Data Exploration and Analysis

COVID-19 Cases/Deaths by Nation

By: Mark Gjuraj

Published: 2020-04-29 (Refactored: 2025-03-20)

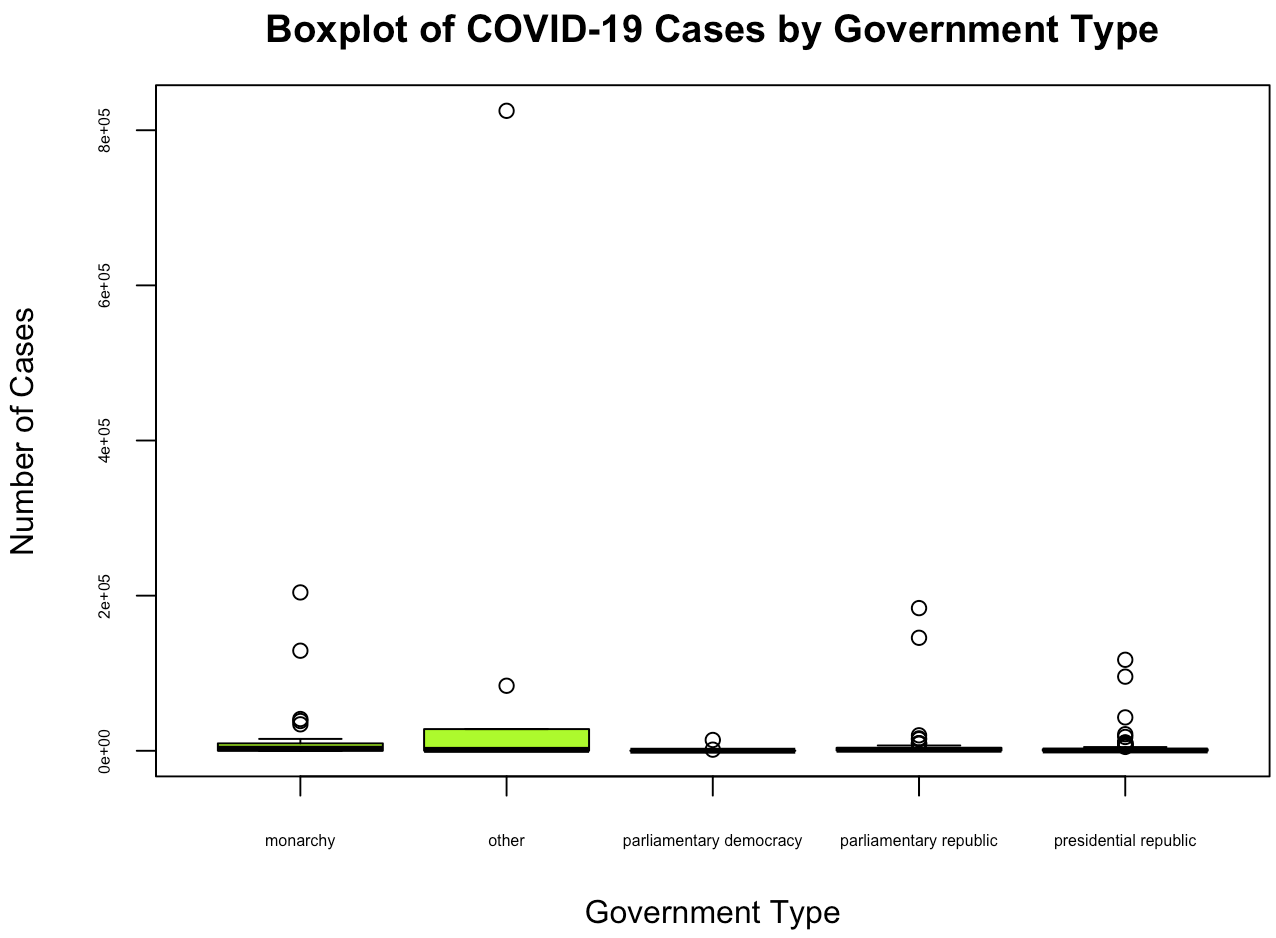
```
db <- read.csv("data/wb-covid-data.csv")
db <- db[, -1]
```

Exploration

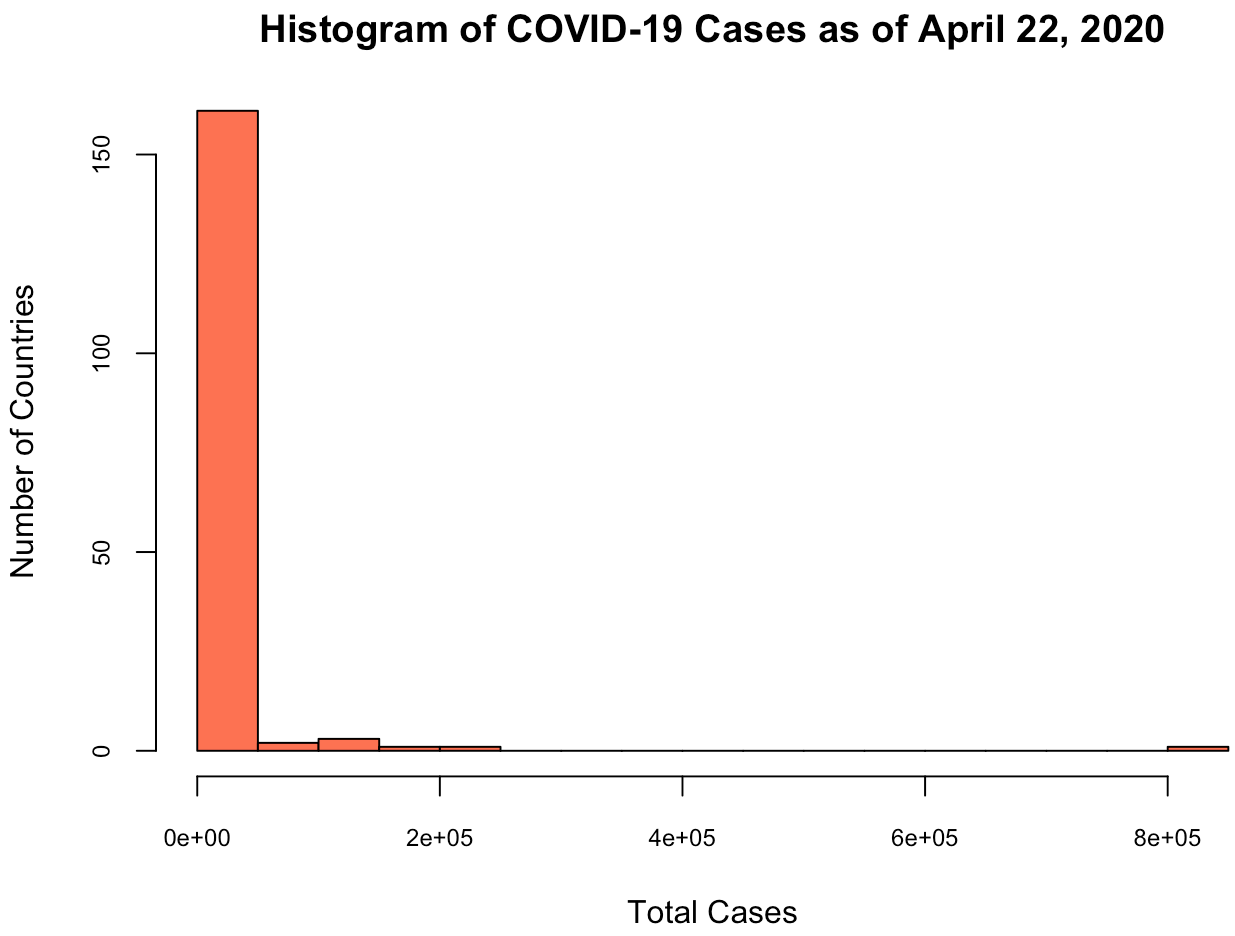
Descriptive Plots

Focus: `CovidCases`

```
par(mar = c(4, 5, 3, 1), cex.axis = 0.50)
boxplot(db$CovidCases ~ db$Govt,
        main = "Boxplot of COVID-19 Cases by Government Type",
        xlab = "Government Type",
        ylab = "Number of Cases",
        col = "greenyellow")
```

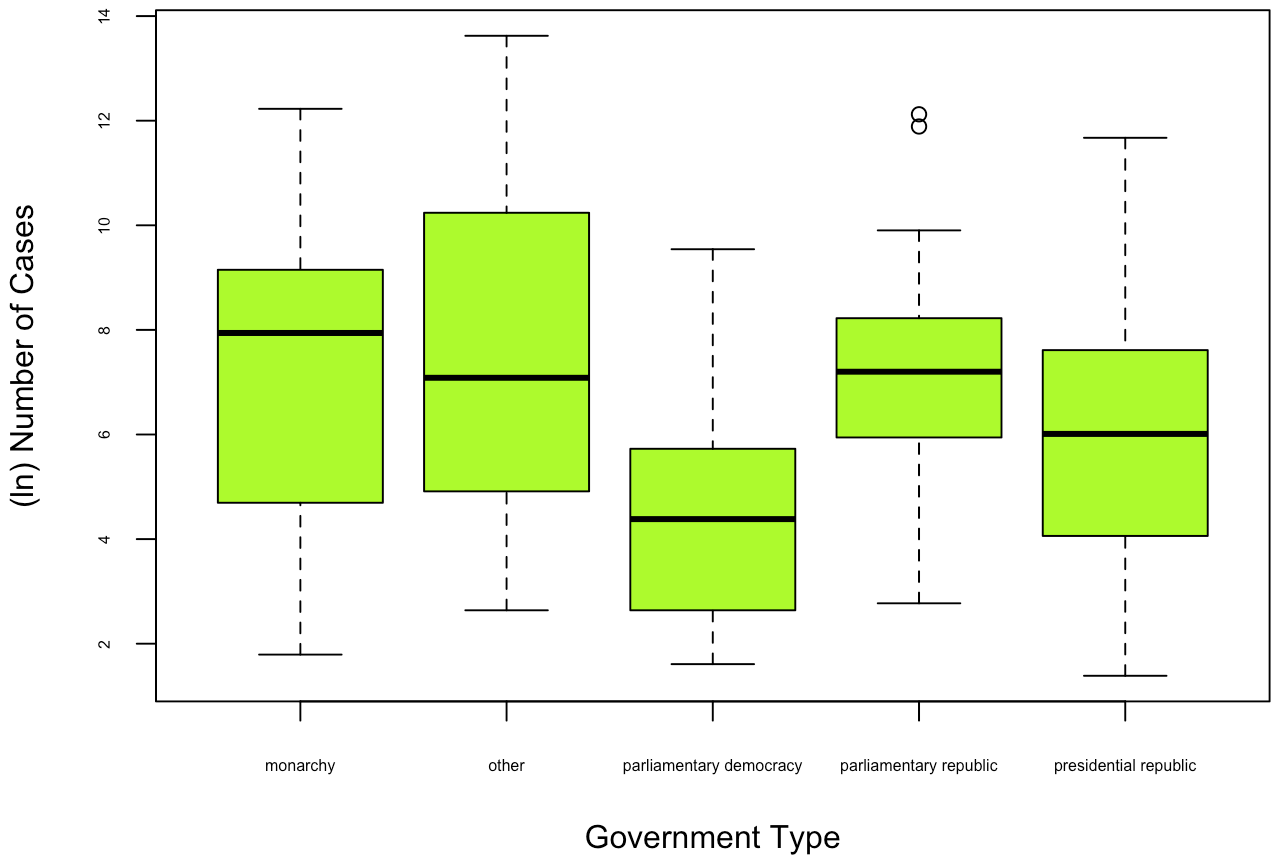


```
par(mar = c(4, 5, 3, 1), cex.axis = 0.75)
hist(db$CovidCases,
     main = "Histogram of COVID-19 Cases as of April 22, 2020",
     xlab = "Total Cases",
     ylab = "Number of Countries",
     col = "coral1",
     breaks = 20)
```



The boxplots indicate that the distribution of `CovidCases` for each factor of `Govt` seem to be heavily right-skewed, and the histogram of `CovidCases` is indicative of an approximately exponential distribution. This suggests a natural log transformation of `CovidCases` . A new feature `logCases` is created.

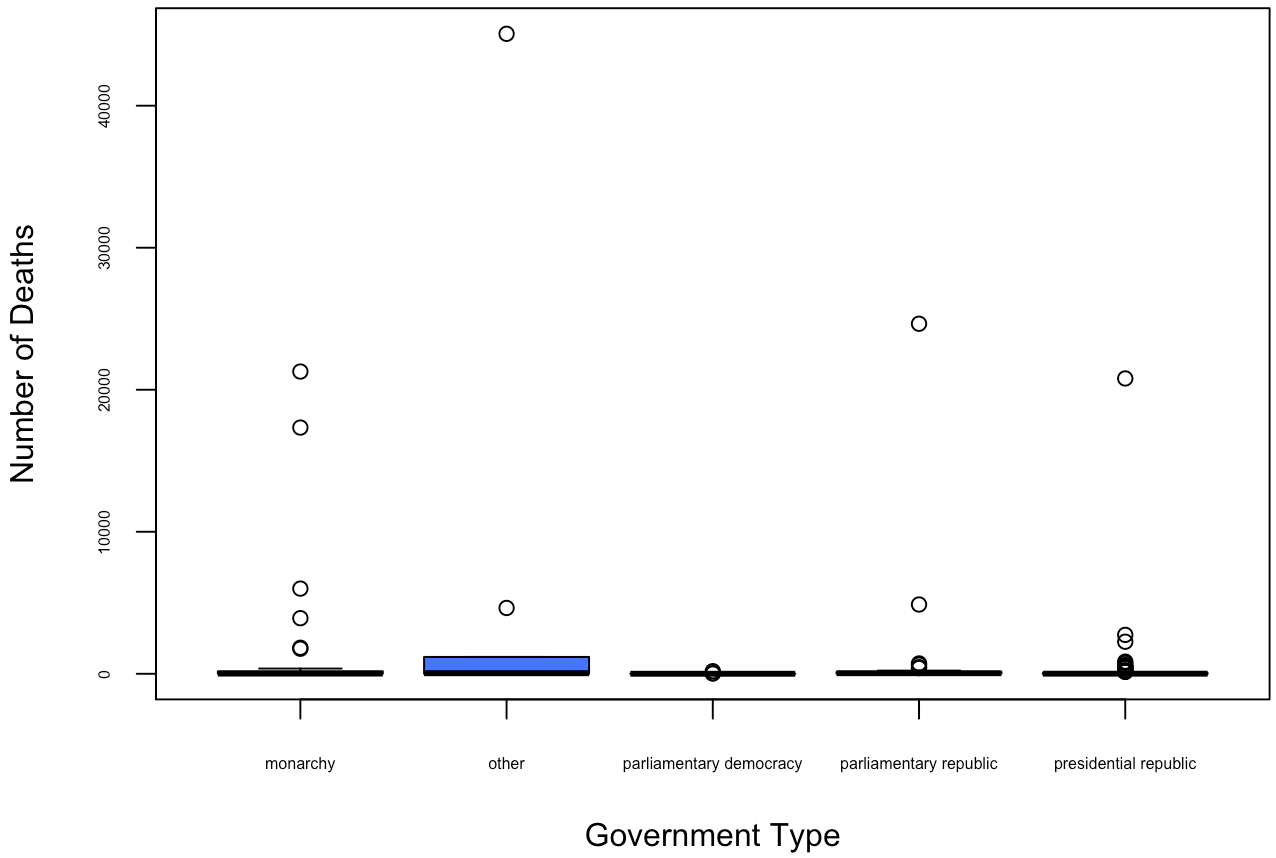
```
par(mar = c(4, 5, 3, 1), cex.axis = 0.50)
boxplot(log(db$CovidCases) ~ db$Govt,
        main = "Boxplot of (ln) COVID-19 Cases By Government Type",
        xlab = "Government Type",
        ylab = "(ln) Number of Cases",
        col = "greenyellow")
```



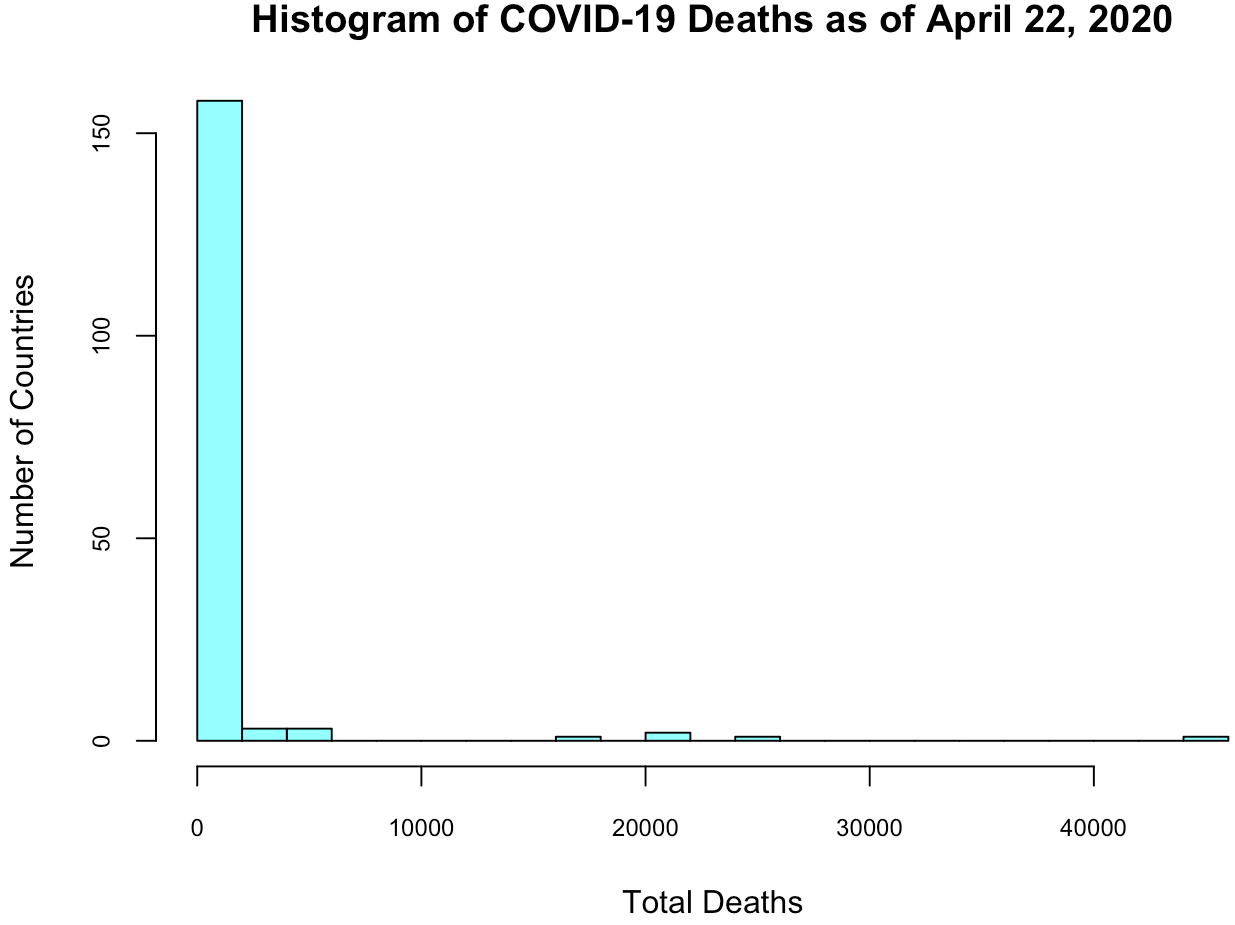

```
db$logCases <- log(db$CovidCases)
```

Focus: `CovidDeaths`

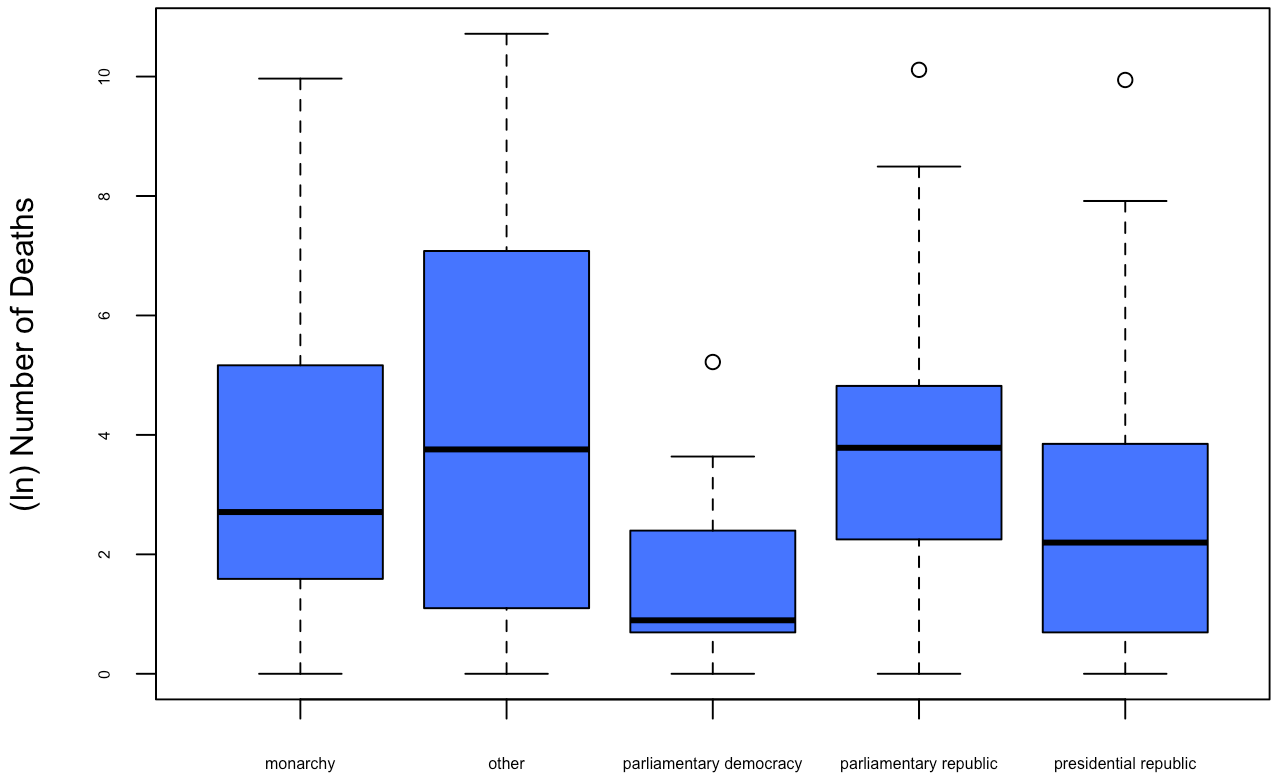
```
par(mar = c(4, 5, 3, 1), cex.axis = 0.50)
boxplot(db$CovidDeaths ~ db$Govt,
        main = "Boxplot of COVID-19 Deaths By Government Type",
        xlab = "Government Type",
        ylab = "Number of Deaths",
        col = "royalblue1")
```




```
par(mar = c(4, 5, 3, 1), cex.axis = 0.75)
hist(db$CovidDeaths,
     main = "Histogram of COVID-19 Deaths as of April 22, 2020",
     xlab = "Total Deaths",
     ylab = "Number of Countries",
     col = "darkslategray1",
     breaks = 20)
```




```
db$logDeaths <- log(db$CovidDeaths + 1)
par(mar = c(4, 5, 3, 1), cex.axis = 0.50)
boxplot(db$logDeaths ~ db$Govt,
        main = "Boxplot of (ln) COVID-19 Deaths By Government Type",
        xlab = "Government Type",
        ylab = "(ln) Number of Deaths",
        col = "royalblue1")
```



A similar distribution for `CovidDeaths` is evident. Note: applying a natural log transformation to `CovidDeaths` results in a handful of "-infinity" values, due to the fact that some nations have a death count of 0. To remedy this, `CovidDeaths` is incremented by 1 before creating `logDeaths` .

Correlation, Linearity, Multicollinearity

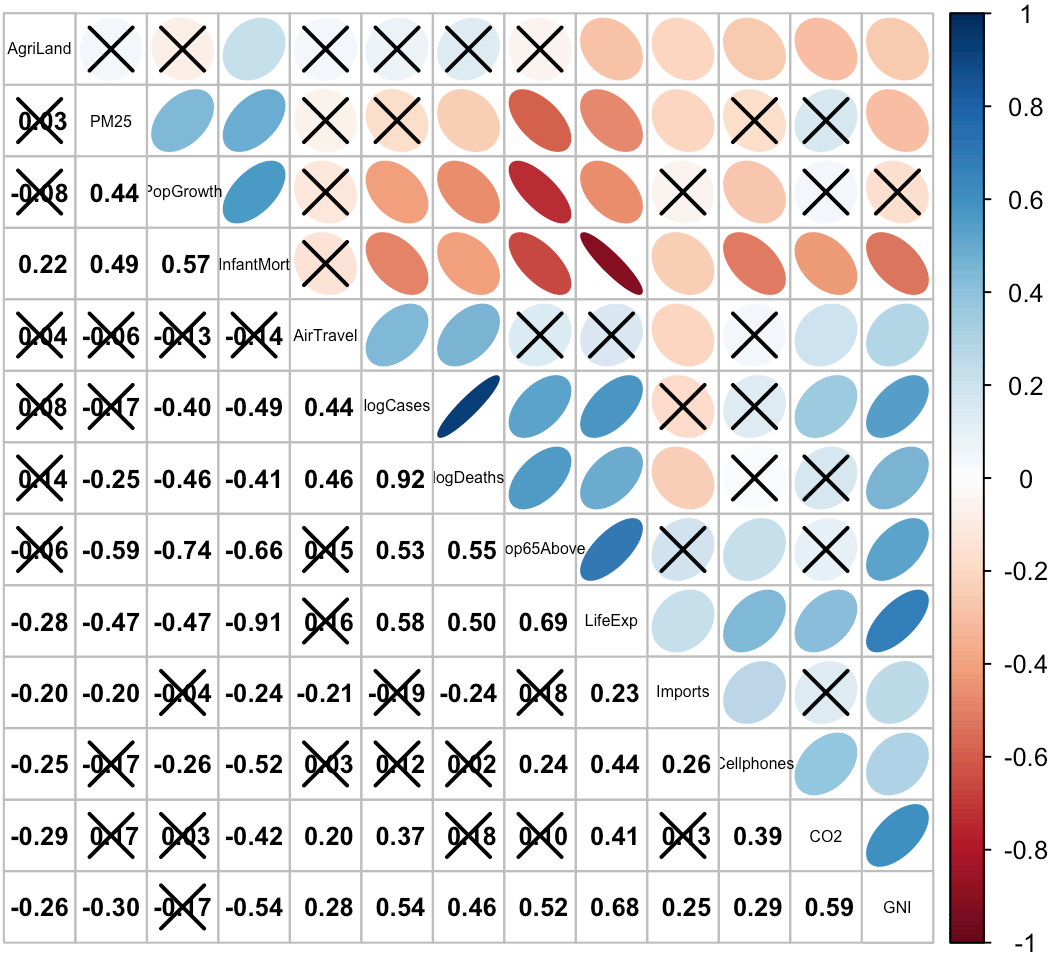
A smaller dataframe `db_0` is created with some of the continuous variables at our disposal. These will be the continuous variables utilized in the multiple regression model conducted later in this report, where we seek to predict `logCases` .

```
db_0 <- db[, c("Country", "logCases", "AgriLand", "CO2", "Imports", "GNI", "PopGrowth", "PM25",
"Pop65Above", "Cellphones", "InfantMort", "AirTravel", "LifeExp", "logDeaths")]
```

```
db_0 <- db_0[complete.cases(db_0), ]
cor1 <- round(cor(db_0[, -1]), 3)
library(corrplot)
```

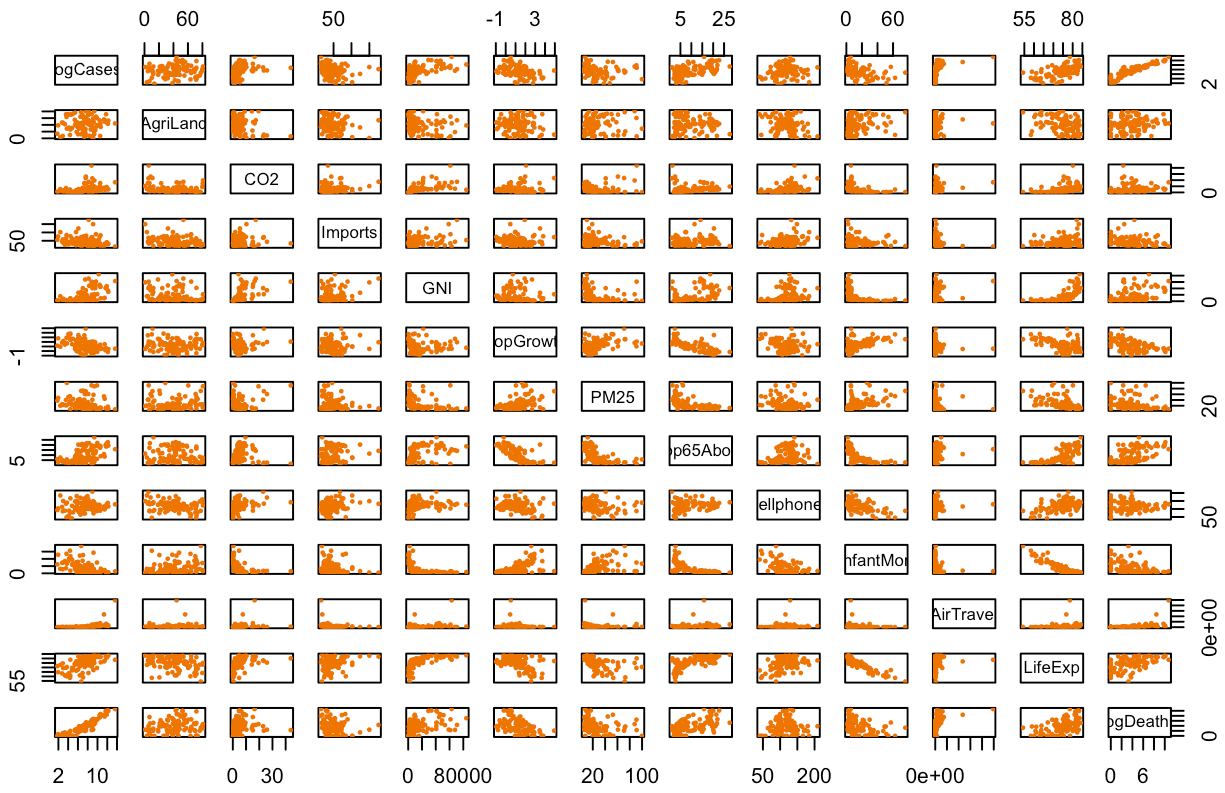
```
## corrplot 0.95 loaded
```

```
sigcor1 <- cor.mtest(db_0[, -1], conf.level = 0.95)
corrplot.mixed(cor1,
               lower.col = "black",
               upper = "ellipse",
               tl.col = "black",
               number.cex = 0.8,
               order = "hclust",
               tl.pos = "d",
               tl.cex = 0.5,
               p.mat = sigcor1$p,
               sig.level = 0.05)
```



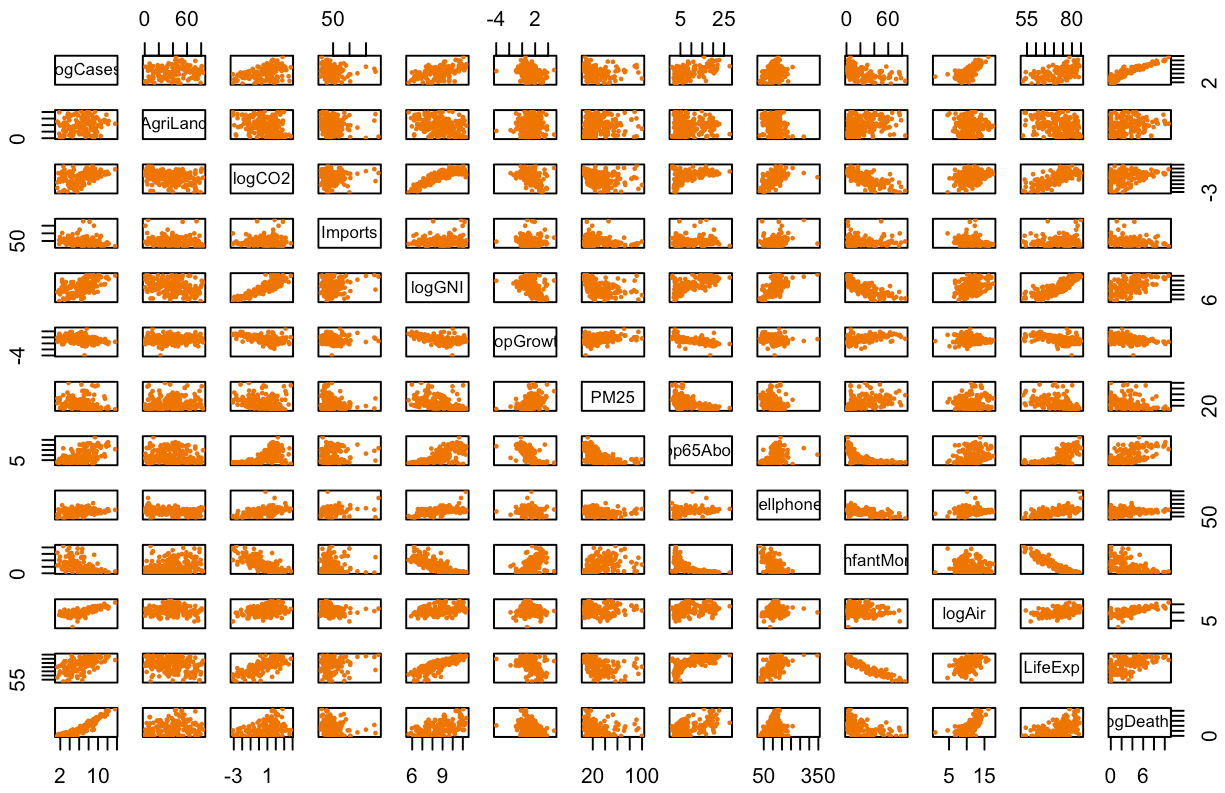
The correlation plot illustrates the strength of each pairwise correlation — including the p-values resulting from parametric tests of each correlation's significance — and an "X" is drawn through the plots whose variables do not have a statistically significant non-zero correlation.

```
plot(db_0[, -1],
     main = "Matrix Plot of Subset of WB Data",
     pch = 16,
     cex = 0.5,
     col = "darkorange2")
```



The top row of the matrix plot indicates that there are a few predictor variables that should be transformed, as some of the plots demonstrate non-linear relationships.

```
db$logCO2 <- log(db$CO2)
db$logGNI <- log(db$GNI)
db$logAir <- log(db$AirTravel)
db_0 <- db[, c("Country", "logCases", "AgriLand", "logCO2", "Imports", "logGNI", "PopGrowth", "PM
25", "Pop65Above", "Cellphones", "InfantMort", "logAir", "LifeExp", "logDeaths")]
plot(db_0[, -1], main = "Matrix Plot of Subset of WB Data", pch = 16, cex = 0.5,
     col = "darkorange2")
```



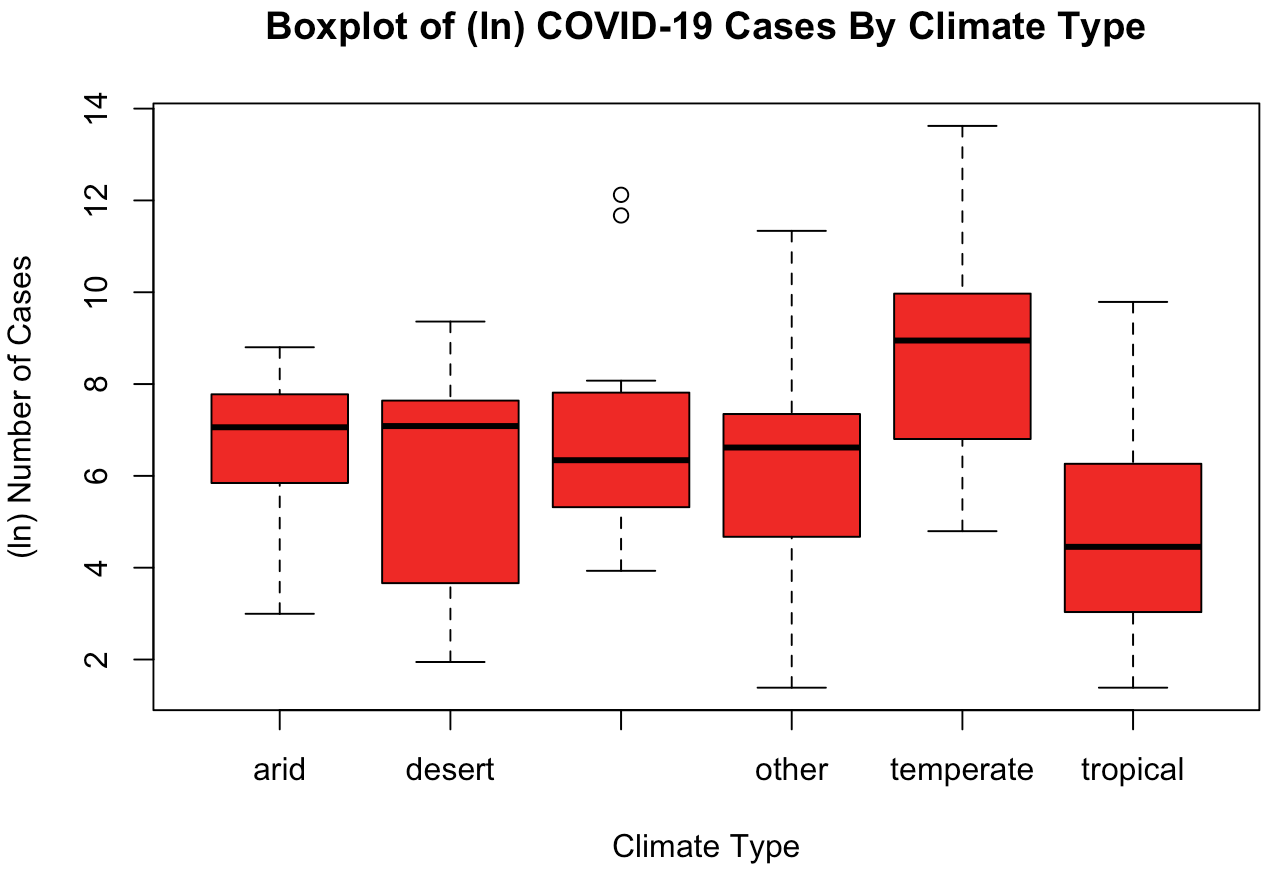
The uppermost row of the matrix plot is now satisfactory; all of the scatterplots appear to resemble either lines or random noise.

Notably, there are a few instances of multicollinearity observable throughout the rest of the plots. While this does violate the assumption that the predictors in a multiple linear regression model are uncorrelated with one another, we accept the reality that some of our variables will compete in explaining the variability in `logCases`, and progress onward.

Analysis

Two Sample T-test

```
boxplot(db$logCases ~ db$Climate,
        main = "Boxplot of (ln) COVID-19 Cases By Climate Type",
        xlab = "Climate Type",
        ylab = "(ln) Number of Cases",
        col = "firebrick2")
```



The boxplots above indicate that there could be a statistically significant difference in mean `logCases` for the `Climate` factors “temperate” and “tropical”.

```
db_temp <- db[, c("Climate","CovidCases","logCases")]
db_temp <- na.omit(db_temp[db_temp$Climate == c('temperate','tropical'),])
```

```
## Warning in db_temp$Climate == c("temperate", "tropical"): longer object length
## is not a multiple of shorter object length
```

```
t.test(CovidCases ~ Climate, data = db_temp)
```

```
##
##  Welch Two Sample t-test
##
## data:  CovidCases by Climate
## t = 1.707, df = 22.007, p-value = 0.1019
## alternative hypothesis: true difference in means between group temperate and group tropical is
not equal to 0
## 95 percent confidence interval:
##  -13299.41 137088.01
## sample estimates:
## mean in group temperate  mean in group tropical
##                63242.61                 1348.31
```

```
(logtt <- t.test(logCases ~ Climate, data = db_temp))
```

```
##
##  Welch Two Sample t-test
##
## data:  logCases by Climate
## t = 6.0883, df = 47.849, p-value = 1.858e-07
## alternative hypothesis: true difference in means between group temperate and group tropical is
not equal to 0
## 95 percent confidence interval:
##  2.552416 5.069807
## sample estimates:
## mean in group temperate  mean in group tropical
##                8.893943                5.082831
```

The two-sample t-test shows that there is no evidence of a significant difference in mean `CovidCases` between the `Climate` types.

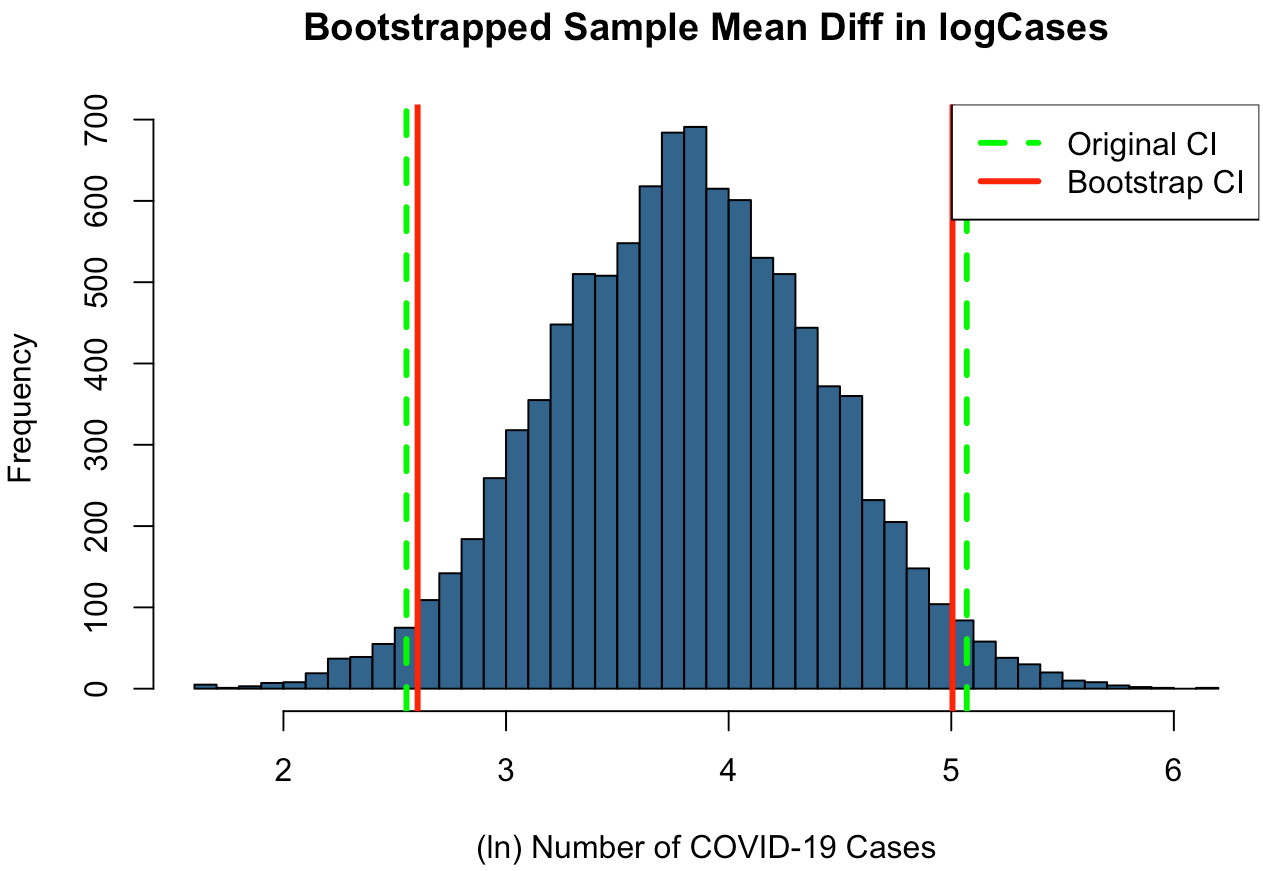
However, when we consider `logCases`, the t-test results in a p-value of approximately 0.0000001, which is less than any reasonable threshold. Thus, we can reject the null hypothesis and accept the alternative hypothesis — the mean `logCases` for countries with tropical and temperate climates are statistically significantly different.

Bootstrapped Confidence Interval

By resampling the difference in the sample means, we can obtain a bootstrapped confidence interval for the true difference in mean `logCases` between nations with “temperate” and “tropical” climates, and subsequently compare this with the theoretical confidence interval from the two-sample t-test.

```
N <- 10000
diffcc <- rep(NA, N)
for (i in 1:N) {
  s_temperate <- sample(db_temp$logCases[db_temp$Climate == "temperate"],
                        sum(db_temp$Climate == "temperate"),
                        replace = TRUE)
  s_tropical <- sample(db_temp$logCases[db_temp$Climate == "tropical"],
                       sum(db_temp$Climate == "tropical"),
                       replace = TRUE)
  diffcc[i] <- mean(s_temperate) - mean(s_tropical)
}
bootci <- quantile(diffcc, c(0.025, 0.975))
ttest_CI <- logtt$conf.int
```

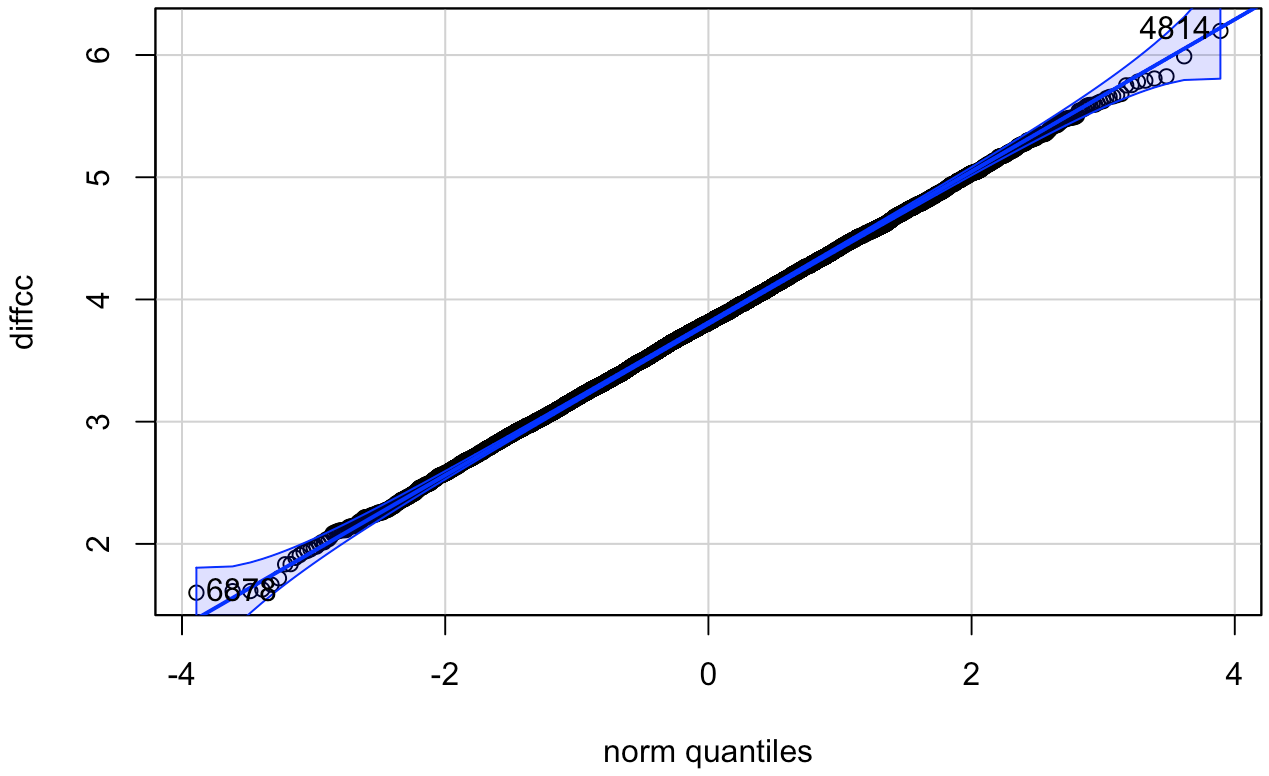
```
hist(diffcc,
     col = "steelblue4",
     main = "Bootstrapped Sample Mean Diff in logCases",
     xlab = "(ln) Number of COVID-19 Cases",
     breaks = 50)
abline(v = bootci,
       lwd = 3,
       col = "red",
       lty = 1)
abline(v = ttest_CI, lwd = 3, col = "green", lty = 2)
legend("topright",
       c("Original CI","Bootstrap CI"),
       lwd = 3,
       col = c("green", "red"),
       lty = c(2,1))
```




```
library(car)
```

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

```
qqPlot(diffcc)
```




```
## [1] 4814 6878
```

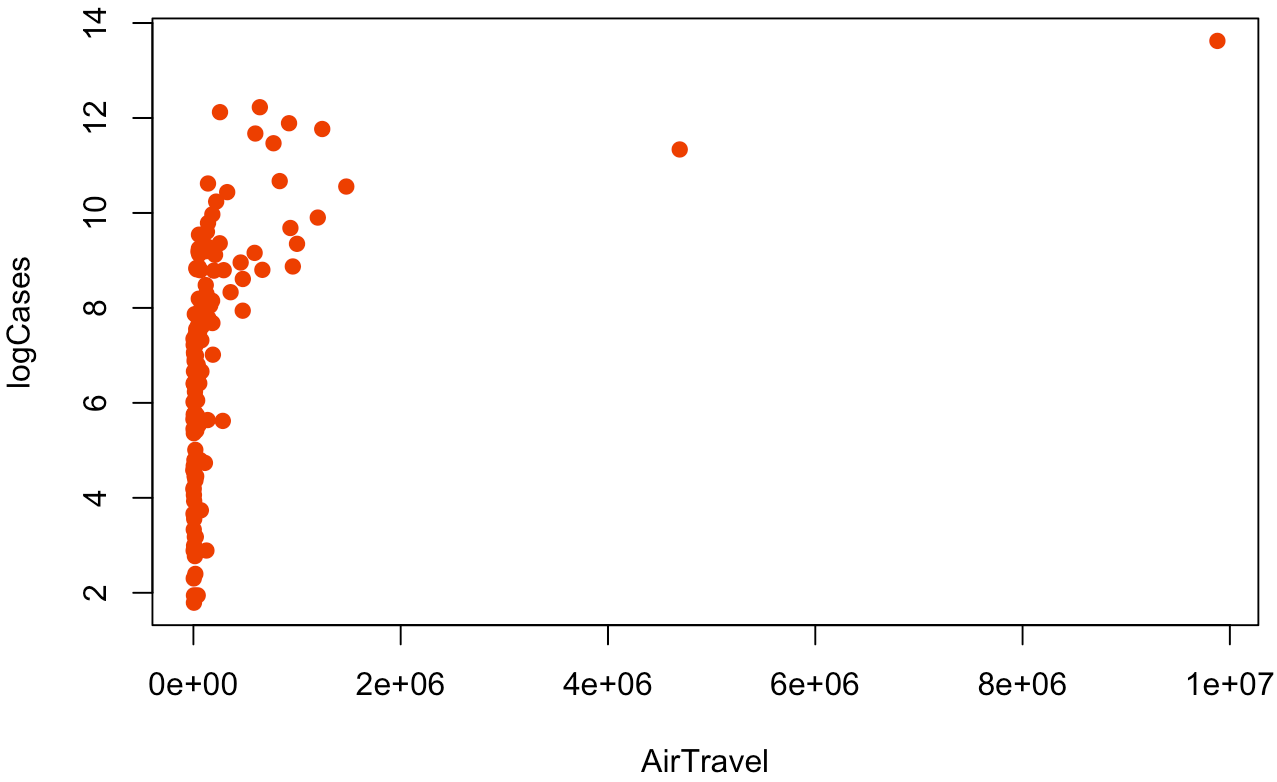
Evidently, as we'd expect (thanks to the Central Limit Theorem) the bootstrapped sample mean differences in `logCases` between nations with "temperate" and "tropical" climates are approximately normally distributed.

The bootstrapped and original (derived from the parametric test) confidence intervals are superimposed onto the histogram, and the bootstrapped confidence interval appears to be slightly more narrow.

Permutation Test: Correlation

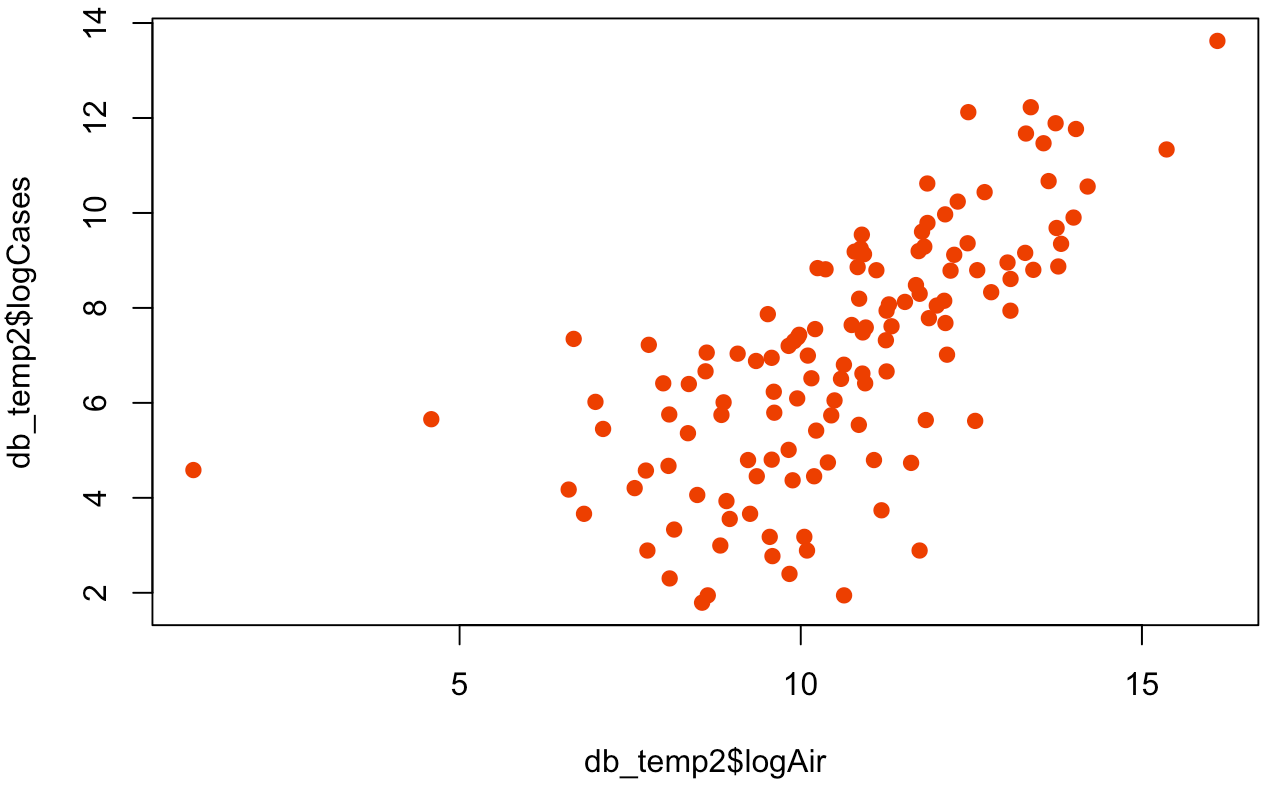
With a Permutation Test, we can determine whether or not the true correlation between two variables is statistically significantly different from 0 (no linear relationship). The variables we chose to assess are `logCases` and `logAir`.

```
db_temp2 <- na.omit(db[,c('AirTravel',"logCases", "logAir")])
plot(logCases ~ AirTravel,
     data = db_temp2,
     pch = 19,
     col = "orangered2")
```



Recall our natural log transformation of the variable `AirTravel` ; computing the correlation of this set of data would not be proper, as these variables do not have a linear relationship.

```
plot(db_temp2$logCases ~ db_temp2$logAir,
     pch = 19,
     col = "orangered2")
```




```
(obs_cor <- cor(db_temp2$logAir, db_temp2$logCases))
```

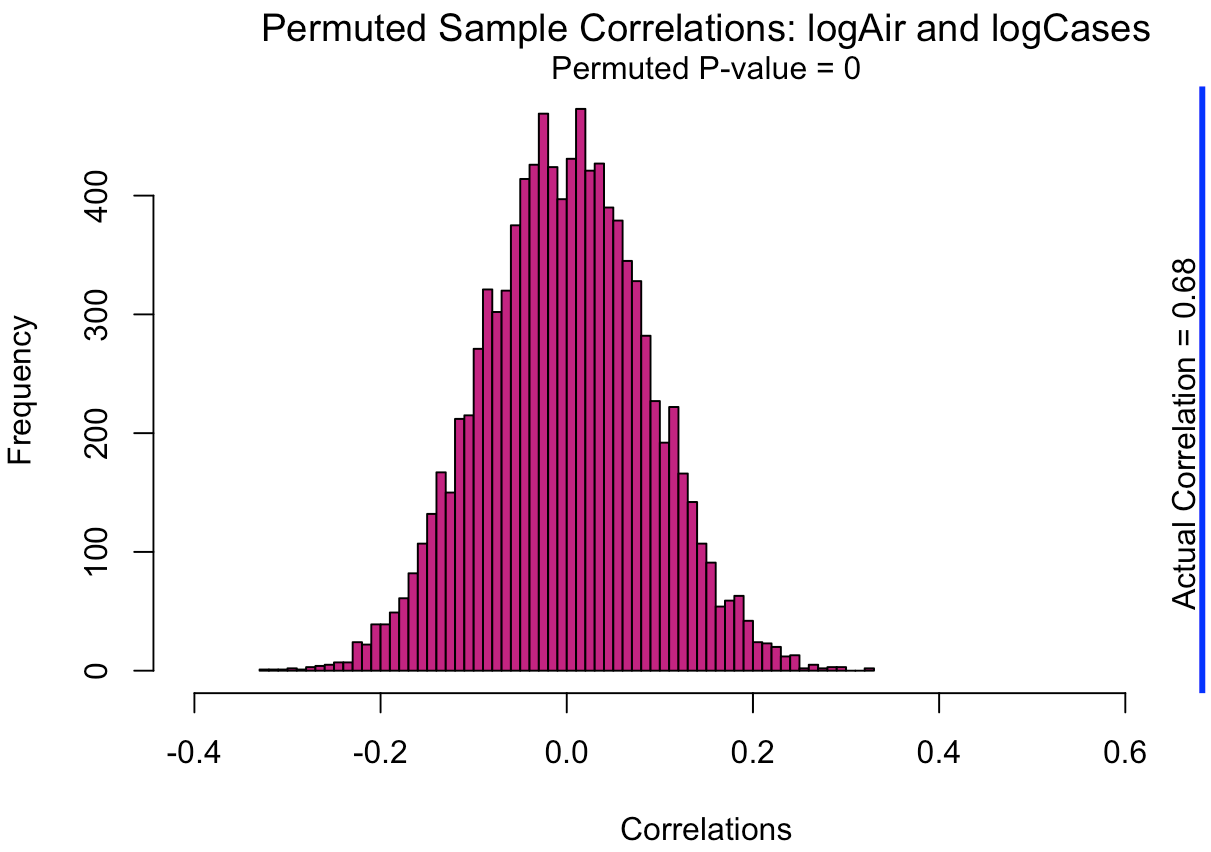
```
## [1] 0.6826924
```

The actual, observed correlation is approximately 0.683, which is a moderately positive correlation. If we assume that the null hypothesis — the true correlation between `logAir` and `logCases` equals 0 — is true, then we can randomize (permute) the values of `logAir` that are associated with each observation of `logCases` (without replacement).

```
n_samp <- 10000
corResults <- rep(NA, n_samp)
for(i in 1:n_samp){
  corResults[i] <- cor(db_temp2$logAir, sample(db_temp2$logCases))
}
# P-value (two-sided) for correlation
(truecor <- mean(abs(corResults) >= abs(obs_cor)))
```

```
## [1] 0
```

The distribution of these 10000 permuted correlations can be plotted using a histogram.



The “permuted p-value” of the correlation between `logAir` and `logCases` is 0. Thus, the probability of observing the actual correlation (or something more extreme) in a distribution under the null hypothesis is approximately 0. This is less than any reasonable threshold, so we reject this null hypothesis, concluding that there is a statistically significant correlation between a nation’s (natural log) Air Travel activity and (natural log) COVID-19 case count.

Multiple Regression

The dataframe `WB_1` is created with the continuous variables we examined in the descriptive plots above. In addition, we include the categorical variables `Govt` and `Climate`.

```
db_1 <- db[, c("logCases", "AgriLand", "logCO2", "Imports", "logGNI", "PopGrowth", "PM25", "Pop65
Above", "Cellphones", "InfantMort", "logAir", "LifeExp", "logDeaths", "Govt", "Climate")]
db_1 <- db_1[complete.cases(db_1), ]
```

We seek to predict `logCases`, and we arrive at our final model using Backwards Stepwise Regression. With the `Anova()` function, the p-values quantifying the overall significance of each of the variables are given. We discard the most insignificant predictors (largest p-values) until we arrive at a model whose predictors are all statistically significant at the 0.05 threshold.

```
m1 <- lm(logCases ~.,
         data = db_1)
Anova(m1, type = 3)
```

```
## Anova Table (Type III tests)
##
## Response: logCases
##              Sum Sq Df  F value  Pr(>F)
## (Intercept)   0.540  1   0.7628 0.38503
## AgriLand      0.903  1   1.2749 0.26218
## logCO2        2.235  1   3.1553 0.07944 .
## Imports       0.119  1   0.1678 0.68313
## logGNI        0.185  1   0.2617 0.61036
## PopGrowth     0.795  1   1.1221 0.29261
## PM25          4.622  1   6.5255 0.01251 *
## Pop65Above    0.439  1   0.6202 0.43327
## Cellphones    0.000  1   0.0000 0.99801
## InfantMort    0.005  1   0.0068 0.93460
## logAir        1.517  1   2.1411 0.14727
## LifeExp       2.303  1   3.2509 0.07510 .
## logDeaths   100.393  1 141.7321 < 2e-16 ***
## Govt          2.578  4   0.9099 0.46231
## Climate       1.835  5   0.5181 0.76185
## Residuals    57.374 81
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Remove Cellphones
m2 <- lm(logCases ~ AgriLand + logCO2 + Imports + logGNI + PopGrowth + PM25 + Pop65Above + Infant
Mort + logAir + LifeExp + logDeaths + Govt + Climate,
         data = db_1)
Anova(m2, type = 3)
```

```
## Anova Table (Type III tests)
##
## Response: logCases
##              Sum Sq Df  F value  Pr(>F)
## (Intercept)   0.542  1   0.7744 0.38144
## AgriLand      0.904  1   1.2921 0.25898
## logCO2        2.236  1   3.1963 0.07750 .
## Imports       0.119  1   0.1706 0.68066
## logGNI        0.194  1   0.2767 0.60028
## PopGrowth     0.822  1   1.1754 0.28147
## PM25          4.678  1   6.6853 0.01149 *
## Pop65Above    0.439  1   0.6279 0.43043
## InfantMort    0.005  1   0.0070 0.93370
## logAir        1.520  1   2.1721 0.14436
## LifeExp       2.303  1   3.2912 0.07331 .
## logDeaths   105.796  1 151.2052 < 2e-16 ***
## Govt          2.712  4   0.9689 0.42914
## Climate       1.858  5   0.5312 0.75205
## Residuals    57.374 82
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Remove InfantMort
m3 <- lm(logCases ~ AgriLand + logCO2 + Imports + logGNI + PopGrowth + PM25 + Pop65Above + logAir
+ LifeExp + logDeaths + Govt + Climate,
         data = db_1)
Anova(m3, type = 3)
```

```
## Anova Table (Type III tests)
##
## Response: logCases
##              Sum Sq Df  F value    Pr(>F)
## (Intercept)   1.153  1   1.6674  0.200190
## AgriLand      0.902  1   1.3050  0.256590
## logCO2        2.478  1   3.5849  0.061793 .
## Imports       0.119  1   0.1722  0.679260
## logGNI        0.189  1   0.2732  0.602608
## PopGrowth     1.031  1   1.4908  0.225546
## PM25          4.711  1   6.8148  0.010724 *
## Pop65Above    0.449  1   0.6502  0.422349
## logAir        1.535  1   2.2197  0.140049
## LifeExp       5.673  1   8.2062  0.005286 **
## logDeaths   111.039  1 160.6195 < 2.2e-16 ***
## Govt          2.711  4   0.9804  0.422813
## Climate       1.861  5   0.5383  0.746708
## Residuals    57.379 83
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Remove Climate
m4 <- lm(logCases ~ AgriLand + logCO2 + Imports + logGNI + PopGrowth + PM25 + Pop65Above + logAir
+ LifeExp + logDeaths + Govt,
         data = db_1)
Anova(m4, type = 3)
```

```
## Anova Table (Type III tests)
##
## Response: logCases
##              Sum Sq Df  F value    Pr(>F)
## (Intercept)   0.872  1   1.2949  0.258233
## AgriLand      0.708  1   1.0518  0.307907
## logCO2        3.356  1   4.9852  0.028102 *
## Imports       0.166  1   0.2471  0.620374
## logGNI        0.149  1   0.2214  0.639133
## PopGrowth     0.901  1   1.3391  0.250321
## PM25          4.386  1   6.5151  0.012420 *
## Pop65Above    0.302  1   0.4490  0.504562
## logAir        2.125  1   3.1567  0.079072 .
## LifeExp       4.701  1   6.9833  0.009738 **
## logDeaths   111.325  1 165.3721 < 2.2e-16 ***
## Govt          2.437  4   0.9049  0.464764
## Residuals    59.240 88
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Remove logGNI
m5 <- lm(logCases ~ AgriLand + logCO2 + Imports + PopGrowth + PM25 + Pop65Above + logAir + LifeEx
p + logDeaths + Govt,
         data = db_1)
Anova(m5, type = 3)
```

```
## Anova Table (Type III tests)
##
## Response: logCases
##              Sum Sq Df  F value    Pr(>F)
## (Intercept)   1.302  1   1.9516  0.165893
## AgriLand      0.902  1   1.3512  0.248185
## logCO2        4.990  1   7.4783  0.007535 **
## Imports       0.170  1   0.2543  0.615299
## PopGrowth     0.755  1   1.1317  0.290286
## PM25          4.665  1   6.9903  0.009685 **
## Pop65Above    0.195  1   0.2925  0.589960
## logAir        2.065  1   3.0946  0.081988 .
## LifeExp       4.758  1   7.1299  0.009011 **
## logDeaths   112.591  1 168.7286 < 2.2e-16 ***
## Govt          2.358  4   0.8833  0.477329
## Residuals    59.389 89
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Remove Imports
m6 <- lm(logCases ~ AgriLand + logCO2 + PopGrowth + PM25 + Pop65Above + logAir + LifeExp + logDea
ths + Govt,
         data = db_1)
Anova(m6, type = 3)
```

```
## Anova Table (Type III tests)
##
## Response: logCases
##              Sum Sq Df  F value    Pr(>F)
## (Intercept)   1.365  1   2.0624  0.154438
## AgriLand      0.885  1   1.3368  0.250660
## logCO2        4.823  1   7.2875  0.008294 **
## PopGrowth     0.680  1   1.0273  0.313506
## PM25          4.944  1   7.4716  0.007546 **
## Pop65Above    0.151  1   0.2283  0.633976
## logAir        2.228  1   3.3667  0.069831 .
## LifeExp       4.628  1   6.9937  0.009651 **
## logDeaths   127.002  1 191.9150 < 2.2e-16 ***
## Govt          2.461  4   0.9297  0.450446
## Residuals    59.559 90
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Remove Pop65ABove
m7 <- lm(logCases ~ AgriLand + logCO2 + PopGrowth + PM25 + logAir + LifeExp + logDeaths + Govt,
         data = db_1)
Anova(m7, type = 3)
```

```
## Anova Table (Type III tests)
##
## Response: logCases
##              Sum Sq Df  F value    Pr(>F)
## (Intercept)   1.436  1   2.1882  0.142527
## AgriLand      0.864  1   1.3173  0.254092
## logCO2        4.862  1   7.4105  0.007771 **
## PopGrowth     0.529  1   0.8060  0.371662
## PM25          5.027  1   7.6610  0.006836 **
## logAir        2.128  1   3.2426  0.075061 .
## LifeExp       5.677  1   8.6521  0.004143 **
## logDeaths   142.016  1 216.4374 < 2.2e-16 ***
## Govt          2.329  4   0.8874  0.474819
## Residuals    59.710 91
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Remove Govt
m8 <- lm(logCases ~ AgriLand + logCO2 + PopGrowth + PM25 + logAir + LifeExp + logDeaths,
         data = db_1)
Anova(m8, type = 3)
```

```
## Anova Table (Type III tests)
##
## Response: logCases
##              Sum Sq Df  F value    Pr(>F)
## (Intercept)   0.725  1   1.1099  0.294785
## AgriLand      0.882  1   1.3510  0.248007
## logCO2        4.004  1   6.1316  0.015048 *
## PopGrowth     0.665  1   1.0188  0.315363
## PM25          4.182  1   6.4036  0.013033 *
## logAir        2.430  1   3.7206  0.056731 .
## LifeExp       4.819  1   7.3800  0.007837 **
## logDeaths   149.002  1 228.1662 < 2.2e-16 ***
## Residuals    62.039 95
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Remove PopGrowth
m9 <- lm(logCases ~ AgriLand + logCO2 + PM25 + logAir + LifeExp + logDeaths,
         data = db_1)
Anova(m9, type = 3)
```

```
## Anova Table (Type III tests)
##
## Response: logCases
##              Sum Sq Df  F value    Pr(>F)
## (Intercept)   0.479  1   0.7335  0.393879
## AgriLand      0.677  1   1.0364  0.311209
## logCO2        3.777  1   5.7825  0.018104 *
## PM25          5.327  1   8.1558  0.005261 **
## logAir        2.944  1   4.5075  0.036318 *
## LifeExp       4.330  1   6.6291  0.011560 *
## logDeaths   156.110  1 239.0039 < 2.2e-16 ***
## Residuals    62.704 96
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Remove AgriLand
```

```
m10 <- lm(logCases ~ logCO2 + PM25 + logAir + LifeExp + logDeaths,
          data = db_1)
summary(m10)
```

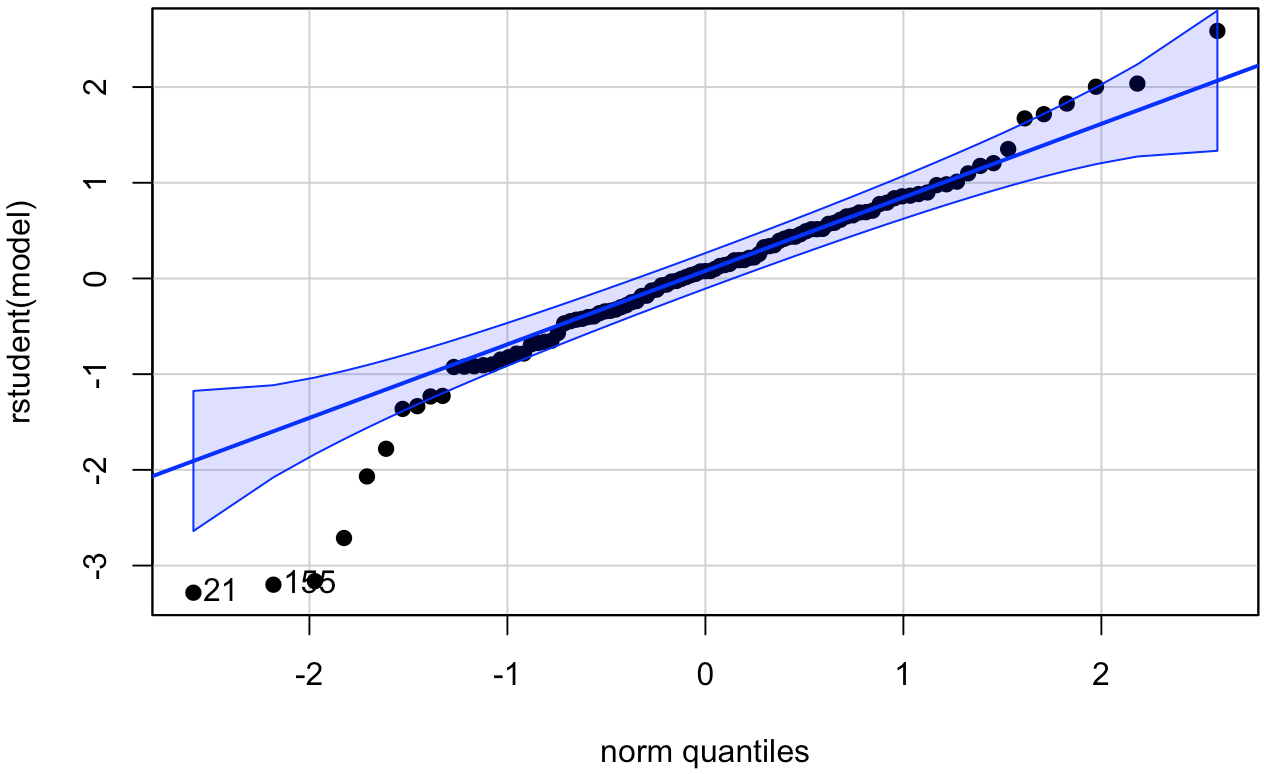
```
##
## Call:
## lm(formula = logCases ~ logCO2 + PM25 + logAir + LifeExp + logDeaths,
##     data = db_1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.45767 -0.34969  0.06048  0.47204  1.97521
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.784801   1.432069  -0.548   0.5849
## logCO2       0.196276   0.087293   2.248   0.0268 *
## PM25         0.012842   0.004634   2.771   0.0067 **
## logAir       0.131803   0.060399   2.182   0.0315 *
## LifeExp      0.046186   0.019284   2.395   0.0185 *
## logDeaths    0.712663   0.043763  16.285   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8083 on 97 degrees of freedom
## Multiple R-squared:  0.898,  Adjusted R-squared:  0.8927
## F-statistic: 170.7 on 5 and 97 DF,  p-value: < 2.2e-16
```

After removing `Cellphones`, `InfantMort`, `Climate`, `logGNI`, `Imports`, `Pop65Above`, `Govt`, `PopGrowth`, and `AgriLand`, we arrive at our final model `m10`, which predicts `logCases` with the continuous variables `logCO2`, `PM25`, `logAir`, `LifeExp`, and `logDeaths`.

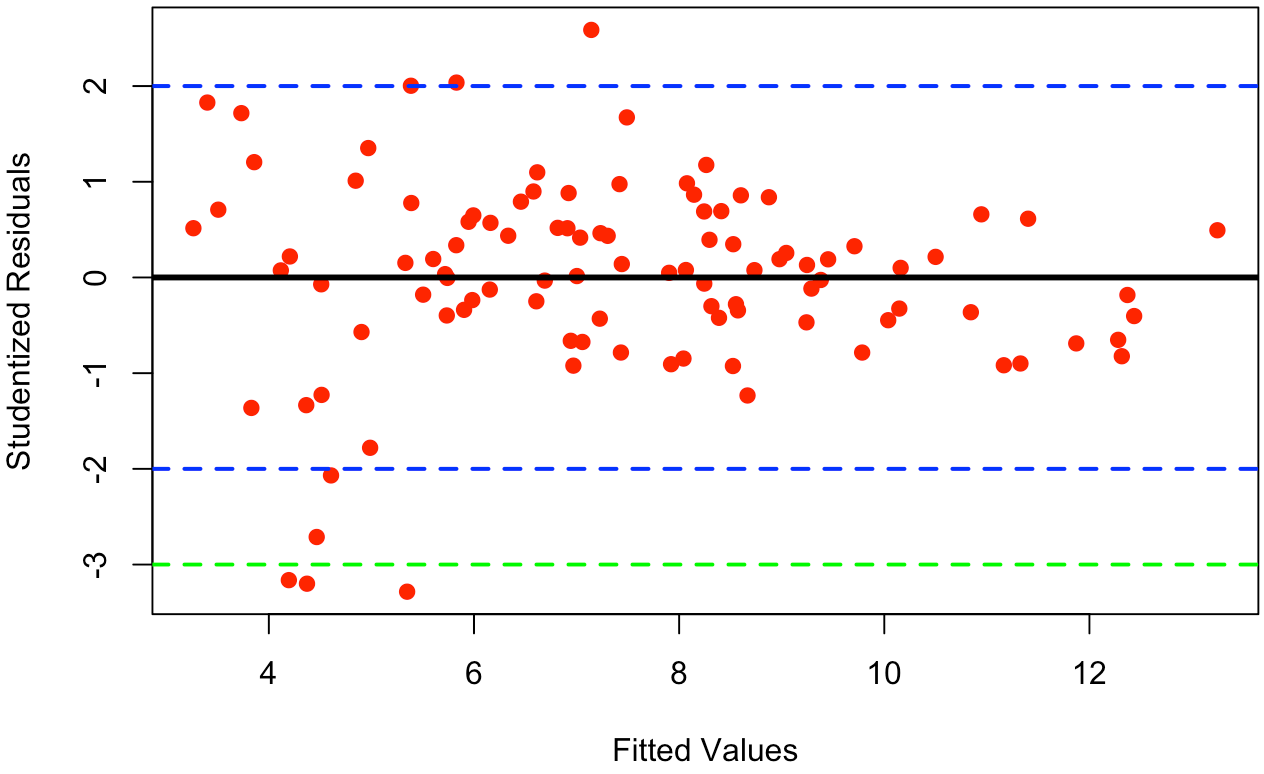
```r
resplots <- function(model, label) {
  #Normal quantile plot of studentized residuals
  qqPlot(rstudent(model),
         pch = 19,
         col = "black",
         main = paste("NQ Plot of Studentized Residuals, ", label))
  #plot of fitted vs. studentized residuals
  plot(rstudent(model) ~ model$fitted.values,
       pch = 19,
       col = 'red',
       xlab = "Fitted Values",
       ylab = "Studentized Residuals",
       main = paste("Fits vs. Studentized Residuals, ", label))
  abline(h = 0, lwd = 3)
  abline(h = c(2,-2), lty = 2, lwd = 2, col="blue")
  abline(h = c(3,-3), lty = 2, lwd = 2, col="green")
}

resplots(m10, label = "(ln) COVID-19 Cases")
```

NQ Plot of Studentized Residuals, (ln) COVID-19 Cases



Fits vs. Studentized Residuals, (ln) COVID-19 Cases



Summary

In this project, we wanted to analyze how different demographics of countries relate to their respective COVID-19 situations. To do this, we used up-to-date World Bank data, as well as categorical variables that we scraped ourselves.

After cleaning our data, we created descriptive plots for some of the variables, and performed the appropriate transformations on the variables that showcased nonlinear trends with `logCases`.

We then created matrix plots and correlation plots with some of the continuous variables to assess possible relationships between the variables; we tried to choose variables for these analyses that would not be subject to multicollinearlity (but we encountered this issue regardless).

Subsequently, we performed a two-sample t-test and bootstrapped the difference in the sample means (`logCases`) of two climates (tropical and temperate). Not to mention, a permutation test was done in an attempt to explore the strength of the linear relationship between (ln) COVID-19 cases and (ln) airline activity.

Finally, we performed a multiple regression analysis with the possible predictors of (ln) COVID-19 cases, using the backwards-stepwise regression method. This was performed using the same continuous variables from our matrix plot analysis, plus the two categorical variables that we scraped.

Our model ended up with five significant predictors: `logCO2`, `PM25`, `logAir`, `LifeExp`, and `logDeaths`, all with p-values of less than alpha = 0.05 and positive coefficients. The overall model has strong predictive power, with 97 degrees of freedom, and a multiple r-squared value of 0.898. This means that roughly 89.8% of the variation in (ln) COVID-19 cases by country can be explained by our model with these five predictors.

Furthermore, the residual plots of our final model satisfy our regression assumptions, albeit with a slight truncation in the approximately normal distribution, as apparent on the Normal Quantile Plot (towards the lower end).

The residuals on the Fits versus Residuals plot possess a relatively constant variance for all fitted values of our Y variable `logCases`, meaning there is no evidence of heteroskedasticity. Notably, there are a handful of outliers hovering in-and-around the studentized residual of -3 on the residual plot, but we decided to ignore them.

Based on the little we know about this baffling virus, some of significant predictors we arrived at make intuitive sense, such as LifeExp, logDeaths, and logAir. The other predictors — logCO2 and PM25 — were a bit surprising.

Countries with higher values of these variables might have more air pollution due to being more industrial / developed; the densely-populated cities that grow around such industrial centers may thus have a relationship to the COVID-19 incidences in these regions.