Homework Week 11

w203: Statistics for Data Science

Adam Yang

Get familiar with the data

You receive a data set from World Bank Development Indicators.

- Load the data using load and see what is loaded by using ls(). You should see Data which is the data frame including data, and Definitions which is a data frame that includes variable names.

```
load("Week11.Rdata")
ls()
```

```
## [1] "Data"        "Definitions"
```

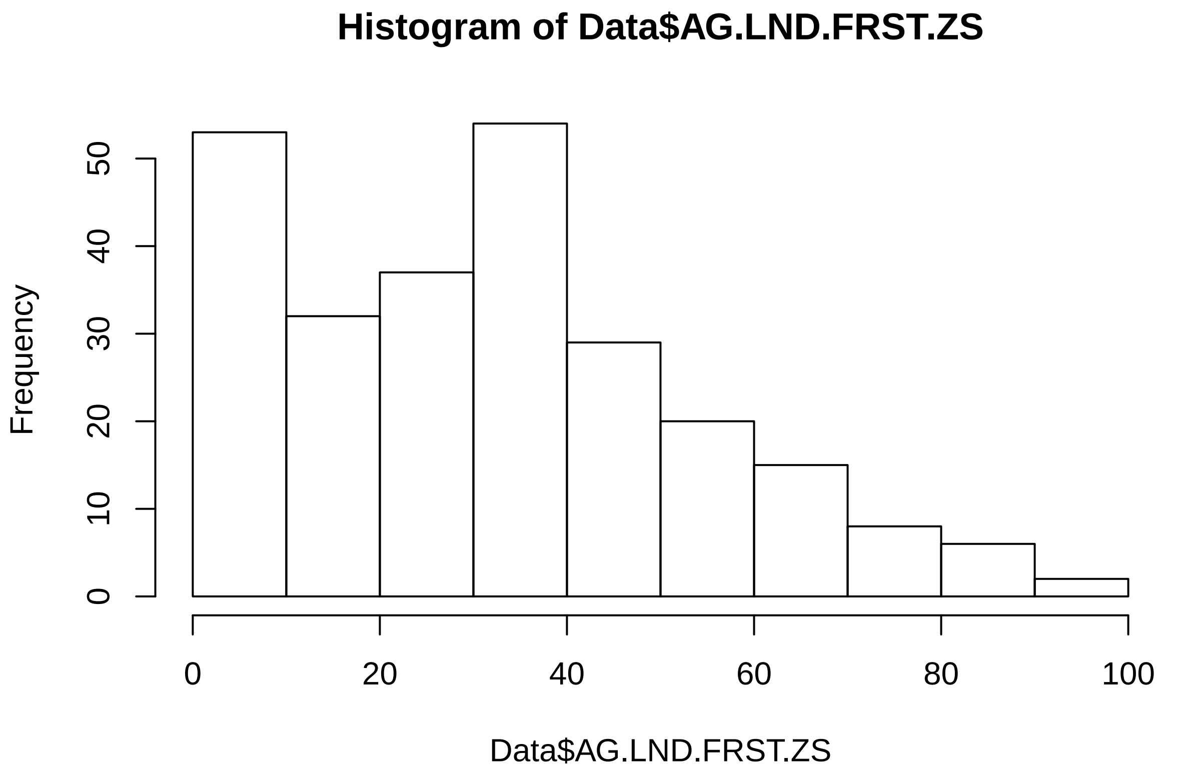
- Look at the variables, read their descriptions, and take a look at their histograms. Think about the transformations that you may need to use for these variables in the section below.

Most of the columns would benefit from a log transformation. I believe the log of the variables would yield a much better linear model.

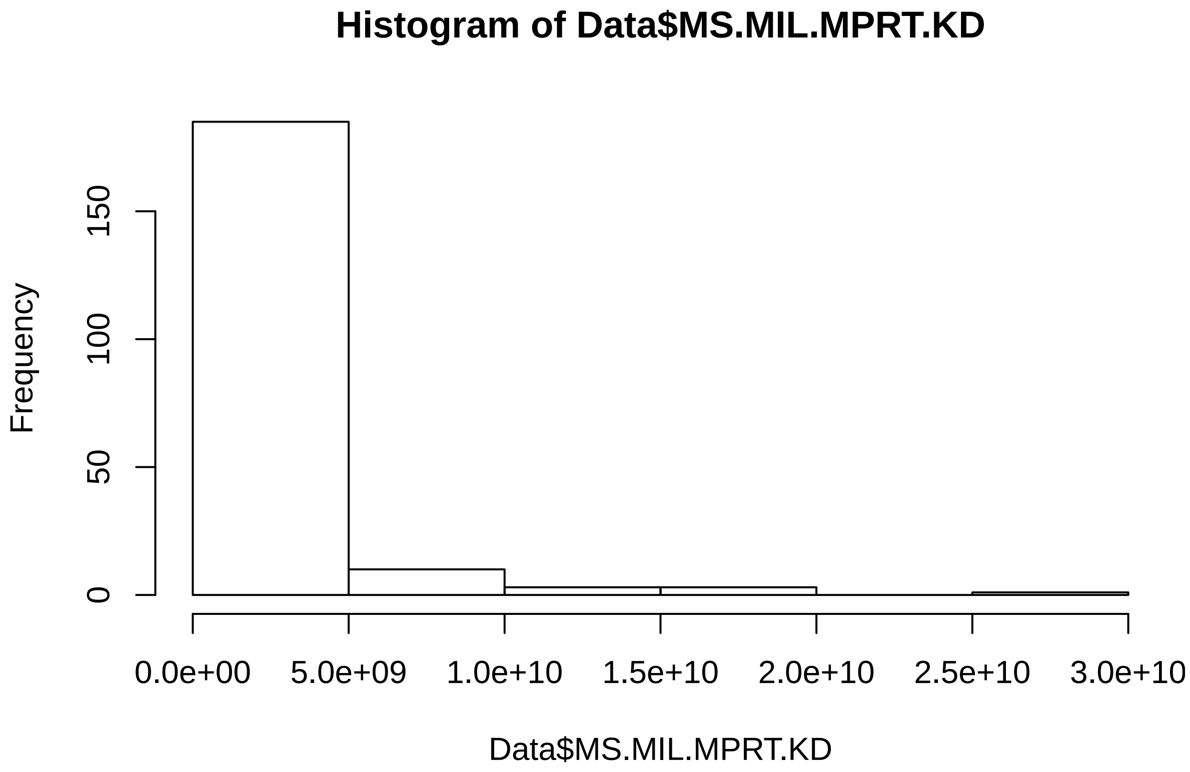
```
head(Definitions, 11)
```

```
##          Series.Code
## 1     AG.LND.FRST.ZS
## 2  MS.MIL.XPND.GD.ZS
## 3     MS.MIL.XPND.ZS
## 4     NY.GDP.MKTP.CD
## 5     NY.GDP.PCAP.CD
## 6  NY.GDP.PETR.RT.ZS
## 7     MS.MIL.XPRT.KD
## 8  TX.VAL.AGRI.ZS.UN
## 9     MS.MIL.MPRT.KD
## 10    NE.IMP.GNFS.CD
## 11    NE.EXP.GNFS.CD
##                                                    Series.Name
## 1                                  Forest area (% of land area)
## 2                                 Military expenditure (% of GDP)
## 3     Military expenditure (% of central government expenditure)
## 4                                              GDP (current US$)
## 5                                  GDP per capita (current US$)
## 6                                          Oil rents (% of GDP)
## 7                        Arms exports (SIPRI trend indicator values)
## 8  Agricultural raw materials exports (% of merchandise exports)
## 9                        Arms imports (SIPRI trend indicator values)
## 10                       Imports of goods and services (current US$)
## 11                       Exports of goods and services (current US$)
```

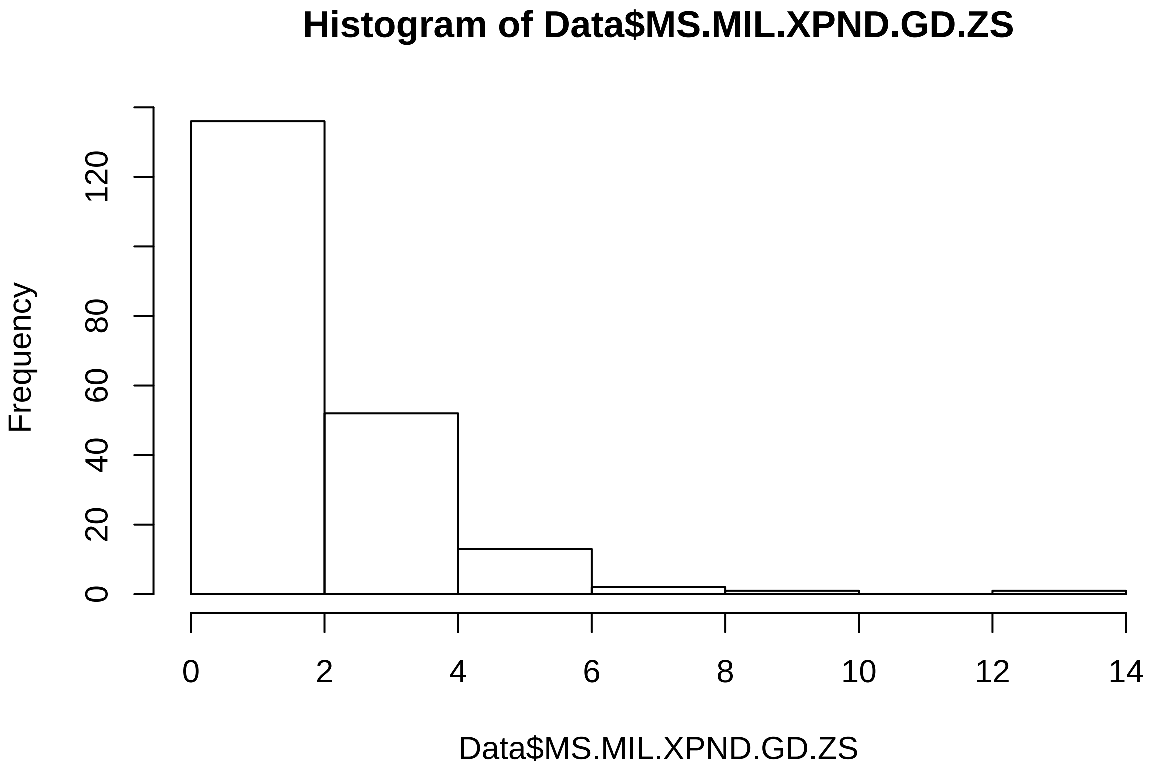
```
hist(Data$AG.LND.FRST.ZS) # Maybe take sqrt
```



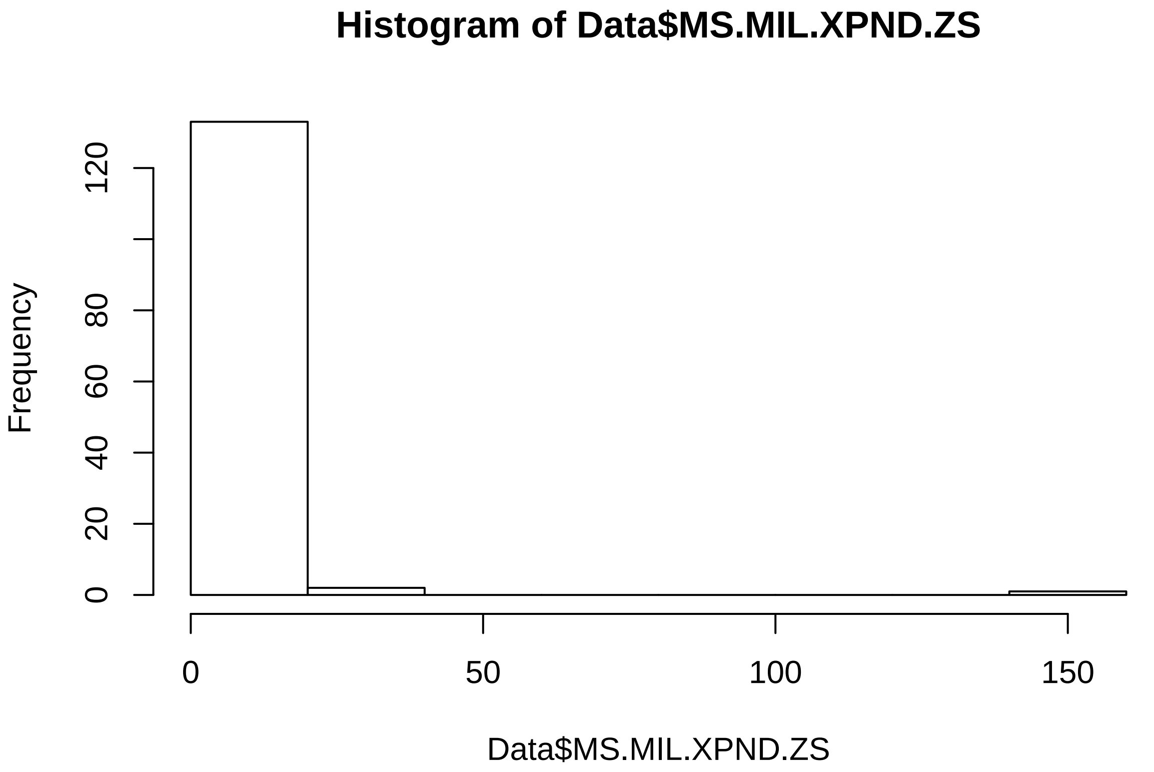
```r
hist(Data$MS.MIL.MPRT.KD) # Log This
```



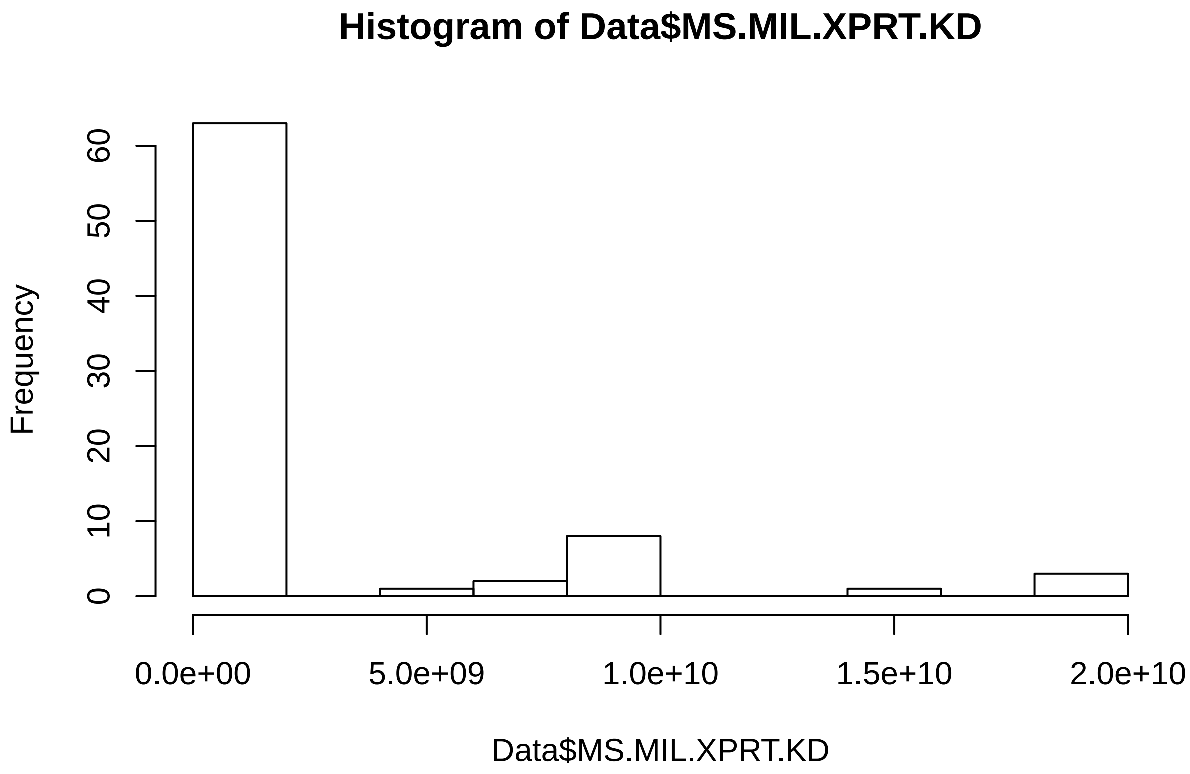
```r
hist(Data$MS.MIL.XPND.GD.ZS) # Maybe take sqrt
```



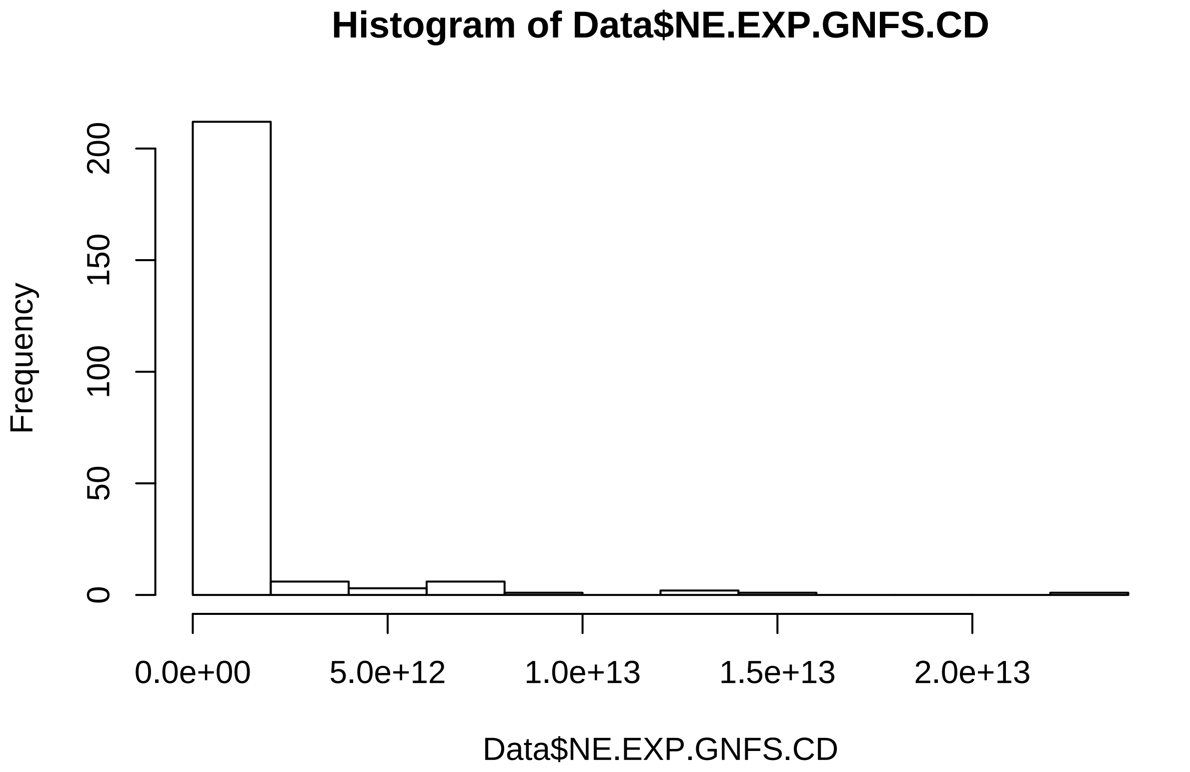
```
hist(Data$MS.MIL.XPND.ZS) # Log This (Maybe sqrt)
```



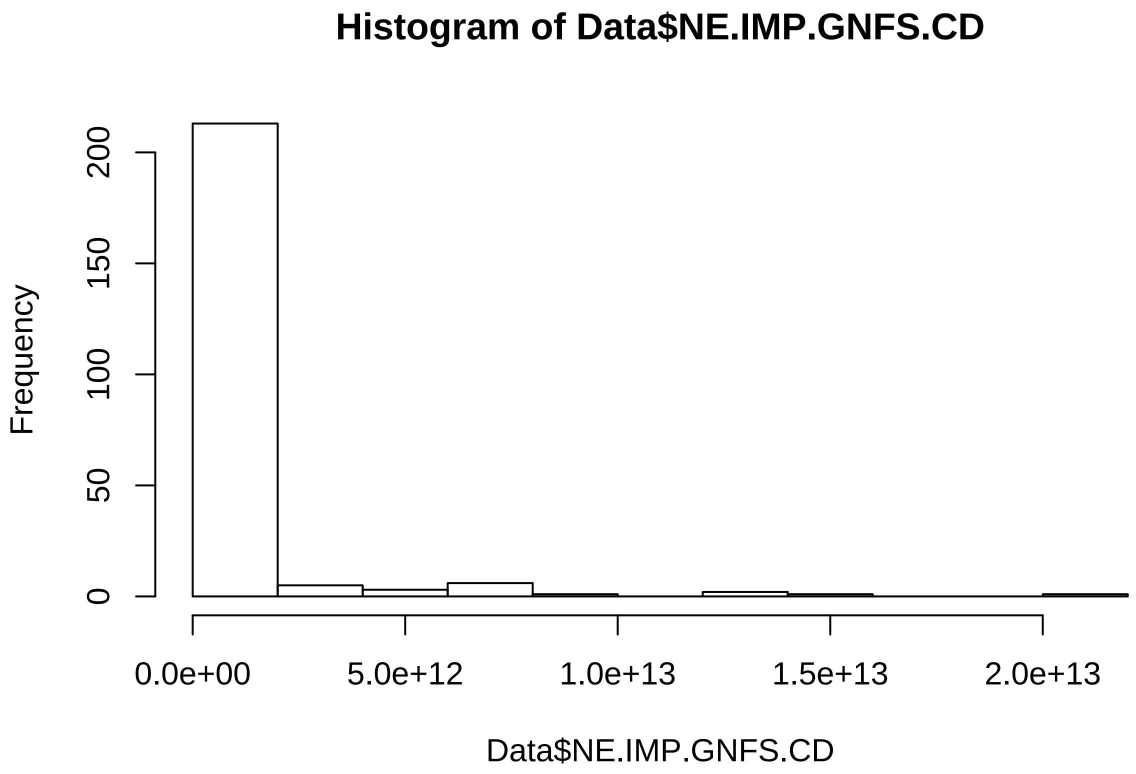
```
hist(Data$MS.MIL.XPRT.KD) # Log This
```



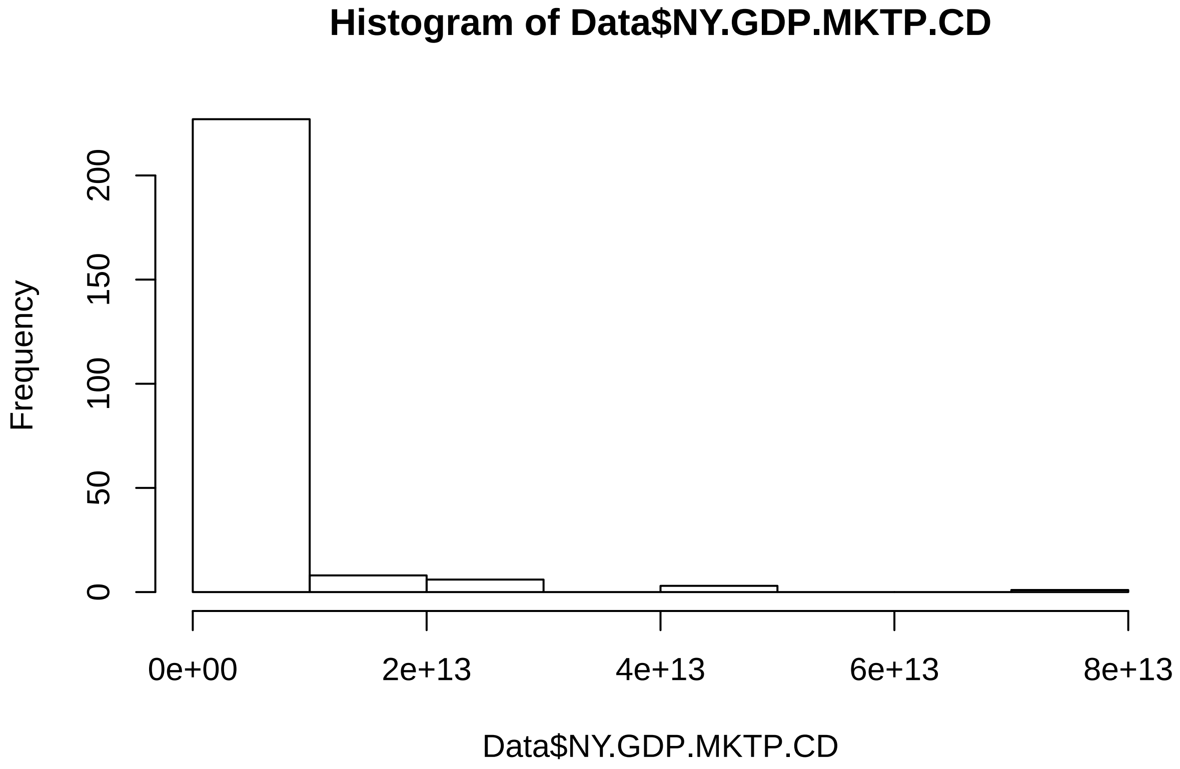
```r
hist(Data$NE.EXP.GNFS.CD) # Log This
```



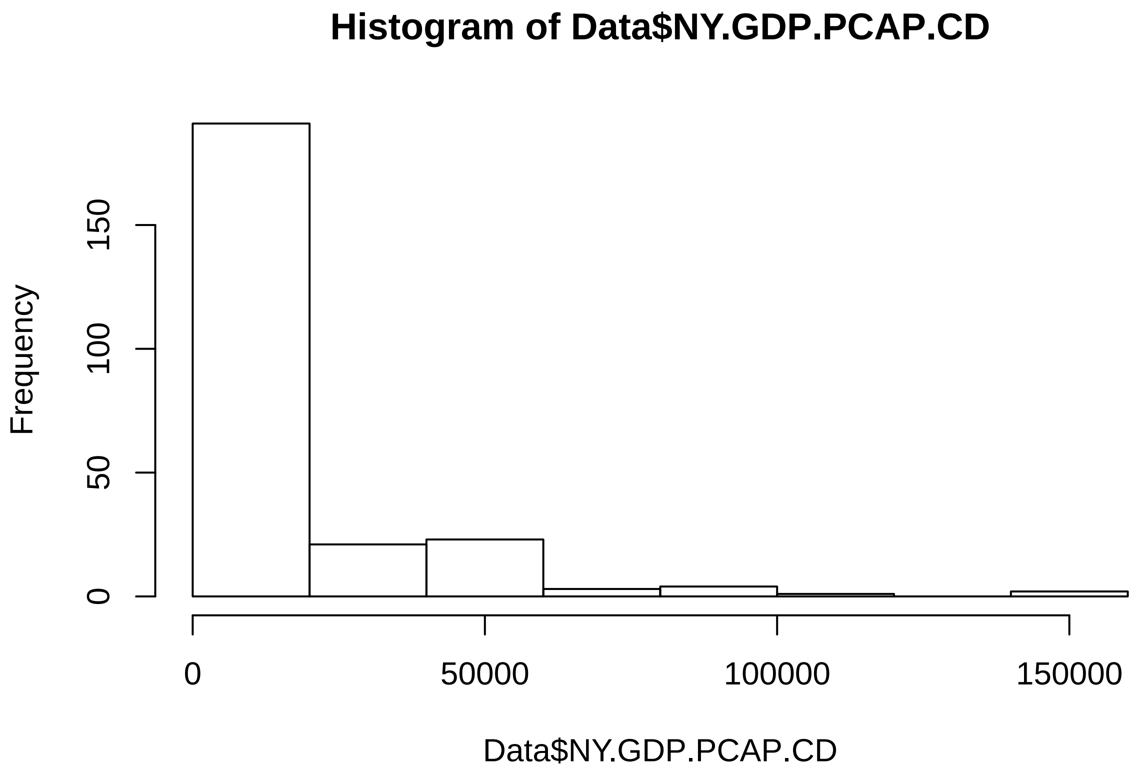
```r
hist(Data$NE.IMP.GNFS.CD) # Log This
```



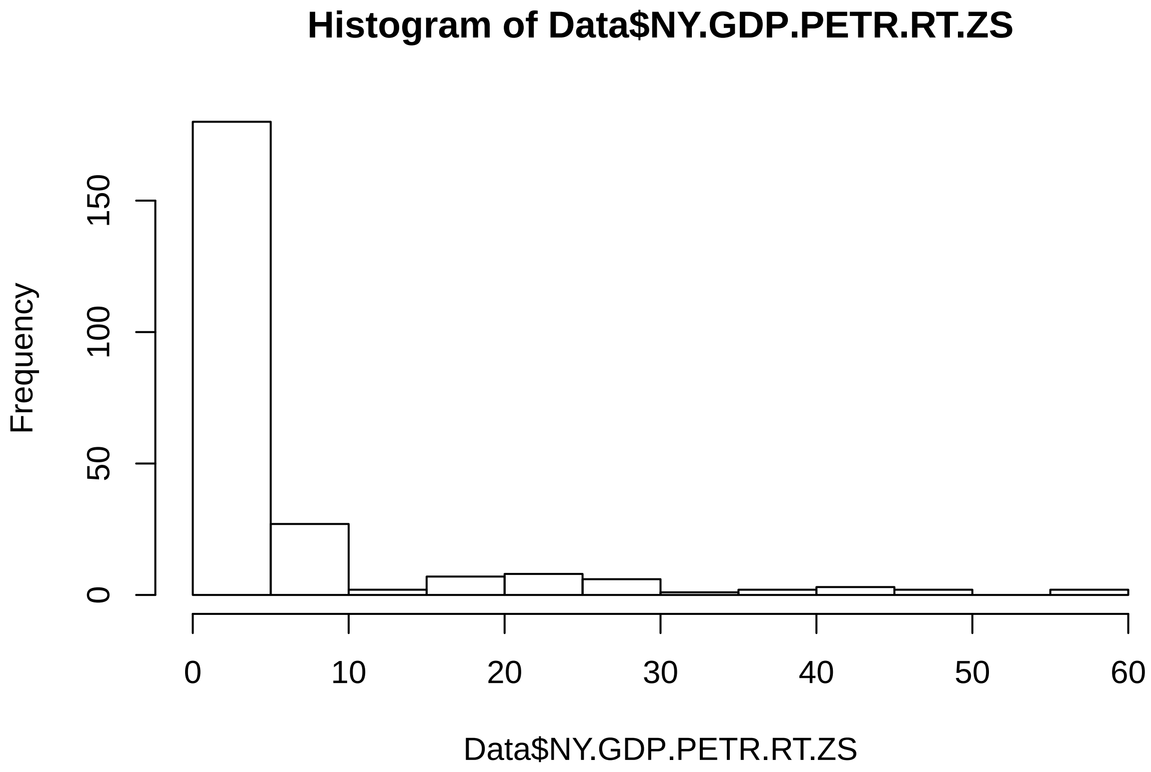
```
hist(Data$NY.GDP.MKTP.CD) # Log This
```



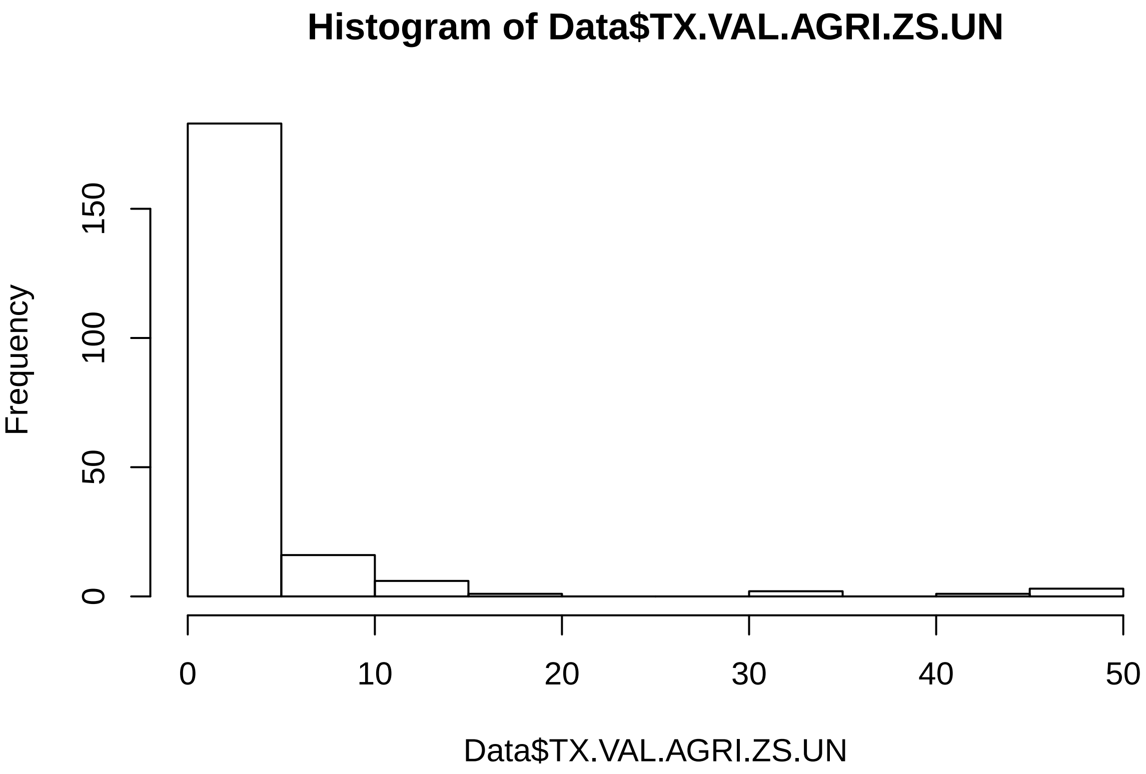
```
hist(Data$NY.GDP.PCAP.CD) # Log This
```



```
hist(Data$NY.GDP.PETR.RT.ZS) # Log This (Maybe sqrt)
```



```
hist(Data$TX.VAL.AGRI.ZS.UN) # Log This (Maybe sqrt)
```



- Run: `apply(!is.na(Data[,-(1:2)] ) , MARGIN= 2, mean )` and explain what it is showing.

```
apply(!is.na(Data[-(1:2)]), MARGIN = 2, mean)
```

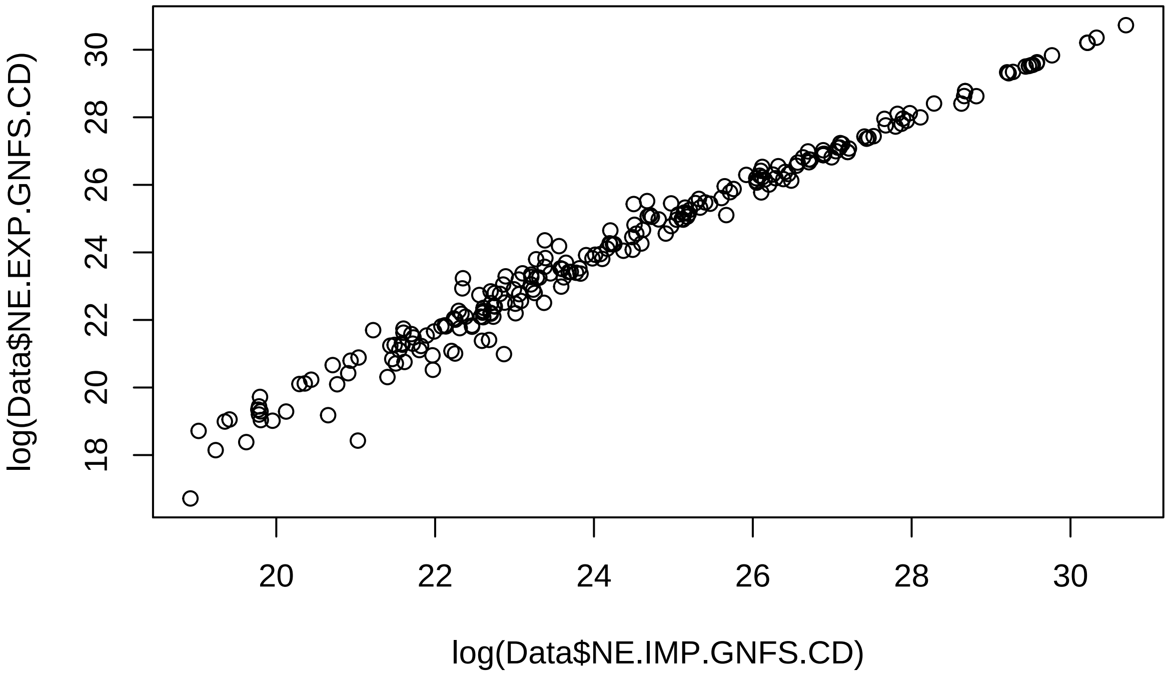
```
##     AG.LND.FRST.ZS     MS.MIL.MPRT.KD MS.MIL.XPND.GD.ZS     MS.MIL.XPND.ZS
##          0.9696970          0.7651515          0.7765152          0.5151515
##     MS.MIL.XPRT.KD     NE.EXP.GNFS.CD     NE.IMP.GNFS.CD     NY.GDP.MKTP.CD
##          0.2954545          0.8787879          0.8787879          0.9280303
##     NY.GDP.PCAP.CD NY.GDP.PETR.RT.ZS TX.VAL.AGRI.ZS.UN
##          0.9280303          0.9090909          0.8030303
```

The code is looking at each column of the data set but excluding the first 2 columns. If the value of the row is NaN, then we assign 0 to that row, otherwise assign a 1. Then we take the mean of all the values in the row. Essentially, we are calculating the percentage of values in the row that is not NaN.

- Can you include both `NE.IMP.GNFS.CD` and `NE.EXP.GNFS.CD` in the same OLS model? Why?

One of the assumptions of OLS modeling is that we have no perfect multicolinearity. It might be the case where the imports of a country are directly related to the exports and therefore there would be colinearity between the two variables. The plot of the logs of the 2 variables (shown below) suggests that there is a pretty strong linear relationship between the two. Therefore, including both of these variables would violate the “no perfect multicolinearity”" assumption.

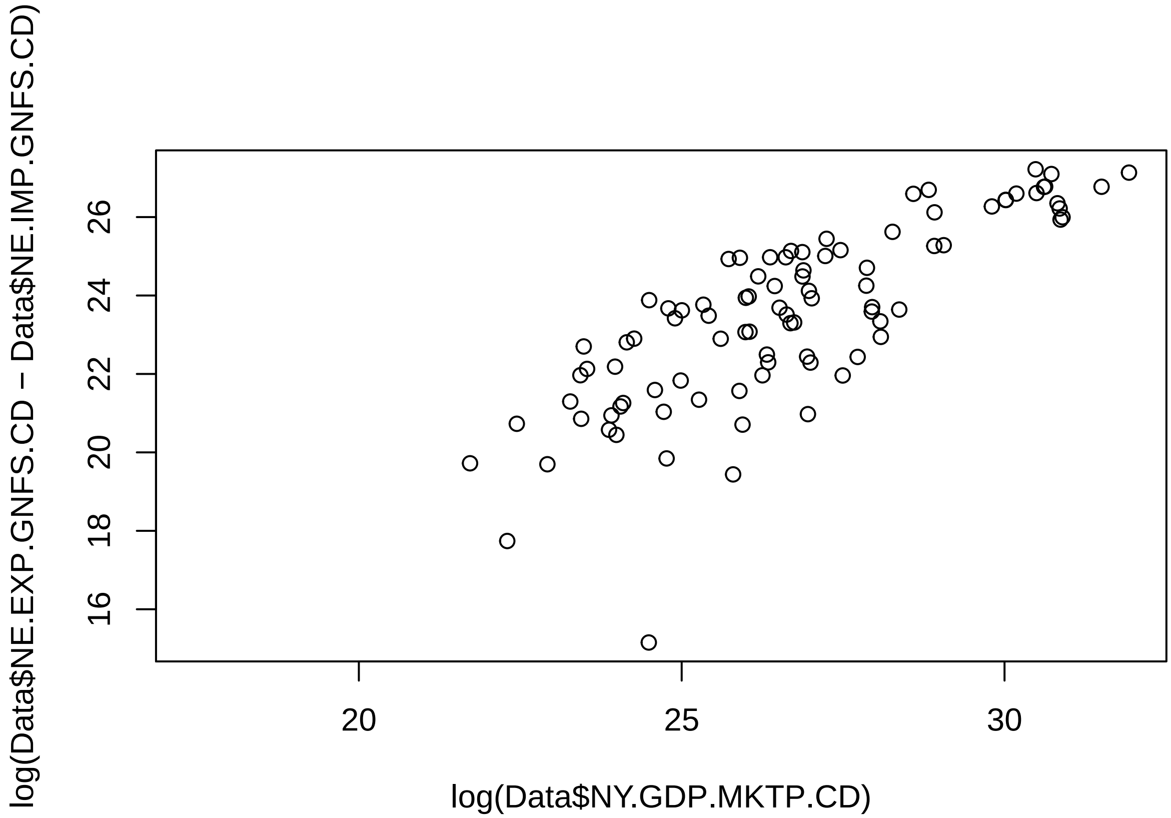
```
plot(log(Data$NE.IMP.GNFS.CD), log(Data$NE.EXP.GNFS.CD))
```



Another thing that can be problematic is that I believe the GDP is a linear combination of import and exports. I belive part of the GDP formula is exports minus imports so having both of these variables in the same model with GDP can possibly violate our “no perfect multicolinearity” assumption as well. The plot below shows the log of GDP vs the log of exports minus imports and it shows a decent linear relationship.

```
plot(log(Data$NY.GDP.MKTP.CD), log(Data$NE.EXP.GNFS.CD - Data$NE.IMP.GNFS.CD))
```

```
## Warning in log(Data$NE.EXP.GNFS.CD - Data$NE.IMP.GNFS.CD): NaNs produced
```



- Rename the variable named `AG.LND.FRST.ZS` to forest. This is going to be our dependent variable.

```
colnames(Data)[3] <- "forest"
```

Decribe a model for that predicts forest

- Write a model with two explanatory variables.

The first 2 variables I chose are:

1. TX.VAL.AGRI.ZS.UN - Agricultural raw materials exports (% of merchandise exports)
2. NY.GDP.PCAP.CD - GDP per capita (current US$)

Therefore, the model would be: $forest = \beta_0 + \beta_1(AgricultureExports) + \beta_2(GDPPerCapita) + u$

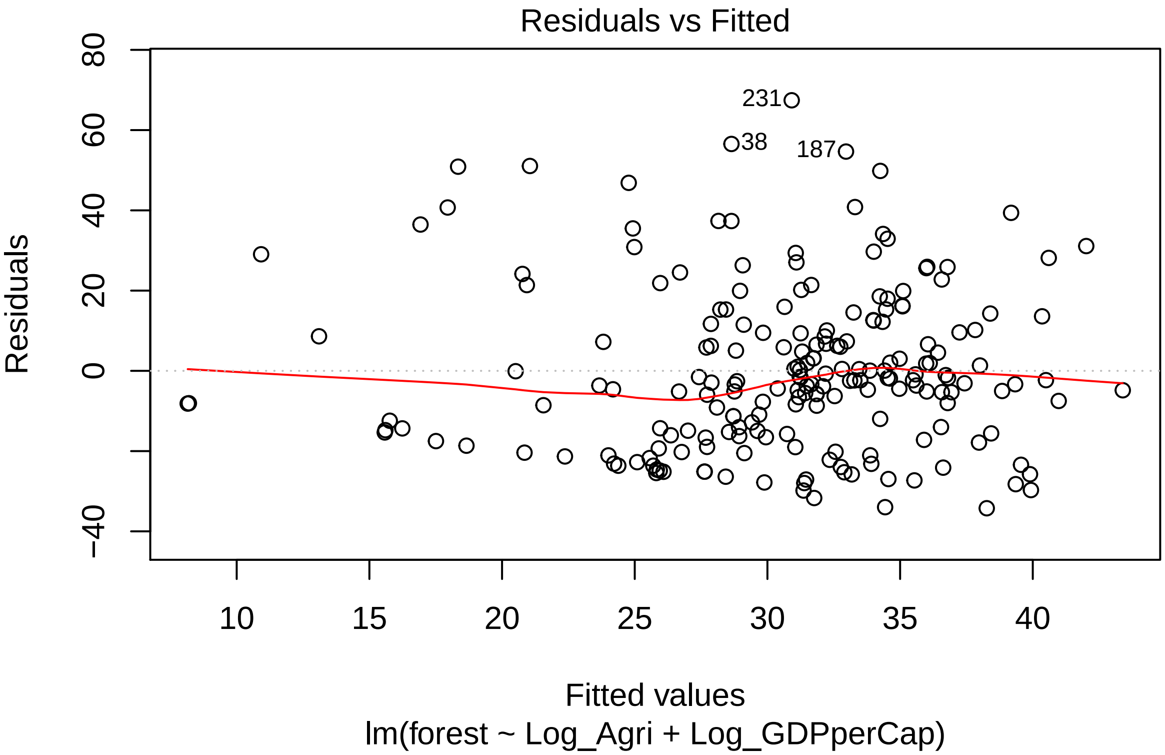
```
Data2 <- data.frame("forest" = Data$forest,
                    "TX.VAL.AGRI.ZS.UN" = Data$TX.VAL.AGRI.ZS.UN,
                    "NY.GDP.PCAP.CD" = Data$NY.GDP.PCAP.CD)



# Take log for Agriculture Exports variable.
Data2["Log_Agri"] <- log(Data2$TX.VAL.AGRI.ZS.UN)
# Make sure there aren't any -Inf after taking the log.
Data2$Log_Agri[which(Data2$Log_Agri==-Inf)] = NaN

# Take log for GDP per Capita variable.
Data2["Log_GDPperCap"] <- log(Data2$NY.GDP.PCAP.CD)
# Make sure there aren't any -Inf after taking the log.
Data2$Log_GDPperCap[which(Data2$Log_GDPperCap==-Inf)] = NaN

Data2 <- na.omit(Data2)
```

```
model1 <- lm(forest~Log_Agri+Log_GDPperCap, data = Data2)
model1
```

```
##
## Call:
## lm(formula = forest ~ Log_Agri + Log_GDPperCap, data = Data2)
##
## Coefficients:
##   (Intercept)       Log_Agri  Log_GDPperCap
##         8.133          3.438          2.541
```

`* Create a residuals versus fitted values plot and assess whether your coefficients are unbiased.`

```
plot(model1,1)
```



It looks like the variables are hovering pretty well around the zero line which suggests pretty decent linearity. The positive residuals seem to be a bit larger than the negative residuals. Maybe this is an indication of skewed errors. Some of the positive residuals might be outliers. Furthermore, the residuals seem to be quite far from the 0 line. Therefore, I think our coefficients are decently unbiased, but there might be room for improvement.

`* How many observations are being used in your analysis?`

```
length(model1$residuals)
```

```
## [1] 206
```

There are 206 observations used on our analysis.

`* Are the countries that are dropping out dropping out by random chance? If not, what would this` do to our inference?

```
Countries <- Data$Country.Name
# Countries that have NaN for Agriculture Exports
Countries[is.na(Data$TX.VAL.AGRI.ZS.UN)]
```

```
##  [1] American Samoa
##  [2] Andorra
##  [3] Angola
##  [4] British Virgin Islands
##  [5] Cayman Islands
##  [6] Chad
##  [7] Channel Islands
##  [8] Congo, Dem. Rep.
##  [9] Cuba
## [10] Curacao
## [11] Djibouti
## [12] Equatorial Guinea
## [13] Eritrea
## [14] Faroe Islands
## [15] Fragile and conflict affected situations
## [16] Gabon
## [17] Gibraltar
## [18] Grenada
## [19] Guam
## [20] Guinea-Bissau
## [21] Haiti
## [22] Isle of Man
## [23] Korea, Dem. People’s Rep.
## [24] Kosovo
## [25] Lao PDR
## [26] Least developed countries: UN classification
## [27] Liberia
## [28] Liechtenstein
## [29] Low income
## [30] Marshall Islands
## [31] Micronesia, Fed. Sts.
## [32] Monaco
## [33] Montenegro
## [34] Nauru
## [35] Northern Mariana Islands
## [36] Not classified
## [37] Pre-demographic dividend
## [38] Puerto Rico
## [39] San Marino
## [40] Serbia
## [41] Seychelles
## [42] Sint Maarten (Dutch part)
## [43] Somalia
## [44] South Sudan
## [45] St. Martin (French part)
## [46] Swaziland
## [47] Tajikistan
## [48] Turkmenistan
## [49] Tuvalu
## [50] Uzbekistan
## [51] Virgin Islands (U.S.)
## [52] West Bank and Gaza
## 267 Levels:  Afghanistan Albania Algeria American Samoa Andorra ... Zimbabwe
```

```
# Countries that have NaN for GDP
Countries[is.na(Data$NY.GDP.PCAP.CD)]
```

```
##  [1] American Samoa            British Virgin Islands
##  [3] Cayman Islands            Channel Islands
##  [5] Curacao                   French Polynesia
##  [7] Gibraltar                 Guam
##  [9] Korea, Dem. People’s Rep. Nauru
## [11] New Caledonia             Northern Mariana Islands
## [13] Not classified            San Marino
## [15] Sint Maarten (Dutch part) St. Martin (French part)
## [17] Syrian Arab Republic      Turks and Caicos Islands
## [19] Virgin Islands (U.S.)
## 267 Levels:  Afghanistan Albania Algeria American Samoa Andorra ... Zimbabwe
```

There might a number of reasons why we are missing data from certain countries. Maybe some countries don’t keep track of their GDP or their Agriculture exports. Maybe some countries just refused to share the information because they have an unfriendly relationship with the UN such as North Korea. However, judging by the countries listed above, I don’t believe countries are dropping out by random chance. There are reasons that specific countries are dropping out due to the lack of data rather than random loss of data. Because countries are dropping out without random chance, our sampling cannot be considered perfectly random. It can cause a bias to our model.

- Now add a third variable.

The third variable I will add is: NE.EXP.GNFS.CD - Exports of goods and services (current US$)

```
Data3 <- data.frame("forest" = Data$forest,
                    "TX.VAL.AGRI.ZS.UN" = Data$TX.VAL.AGRI.ZS.UN,
                    "NY.GDP.PCAP.CD" = Data$NY.GDP.PCAP.CD,
                    "NE.EXP.GNFS.CD" = Data$NE.EXP.GNFS.CD)

# Take log for Agriculture Exports variable.
Data3["Log_Agri"] <- log(Data3$TX.VAL.AGRI.ZS.UN)
# Make sure there aren't any -Inf after taking the log.
Data3$Log_Agri[which(Data3$Log_Agri==-Inf)] = NaN

# Take log for GDP per Capita variable.
Data3["Log_GDPperCap"] <- log(Data3$NY.GDP.PCAP.CD)
# Make sure there aren't any -Inf after taking the log.
Data3$Log_GDPperCap[which(Data3$Log_GDPperCap==-Inf)] = NaN

# Take log for Military Expenditure
Data3["Log_Export"] <- log(Data3$NE.EXP.GNFS.CD)
# Make sure there aren't any -Inf after taking the log.
Data3$Log_Export[which(Data3$NE.EXP.GNFS.CD==-Inf)] = NaN

Data3 <- na.omit(Data3)

model2 <- lm(forest~Log_Agri+Log_GDPperCap+Log_Export, data = Data3)
model2
```

```
##
```

```
## Call:
## lm(formula = forest ~ Log_Agri + Log_GDPperCap + Log_Export,
##     data = Data3)
##
## Coefficients:
##   (Intercept)       Log_Agri  Log_GDPperCap     Log_Export
##        27.577          3.608          4.322         -1.443
```

- Show how you would use the regression anatomy formula to compute the coefficient on your third variable. First, regress the third variable on your first two variables and extract the residuals. Next, regress forest on the residuals from the first stage.

Our regression formula is: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u$.

Lets regress our third variable on our first 2 variables: $x_3 = \gamma_0 + \gamma_1 x_1 + \gamma_2 x_2 + r_1$

The coefficient of the third variable is: $\beta_3 = \frac{cov(y,r_1)}{var(r_1)}$

Therefore, we can extract the residuals out of the our second equation to get r_1 to solve for β_3

```
# Regress the third variable on your first two variables and extract the residuals.
model3 <- lm(Log_Export ~ Log_Agri + Log_GDPperCap, data = Data3)
r1 <- model3$residuals

# Regress forest on the residuals from the first stage.
model4 <- lm(forest ~ r1, data = Data3)
model4
```

```
##
## Call:
## lm(formula = forest ~ r1, data = Data3)
##
## Coefficients:
## (Intercept)           r1
##      30.516       -1.443
```

Our slope coefficient is -1.443 which matches what we got in the previous section.

- Compare your two models.

The slope coefficient for the log of agriculture exports decreased from 3.438 to 3.608 and the slope coefficient for the log of GDP per Capita increased from 2.541 to 4.322.

```
* Do you see an improvement? Explain how you can tell.
```

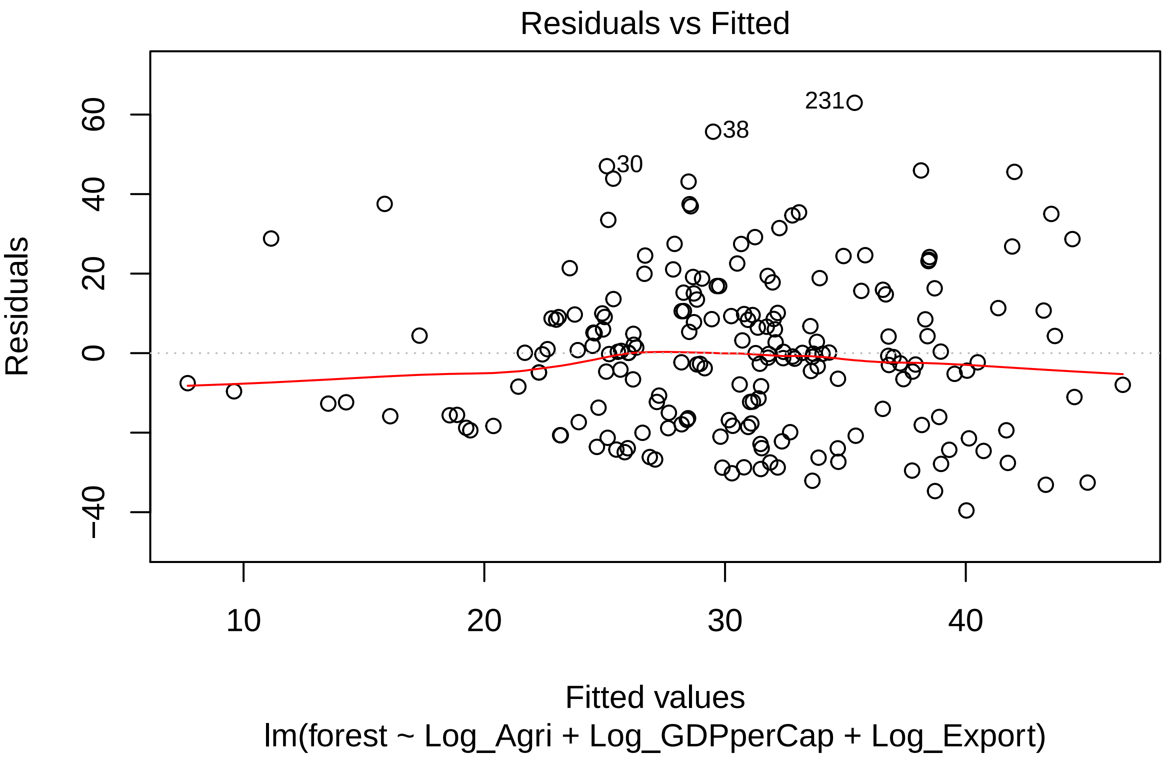
```
summary(model1)$adj.r.square
```

```
## [1] 0.07570573
```

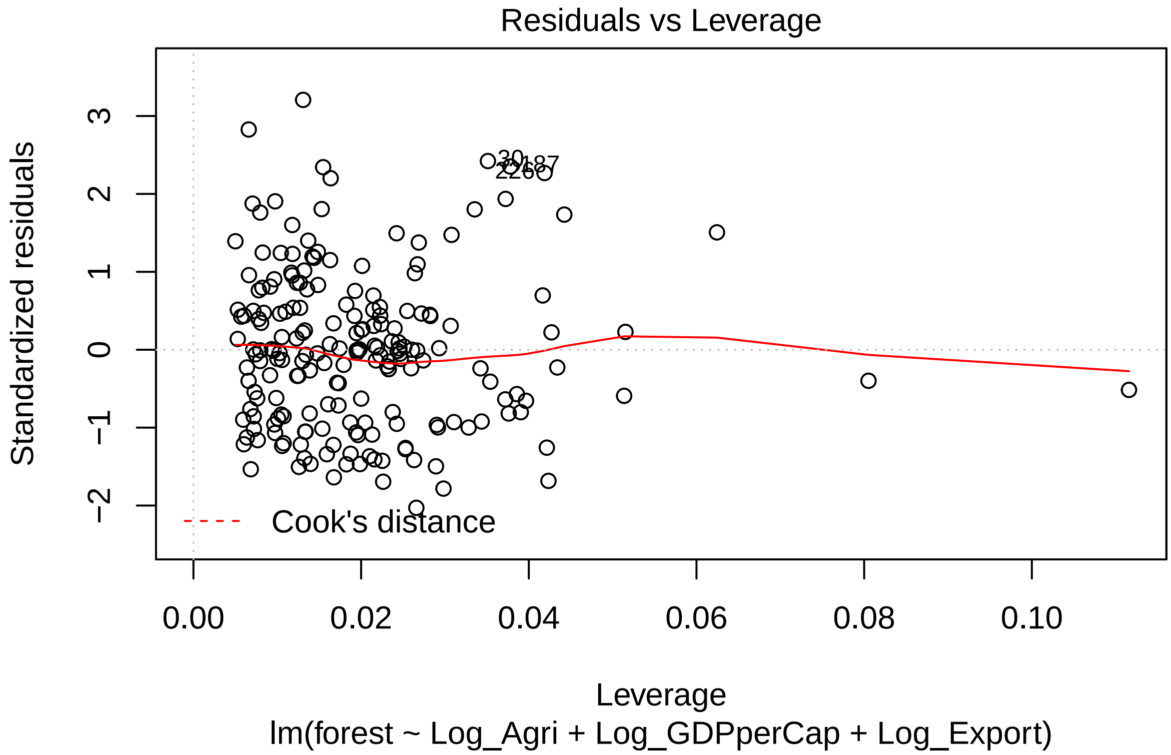
```
summary(model2)$adj.r.square
```

```
## [1] 0.09545353
```

```
plot(model2,1)
```



```r
plot(model2,5)
```



It seems like our residuals vs fitted value graph is quite a bit better. However, there are a couple of outlier points that are getting close to Cook's exclusion distance right above -3 for residuals. This can suggest better causality as the error term doesn't change as we manipulate the inputs. Also, the adjusted R-Squared values have increased a bit from 0.0757 to 0.0954. This suggests that the new model explains a bit more of the variability of the response data around its mean. However, the R-Squared value does tell us if the coefficient estimates and predictions are biased and you can have a high R-Squared value for a model that does not fit the data. Even so, it does seem like we removed some bias from the residuals, so I would say that adding the third export variable does seem to have resulted in an improvement.

Make up a country

- Make up a country named Mediland which has every indicator set at the median value observed in the data.

```
Mediland <- data.frame( "TX.VAL.AGRI.ZS.UN" = median(Data$TX.VAL.AGRI.ZS.UN, na.rm = TRUE),
                        "NY.GDP.PCAP.CD" = median(Data$NY.GDP.PCAP.CD,  na.rm = TRUE),
                        "NE.EXP.GNFS.CD" = median(Data$NE.EXP.GNFS.CD,  na.rm = TRUE))

Mediland["Log_Agri"] <- log(Mediland$TX.VAL.AGRI.ZS.UN)
Mediland["Log_GDPperCap"] <- log(Mediland$NY.GDP.PCAP.CD)
Mediland["Log_Export"] <- log(Mediland$NE.EXP.GNFS.CD)
```

- How much forest would this country have?

```
predict(model1 , data.frame(Log_Agri = Mediland$Log_Agri, Log_GDPperCap = Mediland$Log_GDPperCap, Log_E
```

```
##        1
## 31.77643
```

Take away

- What is the causal story, if any, that you can take away from the above analysis? Explain why.

I think what the negative slope coefficient for the exports of goods and services (current US$) tells us that as a country exports more they might have a more thriving industrial economy set up which would result in fewer forested ares. However, the positive slope coefficient for agricultural raw materials exports (% of merchandise exports) suggests that if the the country focuses more on agricultural exports, they might retain more forested land. The positive slope coefficient for GDP per Capita (current US $), is a bit of a mystery to me. I guess GDP per capita is not the same as raw GDP so it does not necessarily mean a thriving economy. Maybe a country like China would have high GDP but even higher population, so their GDP per capita would be lower. The slope suggests that forested areas has a positive correlation to GDP per Capita which goes slightly against my intuition. Maybe I should've used the raw GDP variable instead of GDP per Capita. If I were to manage a guess, I think maybe countries that have high GDP per Capita isn't only interested in raw economic growth and therefore put efforts into conserving forested areas rather than cutting them down.