Final Project

S&DS 230/530

(NA)

S&DS 230/530 Final Project

Introduction

I wish to explore the relationships among various measures of a country. I considered using the World Bank dataset provided on Canvas, but I wanted to practice scraping data off websites (I like the dynamic element of it), so I decided to do so. In this project, I explore the correlation between life expectancy and median age, the difference in the mean democratic index of countries that recognize same-sex marriage and those that do not, the homicide rate of a country as predicted by military expenditure and continent, and the homicide rate as predicted by a number of continuous variables.

DATA

- Status of same-sex marriage (categorical; recognized, including performed, legal, etc.)
- Continent (categorical)
- Democracy index (continuous; a measure of the state of democracy compiled by the Economist Intelligence Unit)
- CO2 emission (continuous; metric tons)
- Energy use (continuous; KWH)
- Median age (continuous; years)
- Life expectancy at birth (continuous; years)
- Total fertility rate (continuous; children born per woman)
- Infant mortality (continuous; deaths per 1000)
- Homicide rate in 2017 (continuous; homicide deaths per 100000)
- GDP Per Capita (continuous; USD)
- Current account balance (continuous; USD)
- Exports (continuous; USD)
- Imports (continuous; USD)

Data Cleaning Process

The data used in this project are from several sources:
1. The World Factbook
2. The Wikipedia page on the legal status of same-sex marriage
3. This .csv file posted on DataHub by someone called JohnSnowLabs
4. Democracy index information downloaded from here as a .csv file
5. Homicide statistics from 1990 to 2017, downloaded from here, sourced from IHME

The World Factbook

The following variables are obtained via web scraping from the World Factbook: CO2 emission, Energy use, Population, Median age, Life expectancy at birth, Total fertility rate, Infant mortality, GDP Per Capita,

Current account balance, Exports, Imports. Data cleaning mainly included using `gsub()` to remove symbols like ',' and '$' from numerical data.
The following is the code I used to scrape data for CO2 emission in various countries:

```
url7 <- "https://www.cia.gov/library/publications/the-world-factbook/fields/274rank.h
tml"
webpage7 <- read_html(url7)
countries7 <- html_text(html_nodes(webpage7, '#rankOrder a'))
CO2emission <- html_text(html_nodes(webpage7, '.region+ td'))
CO2emission <- as.numeric(gsub(",", "", CO2emission))
countryCO2 <- data.frame(country = countries7, CO2 = CO2emission, stringsAsFactors =
FALSE)
```

I created a dataframe for each variable, then combined each mini dataframe by using `myMerge()`, a function I wrote to merge dataframes by variable "country" so that I could use `Reduce()`.

```
myMerge <- function(x, y) {
  return(merge(x, y, by = "country", all = TRUE))
}

worldFacts <- Reduce(myMerge, list(countryCO2, countryCurrentAcc, countryEnergyUse, c
ountryExports, countryImports, countryInfMort, countryIntUsers, countryFertility, cou
ntryGDP, countryLifeEx, countryMedAge, countryMilExp, countryPop))
```

Status of Same-Sex Marriage

In order to get the data for the status of same-sex marriage in various countries, I created a dataframe `sameSexMar` of countries on the Wikipedia page that recognize same-sex marriage and a variable called `SameSexMarriage` that has a value of "Recognized" for every country initially in the dataframe, then merged this dataframe with the `worldFacts` dataframe I created from the World Factbook data. Next I replaced all the NAs in `SameSexMarriage` in the merged `worldFacts` dataframe with "Not recognized", and made this variable a factor.

```
sameSexMar <- data.frame(country = recognized, SameSexMarriage = c(rep("Recognized",
length(recognized))), stringsAsFactors = FALSE) #create dataframe
worldFacts <- myMerge(worldFacts, sameSexMar) #merge dataframes
worldFacts$SameSexMarriage[is.na(worldFacts$SameSexMarriage)] <- c(rep("Not recognize
d", length(worldFacts$SameSexMarriage[is.na(worldFacts$SameSexMarriage)]))) #assign n
ew value to NAs
worldFacts$SameSexMarriage <- as.factor(worldFacts$SameSexMarriage) #turn variable fr
om string to factor
```

Countries and Continents

The .csv file of the continents of various countries required a lot of cleaning, since many of the country names in this file are different from the names of the same countries in the `worldFacts` dataframe. To facilitate the process, I wrote a function `replaceNames()` that takes in a vector to clean, a vector of things to swap out, and a corresponding vector of things to swap in. (For an example of it running, see the next section.)

```
replaceNames <- function(stringVec, swapOut, swapIn) {
  for (i in 1:length(swapOut)) {
    stringVec <- gsub(swapOut[i], swapIn[i], stringVec)
  }
  return(stringVec)
}
```

An issue I encountered was with “Republic of the Congo” and “Democratic Republic of the Congo”, both of which in the data cleaning process got reduced to “Congo”. I referenced the original country names from the .csv file to determine which Congo was which, and manually replaced them, and then applied `myMerge()` on this new dataframe and the `worldFacts` dataframe.

Democracy Index and Homicide Rate

After reading the 2 .csv files into RStudio, I used `replaceNames()` to clean up the country names in the 2 .csv files, then merged them with the `worldFacts` dataframe using `myMerge()`.

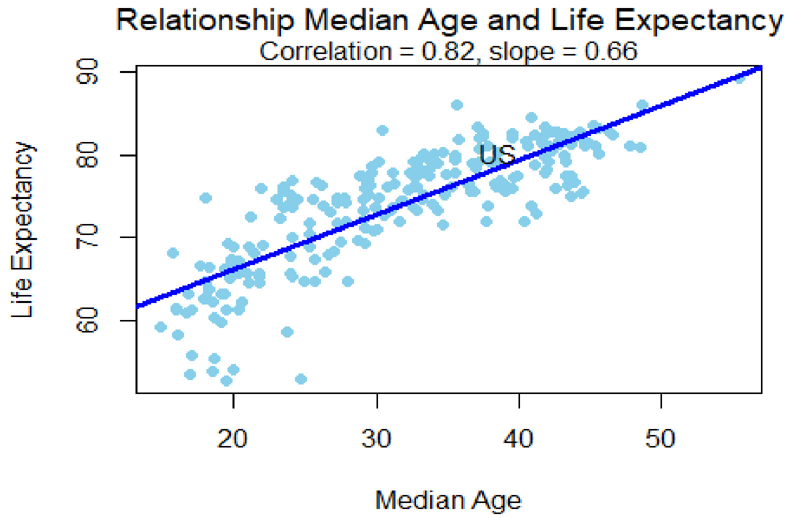
Here is an example of using `replaceNames()` on country names (called `Entity` here) in the `homicide` dataframe:

```
homicide$Entity <- replaceNames(
  homicide$Entity,
  c("Democratic Republic of Congo", "^Congo$", "Czech Republic", "South Korea", "Baha
mas$", "Cape Verde", "Gambia$", "Myanmar", "Timor$", "United States Virgin Islands"),

  c("Congo, Democratic Republic of the", "Congo, Republic of the", "Czechia", "Korea,
 South", "Bahamas, The", "Cabo Verde", "Gambia, The", "Burma", "Timor-Leste", "Virgin
 Islands"))
```

Analysis & Plots

1) Life Expectancy and Median Age (Theoretical and Bootstrapped CIs)

Here I look at the correlation between life expectancy and median age. Here is a plot of the data:

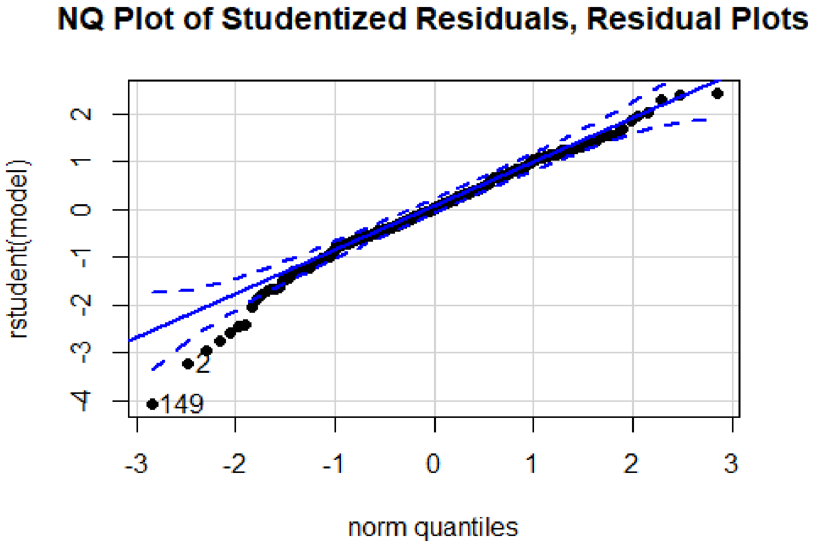


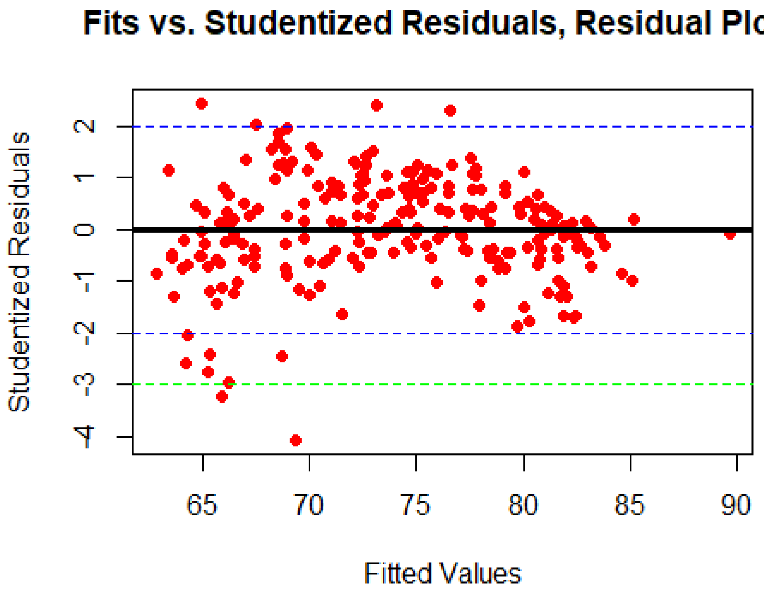
There appears to the a strong positive correlation between median age in a country and the life expectancy at birth in the country. The calculated correlation is 0.8223146.

```
##
## Call:
## lm(formula = lifeExpect ~ medianAge, data = WF4)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.3516  -2.2960   0.0812   2.8789   9.9745
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 53.03456    0.99656   53.22   <2e-16 ***
## medianAge    0.66061    0.03041   21.73   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.162 on 226 degrees of freedom
## Multiple R-squared:  0.6762, Adjusted R-squared:  0.6748
## F-statistic:   472 on 1 and 226 DF,  p-value: < 2.2e-16
```

The correlation appears to be statistically significant, being a lot smaller than 0.05.
Here are the residual plots:





We observe slight nonlinearity to the bottom left of the normal quantile plot of residuals, but otherwise our residual plots indicate that the assumptions behind our model is reasonable.

Bootstrapping

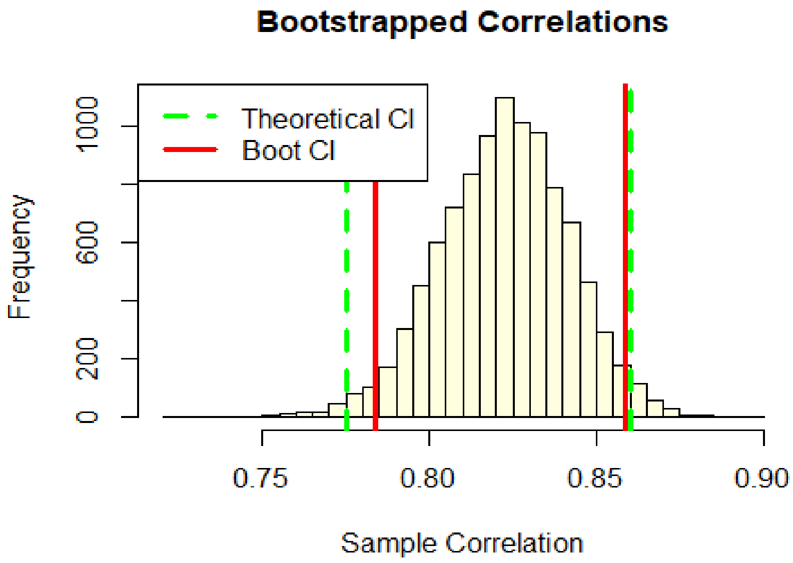
I perform boostrapping with 1000 samples. The code is as follows.

```
N <- nrow(WF4)
n_samp <- 10000 #take 10000 samples
corResults <- rep(NA, n_samp)
bResults <- rep(NA, n_samp)

for(i in 1:n_samp){
  s <- sample(1:N, N , replace = T)
  sVals <- as.numeric(names(table(s)))
  sCounts <-  as.vector(table(s))
  bootAge <-  rep(WF4$medianAge[sVals], sCounts)
  bootLife <-  rep(WF4$lifeExpect[sVals], sCounts)
  corT <- cor(bootAge, bootLife)
  lmT <- lm(bootLife ~ bootAge)
  corResults[i] <-  corT
  bResults[i] <-  lmT$coef[2]
}
```

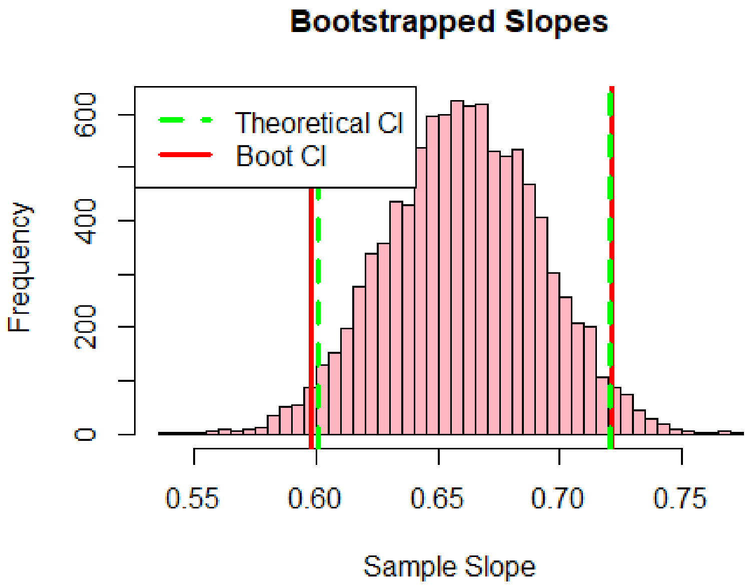
This gives me the following CIs for correlation and slope:

```
## [1] "Bootstrapped CI for correlation:"

##      2.5%     97.5% 
## 0.7838464 0.8584297

## [1] "Bootstrapped CI for slope:"

##      2.5%     97.5% 
## 0.5978332 0.7214113
```

Here is the histogram for bootstrapped correlations. The bootstrapped CI appear to be narrower than the theoretical CI, which is probably due to a few more influential points that are not as often included in the bootstrapped samples.



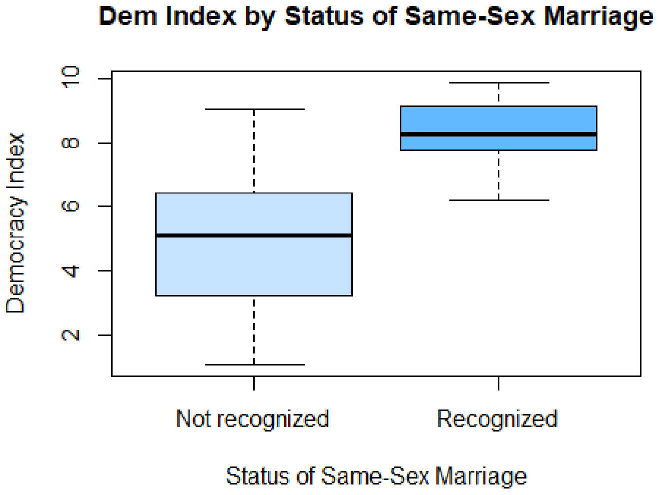
Here is the histogram for bootstrapped slopes. The bootstrapped CI is very close to the theoretical CI.



2) Democracy Index and Status of Same-Sex Marriage (t-test and Permutation Test)

Here I look at the relationship between democracy index and the status of same-sex marriage. In particular, I perform a permutation test to determine whether there is a difference between the mean democracy indices of countries that recognize and don't recognize same-sex marriage.

A boxplot of democracy index by status of same-sex marriage suggests that the democracy index of countries that recognize same-sex marriage seem to be higher than that of countries that don't.



Performing a t-test, we obtain the following results.

```
## 
##  Welch Two Sample t-test
## 
## data:  democracyScore by SameSexMarriage
## t = -14.018, df = 87.03, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.853906 -2.896761
## sample estimates:
## mean in group Not recognized     mean in group Recognized 
##                     4.869333                     8.244667
```

The p-value, 4.865929810^{-24} is very small (< 0.05), and the confidence interval does not include 0, which suggests that the difference between democracy index of countries that recognize same-sex marriage and that of countries that don't is not zero.

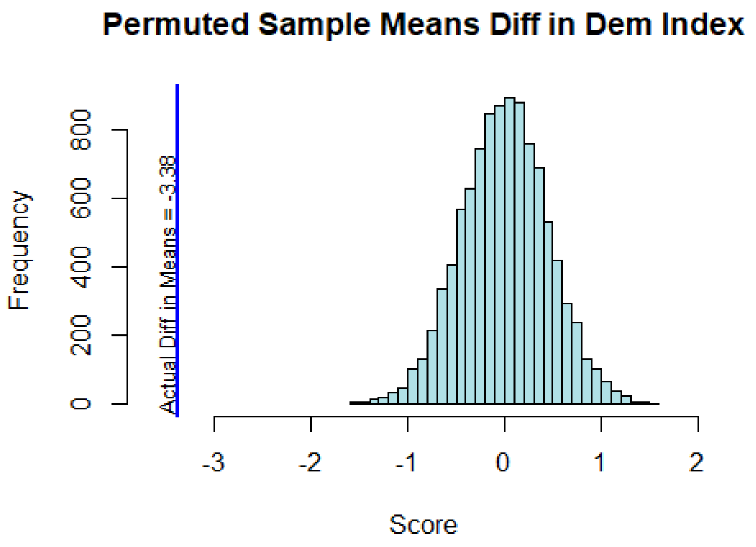
Next I perform a permutation test. The actual difference between the two groups in the sample is:

```
## [1] "Difference between sample means:"

## Not recognized 
##      -3.375333
```

We repeated our permutation many times (10000), as follows.

```
N <- 10000
diffvals <- rep(NA, N)
for (i in 1:N) {
  fakestatus <- sample(WF5$SameSexMarriage)  # default is replace = FALSE
  diffvals[i] <- mean(WF5$democracyScore[fakestatus == "Not recognized"]) -  mean(WF5
$democracyScore[fakestatus == "Recognized"])
}
```

Here is a histogram of the permuted sample means difference from the bootstrapping.



The p-value of the actual difference in means from this bootstrapped distribution is calculated as follows, and we see that it is basically 0.

```
mean(abs(diffvals) >= abs(actualdiff))

## [1] 0
```

Therefore we reject the null hypothesis that there is no difference between the means in our two groups, and conclude that the true difference between the means is not zero. More specifically, the democracy index of countries that recognize same-sex marriage is higher than that of countries that do not.

3) Homicide Rate, Military Expenditure, and Continent (ANCOVA)

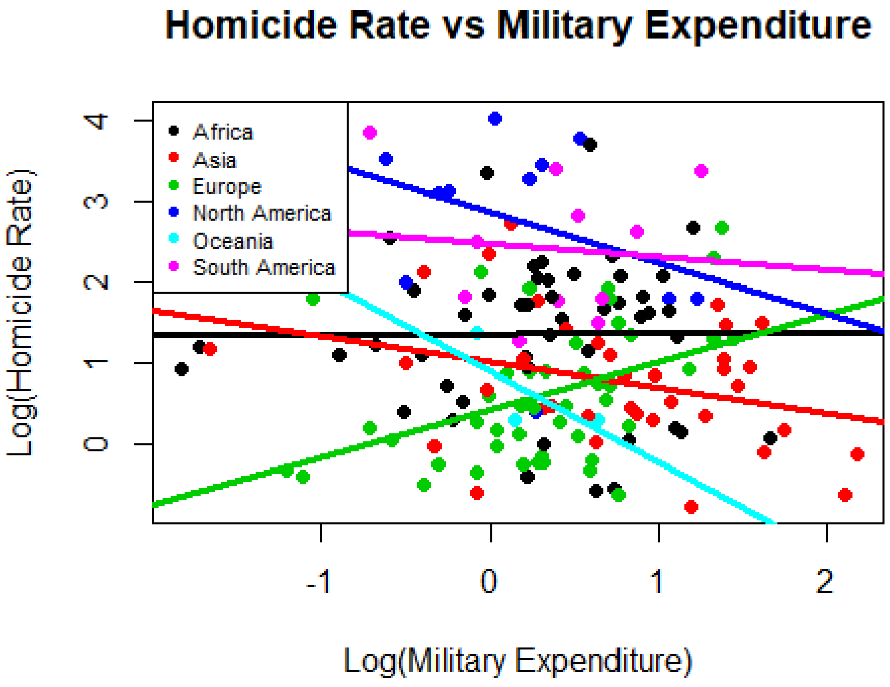
Here I perform ANCOVA for log of homicide rate based on the log of military expenditure and the categorical continent variables.

Applying `Anova()` to the model containing military expenditure, continent, and their interaction, we see that the interaction term is statistically significant at the $\alpha = 0.05$ level.

```
m2 <- lm(logHom ~ logMilExp*Continent_Name, data = WF10)
Anova(m2)

## Anova Table (Type II tests)
##
## Response: logHom
##                           Sum Sq  Df F value    Pr(>F)    
## logMilExp                  0.123   1  0.1661  0.684188    
## Continent_Name            69.840   5 18.8113 2.427e-14 ***
## logMilExp:Continent_Name  12.817   5  3.4521  0.005636 ** 
## Residuals                106.182 143                      
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Here is a visual representation of this model:

```r
plot(logHom ~ logMilExp, data = WF10, col = factor(continent), pch = 19, main = "Homi
cide Rate vs Military Expenditure", ylab = "Log(Homicide Rate)", xlab = "Log(Military
 Expenditure)")
coefs <- coef(m2)
abline(a = coefs[1], b = coefs[2], col = "black", lwd = 3)
for (i in 3:7){
  abline(a = coefs[1] + coefs[i], b = coefs[2] + coefs[i+5], col = (i-1), lwd = 3)
}
legend("topleft", col = 1:6, legend = levels(factor(continent)), pch = 16, cex = 0.7)
```



For all countries except those in Europe and Africa, it seems that lower homicide rate tends to go with higher military expenditure.

4) Homicide Rate (Best Subsets Regression)

Here I use best subsets regression to build a multiple regression model for the log of homicide rate. First, I create a new variable `logHomi` in `worldFacts`, which is the log of homicides (measured as deaths per 100000 people).

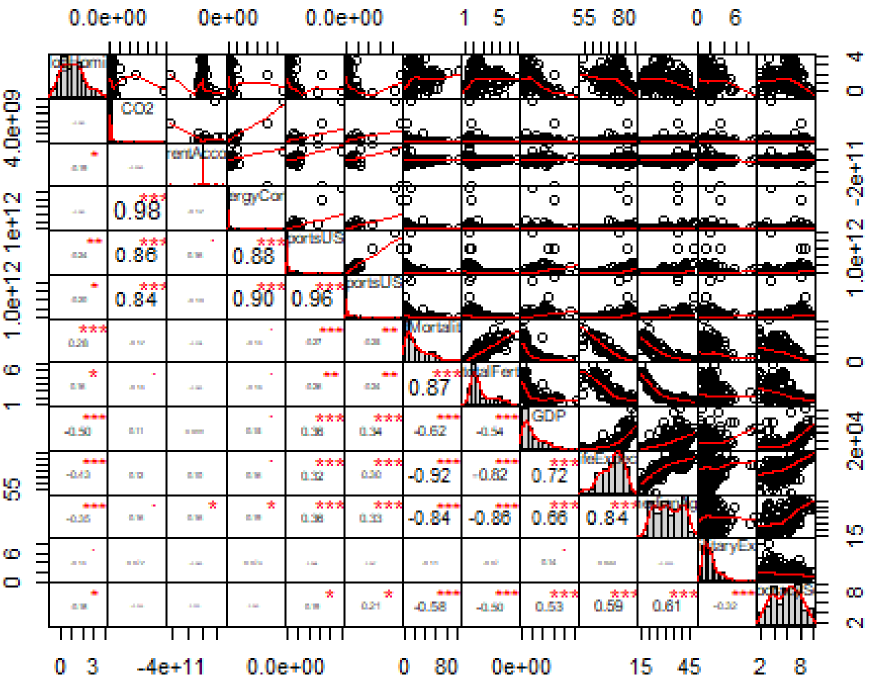
```r
worldFacts$logHomi <- log(worldFacts$Deaths_per_100000)
```

Next I create a new dataframe `WF8` with the relevant (continuous) variables:

```r
#names(worldFacts)
WF8 <- worldFacts[ , c(21, 4:9, 11:15, 19)]
WF8 <- WF8[complete.cases(WF8), ]
names(WF8)

## [1] "logHomi"        "CO2"            "currentAccount" "energyCons"
## [5] "exportsUSD"     "importsUSD"     "infMortality"   "totalFert"
##  [9] "GDP"             "lifeExpect"      "medianAge"       "militaryExp"    
## [13] "democracyScore"
```

I used a matrix plot to check relationships between our variables.

```
pairsJDRS(WF8, labels = names(WF8))
```



Based on the matrix plot, I decided to replace the variables `importsUSD` (imports), `exportsUSD` (exports), `CO2` (CO2 emission), and `energyCons` (energy use) with the log of themselves, due the the nature of the type of data (imports and exports are money data, and CO2 emission and energy use are both right-skewed).

```
WF8$importsUSD <- log(WF8$importsUSD)
WF8$exportsUSD <- log(WF8$exportsUSD)
WF8$CO2 <- log(WF8$CO2)
WF8$energyCons <- log(WF8$energyCons)
```

Next I apply `regsubsets()` to determine the best subsets with various numbers of predictors, and save the summary in a variable called `mod1sum`.

```
mod1 <- regsubsets(logHomi ~ ., data = WF8, nvmax = 12)
mod1sum <- summary(mod1)
```

Best Model According to Adjusted R-Squared

According to adjusted R-squared, the best model is a model with the following predictors:

```
names(WF8)[mod1sum$which[which.max(mod1sum$adjr2), ]][-1]
```

```
##  [1] "CO2"             "currentAccount"  "exportsUSD"      "importsUSD"     
##  [5] "infMortality"    "totalFert"       "GDP"             "lifeExpect"     
##  [9] "medianAge"       "militaryExp"     "democracyScore"
```

The R-squared value for this model is 0.4601, indicating that this model accounts for about 46% of the variability in the log of homicide rate. However, this model has 11 predictors, including ones that are not statistically significant at the $\alpha = 0.05$ level.

Best Model According to BIC

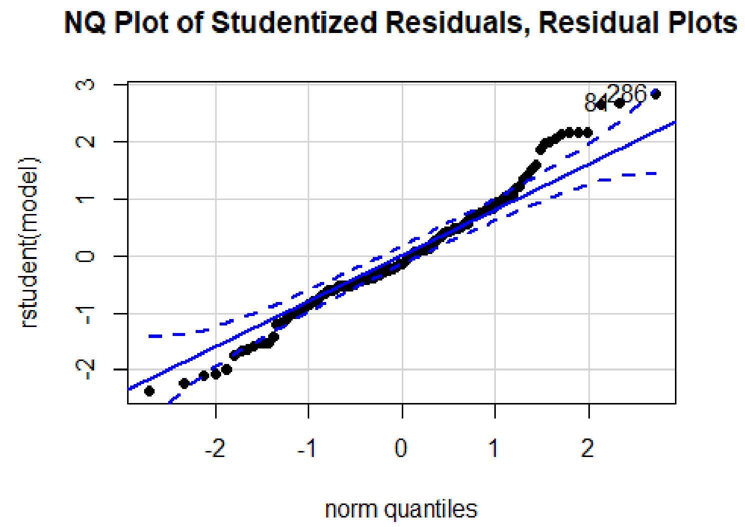
Next I use the Bayesian Information Criterion (BIC) to determine a model. This model includes the following predictors:

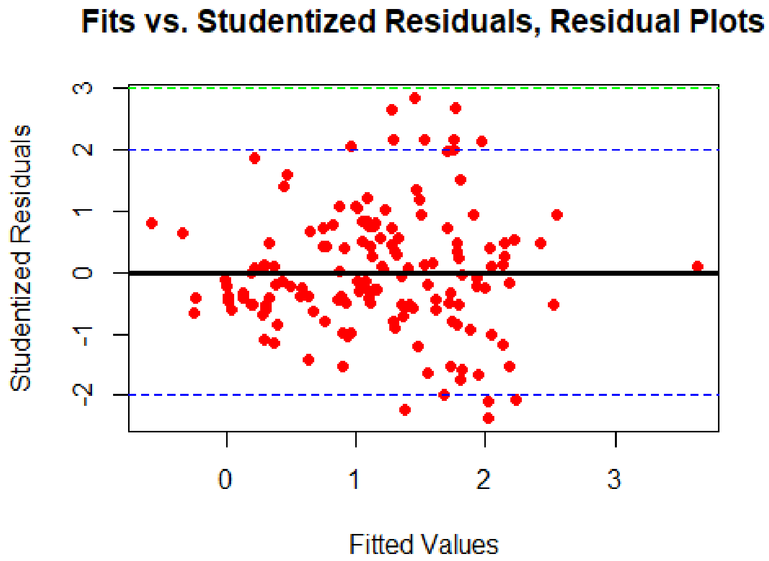
```
names(WF8)[mod1sum$which[which.min(mod1sum$bic), ]][-1]

## [1] "totalFert"      "GDP"            "lifeExpect"     "medianAge"
## [5] "democracyScore"
```

Here is the summary of this model:

```
WFtemp <- WF8[,mod1sum$which[which.min(mod1sum$bic), ]]
lmBIC <- lm(logHomi ~ ., data = WFtemp)
summary(lmBIC)

##
## Call:
## lm(formula = logHomi ~ ., data = WFtemp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.0150 -0.4606 -0.1198  0.4695  2.4078
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     1.027e+01  1.667e+00   6.161 6.87e-09 ***
## totalFert      -6.469e-01  1.219e-01  -5.307 4.11e-07 ***
## GDP            -1.754e-05  5.414e-06  -3.239 0.001490 **
## lifeExpect     -8.275e-02  2.142e-02  -3.863 0.000169 ***
## medianAge      -5.361e-02  1.819e-02  -2.948 0.003735 **
## democracyScore  1.131e-01  4.356e-02   2.597 0.010382 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8797 on 144 degrees of freedom
## Multiple R-squared:  0.3971, Adjusted R-squared:  0.3761
## F-statistic: 18.97 on 5 and 144 DF,  p-value: 1.841e-14
```

From summary information of this model, we see that all predictors are statistically significant at the $\alpha = 0.05$ level, and compared to the model based on adjusted R-squared, this model based on the BIC contains only 5 predictor variables. However, the R-squared value of this model is lower, indicating that this model explains only about 40% of the variation in the log of homicide rate.
Here are the residual plots of the best model according to the BIC:





The residual plots look pretty good. Other than the top-right corner, the normal quantile plot of residuals appears approximately linear, and the fits vs residuals plot does not show signs of heteroskadacity.

Best Model According to the C_p Statistic

Next I try using the C_p statistic to determine a model. This model contains the following predictors:

```
modCP <- min(c(1:length(mod1sum$cp))[mod1sum$cp < c(1:length(mod1sum$cp)) + 1])
names(WF8)[mod1sum$which[modCP, ]][-1]
```

```
##  [1] "CO2"            "currentAccount" "exportsUSD"     "importsUSD"
##  [5] "infMortality"   "totalFert"      "GDP"            "lifeExpect"
##  [9] "medianAge"      "democracyScore"
```

The R-squared value for this model is 0.4533, only slightly lower than that of the model based on adjusted R-squared. However, this model also contains a few predictors that are not statistically significant at the $\alpha = 0.05$ level.

Best Model According to the AIC

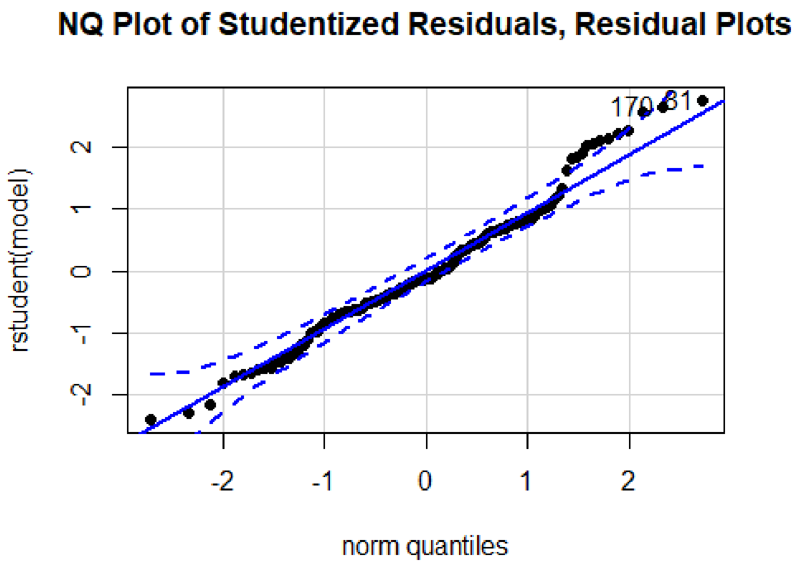
Lastly, I use the Akaike Information Criterion (AIC) to determine a model. This model contains the following 9 predictor variables:

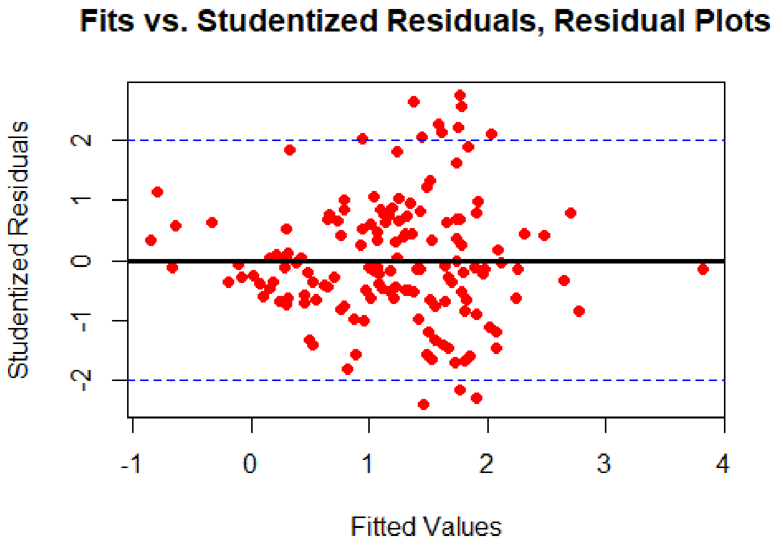
```
npred <- length(mod1sum$bic)
AICvec <- rep(NA, npred)
for (i in 1:npred){
  WFtemp <- WF8[,mod1sum$which[i,]]
  AICvec[i] <- AIC(lm(logHomi ~ .,data = WFtemp))
}
names(WF8[,mod1sum$which[which.min(AICvec), ]])[2:10]

## [1] "currentAccount" "exportsUSD"     "importsUSD"     "infMortality"  
## [5] "totalFert"      "GDP"            "lifeExpect"     "medianAge"     
## [9] "democracyScore"
```

Here is the summary of this model:

```
WFtemp <- WF8[ ,mod1sum$which[which.min(AICvec), ]]
lmAIC <- lm(logHomi ~ ., data = WFtemp)
summary(lmAIC)

## 
## Call:
## lm(formula = logHomi ~ ., data = WFtemp)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -1.9926 -0.5105 -0.1060  0.5318  2.2669 
## 
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)    
## (Intercept)     1.291e+01  2.393e+00   5.392 2.89e-07 ***
## currentAccount -2.744e-12  1.372e-12  -2.000 0.047448 *  
## exportsUSD      2.628e-01  1.125e-01   2.336 0.020929 *  
## importsUSD     -2.876e-01  1.322e-01  -2.176 0.031239 *  
## infMortality   -1.632e-02  1.060e-02  -1.539 0.126104    
## totalFert      -5.565e-01  1.280e-01  -4.347 2.63e-05 ***
## GDP            -1.995e-05  5.735e-06  -3.479 0.000670 ***
## lifeExpect     -1.089e-01  2.918e-02  -3.732 0.000276 ***
## medianAge      -4.864e-02  1.829e-02  -2.659 0.008756 ** 
## democracyScore  1.053e-01  4.269e-02   2.466 0.014867 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.854 on 140 degrees of freedom
## Multiple R-squared:  0.4475, Adjusted R-squared:  0.412 
## F-statistic:  12.6 on 9 and 140 DF,  p-value: 1.607e-14
```

The R-Squared value for this model is 0.4475, indicating that it accounts for about 44.75% of the variation in the log of homicide rate. All but one variable is statistically significant at the $\alpha = 0.05$ level. Here are the residual plots for the best model according to the AIC:





Comparing the R-Squared values and residual plots of the BIC model and this model, I noticed that the normal quantile plot of residuals of this model appear more linear than that of the BIC model, which means that the residuals of this model are more approximately normal than the residuals of the BIC model. The adjusted R-squared model and the C_p statistic model both contain too many predictors, so I decided to not consider them.

Conclusions and Summary

First, we see that life expectancy and median age are strongly and positively correlated, with a statistically significant correlation. This indicates that countries with higher median ages tend to also have higher life expectancies at birth (which is not surprising).

Second, we see that the true difference in the mean democracy index of countries that recognize same-sex marriage and that of countries that do not is not 0. Specifically, the mean democracy index of countries that recognize same-sex marriage is higher than that of countries that do not.

Third, the ANCOVA model for homicide rate, military expenditure, and continent shows that 1) the interaction between continent and log military expenditure is significant in predicting the log homicide rate of a country and 2) that in all continents except Europe and Africa, it seems that lower homicide rate tends to go with higher military expenditure.

Lastly, the best subsets regression gives two models: one based on the BIC and the other based on the AIC. The AIC model has a higher R-squared value, and contains all the predictors that are in the BIC model. These predictors are: total fertility, GDP per capital, life expectancy, median age, and democracy index. On the other hand, the predictors in the BIC model are all statistically significant at the $\alpha = 0.05$ level, although the R-squared value is smaller.

In particular, in both models we see that log homicide rate is negatively correlated to fertility, GDP, life expectancy, and median age; whereas it is positively correlated to democracy index. This would suggest that the higher the democracy index of a country is, the higher the log of its homicide rate is likely to be.