Final Project

Due by 11:59am, Saturday, May 7, 2022

S&DS 230/S&DS 530/ENV 757/PLSC 530

Introduction: Background and Motivation

In this project, we will be looking at the effect of various factors on GDP per capita.

To do this, we will use the data from the World Bank (WDI) which includes yearly data across a variety of variables for 175 countries from 1970-2019. We are primarily interested in GDP per capita and our question of interest: among the variables selected for analysis, which ones have the greatest correlation with gdp per capita? We would also like to see how variables, such as the presence of armed conflict, relates to GDP per capita.

```
fdi <- read.csv("fdi_data.csv")
dim(fdi)
head(fdi)
```

Data Cleaning

Variable Names and Descriptions

We have chosen to look at the following variables:

Categorical variables:

- conflict | Shows whether the country is in an armed conflict with any other country
  - Value of 1 indicates conflict
  - Value of 0 indicates no conflict
- conflict_intensity | Shows the intensity of armed conflict in the country
  - Value of 0 indicates no conflict
  - Value of 1 indicates minor conflict
  - Value of 2 indicates major conflict
- conflict_type | Shows the type of conflict
  - Value of 0 indicates no conflict
  - Value of 1 indicates extrasystemic conflict
  - Value of 2 indicates interstate conflict
  - Value of 3 indicates intrastate conflict
  - Value of 4 indicates internationalized intrastate

Continuous variables:

- gdppc: Gross domestic product per capita (in USD)
- trade_dependence: Total trade (imports + exports) as percentage of GDP
- v2x_polyarchy
  - Value from 0 - 1
  - Electoral democracy index where higher values == stronger democratic institutions
- v2x_rule
  - Value from 0 - 1
  - Rule of law index – measures extent to which laws are fairly enforced – higher values == stronger rule of law
- v2x_gender:
  - Value from 0 - 1
  - Women's political empowerment index – higher values == greater political empowerment for women
- ka_open:
  - Value from 0 - 1
  - Measure of capital account openness – higher values == greater integration with global financial markets

Data cleaning process

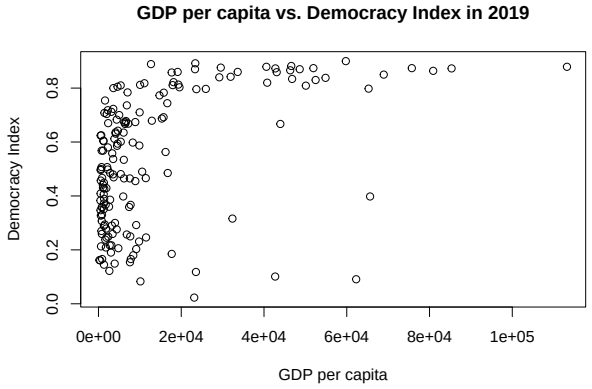
To start, looking at the variables conflict", "conflict_intensity", and "conflict_type", these variables are categorical but are represented as numbers. We will convert these values to booleans/more descriptive values to make it easier to work with and understand the data.

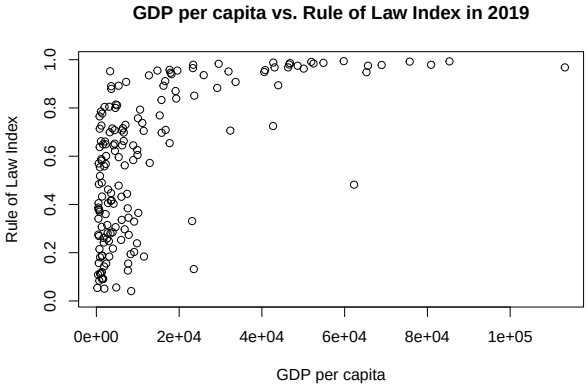
```
fdi$conflict <- as.logical(fdi$conflict)
fdi$conflict_intensity[fdi$conflict_intensity == 0] <- "None"
fdi$conflict_intensity[fdi$conflict_intensity == 1] <- "Minor"
fdi$conflict_intensity[fdi$conflict_intensity == 2] <- "Major"
fdi$conflict_type[fdi$conflict_type == 0] <- "No conflict"
fdi$conflict_type[fdi$conflict_type == 1] <- "Extrasystemic"
fdi$conflict_type[fdi$conflict_type == 2] <- "Interstate"
fdi$conflict_type[fdi$conflict_type == 3] <- "Intrastate"
fdi$conflict_type[fdi$conflict_type == 4] <- "Internationalized intrastate"
```

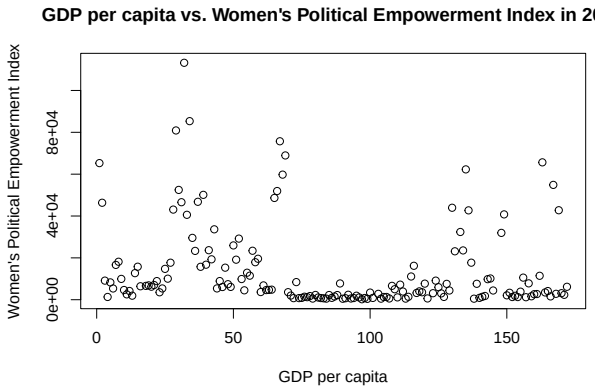
To further narrow down our data for this project, we will only be looking at data from countries in 2019.

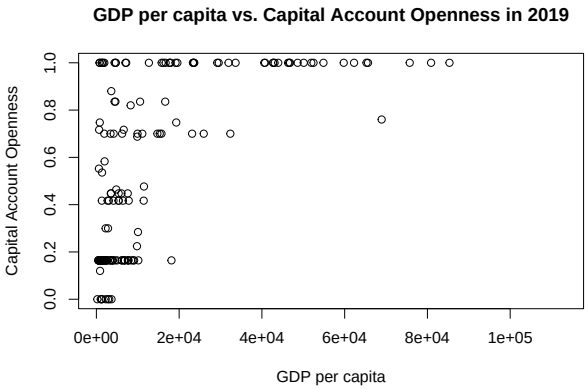
Graphics

Scatterplots

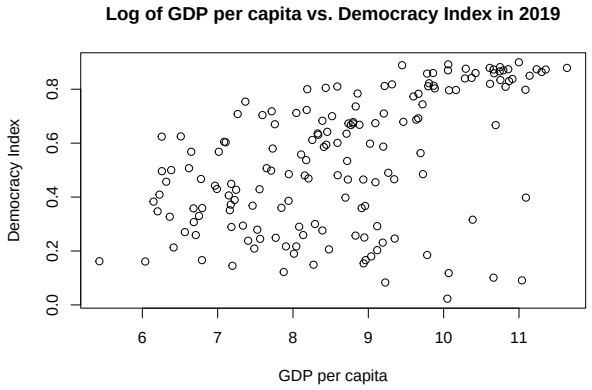


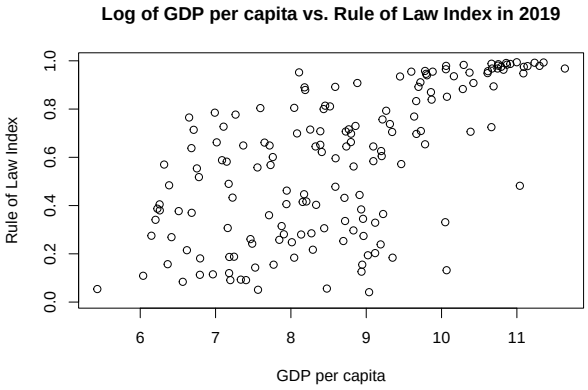


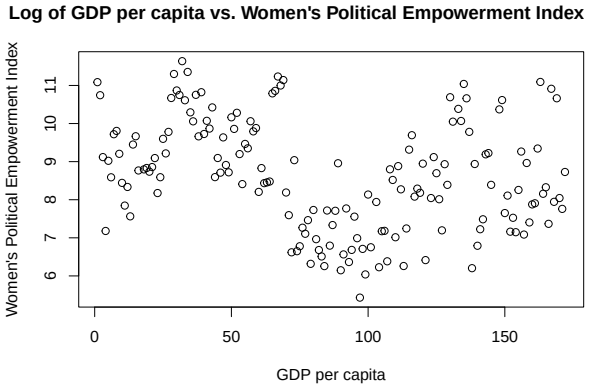


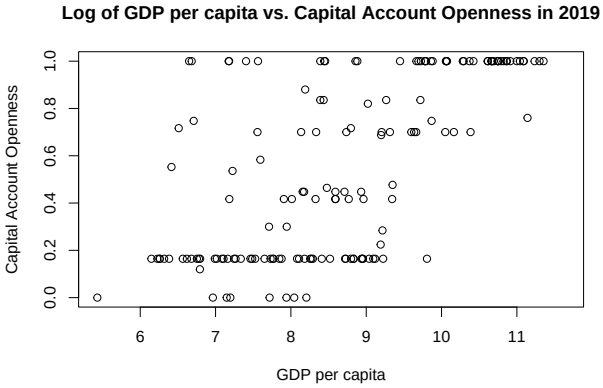


This scatterplot does not convey a lot of useful information due to the outliers on the right side of the plot. To counteract this, we can replot this with the log of the GDP per capita.









After logging the gdppc, it is a bit easier to see that there is a slight positive correlation with gdppc for the democracy index and the rule of law index. Plotting the GDP per capita with the women's empowerment index does not show a significant correlation in any way. The scatter plot with the capital account openness also does not show a significant correlation.

T-Test

Using a t-test, we analyzed if there was a difference in means between countries that were in armed conflict and those which were not. We found that there was a significant difference in means between the two variables.

```
##
##  Welch Two Sample t-test
##
## data:  gdppc_ttest_data$log_gdppc by gdppc_ttest_data$conflict
## t = 5.0539, df = 49.423, p-value = 6.316e-06
## alternative hypothesis: true difference in means between group FALSE and group TRUE is not equal to
## 99 percent confidence interval:
##  0.6138143 1.9986575
## sample estimates:
## mean in group FALSE  mean in group TRUE
##            8.901196            7.594960
```

Because the p-value is 6.316e-06, at an alpha of 0.05 we would conclude that there is a statistically significant difference in GDP per capita between countries that are currently in armed conflict and those that are not.

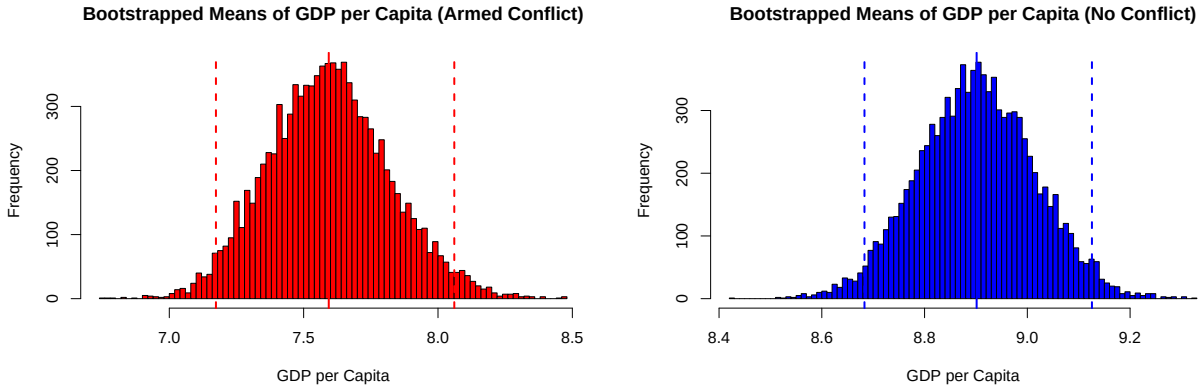
Bootstrap

We then performed a bootstrapped 95% confidence interval:

```
## [1] "The 95% bootstrapped CI for countries with armed conflict is ( 7.174 ,  8.061 )"
```

```
## [1] "The 95% bootstrapped CI for countries with no armed conflict is ( 8.683 ,  9.126 )"
```

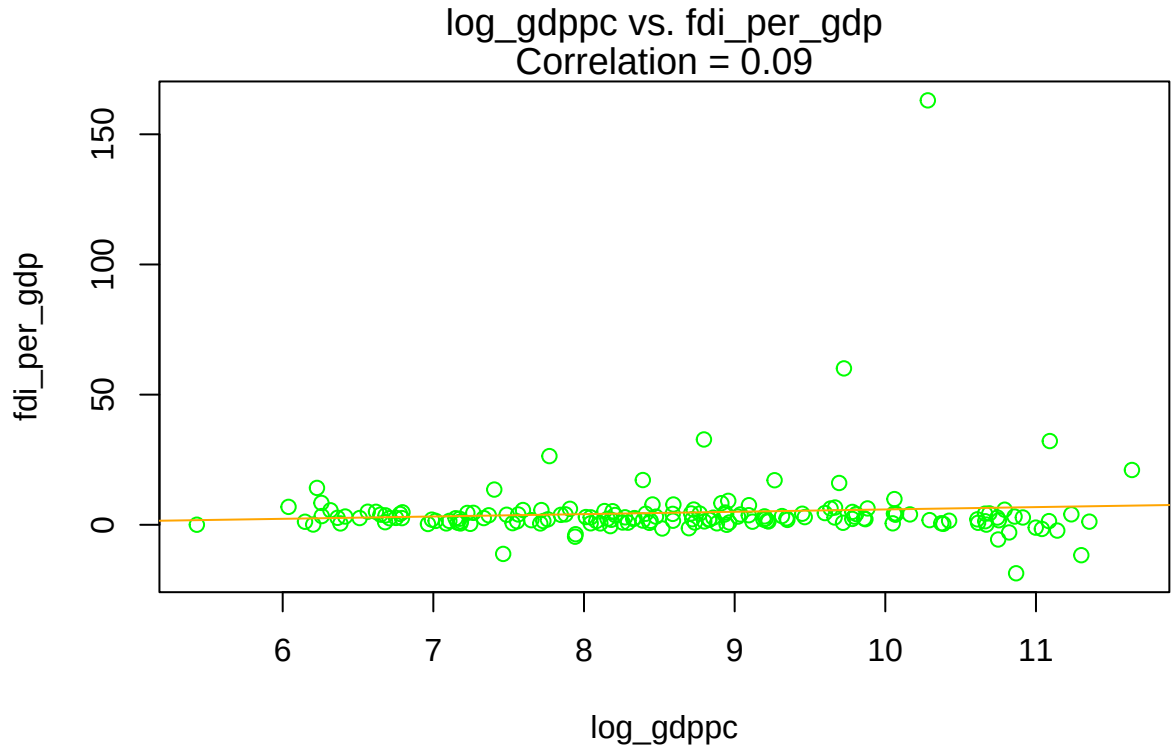
We then created a histogram of the bootstrapped means of GDP per capita for countries with armed conflict and those without.



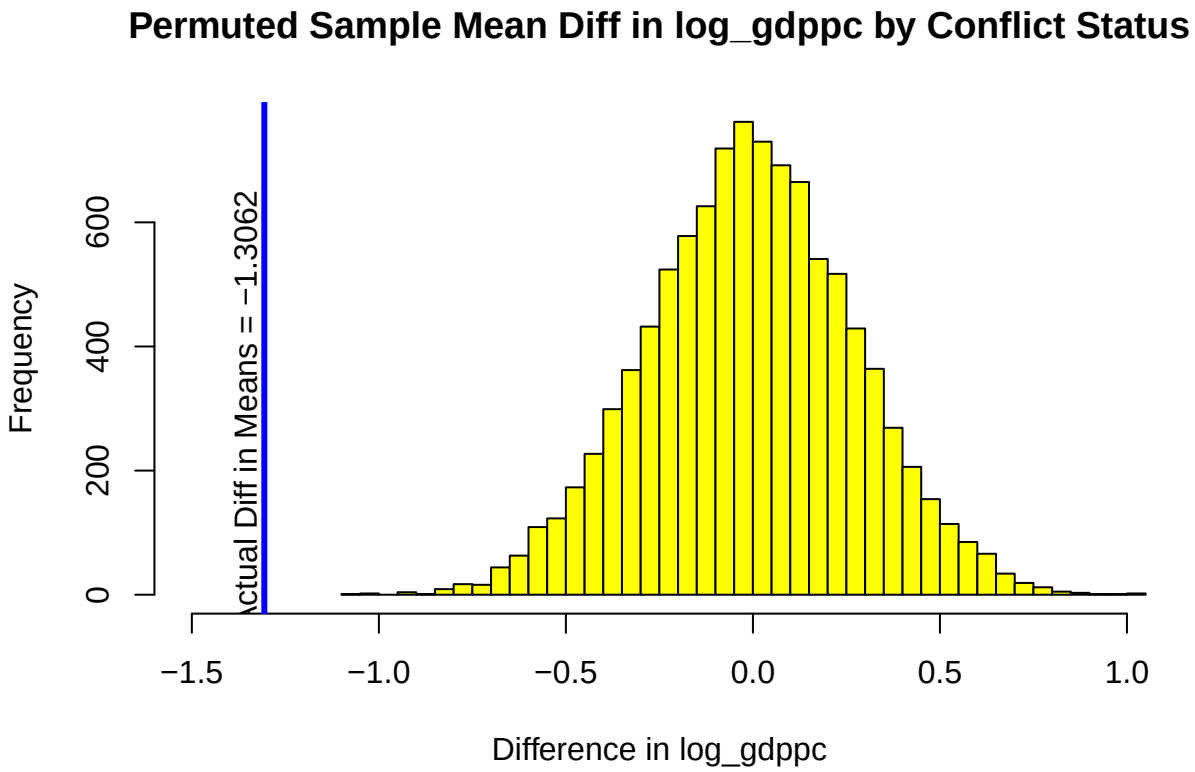
The bootstrapped CIs and histograms indicate the same conclusion as the t-test, which is that there is a significant difference in GDP per Capita between countries that are in armed conflicts versus those that are not.

Correlation

We chose to look at the correlation between FDI Per GDP and GDP Per Capita and found that out of the continuous indicator variables these had a weak correlation of 0.09.



Permutation Test




```
## [1] "P-value: 0"
```

The permuted sample mean difference in log_gdppc by conflict status suggests that the difference in the sample mean is approximately -1.306. The incredibly low p-value that approaches 0.0 suggests that we reject the null hypothesis and conclude that the mean improvement in log_gdppc between groups is statistically significantly different.

Multiple Regression

```
fdi_mr <- fdi_2019[c("gdppc", "v2x_polyarchy", "v2x_rule", "ka_open")]
fdi_mr <- fdi_mr[complete.cases(fdi_mr),]
```

For multiple regression, we focused on predicting GDP per capita based on the following social indices: wto, v2x_polyarchy, v2x_rule, and ka_open, all of which are continuous variables ranging from 0 to 1. v2x_polyarchy represents how democratic a country is, with lower values meaning less democratic. v2x_rule represents how transparently laws are enforced, with lower values being less transparent. ka_open represents how open the state's capital accounts are, with lower values representing a less open economy to world trade.

```r
library(leaps)
bestSubsets <- regsubsets(gdppc ~ ., data=fdi_mr)
(bestSubsetsSummary <- summary(bestSubsets))
```

```
## Subset selection object
## Call: regsubsets.formula(gdppc ~ ., data = fdi_mr)
## 3 Variables  (and intercept)
##               Forced in Forced out
## v2x_polyarchy     FALSE      FALSE
## v2x_rule          FALSE      FALSE
## ka_open           FALSE      FALSE
## 1 subsets of each size up to 3
## Selection Algorithm: exhaustive
##          v2x_polyarchy v2x_rule ka_open
## 1  ( 1 ) " "           " "      "*"
## 2  ( 1 ) " "           "*"      "*"
## 3  ( 1 ) "*"           "*"      "*"
```

We proceed by looking at the "which" matrix created using best subsets regression. According to the matrix, all variables should be included in the model with the largest r-squared, hence a complete model.

```
##   (Intercept) v2x_polyarchy v2x_rule ka_open
## 1        TRUE         FALSE    FALSE    TRUE
## 2        TRUE         FALSE     TRUE    TRUE
## 3        TRUE          TRUE     TRUE    TRUE
```

```
## [1] 3
```

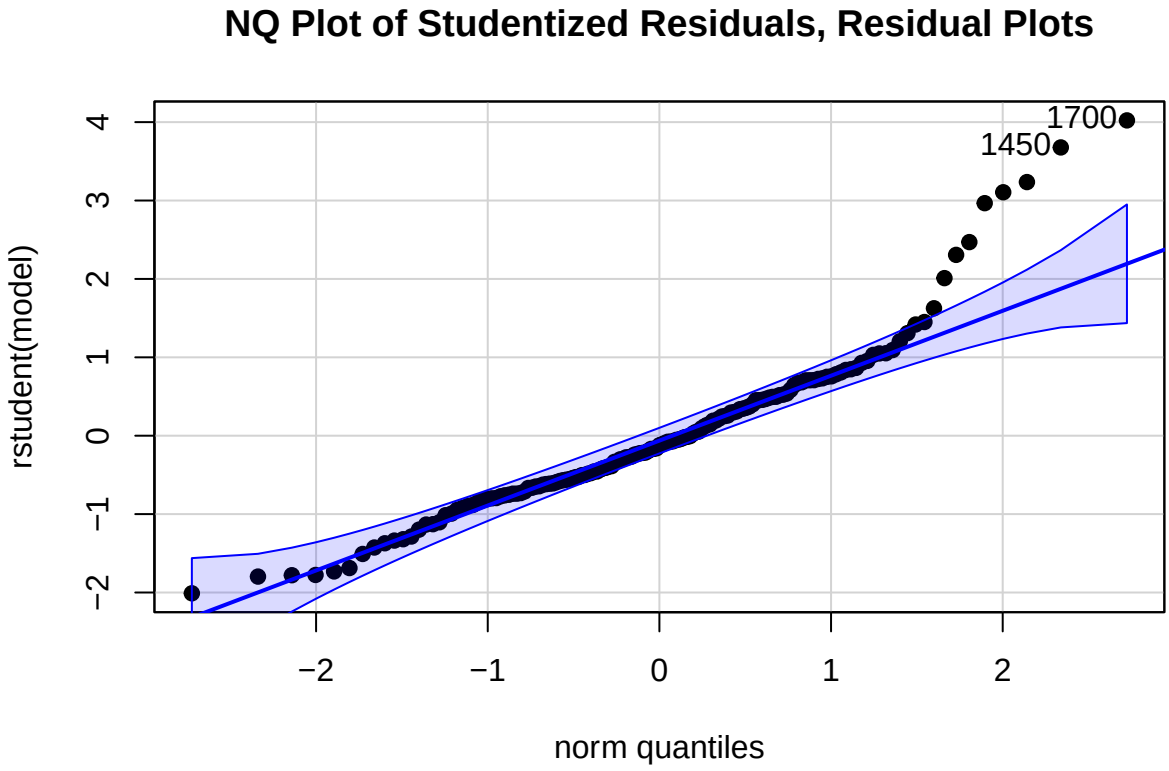
```
##
## Call:
## lm(formula = gdppc ~ ., data = bsTemp)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -27449  -8578  -1686   6853  52806
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -9947       2775  -3.585 0.000455 ***
## v2x_polyarchy    -6485       7880  -0.823 0.411828
## v2x_rule         27569       6850   4.024 9.00e-05 ***
## ka_open          20762       3630   5.719 5.56e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13920 on 151 degrees of freedom
## Multiple R-squared:  0.4712, Adjusted R-squared:  0.4607
## F-statistic: 44.85 on 3 and 151 DF,  p-value: < 2.2e-16
```

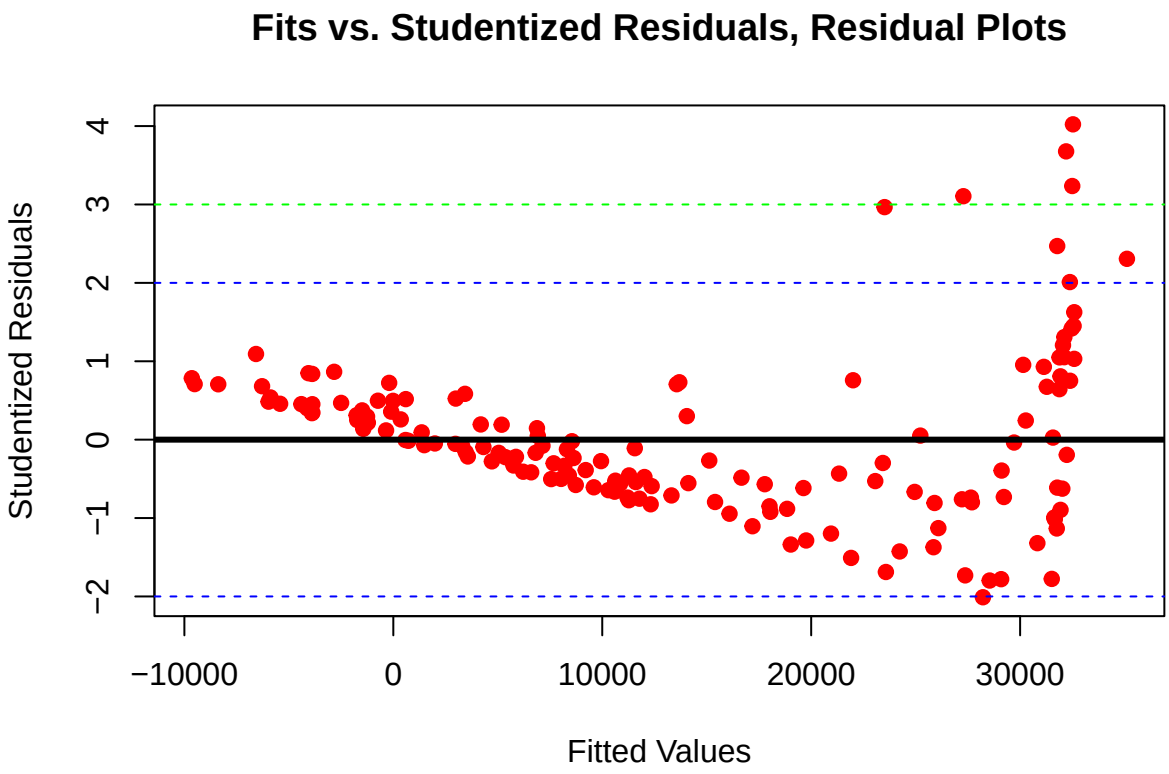
The chosen model is the complete model with all predictors. ka_open, i.e. openness to world trade, is the best predictor of GDP per capita, as shown by its low p-value of 5.56e-08. v2x_rule, i.e. transparency of law enforcement, is also a statistically significant decent predictor of GDP per capita, with a low p-value of 9.00e-05. However, v2x_polyarchy is not a statistically significant predictor of GDP per capita, with a high p-value of 0.411828.

```
##
## Attaching package: 'olsrr'

## The following object is masked from 'package:datasets':
##
##     rivers

## Loading required package: carData
```





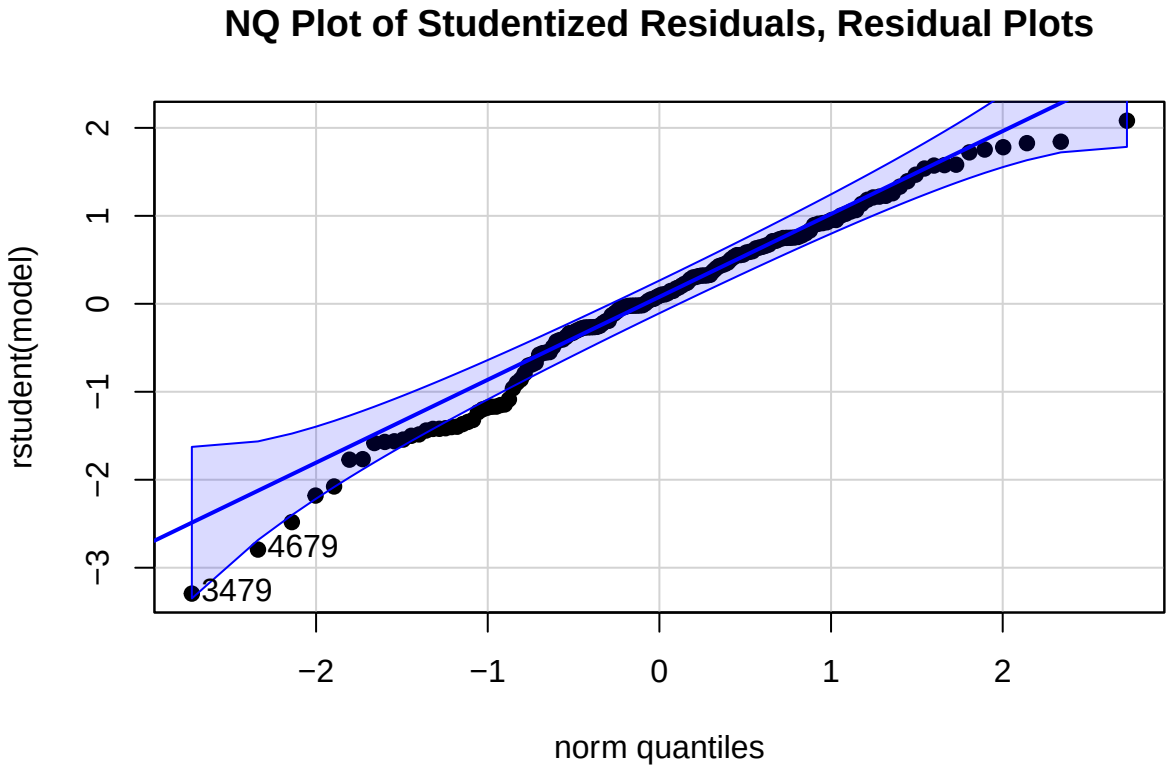
The next step is to check if the conditions for normality and low heteroskedasticity are met. By creating the normal quantile plot of studentized residuals, it is evident that the conditions for normality are not met due to the presence of high outliers. Furthermore, there is high heteroskedasticity as when fitted values increases, there is a greater spread in the studentized residuals. Thus we must take a log transformation of GDP per capita and redo the previous steps.

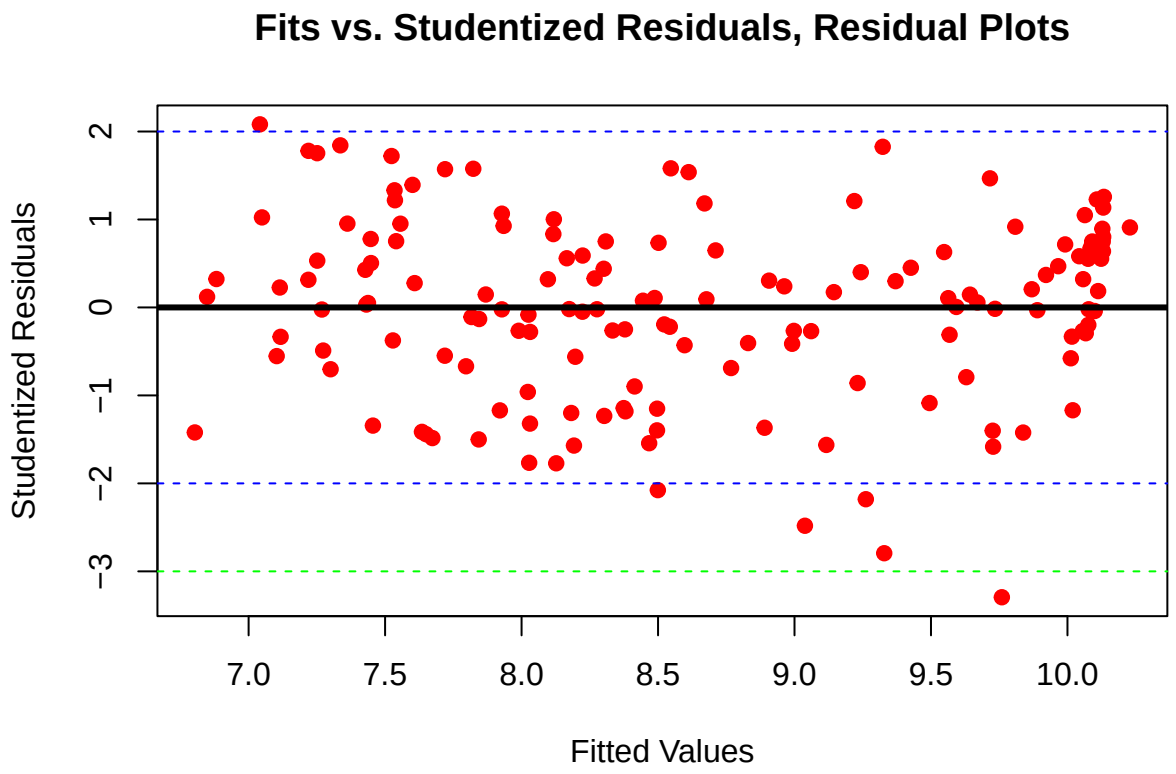
```
## Subset selection object
## Call: regsubsets.formula(log_gdppc ~ ., data = fdi_mr_2)
## 3 Variables  (and intercept)
##                Forced in Forced out
## v2x_polyarchy     FALSE      FALSE
## v2x_rule          FALSE      FALSE
## ka_open           FALSE      FALSE
## 1 subsets of each size up to 3
## Selection Algorithm: exhaustive
##          v2x_polyarchy v2x_rule ka_open
## 1  ( 1 ) " "           " "      "*"
## 2  ( 1 ) " "           "*"      "*"
## 3  ( 1 ) "*"           "*"      "*"

##   (Intercept) v2x_polyarchy v2x_rule ka_open
## 1        TRUE         FALSE    FALSE    TRUE
## 2        TRUE         FALSE     TRUE    TRUE
## 3        TRUE          TRUE     TRUE    TRUE

## [1] 3
```

```
##
## Call:
## lm(formula = log_gdppc ~ ., data = bsTemp2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.11065 -0.54254  0.09071  0.69665  1.99693
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     6.7394     0.1961  34.375  < 2e-16 ***
## v2x_polyarchy  -0.2775     0.5568  -0.498    0.619
## v2x_rule        2.0100     0.4840   4.153 5.47e-05 ***
## ka_open         1.6403     0.2565   6.395 1.90e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9839 on 151 degrees of freedom
## Multiple R-squared:  0.5248, Adjusted R-squared:  0.5153
## F-statistic: 55.58 on 3 and 151 DF,  p-value: < 2.2e-16
```

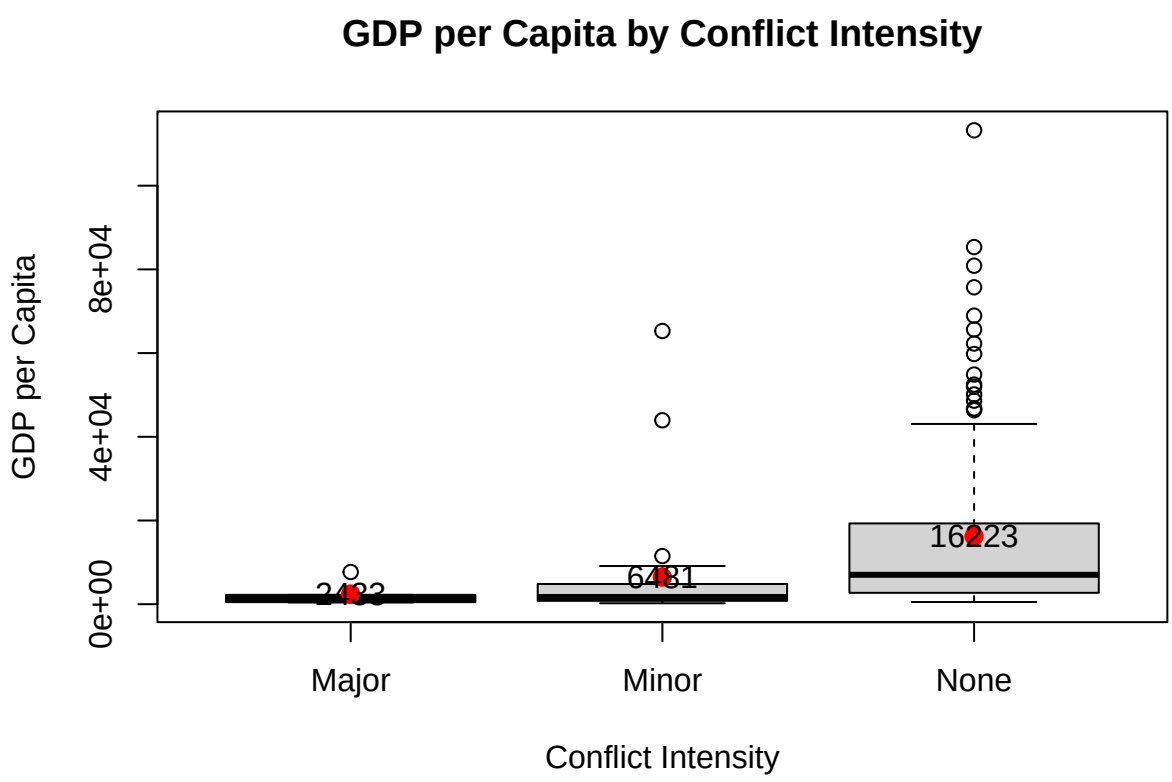




Much better. By taking a log transformation, the points in the normal quantile plot fall in roughly a straight line within the blue upper and lower bounds. Furthermore, the fits vs. studentized residuals plot does not show evidence of heteroskedasticity as the range of studentized residuals is about the same as fitted values increases. Looking at the best subsets results, the complete model with all predictors is the most optimal. After the log transformation, it is still the case that ka_open and v2x_rule are statistically significant predictors of the log of GDP per capita, in this case with very low p-values of 5.47e-05 and 1.90e-09 respectively, whereas v2x_polyarchy is not a statistically significant predictor of log of GDP per capita. The positive coefficients for both v2x_rule and ka_open suggest that as rule of law becomes more transparent and as openness to world trade is higher, GDP per capita is expected to increase as well.

ANOVA

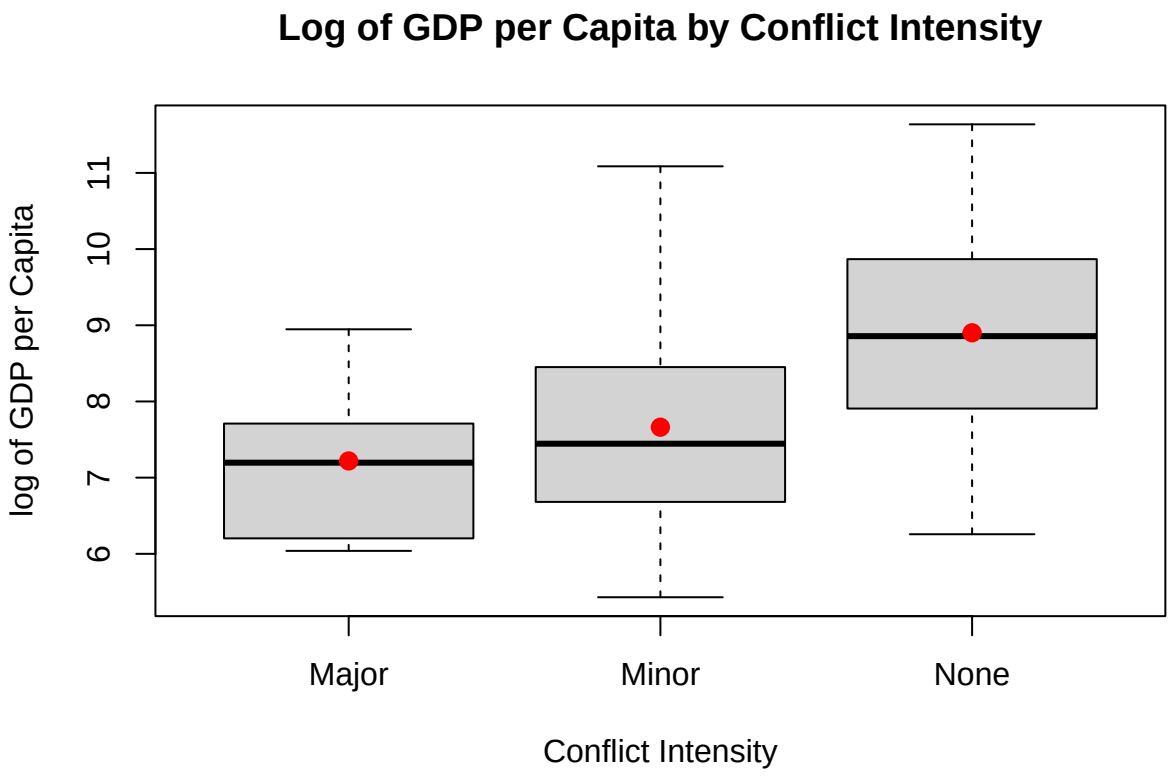
For ANOVA, we decided to analyze whether there is a difference in mean GDP per capita among groups of countries separated by differing levels of conflict intensity. Conflict intensity is a categorical variable that represents the number of deaths as a result of armed conflict. conflict_intensity is 0 (none) if there is no conflict, 1 (minor) if there are between 25 and 999 annual deaths, and 2 (major) if there are 1000 or more.




```
##     Major     Minor      None
##  3027.183 14196.640 20926.709
```

```
## [1] 6.9
```

First step is to see if the assumptions of normality within groups is met, and the ratio of the highest and lowest standard deviations among the groups is less than 2. From the visual information provided by the boxplots, the distribution of GDP per capita is definitely not normal for the conflict levels. There are way too many high outliers for each group. Furthermore, by visual inspection, the countries with no conflict seems to have more than twice the standard deviation of the countries with major conflict. This is confirmed by calculating the ratio of highest to lowest standard deviation among groups, which is 6.9, which is much greater than 2. Hence, a transformation is necessary. We proceed with a transformation: taking the natural log of GDP per capita and using that in place of GDP per capita.




```
##    Major    Minor     None
## 1.188813 1.358691 1.337947
```

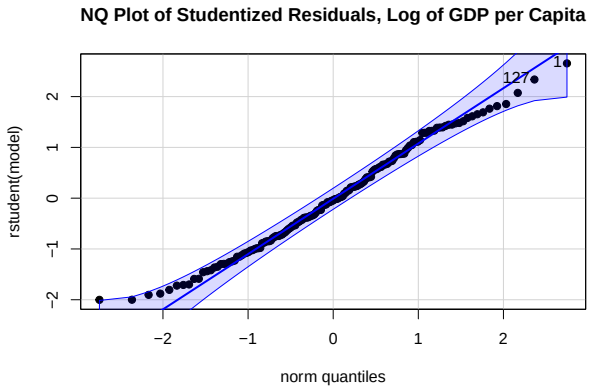
```
## [1] 1.1
```

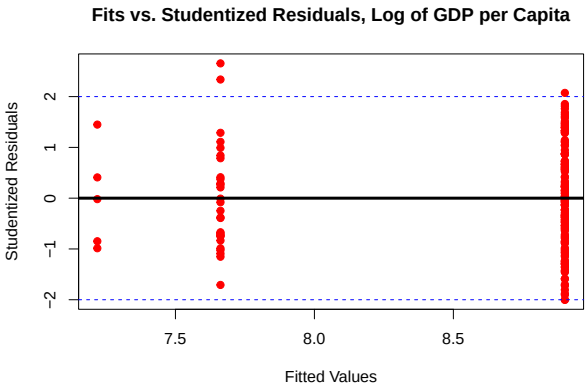
Much better. The ratio of highest to lowest standard deviation is 1.1, which is less than 2. Also, the groups seem to be normally distributed by looking at the boxplots. We proceed with ANOVA.

```
## Call:
##    aov(formula = fdi_anova$log_gdppc ~ fdi_anova$conflict_intensity)
##
## Terms:
##                 fdi_anova$conflict_intensity Residuals
## Sum of Squares                      45.94603 291.78985
## Deg. of Freedom                            2       163
##
## Residual standard error: 1.337954
## Estimated effects may be unbalanced
```

```
##                               Df Sum Sq Mean Sq F value   Pr(>F)
## fdi_anova$conflict_intensity   2  45.95   22.97   12.83 6.67e-06 ***
## Residuals                    163 291.79    1.79
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

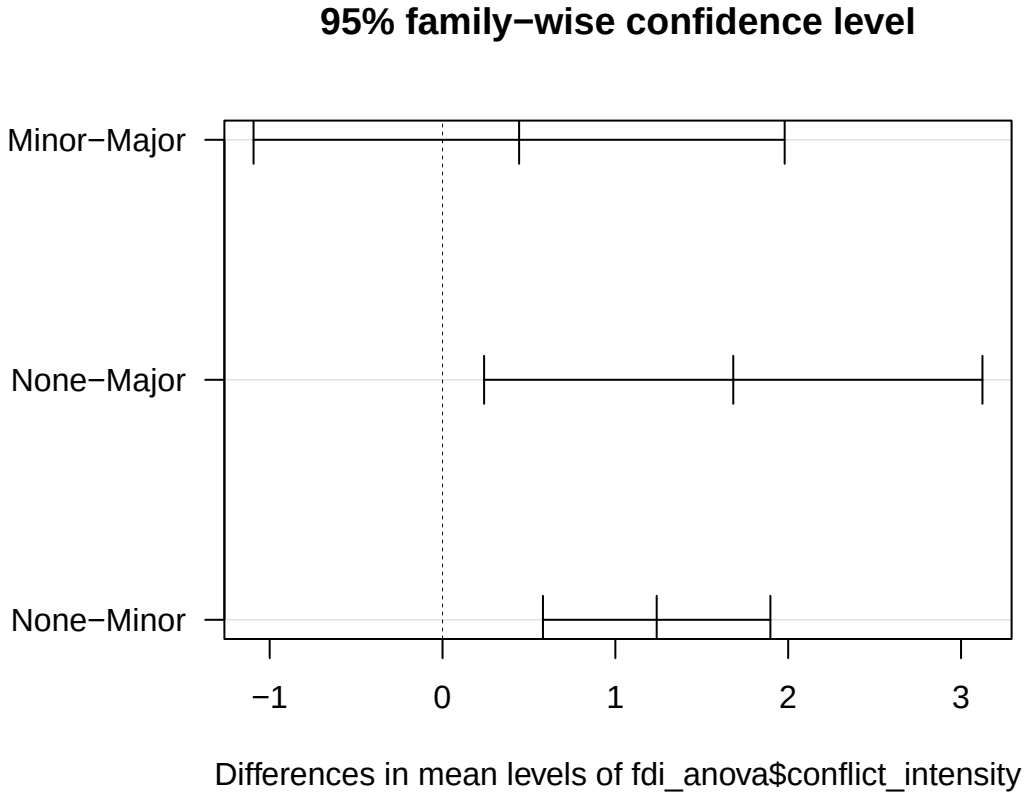
The log of GDP per capita is statistically significantly different across conflict intensity groups, as evidenced by the extremely low p-value of 1.71e-06 of ANOVA.





With the log transformation, the points are approximately normally distributed as they fall on a line in the normal quantile plot. In addition the fits vs. residuals plots shows no evidence of heteroskedasticity, as the residuals are evenly distributed about 0 and there is no significant increase in spread as fitted values increases. Therefore, our ANOVA results are valid, and there is a statistically significant difference in log of GDP per capita among groups of countries with differing degrees of conflict.

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = fdi_anova$log_gdppc ~ fdi_anova$conflict_intensity)
##
## $`fdi_anova$conflict_intensity`
##                   diff        lwr      upr     p adj
## Minor-Major 0.4431501 -1.0933202 1.979620 0.7741907
## None-Major  1.6822420  0.2405917 3.123892 0.0176099
## None-Minor  1.2390919  0.5810716 1.897112 0.0000462
```



We proceed with a Tukey confidence interval plot. Based on the diagram above, there is a statistically significant difference in log of GDP per capita between countries with no conflict and major conflict, and countries with no conflict and minor conflict, as the confidence interval for the difference in means of those comparison groups does not intersect 0. There, however, is not a statistically significant difference between countries with minor and major conflict, as the confidence interval for the difference of means of minor-major intersects 0.

Conclusion and Summary