hw11

K Iwasaki

July 23, 2017

Get familiar with the data

```
Definitions
```

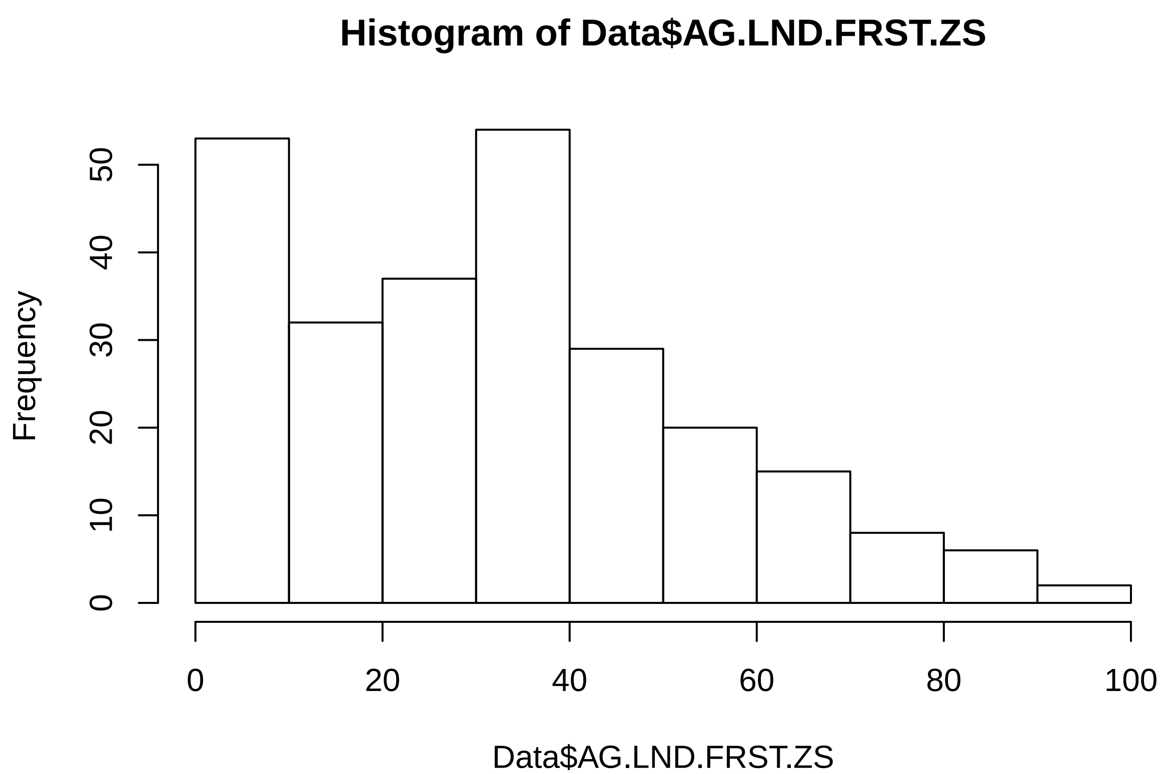
```
##          Series.Code
## 1     AG.LND.FRST.ZS
## 2  MS.MIL.XPND.GD.ZS
## 3     MS.MIL.XPND.ZS
## 4     NY.GDP.MKTP.CD
## 5     NY.GDP.PCAP.CD
## 6  NY.GDP.PETR.RT.ZS
## 7     MS.MIL.XPRT.KD
## 8  TX.VAL.AGRI.ZS.UN
## 9     MS.MIL.MPRT.KD
## 10    NE.IMP.GNFS.CD
## 11    NE.EXP.GNFS.CD
##                                                        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$)
## 11                     Exports of goods and services (current US$)
```

```
head(Data)
```

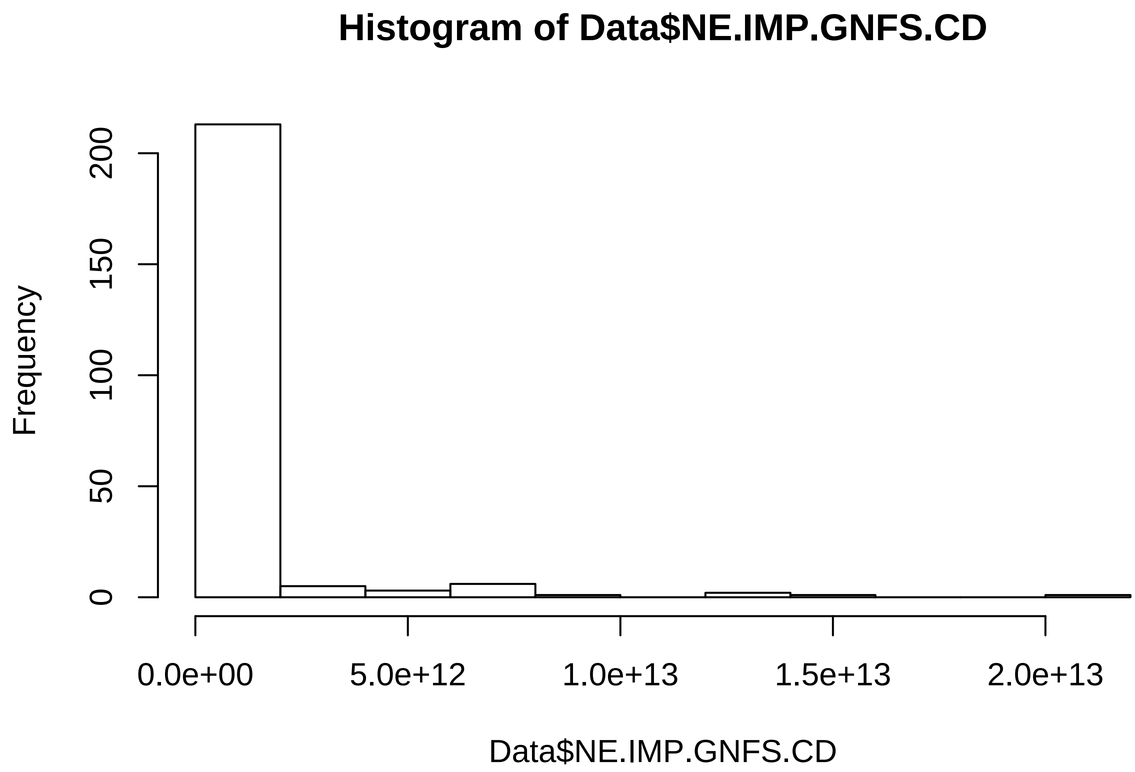
```
##      Country.Name Country.Code AG.LND.FRST.ZS MS.MIL.MPRT.KD
## 1    Afghanistan          AFG       2.067825      359166667
## 2        Albania          ALB      28.244526        9000000
## 3        Algeria          DZA       0.813271      721500000
## 4 American Samoa          ASM      88.133333            NaN
## 5        Andorra          ADO      34.042553            NaN
## 6         Angola          AGO      46.657576       31333333
##   MS.MIL.XPND.GD.ZS MS.MIL.XPND.ZS MS.MIL.XPRT.KD NE.EXP.GNFS.CD
## 1          1.375170       3.183401            NaN     1304521083
## 2          1.413202            NaN              0     3955082222
## 3          4.843526      14.512495            NaN    70304960460
## 4               NaN            NaN            NaN            NaN
## 5               NaN            NaN            NaN            NaN
## 6          4.187594      14.098817            NaN    59957802009
##   NE.IMP.GNFS.CD NY.GDP.MKTP.CD NY.GDP.PCAP.CD NY.GDP.PETR.RT.ZS
## 1     8529983326    18949924158        626.788          0.000000
## 2     6365588048    12442032457       4291.004          4.101974
```

```
## 3    59880526175   193388057520       5114.370          22.388953
## 4            NaN            NaN            NaN                NaN
## 5            NaN     3292207861      40935.583           0.000000
## 6    44133763534   109385918387       4730.046          39.340237
##   TX.VAL.AGRI.ZS.UN
## 1        4.79343482
## 2        2.20095479
## 3        0.01595214
## 4               NaN
## 5               NaN
## 6               NaN
```

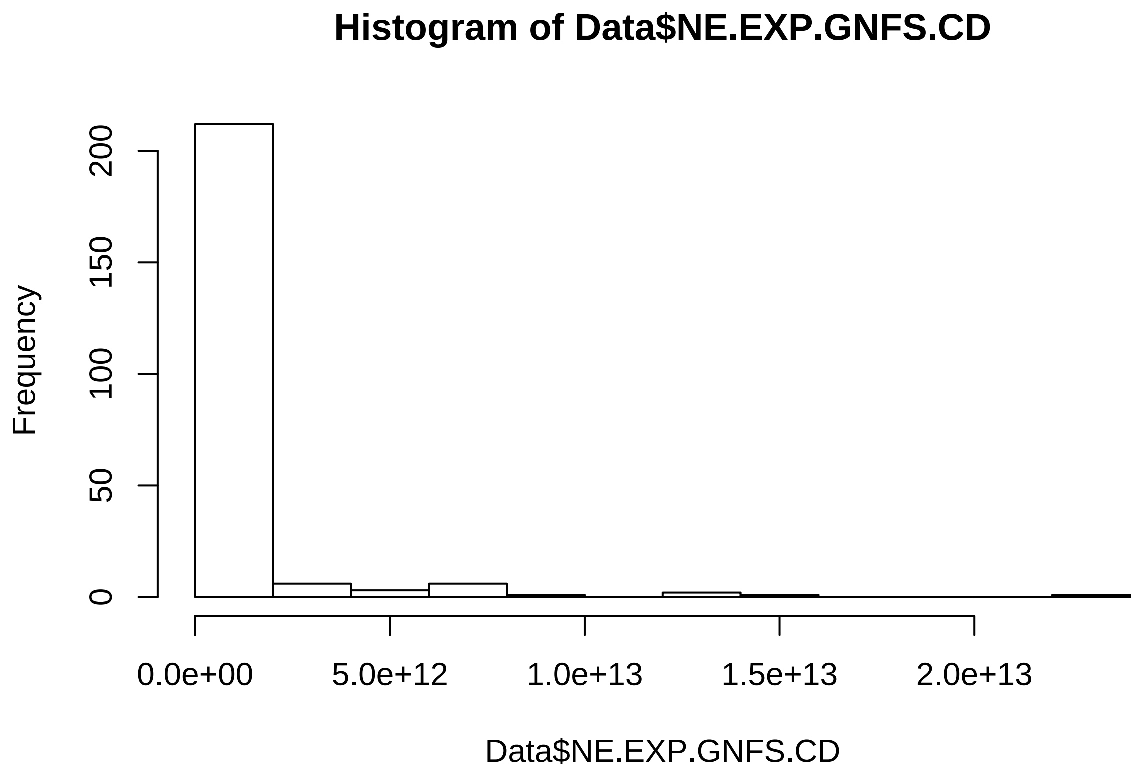
```r
hist(Data$AG.LND.FRST.ZS)

hist(Data$AG.LND.FRST.ZS)
```



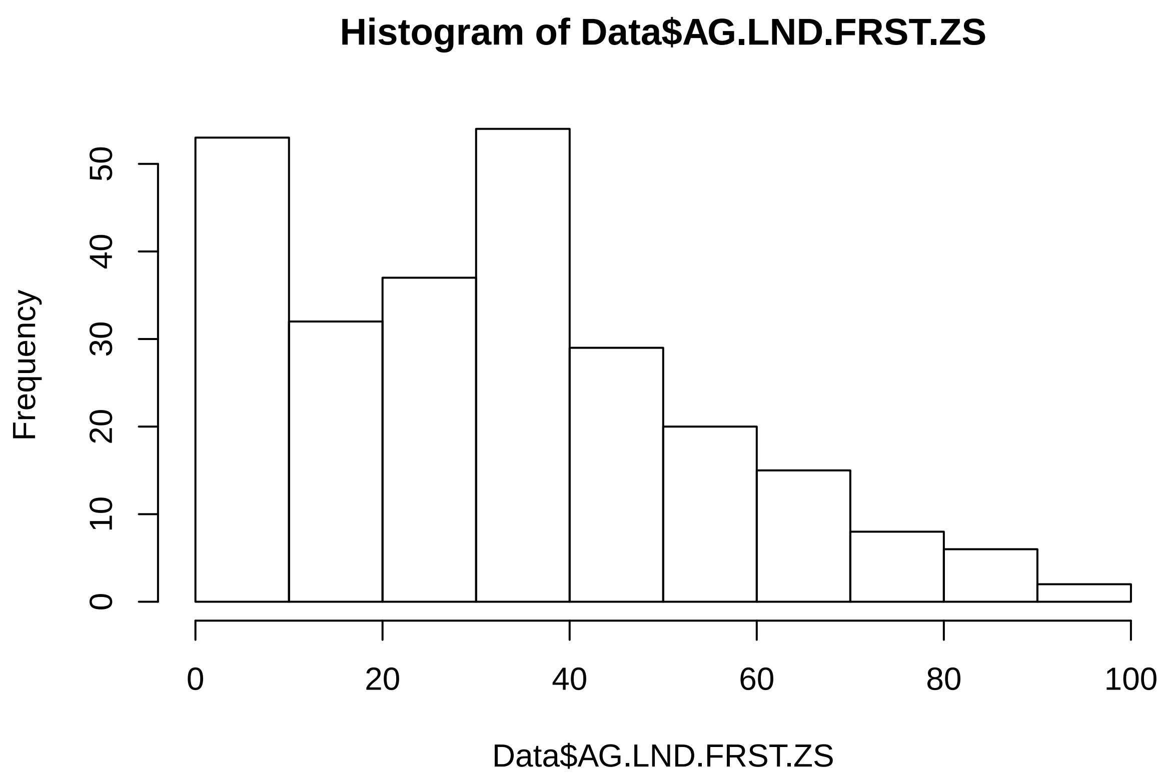

```r
hist(Data$NE.IMP.GNFS.CD) # Imports of goods and services (current US$)
```




```r
hist(Data$NE.EXP.GNFS.CD) # Exports of goods and services (current US$)
```




```r
hist(Data$AG.LND.FRST.ZS) # Forest area (% of land area)
```



Notice there are many NA values in some columns

```r
summary(Data)
```

```
##          Country.Name  Country.Code AG.LND.FRST.ZS  MS.MIL.MPRT.KD
##  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
##  MS.MIL.XPND.GD.ZS MS.MIL.XPND.ZS    MS.MIL.XPRT.KD
##  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
##  NE.EXP.GNFS.CD      NE.IMP.GNFS.CD      NY.GDP.MKTP.CD
##  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
##  NY.GDP.PCAP.CD     NY.GDP.PETR.RT.ZS TX.VAL.AGRI.ZS.UN
##  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
```

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

The line of code computes the percentage of non-NA values in each column. For example, AG.LND.FRST.ZS has 264 - 8 non-NAs out of 264 rows = 0.9696

```
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
```

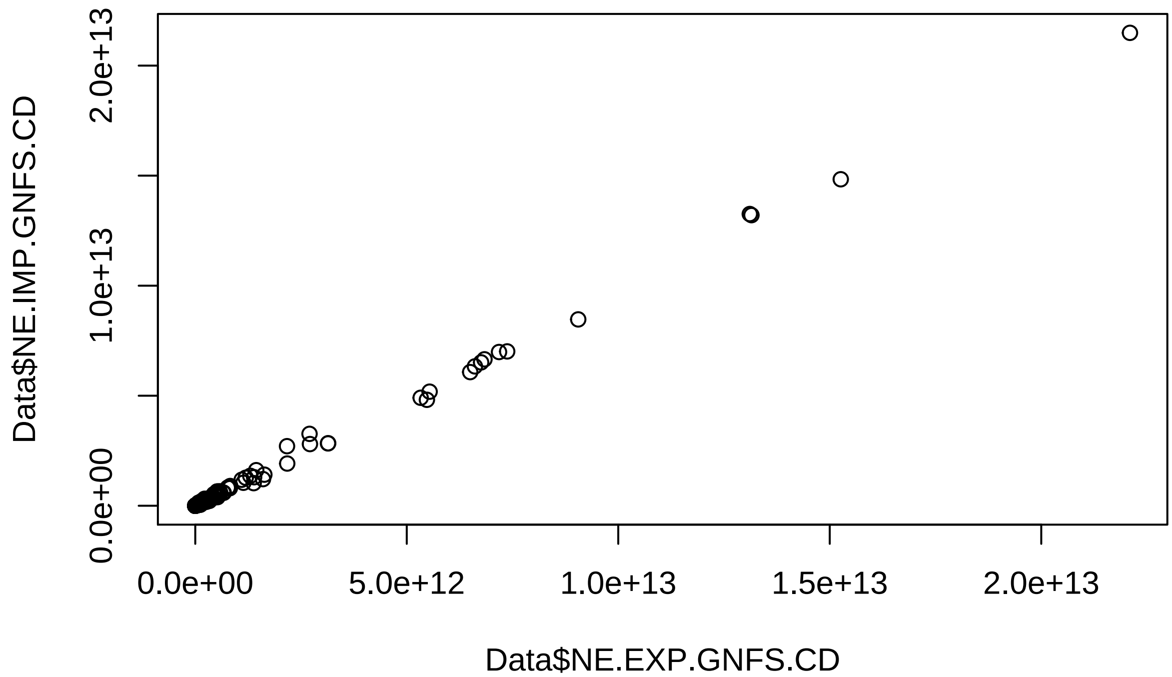
```
nrow(Data) # number of rows 264
```

```
## [1] 264
```

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

No. The plot below shows that these variable have strong linear relationship. This breaks no multicollearity assumption for OLS estimators if they are included in the same model.

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



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

```
colnames(Data)[colnames(Data) == "AG.LND.FRST.ZS"] = "forest"
#colnames(Data)
```

Decribe a model for that predicts forest

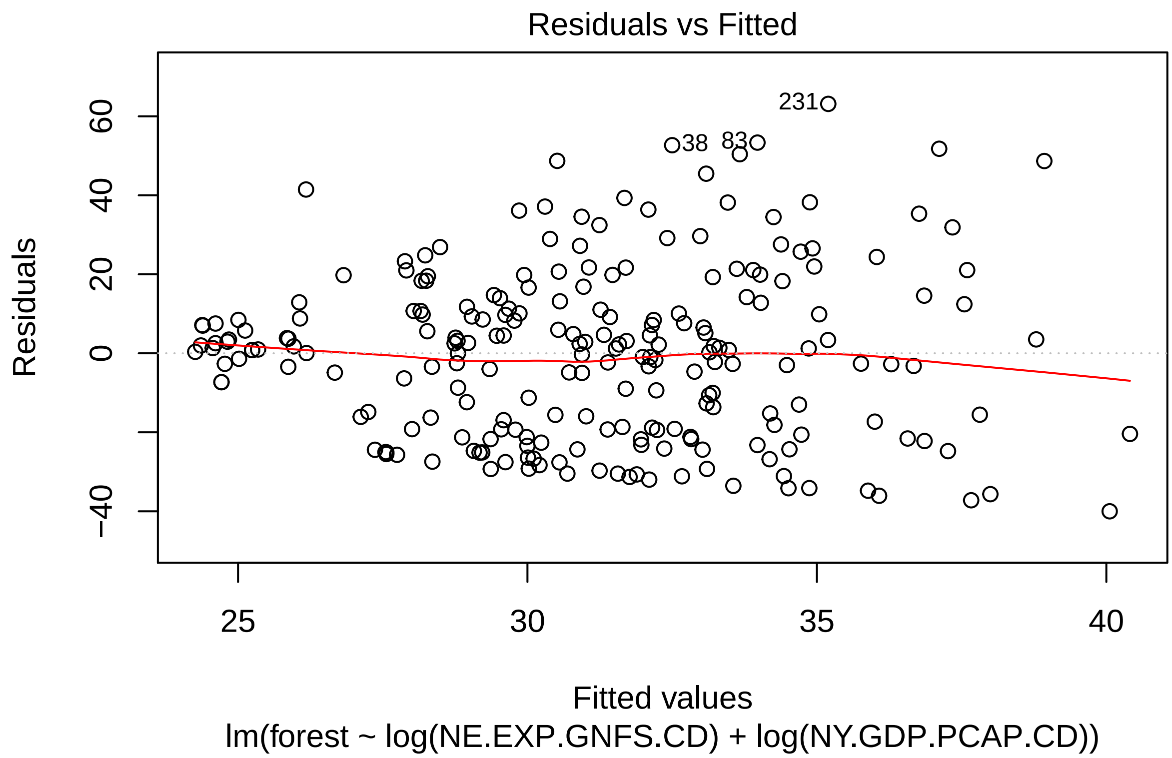
Write a model with two explanatory variables

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

It turns out that the residual plot shows that the mean residual doesn't change with the fitted values. The spread of the residulas are constant. This confirms that two OLS assumptions: 1) error term has a zero conditional mean and 2) Error term has a constant variance (homoskedasticity).

In addition, assume that linearity in model parameters, random sampling and variability, and no perfect collinearity among variables. This OLS estimator is unbiased.

```
model = lm(forest ~ log(NE.EXP.GNFS.CD) + log(NY.GDP.PCAP.CD), Data)
plot(model, which =1)
```



How many observations are being used in your analysis?

```
length(model$residuals)
```

```
## [1] 228
```

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

There are not many duplications for the countries that are dropping out for two variables I selected in the model. Thus, I assume the countries that are dropping out by random chance.

If this is violated, we cannot maintain that the OLS estimator is unbiased.

```
Data$Country.Name[is.na(Data$NE.EXP.GNFS.CD)]
```

```
##  [1] American Samoa              Andorra
##  [3] British Virgin Islands      Cayman Islands
##  [5] Channel Islands             Curacao
##  [7] Djibouti                    French Polynesia
##  [9] Gibraltar                   Greenland
## [11] Guam                        Isle of Man
## [13] Korea, Dem. Peopleâ<U+0080><U+0099>s Rep. Liechtenstein
## [15] Marshall Islands            Micronesia, Fed. Sts.
## [17] Monaco                      Myanmar
## [19] Nauru                       New Caledonia
## [21] Northern Mariana Islands    Not classified
## [23] Papua New Guinea            San Marino
## [25] Sao Tome and Principe       Sint Maarten (Dutch part)
## [27] St. Martin (French part)    Syrian Arab Republic
```

```
## [29] Turks and Caicos Islands    Tuvalu
## [31] Virgin Islands (U.S.)       Yemen, Rep.
## 267 Levels:  Afghanistan Albania Algeria American Samoa Andorra ... Zimbabwe
```

```
Data$Country.Name[is.na(Data$NY.GDP.PCAP.CD)]
```

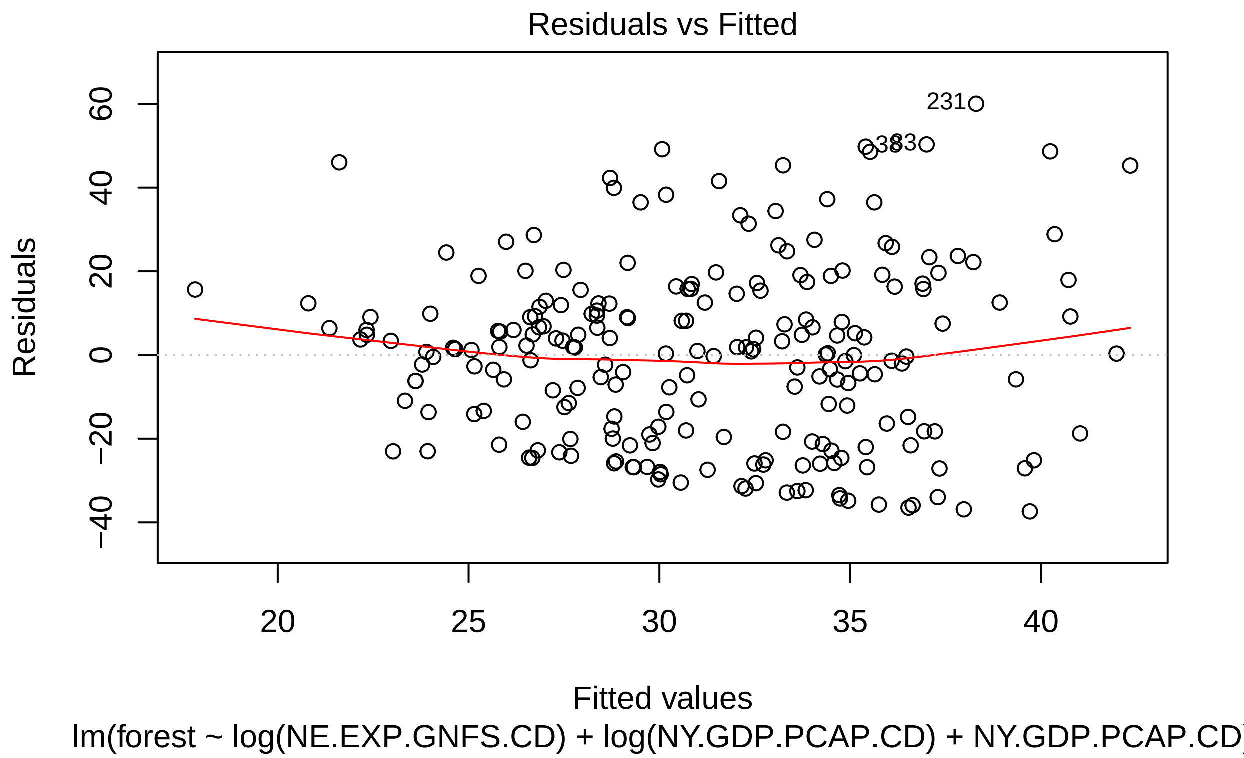
```
##  [1] American Samoa              British Virgin Islands
##  [3] Cayman Islands              Channel Islands
##  [5] Curacao                     French Polynesia
##  [7] Gibraltar                   Guam
##  [9] Korea, Dem. Peopleâ<U+0080><U+0099>s Rep. Nauru
## [11] New Caledonia               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
```

Now add a third variable

```
model2 = lm(forest ~ log(NE.EXP.GNFS.CD) + log(NY.GDP.PCAP.CD) + NY.GDP.PCAP.CD, Data)
model2
```

```
##
## Call:
## lm(formula = forest ~ log(NE.EXP.GNFS.CD) + log(NY.GDP.PCAP.CD) +
##     NY.GDP.PCAP.CD, data = Data)
##
## Coefficients:
##         (Intercept)  log(NE.EXP.GNFS.CD)  log(NY.GDP.PCAP.CD)
##          16.7013095           -1.2270657            5.5757211
##      NY.GDP.PCAP.CD
##          -0.0002964
```

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



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.

```
fs = lm(NY.GDP.PCAP.CD ~ log(NE.EXP.GNFS.CD) + log(NY.GDP.PCAP.CD), Data)
fs
```

```
##
## Call:
## lm(formula = NY.GDP.PCAP.CD ~ log(NE.EXP.GNFS.CD) + log(NY.GDP.PCAP.CD),
##     data = Data)
##
## Coefficients:
##         (Intercept)  log(NE.EXP.GNFS.CD)  log(NY.GDP.PCAP.CD)
##            -74033.1               -337.6              11109.7
```

```
y = Data$forest[!is.na(Data$NY.GDP.PCAP.CD) & !is.na(Data$NE.EXP.GNFS.CD) & !is.na(Data$NY.GDP.PCAP.CD)]

ra = lm(y ~ fs$residuals)
ra
```

```
##
## Call:
## lm(formula = y ~ fs$residuals)
##
## Coefficients:
```

```
## (Intercept) fs$residuals
##  31.0633991   -0.0003029
```

Compare your two models. Do you see an improvement? Explain how you can tell.

Use Akaike information criterion (AIC) for the assessment. Since AIC score decreases from model1 to model2 which is with additional variable. Model1 is better.

```r
AIC(model)
```

```
## [1] 2051.792
```

```r
AIC(model2)
```

```
## [1] 2048.395
```

Make up a country

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

```r
Mediland = apply(Data[,-(1:3)] , MARGIN= 2, mean, na.rm = TRUE)
str(Mediland)
```

```
##  Named num [1:10] 1.30e+09 2.00 8.95 2.27e+09 7.81e+11 ...
##  - attr(*, "names")= chr [1:10] "MS.MIL.MPRT.KD" "MS.MIL.XPND.GD.ZS" "MS.MIL.XPND.ZS" "MS.MIL.XPRT.KI
```

How much forest would this country have?

```r
x = data.frame(NE.EXP.GNFS.CD = Mediland[5], NY.GDP.PCAP.CD = Mediland[7])
predict(model, x)
```

```
## NE.EXP.GNFS.CD
##       74.34863
```

Take away

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

In a causal/structural approach, we believe that if we could just measure all the factors that are out there and put them into a regression equation (in the right way), our parameters will have a causal interpretation.

In this example, the models tells that increase in GDP per capita causes forest to increase and increase in Exports of goods and services causes forest to descrease.