HW week 11

w203: Statistics for Data Science

Ted Pham

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.
- 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.
- Run: `apply(!is.na(Data[,-(1:2)] ) , MARGIN= 2, mean )` and explain what it is showing.
- Can you include both `NE.IMP.GNFS.CD` and `NE.EXP.GNFS.CD` in the same OLS model? Why?
- Rename the variable named `AG.LND.FRST.ZS` to `forest`. This is going to be our dependent variable.

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

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

```r
names(Data)
```

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

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

```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
```

Decribe a model for that predicts `forest`

Include all variables except country name and code in the model. Find the 2 with the lowest p-value are and the fewest # of NA's

Agricultural raw materials exports and GDP (NY.GDP.MKTP.CD)

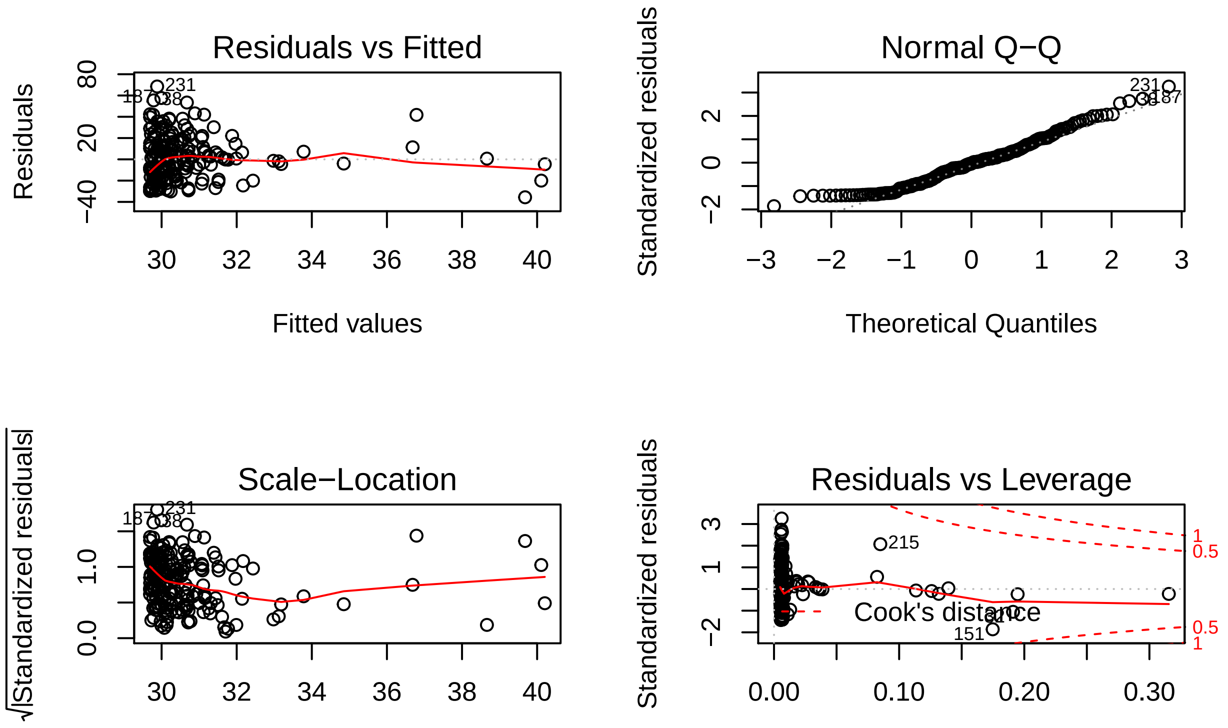
```r
lm.test = lm(AG.LND.FRST.ZS~ . -Country.Name - Country.Code,data = Data)
summary(lm.test)
```

```
##
## Call:
## lm(formula = AG.LND.FRST.ZS ~ . - Country.Name - Country.Code,
##     data = Data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -26.739  -6.407   1.159   6.193  40.749
```

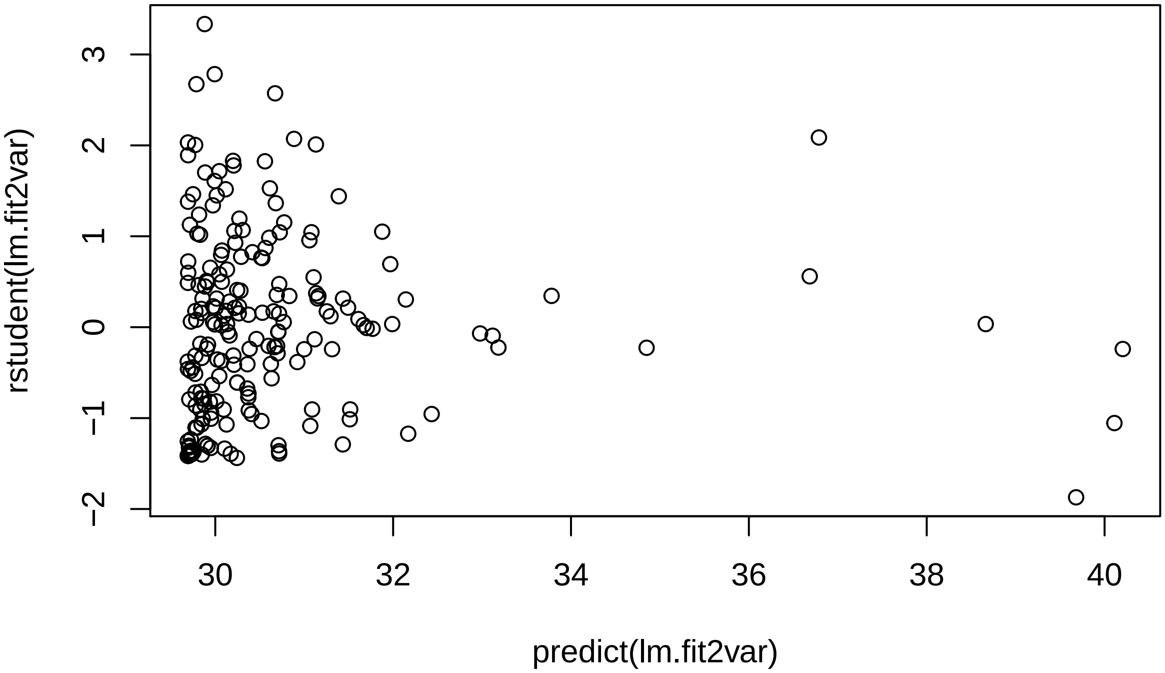
```
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          2.888e+01  7.141e+00   4.044 0.000214 ***
## MS.MIL.MPRT.KD      -1.065e-09  1.837e-09  -0.580 0.565256
## MS.MIL.XPND.GD.ZS -3.318e+00  2.803e+00  -1.184 0.242903
## MS.MIL.XPND.ZS       2.735e-02  2.211e-01   0.124 0.902152
## MS.MIL.XPRT.KD       2.278e-09  1.671e-09   1.363 0.179895
## NE.EXP.GNFS.CD       3.236e-11  4.025e-11   0.804 0.425877
## NE.IMP.GNFS.CD      -4.081e-11  5.094e-11  -0.801 0.427441
## NY.GDP.MKTP.CD       1.872e-12  3.394e-12   0.552 0.584065
## NY.GDP.PCAP.CD       3.022e-05  1.088e-04   0.278 0.782621
## NY.GDP.PETR.RT.ZS -3.486e-01  1.012e+00  -0.344 0.732243
## TX.VAL.AGRI.ZS.UN  2.920e+00  1.143e+00   2.555 0.014231 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.26 on 43 degrees of freedom
##   (210 observations deleted due to missingness)
## Multiple R-squared:  0.2966, Adjusted R-squared:  0.133
## F-statistic: 1.813 on 10 and 43 DF,  p-value: 0.0871
```

- Write a model with two explanatory variables.
  - Create a residuals versus fitted values plot and assess whether your coefficients are unbiased.

```
lm.fit2var = lm(AG.LND.FRST.ZS ~ TX.VAL.AGRI.ZS.UN + NY.GDP.MKTP.CD, data = Data)
par(mfrow=c(2,2))
plot(lm.fit2var)
```



```r
plot(predict(lm.fit2var),rstudent(lm.fit2var))
```



```r
summary(lm.fit2var)
```

```
##
## Call:
## lm(formula = AG.LND.FRST.ZS ~ TX.VAL.AGRI.ZS.UN + NY.GDP.MKTP.CD,
##     data = Data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -35.654 -17.254  -0.691  11.511  68.462
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       2.969e+01  1.724e+00  17.222   <2e-16 ***
## TX.VAL.AGRI.ZS.UN 2.144e-01  2.026e-01   1.058    0.291
## NY.GDP.MKTP.CD    6.552e-14  1.671e-13   0.392    0.695
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21.1 on 203 degrees of freedom
##   (58 observations deleted due to missingness)
## Multiple R-squared:  0.005934,   Adjusted R-squared:  -0.00386
## F-statistic: 0.6059 on 2 and 203 DF,  p-value: 0.5466
```

```r
#Check for unbiasness
#Get the rows used in the fitting
data_use = model.frame(lm.fit2var)
print(sum(data_use$TX.VAL.AGRI.ZS.UN*residuals(lm.fit2var)))
```

```
## [1] 2.600309e-12
```

```
print(sum(data_use$NY.GDP.MKTP.CD*residuals(lm.fit2var)))
```

```
## [1] -0.4927673
```

The sum(X*u)'s are roughly equal to zero so we can assume that the estimators are unbiased. Also by looking at the standardized residuals vs fitted, we can also see that the residuals are randomly distributed around zero, indicative of unbiasness.

- How many observations are being used in your analysis?
- Are the countries that are dropping out dropping out by random chance? If not, what would this do to our inference?

The number of observations used in the regression is 206.

```
length(residuals(lm.fit2var))
```

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

```
#get the list of country ommitted
k = data_use$AG.LND.FRST.ZS
m = data_use$TX.VAL.AGRI.ZS.UN
l = subset(Data,!Data$AG.LND.FRST.ZS %in% k)
data_used = subset(Data, Data$AG.LND.FRST.ZS %in% k & Data$TX.VAL.AGRI.ZS.UN %in% m)
l$Country.Name
```

```
##  [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] French Polynesia
## [17] Gabon
## [18] Grenada
## [19] Guam
## [20] Guinea-Bissau
## [21] Haiti
## [22] Hong Kong SAR, China
## [23] Isle of Man
## [24] Korea, Dem. People's Rep.
## [25] Kosovo
## [26] Lao PDR
## [27] Least developed countries: UN classification
## [28] Liberia
## [29] Liechtenstein
## [30] Low income
## [31] Macao SAR, China
## [32] Marshall Islands
## [33] Micronesia, Fed. Sts.
## [34] Monaco
## [35] Montenegro
## [36] New Caledonia
## [37] Northern Mariana Islands
## [38] Not classified
## [39] Pre-demographic dividend
## [40] Puerto Rico
## [41] Serbia
## [42] Seychelles
## [43] Sint Maarten (Dutch part)
## [44] Somalia
## [45] South Sudan
## [46] St. Martin (French part)
## [47] Swaziland
## [48] Syrian Arab Republic
## [49] Tajikistan
## [50] Turkmenistan
## [51] Turks and Caicos Islands
## [52] Tuvalu
## [53] Uzbekistan
## [54] Virgin Islands (U.S.)
## [55] West Bank and Gaza
## 267 Levels:  Afghanistan Albania Algeria American Samoa Andorra ... Zimbabwe
```

The list of countries ommitted resulted from the NA's from the three columns extracted from Data for the analysis. These appears to be the territories of large countries or small countries without accurate reported GDP. The number of countries omitted was minimized so it would affect minimally on our model.

```
  summary(Data$MS.MIL.XPND.ZS)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.000   4.074   6.746   8.947  10.470 144.900     128
```

- Now add a third variable.
- Show how you would use the regression anatomy formula to compute the coefficient on your third variable. First, regress the third

```
model3var1= lm(data_used$MS.MIL.XPND.GD.ZS~lm.fit2var$residuals)
summary(model3var1)
```

```
##
## Call:
## lm(formula = data_used$MS.MIL.XPND.GD.ZS ~ lm.fit2var$residuals)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6356 -0.7866 -0.2268  0.4895 10.0082
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           1.962902   0.107949  18.184  < 2e-16 ***
## lm.fit2var$residuals -0.027494   0.005541  -4.962  1.6e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 1.457 on 181 degrees of freedom
##   (23 observations deleted due to missingness)
## Multiple R-squared:  0.1197, Adjusted R-squared:  0.1149
## F-statistic: 24.62 on 1 and 181 DF,  p-value: 1.603e-06
```

variable on your first two variables and extract the residuals. Next, regress forest on the residuals from the first stage.

```
lm.fit3var = lm(AG.LND.FRST.ZS ~ TX.VAL.AGRI.ZS.UN + NY.GDP.MKTP.CD + MS.MIL.XPND.GD.ZS, data = Data)
summary(lm.fit3var)
```

```
##
## Call:
## lm(formula = AG.LND.FRST.ZS ~ TX.VAL.AGRI.ZS.UN + NY.GDP.MKTP.CD +
##     MS.MIL.XPND.GD.ZS, data = Data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -36.70 -12.67  -1.59  10.40  51.86
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        3.773e+01  2.430e+00  15.526  < 2e-16 ***
## TX.VAL.AGRI.ZS.UN  4.538e-02  1.873e-01   0.242    0.809
## NY.GDP.MKTP.CD     1.404e-13  1.469e-13   0.956    0.340
## MS.MIL.XPND.GD.ZS -4.466e+00  8.858e-01  -5.042 1.12e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.38 on 179 degrees of freedom
##   (81 observations deleted due to missingness)
## Multiple R-squared:  0.1288, Adjusted R-squared:  0.1142
## F-statistic: 8.819 on 3 and 179 DF,  p-value: 1.746e-05
```

- Compare your two models.

```
AIC(lm.fit3var) > AIC(model3var1)
```

```
## [1] TRUE
```

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

Did not see an improvement because the AIC of the 2nd model is larger than the one obtained from regress the residuals. ### 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?

Mediland would have 30.95% forest area

```
mediland = data.frame("TX.VAL.AGRI.ZS.UN"=1.60804,"NY.GDP.MKTP.CD" = 5.262e+10,"MS.MIL.XPND.GD.ZS" =1.53
predict(lm.fit3var,newdata = mediland, interval = "confidence")
```

```
##        fit      lwr      upr
## 1 30.95471 27.87832 34.03111
```

Take away

- Agricultural raw materials exports (% of merchandise exports) is representative of % forest area.
- Many NA's values affect the strength of multiregression and not all variables are relevant to the models in common sense.
- Must be wary about the fitness of the data for linear regression