HW week 11

w203: Statistics for Data Science

w203 teaching team

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 `Descriptions` which is a data frame that includes variable names.

```r
getwd()
```

```
## [1] "C:/Users/Melwin/Desktop/Data Science files/UC Berkeley/W203 Stats/W203_Assignments/HW11"
```

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

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

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

- 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.

```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
```

```r
cor(Data[,-(1:2)], use="complete.obs")
```

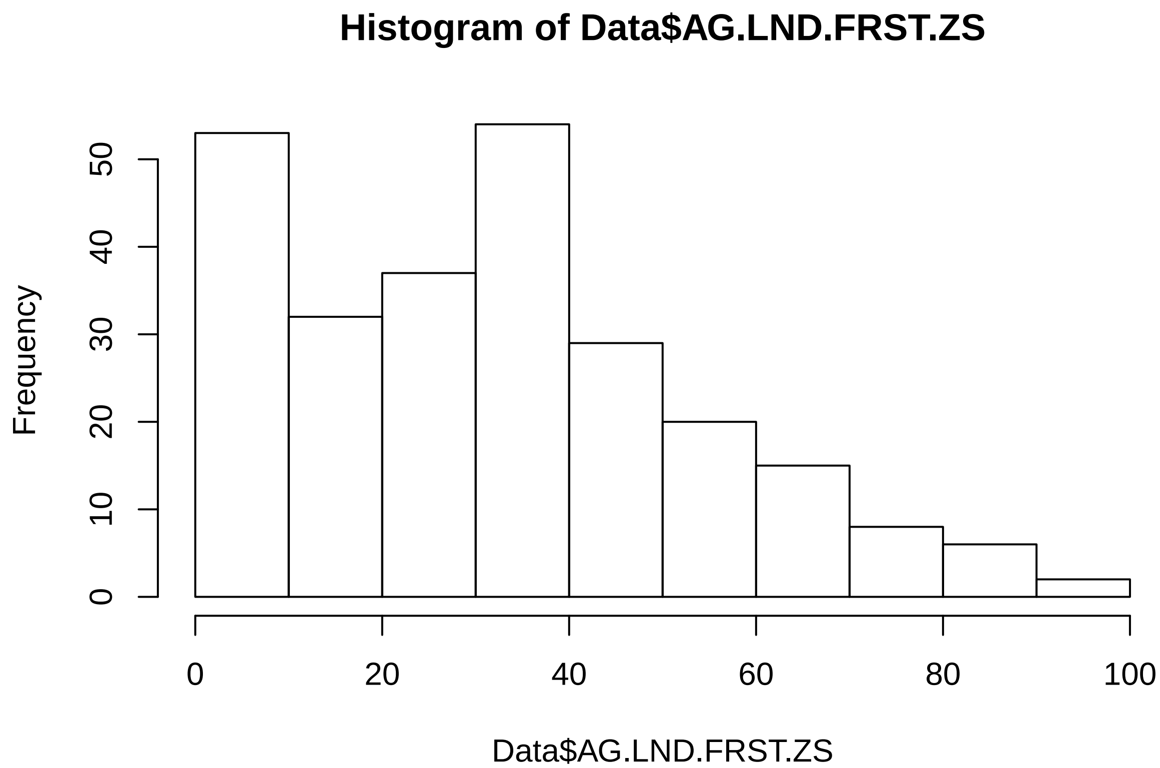
```
##                   AG.LND.FRST.ZS MS.MIL.MPRT.KD MS.MIL.XPND.GD.ZS
## AG.LND.FRST.ZS        1.00000000    -0.03998654       -0.25220161
## MS.MIL.MPRT.KD       -0.03998654     1.00000000        0.19155995
## MS.MIL.XPND.GD.ZS    -0.25220161     0.19155995        1.00000000
## MS.MIL.XPND.ZS       -0.24280966     0.08337472        0.61711211
## MS.MIL.XPRT.KD        0.14881941     0.73559833        0.24571779
## NE.EXP.GNFS.CD        0.08781793     0.82433388        0.08998635
## NE.IMP.GNFS.CD        0.08486420     0.82757634        0.10165348
## NY.GDP.MKTP.CD        0.08539308     0.82040039        0.15307625
## NY.GDP.PCAP.CD        0.11106271    -0.06158964       -0.11782788
## NY.GDP.PETR.RT.ZS    -0.05459529     0.02889363        0.45098282
## TX.VAL.AGRI.ZS.UN     0.38927867    -0.06947298       -0.23266049
##                   MS.MIL.XPND.ZS MS.MIL.XPRT.KD NE.EXP.GNFS.CD
## AG.LND.FRST.ZS       -0.24280966     0.14881941     0.08781793
## MS.MIL.MPRT.KD        0.08337472     0.73559833     0.82433388
## MS.MIL.XPND.GD.ZS     0.61711211     0.24571779     0.08998635
## MS.MIL.XPND.ZS        1.00000000    -0.01281551    -0.03256410
## MS.MIL.XPRT.KD       -0.01281551     1.00000000     0.91161535
## NE.EXP.GNFS.CD       -0.03256410     0.91161535     1.00000000
## NE.IMP.GNFS.CD       -0.03098878     0.91677341     0.99886225
## NY.GDP.MKTP.CD       -0.02014183     0.92999254     0.97489084
## NY.GDP.PCAP.CD        0.01723753     0.10576651     0.14709980
## NY.GDP.PETR.RT.ZS     0.70162419     0.11558163    -0.04885716
## TX.VAL.AGRI.ZS.UN    -0.17232007    -0.06781204    -0.07914586
```

```
##                    NE.IMP.GNFS.CD NY.GDP.MKTP.CD NY.GDP.PCAP.CD
## AG.LND.FRST.ZS         0.08486420     0.08539308    0.111062709
## MS.MIL.MPRT.KD         0.82757634     0.82040039   -0.061589639
## MS.MIL.XPND.GD.ZS      0.10165348     0.15307625   -0.117827876
## MS.MIL.XPND.ZS        -0.03098878    -0.02014183    0.017237530
## MS.MIL.XPRT.KD         0.91677341     0.92999254    0.105766507
## NE.EXP.GNFS.CD         0.99886225     0.97489084    0.147099799
## NE.IMP.GNFS.CD         1.00000000     0.98389962    0.149148299
## NY.GDP.MKTP.CD         0.98389962     1.00000000    0.162137440
## NY.GDP.PCAP.CD         0.14914830     0.16213744    1.000000000
## NY.GDP.PETR.RT.ZS     -0.05525580    -0.05063575   -0.004316487
## TX.VAL.AGRI.ZS.UN     -0.07384466    -0.04944996    0.028044168
##                    NY.GDP.PETR.RT.ZS TX.VAL.AGRI.ZS.UN
## AG.LND.FRST.ZS          -0.054595289        0.38927867
## MS.MIL.MPRT.KD           0.028893630       -0.06947298
## MS.MIL.XPND.GD.ZS        0.450982821       -0.23266049
## MS.MIL.XPND.ZS           0.701624189       -0.17232007
## MS.MIL.XPRT.KD           0.115581635       -0.06781204
## NE.EXP.GNFS.CD          -0.048857161       -0.07914586
## NE.IMP.GNFS.CD          -0.055255804       -0.07384466
## NY.GDP.MKTP.CD          -0.050635754       -0.04944996
## NY.GDP.PCAP.CD          -0.004316487        0.02804417
## NY.GDP.PETR.RT.ZS        1.000000000       -0.08090071
## TX.VAL.AGRI.ZS.UN       -0.080900705        1.00000000
```

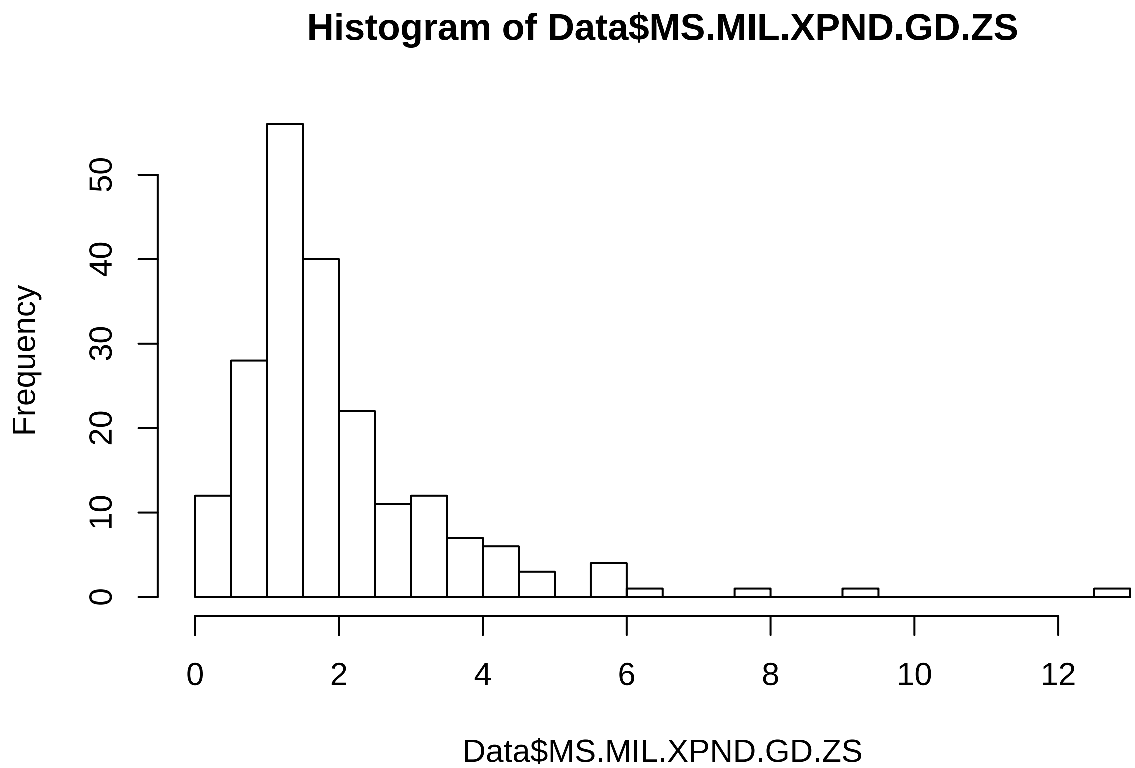
> Here we see that there is a high correlation of AG.LND.FRST.ZS(forest) with TX.VAL.AGRI.ZS.UN, MS.MIL.XPND.GD.ZS , MS.MIL.XPND.ZS , MS.MIL.XPRT.KD and NY.GDP.PCAP.CD. Since we need the independent variables to be not correlated we cannot use MS.MIL.XPND.GD.ZS and MS.MIL.XPND.ZS together, since they have high correlation. Also MS.MIL.XPND.ZS and MS.MIL.XPRT.KD have a lot of na values(from summary). Using these variables reduces our confidence in the model

We examine the some of the variables using histogram

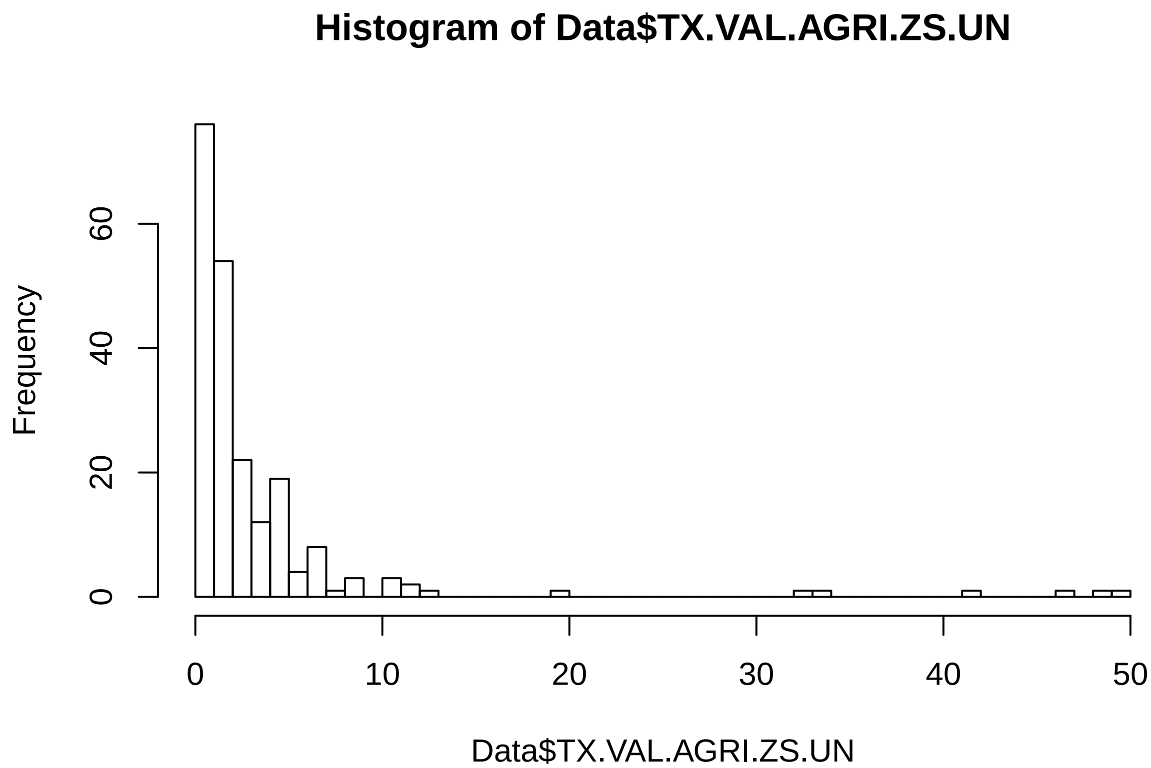
```r
hist(Data$AG.LND.FRST.ZS)
```



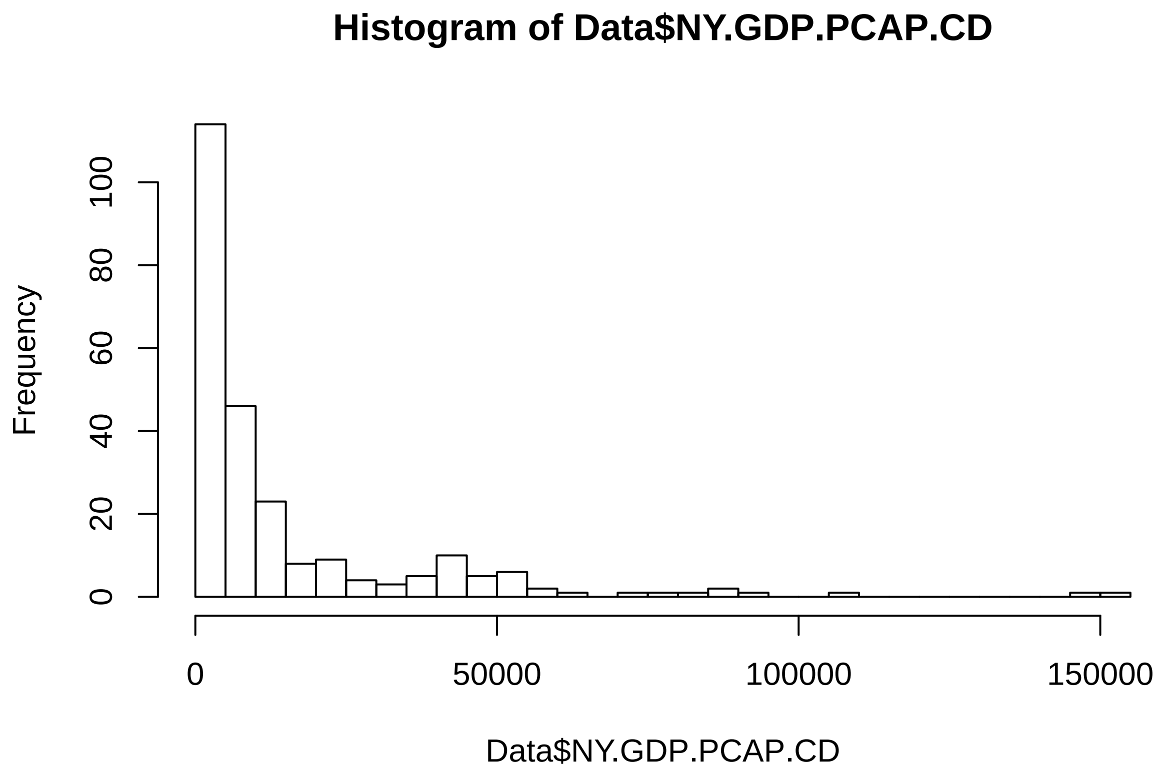

```r
hist(Data$MS.MIL.XPND.GD.ZS, breaks = 35)
```




```
hist(Data$TX.VAL.AGRI.ZS.UN,breaks =50)
```




```
hist(Data$NY.GDP.PCAP.CD,breaks =50)
```



NOTE- %%%% We see there are left skewd, so a log tranformation can help tranform the data to a normal distribution. But there are "0" present in the data. Since log(0) is infinity we will have to either change the 0 values to something else or omit these data points. Both approaches will be determinantal to the performance of our model.

HENCE WE CHOOSE NOT TO APPLY ANY LOG TRANSFORMATION TO THIS DATA

%%%

- Run: `apply(!is.na(Data[,-(1:2)] ) , MARGIN= 2, mean )` and explain what it is showing. >> This above command shows the ratio of the na data to actual data points. This indicates the level of confidence we can have on each of the variable, since we will be omitting all the na values while running our regression. We can arrive at the same answer using column means

```
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
```

```
colMeans(!is.na(Data[,-(1:2)] ))
```

```
##    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
```

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

```
cor(Data$NE.EXP.GNFS.CD,Data$NE.IMP.GNFS.CD,use="complete.obs")
```

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

> Here we see that there is 99% correlation between Data$NE.EXP.GNFS.CD and Data$NE.IMP.GNFS.CD. These will not satisify our no Multicolinearity assumption if used together in our model.

- Rename the variable named `AG.LND.FRST.ZS` to `forest.` This is going to be our dependent variable. >> Here we have renamed the AG.LND.FRST.zs column to forest

```
colnames(Data)[3]="forest"
```

Decribe a model for that predicts `forest`

- Write a model with two explanatory variables.

> Here we have a model with two variable. We choose TX.VAL.AGRI.ZS.UN i.e.total argiculuture export(% of total GDP) and MS.MIL.XPND.GD.ZS military expenditure(% of total expenditure).

```
model1 = lm(forest ~ MS.MIL.XPND.GD.ZS + TX.VAL.AGRI.ZS.UN , data =Data, na.action = na.omit)
model1
```

```
##
## Call:
## lm(formula = forest ~ MS.MIL.XPND.GD.ZS + TX.VAL.AGRI.ZS.UN,
##     data = Data, na.action = na.omit)
##
## Coefficients:
##       (Intercept)  MS.MIL.XPND.GD.ZS  TX.VAL.AGRI.ZS.UN
##          38.17411           -4.42881            0.02945
```

```
summary(model1)$r.squared
```

```
## [1] 0.1243265
```

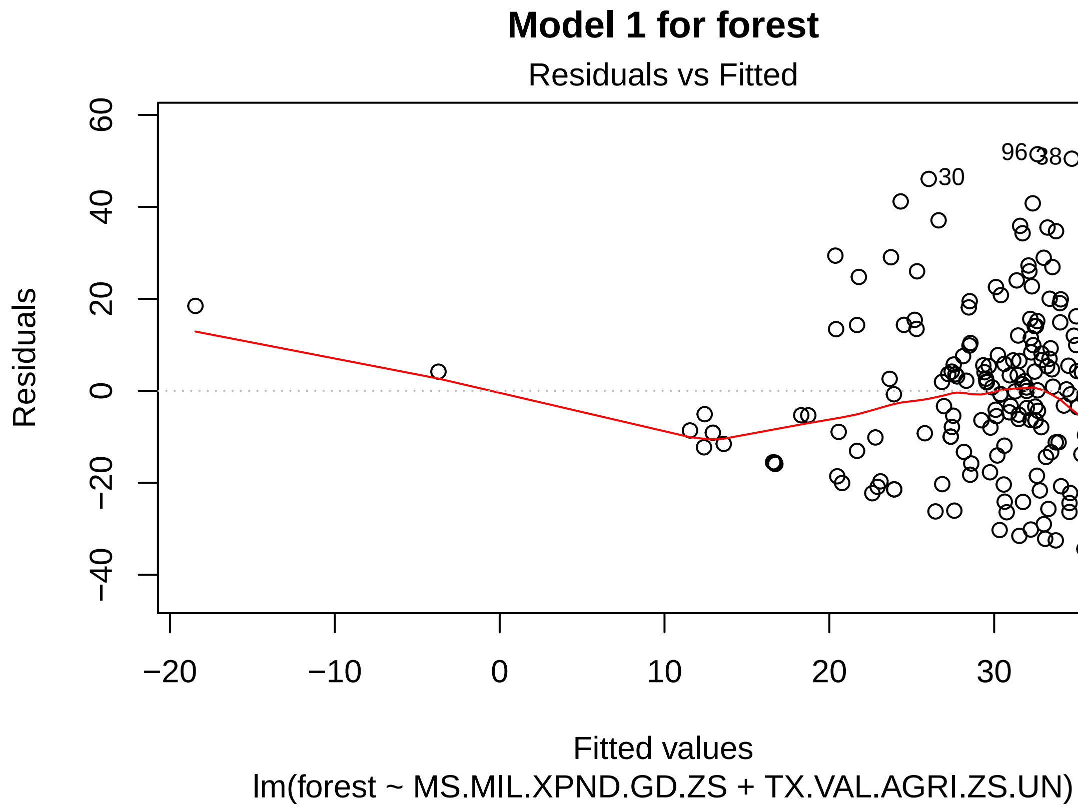
```
AIC(model1)
```

```
## [1] 1589.811
```

> We see that the model can explain 12% of the data.

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

```r
plot(model1, which = 1, main = "Model 1 for forest")
```




```r
summary(model1$residuals)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -37.1400 -13.1600  -0.1219   0.0000  10.1700  51.4500
```

```r
sum(model1$residuals)
```

```
## [1] -6.794565e-14
```

we see that at the extreme left and extreme right side it is off, but there are very few data points present in these areas. Thus we can ignore them for now. Overall the line is almost centered at the center. We can verify this by taking the mean or the sum of the residuals. Thus proving E(u_hat)~=0. Satisfying the 4th assumption.

We are using variables linear in u, thus we satisfy assumption 1 Also we are using data from world bank development indicator data set. They are reliable source and thus we can assume random sampling, satisfying assumption 2 We chose variables which have low co-relation between themselves, thus satisfying assumption 3 Since we satisfy all four assumptions we can say the estimators are unbaised

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

```r
nrow(Data)
```

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

```r
nobs(model1)
```

```
## [1] 183
```

We see there are 183 observations in the plot of 265 available

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

```r
dropped_obs = Data[ is.na(Data$TX.VAL.AGRI.ZS.UN) | is.na(Data$MS.MIL.XPND.GD.ZS) | is.na(Data$forest),]
nrow(dropped_obs)
```

```
## [1] 81
```

```r
head(dropped_obs[is.na(dropped_obs$MS.MIL.XPND.GD.ZS),])
```

```
##           Country.Name Country.Code    forest MS.MIL.MPRT.KD
## 4       American Samoa          ASM 88.133333            NaN
## 5              Andorra          ADO 34.042553            NaN
## 7  Antigua and Barbuda          ATG 22.272727            NaN
## 11               Aruba          ABW  2.333333            NaN
## 15        Bahamas, The          BHS 51.448551       25500000
## 18            Barbados          BRB 14.651163            NaN
##    MS.MIL.XPND.GD.ZS MS.MIL.XPND.ZS MS.MIL.XPRT.KD NE.EXP.GNFS.CD
## 4                NaN            NaN            NaN            NaN
## 5                NaN            NaN            NaN            NaN
## 7                NaN            NaN            NaN      547002901
## 11               NaN            NaN            NaN     1673388268
## 15               NaN            NaN              0     3614441133
## 18               NaN            NaN            NaN     1713416667
##    NE.IMP.GNFS.CD NY.GDP.MKTP.CD NY.GDP.PCAP.CD NY.GDP.PETR.RT.ZS
## 4             NaN            NaN            NaN               NaN
## 5             NaN     3292207861       40935.58         0.0000000
## 7       694336278     1198169901       13377.30         0.0000000
## 11     2040081006     2526083799       24821.46         0.0000000
## 15     4626214301     8310081905       22163.01         0.0000000
## 18     2152833333     4367166667       15488.38         0.6382426
##    TX.VAL.AGRI.ZS.UN
## 4                NaN
## 5                NaN
## 7          1.7765806
## 11         0.0784300
## 15         0.5151958
## 18         0.4154393
```

```r
head(dropped_obs[is.na(dropped_obs$TX.VAL.AGRI.ZS.UN),])
```

```
##              Country.Name Country.Code    forest MS.MIL.MPRT.KD
## 4          American Samoa          ASM 88.133333            NaN
## 5                 Andorra          ADO 34.042553            NaN
## 6                  Angola          AGO 46.657576       31333333
## 29 British Virgin Islands          VGB 24.200000            NaN
## 39         Cayman Islands          CYM 52.916667            NaN
## 42                   Chad          TCD  4.122856       31800000
##    MS.MIL.XPND.GD.ZS MS.MIL.XPND.ZS MS.MIL.XPRT.KD NE.EXP.GNFS.CD
## 4                NaN            NaN            NaN            NaN
## 5                NaN            NaN            NaN            NaN
## 6           4.187594       14.09882            NaN    59957802009
```

```
## 29               NaN            NaN            NaN             NaN
## 39               NaN            NaN            NaN             NaN
## 42          4.250259            NaN            NaN      4293639852
##    NE.IMP.GNFS.CD NY.GDP.MKTP.CD NY.GDP.PCAP.CD NY.GDP.PETR.RT.ZS
## 4             NaN            NaN            NaN               NaN
## 5             NaN     3292207861     40935.5826           0.00000
## 6     44133763534   109385918387      4730.0456          39.34024
## 29            NaN            NaN            NaN               NaN
## 39            NaN            NaN            NaN               NaN
## 42     4994583249    12157171819       940.4099          25.75117
##    TX.VAL.AGRI.ZS.UN
## 4                NaN
## 5                NaN
## 6                NaN
## 29               NaN
## 39               NaN
## 42               NaN
```

```r
head(dropped_obs[is.na(dropped_obs$forest),])
```

```
##             Country.Name Country.Code forest MS.MIL.MPRT.KD
## 54               Curacao          CUW    NaN            NaN
## 101 Hong Kong SAR, China          HKG    NaN            NaN
## 125               Kosovo          KSV    NaN          1e+06
## 145     Macao SAR, China          MAC    NaN            NaN
## 163               Monaco          MCO    NaN            NaN
## 181       Not classified          INX    NaN            NaN
##     MS.MIL.XPND.GD.ZS MS.MIL.XPND.ZS MS.MIL.XPRT.KD NE.EXP.GNFS.CD
## 54                NaN            NaN            NaN            NaN
## 101               NaN            NaN            NaN   590892102811
## 125         0.7143466            NaN            NaN     1257311415
## 145               NaN            NaN            NaN    37805124584
## 163               NaN            NaN            NaN            NaN
## 181               NaN            NaN            NaN            NaN
##     NE.IMP.GNFS.CD NY.GDP.MKTP.CD NY.GDP.PCAP.CD NY.GDP.PETR.RT.ZS
## 54             NaN            NaN            NaN               NaN
## 101   585019101368   269446999890      37567.166                 0
## 125     3468176007     6644439522       3689.738                 0
## 145    14267673956    43518589180      77051.078                 0
## 163            NaN     5712779596     154286.419                 0
## 181            NaN            NaN            NaN               NaN
##     TX.VAL.AGRI.ZS.UN
## 54                NaN
## 101         3.1290150
## 125               NaN
## 145         0.1172368
## 163               NaN
## 181               NaN
```

```r
print(" No of row with TX.VAL.AGRI.ZS.UN and MS.MIL.XPND.GD.ZS as null")
```

```
## [1] " No of row with TX.VAL.AGRI.ZS.UN and MS.MIL.XPND.GD.ZS as null"
```

```r
nrow(dropped_obs[is.na(dropped_obs$TX.VAL.AGRI.ZS.UN) | is.na(dropped_obs$MS.MIL.XPND.GD.ZS),])
```

```
## [1] 81
```

looking at the country names we see that most of the countries are small island countries or union territory of bigger countries like US, China or UK. Most of the countries do not have a decicated military or are incapable of agriculture. Thus we see NA either with MS.MIL.XPND.GD.ZS(Military expenditure) or in TX.VAL.AGRI.ZS.UN(Agricultural export). There are other countries like Syria and Korea that may not be reporting their military expenditure.

- Now add a third variable.'

```
model2 = lm(forest ~ MS.MIL.XPND.GD.ZS + TX.VAL.AGRI.ZS.UN +NY.GDP.PCAP.CD , data =Data, na.action = na
model2
```

```
##
## Call:
## lm(formula = forest ~ MS.MIL.XPND.GD.ZS + TX.VAL.AGRI.ZS.UN +
##     NY.GDP.PCAP.CD, data = Data, na.action = na.omit)
##
## Coefficients:
##       (Intercept)  MS.MIL.XPND.GD.ZS  TX.VAL.AGRI.ZS.UN
##         3.778e+01         -4.421e+00          4.026e-02
##    NY.GDP.PCAP.CD
##         2.315e-05
```

```
summary(model2)$r.squared
```

```
## [1] 0.1248393
```

- 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.

```
third_var = lm(NY.GDP.PCAP.CD ~ MS.MIL.XPND.GD.ZS + TX.VAL.AGRI.ZS.UN , data= model2$model,na.action = r
forest=model2$model$forest
beta_3= cov(forest,third_var$residuals)/var(third_var$residuals)
## value from regression anatomy
print("Values from regression anatomy: ")
```

```
## [1] "Values from regression anatomy: "
```

```
beta_3
```

```
## [1] 2.314945e-05
```

```
## value from model
print("Values from model coefficient: ")
```

```
## [1] "Values from model coefficient: "
```

```
model2$coefficients[4]
```

```
## NY.GDP.PCAP.CD
##   2.314945e-05
```

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

    We will take the AIC as well as use the Stargazer for evalualting which model is better(AIC in my statgazer plot is not working).

```
AIC(model1)
```

```
## [1] 1589.811
```

```r
AIC(model2)
```

```
## [1] 1591.704
```

```r
library(stargazer)
```

```
##
## Please cite as:
```

```
##  Hlavac, Marek (2015). stargazer: Well-Formatted Regression and Summary Statistics Tables.
```

```
##  R package version 5.2. http://CRAN.R-project.org/package=stargazer
```

```r
stargazer(model1, model2, type = "latex", report = "vc",
          header = FALSE,
          title = "Linear Models Predicting College GPA",
          keep.stat = c( "rsq", "n","aic"),
          omit.table.layout = "n")
```

Table 1: Linear Models Predicting College GPA

	Dependent variable:	
	forest	
	(1)	(2)
MS.MIL.XPND.GD.ZS	−4.429	−4.421
TX.VAL.AGRI.ZS.UN	0.029	0.040
NY.GDP.PCAP.CD		0.00002
Constant	38.174	37.784
Observations	183	183
R^2	0.124	0.125

Looking at the AIC we can see that the model1 is better than model2

Make up a country

- 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?

```r
predict(model1,data.frame(MS.MIL.XPND.GD.ZS=mean(Data$MS.MIL.XPND.GD.ZS,na.rm = TRUE),TX.VAL.AGRI.ZS.UN
```

```
##        1
## 29.43429
```

We see that the forest comes to 29.43 which is not equal to the mean of the forest.

Take away

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

We cannot say that it is a causal relationship. Here we are just trying to fit a line that best fits the model. There might be are other variables such a rainfall, weather conditions that are important for forest developments but are not included in this model. These are captured as error .i.e u . Also in the model see some negative fitted values. These is impossible values for forest area thus we indicating that the error in the model is high.