Problem Set 1

Sébastien Annan-Phan, Alejandro Favela, Matthew Tadruno

March 14, 2018

Questions 1 and 2

Reading and cleaning the data.

```r
wb_data <-read_csv(paste0(dir_data, "ps1_raw.csv"), col_types=cols())

#Assigning column names
colnames(wb_data) <-
  c("year", "year_code", "country",
    "country_code", "CO2", "GDP", "POP")

#dropping year code
wb_data <- select(wb_data, -year_code)

#drop ".." observations
vars<-names(wb_data)[4:6]
for (var in vars) {
  wb_data <- wb_data[!wb_data[var] == "..", ]
}

wb_data <- na.omit(wb_data)
wb_data[4:6] <-lapply(wb_data[4:6], as.numeric)
```

Question 3

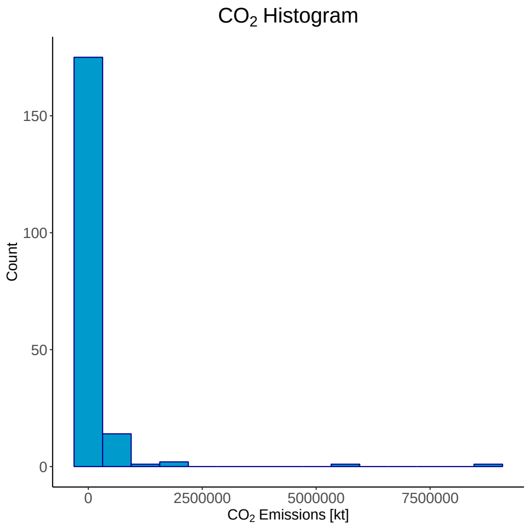
Calculate a table of showing the sample mean, standard deviation, minimum and maximum for each series.

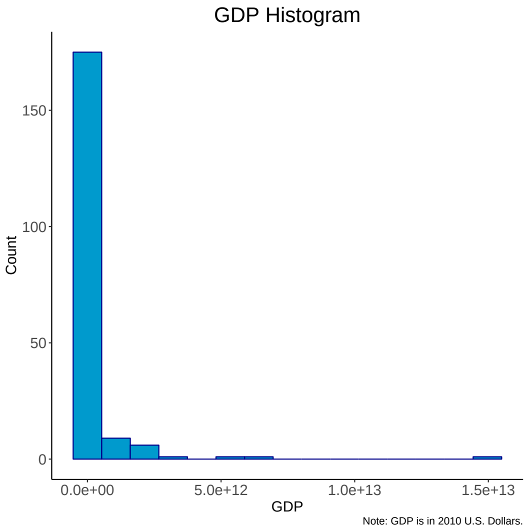
Table 1: Summary Statistics

	CO_2 (kt)	GDP (2010 USD)	Population
Mean	162316.561	3.363660e+11	35220247
S.D.	763012.353	1.297364e+12	134642965
Min	7.334	3.182352e+07	10025
Max	8776040.416	1.496437e+13	1337705000

Question 4

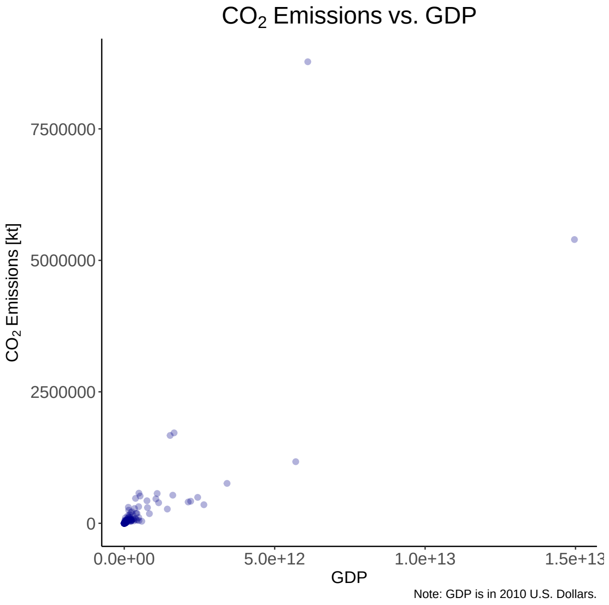
Create a histogram for CO_2 and GDP (15 buckets).





Question 5

Plot CO_2 against GDP.



Question 6

Create a new variable “Per capita CO_2 emissions” called *CO2pc*.

```r
wb_data$CO2pc<-wb_data$CO2/wb_data$POP
```

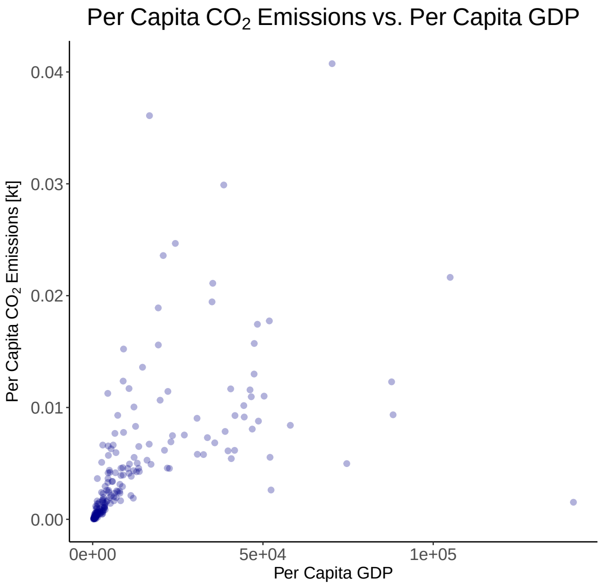
Question 7

Create a new variable “Per capita GDP” called *GDPpc*.

```r
wb_data$GDPpc<-wb_data$GDP/wb_data$POP
```

Question 8

Plot *CO2pc* against *GDPpc*.



Per Capita CO_2 Emissions vs. Per Capita GDP

Note: Per Capita GDP is in 2010 U.S. Dollars.

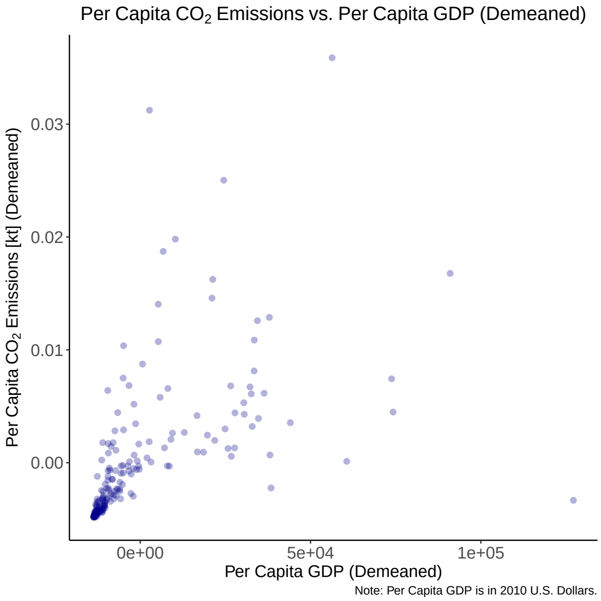
Question 9

Create demeaned variables of *CO2pc* and *GDPpc* called *CO2pcdev* and *GDPpcdev* by subtracting the sample mean from each observation.

```
wb_data$CO2pcdev<-wb_data$CO2pc-mean(wb_data$CO2pc)
wb_data$GDPpcdev<-wb_data$GDPpc-mean(wb_data$GDPpc)
```

Question 10

Plot *CO2pcdev* against *GDPpcdev*.



Per Capita CO_2 Emissions vs. Per Capita GDP (Demeaned)

Note: Per Capita GDP is in 2010 U.S. Dollars.

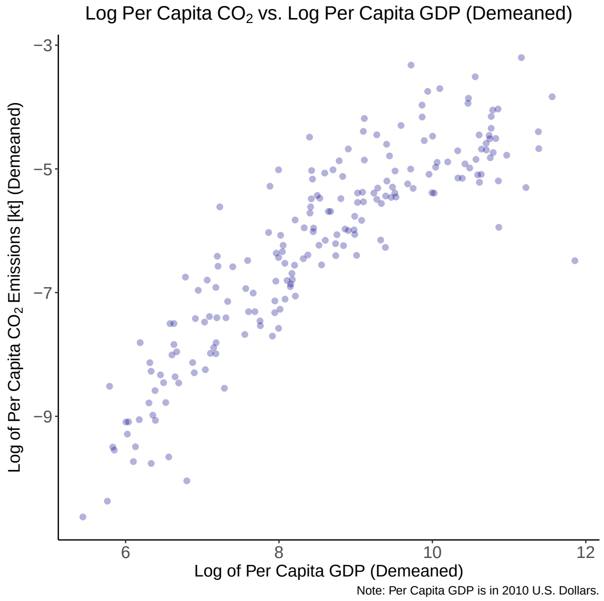
Question 11

Create the variables *CO2pcln* and *GDPpcln* by taking natural logs of *CO2pc* and *GDPpc*.

```r
wb_data$CO2pcln<-log(wb_data$CO2pc)
wb_data$GDPpcln<-log(wb_data$GDPpc)
```

Question 12

Plot *CO2pcln* and *GDPpcln*.


Log Per Capita CO_2 vs. Log Per Capita GDP (Demeaned)

Note: Per Capita GDP is in 2010 U.S. Dollars.

Question 13

Export your data as a comma delimited ascii file.

```r
write.csv(wb_data, file = "wb_formatted.csv")
#default ascii
```

Custom OLS Function

```
b_ols <- function(data, y, X, intercept=NULL) {
  # This function takes as inputs:
  # 1. A data frame, 'data'
  # 2. A dependent variable, y
  # 3. A list of X variables c("x1", "x2", ...)
  # 4. An optional 4th argument, T, to include an intercept
  require(dplyr)

  # Select y variable data from 'data'
  y_data <- subset(data, select=c(y))
  # Select X variable data from 'data'
  X_data <- select_(data, .dots = X)

  if(is.null(intercept)) {
    # Convert y_data to matrices
    y_data <- as.matrix(y_data)
    # Convert X_data to matrices
    X_data <- as.matrix(X_data)
    # Calculate beta hat
    beta_hat <- solve(t(X_data)%*%X_data)%*%t(X_data)%*%y_data
  }

  else{
    # Convert y_data to matrices
    y_data <- as.matrix(y_data)
    # Add a column of ones to X_data
    X_data <- mutate_(X_data, "ones" = 1)
    # Move the intercept column to the front (this is cool)
    X_data <- select_(X_data, "ones", .dots = X)
    # Convert X_data to matrices
    X_data <- as.matrix(X_data)
    # Calculate beta hat
    beta_hat <- solve(t(X_data)%*%X_data)%*%t(X_data)%*%y_data
    # Change the name of 'ones' to 'intercept'
    rownames(beta_hat) <- c("intercept", X)
  }

  y_hat<-X_data%*%beta_hat #Predicted values
  residual<-y_data - y_hat

  SST<-sum((y_data)^2) #Total sum of squares
  SSM<-sum((y_hat)^2) #Regression sum of squares
  SSR<-sum((y_data-y_hat)^2) #Error sum of squares

  SSM_demean <- sum((y_hat - mean(y_data))^2)
  SST_demean <- sum((y_data - mean(y_data))^2)
  n <- dim(X_data)[1]
  k <- dim(X_data)[2]
  dof <- n - k
  R_uc <- 1 - (SSR/SST)
  R <- 1- SSR/SST_demean
  R_adj <- 1 - (1 - R) * ((n-1)/(n-k))
  AIC <- log(SSR/n) + (2*k)/n
  SIC <- log(SSR/n) + (k/n) * log(n)
  s2 <- SSR/(n-k)

  assign("y_hat", y_hat, .GlobalEnv)
  assign("residual", residual, .GlobalEnv)
  assign("beta_hat", beta_hat, .GlobalEnv)
  assign("R", R, .GlobalEnv)
  assign("R_uc", R_uc, .GlobalEnv)
  assign("R_adj", R_adj, .GlobalEnv)
  assign("AIC", R_uc, .GlobalEnv)
  assign("SIC", SIC, .GlobalEnv)
  assign("s2", s2, .GlobalEnv)
  assign("n", n, .GlobalEnv)
  assign("dof", dof, .GlobalEnv)


  fit<-data.frame(n, dof, round(R, 4), round(R_uc, 4),
       round(R_adj, 4), round(AIC, 4), round(SIC, 4), round(s2, 4))
  assign("fit", fit, .GlobalEnv)


  return(c(beta_hat, R, R_uc, R_adj, AIC, SIC, s2))
}
```

Question 14

Regress *CO2pc* on *GDPpc* without an intercept...

```
#Regress CO2pc on GDPpc without an intercept
b_ols(data = wb_data, y="CO2pc", X=c("GDPpc"))[1]

## [1] 2.233062e-07

# Regress CO2pc*1000 on GDPpc without an intercept
wb_data$CO2pc<-wb_data$CO2pc*1000
b_ols(data = wb_data, y="CO2pc", X=c("GDPpc"))[1]

## [1] 0.0002233062

#Regress CO2pc*1000 on GDPpc/1000 without an intercept
wb_data$GDPpc<-wb_data$GDPpc/1000
b_ols(data = wb_data, y="CO2pc", X=c("GDPpc"))[1]

## [1] 0.2233062
```

Multiplying the dependent variable by 1000 increases the estimate of beta by a factor of 1000; dividing the independent variable by 1000 also increases the estimate of beta by a factor of 1000. Together, these give an estimate 1,000,000 times larger than the original regression. While both the sum of squares and the sum of squared residuals change, they do so proportionally, so the R^2 is unchanged after rescaling.

Question 15

```
b_ols(data = wb_data, y="CO2pc", X=c("GDPpc"))[1]

## [1] 0.2233062
```

Table 2: Measures of Fit

n	D.O.F.	R^2	R^2_{uc}	$\bar{R}^2$	AIC	SIC	s^2
194	193	0.1896	0.4905	0.1896	3.4898	3.5066	32.6108

On the graph there are outliers, which means that for certain observations the predicted values are far from the realized values. The fact that there are more outliers on the right-hand side of the graph (higher values of *GDPpc*) suggests that the homoskedasticity assumption does not hold. Additionally, there may be a non-linear relationship between per capita CO_2 emissions and GDP per capita.
The countries with high residuals include Trinidad and Tobago, and Qatar; low-residual countries include Liechtenstein and Switzerland.

Question 16

```
b_ols(data = wb_data, y="CO2pc", X=c("GDPpc"), T)[1:2]

## [1] 0.1668361 2.5314095

#coefficient on GDPpc and intercept, respectively.
```

Table 3: Measures of Fit

n	D.O.F.	R^2	R^2_{uc}	$\bar{R}^2$	AIC	SIC	s^2
194	192	0.2998	0.5598	0.2961	3.354	3.3876	28.3238

Graphing the residuals shows that including intercept improves fit; it reduces the residuals of the outliers at the tails (low and high *GDPpc*). This is reflected in a higher R^2 value.

Question 17

```r
b_ols(data = wb_data, y="CO2pc", X=c("GDPpc", "GDPpc2"), T)[1:3]

## [1]  0.40118282 -0.00280706  1.01533420

#coefficients on GDPpc and GDPpc2dev, and intercept, respectively.
```

Table 4: Measures of Fit

n	D.O.F.	R^2	R^2_{uc}	$\bar{R}^2$	AIC	SIC	s^2
194	191	0.4509	0.6548	0.4451	3.1212	3.1717	22.3273

Overall, the average observation has residuals that are much closer to zero. In particular, the fit has improved on the left tail of the distribution beacuse of the non-linear relationship between CO_2 and GDP. On the right tail, however, we have an outlier for which we overcorrected in the previous model. The Environmental Kuznet's curve provides economic rationale for including *GDPpc2* in our model: as countries start to develop they increase pollution until a certain threshold, after which pollution begins to decline as concerns of environmental quality outweigh marginal gains from pollution.

Question 18

```r
wb_data$CO2pcdev<-wb_data$CO2pc-mean(wb_data$CO2pc)
wb_data$GDPpcdev<-wb_data$GDPpc-mean(wb_data$GDPpc)
wb_data$GDPpc2dev<-wb_data$GDPpc2-mean(wb_data$GDPpc2)
b_ols(data = wb_data, y="CO2pcdev", X=c("GDPpcdev", "GDPpc2dev"))[1:2]

## [1]  0.40118282 -0.00280706

#coefficients on GDPpc and GDPpc2, respectively.
```

Using FWT, we can produce the same estimates for the coefficients on *GDPpc* and *GDPpc2* as in question 17. We do not estimate the intercept with this method. This illustrates the fact that running a regression without an intercept using demeaned variables is equivalent to running one with an intercept and unaltered variables.

Question 19

```
#no intercept, generate residuals
b_ols(data = wb_data, y="CO2pc", X=c("GDPpc"))

## [1]  0.2233062  0.1895649  0.4904832  0.1895649  3.4897836  3.5066283
## [7] 32.6107610

wb_data$residual1<-residual #save residuals
wb_data$i<-1
b_ols(data = wb_data, y="i", X=c("GDPpc")) #no intercept

## [1]  0.02230778        -Inf  0.31166872        -Inf -0.36317577 -0.34633114
## [7]  0.69189776

wb_data$residual2<-residual
b_ols(data = wb_data, y="GDPpc2", X=c("GDPpc")) #no intercept

## [1] 7.143648e+01 7.432497e-01 7.674316e-01 7.432497e-01 1.379389e+01
## [6] 1.381073e+01 9.735901e+05

wb_data$residual3<-residual

#no intercept
b_ols(data = wb_data, y="residual1", X=c("residual2", "residual3"))[1:2]

## [1]  1.01533420 -0.00280706

#intercept and coefficient on GDPpcdev2, respectively.
```

Here we replicate the estimate of the intercept and the coefficient *GDPpc2* from question 17, but ignore *GDPpc*.