, data = un3.na.o))
e_X = residuals(lm(log(Fertility) ~ log(PPgdp) + Pop +
                     Change + Frate + Purban, data = un3.na.o))
lm(e_Y ~ e_X)[['coefficients']][2]
```

```
##       e_X
## -19.51059
```

```
un3.lm2[['coefficients']][2]
```

```
## log(Fertility)
##      -19.51059
```

The slope values are the same.