HW2 STA521 Fall18

[Your Name Here, netid and github username here]

Due September 23, 2018 5pm

Backgound Reading

Readings: Chapters 3-4 in Weisberg Applied Linear Regression

This exercise involves the UN data set from `alr3` package. Install `alr3` and the `car` packages and load the data to answer the following questions adding your code in the code chunks. Please add appropriate code to the chunks to suppress messages and warnings as needed once you are sure the code is working properly and remove instructions if no longer needed. Figures should have informative captions. Please switch the output to pdf for your final version to upload to Sakai. **Remove these instructions for final submission**

Exploratory Data Analysis

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

```
library(alr3)
```

```
## Warning: package 'alr3' was built under R version 3.4.4

## Loading required package: car

## Warning: package 'car' was built under R version 3.4.4

## Loading required package: carData
```

```
data(UN3, package="alr3")
help(UN3)
```

```
## starting httpd help server ...

##  done
```

```
library(car)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following object is masked from 'package:car':
##
##     recode

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(knitr)
library(ggplot2)
library(GGally)
```

```
## Warning: package 'GGally' was built under R version 3.4.4
```

```
##
## Attaching package: 'GGally'
```

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

```r
library(car)
```

1. Create a summary of the data. How many variables have missing data? Which are quantitative and which are qualtitative?

```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. Thus, all the data is quatitative.

Unfortuntately, all of the variables with the exception of Purban contain missing values (NA).

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() %>%kable(, col.names = c("Mean", "Standard Deviation"), align = 'l')
```

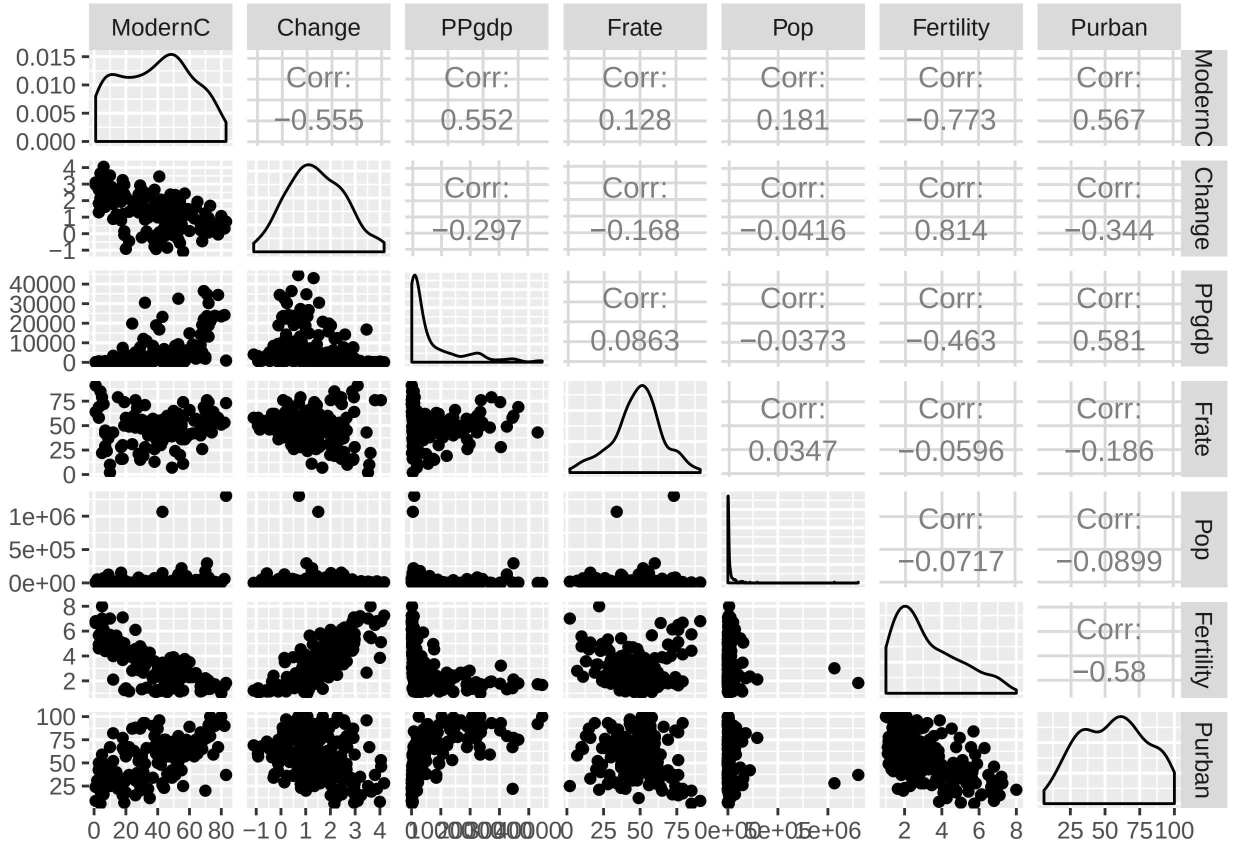
	Mean	Standard Deviation
ModernC	38.717105	2.263661e+01
Change	1.418373	1.133133e+00
PPgdp	6527.388060	9.325189e+03

	Mean	Standard Deviation
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?

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

```
## Warning: Removed 58 rows containing non-finite values (stat_density).

## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 58 rows containing missing values

## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 60 rows containing missing values

## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 82 rows containing missing values

## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 58 rows containing missing values

## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 60 rows containing missing values

## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 58 rows containing missing values

## Warning: Removed 58 rows containing missing values (geom_point).

## Warning: Removed 1 rows containing non-finite values (stat_density).

## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 10 rows containing missing values

## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 43 rows containing missing values

## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 2 rows containing missing values

## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 11 rows containing missing values

## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removing 1 row that contained a missing value

## Warning: Removed 60 rows containing missing values (geom_point).

## Warning: Removed 10 rows containing missing values (geom_point).

## Warning: Removed 9 rows containing non-finite values (stat_density).

## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 50 rows containing missing values

## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 11 rows containing missing values
## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 17 rows containing missing values

## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 9 rows containing missing values

## Warning: Removed 82 rows containing missing values (geom_point).

## Warning: Removed 43 rows containing missing values (geom_point).

## Warning: Removed 50 rows containing missing values (geom_point).

## Warning: Removed 43 rows containing non-finite values (stat_density).

## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 43 rows containing missing values

## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 49 rows containing missing values

## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 43 rows containing missing values

## Warning: Removed 58 rows containing missing values (geom_point).

## Warning: Removed 2 rows containing missing values (geom_point).

## Warning: Removed 11 rows containing missing values (geom_point).

## Warning: Removed 43 rows containing missing values (geom_point).

## Warning: Removed 2 rows containing non-finite values (stat_density).

## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 11 rows containing missing values

## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 2 rows containing missing values

## Warning: Removed 60 rows containing missing values (geom_point).

## Warning: Removed 11 rows containing missing values (geom_point).

## Warning: Removed 17 rows containing missing values (geom_point).

## Warning: Removed 49 rows containing missing values (geom_point).

## Warning: Removed 11 rows containing missing values (geom_point).

## Warning: Removed 10 rows containing non-finite values (stat_density).

## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 10 rows containing missing values

## Warning: Removed 58 rows containing missing values (geom_point).

## Warning: Removed 1 rows containing missing values (geom_point).

## Warning: Removed 9 rows containing missing values (geom_point).

## Warning: Removed 43 rows containing missing values (geom_point).

## Warning: Removed 2 rows containing missing values (geom_point).

## Warning: Removed 10 rows containing missing values (geom_point).
```



As seen in the message, each plot contained missing values that were removed by the ggpairs plotting function.

From the above plots, we notice several things.

(1) The marginal relationship between ModernC and Change is approximately linear and negatively correlated. 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 general tend to have lower wealth/education could lead to speculation that the relationship between change and PPgdp is important in predicting ModernC.

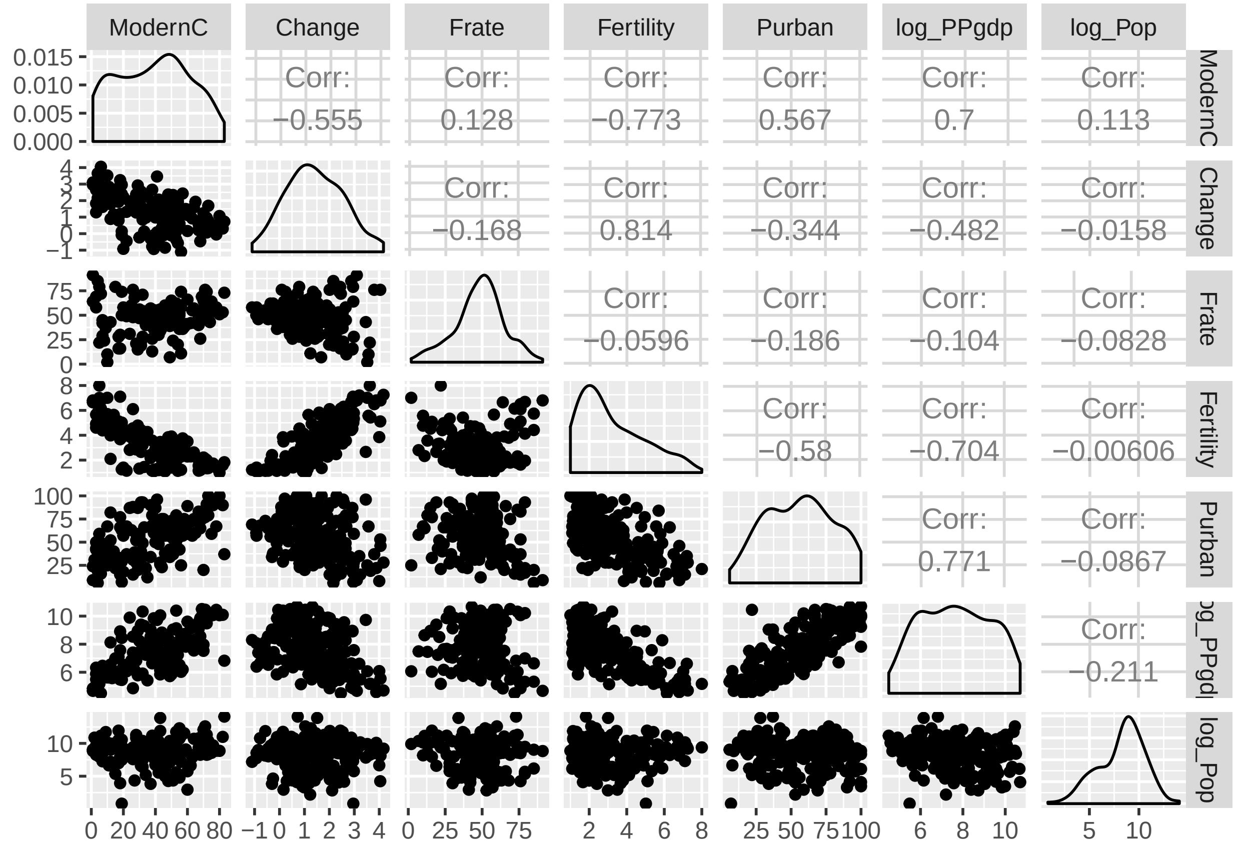
(2) Most of the populations are under 18,913,500. However, two populations are extremely large, with the largest being 1,304,196,000 (China). Consequentially, a logarithmic transformation of population could bring these orders of magnitude into scale.

(3) Another candidate for a logarithmic transformation is PPgdp (GDP per capita), as several of the richer countries have much larger values than others.

After transforming the variables, we get the following pair plots.

```
UN3 %>% mutate(log_PPgdp = log(PPgdp), log_Pop = log(Pop)) %>% select(-Pop, -PPgdp) %>% ggpairs(, progre
```

```
## Warning: Removed 58 rows containing non-finite values (stat_density).

## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 58 rows containing missing values

## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 82 rows containing missing values

## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 60 rows containing missing values
## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 58 rows containing missing values

## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 60 rows containing missing values

## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 58 rows containing missing values

## Warning: Removed 58 rows containing missing values (geom_point).

## Warning: Removed 1 rows containing non-finite values (stat_density).

## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 43 rows containing missing values

## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 11 rows containing missing values

## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removing 1 row that contained a missing value

## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 10 rows containing missing values

## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 2 rows containing missing values

## Warning: Removed 82 rows containing missing values (geom_point).

## Warning: Removed 43 rows containing missing values (geom_point).

## Warning: Removed 43 rows containing non-finite values (stat_density).

## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 49 rows containing missing values

## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 43 rows containing missing values

## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 50 rows containing missing values

## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 43 rows containing missing values

## Warning: Removed 60 rows containing missing values (geom_point).

## Warning: Removed 11 rows containing missing values (geom_point).

## Warning: Removed 49 rows containing missing values (geom_point).

## Warning: Removed 10 rows containing non-finite values (stat_density).

## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 10 rows containing missing values

## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 17 rows containing missing values

## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 11 rows containing missing values

## Warning: Removed 58 rows containing missing values (geom_point).

## Warning: Removed 1 rows containing missing values (geom_point).
## Warning: Removed 43 rows containing missing values (geom_point).

## Warning: Removed 10 rows containing missing values (geom_point).

## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 9 rows containing missing values

## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 2 rows containing missing values

## Warning: Removed 60 rows containing missing values (geom_point).

## Warning: Removed 10 rows containing missing values (geom_point).

## Warning: Removed 50 rows containing missing values (geom_point).

## Warning: Removed 17 rows containing missing values (geom_point).

## Warning: Removed 9 rows containing missing values (geom_point).

## Warning: Removed 9 rows containing non-finite values (stat_density).

## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 11 rows containing missing values

## Warning: Removed 58 rows containing missing values (geom_point).

## Warning: Removed 2 rows containing missing values (geom_point).

## Warning: Removed 43 rows containing missing values (geom_point).

## Warning: Removed 11 rows containing missing values (geom_point).

## Warning: Removed 2 rows containing missing values (geom_point).

## Warning: Removed 11 rows containing missing values (geom_point).

## Warning: Removed 2 rows containing non-finite values (stat_density).
```



With these transformations, most of the relationships appear more linear. From a marginal perspective, all of the variables except for log(PPgdp) 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 variables individually.

While not all the points fit very closely along a straight line, the plots do not raise any immediate concerns about outliers. Transforming population logarithmically 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 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?

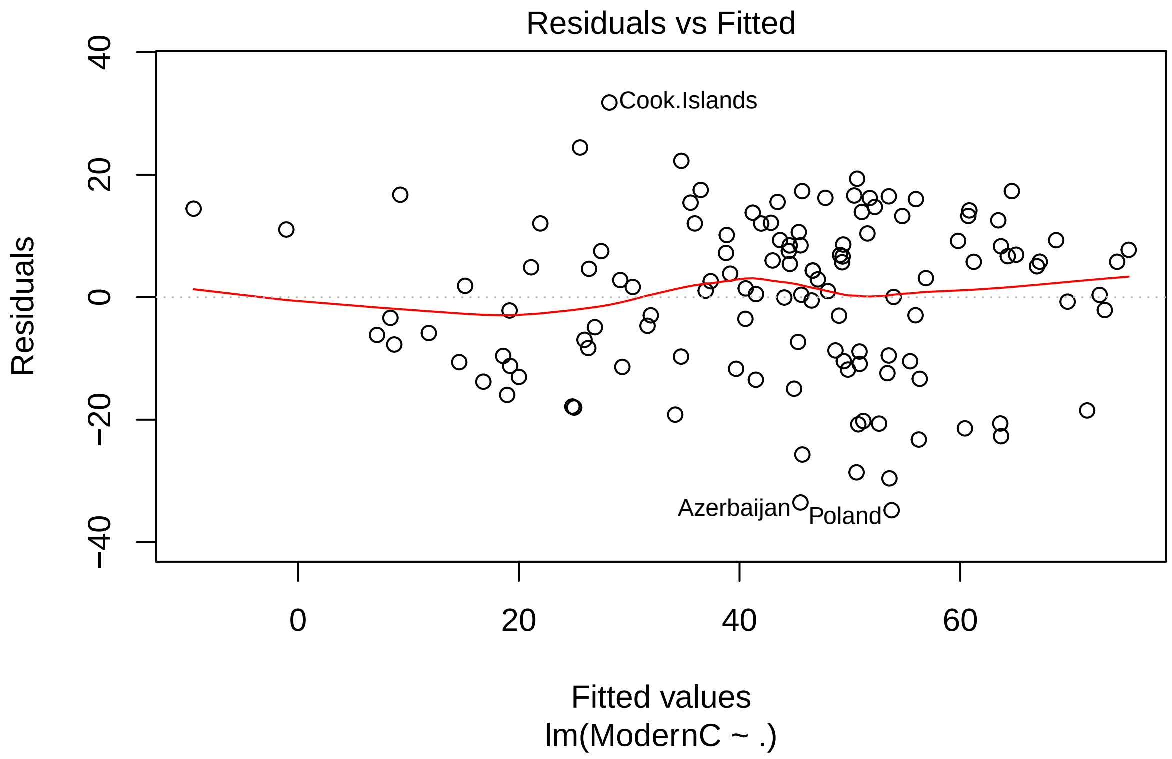
```r
UN3_mod_simple <- lm(ModernC ~., data = UN3)
summary(UN3_mod_simple)
```

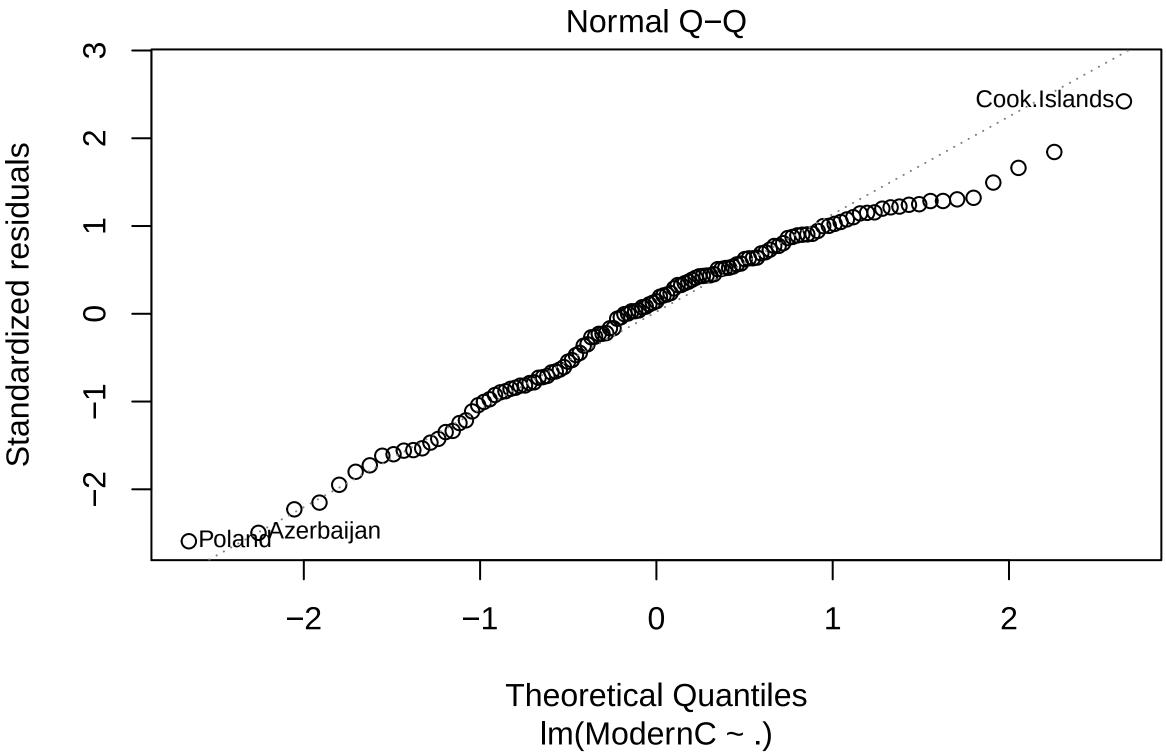
```
##
## Call:
## lm(formula = ModernC ~ ., data = UN3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -34.781  -9.698   1.858   9.327  31.791
##
## Coefficients:
```

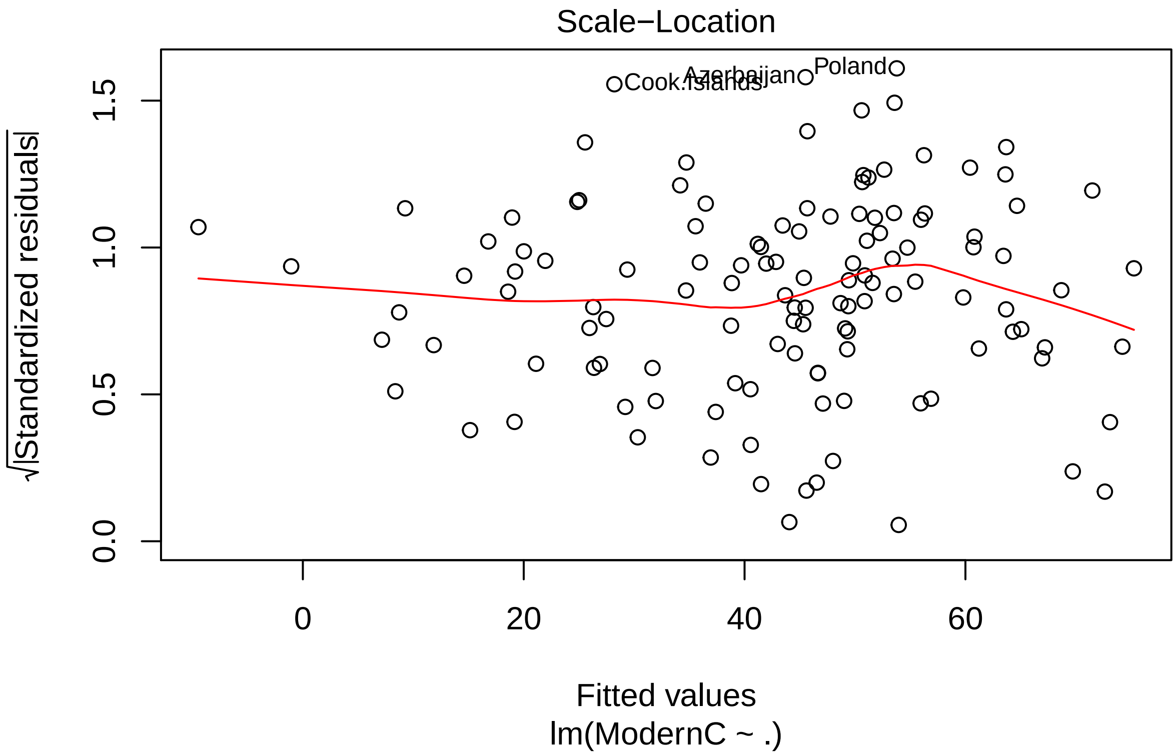
```
##                  Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  5.529e+01  9.467e+00   5.841 4.69e-08 ***
## Change       5.268e+00  2.088e+00   2.524  0.01294 *  
## PPgdp        5.301e-04  1.770e-04   2.995  0.00334 ** 
## Frate        1.232e-01  8.060e-02   1.529  0.12901    
## Pop          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
```

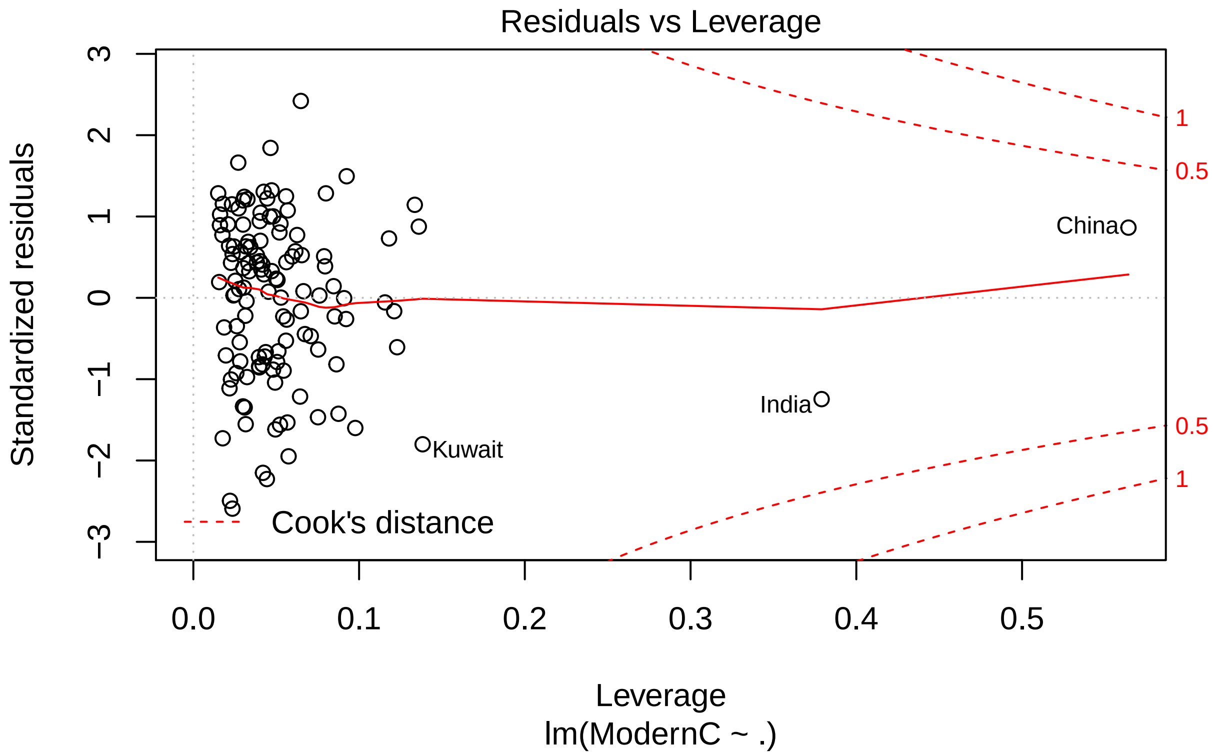
Because the lm function in R removes observations with missing data as a default, there were 85 observations that were deleted. Considering there were only 210 observations to start with, we are only left with 125 (60%) of the observations.

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



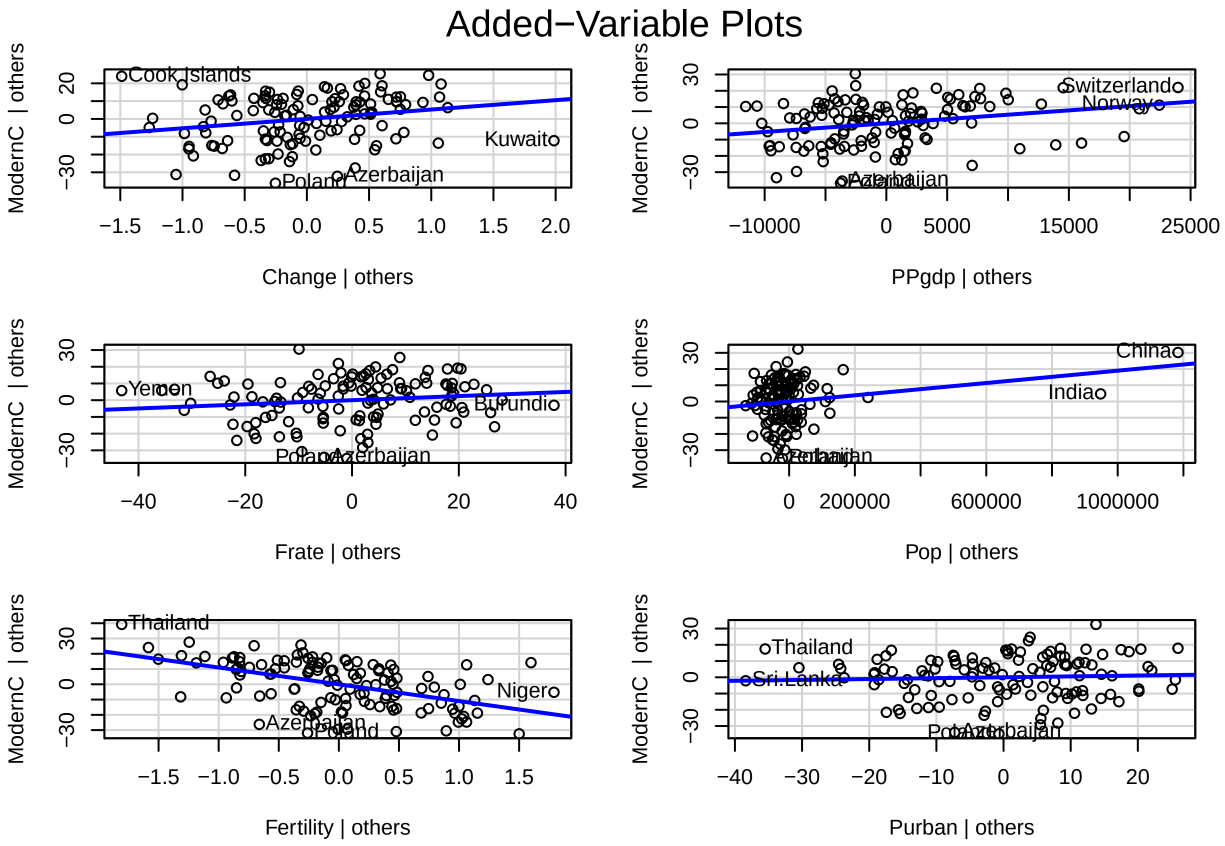
From the Residuals vs Fitted plot, we see that the variance of the residuals is roughtly constant and the dispersion around 0 is not extreme for any one point in relation to the others. Poland and Azerbaikan had fitted values of ModernC the furthest above their actual values, while the Cook Islands had a fitted value the further below its actual value. The red fitted line to the residuals shows that, while there is some curvature to the residuals, there is no clear, discernable pattern.

When we examine the Normal Q-Q plot, we see that the residuals definitely do not follow a theoretical normal distribution. The residuals on the negative end follow the theoretical quantiles, but the positive residuals are not as dispersed on the upper end as would eb expected from normally distributed errors. Thus, our assumption of normally distributed errors appears to be incorrect.

In the Residuals vs Leverage plot, we see that China and India has the largest leveral values, but neitiher are very influential based on their Cook's Distance. There is also no discernable 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?

```
avPlots(UN3_mod_simple)
```



Looking at the added variabke plots, we can see the amount of variation unexplained by the model compared to the variation in the predictor 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. Most of the relationships appear linear with the exception of population, which has a cluster of values with populations significantly smaller than the largest several countries. Because of the distribution of populations, log transforming this variable might be an appropriate solution.

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.

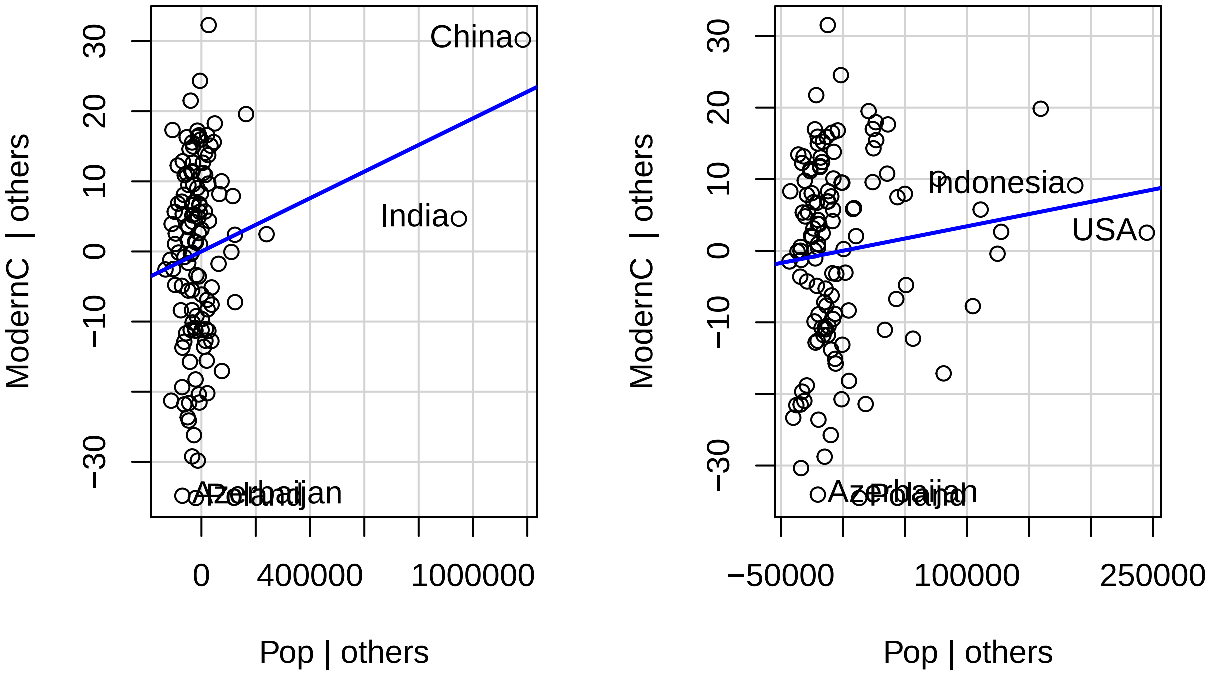
```
UN3_mod_no_chi_ind <- lm(ModernC ~ ., data = UN3[UN3$Pop < 1000000,])
china_india <- UN3[UN3$Pop > 1000000,] %>% filter(is.na(Pop) == "FALSE")
p_wo_chi_ind <- as.vector(predict(UN3_mod_no_chi_ind, newdata = china_india))
p_w_chi_ind <- as.vector(predict(UN3_mod_simple, newdata = china_india))
data.frame("With_China_India" = 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, greatly influence the fit of the model.

This is also seen clearly in the added variable plots

```
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")
```



```
par(mfrow = c(1,1))
```

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

```
## 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             0.33925             -1.2111  0.22587
## Frate           -38.16826              0.9287  0.35303
## Pop               0.47142             -0.9180  0.35864
## Fertility         1.42133             -1.9578  0.05026 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## iterations =  26
```

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.