Assignment 4

Giang Vu

3/24/2021

Assignment 4

1.

Data was imported below.

```
#set working dir.
setwd("/Users/giangvu/Desktop/STAT 2301 - Statistical Computing and Data Science/HW/HW4")
#import dataset
debt <- read.csv("debt.csv",header = T,as.is = T,sep = ",")
dim(debt)
```

```
## [1] 1171    4
```

```
head(debt)
```

```
##     Country Year    growth     ratio
## 1 Australia 1946 -3.557951 190.41908
## 2 Australia 1947  2.459475 177.32137
## 3 Australia 1948  6.437534 148.92981
## 4 Australia 1949  6.611994 125.82870
## 5 Australia 1950  6.920201 109.80940
## 6 Australia 1951  4.272612  87.09448
```

a.

The function mean.growth() is defined as follows.

```
#define mean.growth()
mean.growth <- function(df){
  return(signif(mean(df$growth),3))
}
```

b.

After applying the function above to data split by countries, we obtained the results of average GDP growth rates as follows.

```r
#use daply()
avgGDP.country <- daply(debt, .(Country), mean.growth)
#check results
avgGDP.country["Australia"]
```

```
## Australia
##      3.72
```

```r
avgGDP.country["Netherlands"]
```

```
## Netherlands
##        3.03
```

```r
#present full result
avgGDP.country
```

```
##   Australia     Austria     Belgium      Canada     Denmark     Finland
##        3.72        4.44        3.18        3.65        2.66        3.57
##      France     Germany      Greece     Ireland       Italy       Japan
##        3.78        3.31        2.93        3.93        3.25        4.45
## Netherlands New Zealand      Norway    Portugal       Spain      Sweden
##        3.03        3.07        3.83        4.00        3.20        3.07
##          UK          US
##        2.41        3.00
```

2.

Below is the average GDP growth rate for each year (averaging over countries), and a corresponding plot.

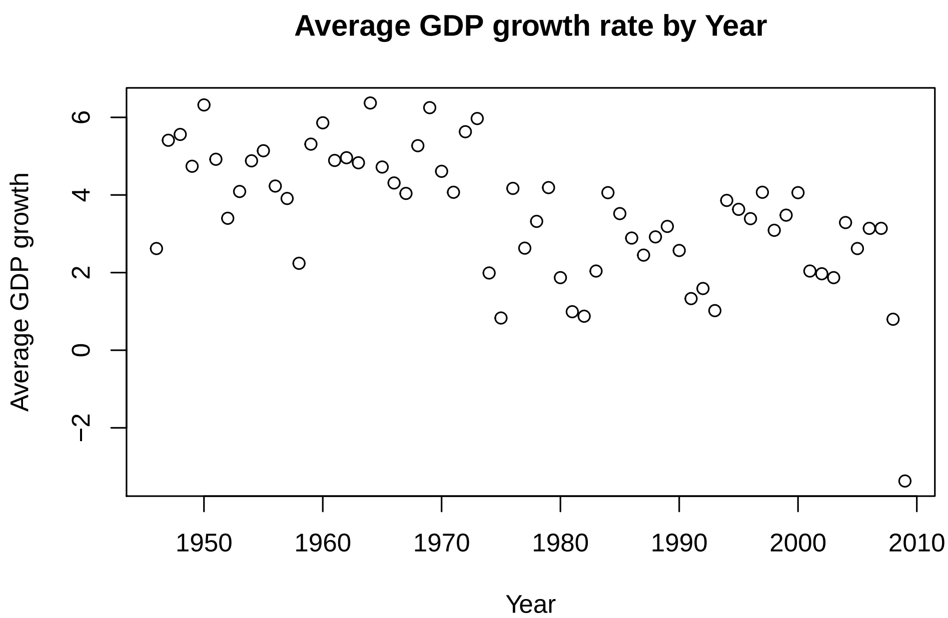
```r
#use daply()
avgGDP.year <- daply(debt, .(Year), mean.growth)
#check results
avgGDP.year["1972"]
```

```
## 1972
## 5.63
```

```r
avgGDP.year["1989"]
```

```
## 1989
## 3.19
```

```r
#plot
plot(y=avgGDP.year,x=c(1946:2009),ylab = "Average GDP growth", xlab = "Year",
     main = "Average GDP growth rate by Year")
```



3.

a.

The correlation coefficient between GDP growth and debt ratio over the whole dataset (all countries, all years) is -0.1995.

```
#corr btw GDP and debt ratio overall
signif(cor(debt$growth,debt$ratio),4)
```

```
## [1] -0.1995
```

b.

Below is the correlation coefficients for each country, and a histogram of them.

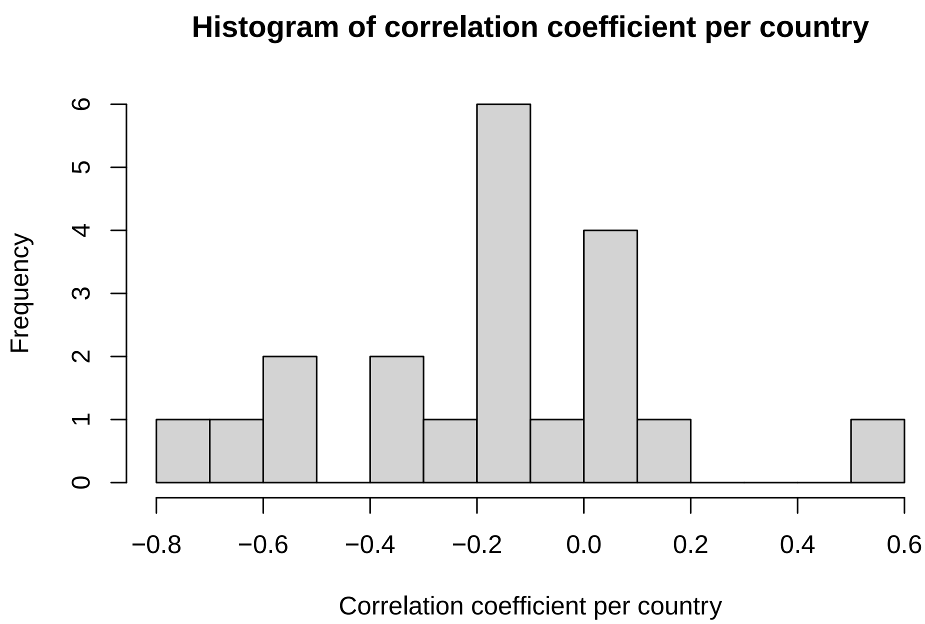
```
#a custom function to calculate corr btw GDP and debt ratio for a dataframe
corr.fcn <- function(df){
  signif(cor(df$growth,df$ratio),3)
}

#apply to data split by country
corr.country <- daply(debt,.(Country),corr.fcn)
corr.country
```

```
##   Australia     Austria     Belgium      Canada     Denmark     Finland
##    0.025200   -0.253000   -0.192000    0.075000   -0.168000    0.000581
##      France     Germany      Greece     Ireland       Italy       Japan
```

```
##   -0.502000   -0.576000   -0.093500   -0.140000   -0.645000   -0.702000
## Netherlands New Zealand      Norway    Portugal       Spain      Sweden
##   -0.199000    0.161000    0.563000   -0.352000    0.081400   -0.161000
##          UK          US
##   -0.137000   -0.341000
```

```
#plot historgram
hist(corr.country, breaks = 10,xlab = "Correlation coefficient per country",
     main = "Histogram of correlation coefficient per country")
```



c.

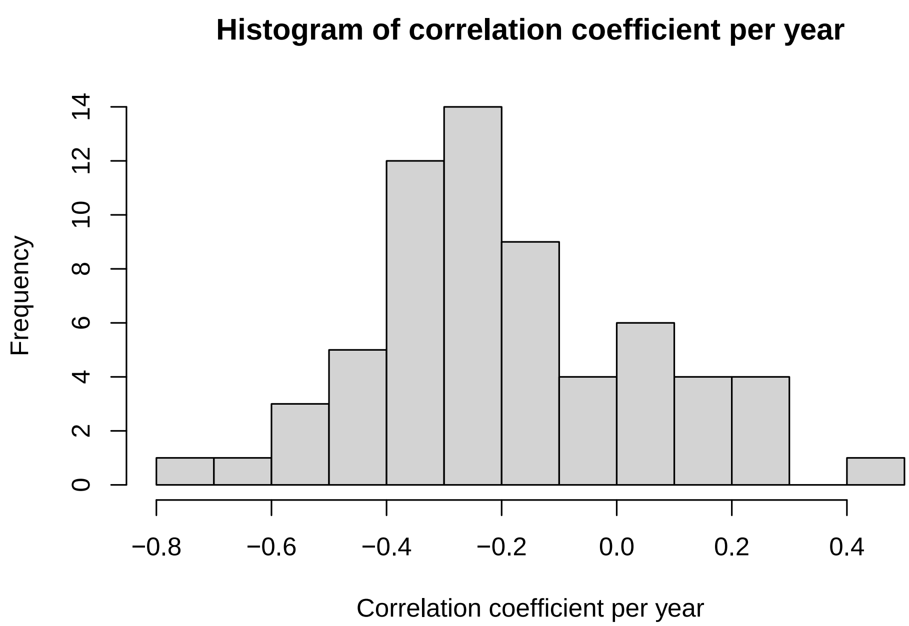
Below is the correlation coefficients for each year, and a histogram of them.

```
#apply to data split by year
corr.year <- daply(debt,.(Year),corr.fcn)
corr.year
```

```
##     1946     1947     1948     1949     1950     1951     1952     1953
## -0.62000 -0.27400 -0.34000 -0.20000  0.03980 -0.41600 -0.27700 -0.20500
##     1954     1955     1956     1957     1958     1959     1960     1961
## -0.27500 -0.22700 -0.45800 -0.75500 -0.45400 -0.28500 -0.50400 -0.53900
##     1962     1963     1964     1965     1966     1967     1968     1969
## -0.38300  0.12800 -0.36100 -0.31100 -0.31100 -0.27800 -0.18100 -0.25000
##     1970     1971     1972     1973     1974     1975     1976     1977
## -0.51200  0.00872 -0.19600  0.11400  0.26000  0.27100 -0.17100  0.16400
##     1978     1979     1980     1981     1982     1983     1984     1985
##  0.43100 -0.42900 -0.12700  0.03040  0.23900 -0.36200 -0.15600 -0.44900
##     1986     1987     1988     1989     1990     1991     1992     1993
## -0.35800 -0.06890  0.07970  0.06640  0.15600  0.20200 -0.00222 -0.37200
```

```
##      1994      1995      1996      1997      1998      1999      2000      2001
## -0.22400  0.05190 -0.35700 -0.11100 -0.26500 -0.25800 -0.13400 -0.23800
##      2002      2003      2004      2005      2006      2007      2008      2009
## -0.34900 -0.06790 -0.17100 -0.31400 -0.19600 -0.34400 -0.09450 -0.20500
```

```r
#plot historgram
hist(corr.year, breaks = 10,xlab = "Correlation coefficient per year",
     main = "Histogram of correlation coefficient per year")
```



d.

Looking at the histograms I could see there is a country and a year that has higher correlation than the rest, which I checked to be Norway and year 1978, respectively.

```r
which.max(corr.country)
```

```
## Norway
##     15
```

```r
which.max(corr.year)
```

```
## 1978
##   33
```

4.

The estimates for the intercept and slope of the linear model are 4.280 and -0.018, respectively.
The scatterplot of GDP growth versus debt ratio was also generated, with the red line being the fitted

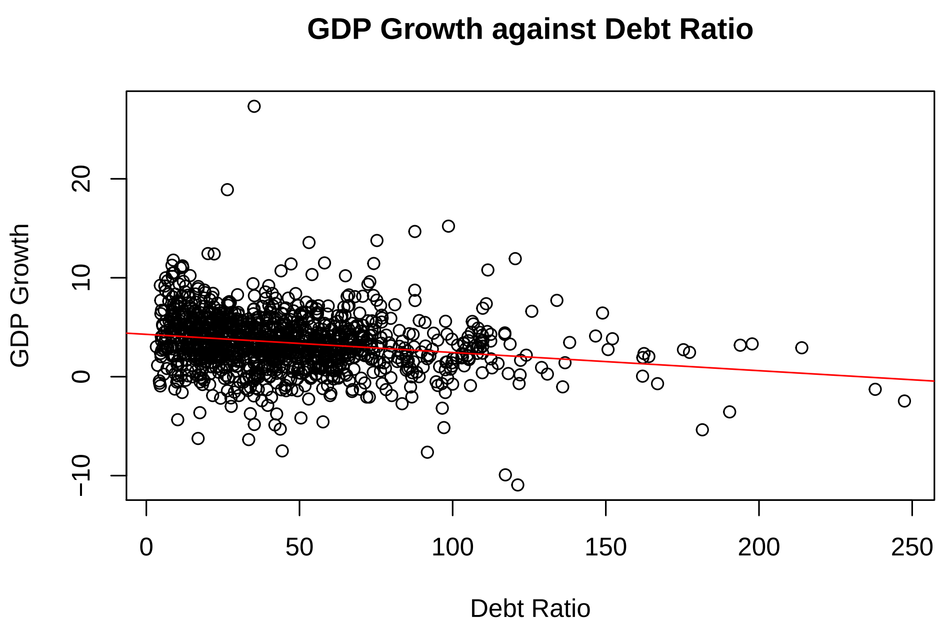
regression line.

```r
#fit linear model
debt.fit <- lm(growth ~ ratio, data = debt)
debt.fit
```

```
##
## Call:
## lm(formula = growth ~ ratio, data = debt)
##
## Coefficients:
## (Intercept)        ratio
##     4.27929     -0.01836
```

```r
#scatterplot
plot(x=debt$ratio,y=debt$growth,
     xlab = "Debt Ratio", ylab = "GDP Growth",
     main = "GDP Growth against Debt Ratio")

#fitted line
abline(debt.fit,col="red")
```



5.

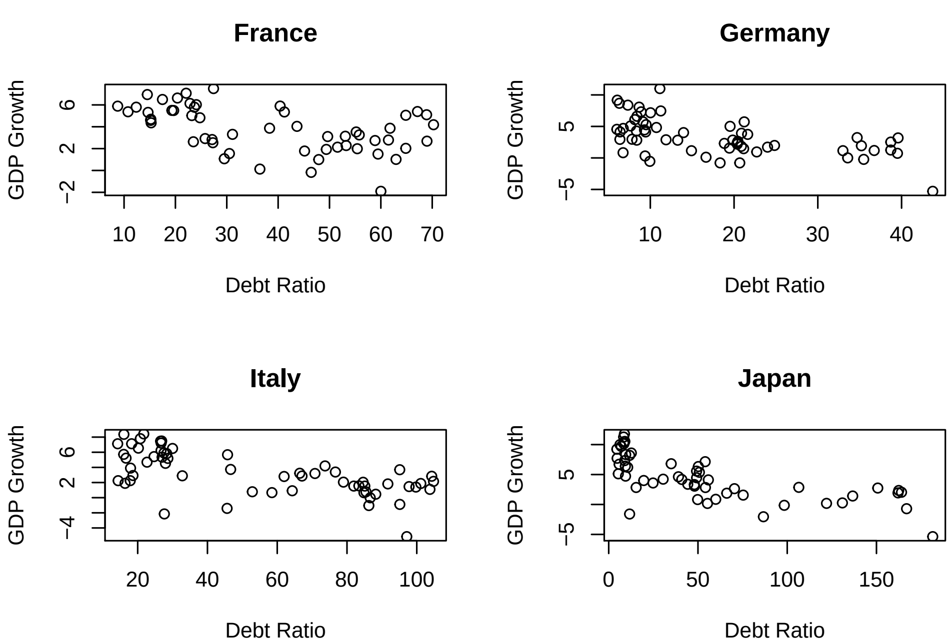
The four countries that have correlation smaller than -0.5 are France, Germany, Italy and Japan.
I then defined a custom function to plot GDP growth against debt ratio, and applied this function to data of the four countries split by country.

```r
#check countries
corr.country < -0.5
```

```
##   Australia     Austria     Belgium      Canada     Denmark     Finland
##       FALSE       FALSE       FALSE       FALSE       FALSE       FALSE
##      France     Germany      Greece     Ireland       Italy       Japan
##        TRUE        TRUE       FALSE       FALSE        TRUE        TRUE
## Netherlands New Zealand      Norway    Portugal       Spain      Sweden
##       FALSE       FALSE       FALSE       FALSE       FALSE       FALSE
##          UK          US
##       FALSE       FALSE
```

```r
#define plot fcn
plotxy <- function(df){
  plot(x=df$ratio,y=df$growth,
       xlab = "Debt Ratio", ylab = "GDP Growth",
       main = unique(df$Country))
}

#set grid for plots
par(mfrow=c(2,2))
#create plots
d_ply(debt[debt$Country %in% c("France","Germany","Italy","Japan"),c(1,3,4)],
      .(Country),
      plotxy)
```



6.

a.

Data for France is filtered out in a 54x4 dataframe.

```
#filter out france data
debt.fr <- debt[debt$Country=="France",]
dim(debt.fr)
```

```
## [1] 54  4
```

b.

A new column is added using a loop that checks the difference between the "Year" value of pairs of consecutive rows. If the difference is equal to 1 then the new column entry takes the "growth" value of the next row, and "NA" otherwise.

```
#empty vector for new column
next.growth <- c()

#a loop to check difference between year of consecutive rows, fill in empty vector with next growth and
for (i in 1:nrow(debt.fr)){
  if (i < nrow(debt.fr) & debt.fr[i+1,2]-debt.fr[i,2]==1) {
    next.growth[i] <- debt.fr[i+1,3]

  } else {
    next.growth[i] <- NA
  }
}

#add vector into df as a column
debt.fr$next.growth <- next.growth

#check result
signif(debt.fr[debt.fr$Year==1971,5],4)
```

```
## [1] 5.886
```

```
signif(debt.fr[debt.fr$Year==1972,5],4)
```

```
## [1] NA
```

7.

I defined a custom function that is essentially the loop procedure in previous part, and applied this function on the data split by country.

```r
#define function
next.fcn <- function(df){
  next.vec <- c()
  for(i in 1:nrow(df)){
    if (i < nrow(df) & df[i+1,2]-df[i,2]==1) {
    next.vec[i] <- df[i+1,3]

  } else {
    next.vec[i] <- NA
  }
  }
  df$next.growth <- next.vec
  return(df)
}

#apply to split data
debt.next <- ddply(debt, .(Country), next.fcn)

#check result
signif(debt.next[debt.next$Year==2009 & debt.next$Country=="France",5],4)
```

```
## [1] NA
```

8.

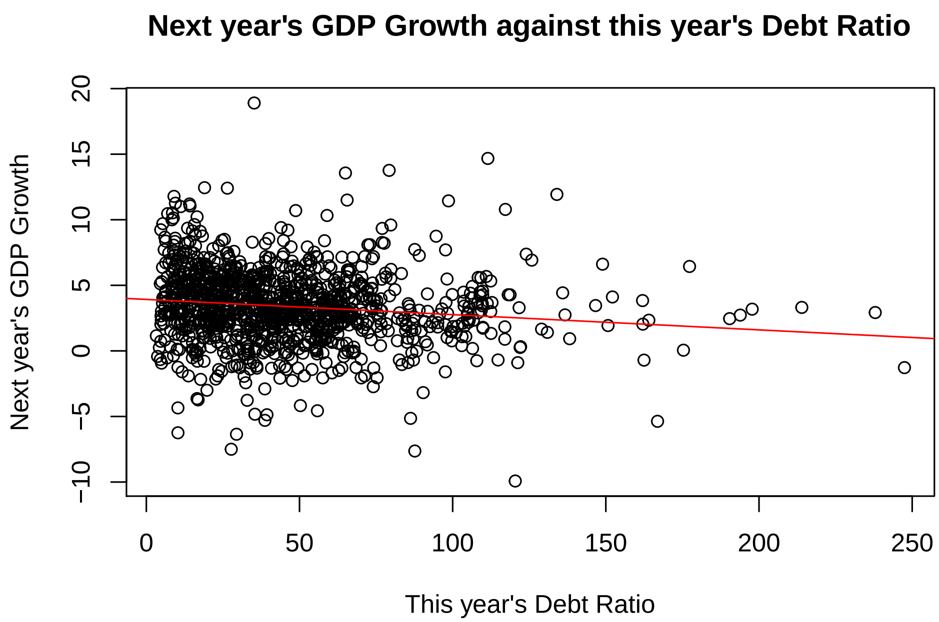
The estimates for the intercept and slope of the linear model are 3.925 and -0.012, respectively.
The scatterplot of next year's GDP growth versus this year's debt ratio was also generated, with the red line being the fitted regression line.
I think there isn't too much difference comparing the two models. This new model even fits a bit worse than the original model in question 4, because from looking at the plot, we can see most of the points stay packed together and closer to the line in the original model, but somewhat more spread out in this new one.

```r
#fit linear model
debt.fit2 <- lm(next.growth ~ ratio, data = debt.next)
debt.fit2
```

```
##
## Call:
## lm(formula = next.growth ~ ratio, data = debt.next)
##
## Coefficients:
## (Intercept)        ratio
##     3.92472     -0.01161
```

```r
#scatterplot
plot(x=debt.next$ratio,y=debt.next$next.growth,
     xlab = "This year's Debt Ratio", ylab = "Next year's GDP Growth",
     main = "Next year's GDP Growth against this year's Debt Ratio")

#fitted line
abline(debt.fit2,col="red")
```



9.

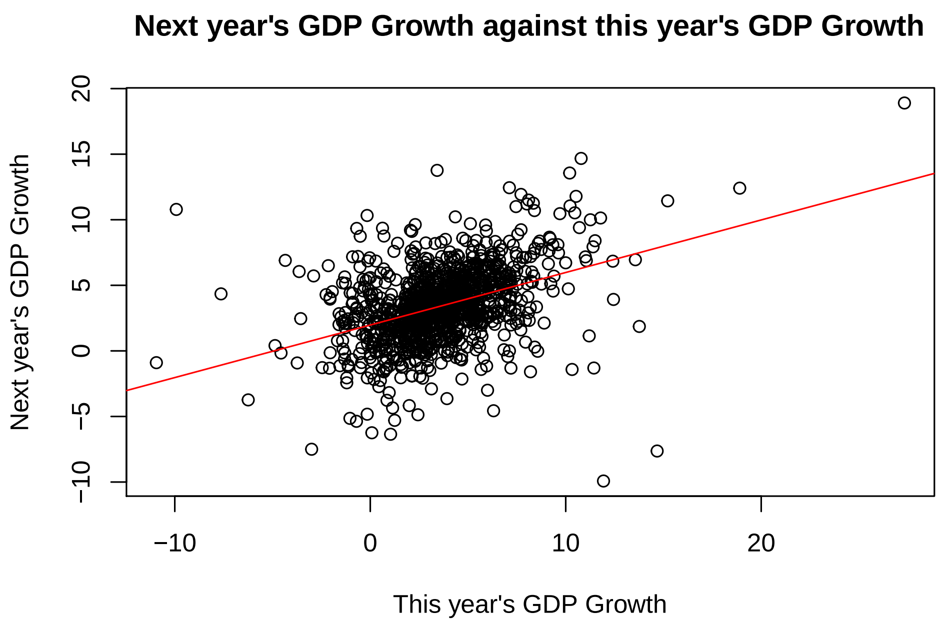
The estimates for the intercept and slope of the linear model are 1.971 and 0.401, respectively.
The scatterplot of next year's GDP growth versus this year's GDP growth was also generated, with the red line being the fitted regression line.
Looking at this model and the model in question 8, we can see that current GDP might be a better predictor for future GDP than debt ratio. The correlation seems a litle stronger in this new model compared to model in question 8.

```r
#fit linear model
debt.fit3 <- lm(next.growth ~ growth, data = debt.next)
debt.fit3
```

```
##
## Call:
## lm(formula = next.growth ~ growth, data = debt.next)
##
## Coefficients:
## (Intercept)       growth
##      1.9711       0.4007
```

```r
#scatterplot
plot(x=debt.next$growth,y=debt.next$next.growth,
     xlab = "This year's GDP Growth", ylab = "Next year's GDP Growth",
     main = "Next year's GDP Growth against this year's GDP Growth")

#fitted line
abline(debt.fit3,col="red")
```



Next year's GDP Growth against this year's GDP Growth
Next year's GDP Growth
This year's GDP Growth
−10
0
10
20
−10
−5
0
5
10
15
20