Final Project

Michael Chau and Timmy Luz

May 7, 2019

Background and Motivation:

Our goal for this analysis is to determine the relationship between several pairs of variables from the 2016 World Bank dataset on country data. Specifically, we seek to determine if there are relationships between fertility and GNI, Life expectancy and CO2 emissions, Export Ratio and Life Expectancy, and Deforestation and CO2. Overall, we believe these combinations of variables will help us answer pressing and relevant questions about our changing world. Whether this be through determining an unlikely relationship, or finding an outcome that challenges a misconception, we wanted to use our data to accrue a real understanding of the world in this day and age.

DATA VARIABLES: Country - Name of Country Fertility16 - Fertility rate of a certain country in 2016 GNI - Gross national income per capita in US dollars LifeExp - Life expectancy in years CO2 - Carbon Dioxide emissions, metric tons per capita NetExport - Percentage of GDP made up by net exports NetForest - Change in forest land area in sq kms between 1994 and 2014 HDI - Human Development Index

```
WBData <- read.csv("/Users/michaelchau/Downloads/WB.2016.csv")
WBData2 <- WBData[,c("Country", "Fertility16", "GNI", "LifeExp", "CO2", "Exports", "Imports", "Forest14", "Forest94")]
dim(WBData2)
```

```
## [1] 217   9
```

```
names(WBData2)
```

```
## [1] "Country"     "Fertility16" "GNI"         "LifeExp"     "CO2"
## [6] "Exports"     "Imports"     "Forest14"    "Forest94"
```

```
attach(WBData2)
```

```
## The following object is masked from package:datasets:
##
##     CO2
```

Creating Variables we need

```
WBData2$NetExport <- Exports - Imports
WBData2$NetForest <- Forest14 - Forest94
WBData2$Sum <- (LifeExp *.6) + (GNI *.4)
WBData2$HDI <- NA
WBData2$HDI[WBData2$Sum > 0] <- "Low HDI"
WBData2$HDI[WBData2$Sum > 600] <- "Medium HDI"
WBData2$HDI[WBData2$Sum > 1670] <- "High HDI"
WBData2$HDI[WBData2$Sum > 3000] <- "Very High HDI"
```

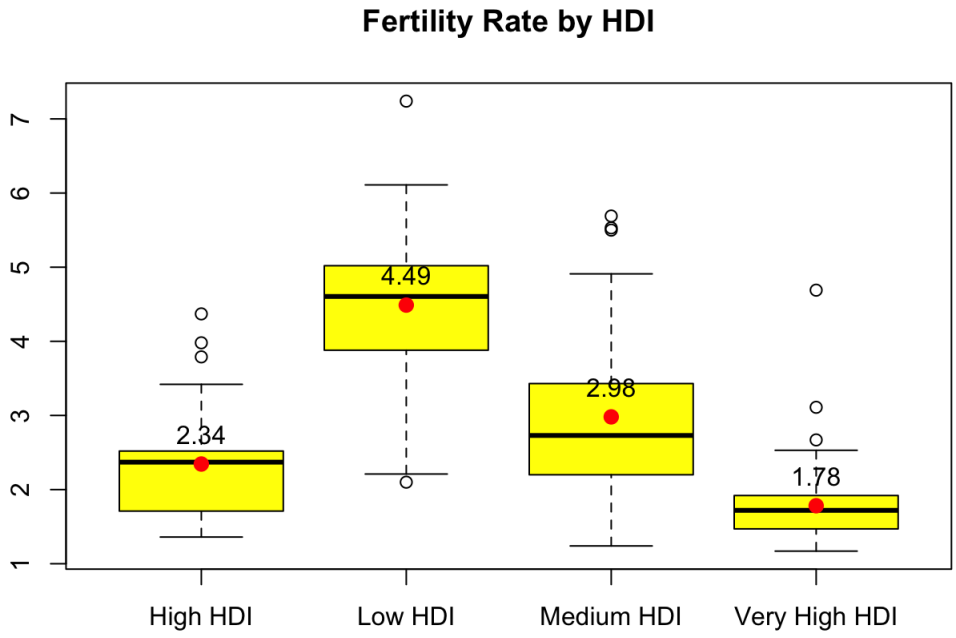
Data Cleaning

```
WBData2 <- na.omit(WBData2)
WBData2[,c(-1,-13)] <- round(WBData2[,c(-1,-13)], 2)
```

Our data cleaning process consisted of manipulating variables into more adequate ones for the hypotheses we wished to test. We first created a new data set (WBData2) that contained only the variables we wished to analyze. We next had to establish variables that were unique to our purposes. We first created a variable NetExport, exports minus imports. This, and its components, are expressed as a percentage of GDP. The intuition behind combining them was that net exports, not its components, is a factor of the economic GDP equation (GDP = Consumption + Investment + Gov Spending + Net Exports). The original data had several countries with 70 or 80 percent values in both components columns. This was deceiving as if a country that has 70% in both columns it means that zero percent of its GDP is made up by net exports. The next variable we created was NetForest, which was the difference in sq kilometers of forest between 2014 and 1994, which was useful to compare to CO2 emissions, as forests sequester carbon. Finally we created a variable for HDI, which was calculated using information from the United Nations Development Programme (UNDP). After creating new variables we omitted NAs and rounded all continuous variables.

Fertility and HDI

```
boxplot(WBData2$Fertility16 ~ WBData2$HDI, col = 'yellow', main = 'Fertility Rate by HDI')
means <- tapply(WBData2$Fertility16, WBData2$HDI, mean)
points(means, col = "red", pch = 19, cex = 1.2)
text(x=c(1:4), y=means+.4, labels = round(means,2))
```




```
sds <- tapply(WBData2$Fertility16, WBData2$HDI, sd)
round(max(sds)/min(sds),1)
```

```
## [1] 2.2
```

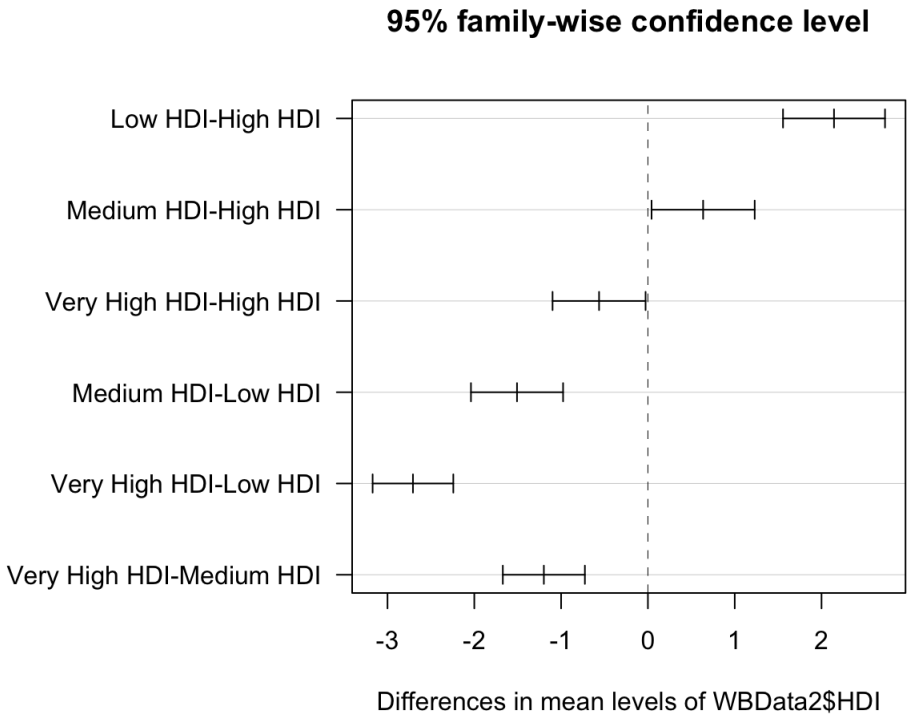
```
aov1 <- aov(WBData2$Fertility16 ~ WBData2$HDI)
summary.aov(aov1)
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## WBData2$HDI    3  182.7   60.91   78.87 <2e-16 ***
## Residuals    161  124.3    0.77
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
TukeyHSD(aov1)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = WBData2$Fertility16 ~ WBData2$HDI)
##
## $`WBData2$HDI`
##                                diff         lwr         upr     p adj
## Low HDI-High HDI          2.1438842  1.55637515  2.73139327 0.0000000
## Medium HDI-High HDI       0.6360333  0.04208351  1.22998316 0.0306605
## Very High HDI-High HDI   -0.5626788 -1.09845672 -0.02690086 0.0354768
## Medium HDI-Low HDI       -1.5078509 -2.03846512 -0.97723663 0.0000000
## Very High HDI-Low HDI    -2.7065630 -3.17114138 -2.24198461 0.0000000
## Very High HDI-Medium HDI -1.1987121 -1.67140925 -0.72601499 0.0000000
```

```
par(mar=c(5,15,4,2))
plot(TukeyHSD(aov1), las = 1)
```



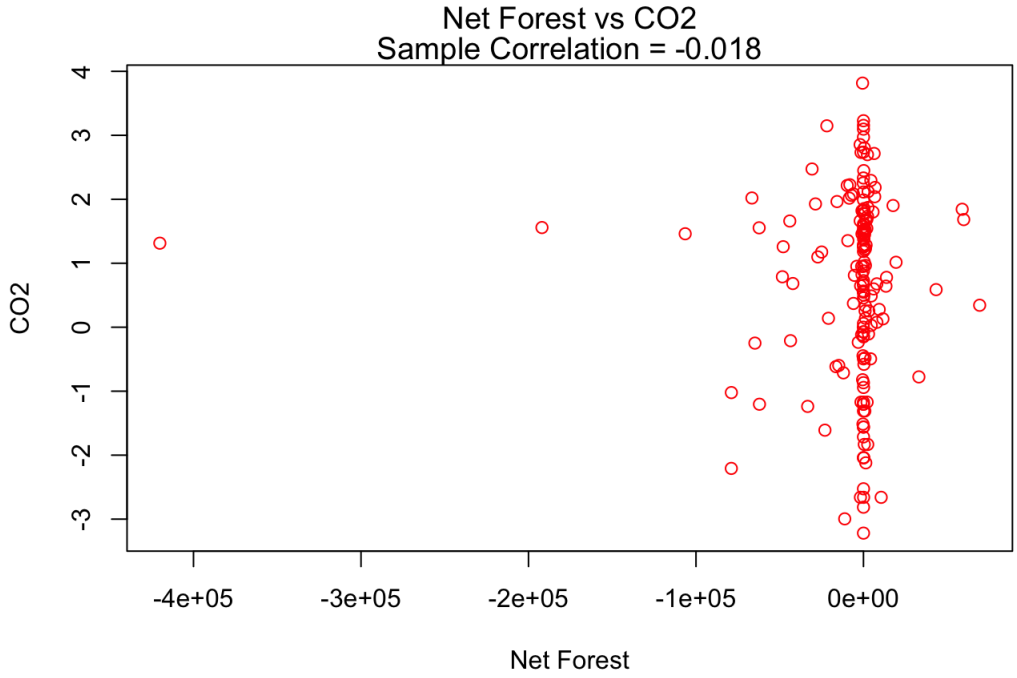
Our first analysis was between fertility rate in 2016 and our previously created categorical HDI variable. We found lower HDI groups to have higher mean fertility rates than higher HDI groups, Implying fertility falls with development. Our tukey test, t-tests which correct for family wise error rate, found that there was a statistically significant difference in mean fertility between all pairs (Low HDI and High HDI, etc.)

CO2 and NET FOREST

```
library(car)
```

```
## Loading required package: carData
```

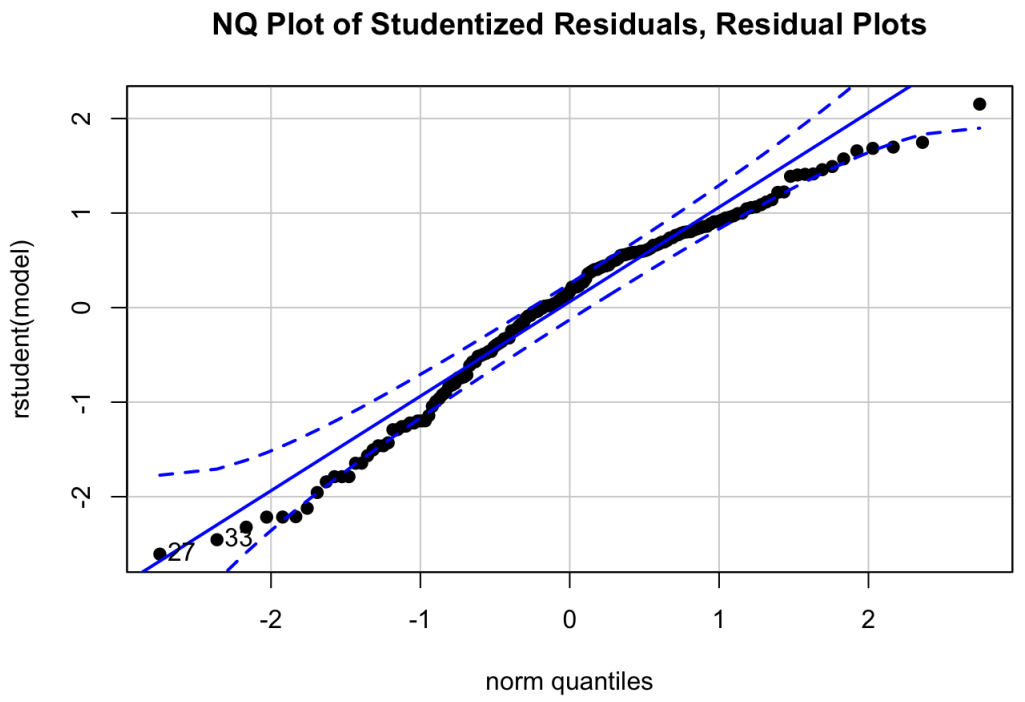
```
source("http://www.reuningscherer.net/s&ds230/Rfuncs/regJDRS.txt")

plot(log(WBData2$CO2) ~ WBData2$NetForest, col="red", xlab = "Net Forest", ylab = "CO2")
mtext(paste("Sample Correlation =", round(cor(log(WBData2$CO2),WBData2$NetForest),3)), cex=1.2)
mtext("Net Forest vs CO2", cex=1.2, line = 1)
```

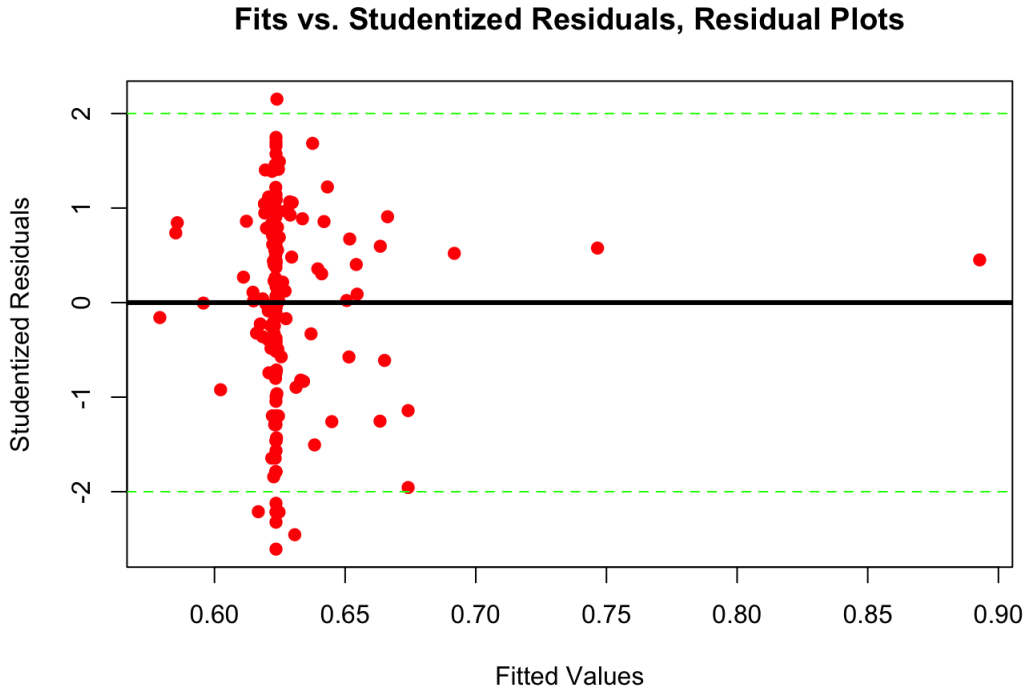



```
lm1 <- lm(log(WBData2$CO2) ~ WBData2$NetForest)
summary(lm1)
```

```
##
## Call:
## lm(formula = log(WBData2$CO2) ~ WBData2$NetForest)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8424 -0.9136  0.2560  1.0975  3.1921
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         6.235e-01  1.191e-01   5.235 5.02e-07 ***
## WBData2$NetForest -6.410e-07  2.852e-06  -0.225    0.822
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.504 on 163 degrees of freedom
## Multiple R-squared:  0.0003097,  Adjusted R-squared:  -0.005823
## F-statistic: 0.05049 on 1 and 163 DF,  p-value: 0.8225
```

```
myResPlots2(lm1)
```


NQ Plot of Studentized Residuals, Residual Plots
rstudent(model)
norm quantiles
27
33
-2
-1
0
1
2

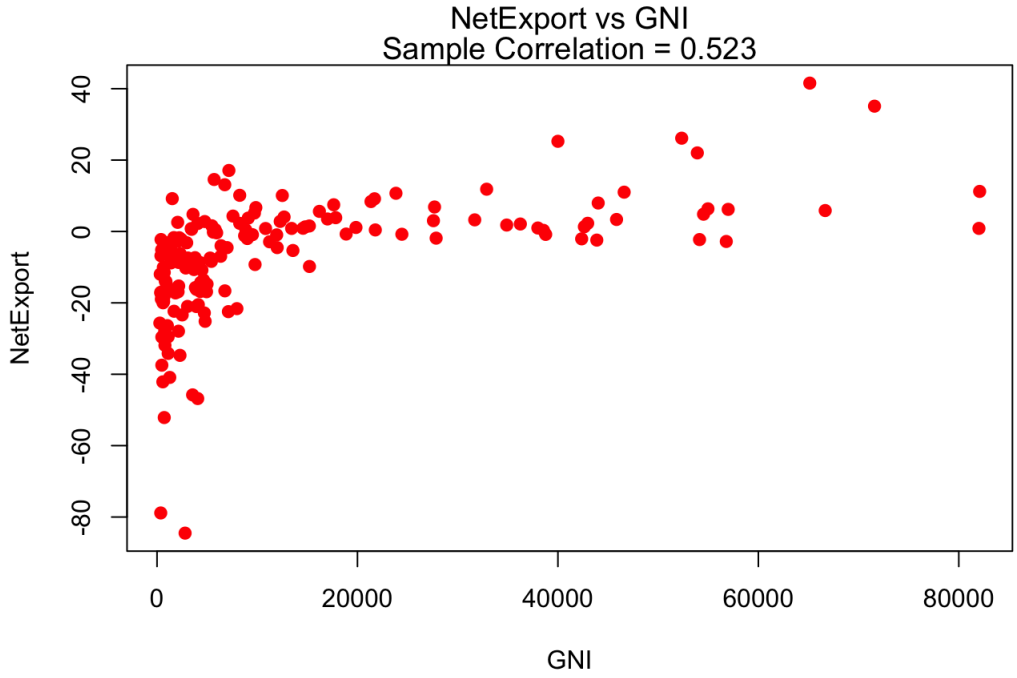

Fits vs. Studentized Residuals, Residual Plots
Studentized Residuals
Fitted Values
0.60
0.65
0.70
0.75
0.80
0.85
0.90

Forest type	Deforestation (Emissions)	Growth and Management	Net Change
Managed Forest	-19.8	42.8	23.0
Unmanaged Forest	-134.3	136.8	2.5
Total	-154.1	179.7	25.6

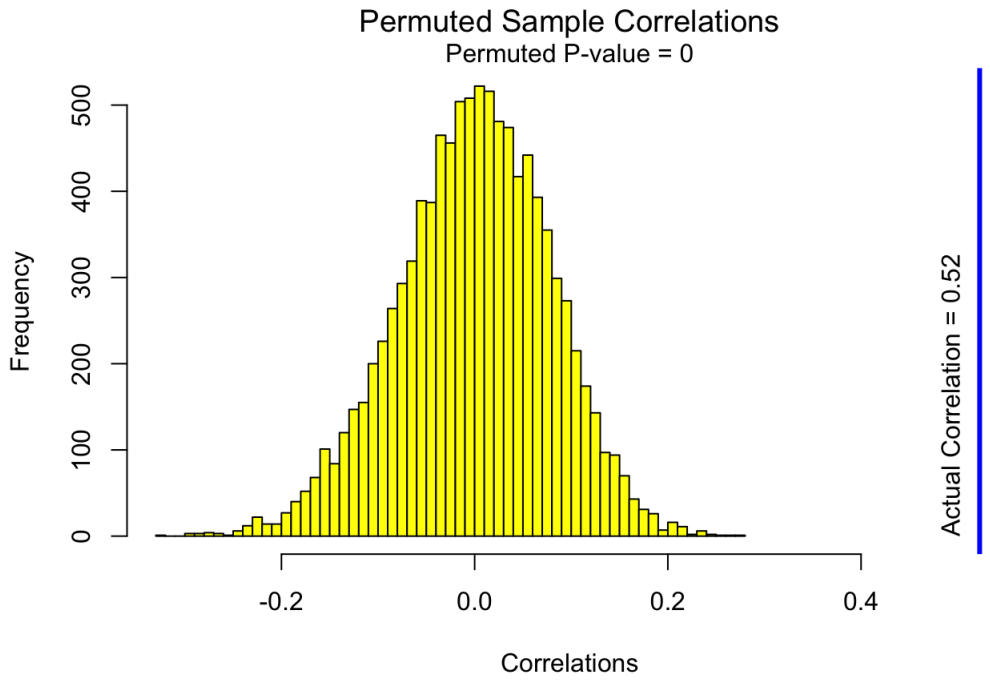
We next ran a linear regression of CO2 predicted by NetForest, our change in forest area between 2014 and 1994. We found there was a negligible regression. It is a widely held misconception that the forestry industry and deforestation are a contributor to CO2 emissions. While it is true that forests sequester carbon and cutting them down releases the gas, we found a very small regression between the two variables. We believe this is because of a change the forestry industry went through in the 1990s when its switched from cutting down old growth forests to cutting down managed, renewable forests. Managed forests are able to sequester much more carbon as there are less gaps in the tree cover. Thus, less land area is needed to hold the same amount of carbon. So, better managed forests are able to take carbon out of the atmosphere. In fact, it is estimated by Yale professor of economics Robert Mendelsohn that the forestry industry has contributed a net negative, or an abatement, to greenhouse gas pollution.

NET EXPORT and GNI

```
options(scipen=999)

myCor <- function(x,y){
  plot(x,y,pch=19, col="red", xlab = "GNI", ylab = "NetExport")
  mtext(paste("Sample Correlation =", round(cor(x,y),3)), cex=1.2)
}
x <- WBData2$GNI
y <- WBData2$NetExport
myCor(x,y)
mtext("NetExport vs GNI", cex=1.2, line = 1)
```




```
options(scipen=999)
permCor <- function(x, y, n_samp = 10000, plot = T){
   corResults <- rep(NA, n_samp)
   for (i in 1:n_samp){
      corResults[i] <- cor(x, sample(y))
   }
   pval <- mean(abs(corResults) >= abs(cor(x,y)))
   if (plot == T){
      #Make histogram of permuted correlations
      hist(corResults, col = "yellow", main = "", xlab = "Correlations", breaks = 50,
           xlim = range(corResults,cor(x,y)))
      mtext("Permuted Sample Correlations", cex = 1.2, line = 1)
      mtext(paste("Permuted P-value =",round(pval,5)), cex = 1, line = 0)
      abline(v = cor(x,y), col="blue", lwd=3)
      text(cor(x,y)-.03, 0,paste("Actual Correlation =", round(cor(x,y),2)),srt = 90, adj = 0)
   }
   if (plot == F){
      return(round(pval,5))
   }
}
permCor(x,y)
```




```
cor.test(x,y)
```

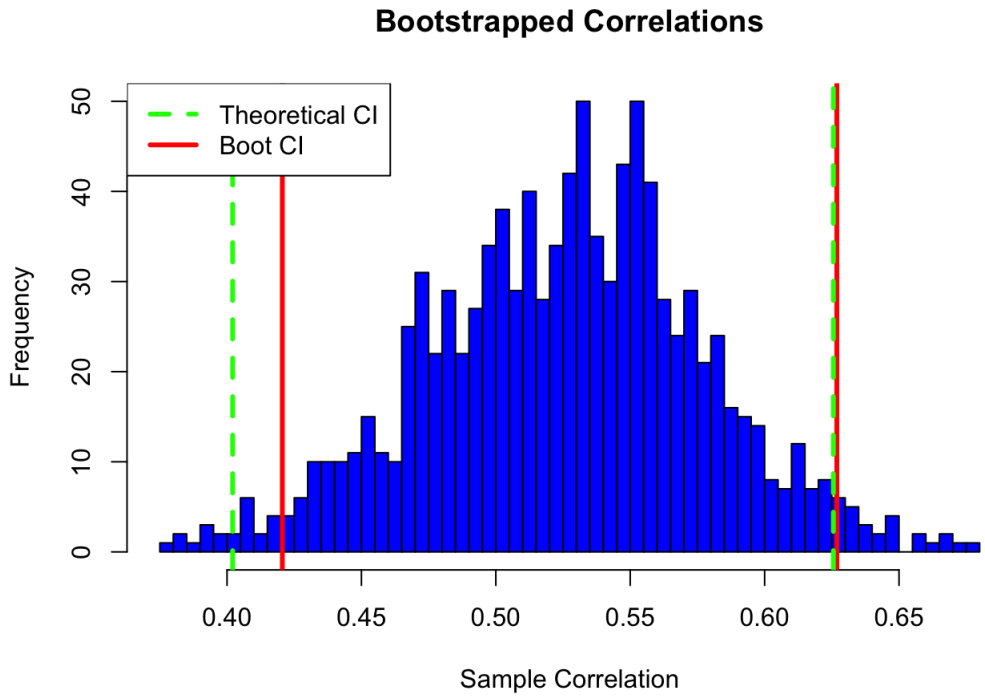
```
##
##  Pearson's product-moment correlation
##
## data:  x and y
## t = 7.8292, df = 163, p-value = 0.0000000000005908
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.4020951 0.6255820
## sample estimates:
##       cor
## 0.5227634
```

Bootstrap Correlation for Net Export and GNI

```r
cor1test <- cor.test(WBData2$NetExport, WBData2$GNI)$conf.int

#get number of rows in dataset
N <- nrow(WBData2)

#Specify how many boostrap samples to take
n_samp <- 1000

corResults <- rep(NA, n_samp)

for(i in 1:n_samp){
  #get vector of rows in our fake sample
  s <- sample(1:N, N , replace=T)

  #Calculate unique rows in our sample
  sVals <- as.numeric(names(table(s)))

  #Calculate how many times each row shows up
  sCounts <-  as.vector(table(s))

  #Get bootstrapped GNI and NetExport values
  fakeGNI <-  rep(WBData2$GNI[sVals], sCounts)
  fakeNetExport <-  rep(WBData2$NetExport[sVals], sCounts)

  #Get bootstrapped correlation and regression slope
  corResults[i] <- cor(fakeGNI, fakeNetExport)
}
(ci_r <- quantile(corResults, c(.025, .975)))
```

```
##      2.5%     97.5%
## 0.4205789 0.6268695
```

```r
hist(corResults, col = "blue", main = "Bootstrapped Correlations", xlab = "Sample Correlation", breaks = 50)
abline(v = ci_r, lwd = 3, col = "red")
abline(v = cor1test, lwd = 3, col = "green", lty = 2)
legend("topleft", c("Theoretical CI","Boot CI"), lwd=3, col = c("green","red"), lty = c(2,1))
```



Furthermore, we also decided to examine the relationship between GNI and NetExport, which is the difference between the variables export and import. This topic has recently received much more attention from our trade war with China. It is commonly believed that a positive net export translates to an overall stronger economy. Looking at our results, there does appear to be a moderately strong correlation between GNI and NetExport without any significant outliers in the correlation plot. Our permutation test gives us a permuted p-value of 0.000 so we can conclude that there is a statistically significant non-zero correlation between GNI and NetExport. Looking at our bootstrapped

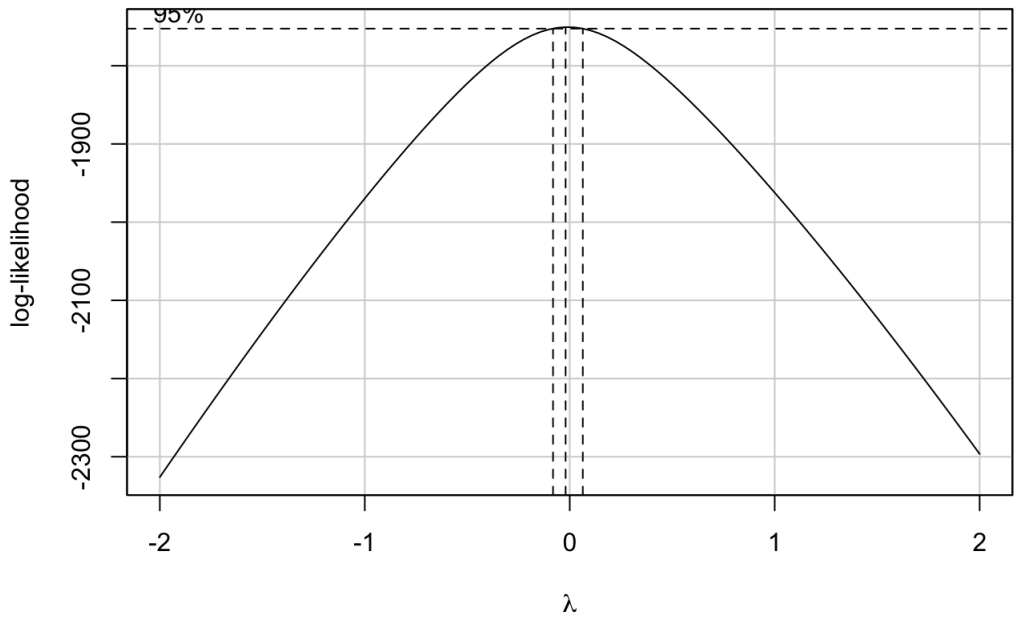
correlation plot, it does appear that the gap between the left side theoretical CI and Boot CI is wider than the right. This can be attributed to several minor outliers on the left side of the correlation plot. However, we think that these minor outliers are not significant enough to remove from the correlation because they play an important role in the overall data.

MULTIPLE REGRESSION

```
GNIpred <- lm(WBData2$GNI ~ WBData2$LifeExp + WBData2$NetExport + WBData2$NetForest + WBData2$Fertility16 +
WBData2$CO2)

#Get summary information for model
summary(GNIpred)
```

```
##
## Call:
## lm(formula = WBData2$GNI ~ WBData2$LifeExp + WBData2$NetExport +
##     WBData2$NetForest + WBData2$Fertility16 + WBData2$CO2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -20783  -7610  -3089   4655  54286
##
## Coefficients:
##                           Estimate     Std. Error t value       Pr(>|t|)
## (Intercept)         -107666.058208   18717.162772  -5.752 0.000000043960
## WBData2$LifeExp        1482.225262     221.264133   6.699 0.000000000344
## WBData2$NetExport       221.264020      64.193604   3.447       0.000726
## WBData2$NetForest         0.008442       0.022148   0.381       0.703592
## WBData2$Fertility16    4077.207351    1259.470558   3.237       0.001469
## WBData2$CO2            1095.289918     182.697141   5.995 0.000000013189
##
## (Intercept)         ***
## WBData2$LifeExp     ***
## WBData2$NetExport   ***
## WBData2$NetForest
## WBData2$Fertility16 **
## WBData2$CO2         ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11590 on 159 degrees of freedom
## Multiple R-squared:  0.6068, Adjusted R-squared:  0.5944
## F-statistic: 49.07 on 5 and 159 DF,  p-value: < 0.00000000000000022
```

```
#Run a Box Cox procedure on model
boxCox(GNIpred)
```




```
#Transform regression according to box-cox
LogGNIpred <- lm(log(WBData2$GNI) ~ WBData2$LifeExp + WBData2$NetExport + WBData2$NetForest + WBData2$Fertil
ity16 + WBData2$CO2)
summary(LogGNIpred)
```

```
##
## Call:
## lm(formula = log(WBData2$GNI) ~ WBData2$LifeExp + WBData2$NetExport +
##     WBData2$NetForest + WBData2$Fertility16 + WBData2$CO2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.63134 -0.39270 -0.05041  0.38673  1.81515
##
## Coefficients:
##                           Estimate     Std. Error t value
## (Intercept)          1.03461161113  0.99429625286   1.041
## WBData2$LifeExp      0.10581462587  0.01175403031   9.002
## WBData2$NetExport    0.01747301456  0.00341010336   5.124
## WBData2$NetForest    0.00000009685  0.00000117653   0.082
## WBData2$Fertility16 -0.08535062221  0.06690580574  -1.276
## WBData2$CO2          0.06218684946  0.00970526812   6.408
##                                Pr(>|t|)
## (Intercept)                       0.300
## WBData2$LifeExp     0.000000000000000645 ***
## WBData2$NetExport   0.000000857610995528 ***
## WBData2$NetForest                 0.935
## WBData2$Fertility16               0.204
## WBData2$CO2         0.000000001599511249 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6158 on 159 degrees of freedom
## Multiple R-squared:  0.8309, Adjusted R-squared:  0.8256
## F-statistic: 156.3 on 5 and 159 DF,  p-value: < 0.00000000000000022
```

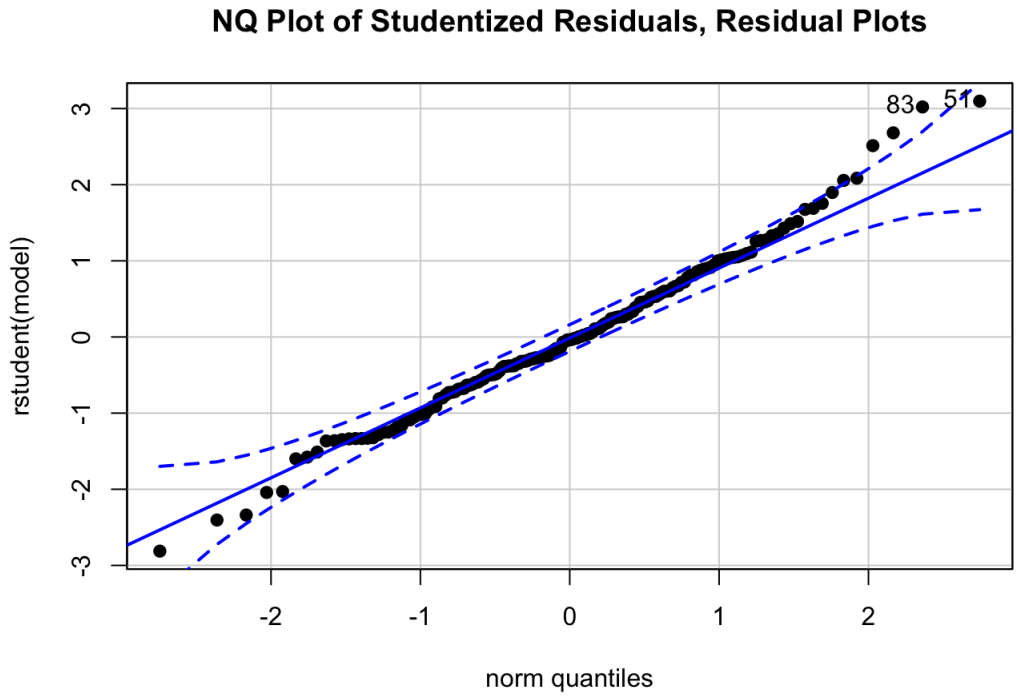
```
#Backwards stepwise regression
LogGNIpred2 <- lm(log(WBData2$GNI) ~ WBData2$LifeExp + WBData2$NetExport + WBData2$Fertility16 + WBData2$CO2
)
summary(LogGNIpred2)
```

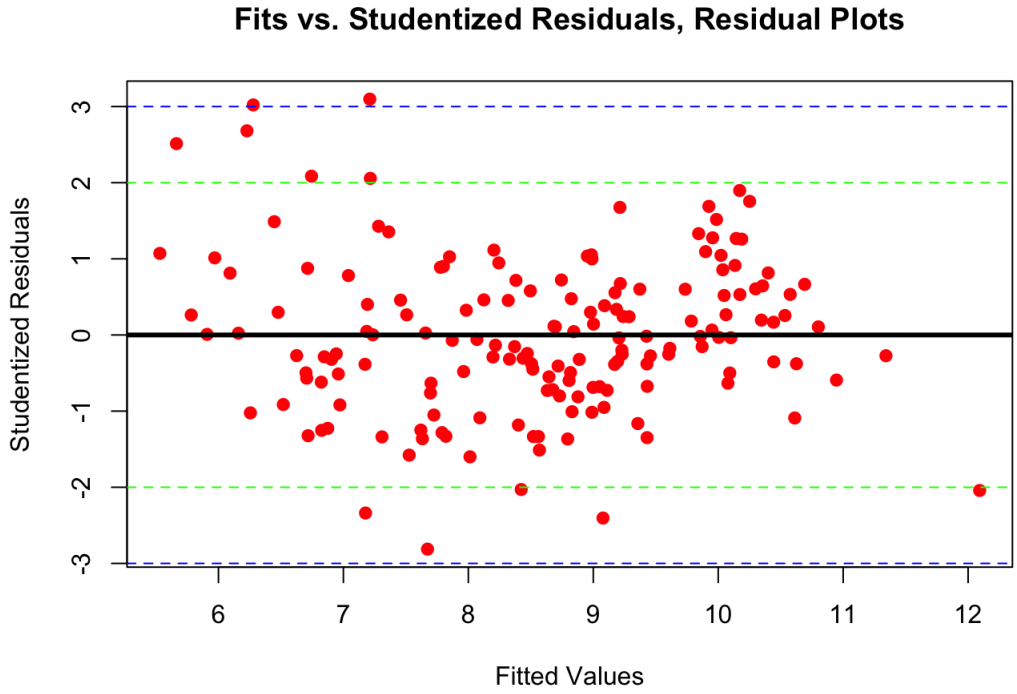
```
##
## Call:
## lm(formula = log(WBData2$GNI) ~ WBData2$LifeExp + WBData2$NetExport +
##     WBData2$Fertility16 + WBData2$CO2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.63019 -0.39167 -0.05002  0.38757  1.80983
##
## Coefficients:
##                      Estimate Std. Error t value             Pr(>|t|)
## (Intercept)          1.042197   0.986939   1.056                0.293
## WBData2$LifeExp      0.105724   0.011666   9.062 0.000000000000000432 ***
## WBData2$NetExport    0.017475   0.003399   5.140 0.000000790717736220 ***
## WBData2$Fertility16 -0.086015   0.066211  -1.299                0.196
## WBData2$CO2          0.062190   0.009675   6.428 0.000000001419938750 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6139 on 160 degrees of freedom
## Multiple R-squared:  0.8309, Adjusted R-squared:  0.8267
## F-statistic: 196.5 on 4 and 160 DF,  p-value: < 0.00000000000000022
```

```
#Next Step
LogGNIpred3 <- lm(log(WBData2$GNI) ~ WBData2$LifeExp + WBData2$NetExport + WBData2$CO2)
summary(LogGNIpred3)
```

```
##
## Call:
## lm(formula = log(WBData2$GNI) ~ WBData2$LifeExp + WBData2$NetExport +
##     WBData2$CO2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.68170 -0.38483 -0.02214  0.36921  1.80669
##
## Coefficients:
##                    Estimate Std. Error t value             Pr(>|t|)
## (Intercept)       -0.051812   0.515756  -0.100                 0.92
## WBData2$LifeExp    0.117658   0.007207  16.325 < 0.0000000000000002 ***
## WBData2$NetExport  0.017842   0.003395   5.255        0.00000046291 ***
## WBData2$CO2        0.062551   0.009692   6.454        0.00000000122 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6152 on 161 degrees of freedom
## Multiple R-squared:  0.8291, Adjusted R-squared:  0.8259
## F-statistic: 260.4 on 3 and 161 DF,  p-value: < 0.00000000000000022
```

```
#Get residual plots
myResPlots2(LogGNIpred3)
```





Our multiple regression analysis predicted the natural log of GNI per capita based on the majority of the variables in our data set. We decided to omit our HDI variable for this regression as GNI itself was a component of that variable. Our backwards, stepwise regression showed that life expectancy, net exports, and CO2 emissions were all significant predictors of GNI. Each of these predictors has a positive coefficient, meaning that as the value of each predictor increased, so did the mean of logGNI. We were surprised to see that Fertility was not a significant predictor of GNI, especially after we found it to be related to our HDI variable, of which GNI was a component. This implies that the relation may have been due mostly to Life Expectancy, the other part of our HDI variable. Overall, the model had a high adjusted r^2 value of .826, meaning that ~83% of the variance in GNI can be explained by this model.

Conclusion and Summary:

Overall, we were able to determine that there are differences in mean fertility rates by HDI using ANOVA and tukey. Next, we discover that there was not a significant regression for CO2 predicted by NetForest. This is despite the fact that there is a widely held belief that deforestation is a large contributor to CO2 emissions. We believe that other factors such as increased usage of managed forests, which sequester more carbon, may account for this. In our analysis of NetExport and GNI, we determined that was a moderately strong correlation between the two variables. To further investigate this correlation, we ran a permutation test, where we were able to determine that there is a statistically significant non-zero correlation between GNI and NetExport. Finally, in our multiple regression model, using backwards, stepwise regression, we observed that LifeExp, NetExport, and CO2 emissions were all statistically significant predictors of GNI.

Our original intention in examining this dataset was to analyze and answer relevant questions that have great significance in society. Overall, we believe that we were able to accomplish this goal and discover some surprising conclusions along the way. We hope that this analysis can inspire further dialogue and examination into this topic with real-world implications.