HW2 STA521 Fall18

Michael Valancius, mfv7, mvalancius

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 qualitative?

```r
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
```

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

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

ModernC, Change, PPgdp, Frate, Pop, Fertility, and Purban are all quantitative variables. Unfortunately, all of the variables with the exception of Purban contain missing values (NA). As noted in the details of the UN3 data set, missing values are more predominant in less developed countries. Thus, there is a bias in the data, and conclusions drawn from the analysis below should be limited to the countries (or perhaps similar countries) with full data in the data set.

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

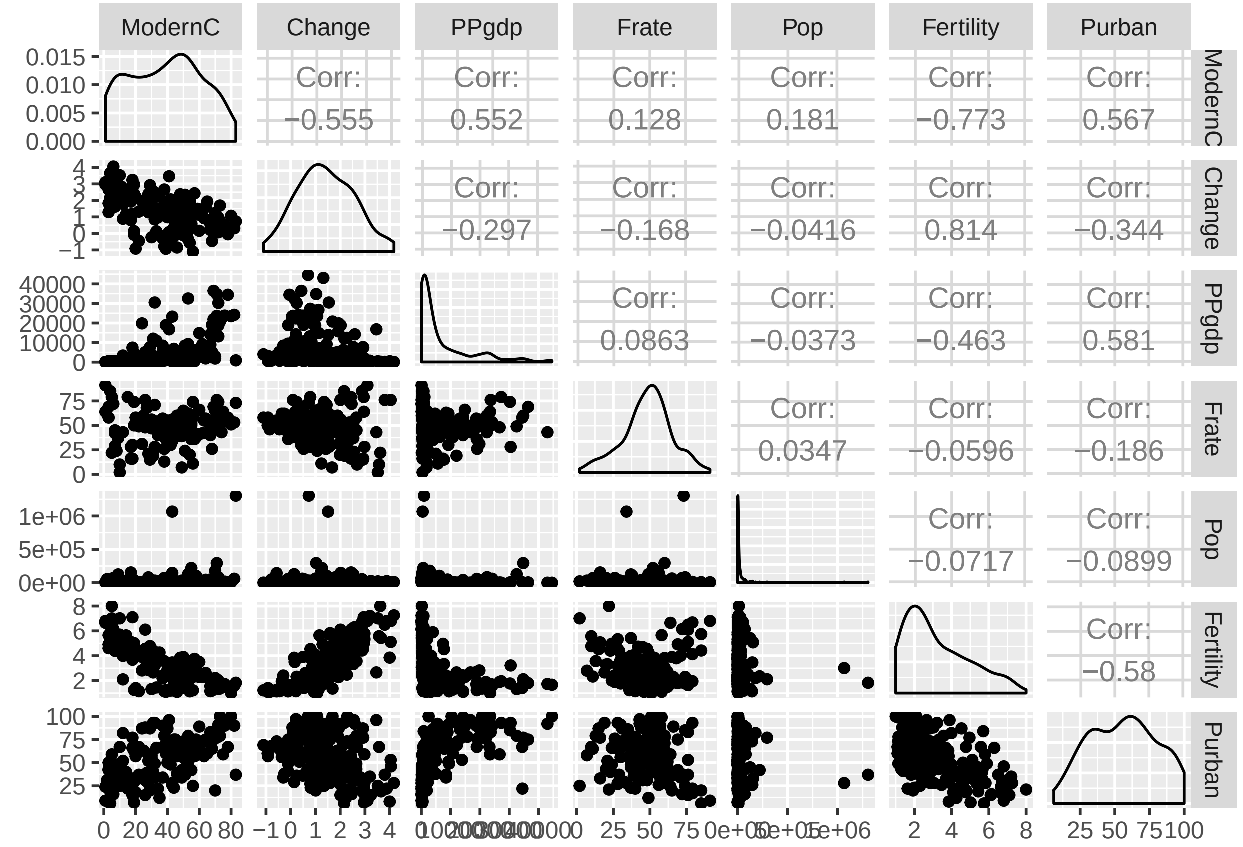
```r
UN3_means <- summarise_all(UN3, mean, na.rm = TRUE)
UN3_sds <- summarise_all(UN3, sd, na.rm = TRUE)
rbind(UN3_means, UN3_sds) %>% t() %>% apply(., MARGIN = c(1, 2), function(x) round(x,
    2)) %>% kable(, col.names = c("Mean", "Standard Deviation"), align = "l")
```

	Mean	Standard Deviation

	Mean	Standard Deviation
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 scatter plots 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
ggpairs(UN3, progress = FALSE)
```



In order to plot the pairwise comparisons of the variables, the ggpairs function automatically removed data that had missing values for either of the two variables in each plot.

From the above plots, we notice several things.

- The strongest relationship in the data, with a correlation of 0.814, is between fertility and change. Higher rates of population growth are associated with higher fertility rates. This is expected since birth rate (fertility) is a major driver of population growth.

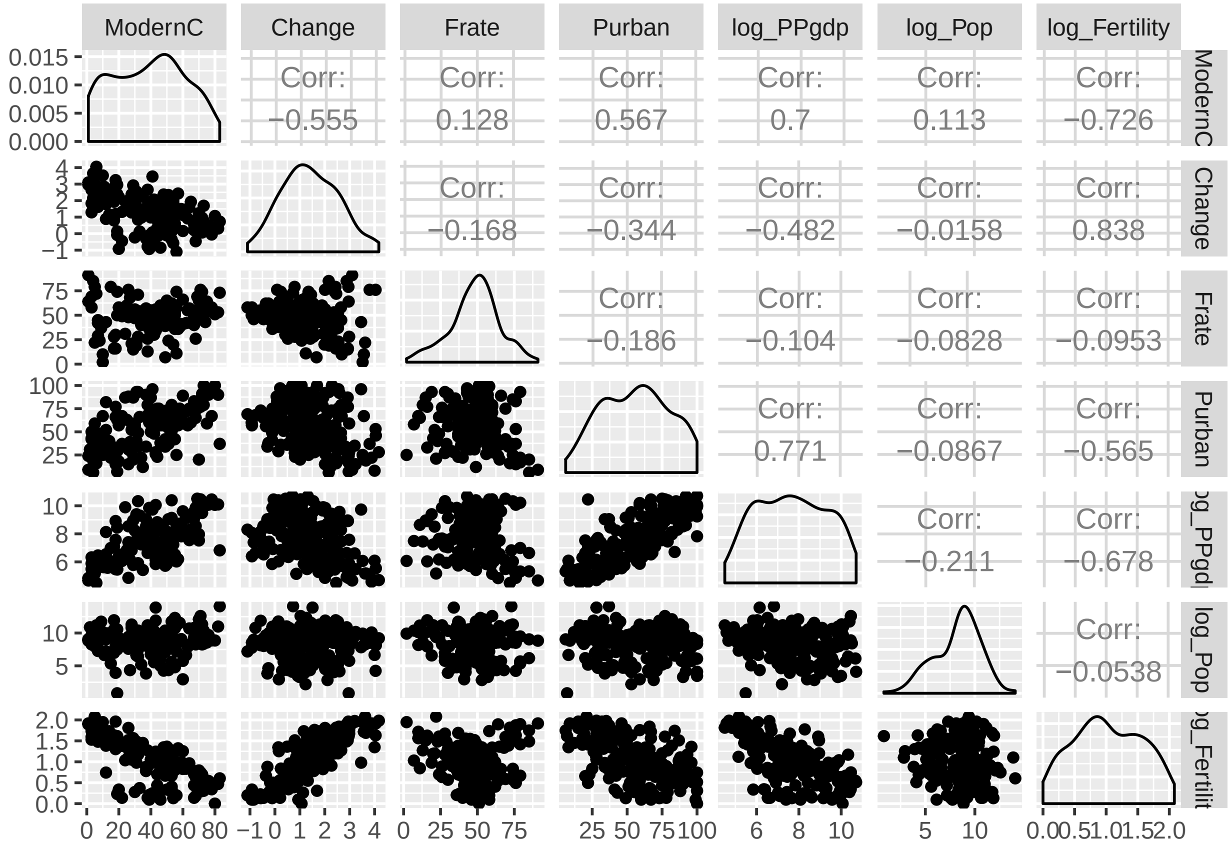
- Our response variable of interest, ModernC, has approximately linear relationships with Change and Purban individually. Meanwhile, the plot between fertility and ModernC shows some curvature. Without controlling for other variables, areas with lower rates of population growth have higher levels of modern contraception use. Knowing that areas of high growth also generally tend to have lower wealth/education could lead to speculation that the relationship between Change and PPgdp is important in predicting ModernC.

- The relationship of ModernC with PPgdp appears to be better modeled with a curve than a straight line, suggesting a logarithmic or power transformation might better approximate a linear relationship. Additionally, the relationship of ModernC with population is unclear do to some potential outliers. From the plot of population with itself, we see that it exhibits a heavy positive skew, with several countries (especially China and India) having much larger populations that the median. Consequentially, a logarithmic transformation of population could bring these orders of magnitude into scale. Only removing these values would lose important information. Since population is not evenly distributed around its mode, removing large values glosses over the effect of larger countries, whose greater magnitude is better expressed in terms of multiplication than addition.

In addition to checking for a linear relationship between the response and the individual predictors, in a multivariate setting, a good general procedure is to transform the predictors so that they exhibit a close to linear relationship. Many of the predictors plotted against each other have relationships that do not appear linear, and thus utilizing transformations is appropriate. We will further investigate the predictor relationships after applying the transformations.

While the appropriate transformations will be discussed further in a later question, I wanted to plot the pairs after applying a log transformation to PPgdp, Fertility, and Population. Justification for these transformations is graphical. Population has a distribution that spans many orders of magnitude, while PPgdp and Fertility span several orders of magnitude and have non linear relationships with ModernC based on the plotted samples.

```
UN3 %>% mutate(log_PPgdp = log(PPgdp), log_Pop = log(Pop), log_Fertility = log(Fertility)) %>%
    select(-Pop, -PPgdp, -Fertility) %>% ggpairs(, progress = FALSE)
```



With these transformations, most of the relationships appear more linear. All of the variables, except for log(Pop), individually have a relationship with MordernC. It is worth noting that, on its own, Frate does not have much of a relationship with any of the other predictors individually.

Additionally, many of the plots of predictors on predictors now demonstrate a closer to linear relationship, suggesting that further transformations may be unnecessary.

While not all the points fit very closely along a straight line, the plots do not raise any immediate concerns about outliers. Transforming population using logarithms helped account for the huge populations of China and India.

Model Fitting

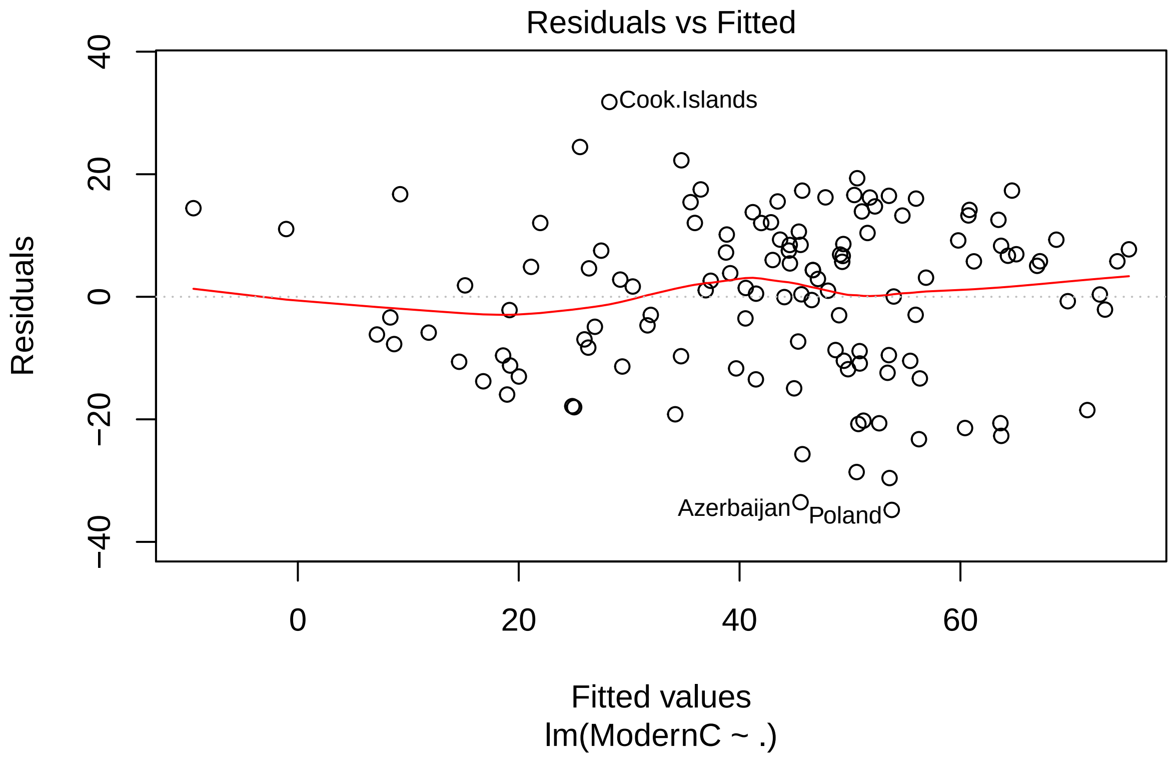
4. Use the `lm()` function to perform a multiple linear regression with `ModernC` as the response and all other variables as the predictors, using the formula `ModernC ~ .`, where the `.` includes all remaining variables in the data frame. Create diagnostic residual plot from the linear model object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

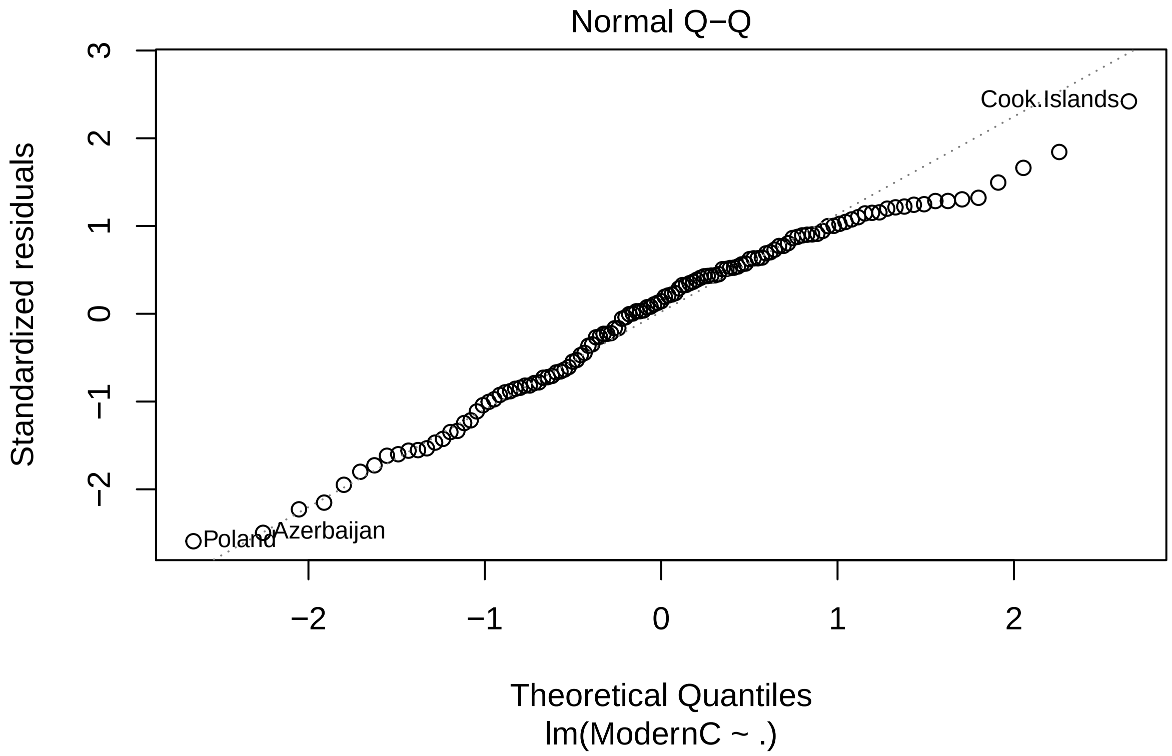
```r
UN3_mod_simple <- lm(ModernC ~., data = UN3)
summary(UN3_mod_simple)
```

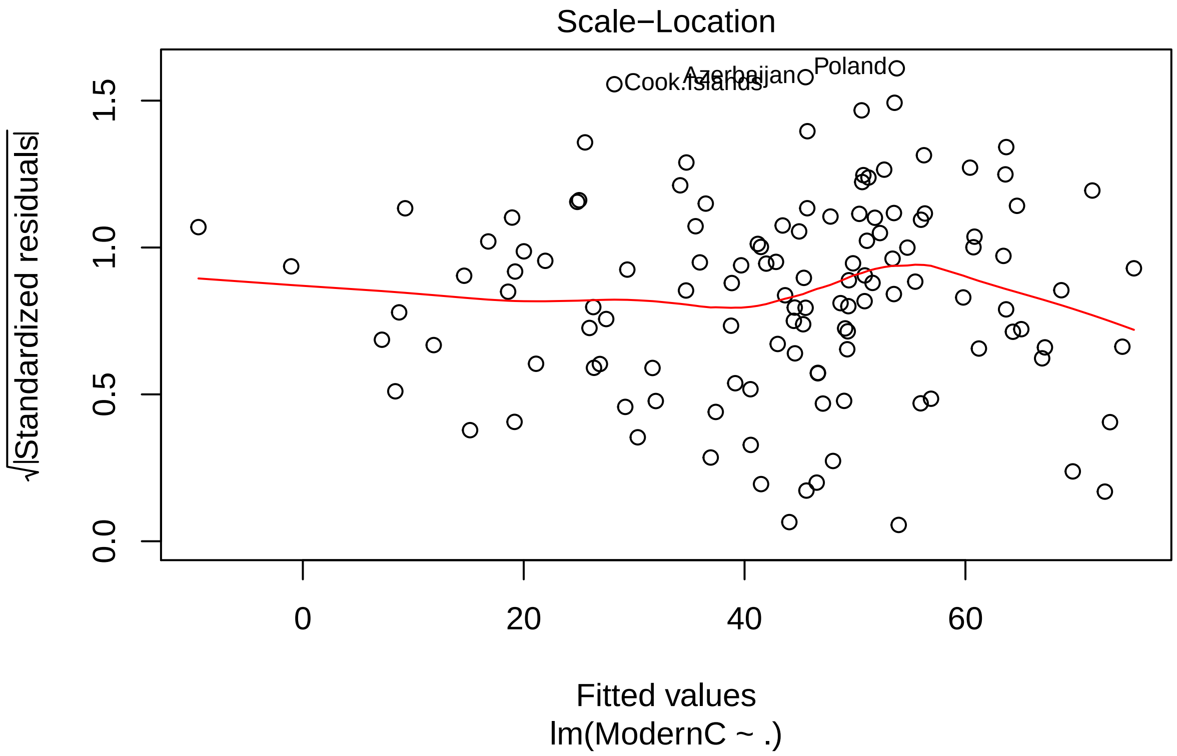
```
##
## Call:
## lm(formula = ModernC ~ ., data = UN3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -34.781  -9.698   1.858   9.327  31.791
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.529e+01  9.467e+00   5.841 4.69e-08 ***
## Change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

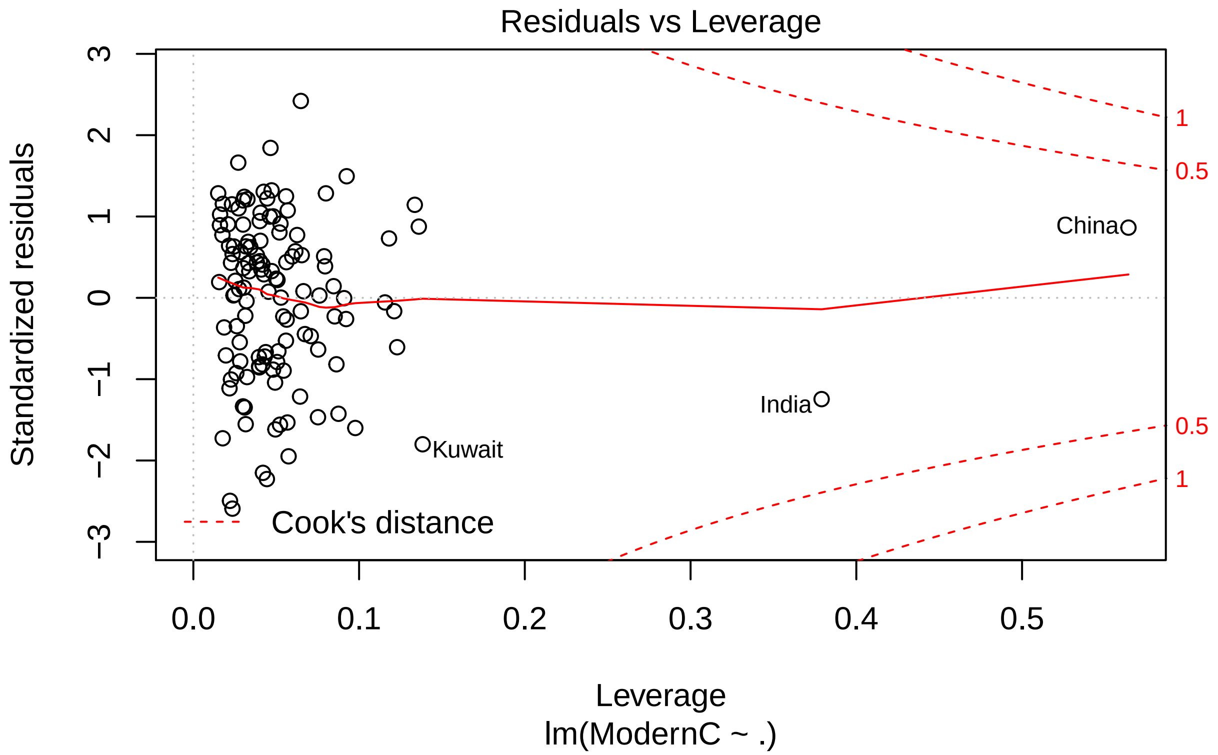
Because the lm function in R removes observations with missing data as a default, there were 85 observations that were deleted. Considering there were only 210 observations to start with, we are only left with 125 (60%) of the observations.

```
plot(UN3_mod_simple)
```




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


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

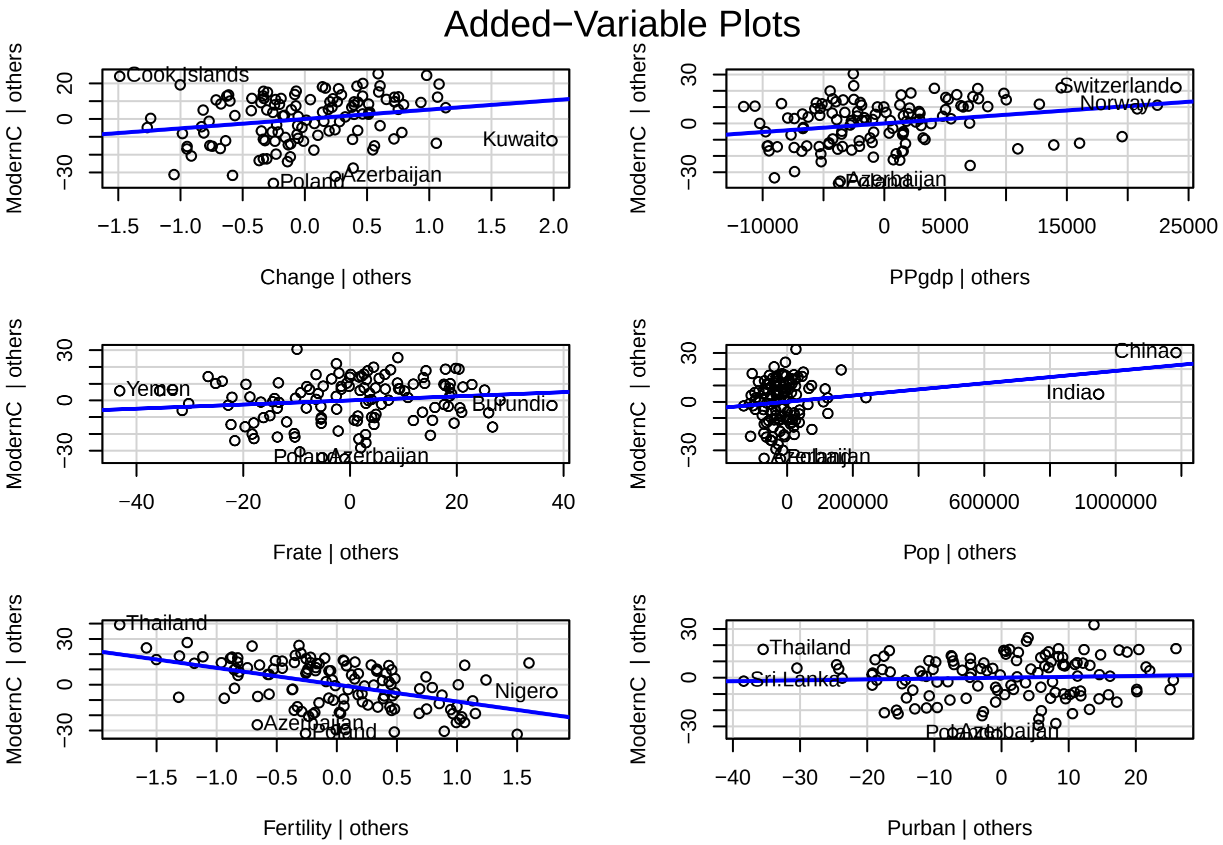


In a multiple linear regression setting, interpretations of residual plots are dependent upon the presence of a somewhat linear relationship among the predictors. Since this model was fit without any transformations, that assumption is questionable, as noted in question 3.. So even though the residuals display a roughly constant variance with no clear patterns, interpreting the residuals vs fitted plots is inconclusive. Similarly, even though the pattern in the Q-Q plot indicates that the residuals do not follow a theoretical normal distribution, we need to transform the variables and re-run the model to make stronger conclusions about the fit.

In the Residuals vs Leverage plot, we see that China and India has the largest leverage values, but neither are very influential based on their Cook's Distance. There is also no discernible pattern between leverage and the size of the standardized residual.

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?

```r
avPlots(UN3_mod_simple)
```



Looking at the added variable plots, we can see the amount of variation unexplained by the model plotted against the variation in the added variable that is not explained by the other variables.

Based on the added variable plots, it appears that Purban (slope of around 0) has very little influence on ModernC after the other variables have been accounted for. Meanwhile, fertility has the strongest (albeit negative) relationship with ModernC after accounting for the other variables. Both fertility and PPgdp seem to have some curvature in the relationships, and thus transformations should be investigated.

The most interesting situation involves population, which has a cluster of localities with populations significantly smaller than the largest several countries. Because of the distribution of populations, log transforming this variable might be an appropriate solution. However, it also could be possible that there is no real relationship between population and ModernC after controlling for the other variables and that a few of the countries just represent outliers.

Both India and China have the potential to influence the regression line based on their population level. We can test how much their predicted points change based on whether China and India were included in the training data or not.

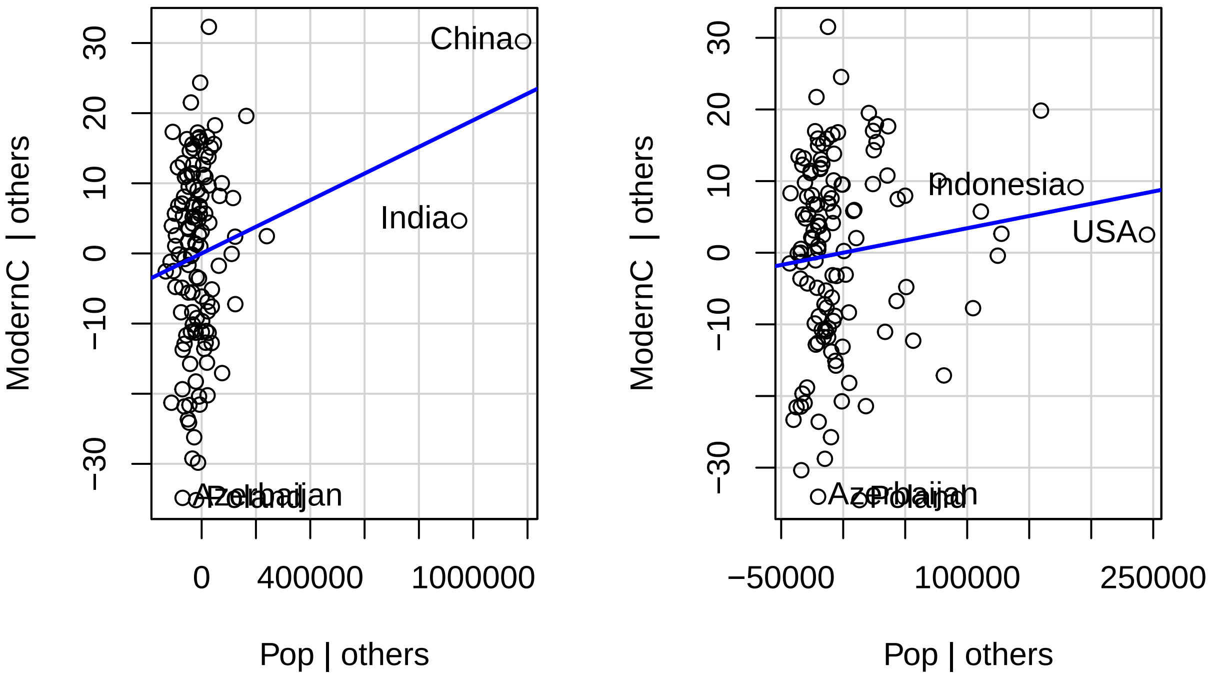
```
UN3_mod_no_chi_ind <- lm(ModernC ~ ., data = UN3[UN3$Pop < 1000000,])
china_india <- UN3[UN3$Pop > 1000000,] %>% filter(is.na(Pop) == "FALSE")
p_wo_chi_ind <- as.vector(predict(UN3_mod_no_chi_ind, newdata = china_india))
p_w_chi_ind <- as.vector(predict(UN3_mod_simple, newdata = china_india))
data.frame("With_China_India" = p_w_chi_ind, "Without_China_inda" = p_wo_chi_ind) %>% t() %>% kable(, c
```

	China	India
With_China_India	75.26000	56.32611
Without_China_inda	94.50408	72.47696

As seen, China and India both have significantly smaller predicted values when the model was trained with their observations. This is evidence that the inclusion of these two observations, which have population values far larger than any other population values, influence the fit of the model.

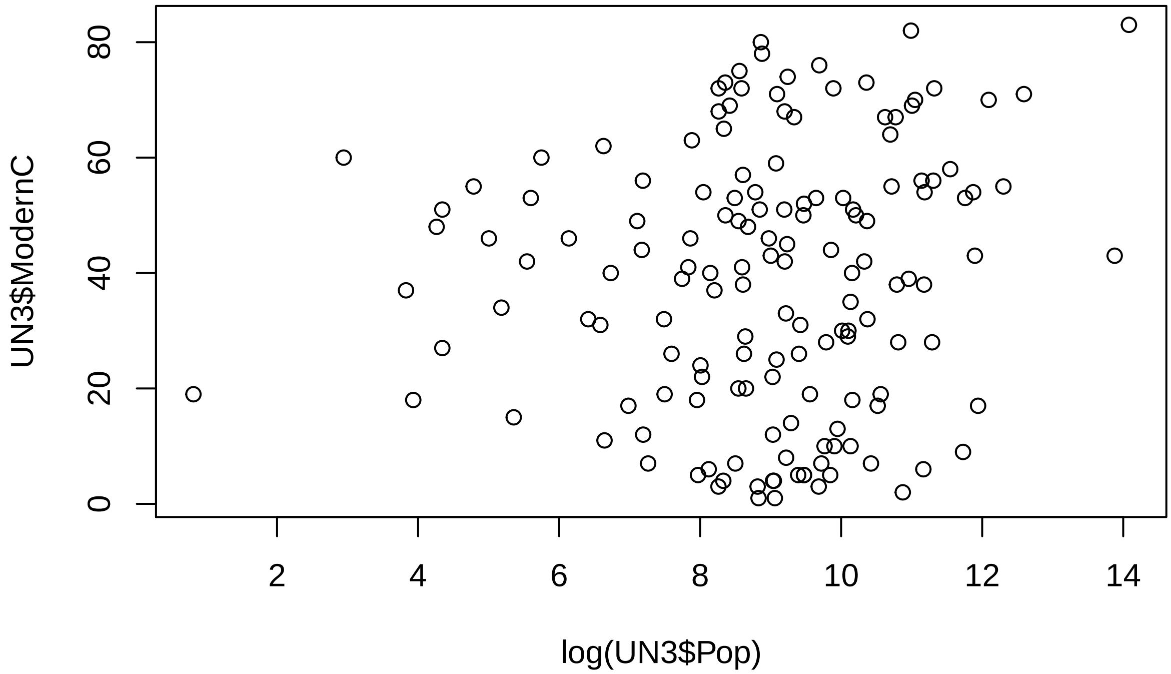
Using added variable plots constructed with and without China and India allows us to get a better glimpse into whether there is a relationship that needs transforming or if there is no relationship and China and India are just outliers.

```r
par(mfrow = c(1,2))
avPlots(UN3_mod_simple, terms = ~Pop, main = "Added Variable Plot with China, India")
avPlots(UN3_mod_no_chi_ind, terms = ~Pop, main = "Added Variable Plot without China, India")
```



The added variable plot on the right seems to suggest that there is some relationship between population and ModernC after accounting for the other variables, but that it needs a transformation to make the relationship linear.

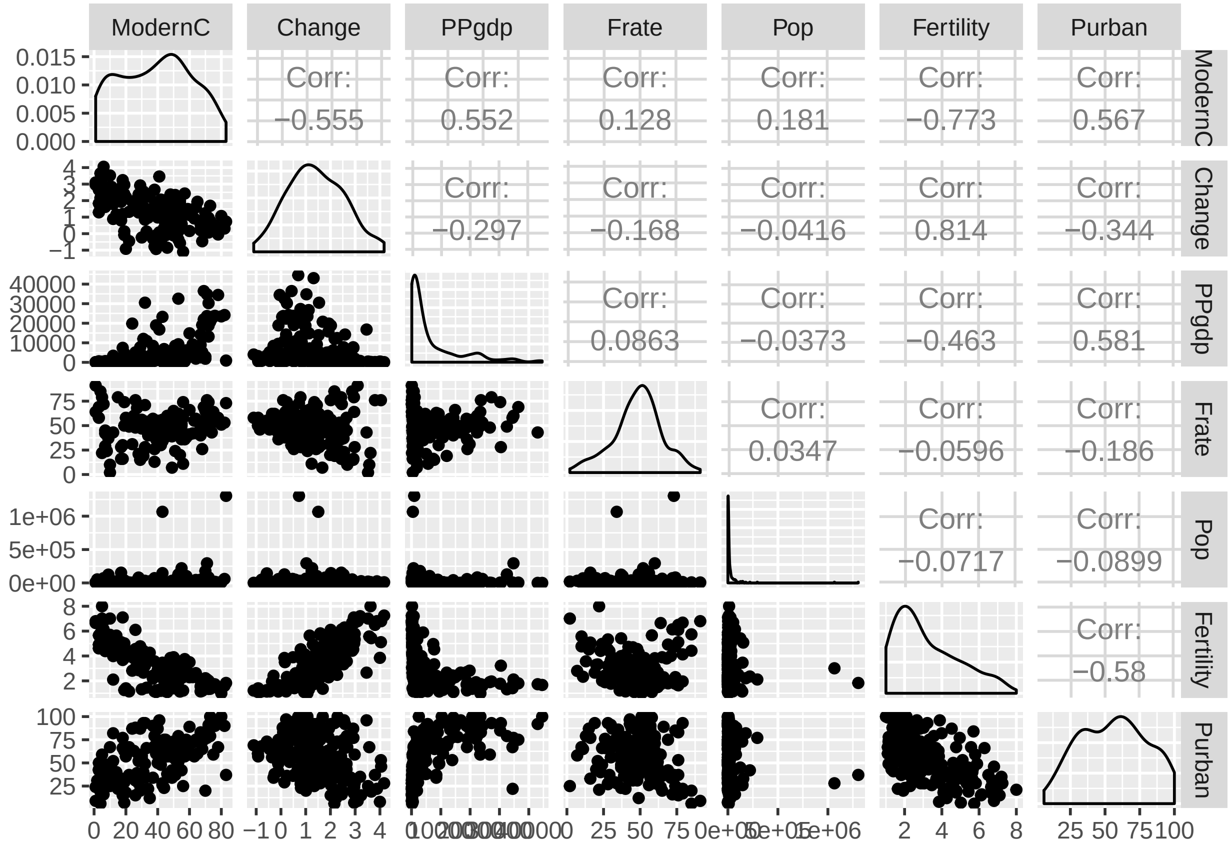
```r
plot(x = log(UN3$Pop), y = UN3$ModernC)
```



As seen in the above plot, although there is not a strong relationship between just ModernC and log(Population), the transformation removes the non linear trend previously exhibited.

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.

```r
ggpairs(UN3, progress = FALSE)
```



Based on their nonlinear relationships with ModernC, the three variables that appear to be in need of transformations are population, fertility, and PPgdp.

(1) Population

Comparing the plots of ModernC against population and then the log of population, we see that the transformation better models the majority of the data, although the relationship is still not very strong. The flexible loess regression line offers a comparison to the linear regression line and demonstrates the stark contrast between the fit of the two plots.

(2) Fertility

Above are the plots of ModernC with Fertility, the log of Fertility, and 1/Fertility. While the log transformation improves the linear relationship the most, there is still some curvature for lower values of fertility. Checking the results of the Box-Tidwell method could produce a more suitable transformation.

(3) PPgdp

Before the transformation, most of the PPgdp values are clustered below 200. However, several are many multiples larger than the smallest. Thus, it is a natural candidate for a logarithmic transformation. The second plot demonstrates that the transformation produces a much closer to linear relationship,

While the three logarithmic transformations appear to resolve the lack of linearity in the predictors from a graphical analysis, it is useful to compare these with the recommended transformations produced by the BoxTidwell method.

```
UN3_shiftChange <- mutate(UN3, Change = Change + 100) %>% na.omit()
boxTidwell(ModernC ~ PPgdp + Fertility + Pop, ~Change + Purban + Frate, data = UN3_shiftChange)
```

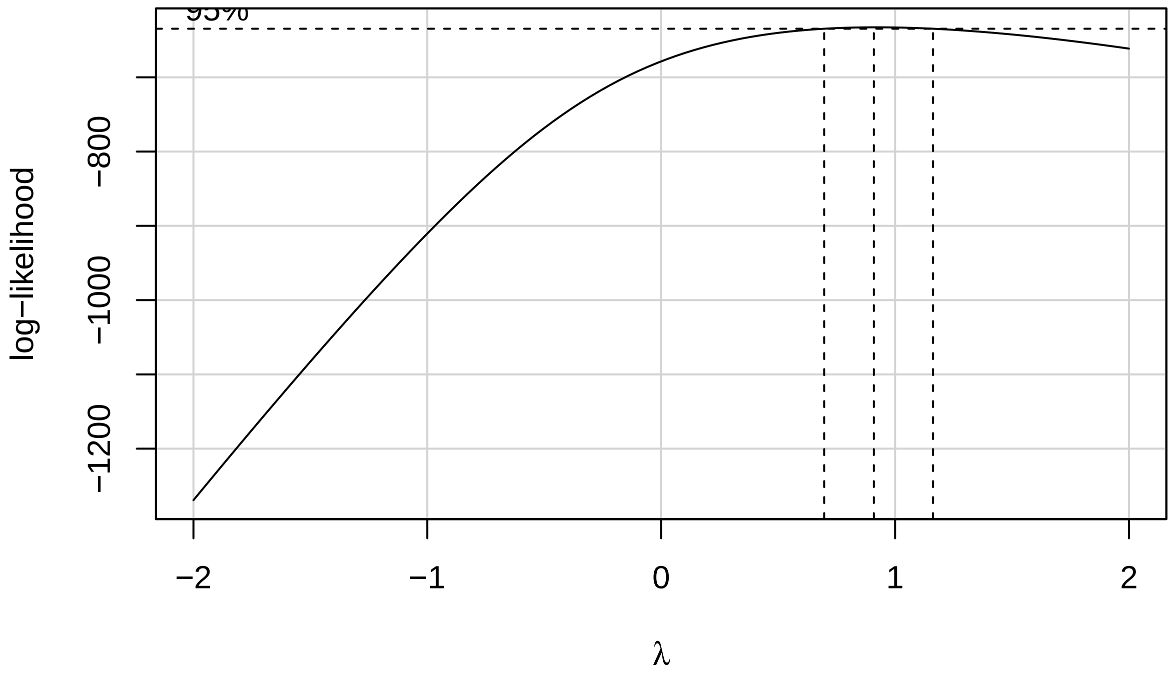
```
##            MLE of lambda Score Statistic (z) Pr(>|z|)
## PPgdp         -0.035767             -1.2324   0.2178
## Fertility      1.346874             -1.7985   0.0721 .
## Pop            0.374984             -0.9042   0.3659
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## iterations =  22
```

None of the maximum likelihood estimates for the powers are statistically significant at the 0.05 level. Given that there is no strong statistical or previously known justification for these power transformations and due to the improvement in the linear relationships with the log transformation, the latter transformation is preferred,

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

Since we log transformed Pop, Fertility, and PPgdp, we include them in the BoxCox method.

```
UN3_naomit <- na.omit(UN3)
UN3_mod_tran <- lm(ModernC ~ Purban + Change + Frate + log(Pop) + log(Fertility) +
    log(PPgdp), data = UN3_naomit)
boxCox(UN3_mod_tran)
```



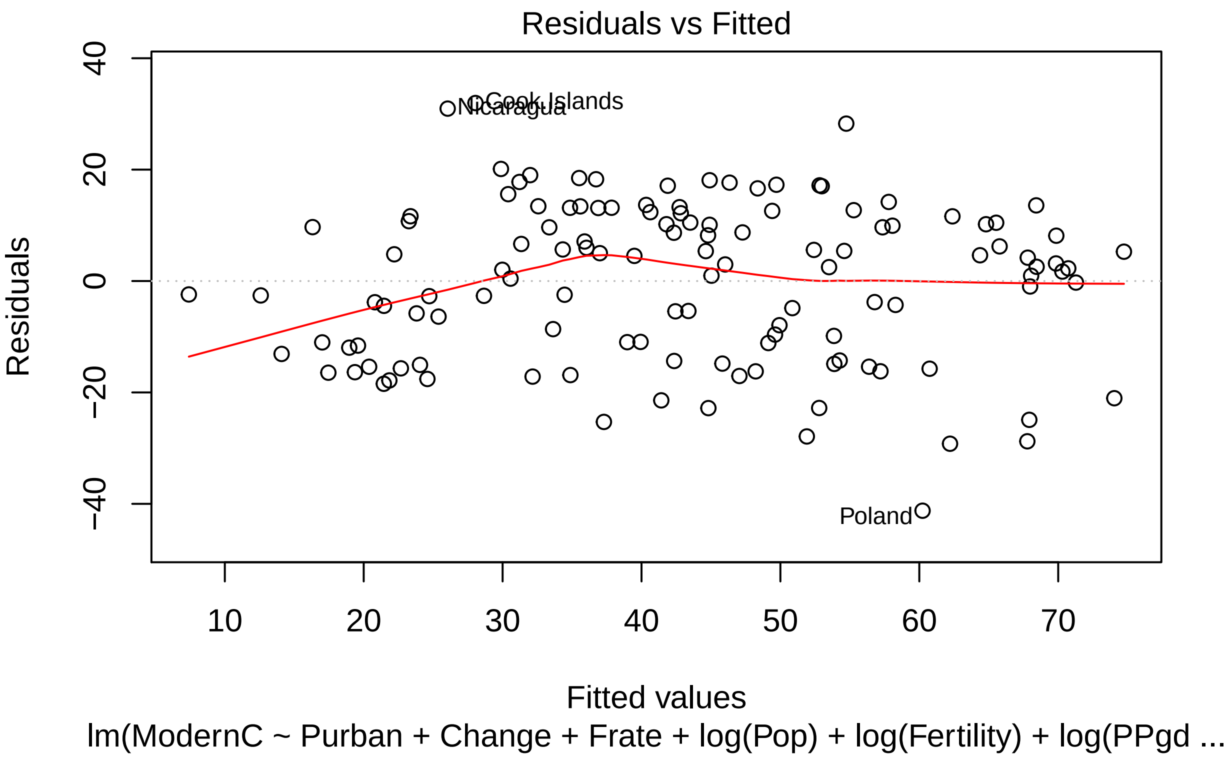
Based on the maximum likelihood estimate for lambda, the power to which the response variable would be raised, we see that the 95% confidence interval contains (and is almost centered at) 1, which would involve no transformation. Thus, we do not transform ModernC.

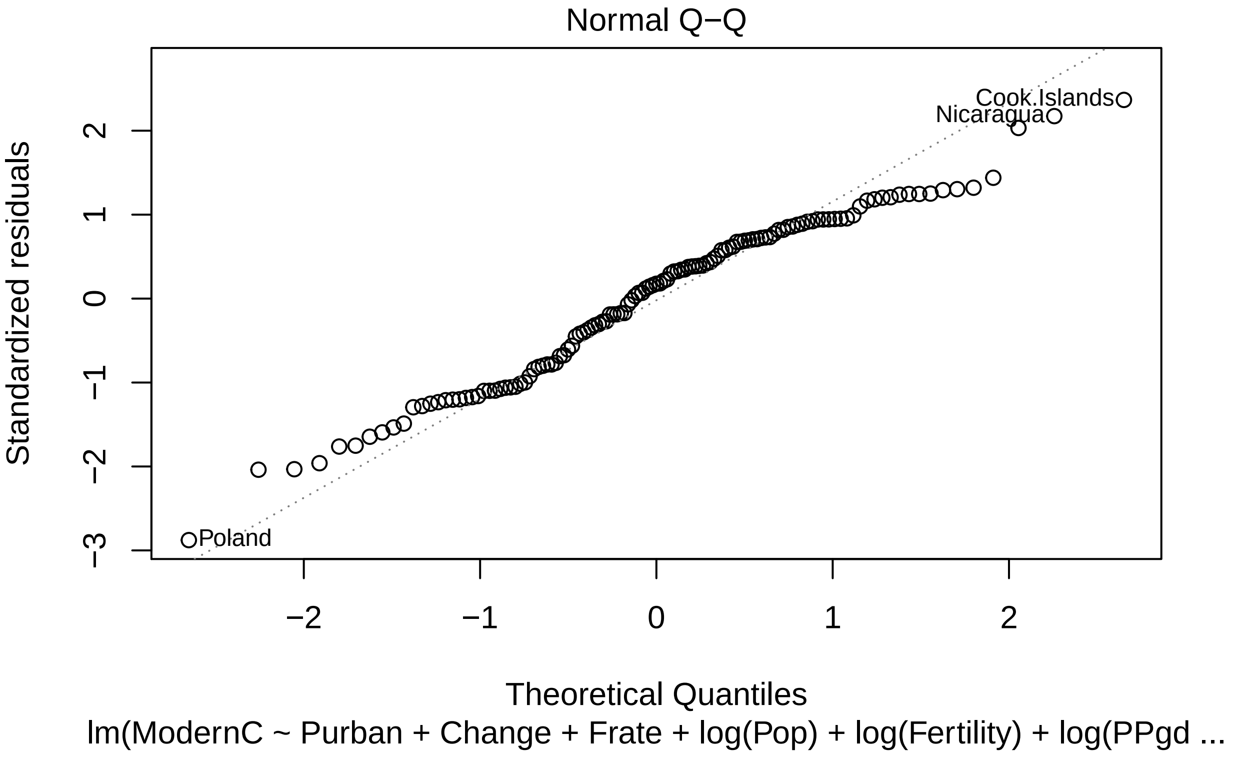
8. Fit the regression using the transformed variables. Provide residual plots and added variables plots and comment. If you feel that you need additional transformations of either the response or predictors, repeat any steps until you feel satisfied.

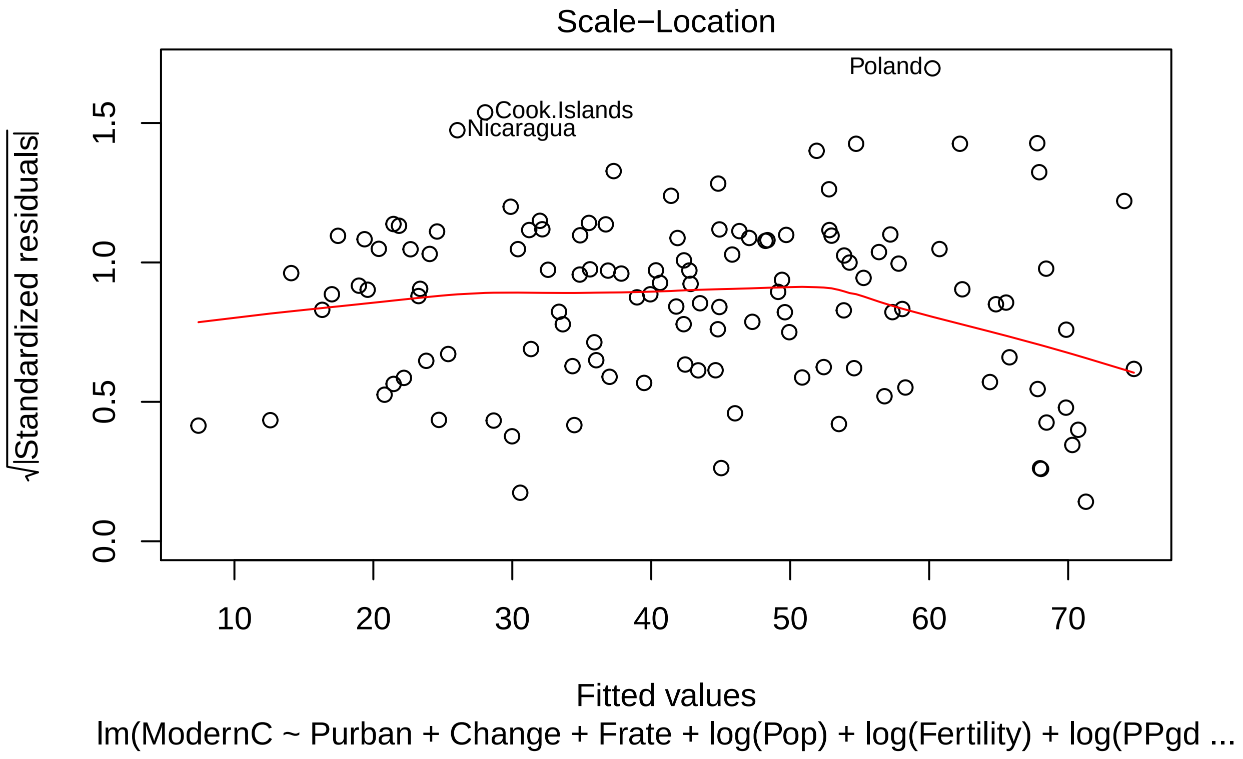
We have the following model:

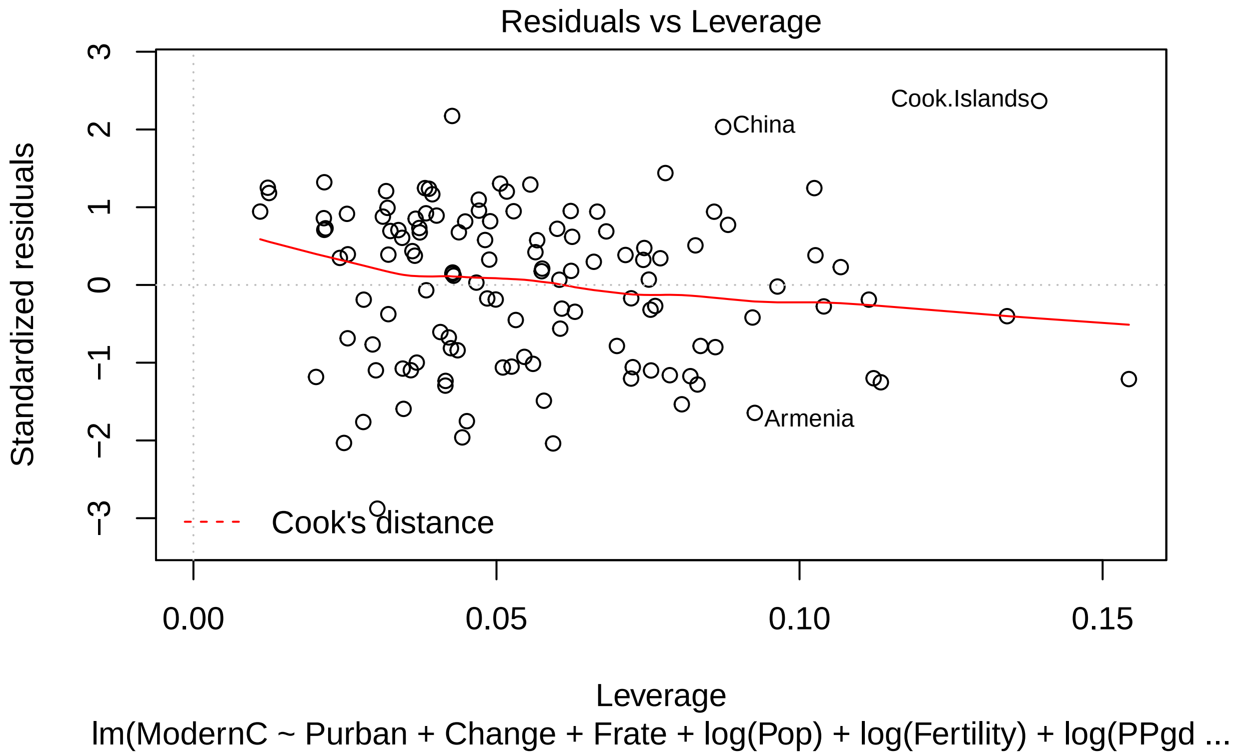
$$ModernC = \beta_0 + \beta_1 log(Pop) + \beta_2 log(Fertility) + \beta_3 log(PPgdp) + \beta_4 Purban + \beta_5 Change + \beta_6 Frate + \epsilon$$

```r
plot(UN3_mod_tran)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(ModernC ~ Purban + Change + Frate + log(Pop) + log(Fertility) + log(PPgd ...
Cook.Islands
Nicaragua
Poland
−3
−2
−1
0
1
2
−2
−1
0
1
2

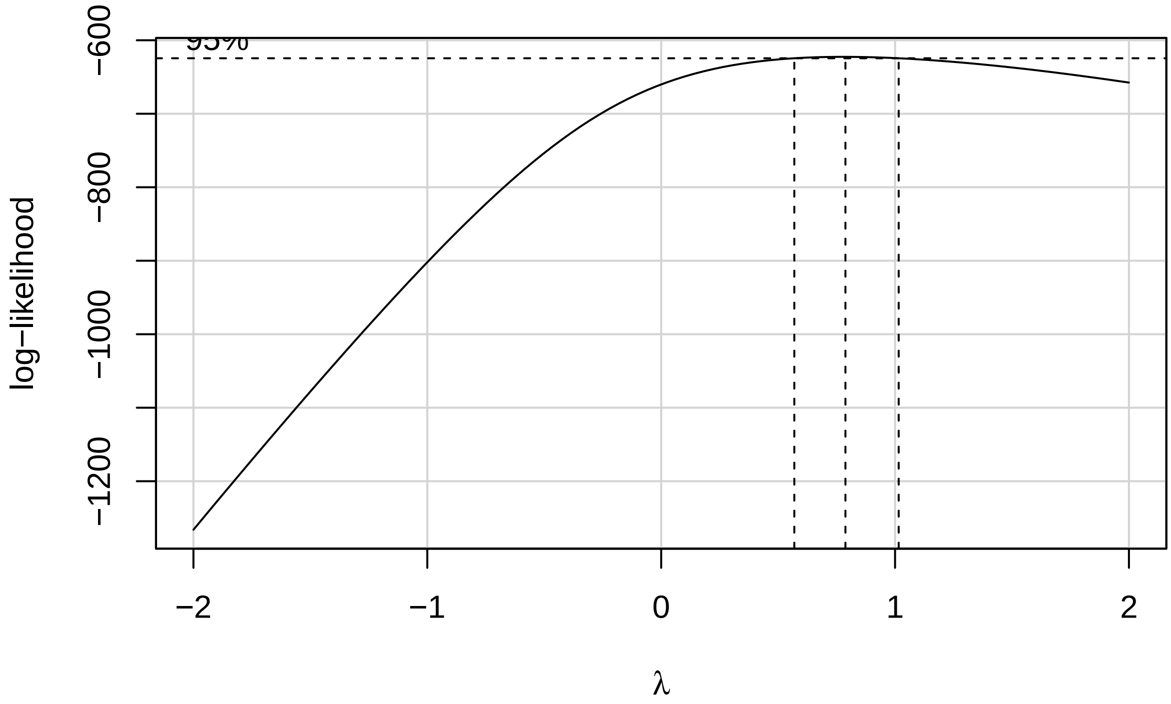

Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
10
20
30
40
50
60
70
Fitted values
lm(ModernC ~ Purban + Change + Frate + log(Pop) + log(Fertility) + log(PPgd ...
Poland
Cook.Islands
Nicaragua



Since the variables with the transformations have approximately linear relationships with each other, we can now analyze the residual plots. In the Fitted Values vs. Standardized Residuals plot, we see no noticeable trend in the residuals and the variance appears constant. However, the Q-Q plot does give some concern that the residuals are not quite normally distributed, with the thin tails indicating that the residuals are more tightly centered around the 0.

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?

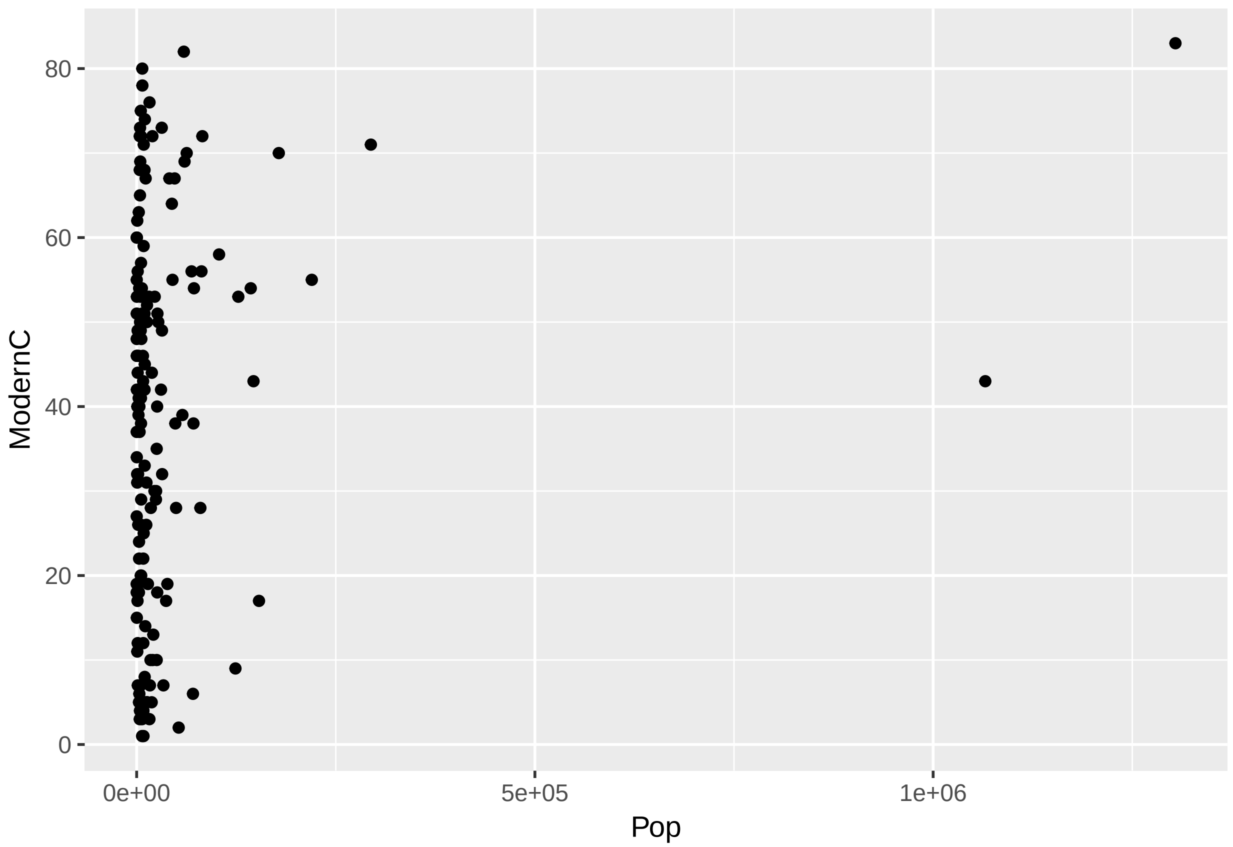
```
boxCox(UN3_mod_simple)
```



Using a model that includes all the predictors with no transformations, we see that the Box-Cox method recommends transforming the response variable to the power of around 0.8. The 95% confidence interval for the estimate contains 1 (no transformation). Because of this and the difficulties in justifying a power of 0.8 theoretically, the optimal procedure is to leave the predictor unchanged. Thus, both when we transformed the variables and when we didn't, the Box cox method provided evidence that a power transformation of the response variable, ModernC, would not be optimal.Thus, the models are the same because we would be repeating the same process as before again.

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.

```
ggplot(UN3, aes(x = Pop, y = ModernC)) + geom_point()
```



In the plot of Population against ModernC, we see that there are two values of population over 150 times larger than the median. These countries are China and India. In general, the population data exhibits a positive skew even when ignoring China and India, as the first plot below shows. This can be seen in the histogram, which shows the population distribution with all localities included.

As mentioned in answer #5, with no transformation applied to Population, the inclusion of China and India influence the model.

```
UN3_outlier <- UN3[UN3$Pop < 1e+06, ]
## Model with no transformations and with outliers
summary(UN3_mod_simple)
```

```
##
## 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
```

```
## Model with no transformations and without outliers
summary(UN3_mod_no_chi_ind)
```

```
##
## Call:
## lm(formula = ModernC ~ ., data = UN3[UN3$Pop < 1e+06, ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -34.981  -9.701   1.708   9.564  31.946
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.583e+01  9.602e+00   5.815 5.46e-08 ***
## Change       5.169e+00  2.096e+00   2.467  0.01510 *
## PPgdp        5.284e-04  1.792e-04   2.949  0.00385 **
## Frate        1.104e-01  8.188e-02   1.349  0.18012
## Pop          3.419e-05  2.702e-05   1.266  0.20821
## Fertility   -1.095e+01  1.758e+00  -6.226 7.89e-09 ***
## Purban       4.847e-02  9.344e-02   0.519  0.60494
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.6 on 116 degrees of freedom
##   (85 observations deleted due to missingness)
## Multiple R-squared:  0.6128, Adjusted R-squared:  0.5928
## F-statistic:  30.6 on 6 and 116 DF,  p-value: < 2.2e-16
```

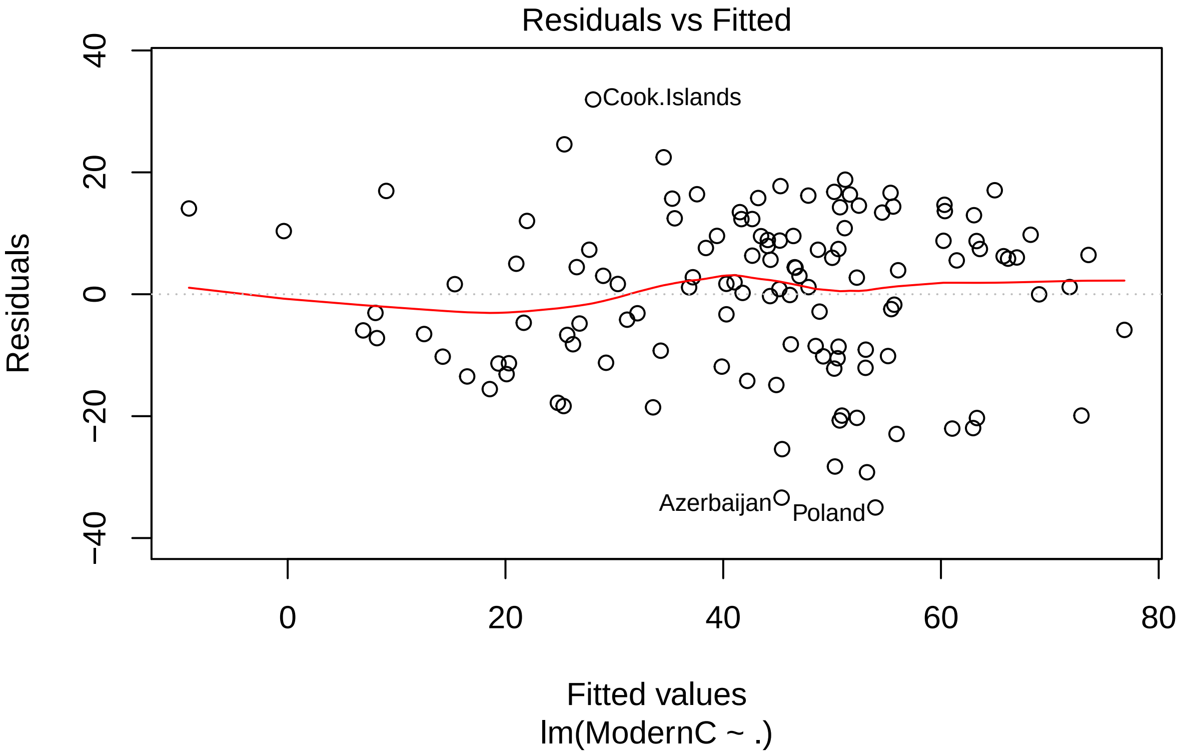
```
## Model with transformations and with outliers.
summary(UN3_mod_tran)
```

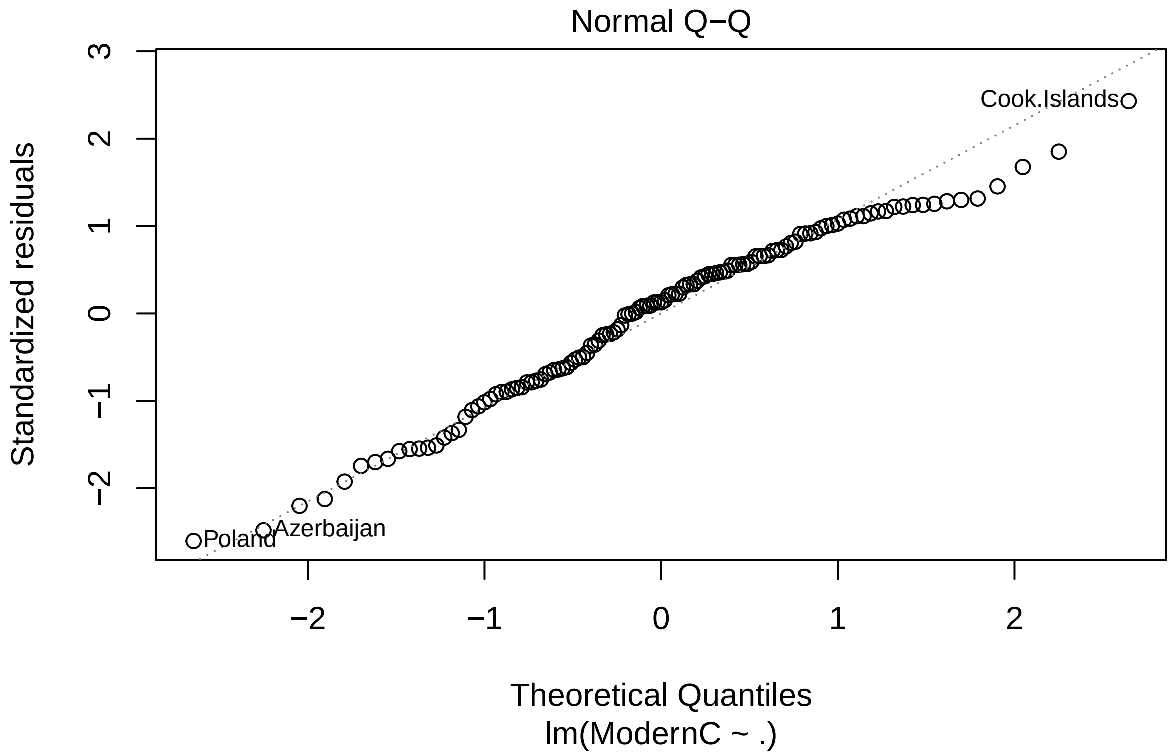
```
##
## Call:
## lm(formula = ModernC ~ Purban + Change + Frate + log(Pop) + log(Fertility) +
##     log(PPgdp), data = UN3_naomit)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -41.235 -11.589   2.498  10.748  31.954
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -15.106490  15.949888  -0.947  0.34551
## Purban          -0.007352   0.106591  -0.069  0.94513
## Change           2.310274   2.560728   0.902  0.36879
## Frate            0.178242   0.083567   2.133  0.03500 *
## log(Pop)          1.595611   0.699289   2.282  0.02430 *
## log(Fertility) -18.237639   6.336680  -2.878  0.00475 **
## log(PPgdp)       6.445713   1.508057   4.274 3.91e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.55 on 118 degrees of freedom
## Multiple R-squared:  0.5615, Adjusted R-squared:  0.5392
## F-statistic: 25.19 on 6 and 118 DF,  p-value: < 2.2e-16
```

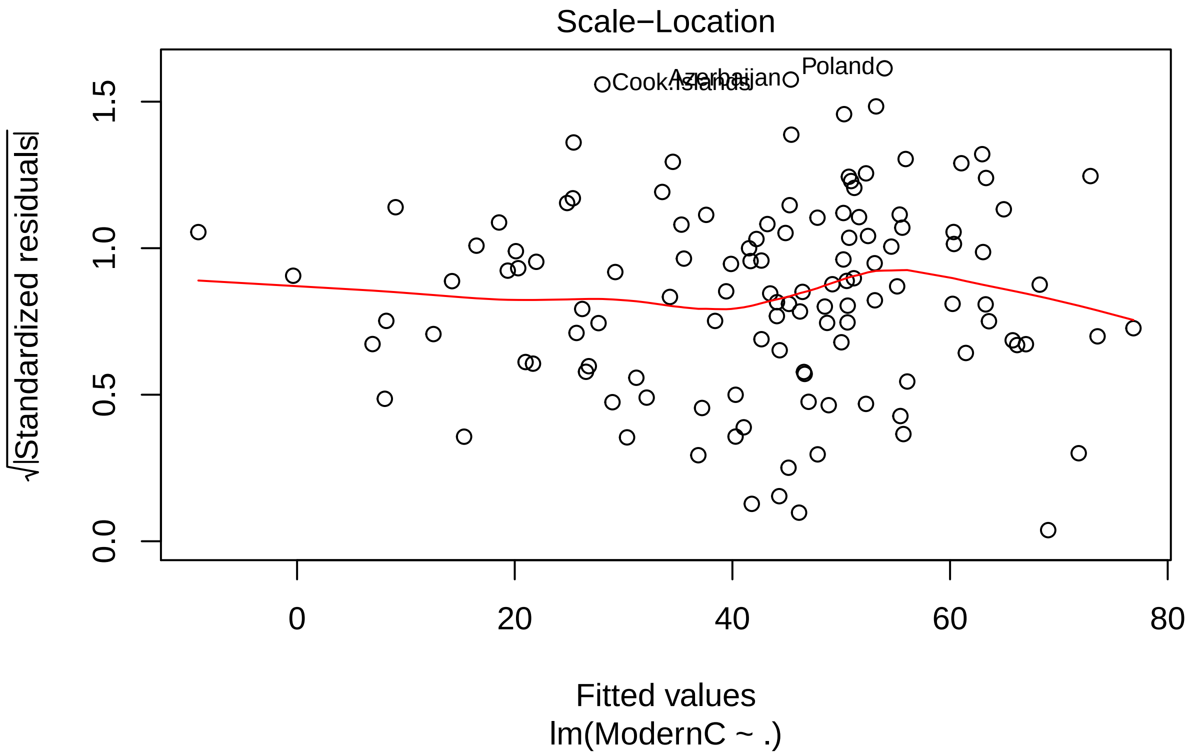
```
## Model with transformations and no outliers.
UN3_mod_no_chi_ind_tran <- lm(ModernC ~ Purban + Change + Frate + log(Pop) +
    log(Fertility) + log(PPgdp), data = UN3[UN3$Pop < 1e+06, ])
summary(UN3_mod_no_chi_ind_tran)
```

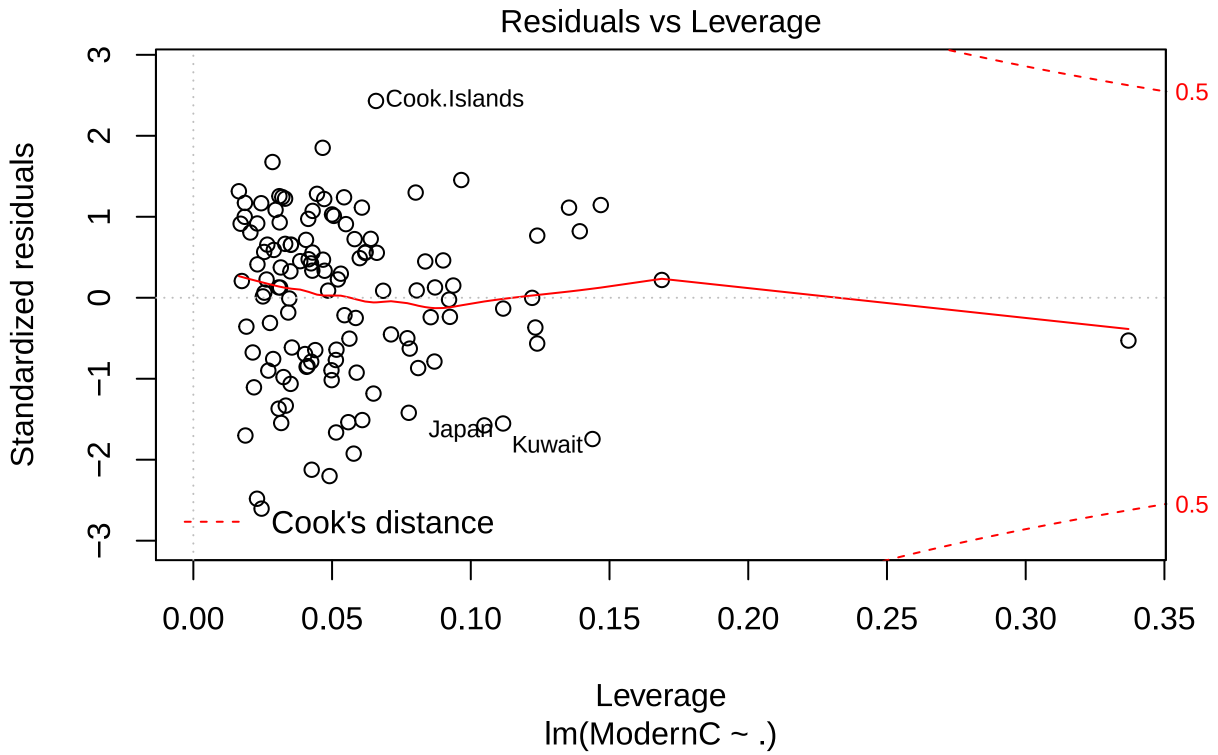
```
##
## Call:
## lm(formula = ModernC ~ Purban + Change + Frate + log(Pop) + log(Fertility) +
##     log(PPgdp), data = UN3[UN3$Pop < 1e+06, ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -39.926 -12.029   2.916  11.167  30.590
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -12.49501   15.82078  -0.790  0.43127
## Purban           0.02189    0.10673   0.205  0.83789
## Change           2.18291    2.53443   0.861  0.39085
## Frate            0.16688    0.08358   1.997  0.04820 *
## log(Pop)         1.18964    0.72500   1.641  0.10353
## log(Fertility) -17.39486    6.27937  -2.770  0.00653 **
## log(PPgdp)       6.31407    1.49317   4.229 4.71e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.39 on 116 degrees of freedom
##   (85 observations deleted due to missingness)
## Multiple R-squared:  0.5663, Adjusted R-squared:  0.5439
## F-statistic: 25.25 on 6 and 116 DF,  p-value: < 2.2e-16
```

```
plot(UN3_mod_no_chi_ind)
```


Residuals vs Fitted
Residuals
Fitted values
lm(ModernC ~ .)
Cook.Islands
Azerbaijan
Poland
40
20
0
−20
−40
0
20
40
60
80

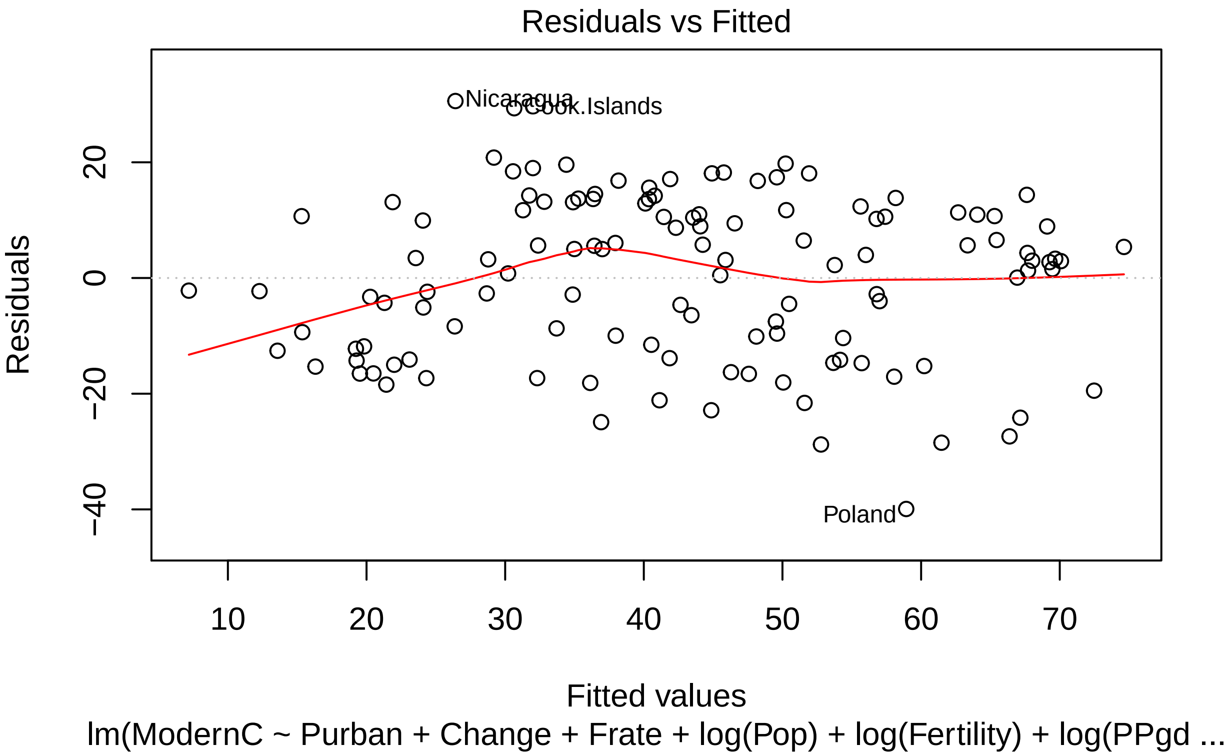

Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(ModernC ~ .)
3
2
1
0
−1
−2
−2
−1
0
1
2
Cook.Islands
Azerbaijan
Poland

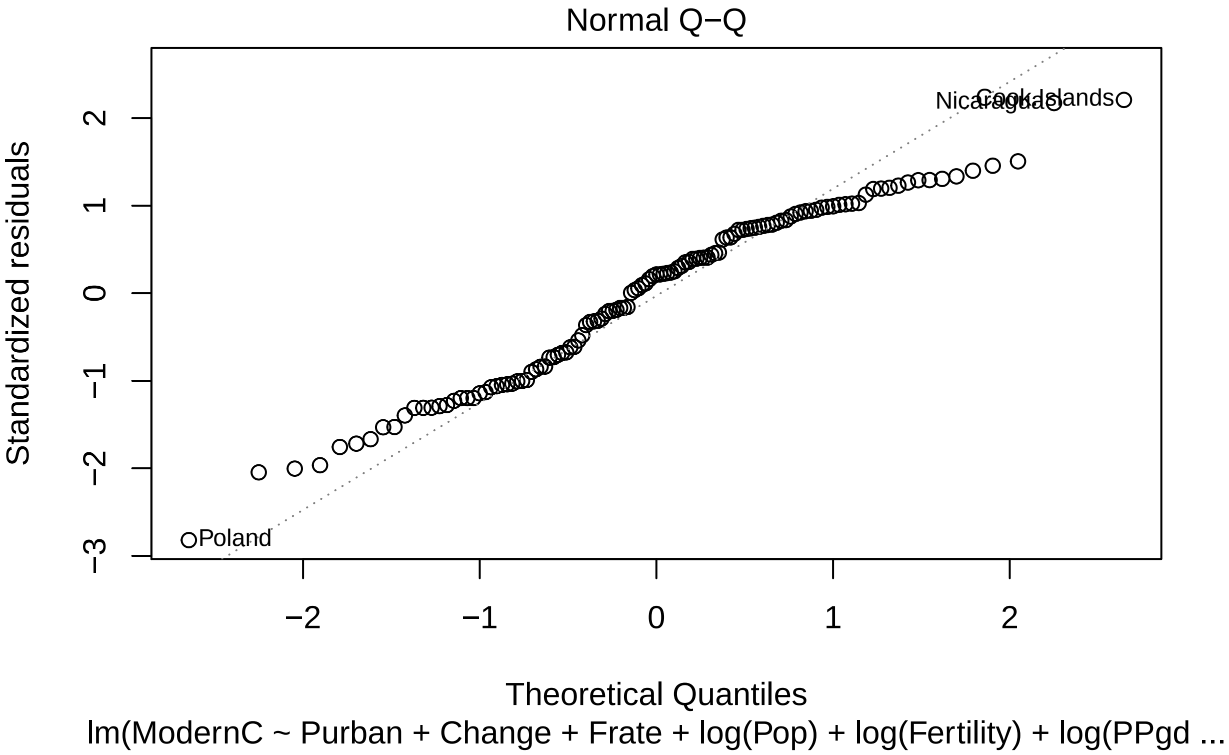

Scale−Location
Cook.Islands
Azerbaijan
Poland
√|Standardized residuals|
0.0
0.5
1.0
1.5
0
20
40
60
80
Fitted values
lm(ModernC ~ .)

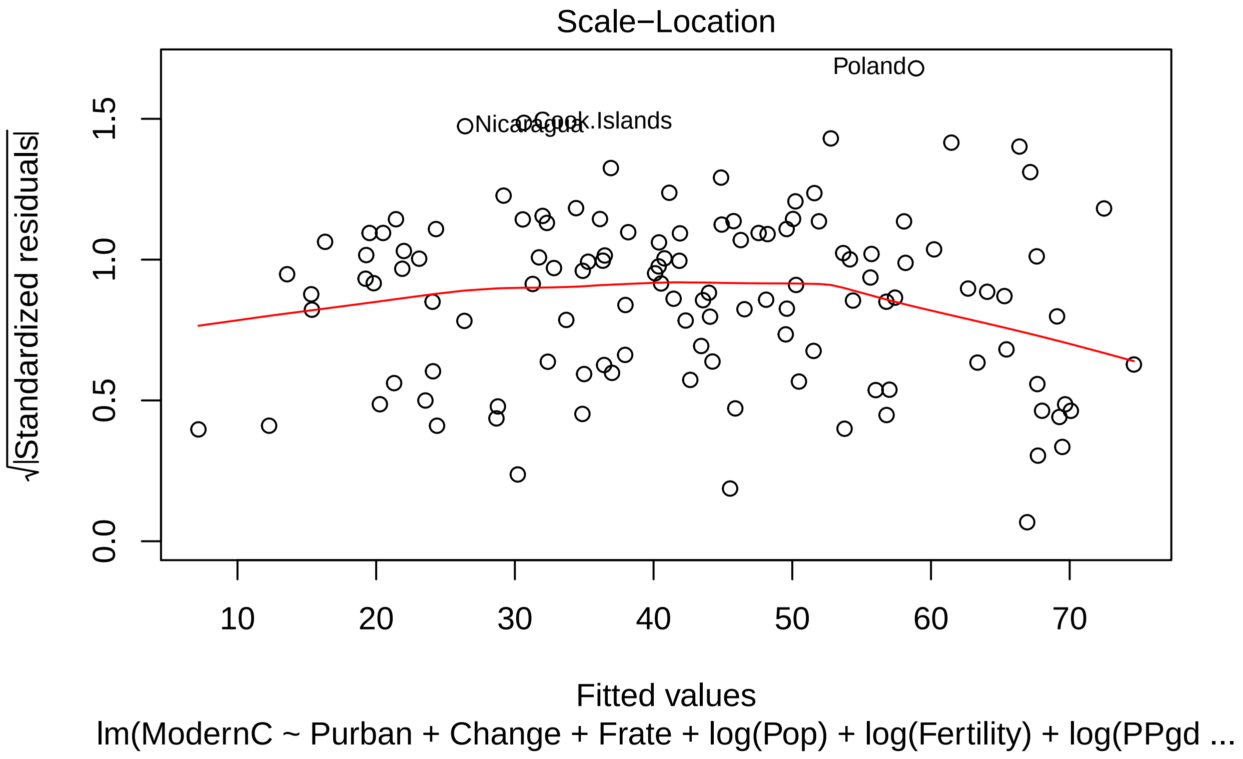


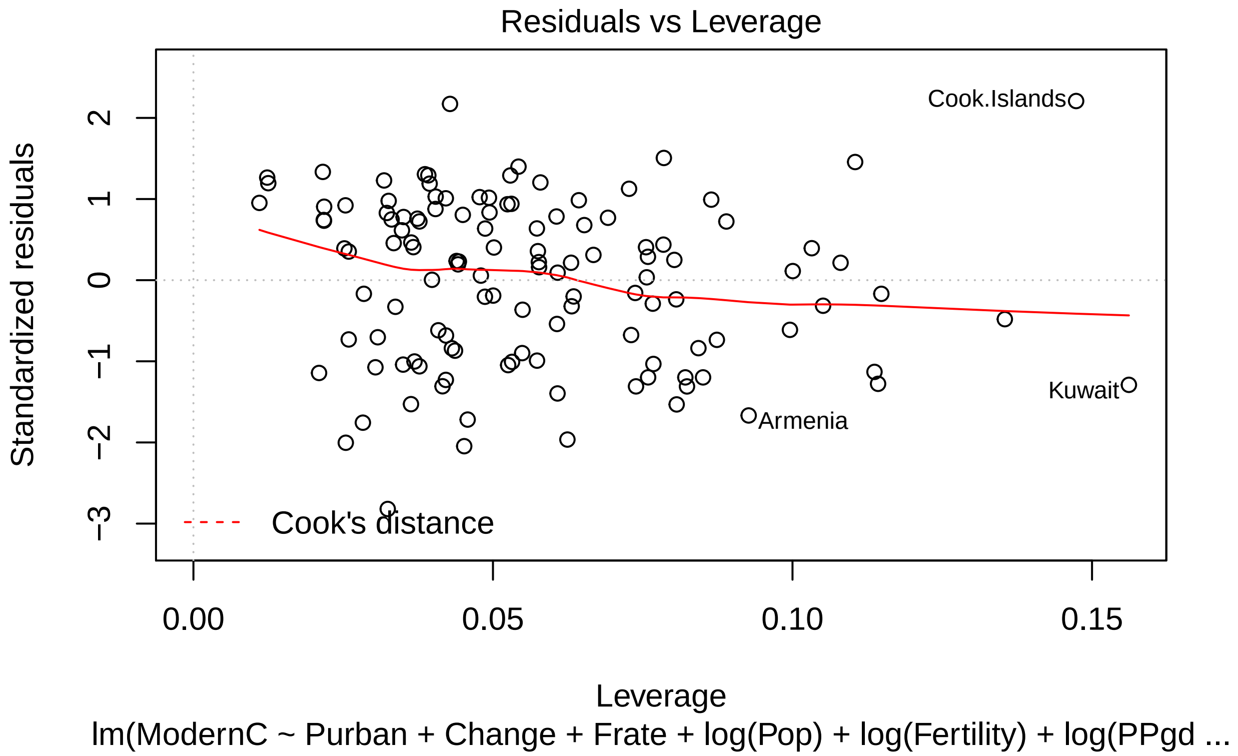
Above are the residual plots with China and India. Below are the plots for the model without China and India.

```
plot(UN3_mod_no_chi_ind_tran)
```


Residuals vs Fitted
Nicaragua
Cook.Islands
Poland
Residuals
20
0
−20
−40
10
20
30
40
50
60
70
Fitted values
lm(ModernC ~ Purban + Change + Frate + log(Pop) + log(Fertility) + log(PPgd ...


Normal Q−Q
Standardized residuals
−3
−2
−1
0
1
2
Poland
Nicaragua
Cook Islands
−2
−1
0
1
2
Theoretical Quantiles
lm(ModernC ~ Purban + Change + Frate + log(Pop) + log(Fertility) + log(PPgd ...


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
Poland
Nicaragua
Cook.Islands
10
20
30
40
50
60
70
Fitted values
lm(ModernC ~ Purban + Change + Frate + log(Pop) + log(Fertility) + log(PPgd ...



Removal of the outliers does not result in any noticeable changes in the residual plots. The Q-Q plot still shows light tails and there is no clear pattern to the residuals. The constant variance assumption still appears suitable. Again, it should be noted that these comments should only apply to the residual plots for the model with the log transformations since the model without them had non linear relationships among the predictors.

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!

Note above that the coefficients for the logarithm variables should be interpreted as the percentage change in ModernC that corresponds to a 10% increase in the predictor (on average and all else equal)

Interpretation of the coefficients:

Intercept: When all the other variables have values of zero, the average percent of unmarried women using a modern method of contraception is -15%. Obviously, this intercept has no meaning, and there is never a situation in which all the other predictors could be zero.

Purban: On average and all else constant, a one percentage point increase in the urban population corresponds to a 0.007 percentage point decrease in modern contraception usage among unmarried women.

Change: On average and all else constant, a one percentage point increase in population growth corresponds to a 2.3102 percentage point increase in modern contraception usage among unmarried women.

Frate: On average and all else constant, a one percentage point increase in females over 15 economically active corresponds to a 0.178 percentage point increase in modern contraception usage among unmarried women.

Pop: On average and all else constant, a 10% increase in population corresponds to a 1.164% increase in the percentage of unmarried women using modern contraception.

Fertility: On average and all else constant, a 10% increase in the expected number of live births per female corresponds to a 0.176% increase in the percentage of unmarried women using modern contraception.

PPgdp: On average and all else constant, a 10% increase in the per capita GDP corresponds to a 1.848% increase in the percentage of unmarried women using modern contraception.

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

Using a linear regression model, we are able to describe, for 125 localities, the linear relationship between the percent of unmarried women using a modern method of contraception and seven predictor variables: population, population growth rate, per capita GDP, percent of females over 15 economically active, expected number of live births per female (year 2000), and percent of population that is urban (year 2001). The final model is statistically significant and explains roughly 56% of the variation in the modern contraception rate. The data set used to create the model presented several challenges. First, the data contained 210 localities, but only 125 of them were not missing values and thus could be used in the regression. If these variables were not missing at random, the model may not be appropriate for extending to other localities or explaining the true relationship. Additionally, the population size of China and India were outliers. However, they were kept in the model for a couple of reasons. First, the logarithm of population produced a fairly linear relationship with modern contraception usage. Additionally, while they were outliers, they contained valuable information about the spread of populations. Populations of countries are not evenly distributed around the mode, and thus removing China and India would have made countries such as Brazil "new" outliers. After giving population, PPgdp, and fertility a logarithmic transformation, all of the predictors had linear relationships with our response. Based on the model, female economic activity, the log of Fertility, and the log of GDP per capita all have statistically significant relationships with modern contraception usage by unmarried women. While fertility has a negative relationship with modern contraception usage for unmarried women, female economic activity and GDP per capita are both positively associated with its usage.

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.*

First, note that in Simple Linear Regression, the intercept, β_0, is given by the following formula:

$$\hat{\beta_0} = \bar{y} - \hat{\beta}_1\bar{x}$$

Since the added variable plot plots two residuals on each other, both $\bar{x}$ and $\bar{y}$ are the average of residuals.

$$\bar{e} = \frac{\sum_{k=0}^{n}\hat{e}}{n}$$

$$\begin{aligned}\sum_{k=0}^{n}\hat{e} &= 1'\hat{e} \\ &= 1'(Y - X\hat{\beta_0}) \\ &= 1'(Y - X(X'X)^{-1}X'Y) \\ &= 1'(I_n - X(X'X)^{-1}X')Y \\ &= 1'(I_n - H)Y = 0Y = 0.\end{aligned}$$

Thus, $\bar{e}$ is 0. Then:

$$\hat{\beta_0} = 0 - \hat{\beta_1}0 = 0$$

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.

The following will demonstrate that the slope of Change in the full model is the same as the slope of the added variable plot.

```
summary(UN3_mod_tran)
```

```
##
## Call:
## lm(formula = ModernC ~ Purban + Change + Frate + log(Pop) + log(Fertility) +
##     log(PPgdp), data = UN3_naomit)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -41.235 -11.589   2.498  10.748  31.954
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -15.106490  15.949888  -0.947  0.34551
## Purban          -0.007352   0.106591  -0.069  0.94513
## Change           2.310274   2.560728   0.902  0.36879
## Frate            0.178242   0.083567   2.133  0.03500 *
## log(Pop)         1.595611   0.699289   2.282  0.02430 *
## log(Fertility) -18.237639   6.336680  -2.878  0.00475 **
## log(PPgdp)       6.445713   1.508057   4.274 3.91e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.55 on 118 degrees of freedom
## Multiple R-squared:  0.5615, Adjusted R-squared:  0.5392
## F-statistic: 25.19 on 6 and 118 DF,  p-value: < 2.2e-16
```

From the summary, we see that the slope of Change is 2.31.

In the added variable plot, we have the residuals from ModernC regressed on all the variables except Change on the Y axis. On the X axis, we have the residuals from Change regressed on the other predictors. Below, we regress these two residuals.

```
## The residuals from ModernC regressed on all the variables except Change
resid_y <- lm(formula = ModernC ~ Purban + Frate + log(Pop) + log(Fertility) +
    log(PPgdp), data = UN3_naomit)

## The residuals from Change regressed on the other predictors
resid_x <- lm(formula = Change ~ Purban + Frate + log(Pop) + log(Fertility) +
    log(PPgdp), data = UN3_naomit)

summary(lm(resid_y$residuals ~ resid_x$residuals))
```

```
##
```

```
## Call:
## lm(formula = resid_y$residuals ~ resid_x$residuals)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -41.235 -11.589   2.498  10.748  31.954
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -3.441e-17  1.275e+00   0.000    1.000
## resid_x$residuals  2.310e+00  2.508e+00   0.921    0.359
##
## Residual standard error: 14.26 on 123 degrees of freedom
## Multiple R-squared:  0.006851,   Adjusted R-squared:  -0.001224
## F-statistic: 0.8484 on 1 and 123 DF,  p-value: 0.3588
```

Thus, since the coefficient is 2.31, we have confirmed that the slope in the added variable plot is the same as the slope in the full model.