Final Project

Aidan Gannon, Jackie Dong, Zach Brown

Introduction

For our final project we decided to provide analysis of World Bank countries, in particular with regards to GNI per capita. We wanted to see the relationship that GNI per capita has with some different variables, focusing both on economic factors and social factors. Ultimately, our findings were illuminating and provide an interesting perspective to take into account when considering future policy.

The libraries we use:

```
library(rvest,     quietly = TRUE, warn.conflicts = FALSE)
library(dplyr,     quietly = TRUE, warn.conflicts = FALSE)
library(tidyverse, quietly = TRUE, warn.conflicts = FALSE)
library(ggplot2,   quietly = TRUE, warn.conflicts = FALSE)
library(car,       quietly = TRUE, warn.conflicts = FALSE)
library(ggpubr,    quietly = TRUE, warn.conflicts = FALSE)
library(corrplot,  quietly = TRUE, warn.conflicts = FALSE)
library(leaps,     quietly = TRUE, warn.conflicts = FALSE)
library(MASS,      quietly = TRUE, warn.conflicts = FALSE)
library(lindia,    quietly = TRUE, warn.conflicts = FALSE)
library(patchwork, quietly = TRUE, warn.conflicts = FALSE)
```

Data

The data we included in our analysis revolves mostly around economic factors, though we included some other interesting variables. We looked at different countries' GNI (Gross National Income) per capita in US dollars. At its most basic, GNI is used to track a nation's wealth year over year. Here are the variables we considered as potential predictors of GNI:

- Rural: percentage of total people living in rural areas, continuous variable
- LifeExp: life expectancy at birth (years), continuous variable
- CO2: carbon dioxide emissions (metric tons per capita), continuous variable
- Diesel: diesel fuel pump price (US$ per liter), continuous variable
- NATO: whether the country is a member of NATO, categorical variable
- HappinessScore: a measure of the happiness of people in a country, continuous variable
- Continent: what continent the country is on, categorical variable
- PressFreedom: measure of how free the press of a country is, categorical variable
- Inequality: measure of wealth inequality in a country, continuous variable

We then web-scaped five of the columns for our dataset. We scrapped the `inequality` column from Wikipedia. It contains scores of the distribution of wealth in countries with a higher score corresponding to greater inequality in wealth distribution. Another added was `NATO`, a categorical variable that denoted whether the country was in NATO. To get that data, we scraped a list of NATO countries from Wikipedia. Next, we added a tiered categorical variable, `PressFreedom`. To acquire this data, we scraped a ranking from here. We decided to include happiness data, column named `HappinessScore`, for each country: how they rank in terms of total happiness and their happiness score. Finally, we added the `Continent` column.

```
wb <- read.csv("http://www.reuningscherer.net/s&ds230/data/WB.2016.csv",
header = TRUE, as.is = TRUE)

url <- "https://en.wikipedia.org/wiki/List_of_countries_by_wealth_inequality"
webpage <- read_html(url)
inequalityIndex <- html_nodes(webpage, 'tr+ tr td:nth-child(4) , .table-na+
td:nth-child(4)')

url2 <- "https://en.wikipedia.org/wiki/Member_states_of_NATO"
webpage2 <- read_html(url2)
memberNATO <- html_text(html_nodes(webpage2, 'td:nth-child(3)'))

url3 <- "https://www.nationsonline.org/oneworld/press_freedom.htm"
webpage3 <- read_html(url3)
pressFreedom <- html_text(html_nodes(webpage3, 'td a'))

Happiness <- read.csv("https://docs.google.com/spreadsheets/d/e/2PACX-
1vQcZs4E3jhQdZEyfUJQTQ3ogqJ2YmHYZ1dCtVH3Xhi_DkbOfEzuMi7FBSPvteL6I4fOcIJWzukBB
EMl/pub?gid=1586447313&single=true&output=csv")
```

Data Cleaning

In this section, we cleaned the data from the additional columns that we web-scraped as well as for the World Bank data itself. For many observations in the World Bank data, the country name was written in strange formats different from the country names from the web-scraped data, so we cleaned the World Bank country names to match our other data.

For inequality, the cleaning involved using gsub to remove trailing/leading white space. Once the data was cleaned, we merged the data frame–consisting of the inequality score and country name–with the World Bank dataset. We then cleaned our NATO data, again using gsub and then merging our data frame with whether a country was in NATO or not with the World Bank data. For press freedom, after getting the rankings we separated countries into three categories–good, mediocre, and bad press freedom. Once the categories were created, we also merged that data with the World Bank data. Happiness was already fairly organized, so we simply renamed some of the columns for easier access and then merged that data into our final dataframe.

Finally, the HappinessScore data came from a csv online, so it was already cleaned. The Happiness dataset also came with a column that labeled each country's region. We cleaned that column to create the Continent column by using gsub to change regions into continents. After the data was all cleaned, we selected what specific columns we wanted and created a new dataset called wb_final with those columns.

```
# Cleaning wb$Country to make dataframe merging easier
wb$Country <- gsub(" Darussalam", "", wb$Country)
wb$Country <- gsub("Syrian Arab Republic", "Syria", wb$Country)
wb$Country <- gsub("Macedonia", "North Macedonia", wb$Country)
wb$Country <- gsub("Timor-Leste", "East Timor", wb$Country)
wb$Country <- gsub("Lao PDR", "Laos", wb$Country)
wb$Country <- gsub("Kyrgyz Republic", "Kyrgyzstan", wb$Country)
wb$Country <- gsub("Russian Federation", "Russia", wb$Country)
wb$Country <- gsub("Congo, Dem. Rep.", "DR Congo", wb$Country)
wb$Country <- gsub("Korea, Rep.", "South Korea", wb$Country)
wb$Country <- gsub("Korea, Dem. People\x92s Rep.", "North Korea", wb$Country)
wb$Country <- gsub("Great Britain", "United Kingdom", wb$Country)
wb$Country <- gsub(",.*", "", wb$Country)

# Data Cleaning for inequality
inequalityIndex <- as.numeric(html_text(inequalityIndex))
countries <- html_text(html_nodes(webpage, 'td:nth-child(1)'))
countries <- countries[1:181]
countries <- gsub("\\..*", "", countries)
countries <- gsub("[(].*", "", countries)
countries <- gsub("^\\s+|\\s+$", "", countries)
countries <- gsub("Great Britain", "United Kingdom", countries)
inequality <- data.frame(Country = countries,inequality = inequalityIndex)
wb_new1 <- merge(wb, inequality, by = "Country")
```

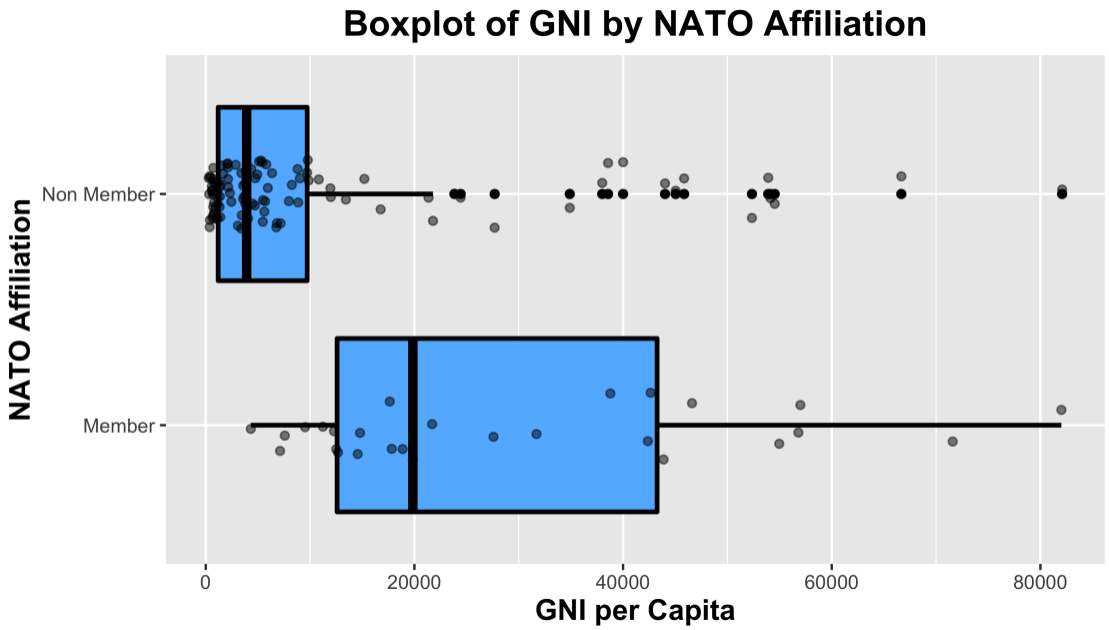
In the previous section, we cleaned the Country column in the World Bank data set and cleaned and merged the inequality data with the World Bank data. We won't show the rest of the data cleaning, but it's in the Rmd file.

```
## [1] 132  12
```

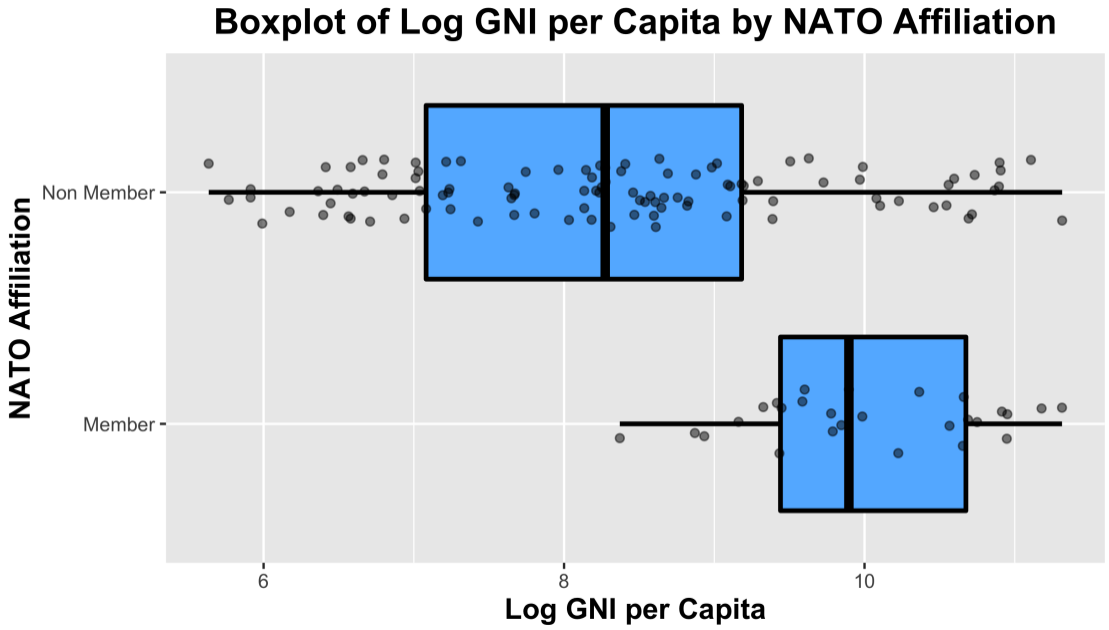
Our final data frame consists of 132 observations with 11 variables each.

Graphics

First, to get a feel for the data, we will look at boxplots of GNI per capita by NATO affiliation.

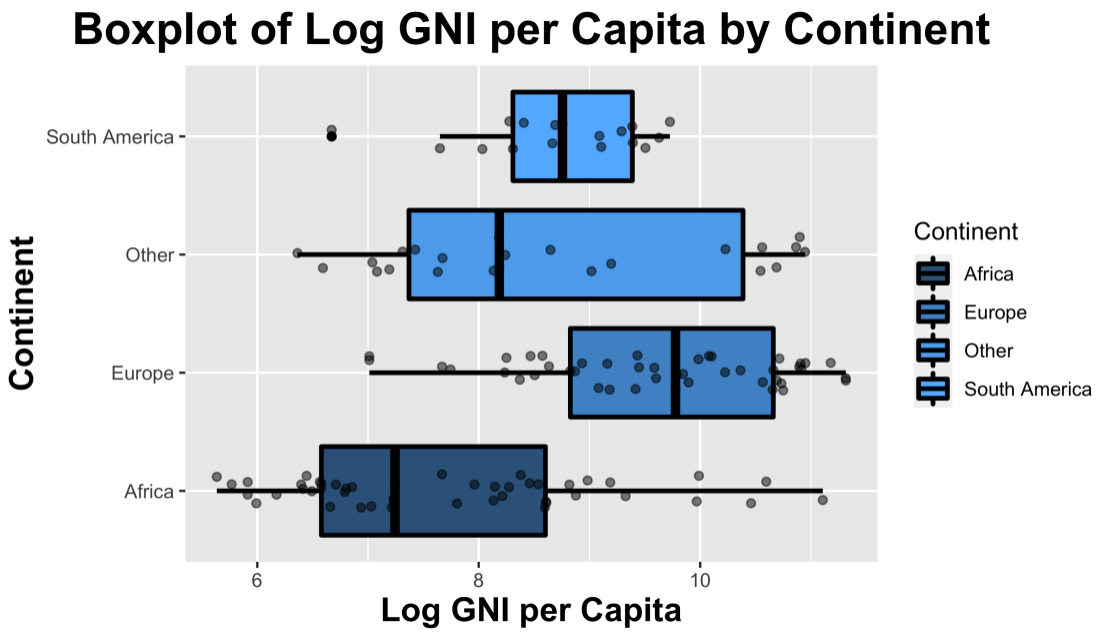


This plot is very right skewed, and the extreme skew makes differences between the two groups hard to see, although it seems that non-NATO countries potentially have higher rates of gun deaths. In order to get a better feel for the data, we'll look at the same boxplot but with log GNI.



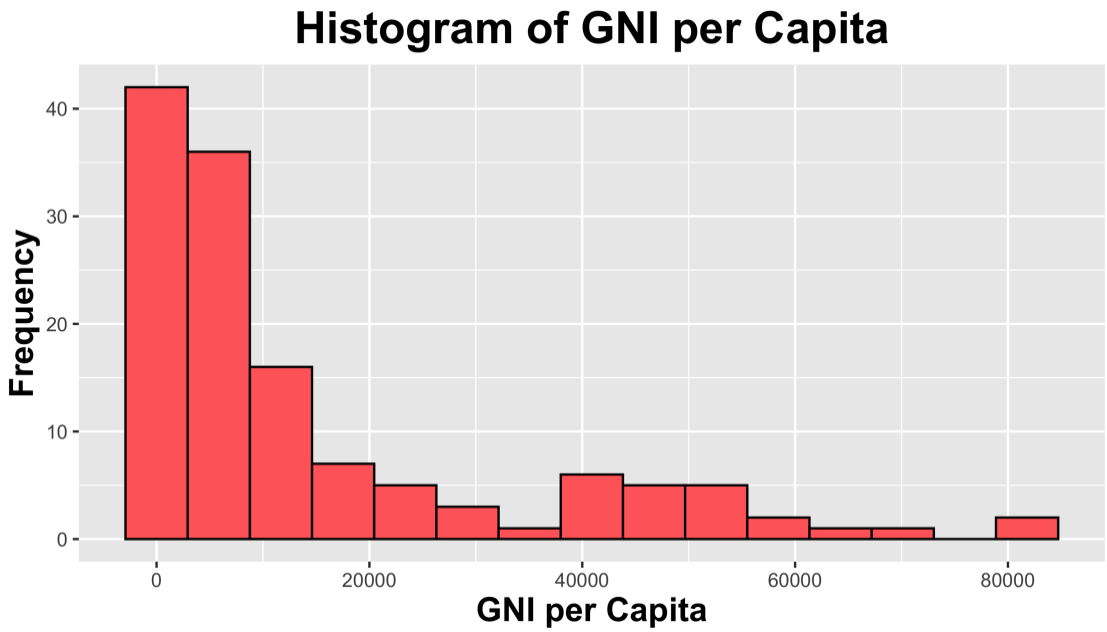
In this plot, the differences between the two groups is much easier to see. It looks like NATO member countries tend to have greater GNI per capita than non-NATO countries.

We'll also look at a boxplot of log GNI per capita by continent, just to get a better sense of how data vary geographically.

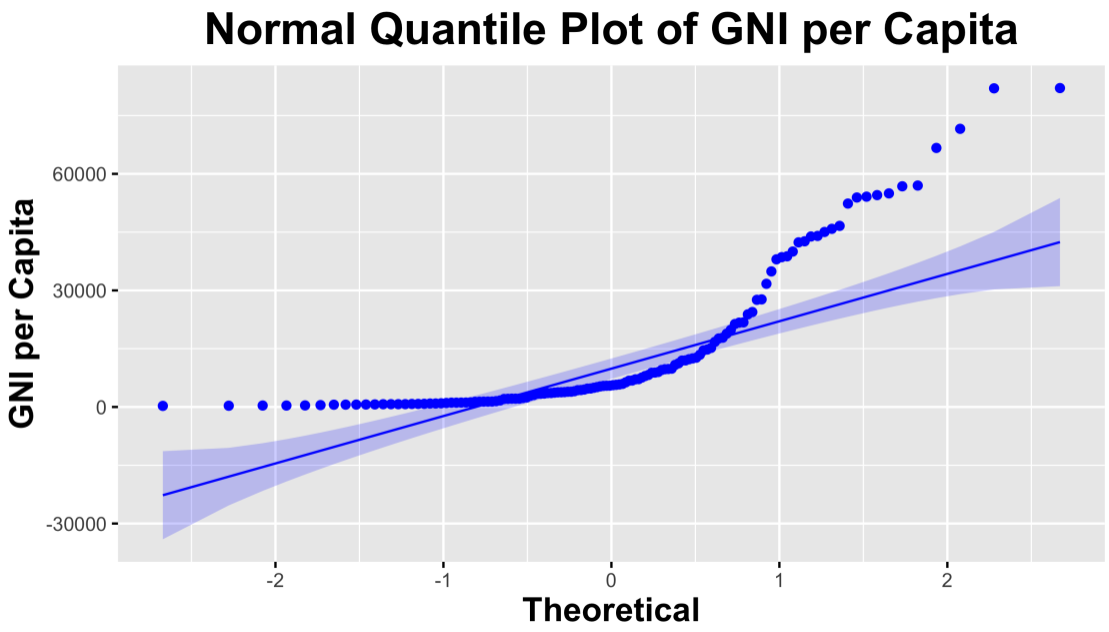


From this boxplot, we see that European countries tend to have a greater GNI per capita than countries on other continents.

Now, let's look at a histogram of GNI per capita.



This histogram of GNI per capita shows a heavy right skew which we will need to correct for later on. The second plot, a normal quantile plot of the same data, confirms this as the data is definitely not normally distributed and indicates a right skew. We'll look at normal quantile plot to test the normality of the distribution of GNI per capita.



The normal quantile plot confirms what we had suspected: the distribution of GNI per capita is not normal.

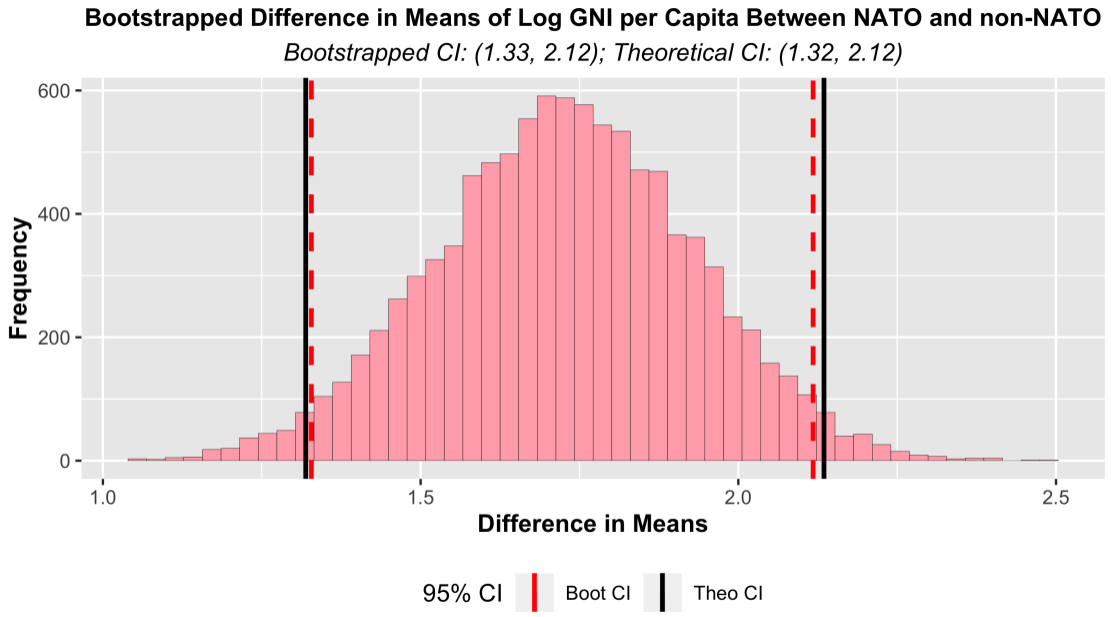
Analysis

Basic Tests

T-test and Bootstrap

Here, we will perform a t-test to compare the mean GNI per capita between NATO countries and non-NATO countries.

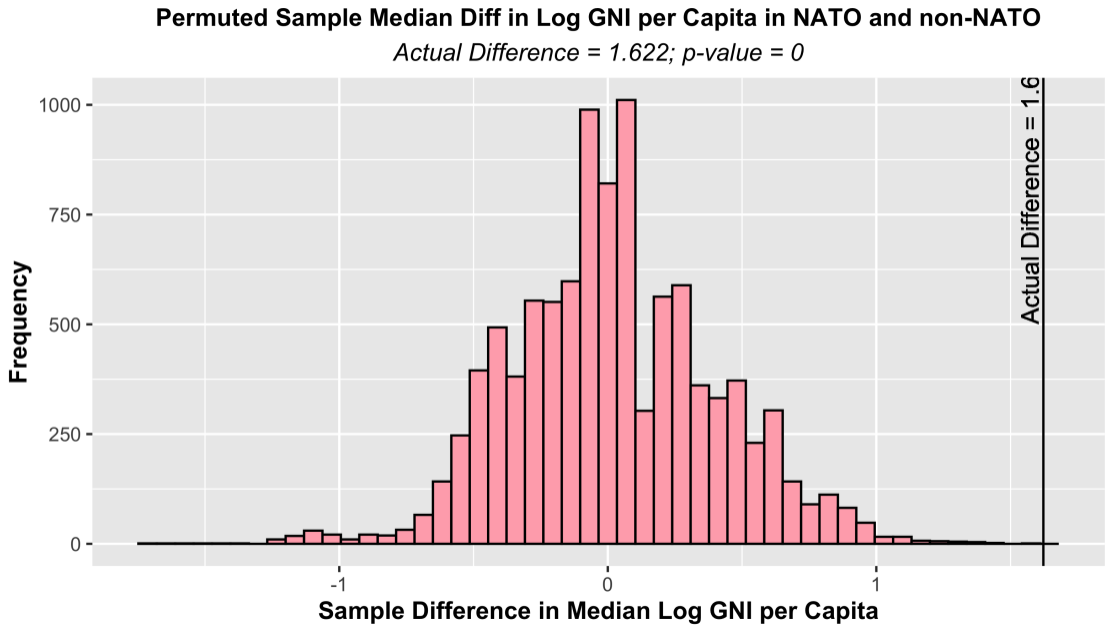
```
##
##  Welch Two Sample t-test
##
## data:  log(wb_final$GNI) by wb_final$NATO
## t = 8.4342, df = 76.116, p-value = 1.604e-12
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.319217 2.134844
## sample estimates:
##     mean in group Member mean in group Non Member
##                10.024256                 8.297225

##  2.5% 97.5%
##  1.33  2.12

## [1] 1.32 2.13
## attr(,"conf.level")
## [1] 0.95
```



The t-test and the confidence intervals have the same conclusion: we can reject the null hypothesis that there is no difference between the mean log rate of gun deaths per 100,000 people between NATO and non-NATO countries. The t-test had a p-value of 0.000000000001604, which is less than our significance level of 0.05. As seen on the plot, both the theoretical 95% CI, taken from the t-test, and the bootstrapped 95% CI had similar bounds. For both confidence intervals, the lower bound is greater than 0, again showing that we can reject the null hypothesis that the difference in means between the groups is 0.

Permutation Test

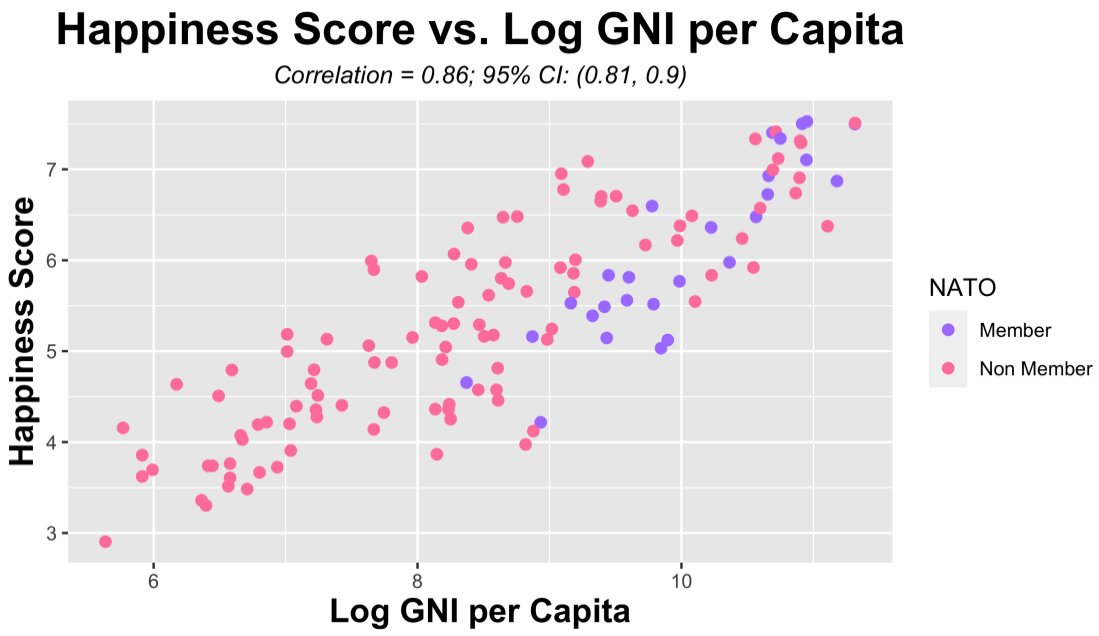
Here, we perform a permutation test to compare the median log GNI per capita across NATO and non-NATO countries.



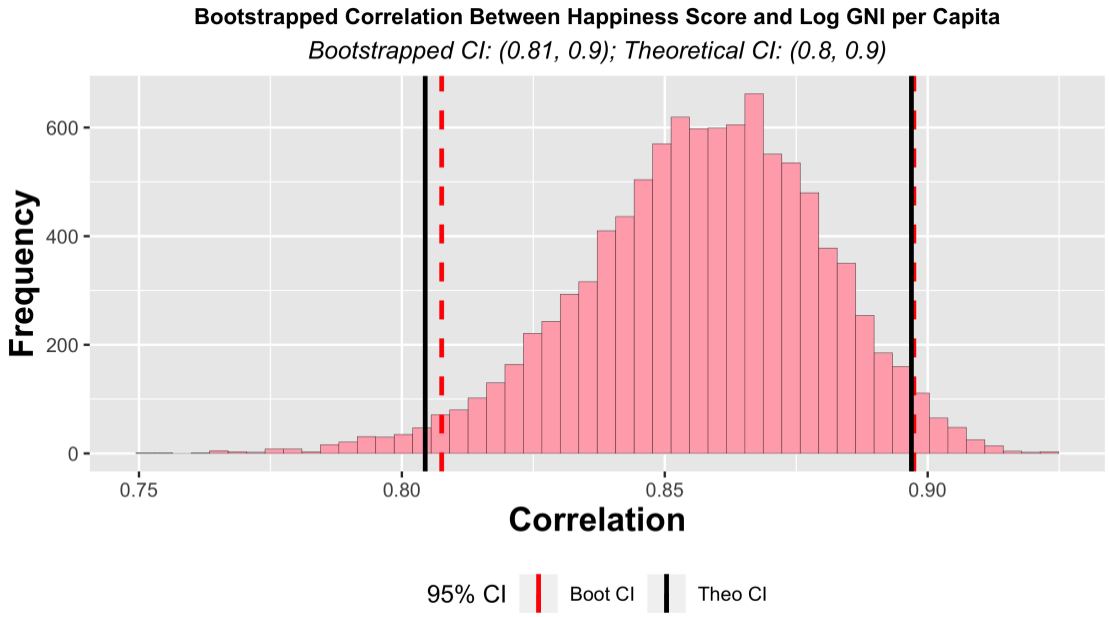
The null hypothesis is that there is no difference between the median GNI per capita for NATO countries and non-NATO countries in 2016. From the permutation test, we calculated the p-value to be 0, which is less than our alpha level of 0.05, meaning that we can reject the null hypothesis. This rejection means that the actual difference in medians (1.622) is not probable given the null hypothesis is true. This finding is in line with what we found in the t-test about the difference in mean log GNI per capita between NATO and non-NATO countries. In both tests, we rejected the null hypothesis.

Correlation

Now, we'll look at the correlation between log GNI per capita and the happiness score.



In this plot, we examined how related a country's happiness score is to its log GNI per capita. We were curious to see this relationship because statistics like GNI and GDP are often criticized for their ubiquity since they are measures of production/income not well-being. The happiness score takes into account these "well-being" metrics such as social support, freedom, life expectancy, and education level. This strong correlation (0.86) between log GNI per capita and happiness score seems to suggest that high production/income is a good indicator of these other gauges that are more related to the happiness and well-being of people. We chose to color the points by NATO status just to get a visual sense of how NATO countries compare to non-NATO countries in terms of happiness and GNI.



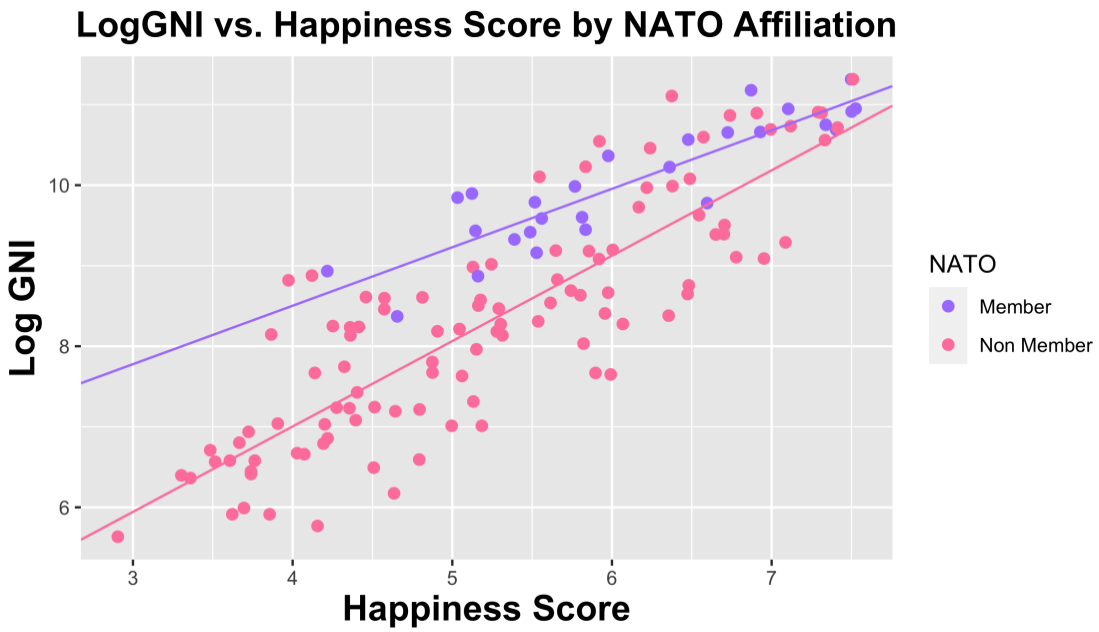
The null hypothesis is that the true correlation between log GNI per capita and Happiness score is 0. In both the cor-test and the confidence intervals we can reject the null hypothesis.. The cor-test had a p-value of 2.2e-16, which is less than our significance level of 0.05. As seen on the plot, both the theoretical 95% CI, taken from the cor-test, and the bootstrapped 95% CI had similar bounds, though the bootstrap CI was slightly smaller. For both confidence intervals, the lower bound is greater than 0, again showing that we can reject the null hypothesis that the difference in means between the groups is 0.

Ancova

After concluding that the median and mean LogGNI per capita between Nato members and non-members is statistically significantly different, we decided to perform an ANCOVA analysis of the interaction effect between Happiness Score and NATO affiliation when predicting the LogGNI of countries. We then visualize our results using a scatter plot and lines of best fit generated from the summary information.

```
##
## Call:
```

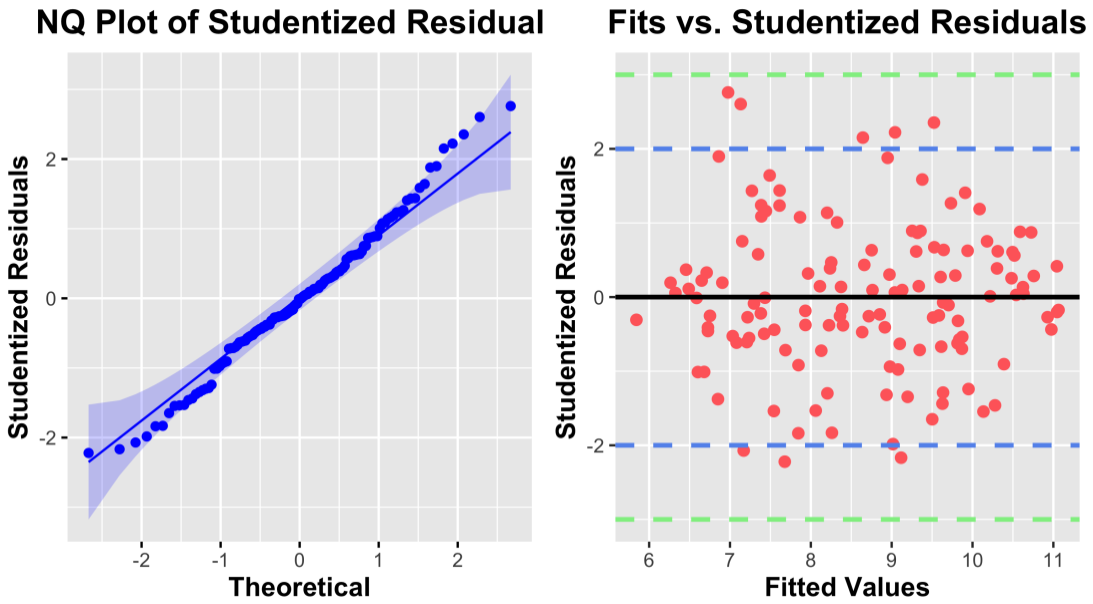
```
## lm(formula = LogGNI ~ HappinessScore + NATO + NATO * HappinessScore,
##     data = wb_final)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.50376 -0.39562 -0.00696  0.42147  1.84233
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      5.5991     0.8712   6.427 2.34e-09 ***
## HappinessScore                   0.7261     0.1413   5.140 1.00e-06 ***
## NATONon Member                  -2.8377     0.9266  -3.063  0.00268 **
## HappinessScore:NATONon Member    0.3345     0.1531   2.185  0.03074 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6915 on 128 degrees of freedom
## Multiple R-squared:  0.7904, Adjusted R-squared:  0.7855
## F-statistic: 160.9 on 3 and 128 DF,  p-value: < 2.2e-16

##                   (Intercept)                 HappinessScore
##                     5.5990721                     0.7261409
##                NATONon Member HappinessScore:NATONon Member
##                    -2.8377067                     0.3345235
```



From our summary statistics of our generalized linear model predicting the Log GNI per capita of countries from Happiness Score, NATO, and the interaction between Happiness Score and NATO, we can see that all three predictive coefficients are statistically significant in our model, as their p-values (1.00e-06, 0.00268, 0.03074 respectively) are all less than our alpha level of 0.05. Additionally, from our summary statistic we can also see conclude

that this is a model of strong fit as our R-squared value is 0.79, which indicates that 79% of the variance in our y-variable (LogGNI per Capita) is accounted for by our regression model.

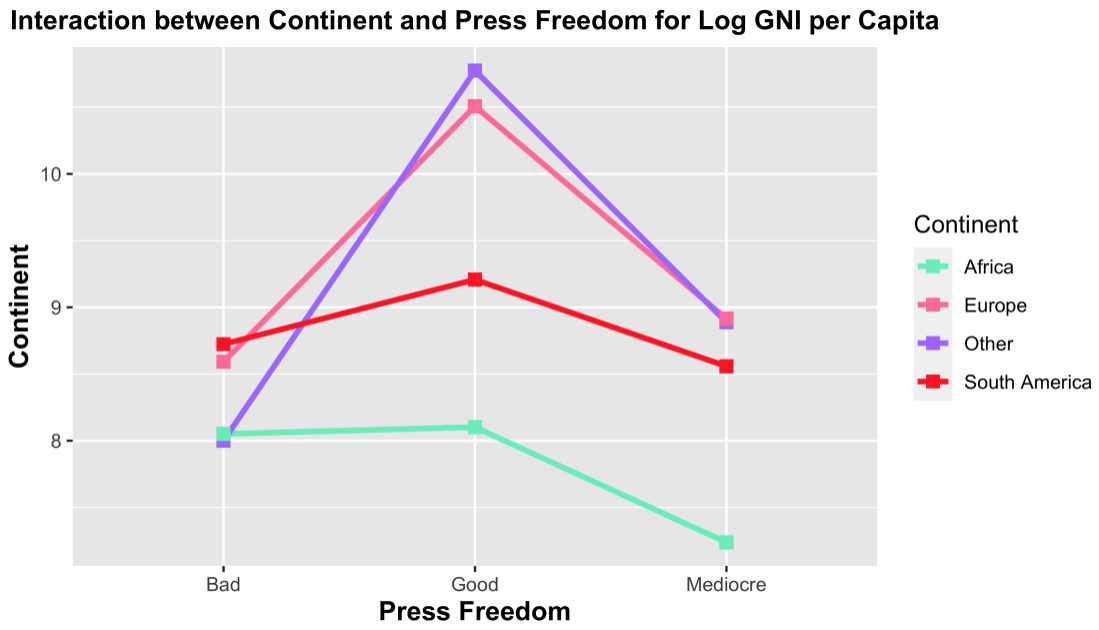
From our coefficients, we can analyze the association between our predictors variables and response variables. Our Happiness Score coefficient tells us that for every one increase in Happiness Score, our predicted Log GNI per capita increases by 0.7621. Our Nato Non Member term indicates that if our country is not a member of NATO, there is a shift in our regression line of -2.84 Log GNI per capita. The interaction term (HappinessScore and non-NATO) has a coefficient of 0.3345, meaning that if a country is not in NATO, increases in their happiness score correspond to greater increases in log GNI per capita compared to countries that are in NATO.



We finally checked if our ANCOVA model assumptions were met using residual plots. Our normal quantile plot of studentized residuals appears to be approximately linear, indicating that our assumption of our distribution of errors coming from an approximately normal distribution is satisfied. Our fits vs. residuals plots also displays no heteroskedasticity as the spread across fitted values seems to be approximately the same. Lastly, there is no evidence of a non-linear pattern in our plot.

Two-way ANOVA

Here, we examine the influence Continent and Press Freedom have on log GNI per capita. The goal is to assess the effects of each variable and determine if there is any interaction between them. We begin with an interaction plot.



From this plot, there does seem to be some interaction between continent and press freedom. The relative mean log GNI per capita based on continent seems to change depending on the level of press freedom. For example, the Other continent (consisting of North America, some of Asia, and Oceania), has the lowest log GNI per capita with Bad press freedom but the highest with Good press freedom.

Now, we will fit a two-way ANOVA model to see if there are significant interaction terms.

```
## Anova Table (Type III tests)
##
## Response: log(wb_final$GNI)
##                                          Sum Sq  Df   F value  Pr(>F)
## (Intercept)                             1555.58   1 1304.4519 < 2e-16 ***
## wb_final$Continent                         2.92   3    0.8168 0.48699
## wb_final$PressFreedom                      7.70   2    3.2287 0.04307 *
## wb_final$Continent:wb_final$PressFreedom  19.65   6    2.7459 0.01551 *
## Residuals                                143.10 120
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the summary, we see that the interaction of Continent and PressFreedom is statistically significant (p-value 0.01551). This finding is consistent with what we found in the interaction plot, but it's interesting to note that according to the two-way ANOVA, Continent is not a statistically significant term.

To save space, we did not include the residual plots. The normal quantile plot is approximately linear, which is reasonable given our assumption of normally distributed errors. The plot of fits vs. studentized residuals has multiple outliers though, which isn't ideal. There does not appear to be much heteroskedasticity on the residual plots, although

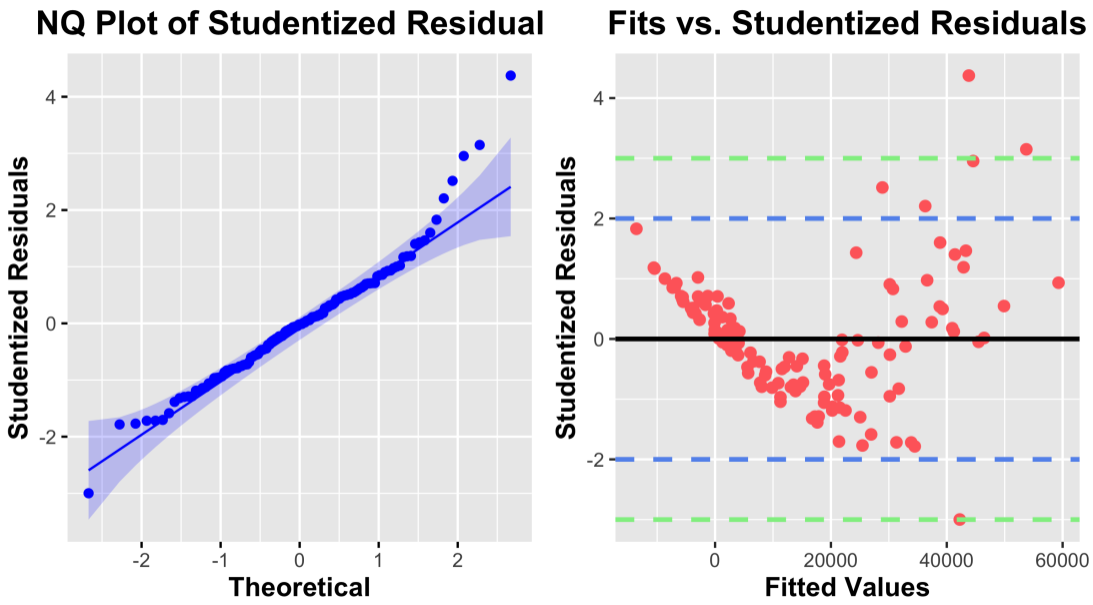
the variance of the residuals seems to decrease as the fitted values increase, suggesting a not equal standard deviation for all the residuals.

Multiple Linear Regression

Finally, we fit a best subsets model based on the Bayesian Information Criterion analysis method. The goal was to see which variables were best at predicting GNI per capita. We chose the best subsets model because we wanted to examine many different possible models and choose the best from there. We used Bayesian Information Criteria because we used that one on homeworks before and because it has a larger penalty for extra predictors than the Akaike information criterion.

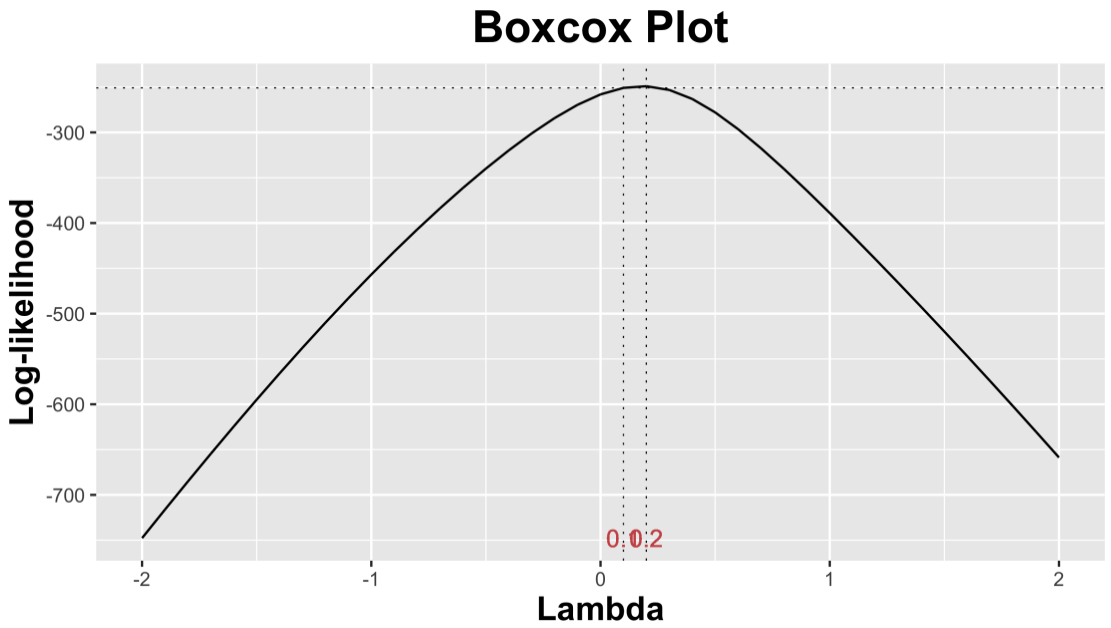
```
## [1] "HappinessScore" "CO2"            "Diesel"

##
## Call:
## lm(formula = GNI ~ ., data = wb_final2temp)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -25481  -6908   -172   5090  38271
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -52839.7     4357.6 -12.126  < 2e-16 ***
## HappinessScore   7578.2      878.2   8.629 2.06e-14 ***
## CO2              1207.2      152.3   7.926 9.63e-13 ***
## Diesel          23656.8     2803.0   8.440 5.86e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9614 on 128 degrees of freedom
## Multiple R-squared:  0.7514, Adjusted R-squared:  0.7456
## F-statistic:   129 on 3 and 128 DF,  p-value: < 2.2e-16
```

From our summary statistics of our fitted model based on the BIC criteria, we can see that the three variables that ended up as significant predictors of GNI per capita were Happiness Score, CO2, and Diesel, as their p-values (2.06e-14, 9.63e-13, 5.86e-14 respectively) are all less than our significance level of .05. Let's check our model assumptions now with some **awesome** residual plots!



Despite a few outliers exceeding a studentized residual of 3, our normal quantile plot for our studentized residuals appears approximately linear, indicating that our model assumption of our model errors coming from an approximately normal distribution is satisfactory. However, there are some evident issues of heteroskedasticity in our plots of fits vs. residuals, as there are large outliers associated with larger fitted values. The most concerning thing is not even the illustrated heteroskedasticity, but rather the non-linear pattern (potentially quadratic). This nullifies our assumption of linearity.

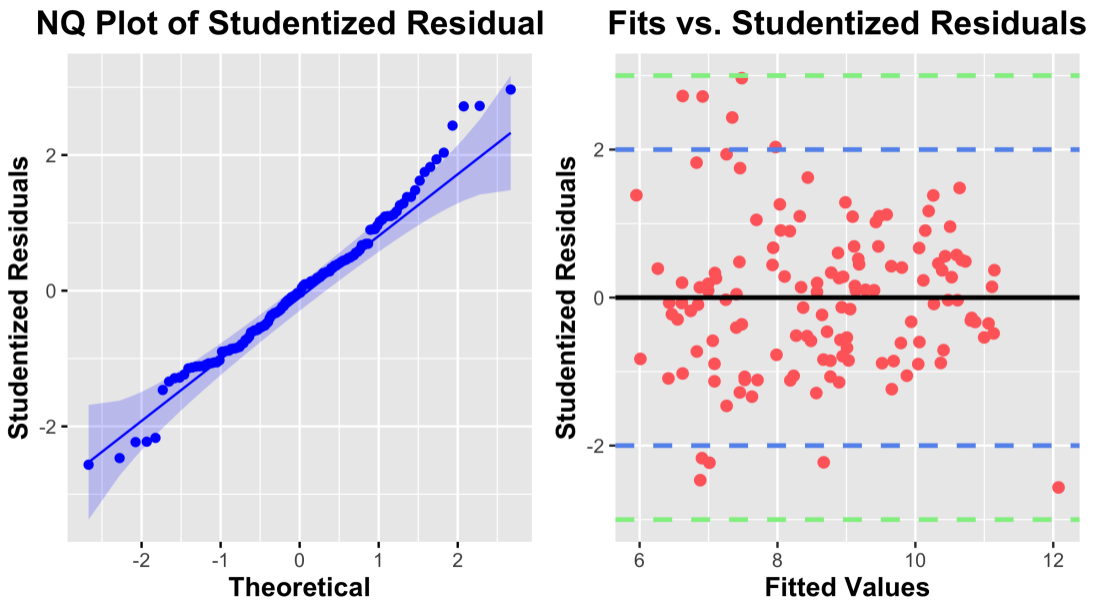
We then decided to perform a box cox transformation of our model because of the evident curvature in the fits vs. residuals plot



From the boxcox, we can see a suggested lambda of approximately 0.2. However, we would rather take a log transformation of our GNI per capita than take the fifth root. We then proceeded to perform a best subsets regression with logGNI per capita and fit the model.

```
## [1] "HappinessScore" "Rural"          "LifeExp"        "CO2"
## [5] "Diesel"

##
## Call:
## lm(formula = LogGNI ~ ., data = wb_boxcoxtemp)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.11163 -0.33156 -0.01413  0.23697  1.33659
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     1.716510   0.626218   2.741  0.00701 **
## HappinessScore  0.352703   0.064580   5.461 2.42e-07 ***
## Rural          -0.014250   0.002766  -5.152 9.68e-07 ***
## LifeExp         0.063881   0.009558   6.684 6.76e-10 ***
## CO2             0.062644   0.007639   8.200 2.36e-13 ***
## Diesel          0.740989   0.140621   5.269 5.75e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4714 on 126 degrees of freedom
## Multiple R-squared:  0.9041, Adjusted R-squared:  0.9003
## F-statistic: 237.6 on 5 and 126 DF,  p-value: < 2.2e-16
```

What we noticed is that after performing our log transformation on GNI, we gained two new significant predictors Rural and Life Expectancy. From our summary statistics of our fitted model based on the BIC criteria, we can see that the five variables that ended up as significant predictors of GNI per capita were Happiness Score, CO2, Rural, LifeExpectancy, and Disel as their p-values (2.42e-07, 9.68e-07,6.76e-10, 2.36e-13, and 5.75e-07 respectively) are all less than our significance level of .05. Also, our R-squared value has improved and is now 0.9041, indicating that 90.41% of the variance in our y-variable (Log GNI per capita) is accounted for by our regression model. This indicates that our model is extremely predictive! Let's check our model assumptions again with some **awesome** residual plots!



Following our transformation, there are no longer residuals with an absolute value of greater than 3 present in either or normal quantile plot or fits vs. residuals plot . Our normal quantile plot for our studentized residuals appears approximately linear, indicating that our model assumption of our model errors coming from an approximately normal distribution is satisfactory. There are no longer issues of heteroskedasticity in our plots of fits vs. residuals and there is no longer any evidence that there is a non-linear pattern. Our Sisyphean assumptions have been met!

Conclusion and Summary

Our analysis of GNI per capita proved to be a really interesting experiment in data cleaning and analysis. We found that Happiness Score, Life Expectancy, C02 Emissions, Diesel Fuel Pump Prices, and the Percentage of People Living in Rural Areas all significantly predict a country's GNI per capita. We were surprised some of these were included, while inequality was left out. Ultimately, this conclusion is mostly logical. As people live longer, they are more likely to work longer, which would increase the GNI of a nation. Similarly, happy people are more productive. The negative relationship between the percentage of people living in a rural area and GNI, is interesting and may speak to higher wages earned by people living in urban areas. C02 emissions, while bad for the environment, do indicate more industrial activity that contributes to GNI. Finally, diesel fuel pump price is the one predictor that is the hardest to explain. Perhaps, countries with higher prices use more diesel, which could also indicate more industrial activity. In any case, this analysis provides an interesting perspective on some of the potential relationships between GNI and socio-economic variables.