Project Report - Inferential Analysis of Global Piracy

May 27, 2022

Authors

Tuomas Rickansrud - trickansrud@ucdavis.edu

Lizzy Stampher - estampher@ucdavis.edu

Emilio Barbosa Valdiosera - ebarbosavaldiosera@ucdavis.edu

Instructor: Emanuela Furfaro

STA141A - Fundamentals of Statistical Data Science

University of California, Davis

Group Contributions

Tuomas Rickansrud: Programming functions and Statistical methods

Lizzy Stampher: Writing the Introduction, Conclusion, Methodologies, and discussion

Emilio Barbosa Valdiosera: Researching and Obtaining data, Visualizations, Formatting the Final output, and Data Descriptions

1. Introduction and Questions

Maritime piracy is a resurgent criminal enterprise with significant human and economic impact. Global trade occurs primarily by sea, and exposure to pirate attacks has increased correspondingly with growth in trade volumes. Contrary to the persistent popular fantasy image, modern pirates are often trained in advanced combat and utilize speedboats, automatic weapons, anti tank missiles, and grenades in their operations. They often target cargo ships and hold crew members hostage for ransom, with the longest known captivities spanning several years.

The goal of this project is to use data collected and compiled by the Journal of Open Humanities Data on pirate attacks to analyze potential risk factors associated with their occurrence. To that end, we seek to answer the following questions:

- Which countries, or regions, see the greatest incidence of pirate attacks?
- Which socioeconomic factors correlate most strongly with the incidence of pirate attacks?
- What other factors, such as type or status of the vessels targeted, correlate most strongly with the incidence of pirate attacks?
- Is piracy globally increasing or decreasing?

2. Dataset Introduction

For this project, we seek to find conclusions regarding the phenomenon of Global Piracy through data recorded on each pirate attack across the globe from January 1993 to December 2020. To accomplish this, we are using two main datasets, along with one smaller descriptive dataset. The first and main dataset is 'pirate_attacks' which entails 16 variables with 7511 observations. These 16 variables are various aspects of individual pirate attacks, such as the location of the attack in longitude and latitude coordinates, the date and time of the attack, and the descriptions of the vessels attacked. The second dataset, labeled as 'country_indicators', contains socioeconomic data on a yearly basis from 1993 to 2020 of every country on the globe and some dependencies. This data is provided so we may conduct analysis on socioeconomic factors against pirate attacks to see if there is a correlation between a country's socioeconomic status and the level of piracy present. It is important to note that only countries that have incidents of pirate attacks will have their economic factors analyzed. For example, the United States currently has one of the largest militaries on the planet, but is not mentioned in this study because there are basically 0 incidences of piracy. Finally, we have included a third dataset of country codes ('country_codes') which works as a reference for the other datasets, as the countries listed in the other two sets are displayed in their ISO standard three letter form.

Descriptive Statistics

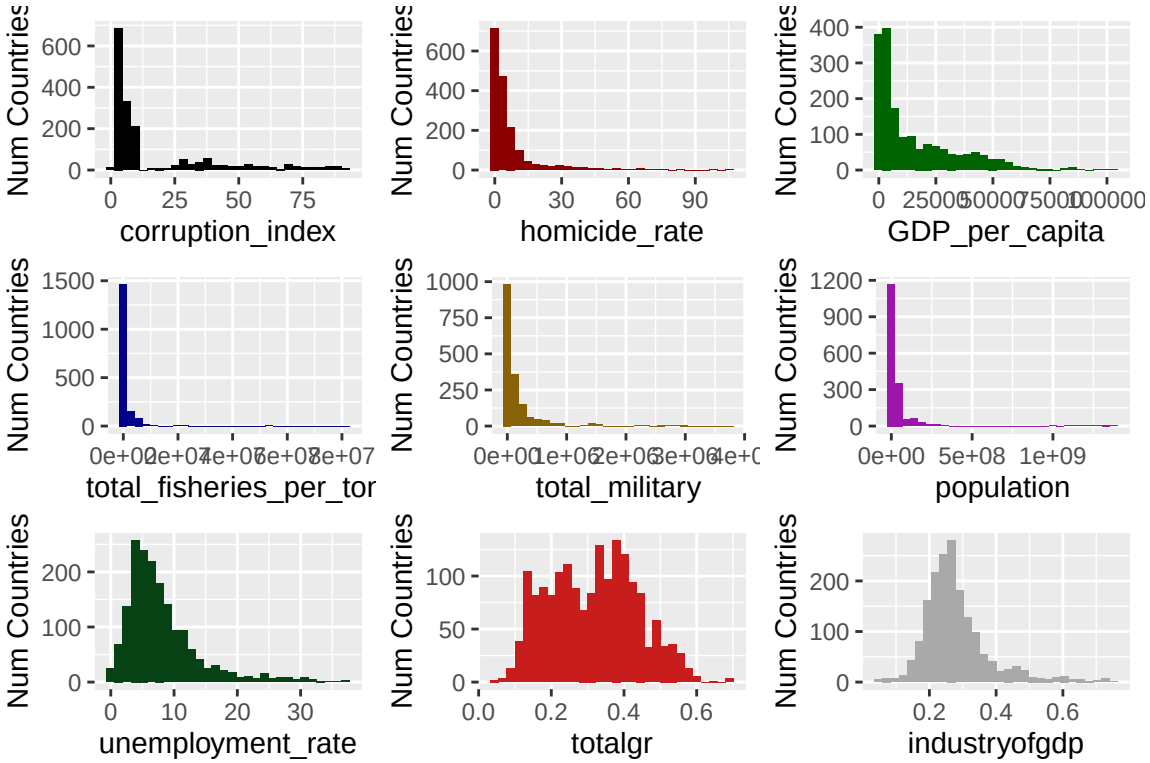
Looking at the distributions of the data reveals some interesting patterns. Six out of the nine variables used in the regression models are heavily right skewed, with the last three being normally distributed, albeit with a slight right skew as well.

Table 1. Descriptive Statistics showing Inter-quartile ranges on all variables used.

corruption_index	homicide_rate	GDP_per_capita	total_fisheries_per_ton	total_military
Min. : 0.40	Min. : 0.000	Min. : 150.2	Min. : 15	Min. : 50
1st Qu.: 3.30	1st Qu.: 1.202	1st Qu.: 2215.8	1st Qu.: 16973	1st Qu.: 16000
Median : 6.10	Median : 2.389	Median : 6727.1	Median : 155846	Median : 47000
Mean :17.48	Mean : 7.100	Mean : 15611.1	Mean : 1304294	Mean : 196619
3rd Qu.:28.00	3rd Qu.: 7.205	3rd Qu.: 24551.0	3rd Qu.: 652204	3rd Qu.: 177450
Max. :92.00	Max. :105.231	Max. :102913.4	Max. :81500000	Max. :3755000

population	unemployment_rate	totalgr	industryofgdp	num_attacks
Min. :2.740e+05	Min. : 0.200	Min. :0.04763	Min. :0.04556	Min. : 0.000
1st Qu.:4.303e+06	1st Qu.: 4.189	1st Qu.:0.21535	1st Qu.:0.21895	1st Qu.: 0.000
Median :1.029e+07	Median : 6.558	Median :0.32261	Median :0.26166	Median : 0.000
Mean :5.225e+07	Mean : 8.030	Mean :0.31481	Mean :0.27959	Mean : 1.892
3rd Qu.:3.541e+07	3rd Qu.: 9.997	3rd Qu.:0.40298	3rd Qu.:0.31068	3rd Qu.: 0.000
Max. :1.379e+09	Max. :37.250	Max. :0.69281	Max. :0.74812	Max. :134.000

Figure 1. Distributions of each variable used.



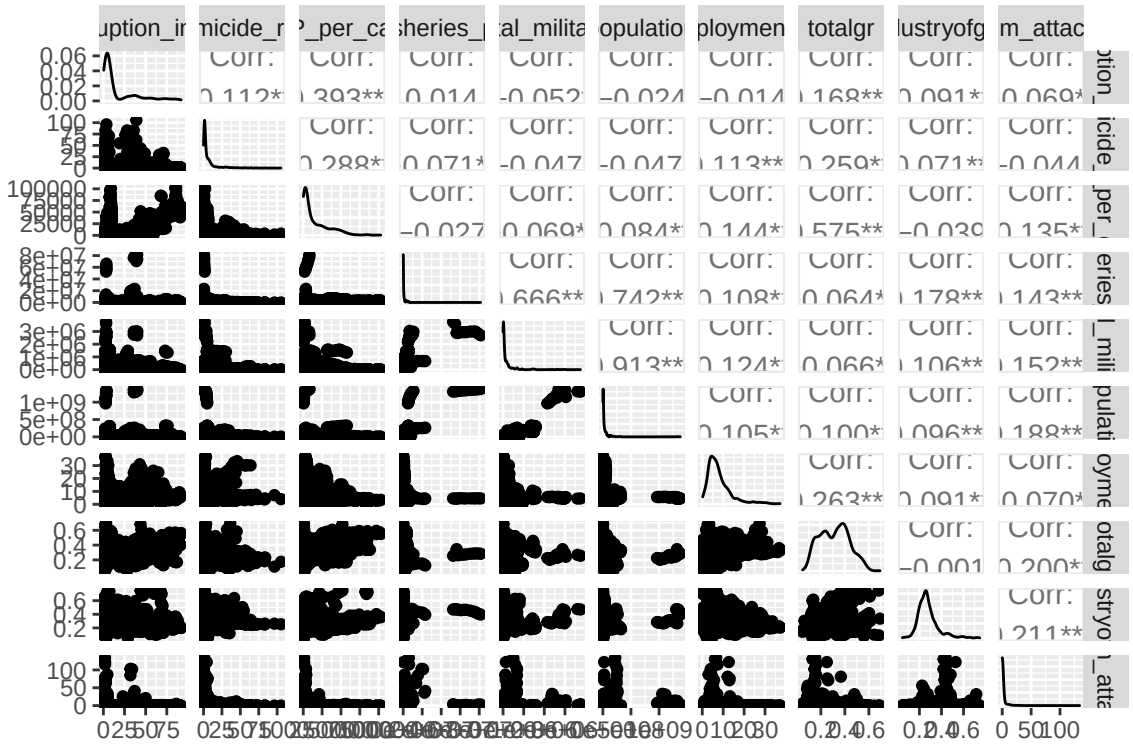
Methodology and Results

A number of methods will be employed to address these research questions. For the matter of frequencies, the data can be tabulated to reveal the greatest incidence of pirate attacks per region (question 1) as well as provide insight into the prevalence of certain categorical factors (question 3). For the matter of modeling socioeconomic variables, multiple linear regression will be applied (question 2), and the global trend of attack occurrences (question 4) can be graphed sequentially.

Data Cleaning

Due to the large size of the data sets and prevalence of missing values, cleaning processes were applied to increase the strength of the variables used in the regression analyses. The number of missing values was counted for each variable, and those with substantial numbers of missing values were removed. Rows containing NAs were also removed, leaving a data set spanning 6294 observations of 12 variables with no missing values. Further, a correlation matrix was used to identify collinear variables and exclude them from the regression model:

Figure 2. Correlation Matrix of all variables used.



Population, fishery yields (total_fisheries_per_ton), and GDP per capita were excluded from the model due to high collinearity. The remaining variables of unemployment rate, GDP pertaining to the industrial sector (industryofgdp), total government revenue (totalgr), number of military personnel (total_military), homicide rate, and corruption index were used in the regression model.

Multiple Linear Regression

To address the question of which socioeconomic factors may contribute to the incidence of pirate attacks, we attempted to fit socioeconomic variables pertaining to the countries nearest each attack site onto the number of attacks occurring each year using multiple linear regression.

Table 2. Summary of Multiple Linear Regression Model.

term	estimate	std.error	statistic	p.value
(Intercept)	1.7754113	0.9154261	1.939437	0.0526107
corruption_index	-0.0075833	0.0096155	-0.788659	0.4304195
homicide_rate	-0.0938222	0.0189793	-4.943402	0.0000008
total_military	0.0000024	0.0000005	4.877836	0.0000012
unemployment_rate	0.0598593	0.0392126	1.526531	0.1270606
totalgr	-17.8550618	1.9411946	-9.197976	0.0000000
industryofgdp	19.9451979	2.1997912	9.066860	0.0000000

R-Squared : 0.1112465

In this model, the corruption index and unemployment rate variables appear to be insignificant. Homicide rate, military size, total government revenue, and GDP from the industrial sector were all considered highly significant, with the largest p-value of the four sitting at 1.17e-06. The R-squared is low, at a value of 0.1112.

Figure 3 Residual Plots for Standard Regression Model

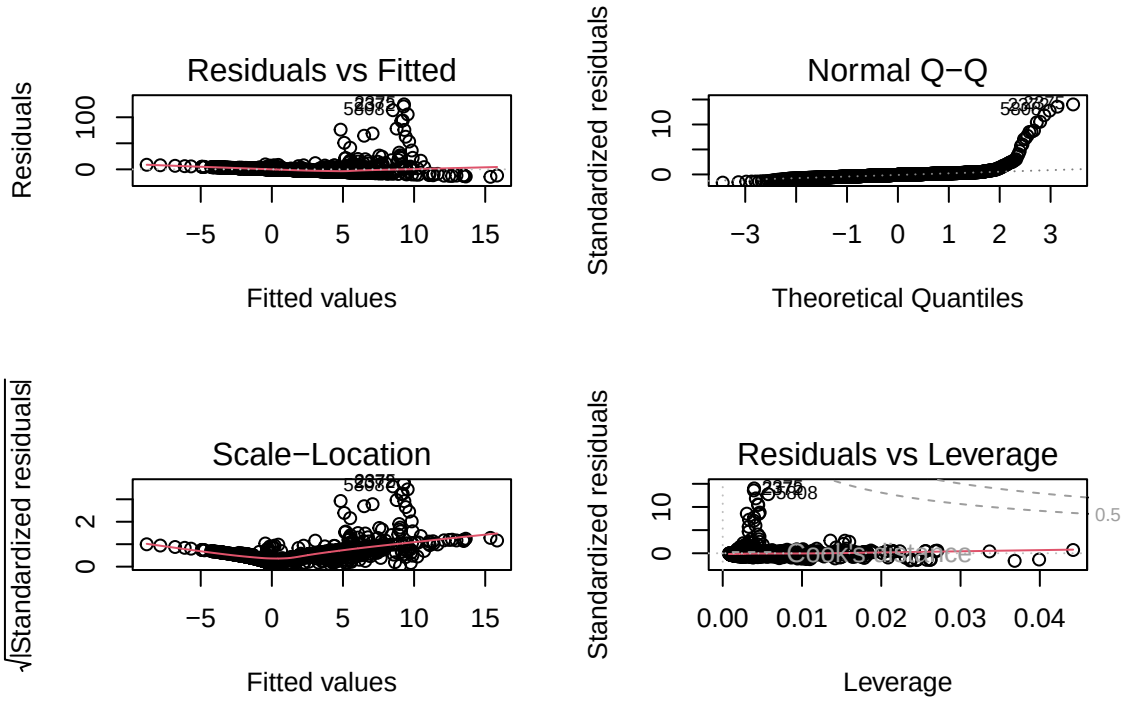


Table 3. Summary of Formal Diagnostic Tests. Tests include Shapiro-Wilks, Durbin-Watson, and Breusch-Pagan.

W Stat	P-Val	Autocorr	DW Stat	P-Val	BP Stat	DF	P-Val
0.3729537	0	0.7893001	0.4212693	0	73.04496	6	0

The residuals vs fitted plot shows a strong indication of autocorrelation and heteroscedasticity with the distinct linear shape through most of the residuals. A Durbin-Watson test confirms autocorrelation in the residuals, with a p-value so small it only outputs 0. The Q-Q plot indicates non-normal and right skewed

residuals. A Shapiro-Wilk normality test further confirms the violation of normality with a p-value smaller than 2.2e-16. The scale-location plot reveals a very interesting cusp-like shape due to the transformation of the residuals restricting them to positive values. It, too, shows a strong indication of autocorrelation and heteroscedasticity. A Breusch-Pagan test of the residuals also reinforces heteroscedasticity, with a p-value of 9.691e-14. There are a handful of outliers and high leverage points, but none of them seem to be of high influence per the Cook's distance criteria. All in all, these residuals show that the model is performing very poorly.

A second regression strategy was implemented by fitting a new model with weights:

Table 4. Summary of the Weighted Linear Regression Model.

term	estimate	std.error	statistic	p.value
(Intercept)	0.8083953	0.0932444	8.669642	0
corruption_index	-0.0017051	0.0002646	-6.443367	0
homicide_rate	-0.0215070	0.0026716	-8.050223	0
total_military	0.0000013	0.0000001	11.813684	0
unemployment_rate	0.0146590	0.0019032	7.702180	0
totalgr	-4.3318782	0.4998049	-8.667138	0
industryofgdp	4.7090850	0.5605500	8.400829	0

R-Squared : 0.0854278

While every variable in this model is highly significant, this model carries the lowest R-squared thus far, at 0.08543. These significant values could simply be the result of model overfitting.

Figure 4. Residual Plots for Weighted Model.

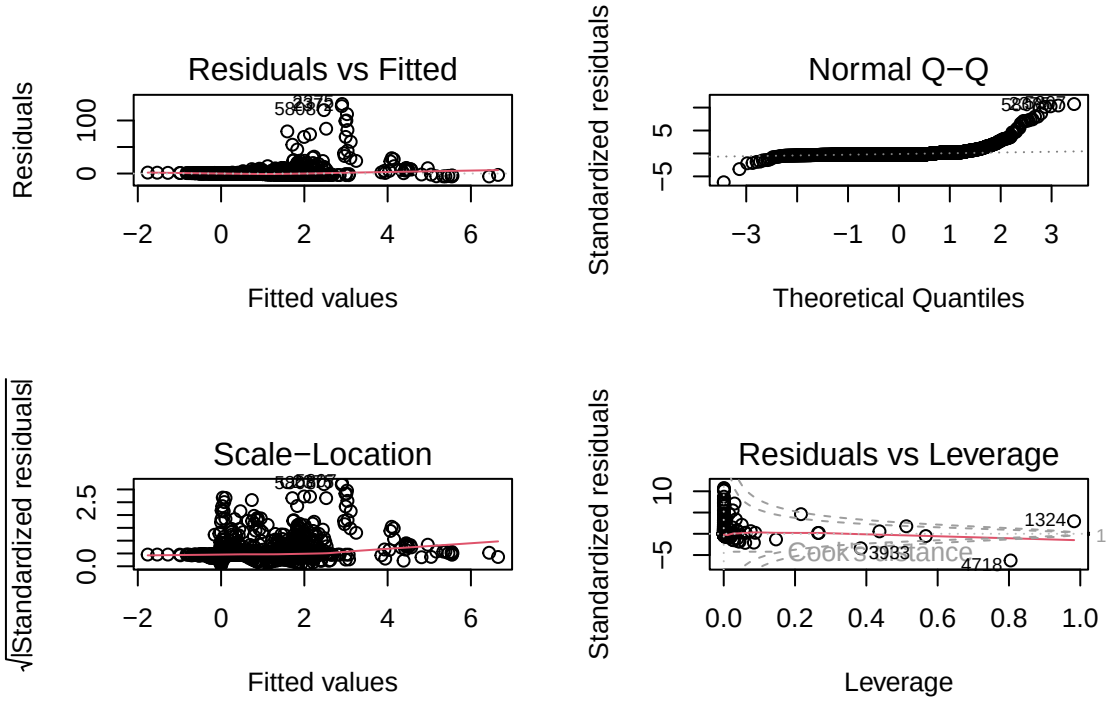


Table 5. Summary of Formal Diagnostic Tests for the Weighted Model. Tests include Shapiro-Wilks, Durbin-Watson, and Breusch-Pagan.

W Statistic	P-Val	Autocorr	DW Stat	P-Val	BP Stat	DF	P-Val
0.381135	0	0.8065116	0.3869576	0	3.486951	6	0.7457049

These plots appear to see no improvement either, and Durbin-Watson and Shapiro-Wilks tests return the same autocorrelated and non-normal results as before. However, the Breusch-Pagan Test returns a surprising p-value of 0.7457, indicating that the residuals are passing as homoscedastic in this case. This result seems dubious given the graph's appearance, but is interesting nonetheless. In fact, the residuals vs leverage plot shows influential points being introduced which were not present in the other models.

Other factors

In order to get a sense of the other factors related to the attacks, frequency tables were generated to assess the proportions of categorical attributes such as type and status of vessels targeted, as well as the type of attack committed.

Table 6. Summary of Vessel types targeted.

Vessel Type	Freq.
Bulk Carrier	322
Product Tanker	252
Container	119
General Cargo	67
Chemical Tanker	53
Tug	45
Crude Oil Tanker	40
LPG Tanker	40
Tanker	30
Fishing Vessel	27

Bulk carriers (28%), product tankers (22%), containers (10%), general cargo ships (6%), and chemical tankers (5%) are the five most commonly targeted vessel types which have been recorded, and together they constitute 71% of all recorded vessel types. However, this portion of the data also contained a very high number of missing values (84% of the data contained no information about the type of vessel targeted) so conclusions about the true proportionality are highly speculative.

The vessel status was also tabulated:

Table 7. Summary of the status of the Vessels while they were attacked

Vessel Status	Freq.
Anchored	3262
Steaming	2587
Berthed	653
Underway	56
Stationary	14
steaming	11
Drifting	10
Moored	2
Bunkering operations	1
Fishing	1

Vessel Status	Freq.
Grounded	1
Towed	1

The three most common reported statuses of targeted vessels are anchored (either at sea or in a harbor) (49%), steaming (vessel is traveling underway at low speeds) (39%), and berthed (vessel is positioned or being prepared for loading/unloading of cargo) (10%). Together they constitute 98% of all recorded vessel statuses at the time of attack. This portion of the data was missing 12% of its information, but is a substantial improvement over the previous vessel type data.

Finally, the types of attacks committed were tabulated:

Table 8. Summary of the types of attacks.

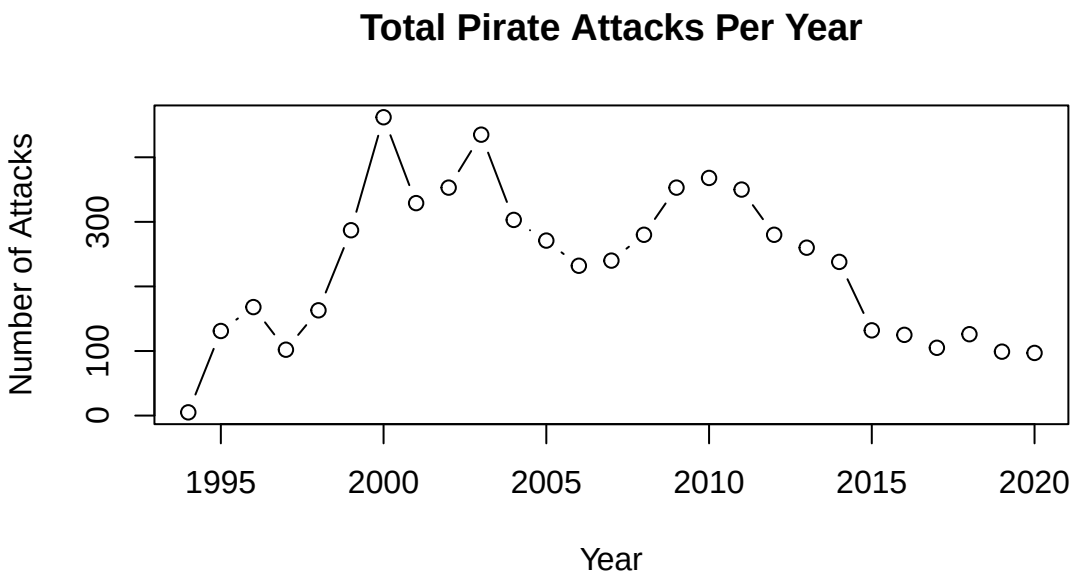
Attack Type	Freq.
Boarded	3421
Attempted	1999
Boarding	1367
Hijacked	511
Fired Upon	73
Suspicious	16
Explosion	3
Detained	1

Of those recorded, the boarding style of attack was the most common (65%). This general description involves an approach along the side of a target vessel, where the attackers are able to board the ship. The second most common attack type is recorded as "attempted" (27%). Presumably, these are attacks which did not fully succeed in their execution. The third most common attack type is hijacking (7%), where the pirates seized control of the entire ship. Together they constitute 99% of total attack types. This data was the most complete of the three sets, with only 1.6% of its values missing, so the highest confidence in true proportionality lies here.

Global trends

To address question 4 about the nature of global trends in pirate attack occurrence, a sequential plot of all pirate attacks per year was obtained:

Figure 5. Total pirate attacks globally over time.



The global number of attacks rises substantially around 2000, dips around 2005 before peaking again in 2010, and finally follows a steady decline through 2020. It appears as though, overall, global piracy has been declining since its initial strong peak in 2000.

Conclusions

While the global trends graph and tabulation data provided an interesting look into the frequency of attacks per region, as well as the nature of the attacks themselves, the multiple linear regression models proved unsuitable. In addition to low R-squared values in these regression models, the distributions of the residuals were highly problematic. One reason for this arises from the time series nature of the data: the sequential measurement of the variables is the most likely explanation for the strong linear trends seen in the residual plots. As such, independence, homoscedasticity, and normality assumptions were all violated and linear regression ultimately yielded poor results. Time series analysis would be a more appropriate means to address this problem of autocorrelation, but due to time, knowledge, and space constraints it could not be implemented beyond the acknowledgement of its likely superiority as a modeling technique.

Bibliography

Asariotis, Regina, Hassiba Benamara, Jennifer Lavelle, and Anita Premti. 2014. "Maritime Piracy. Part i: An Overview of Trends, Costs and Trade-Related Implications."

Benden, Paul, Alan Feng, Christopher Howell, and Giulio Valentino Dalla Riva. 2021. "Crime at Sea: A Global Database of Maritime Pirate Attacks (1993–2020)." *Journal of Open Humanities Data* 7.

Luft, Gal, and Anne Korin. 2004. "Terrorism Goes to Sea." *Foreign Affairs*, 61–71.

Nations, United. 2015. "Longest-Held Hostages in Somalia's History Released." *UNODC*, 1–2.

Appendix

```
knitr::opts_chunk$set(echo = TRUE, fig.width=6, fig.height=4, warning = FALSE)
rm=(list=ls())
country_indicators = read.csv("country_indicators.csv")
pirate_attacks = read.csv("pirate_attacks.csv")
country_codes = read.csv("country_codes.csv")

names(country_indicators)[5] = "GDP_per_capita"
pay = pirate_attacks
pay$date = strtrim(pay$date, 4)
clean.pay = na.omit(pay[,c(-2, -12, -13, -14)])
clean.ci = country_indicators[complete.cases(country_indicators$year),]
fpay = merge(clean.pay, clean.ci, by.x = c("date", "nearest_country"), by.y = c("year", "country"))
attacks_per_year = function(dataset1, dataset2){
  num_attacks = c()
  for (country in unique(na.omit(dataset1$country))){
    for (year in na.omit(dataset1[dataset1$country == country,]$year)){

      num_attacks = append(num_attacks, nrow(dataset2[dataset2$date == year & dataset2$nearest_country =
    }
  }
  return (num_attacks)
}
num_attacks = attacks_per_year(country_indicators, clean.pay)
ci.attacks = cbind(clean.ci, num_attacks)
clean.cia = na.omit(ci.attacks)
clean.cia$year = factor(clean.cia$year)
mod = lm(num_attacks~., data = clean.cia[,c(-1,-2,-6,-8,-5)])
wt = 1 / lm(abs(mod$residuals)~mod$fitted.values)$fitted.values^2

wt.mod = lm(num_attacks~., data = clean.cia[,c(-1,-2,-6,-8,-5)], weights = wt)
knitr::kable(summary(clean.cia[,3:7]))
knitr::kable(summary(clean.cia[,8:12]))
library(ggplot2)
library(gridExtra)
ggp1 = ggplot(clean.cia[,c(-1,-2)]) +
  geom_histogram(mapping = aes(corruption_index), bins = 30, fill = "Black") +
  labs(y = "Num Countries")
ggp2 = ggplot(clean.cia[,c(-1,-2)]) +
  geom_histogram(mapping = aes(homicide_rate), bins = 30, fill = "Dark Red") +
  labs(y = "Num Countries")
ggp3 = ggplot(clean.cia[,c(-1,-2)]) +
  geom_histogram(mapping = aes(GDP_per_capita), bins = 30, fill = "Dark Green") +
  labs(y = "Num Countries")
ggp4 = ggplot(clean.cia[,c(-1,-2)]) +
  geom_histogram(mapping = aes(total_fisheries_per_ton), bins = 30, fill = "Dark Blue") +
  labs(y = "Num Countries")
ggp5 = ggplot(clean.cia[,c(-1,-2)]) +
  geom_histogram(mapping = aes(total_military), bins = 30, fill = "#886207") +
  labs(y = "Num Countries")
ggp6 = ggplot(clean.cia[,c(-1,-2)]) +
  geom_histogram(mapping = aes(population), bins = 30, fill = "#9b16aa") +
```

```
  labs(y = "Num Countries")
ggp7 = ggplot(clean.cia[,c(-1,-2)]) +
  geom_histogram(mapping = aes(unemployment_rate), bins = 30, fill = "#064214") +
  labs(y = "Num Countries")
ggp8 = ggplot(clean.cia[,c(-1,-2)]) +
  geom_histogram(mapping = aes(totalgr), bins = 30, fill = "#c71d1d") +
  labs(y = "Num Countries")
ggp9 = ggplot(clean.cia[,c(-1,-2)]) +
  geom_histogram(mapping = aes(industryofgdp), bins = 30, fill = "Dark Grey") +
  labs(y = "Num Countries")
grid.arrange(ggp1, ggp2, ggp3, ggp4, ggp5, ggp6, ggp7, ggp8, ggp9)
library(GGally)
ggpairs(clean.cia[,c(-1,-2)])
mod = lm(num_attacks~., data = clean.cia[,c(-1,-2,-6,-8,-5)])
summod = summary(mod)
knitr::kable(cbind(broom::tidy(mod)))
library(car)
library(lmtest)
par(mfrow=c(2,2))
plot(mod)
s = shapiro.test(rstandard(mod))
st = cbind(s$statistic, s$p.value)
rownames(st) = c()
colnames(st) = c("W Stat", "P-Val")
dw = durbinWatsonTest(mod)
dwt = cbind(dw$r, dw$dw, dw$p)
rownames(dwt) = c()
colnames(dwt) = c("Autocorr", "DW Stat", "P-Val")
bp = bptest(mod)
bpt = cbind(bp$statistic, bp$parameter, bp$p.value)
rownames(bpt) = c()
colnames(bpt) = c("BP Stat","DF","P-Val")
alltest = cbind(st, dwt, bpt)
knitr::kable(alltest)
wt = 1 / lm(abs(mod$residuals)~mod$fitted.values)$fitted.values^2
wt.mod = lm(num_attacks~., data = clean.cia[,c(-1,-2,-6,-8,-5)], weights = wt)
summod2 = summary(wt.mod)
knitr::kable(broom::tidy(wt.mod))
library(car)
library(lmtest)
par(mfrow = c(2,2))
plot(wt.mod)
s2 = shapiro.test(rstandard(wt.mod))
st2 = cbind(s2$statistic, s2$p.value)
rownames(st2) = c()
colnames(st2) = c("W Statistic", "P-Val")
dw2 = durbinWatsonTest(wt.mod)
dwt2 = cbind(dw2$r, dw2$dw, dw2$p)
rownames(dwt2) = c()
colnames(dwt2) = c("Autocorr", "DW Stat", "P-Val")
bp2 = bptest(wt.mod)
bpt2 = cbind(bp2$statistic, bp2$parameter, bp2$p.value)
rownames(bpt2) = c()
```

```
colnames(bpt2) = c("BP Stat","DF","P-Val")
alltest2 = cbind(st2, dwt2, bpt2)
knitr::kable(alltest2)
knitr::kable(head(sort(table(pirate_attacks$vessel_type), decreasing = T), n = 10L), col.names = c("Ves
knitr::kable(sort(table(pirate_attacks$vessel_status), decreasing = T), col.names = c("Vessel Status", "
knitr::kable(sort(table(pirate_attacks$attack_type), decreasing = T), col.names = c("Attack Type", "Fre
year = c()
num_attacks_per_yr = c()
  for (yr in na.omit(unique(clean.pay$date))){
    year = append(year, yr)
    num_attacks_per_yr = append(num_attacks_per_yr, nrow(clean.pay[clean.pay$date == yr,]))
  }
tapy = cbind(year, num_attacks_per_yr)
par(mfrow = c(1,1))
plot(tapy, type = "b", ylab = "Number of Attacks",
     main = "Total Pirate Attacks Per Year", xlab = "Year")
```