Exploring Country Gender Equality

Names: Kat Berman and Max Yuhas

Introduction

Recent studies in political science have used a gendered political theory lense to suggest a correlation between increased female participation in traditionally male sectors, such as agriculture, economics, and government, with more environmentally friendly carbon emissions. (United Nations Development Program, "Overview of linkages between gender and climate change," 2013). In addition, recent articles in political science have also suggested a relationship between increased democratization and gender equality (Valentine M. Moghdam, "The Gender of Democracy: The Link Between Women's Rights and Democratization in the Middle East," 2008). Based on this information, we were interested in further exploring this correlation and developing our understanding of the relationship between gender equality and other indicator values for countries around the world. In this project, we decided to compare the gender equality index for countries to their data on female participation in the agricultural sector, carbon emissions data, country GDP, democracy level, and region of the world.

Data Scraping and Cleaning

- Country: (Character) List of names of all countries for which we have complete data
- Total Agro Holding: (Numeric) Total number of agricultural holders ("economic unit of agricultural production under single management comprising all livestock kept and all land used wholly or partly for agricultural production purposes, without regard to title, legal form, or size") in a Country (Food and Agriculture Organization of the United Nations, "Gender and Land Rights Database," http://www.fao.org/gender-landrights-database/data-map/statistics/en/)
- Female Agro Holding: (Numeric) The number of female agricultural holders in a country
- Percentage Female: (Numeric) The percentage of female agricultural holders out of all agricultural holders in a country

Note: The previous three data variables have values that are all from random years due to availability of data. We assume that these values do not change drastically from year to year.

- CO2 Emissions: (Numeric) Total annual carbon dioxide emissions for a country (in gigatons)
- GDP (Millions USD): (Numeric) Total annual GDP for each country in millions of dollars
- GDP from Agro (Millions USD): (Numeric) The total GDP for each country from the agricultural sector in millions of USD
- Region: (Factor) Region of the world that the country is in with levels Africa, Asia, Europe, Middle East, North America or South America.
- Global Gender Gap Index: (Numeric) A rank of the countries gender gap based on health, education, economy, and politics to assess gender equality in a country. 1 signifies complete equality for women and 0 is the lowest equality for women. This index is given to countries by the World Economic Forum.
- Democracy Score: (Numeric) A score from one to ten (one the least and ten the most democratic) ranking how democratic a country is. This index is given to countries by the Economist Intelligence Unit.
- Democratic Category: (Factor) A qualitative representation categorizing the countries based on their democracy score with levels full democracy, flawed democracy, hybrid regime, or authoritarian.

First, the data on agricultural land holdings, Carbon Dioxide emissions, and GDP were downloaded from the Food and Agriculture Organization of the United Nations database. Data for Global Gender Gap Index, Democracy Score, and Democracy Category were scraped off of Wikipedia. The main aspects of this data cleaning revolved around merging each individual data set from the different sources onto the master data

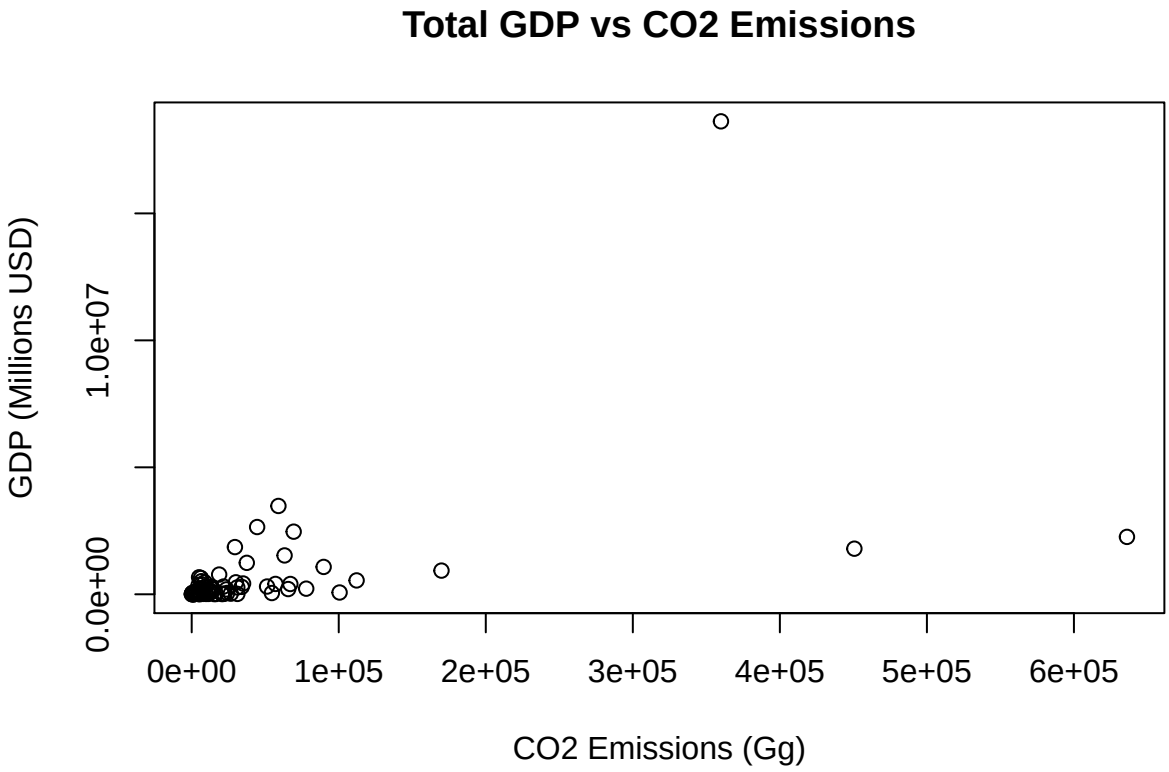
frame. In order to do this, it was necessary to go through the country names from each source and edit those that had alternative spellings so that the same country would be recognized from different data sources (e.g. "United States" vs "United States of America"). In addition, during the data scraping process, we collected a number of variables that were repetitive/not necessary for the final dataframe, so we removed these variables from the final dataset. Also, some countries were the only nations in a region, so we edited the regions for these countries to include them in a larger region (e.g. Egypt was the only country in the "Middle East and Africa", we just included it in the "Middle East" region). After cleaning all of the variables data, we converted types of data of some variables so that Country, Democracy Category, and Region are factors whereas Democracy Score, Total Agro Holding, Female Agro Holding, and Percentage Female are numeric. Finally, we merged all data onto the master data frame and omitted any countries with missing data and attached the variables.

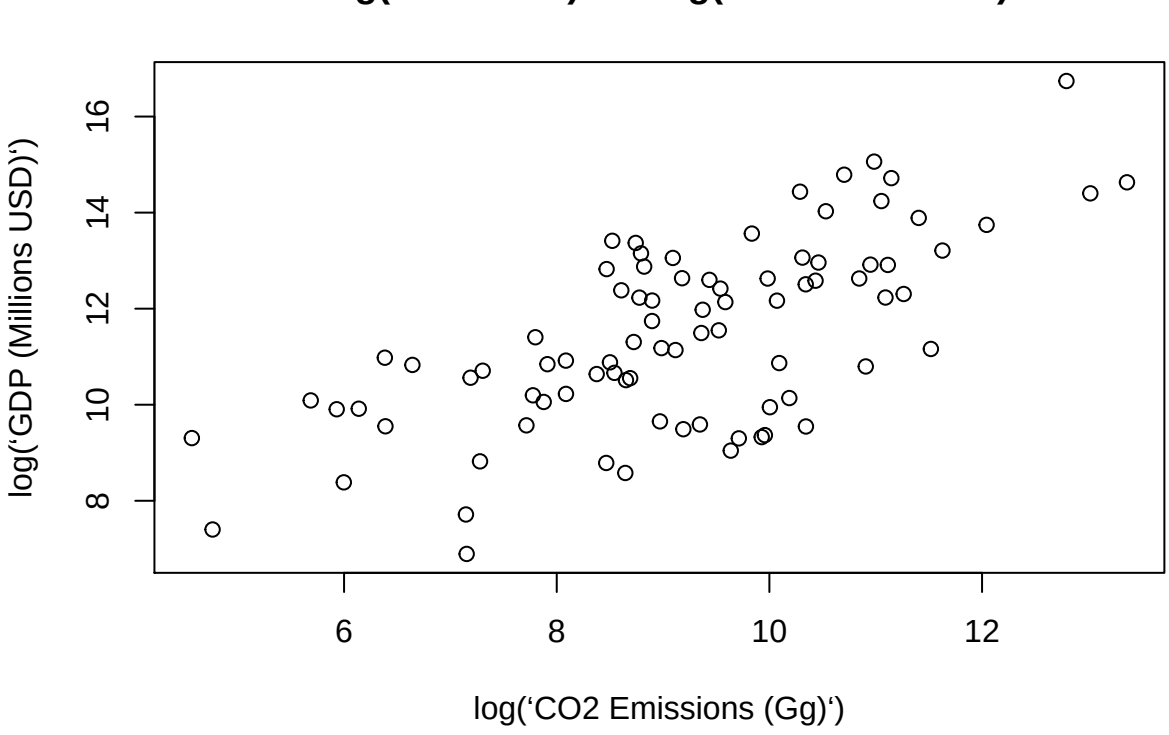
```
head(masterframe)
```

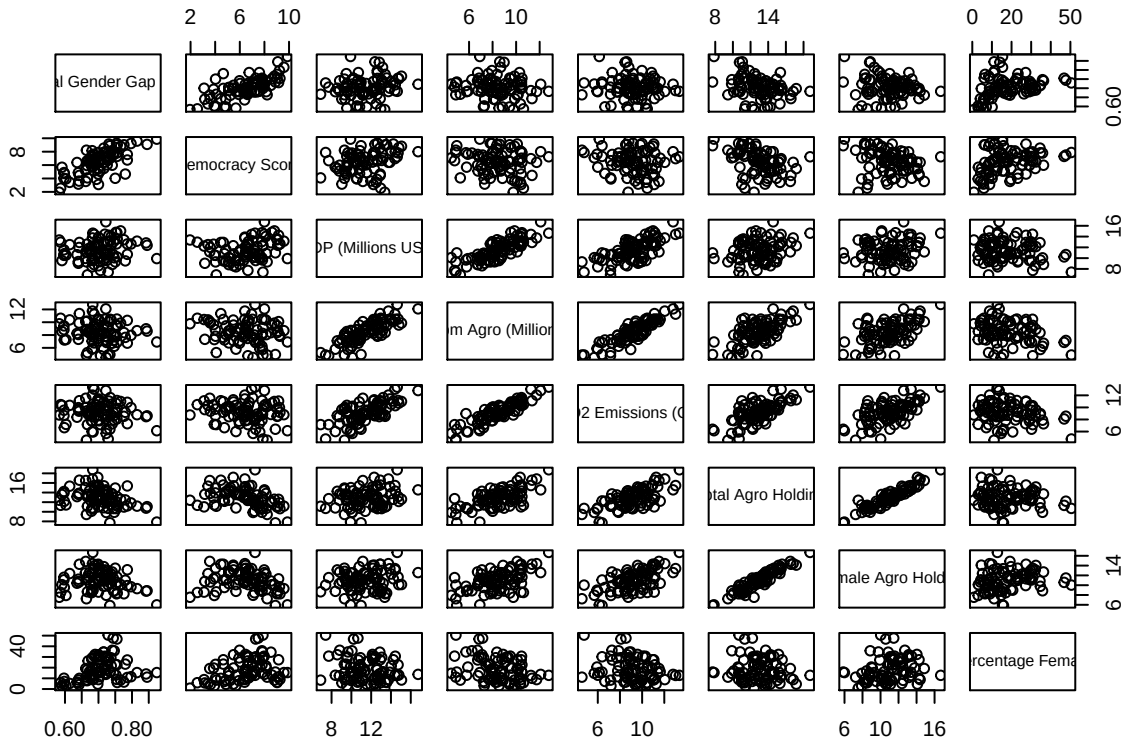
```
##      Country        Region Global Gender Gap Index Democracy Score
## 1    Algeria        Africa                 0.642            3.61
## 2  Argentina South America                 0.735            7.03
## 4    Austria        Europe                 0.716            8.53
## 5 Bangladesh          Asia                 0.698            5.47
## 6    Belgium        Europe                 0.745            7.79
## 7   Botswana        Africa                 0.715            7.84
##   Democracy Category GDP (Millions USD) GDP from Agro (Millions USD)
## 1      Authoritarian          159049.15                   19556.2868
## 2   Flawed democracy          545866.16                   34779.2077
## 4     Full democracy          390799.99                    4305.0097
## 5      Hybrid regime          220836.73                   31017.7290
## 6   Flawed democracy          467955.27                    2887.0560
## 7   Flawed democracy           15566.06                     311.2894
##   CO2 Emissions (Gg) Total Agro Holding Female Agro Holding
## 1          11762.193             1023799               41793
## 2         112150.725              202423               32768
## 4           6782.998              150170               51780
## 5          77885.750            28695763             1322937
## 6           8885.319               42850                6450
## 7           7869.694               50690               17576
##   Percentage Female
## 1               4.1
## 2              16.2
## 4              34.5
## 5               4.6
## 6              15.1
## 7              34.7
```

Graphics, Data Visualization, Testing

Scatterplots

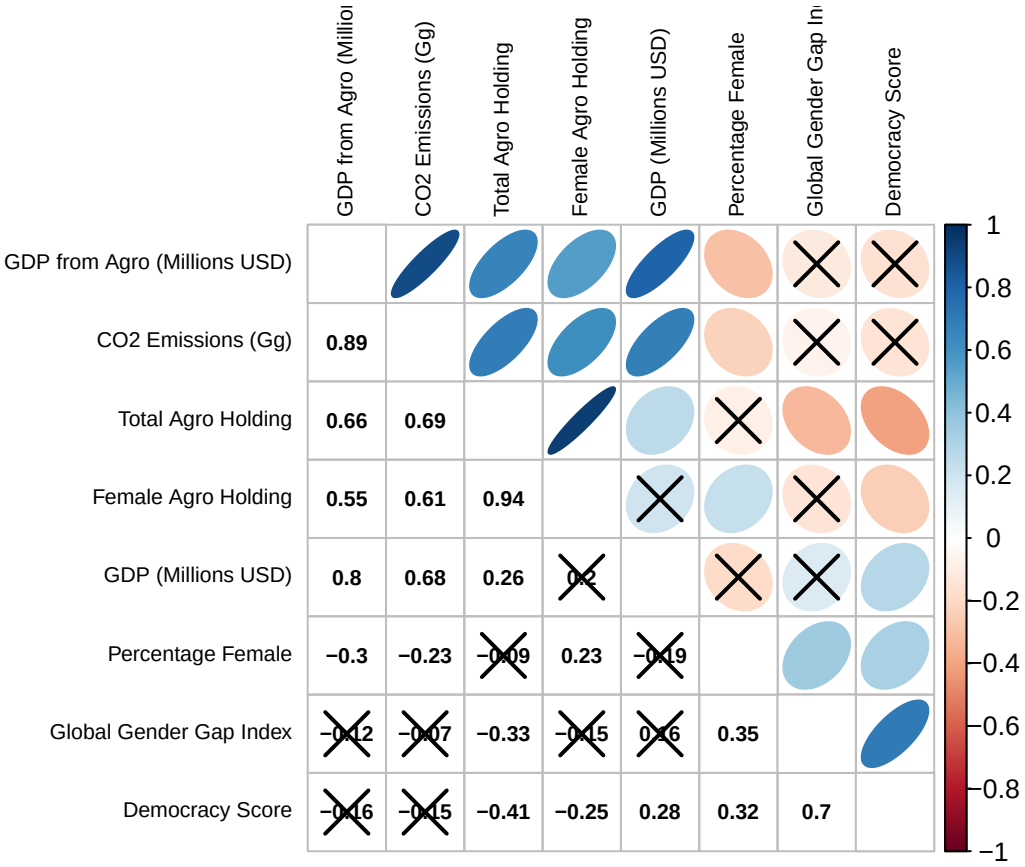

Total GDP vs CO2 Emissions
GDP (Millions USD)
0.0e+00
1.0e+07
0e+00
1e+05
2e+05
3e+05
4e+05
5e+05
6e+05
CO2 Emissions (Gg)


Log(Total GDP) vs Log(CO2 Emissions)
log('GDP (Millions USD)')
8
10
12
14
16
6
8
10
12
log('CO2 Emissions (Gg)')



There is a clear positive trend between democracy score and the global gender gap index score (GGGI). In addition, there is a positive trend between percentage of agricultural landholdings help by females and the democracy score and the GGGI score. However, it is difficult to determine whether there are any strong correlations between many of the variables because most plots have clusters in some areas. Because of this, we expected that transformations would be useful to determine connections between numeric variables. We used a logarithmic transformation on the numeric data variables (except GGGI, Democracy Score, and Percentage Female) to better see trends in the data. Above is an example of Carbon Dioxide Emissions vs Total GDP without and with the log transformation, clearly the log-transformed scatterplot allows us to better see the connection between the data. Thus, for the remainder of the project, we use the log-transform total GDP, GDP from Agro, CO2 emissions, Total Agro Holdings, and Female Agro Holding. With this transformation, we can see a positive correlation between (1) GGGI and Democracy Score, (2) GDP and GDP from Agro, (3) GDP and CO2 emissions, (4) GDP from Agro and Total Agro Holdings, (5) GDP from Agro and Female Agro Holdings, (6) CO2 emissions and Female agro holdings, and (7) total agro holdings and female agro holdings.

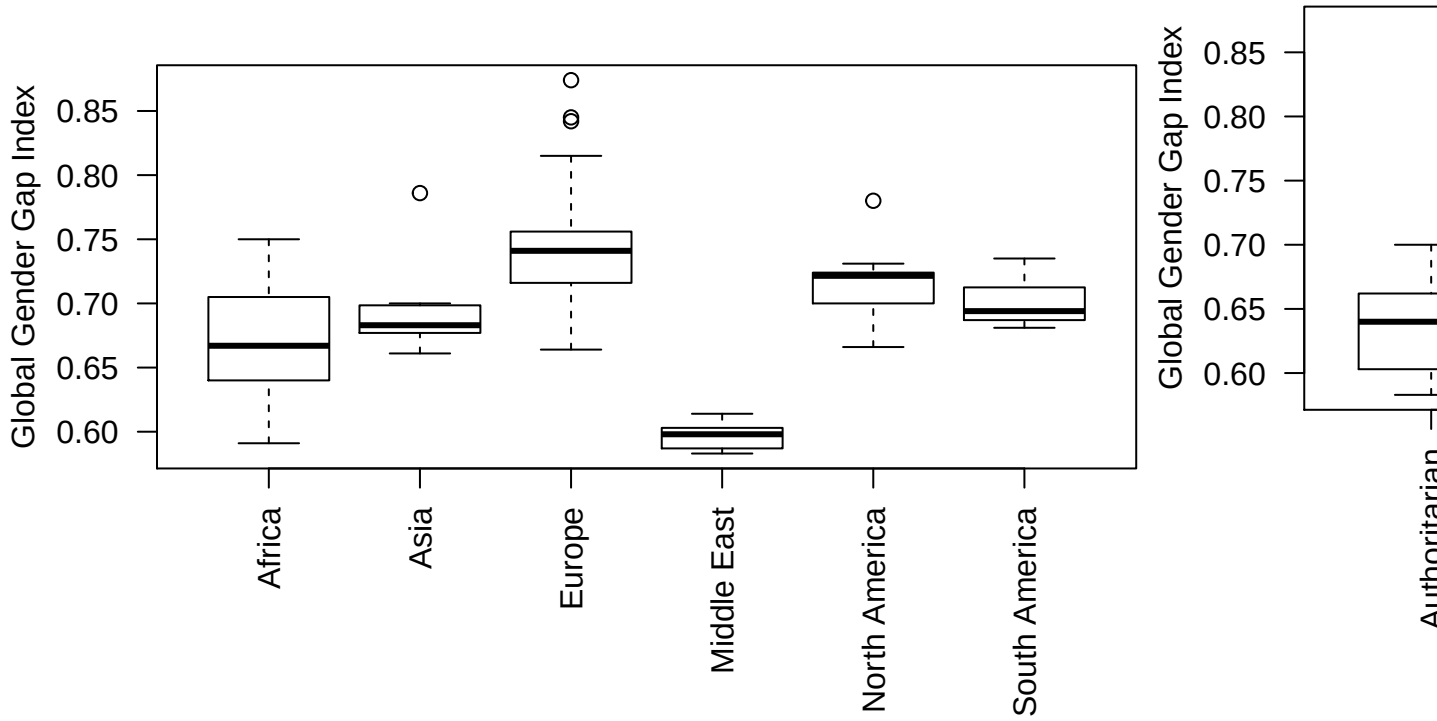
Correlation Plots and Correlation Permutation



Without looking at each individual scatterplot on a larger scale, it can be difficult to tell if there is any significant correlation in some of the scatterplots. This corrplot function gives some added insight on the above scatterplots. It allows us to see correlations between variables and how significant these correlations are (insignificant correlations are crossed out). There is a negative correlation between percentage female and both GDP from Agro and CO2 emissions, as well as a negative correlation between democracy score and total agro holding. In general, the trends seen in the scatterplots above are confirmed by the corrplot (such as high positive correlation between GGGI and Democracy Score).

By doing permutation tests on correlation, we can confirm what we see in the Correlation Plot above. The permutation tests have a null hypothesis of correlation equal to zero and an alternative hypothesis of non-zero correlation. When permuting for Percentage Female vs GGGI, we get a p-value of 0.0017 (less than alpha=0.05). Thus, we reject the null hypothesis(therefore it makes sense that it is not crossed out above as insignificant). However, when we look at Total GDP vs GGGI, we get a p-value of 0.1643 (greater than alpha=0.05). Thus, we fail to reject the null hypothesis(again it makes sense then that it was crossed out above). Therefore, we would expect Percentage Female to be a good predictor of GGGI and GDP to not be a good predictor of GGGI when we look at our regression models.

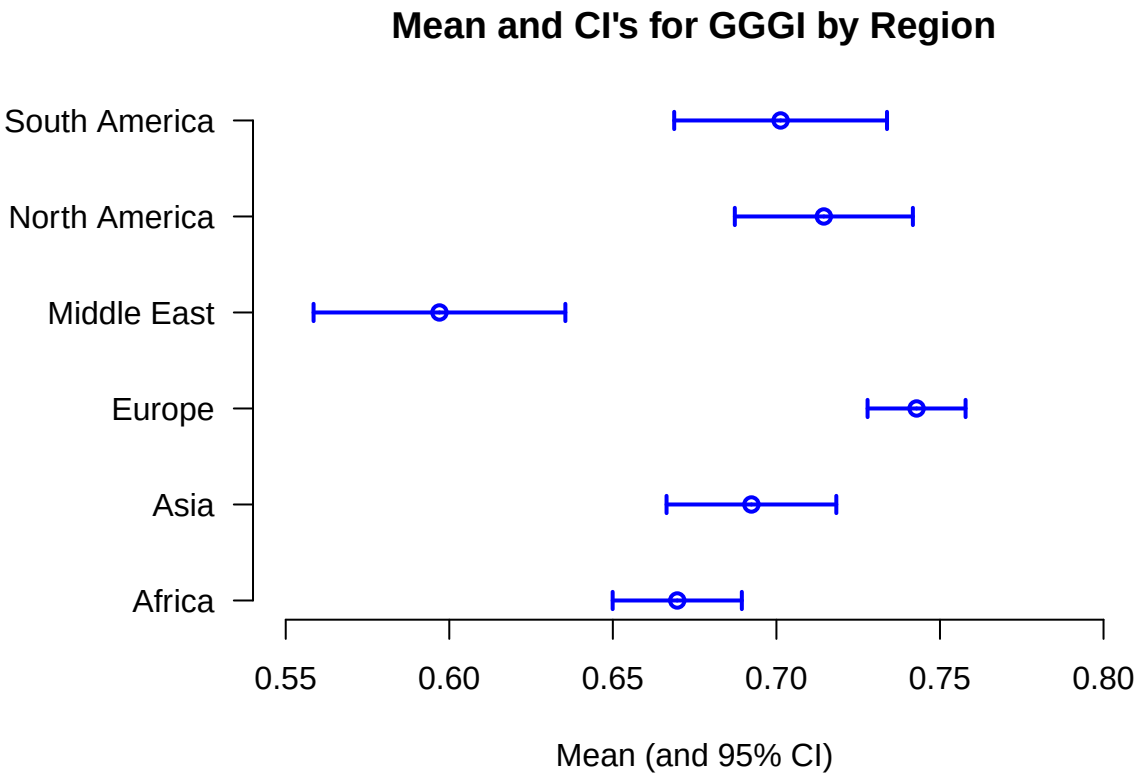
Categorical Variable Analysis (Boxplots, T-tests, Bootstrapping)

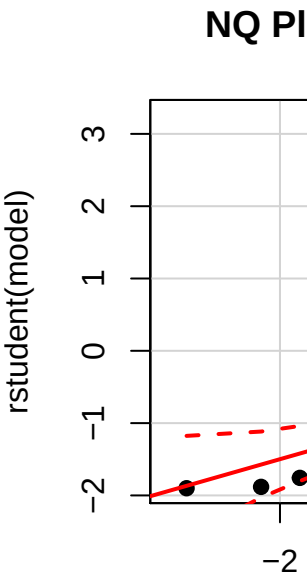


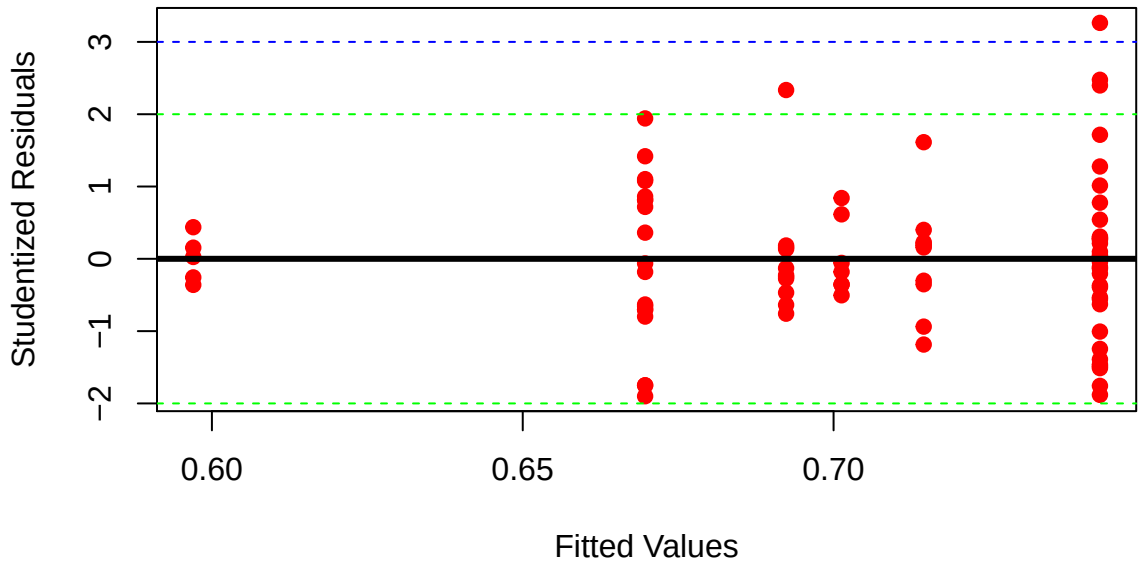
These two boxplots help visualive how GGGI changes for our different categorical variables. For region, we see that GGGI is highest for Europe and lowest for the Middle East. Based on this plot, variance seems to change by region, this may cause some issues with heteroskedasticity later on. For democracy category, GGGI is highest for full democracies and lowest in authoritarian regimes. The variances here all look relatively similar. In general, these graphs make sense based on what we know about different country's political regimes and attitudes towards women.

```
##             Df Sum Sq  Mean Sq F value  Pr(>F)
## Region       5 0.1320 0.026407   14.13 7.3e-10 ***
## Residuals   79 0.1477 0.001869
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  Global Gender Gap Index and Region
##
##               Africa  Asia   Europe  Middle East North America
## Asia          0.6804  -      -       -           -
## Europe        1.3e-06 0.0123 -       -           -
## Middle East   0.0123  0.0011 9.9e-09 -           -
## North America 0.0771  0.7344 0.4386  5.1e-05     -
## South America 0.5114  1.0000 0.1644  0.0011      1.0000
##
## P value adjustment method: holm
```







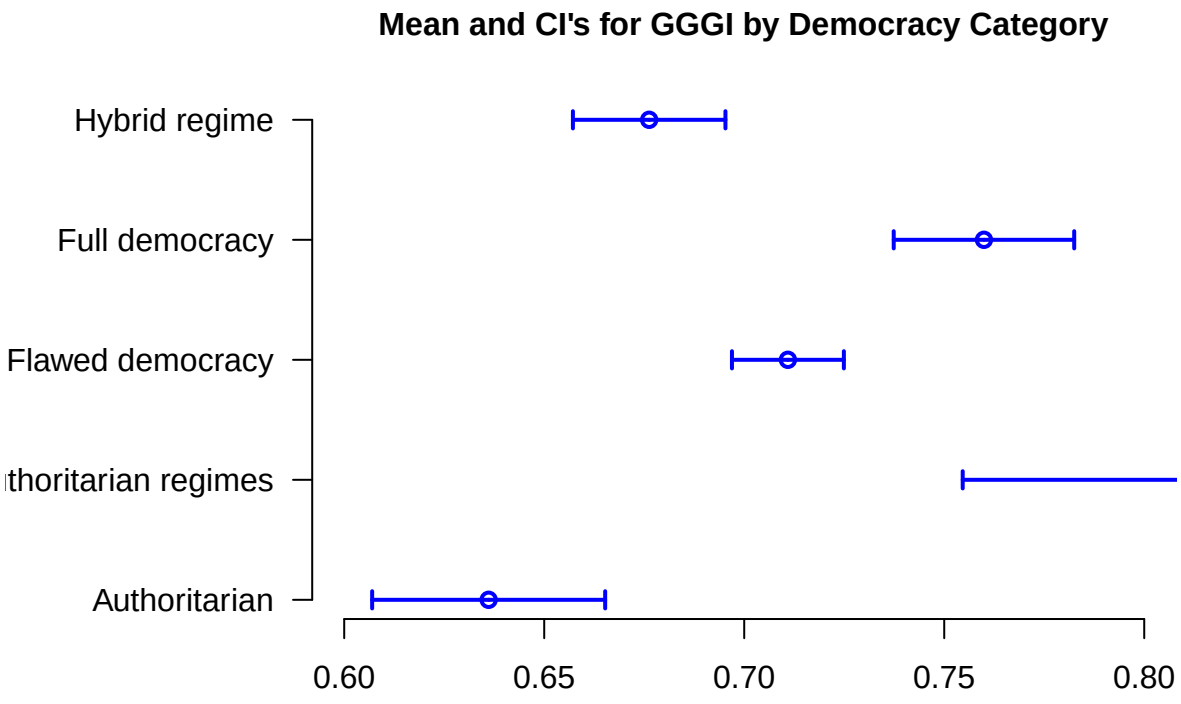

```
##
##  One-way analysis of means (not assuming equal variances)
##
## data:  `Global Gender Gap Index` and Region
## F = 50.346, num df = 5.00, denom df = 25.96, p-value = 1.539e-12

##
##  Kruskal-Wallis rank sum test
##
```

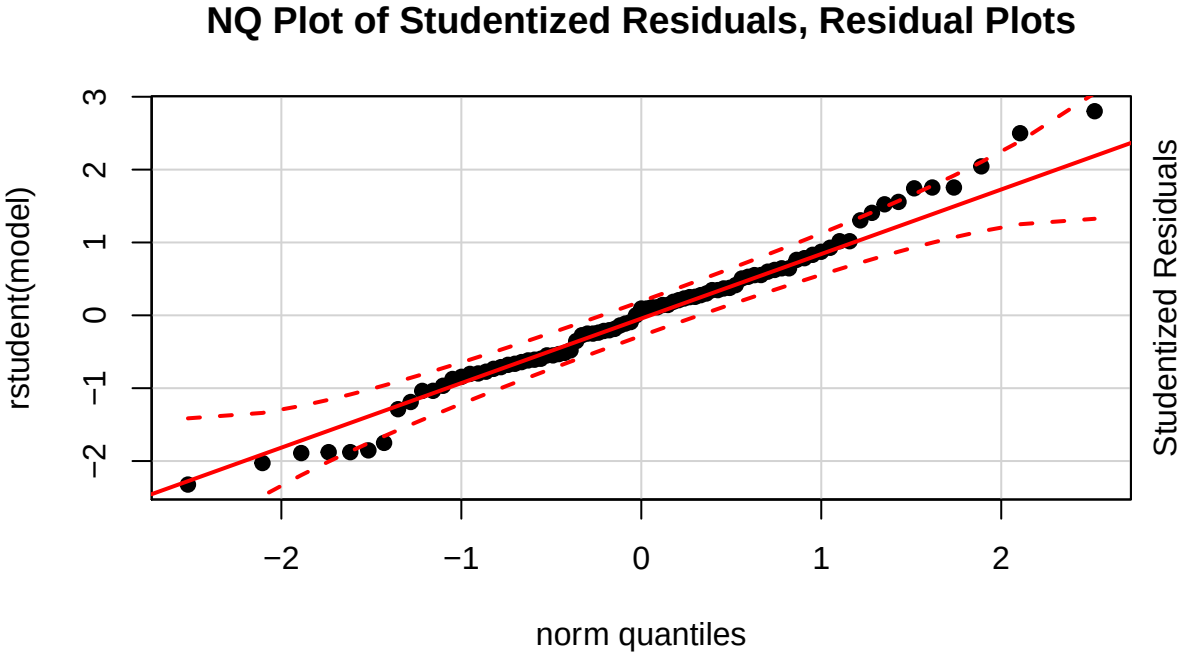
```
## data:  Global Gender Gap Index by Region
## Kruskal-Wallis chi-squared = 35.346, df = 5, p-value = 1.283e-06
```

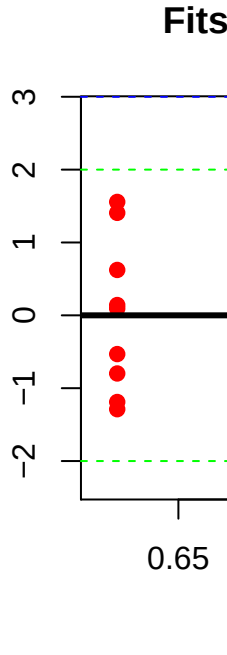
Here we performed a ANOVA model on GGGI by Region. This test shows that there is statistically significant different in mean GGGI by Region (p-value=7.3e-10 < alpha=0.05). Additionally, we did a pairwise t-test to look closer at the different in mean GGGI for each pair of regions. We see that the only statistically significant pairs (alpha=0.05) are Europe-Africa, Europe-Asia, Middle East-Africa, Middle East-Asia, Middle East-Europe, North America-Middle East, and South America-Middle East. We plotted the confidence intervals for mean GGGI and plotted the normal quantiles and residuals by fit. We saw that the residuals are NOT normally distributed and there is strong evidence of heteroskedasticity (ratio of maximum to minimum standard deviations is 4.11). We could have done a Box-Cox procedure, but it does not exactly make sense to transform an index scored from 0 to 1. Instead, we performed a one-way means analysis to confirm the difference in means without assumptions of equal variances (p-value=1.539e-12 < alpha=0.05). Additionally, we performed a Kruskal-Wallis test with no assumption of variance of sample distribution and confirmed the difference in means (p-value=1.283e-06 < alpha=0.05).

```
##                       Df Sum Sq  Mean Sq F value   Pr(>F)
## `Democracy Category`   4 0.1255 0.031368   16.27 8.64e-10 ***
## Residuals             80 0.1542 0.001928
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  Global Gender Gap Index and Democracy Category
##
##                       Authoritarian Authoritarian regimes Flawed democracy
## Authoritarian regimes -             -                     -
## Flawed democracy      -             -                     -
## Full democracy        -             -                     -
## Hybrid regime         -             -                     -
##                       Full democracy
## Authoritarian regimes -
## Flawed democracy      -
## Full democracy        -
## Hybrid regime         -
##
## P value adjustment method: holm
```





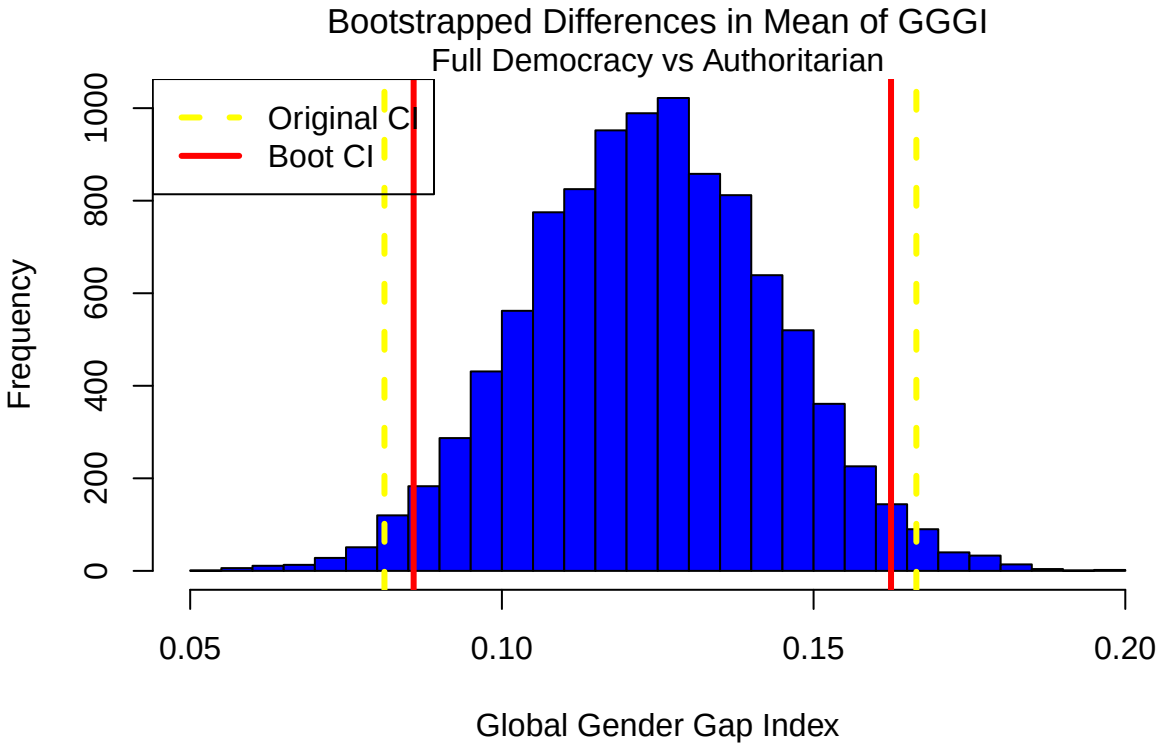


Now we performed a ANOVA model on GGGI by Democracy Category. This test shows that there is statistically significant different in mean GGGI by Democracy Category (p-value=2.21e-10 < alpha=0.05). Additionally, we did a pairwise t-test to look closer at the different in mean GGGI for each pair of democracy styles. We see that all pairs have a statistically significant difference in mean GGGI (alpha=0.05). We plotted the confidence intervals for mean GGGI and plotted the normal quantiles and residuals by fit. We saw that the residuals are normally distributed and there is no evidence of heteroskedasticity (ratio of maximum to minimum standard deviations is 1.77).

```
##          2.5%       97.5%
```

```
## 0.08584333 0.16242222
```

```
## [1] 0.08116244 0.16648200
## attr(,"conf.level")
## [1] 0.95
```



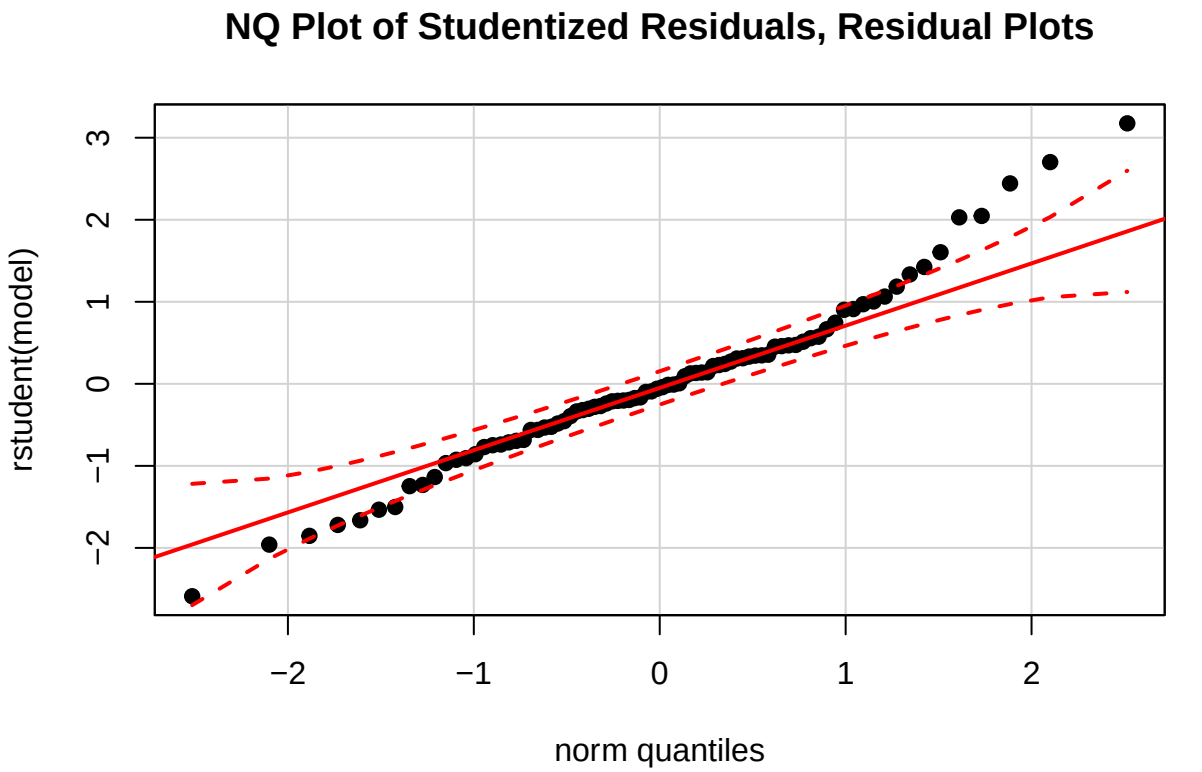
We bootstrapped the difference in mean GGGI for Full Democracies vs Authoritarian regimes. We took 10000 samples and plotted them. The histogram also displays the original t-test and bootstrapped 95% confidence intervals. The bootstrapped confidence interval (0.0953,0.1698) is slightly narrower than then t-test interval (0.0919,0.1738). This bootstrap analysis confirms that the difference in means is statistical significant since both confidence intervals do not include 0.

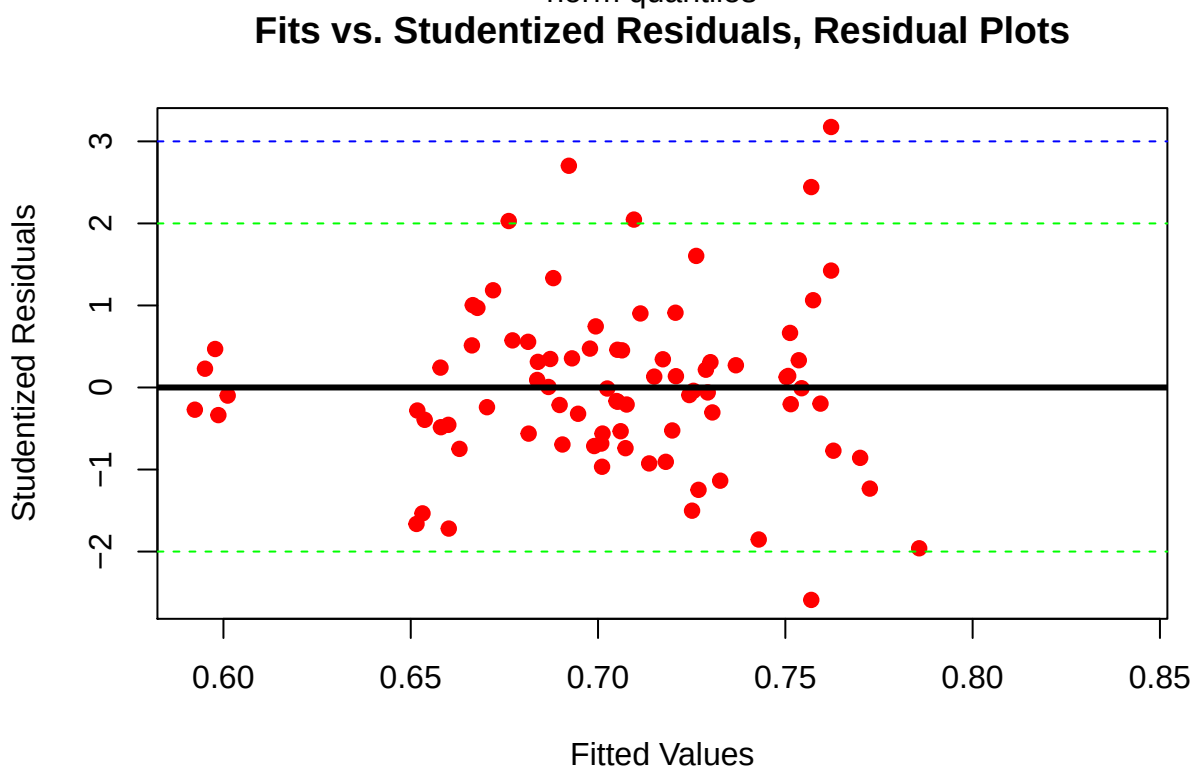
Regression

We made a new data frame to perform regression. We did not need every single variable, because some are rudeundant or unnecessary. We did not include Country, Democracy Score, GDP from Agro, or Female Agro Holding. Country is too specific to tell us anything important, since we want to predict GGGI score based on general economic and politcal information. Each country is directly given a GGGI score, so it would be a perfect (but useless) predictor. We will use democracy category as a categorical variable instead of democracy score. We are interested in how different style regimes affect GGGI, not the exact level of democracy. GDP from Agro is not going to be very important given we have total GDP and Total Agro Holding. Finally, we don't need to include the raw female agro holding numbers if we are using the percentage of female agro holding. This limits the number of potential variables in a reasonable way. We tried to use regression subsets function, but this may not make sense because it actually subsets based on each level for the categorical data.

```
##
## Call:
## lm(formula = `Global Gender Gap Index` ~ Region + `Democracy Category` +
##     `Percentage Female`)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -0.09292 -0.02020 -0.00148  0.01660  0.11179
##
## Coefficients:
##                                          Estimate Std. Error t value
## (Intercept)                             0.6467152  0.0177856  36.362
## RegionAsia                              0.0237824  0.0149082   1.595
## RegionEurope                            0.0375926  0.0146531   2.565
## RegionMiddle East                      -0.0553336  0.0226296  -2.445
## RegionNorth America                     0.0331697  0.0164101   2.021
## RegionSouth America                     0.0129836  0.0186940   0.695
## `Democracy Category`Authoritarian regimes  0.1403336  0.0435269   3.224
## `Democracy Category`Flawed democracy       0.0084158  0.0191072   0.440
## `Democracy Category`Full democracy         0.0589467  0.0218562   2.697
## `Democracy Category`Hybrid regime          0.0010081  0.0179814   0.056
## `Percentage Female`                        0.0012311  0.0004668   2.637
##                                            Pr(>|t|)
## (Intercept)                                 < 2e-16 ***
## RegionAsia                                  0.11492
## RegionEurope                                0.01233 *
## RegionMiddle East                           0.01686 *
## RegionNorth America                         0.04687 *
## RegionSouth America                         0.48952
## `Democracy Category`Authoritarian regimes   0.00188 **
## `Democracy Category`Flawed democracy        0.66089
## `Democracy Category`Full democracy          0.00866 **
## `Democracy Category`Hybrid regime           0.95544
## `Percentage Female`                         0.01018 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03865 on 74 degrees of freedom
## Multiple R-squared:  0.6048, Adjusted R-squared:  0.5514
## F-statistic: 11.32 on 10 and 74 DF,  p-value: 1.761e-11

## Anova Table (Type III tests)
##
## Response: Global Gender Gap Index
##                        Sum Sq Df   F value    Pr(>F)
## (Intercept)          1.97523  1 1322.1700 < 2.2e-16 ***
## Region               0.02465  5    3.3002 0.0095783 **
## `Democracy Category` 0.03430  4    5.7403 0.0004436 ***
## `Percentage Female`  0.01039  1    6.9555 0.0101790 *
## Residuals            0.11055 74
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```





Instead, we decided to just do backwards stepwise regression using ANOVA Type III sum of squares and a significance level of 0.05. We made sure to remove interaction terms before removing main effect terms when performing backwards step regression. In order we removed the following: Total GDP-Total Agro Holding, Region-Total GDP, Total GDP-CO2 Emissions, Total GDP, Percentage Female-Democracy Category, Total Agro Holding, and CO2 Emissions. This left us with a linear model with Region, Democracy Category, and

Percentage Female as the only predictors. This makes sense, as democracy category and region were found to be statistically significant with regards to mean GGGI. Additionally, percentage female and GGGI had a strong positive correlation in the corrplot. Looking at the summary information, Europe, Asia and North America predict the highest GGGI while the Middle East predicts a much lower GGGI. Additionally, Full Democracies have much higher GGGI scores than Authoritarian regimes. Finally, we see that as percentage of female agricultural landholdings increases so does the GGGI. More generally, these predictors make sense because countries that have higher democracy levels and have female holding a higher percentage of agricultural land would likely have higher levels of gender equality. Looking at the residual plots for this model, we see some slight heteroskedasticity and relatively normally distributed residuals (though there is some deviance from the normal at the upper tail). However, these do not appear to be very big issues and, again, it does not really make sense to transform a variable scored from 0 to 1.

Conclusion and Summary

The initial motivation for this project was to extend current political science research connecting gendered participation in industry to climate change and gender equality to democratization. We took data from the Food and Agriculture Organization of the United Nations database and from Wikipedia regarding macroeconomics, carbon dioxide emissions, agricultural landholdings, and democracy. We wanted to see how these factors predict gender equality (specifically the Global Gender Gap Index (GGGI) score calculated by the World Economic Forum). We found percentage of female agricultural landholdings and democracy score were strongly positively correlated, while total agricultural landholdings were negatively correlated. We saw that the mean GGGI scores was statistically different by region and democracy category. Finally, we did a backwards stepwise regression to find the best GGGI predictors. We found that region, democracy category, and percentage of female agricultural landholdings were the bests predictors. Countries in the Middle East and authoritarian regimes both had lower GGGI scores compared to countries that were fully democratic and from Europe, Asia, or North America. A higher percentage of agricultural land held by females predicted a higher GGGI score as well. Logically this make sense given the previous research explained in the introduction and the expectation that the larger role of females industry is suggestive of a greater gender equality.