HW Week 11

W203: Statistics for Data Science

Answer Key

Load data

Load the data and review the definitions:

```
#setwd("~/Desktop/W203/HW11/")
getwd()
```

```
## [1] "/Users/8ps/Google Drive/_UCB_W203/W203_Spring2017/LiveSession/week_11"
```

```
#load("Week11.Rdata")
load("~/Google Drive/_UCB_W203/W203 Update/weekly_materials/week_11/HW11/Week11.Rdata")
objects()
```

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

```
Definitions
```

```
##          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$)
## 11                  Exports of goods and services (current US$)
```

Rename variables

Rename all the variables to shorter, more meaningful and easier to use names and examine the `head()` and `summary()` of the data:

```
oldvars = c("AG.LND.FRST.ZS", "MS.MIL.XPND.GD.ZS", "MS.MIL.XPND.ZS", "NY.GDP.MKTP.CD",
            "NY.GDP.PCAP.CD", "NY.GDP.PETR.RT.ZS", "MS.MIL.XPRT.KD", "TX.VAL.AGRI.ZS.UN",
            "MS.MIL.MPRT.KD", "NE.IMP.GNFS.CD", "NE.EXP.GNFS.CD")
newvars = c("forest", "m_exp_gdp", "m_exp_gov", "gdp", "gdp_pc", "oil", "arms_ex", "agri",
            "arms_im", "imp", "exp")
for (i in 3:13) {names(Data)[i] = newvars[match(names(Data)[i],oldvars)]}
head(Data)
```

```
##     Country.Name Country.Code    forest   arms_im m_exp_gdp m_exp_gov
## 1    Afghanistan          AFG  2.067825 359166667  1.375170  3.183401
## 2        Albania          ALB 28.244526   9000000  1.413202       NaN
## 3        Algeria          DZA  0.813271 721500000  4.843526 14.512495
## 4 American Samoa          ASM 88.133333       NaN       NaN       NaN
## 5        Andorra          ADO 34.042553       NaN       NaN       NaN
## 6         Angola          AGO 46.657576  31333333  4.187594 14.098817
##   arms_ex         exp          imp          gdp    gdp_pc       oil
## 1     NaN  1304521083  8529983326  18949924158   626.788  0.000000
## 2       0  3955082222  6365588048  12442032457  4291.004  4.101974
## 3     NaN 70304960460 59880526175 193388057520  5114.370 22.388953
## 4     NaN         NaN         NaN          NaN       NaN       NaN
## 5     NaN         NaN         NaN   3292207861 40935.583  0.000000
## 6     NaN 59957802009 44133763534 109385918387  4730.046 39.340237
##         agri
## 1 4.79343482
## 2 2.20095479
## 3 0.01595214
## 4        NaN
## 5        NaN
## 6        NaN
```

```
summary(Data)
```

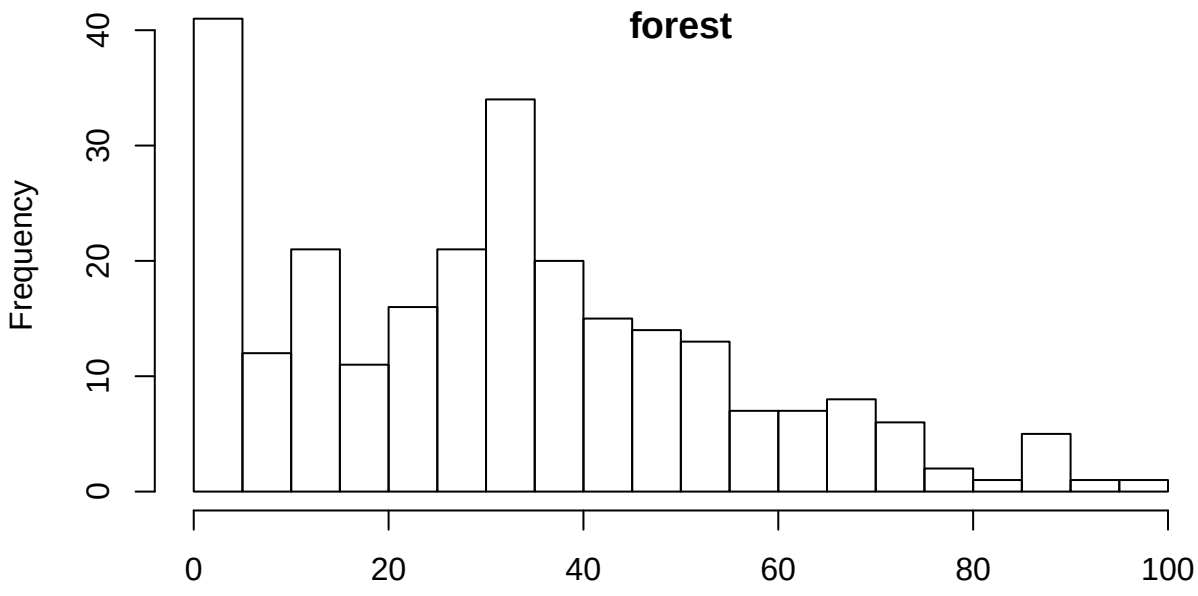
```
##          Country.Name  Country.Code     forest         arms_im
##  Afghanistan   :  1   ABW    :  1   Min.   : 0.00   Min.   :0.000e+00
##  Albania       :  1   ADO    :  1   1st Qu.:12.47   1st Qu.:1.081e+07
##  Algeria       :  1   AFG    :  1   Median :31.11   Median :7.458e+07
##  American Samoa:  1   AGO    :  1   Mean   :31.53   Mean   :1.299e+09
##  Andorra       :  1   ALB    :  1   3rd Qu.:46.00   3rd Qu.:7.234e+08
##  Angola        :  1   ARB    :  1   Max.   :98.34   Max.   :2.804e+10
##  (Other)       :258   (Other):258   NA's   :8       NA's   :62
##    m_exp_gdp        m_exp_gov          arms_ex
##  Min.   : 0.000   Min.   :  0.000   Min.   :0.000e+00
##  1st Qu.: 1.115   1st Qu.:  4.074   1st Qu.:1.800e+07
##  Median : 1.535   Median :  6.746   Median :5.733e+07
##  Mean   : 1.997   Mean   :  8.947   Mean   :2.266e+09
##  3rd Qu.: 2.426   3rd Qu.: 10.467   3rd Qu.:1.434e+09
```

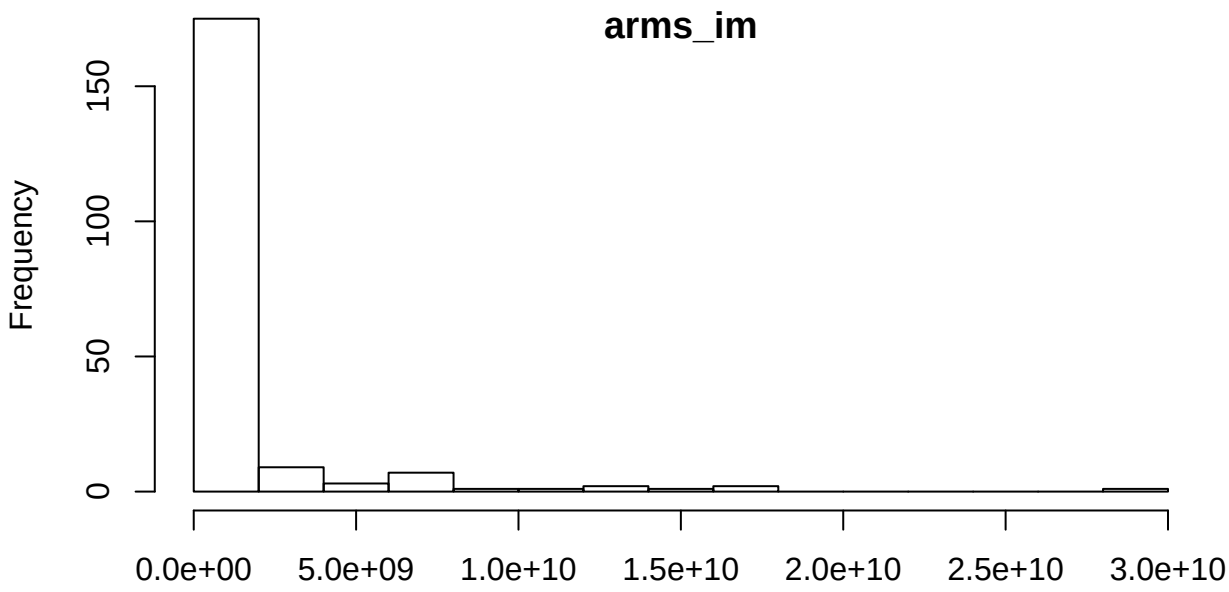
```
##  Max.   :12.787   Max.   :144.906   Max.   :1.816e+10
##  NA's   :59        NA's   :128       NA's   :186
##       exp                 imp                 gdp
##  Min.   :1.817e+07   Min.   :1.646e+08   Min.   :3.744e+07
##  1st Qu.:3.855e+09   1st Qu.:5.594e+09   1st Qu.:8.998e+09
##  Median :2.823e+10   Median :2.904e+10   Median :5.262e+10
##  Mean   :7.813e+11   Mean   :7.589e+11   Mean   :2.469e+12
##  3rd Qu.:2.894e+11   3rd Qu.:2.892e+11   3rd Qu.:5.396e+11
##  Max.   :2.210e+13   Max.   :2.149e+13   Max.   :7.346e+13
##  NA's   :32          NA's   :32          NA's   :19
##      gdp_pc              oil               agri
##  Min.   :   253.4   Min.   : 0.0000   Min.   : 0.00022
##  1st Qu.:  1687.2   1st Qu.: 0.0000   1st Qu.: 0.59231
##  Median :  5785.5   Median : 0.1494   Median : 1.60804
##  Mean   : 14975.8   Mean   : 5.2032   Mean   : 3.47449
##  3rd Qu.: 15065.1   3rd Qu.: 5.0281   3rd Qu.: 3.29650
##  Max.   :154286.4   Max.   :57.7407   Max.   :49.05388
##  NA's   :19         NA's   :24        NA's   :52
```

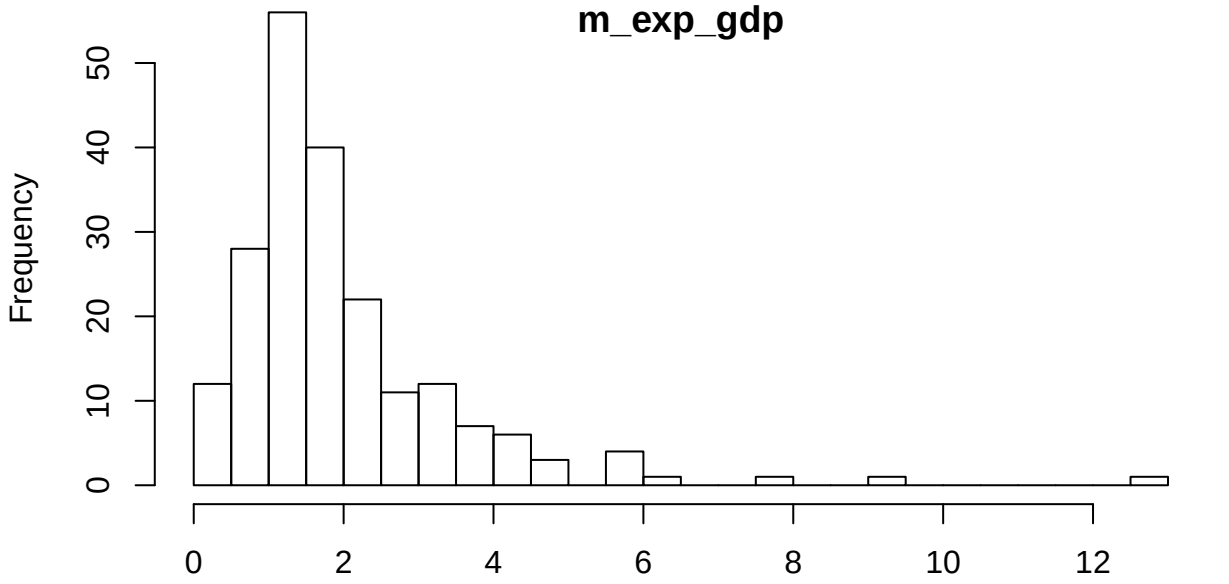
Exploratory data analysis

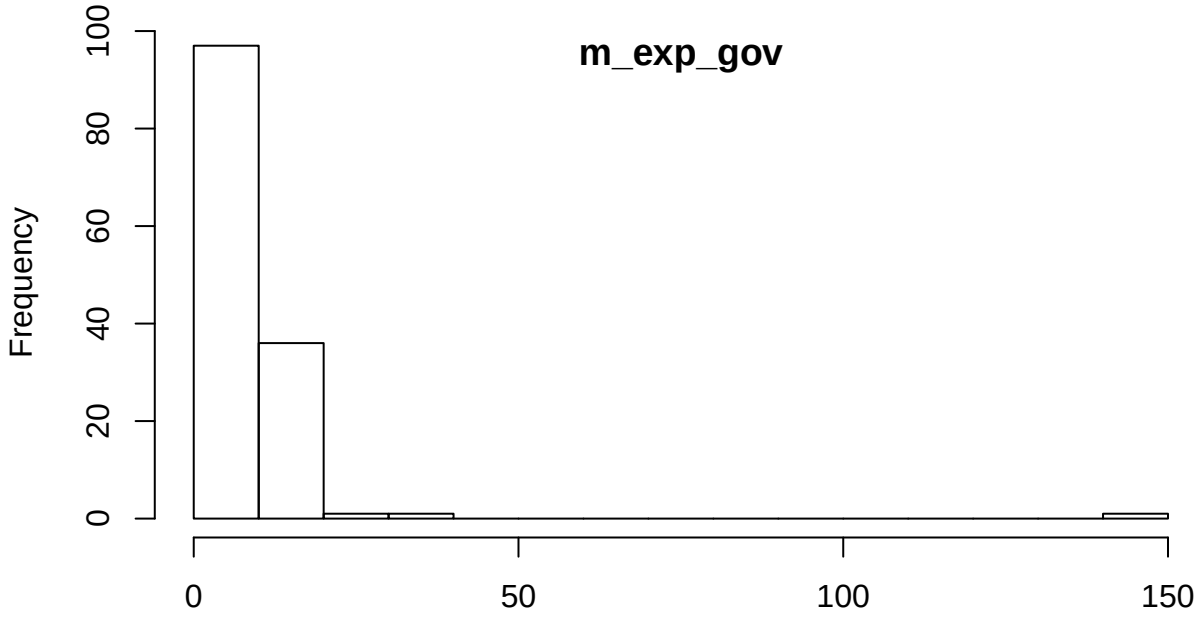
Look at their histograms and think about transformations that you may need to use for these variables in the model section below.

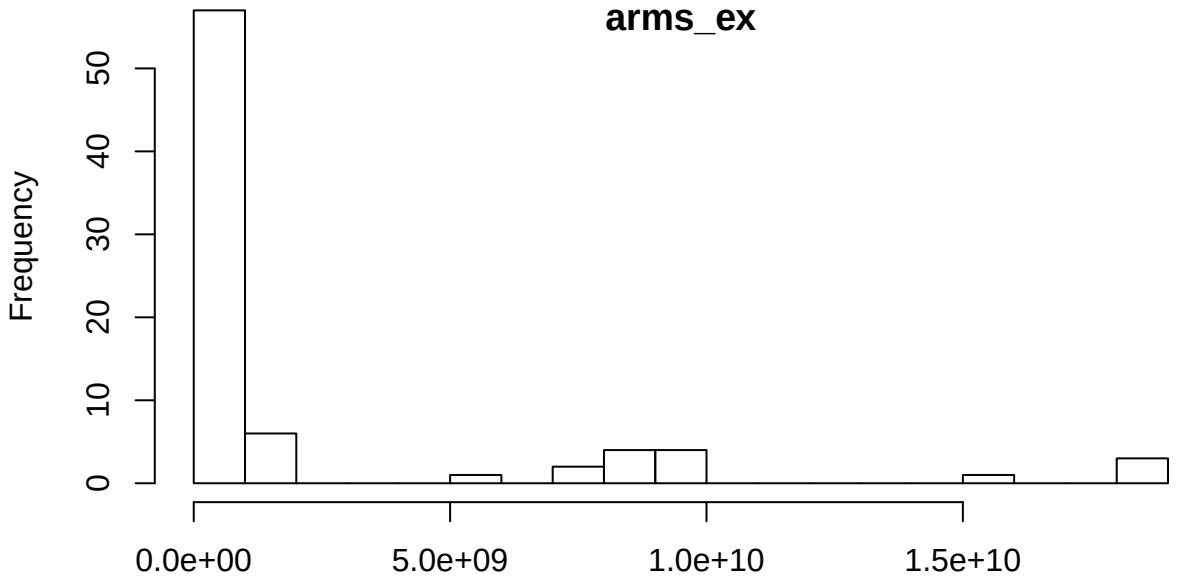
```
for (i in 3:13)
  {
  par(mar=c(5.1,4.1,4.1,1))
  hist(Data[,i],main=" ",xlab=NULL, breaks=20)
  title(names(Data)[i], line = -1)
  }
```

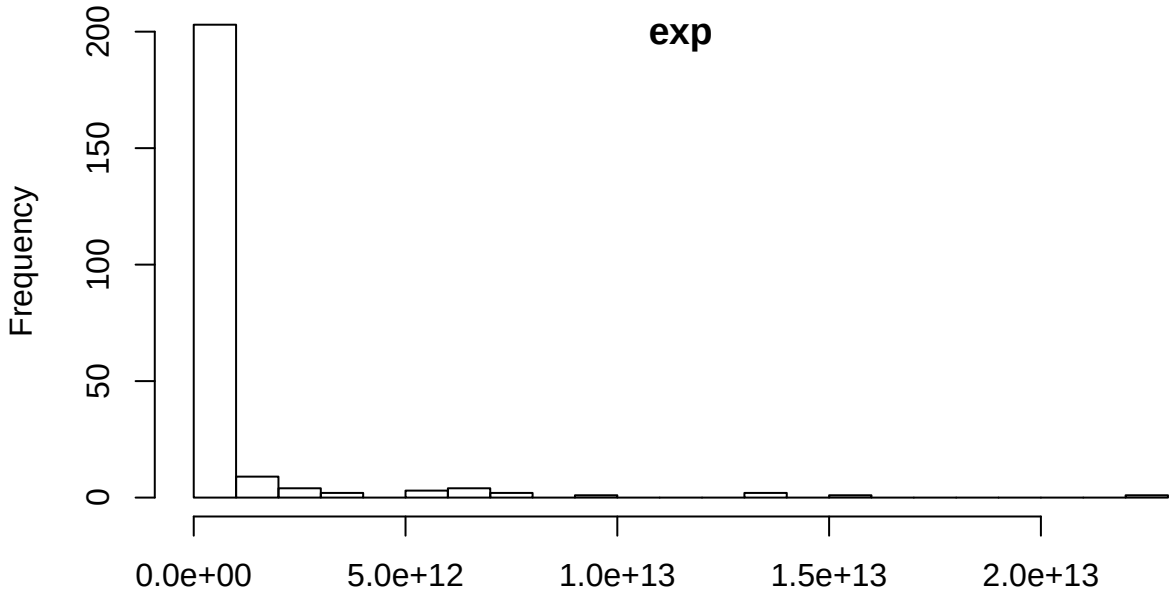


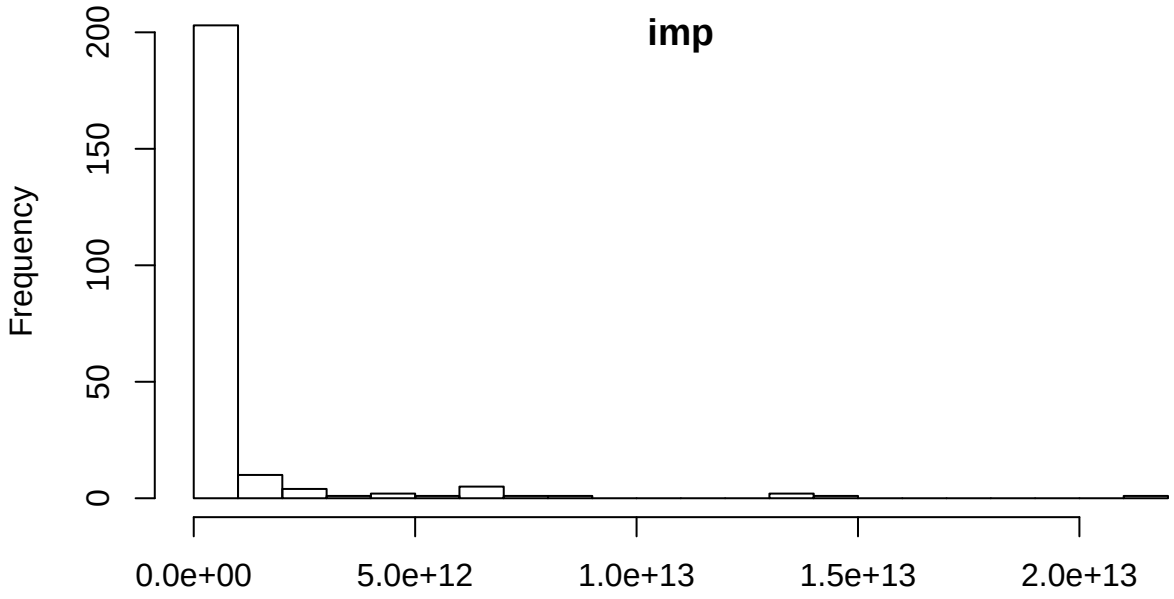

arms_im
Frequency
0
50
100
150
0.0e+00
5.0e+09
1.0e+10
1.5e+10
2.0e+10
2.5e+10
3.0e+10

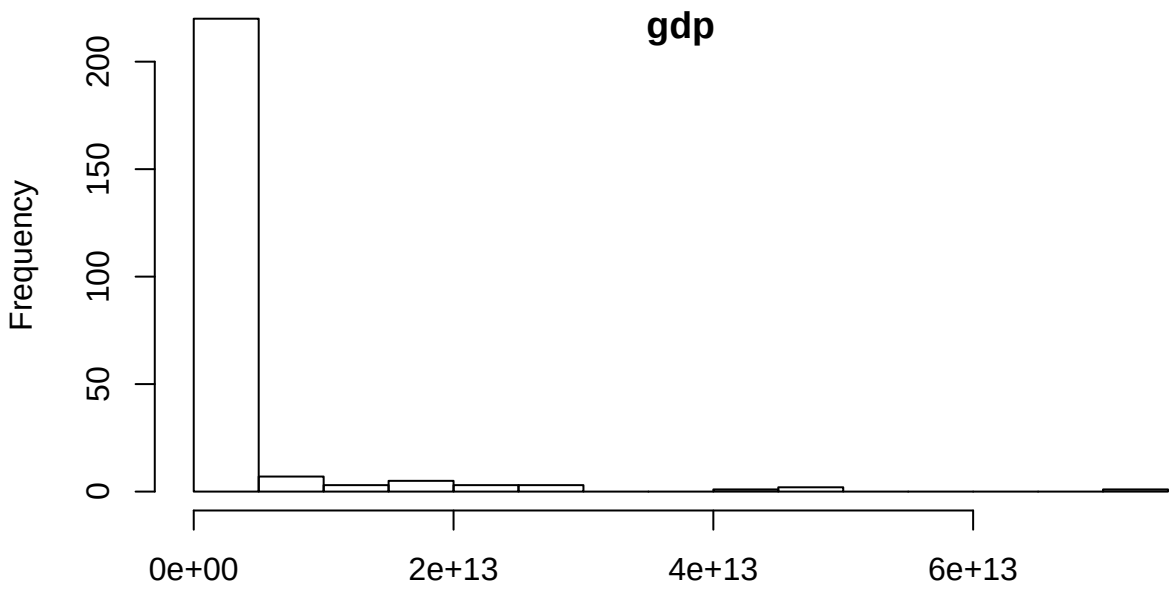

m_exp_gdp
Frequency
0
10
20
30
40
50
0
2
4
6
8
10
12

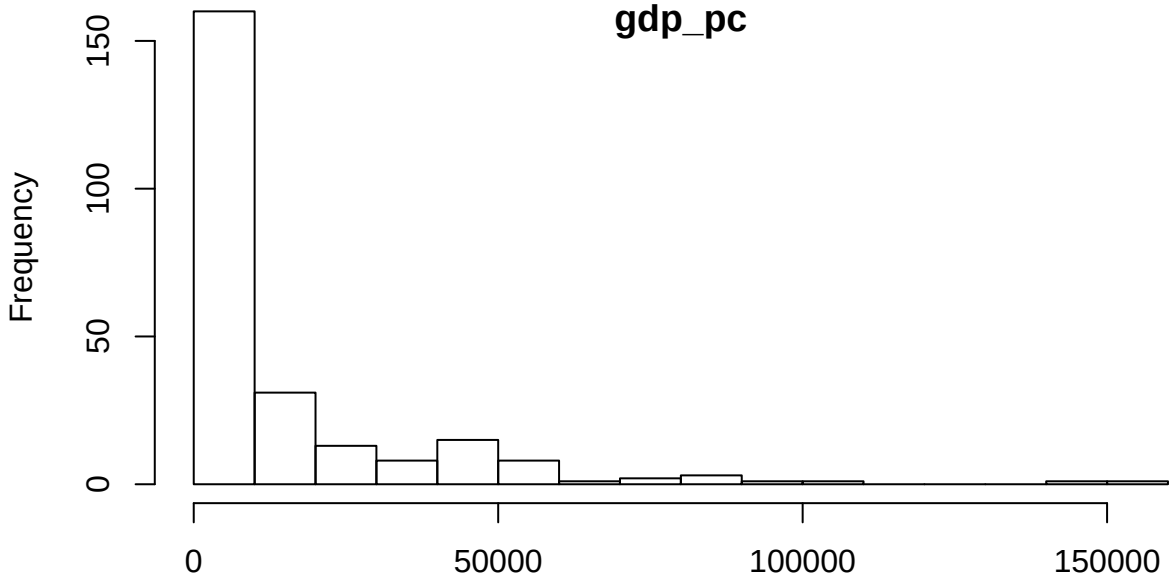

m_exp_gov
Frequency
0
20
40
60
80
100
0
50
100
150

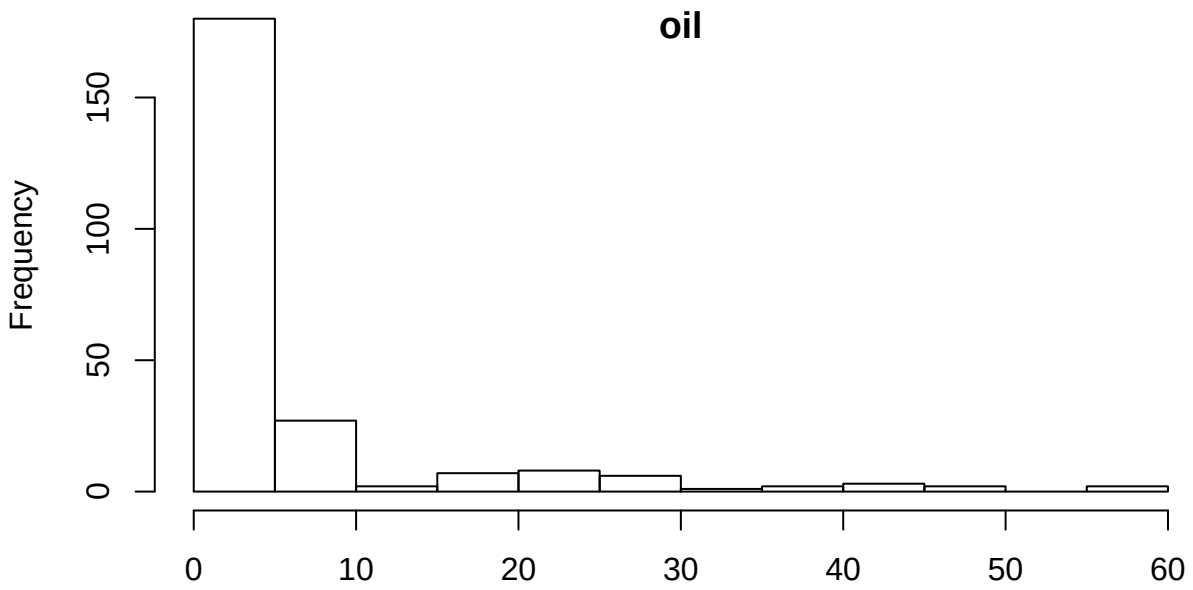

arms_ex
Frequency
0
10
20
30
40
50
0.0e+00
5.0e+09
1.0e+10
1.5e+10

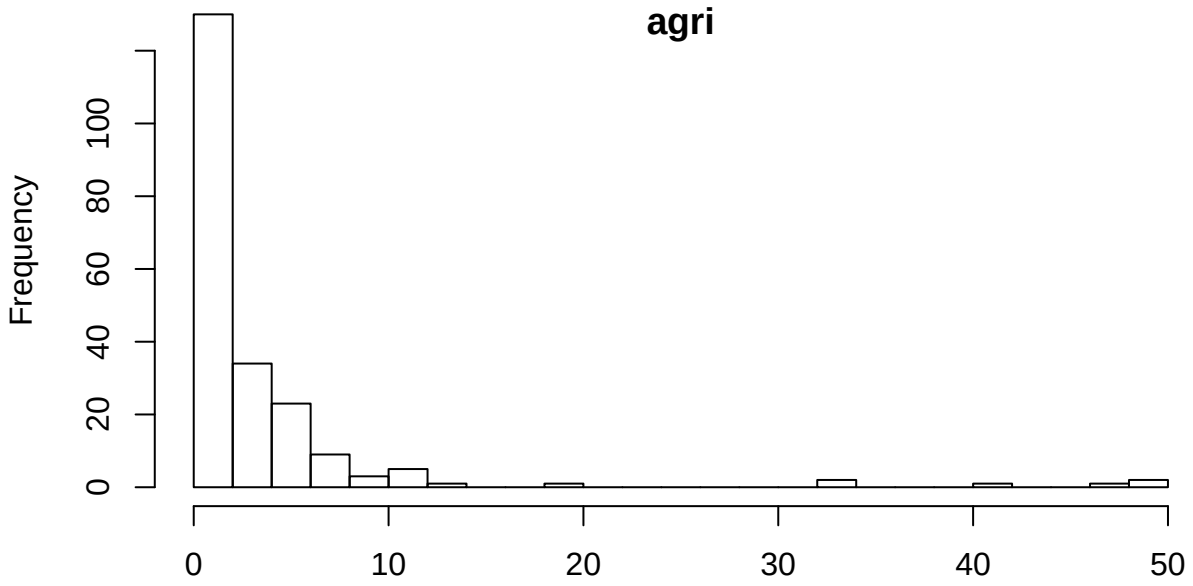

exp
Frequency
0
50
100
150
200
0.0e+00
5.0e+12
1.0e+13
1.5e+13
2.0e+13


imp
Frequency
0
50
100
150
200
0.0e+00
5.0e+12
1.0e+13
1.5e+13
2.0e+13


gdp
Frequency
0
50
100
150
200
0e+00
2e+13
4e+13
6e+13


gdp_pc
Frequency
0
50
100
150
0
50000
100000
150000





Many of the variables have significant positive skews, the dollar value-based variables (`gdp`, `gdp_pc`, `imp`, `exp`) in particular. Given that the dollar value-based variables all have non-zero values log transformations can be used on these variables (notably `agri` is also positively skewed and has non-zero observations, albeit a percentage, so can be log transformed as well).

Proportion of non-null values

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

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

```
##    forest   arms_im m_exp_gdp m_exp_gov   arms_ex       exp       imp
## 0.9696970 0.7651515 0.7765152 0.5151515 0.2954545 0.8787879 0.8787879
##       gdp    gdp_pc       oil      agri
## 0.9280303 0.9280303 0.9090909 0.8030303
```

The `apply()` function is running a `NaN` filter on each column of the data frame, returning `TRUE` if not `NaN` and `FALSE` if `NaN` and then the `mean()` function is performed on each filtered column (i.e. `MARGIN = 2` to calculate the average of each filtered column where `TRUE = 1` and `FALSE = 0`, so the output is the proportion of observations for each variable that are not null values. For example, 96.7% of `forest` observations are not null.

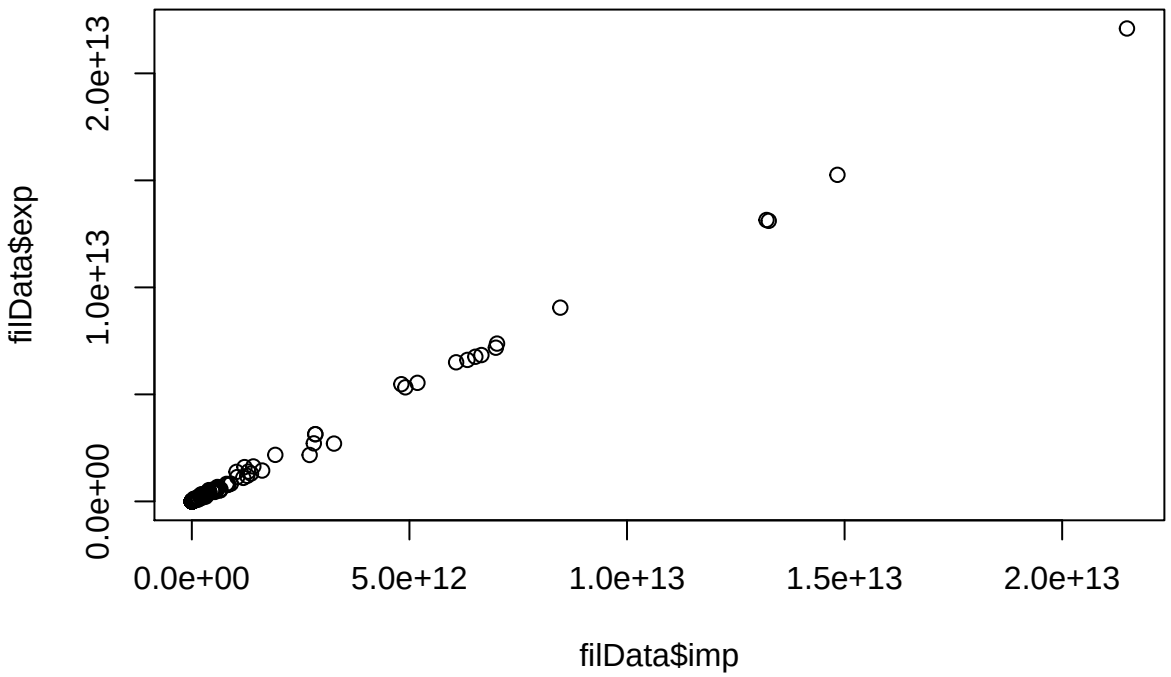
Perfect multicollinearity

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

```r
filter = !is.na(Data$imp) & !is.na(Data$exp)
filData = Data[filter,]
(Cor = cor(filData$imp,filData$exp))
```

```
## [1] 0.9991012
```

```r
plot(filData$imp,filData$exp)
```



No, you can't include both variables in the same OLS model because they have near perfect positive correlation of 0.9991, which would invalidate the assumption that there is no perfect multicollinearity among the explanatory variables.

Development of Hypothesis

The dependent variable is `forest`, which is the percentage of a country's land area that is forest.

There are many factors that could contribute to the percentage of land area of a country that is forest, for example:

a. **Geography** would likely be a key determinant of the natural proportion of forest in a given country because for example countries near: (i) the equator (e.g. Sudan), (ii) the earth's magnetic poles (e.g. Iceland), or (iii) at very high altitude (e.g. Nepal) are either extremely hot or cold, so plant life struggles to grow in these conditions because the landscapes tend to be dominated by deserts and snow.

b. **Population density** would likely be another key determinant because more densely populated countries would require a greater proportion of land to be cleared for housing and agriculture for subsistence.

c. **Economic development** would likely be a another key determinant because greater economic development indicates greater prosperity and hence a greater ability to clear land for agricultural and commercial/industrial development purposes.

These are just a few examples of possible key drivers of forest levels, there are likely others.

Based on these prepositions the variables in the available data set that are more likely to have a cross-sectional relationship with `forest` are the ones that are related to broad economic activity indicators and agricultural activity. While there is no obvious intuitive link to support inclusion of variables related to military activity in a predictive model of `forest`.

Accordingly, variables of interest will come from:

- GDP (current US$) (`gdp`)
- GDP per capita (current US$) (`gdp_pc`)
- Oil rents (% of GDP) (`oil`)
- Agricultural raw materials exports (% of merchandise exports) (`agri`)
- Imports of goods and services (current US$) (`imp`)
- Exports of goods and services (current US$) (`exp`)

Considering the relative intuitive merit of these variables the two variables I include in the model are:

- **GDP (current US$) (`gdp`)**: because it is based on a combination of population size (which to some degree reflects density) and economic development. Because it has significant skew to the right as a result of there being a small number of very large economies globally a log transformation of the variable is necessary. It is expected that sign of the coefficient on this variable should be negative reflecting that more populous and economically developed countries likely have relatively lower area of forest and conversely.
- **Agricultural raw materials exports (% of merchandise exports) (`agri`)**: because as a proxy for the proportion of economic activity dependent on cleared land it may relate to the proportion of the country's overall cleared land. Because it has significant skew to the right as a result of there being a small number of countries with agriculture dominating their exports a log transformation of the variable is used. It is expected that the sign of this variable will be negative reflecting that higher proportions of agricultural economic activity will lead to relatively lower area of forest and conversely.

A model that predicts `forest`

A two variable model

My predictive model of `forest` with two explanatory variables is as follows:

$$forest = \beta_0 + \beta_1 log(gdp) + \beta_2 log(agri) + u$$

```
model2 = lm(forest ~ log(gdp) + log(agri), data = Data)
summary(model2)
```

```
##
## Call:
## lm(formula = forest ~ log(gdp) + log(agri), data = Data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -38.785 -17.490  -0.683  10.371  65.958
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  49.7331    12.3778   4.018 8.27e-05 ***
## log(gdp)     -0.7640     0.4844  -1.577 0.116294
## log(agri)     2.8269     0.7605   3.717 0.000261 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.4 on 203 degrees of freedom
##   (58 observations deleted due to missingness)
## Multiple R-squared:  0.07128,    Adjusted R-squared:  0.06213
## F-statistic:  7.79 on 2 and 203 DF,  p-value: 0.0005502
```

As you can see despite some reasonably sound intuitive development of the hypotheses the model has relatively poor predictive power, an adjusted r-squared of 0.0621 implies that only 6.2% of the variability in `forest` is explained by the model.

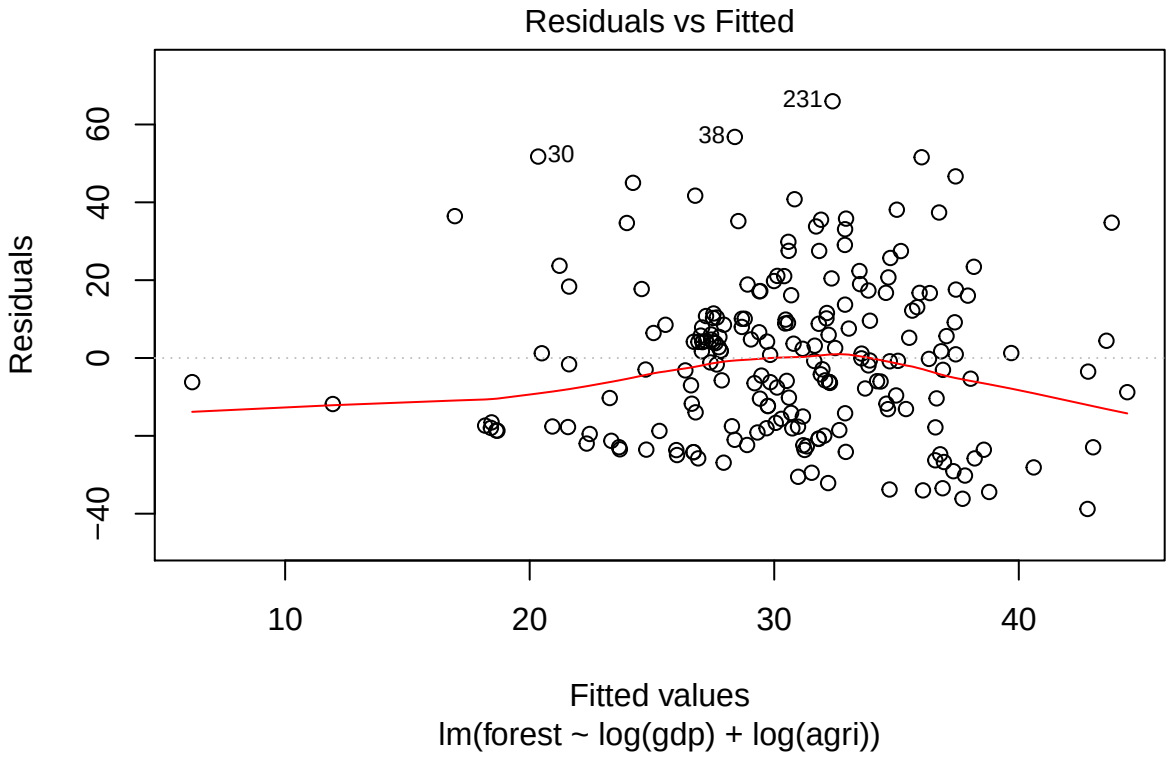
The t-stat and p-value on the `log(gdp)` coefficient indicate that it is not statistically significant, albeit the sign of the coefficient is at least in the expected direction.

While the t-stat and p-value of`agri` indicate it has high statistical significance the sign of `agri` is opposite to expected intuition, instead it is indicating that higher agricultural exports is predictive of more `forest`, which is non-sensical.

Overall this model appears to be a poor predictor of the proportion of a country's land that is forest, which given reasonable supporting intuition I find this a little surprising.

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

```
plot(model2, which=1)
```



The line is not overly horizontal particularly where the observations are more dense, so this suggests that the coefficients are biased because the unobserved variables in `u` are likely correlated with `log(gdp)` and/or `log(agri)`. For example, **geography** is a key variable that is missing from the model.

How many observations are being used in your analysis?

The number of observations being used in my analysis is 206, having lost 58 observations that were missing from at least one of the explanatory variables or the dependent variable.

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

Countries that have no `gdp` data include:

```
filter_gdp = is.na(Data$gdp)
Data$Country.Name[filter_gdp]
```

```
##  [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
```

Notably the majority of these countries appear to be territories of other countries in the sample, e.g. New Caledonia is a territory of France, so it is likely that these country's GDP is contained within the parent country GDP. As such the parent countries explanatory variables will be inflated or upwardly biased relative to the dependent variable, which will bias the coefficient on `log(gdp)`.

Countries that have no `agri` data (and are not in `gdp` because they have no `agri` data for the same reason that they have no `gdp` data) include:

```
filter_agri = is.na(Data$agri) & !filter_gdp
Data$Country.Name[filter_agri]
```

```
##  [1] Andorra
##  [2] Angola
##  [3] Chad
##  [4] Congo, Dem. Rep.
##  [5] Cuba
##  [6] Djibouti
##  [7] Equatorial Guinea
##  [8] Eritrea
##  [9] Faroe Islands
## [10] Fragile and conflict affected situations
## [11] Gabon
## [12] Grenada
## [13] Guinea-Bissau
## [14] Haiti
## [15] Isle of Man
## [16] Kosovo
## [17] Lao PDR
## [18] Least developed countries: UN classification
## [19] Liberia
## [20] Liechtenstein
## [21] Low income
## [22] Marshall Islands
## [23] Micronesia, Fed. Sts.
## [24] Monaco
## [25] Montenegro
## [26] Pre-demographic dividend
## [27] Puerto Rico
## [28] Serbia
## [29] Seychelles
## [30] Somalia
## [31] South Sudan
## [32] Swaziland
## [33] Tajikistan
## [34] Turkmenistan
## [35] Tuvalu
## [36] Uzbekistan
## [37] West Bank and Gaza
## 267 Levels:  Afghanistan Albania Algeria American Samoa Andorra ... Zimbabwe
```

Most of these countries are poor, so probably don't have agricultural exports or just don't have the data available, so their absence from the data set, which doesn't appear to be random, will also likely be causing a bias in the results.

Add a third variable.

As the model has been fairly unsuccessful with the variables that most match the developed hypothesis I now add the military variable, `m_exp_gdp` (i.e. military expenditure as a % of GDP), which has no hypothetical support. So the new three variable predictive model of `forest` is:

$$forest = \beta_0 + \beta_1 log(gdp) + \beta_2 log(agri) + \beta_3 m\_exp\_gdp + u$$

```
model3 = lm(forest ~ log(gdp) + log(agri) + m_exp_gdp, data = Data)
summary(model3)
```

```
##
## Call:
## lm(formula = forest ~ log(gdp) + log(agri) + m_exp_gdp, data = Data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -33.801 -12.033  -0.355  10.065  54.652
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  26.5000    13.2965   1.993 0.047779 *
## log(gdp)      0.3768     0.5125   0.735 0.463155
## log(agri)     1.9416     0.8146   2.383 0.018199 *
## m_exp_gdp    -3.6589     0.9448  -3.873 0.000151 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.12 on 179 degrees of freedom
##   (81 observations deleted due to missingness)
## Multiple R-squared:  0.1537, Adjusted R-squared:  0.1395
## F-statistic: 10.83 on 3 and 179 DF,  p-value: 1.415e-06
```

Notably the coefficient on `m_exp_gdp` from the three variable model is -3.6589.

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.

```
# create a data frame with rows of data where there are no nulls
filter = !is.na(Data$gdp) & !is.na(Data$agri) & !is.na(Data$m_exp_gdp) & !is.na(Data$forest)
filData = Data[filter,]
# perform regression anatomy formula approach
model_m_exp_gdp = lm(m_exp_gdp ~ log(gdp) + log(agri), data = filData)
model_forest = lm(forest ~ model_m_exp_gdp$residuals, data = filData)
summary(model_forest)
```

```
##
## Call:
## lm(formula = forest ~ model_m_exp_gdp$residuals, data = filData)
```

```
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -41.67 -14.73  -0.45  12.27  54.47
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 29.4360     1.3956  21.092  < 2e-16 ***
## model_m_exp_gdp$residuals  -3.6589     0.9844  -3.717 0.000269 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.88 on 181 degrees of freedom
## Multiple R-squared:  0.07092,    Adjusted R-squared:  0.06579
## F-statistic: 13.82 on 1 and 181 DF,  p-value: 0.0002686
```

The coefficient on the `residuals` in the regression of `forest` on the `residuals` from the regression of `m_exp_gdp` on both `log(gdp)` and `log(agri)` is '-3.6589, which is exactly the same as the coefficient from the three variable model.

Compare your two models.

The most striking feature of the three variable model is that the new variable, `m_exp_gdp`, is **highly statistically significant** given an absolute t-statistic close to 4. On the face of it this is completely at odds with the developed hypothesis.

The negative sign on the coefficient indicates that as `military spending` rises as a proportion of GDP that `forest` area declines, which doesn't make any intuitive sense. More on this shortly.

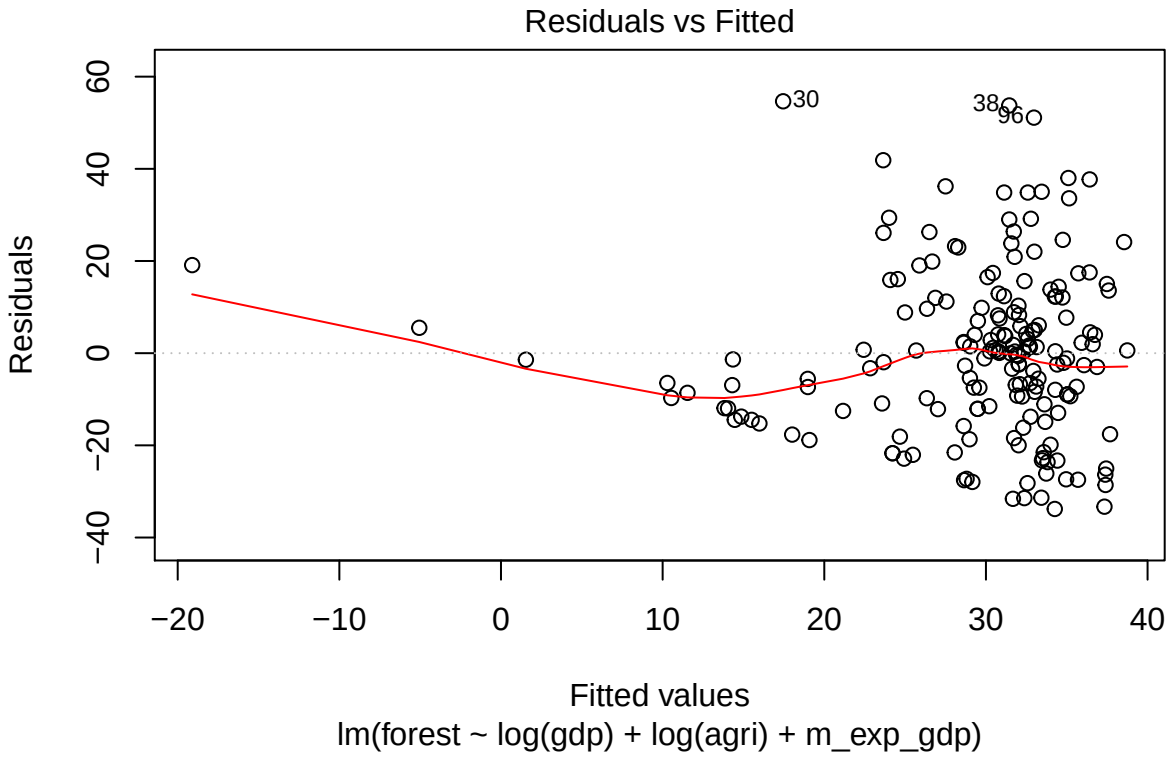
Do you see an improvement? Explain how you can tell.

There is some improvement in the model from adding the third variable.

The third variable has been found to be statistically significant, so you would expect that the proportion of the variability of `forest` explained by the model would have risen, which it has the adjusted r-squared (which adjusts for the fact that the r-squared will always rise when a new variable is added) to 0.1395 from 0.0621.

A residuals versus fitted values plot indicates that the coefficients are no less unbiased under the three variable model compared to the two variable model:

```
plot(model3, which=1)
```



Alternatively, the **Akaike Information Criterion (AIC)** is a measure of the relative quality of statistical models for a given set of data. Given a set of candidate models for the data, the preferred model is the one with the minimum AIC value.

```
(AIC2 = AIC(model2))
```

```
## [1] 1831.946
```

```
(AIC3 = AIC(model3))
```

```
## [1] 1585.572
```

The three variable model has a lower AIC, so based on this measure it also indicates that the three variable model is an improvement on the variable model.

Why is `m_exp_gdp` statistically significant?

As mentioned above, on the face of it there doesn't appear to be any sound reason why the proportion of GDP spent on military activity would explain variability in `forest` and in particular a negative relationship. So why is it the case?

To determine this I investigate the countries with the highest spend on military activities:

```
newData = Data[order(Data$m_exp_gdp, decreasing=T),]
head(newData$Country.Name,20)
```

```
##  [1] Oman                        Saudi Arabia
##  [3] South Sudan                 Arab World
```

```
##  [5] Libya                       Israel
##  [7] United Arab Emirates        Middle East & North Africa
##  [9] Yemen, Rep.                 Jordan
## [11] Algeria                     Azerbaijan
## [13] Lebanon                     Chad
## [15] Angola                      Russian Federation
## [17] United States               Iraq
## [19] Armenia                     Bahrain
## 267 Levels:  Afghanistan Albania Algeria American Samoa Andorra ... Zimbabwe
```

As you can see the countries that spend the most on military activities happen to be countries that have substantial areas of desert and hence low `forest` area, which is a spurious relationship, i.e. `m_exp_gdp` appears to somewhat proxy **geography** from the developed hypothesis despite there being no reason to suggest that there should be a relationship between proprotion of desert and high levels of spending on military activities.

Make up a country named `Mediland` which has every indicator set at the median value observed in the data. How much forest would this country have?

```
# get median inputs for the 3 variables and create an inputs vector
m0 = 1
m1 = log(median(Data$gdp[!is.na(Data$gdp)]))
m2 = log(median(Data$agri[!is.na(Data$agri)]))
m3 = median(Data$m_exp_gdp[!is.na(Data$m_exp_gdp)])
(inputs = c(m0,m1,m2,m3))
```

```
## [1]  1.0000000 24.6864542  0.4750145  1.5351621
```

```
# get the 4 coefficients of the three variable model
(coef3 = model3$coefficients)
```

```
## (Intercept)    log(gdp)   log(agri)   m_exp_gdp
##  26.5000235   0.3768077   1.9416430  -3.6589304
```

```
# calculate the predicted forest level for Mediland
(forest_medi = sum(inputs * coef3))
```

```
## [1] 31.10733
```

The predicted level of forest in Mediland is 31.1%.

Take away

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

The primary takeaway from the analysis is that blind data mining may find statistical significance, but the findings may be spurious without soundly developed hypotheses.