Lab1

Greg Ceccarelli

June 10, 2015

Part 1. Multiple Choice (5 points each)

1. e
2. a
3. d
4. b
5. d
6. c
7. c
8. d
9. b

Part 2a. Variable Manipulations (10 points each)

10. Load the data found in the file, GDPWorldBank.csv into a new dataframe. Notice that this dataset includes GDP statistics for the years 2010, 2011, and 2012.Create a new variable, gdp_growth, that equals the increase in GDP from 2011 to 2012 (the nominal increase, not a fractional one). What is the mean of your new variable?

```
#set up working envrionment
setwd('/Users/ceccarelli/MIDS/DATASCI_W203/Assignments/Labs/Lab 1')
rm( list = ls() )

#load libraries
library(calibrate) # used to label xy points in eventual scatter
```

```
## Loading required package: MASS
```

```
#define function to omit row values where only certain columns are NA
data.complete <- function(data, desiredCols) {
  completeVec <- complete.cases(data[, desiredCols])
  return(data[completeVec, ])
}

##force R not to use scientific notation
options("scipen"=10, "digits"=4)

##load GDP data as a data frame
gdp.data <- read.csv("GDP_World_Bank1.csv", sep=",", header = TRUE)

#count number of rows with incomplete entries
nrow(gdp.data[!complete.cases(gdp.data),])
```

```
## [1] 39
```

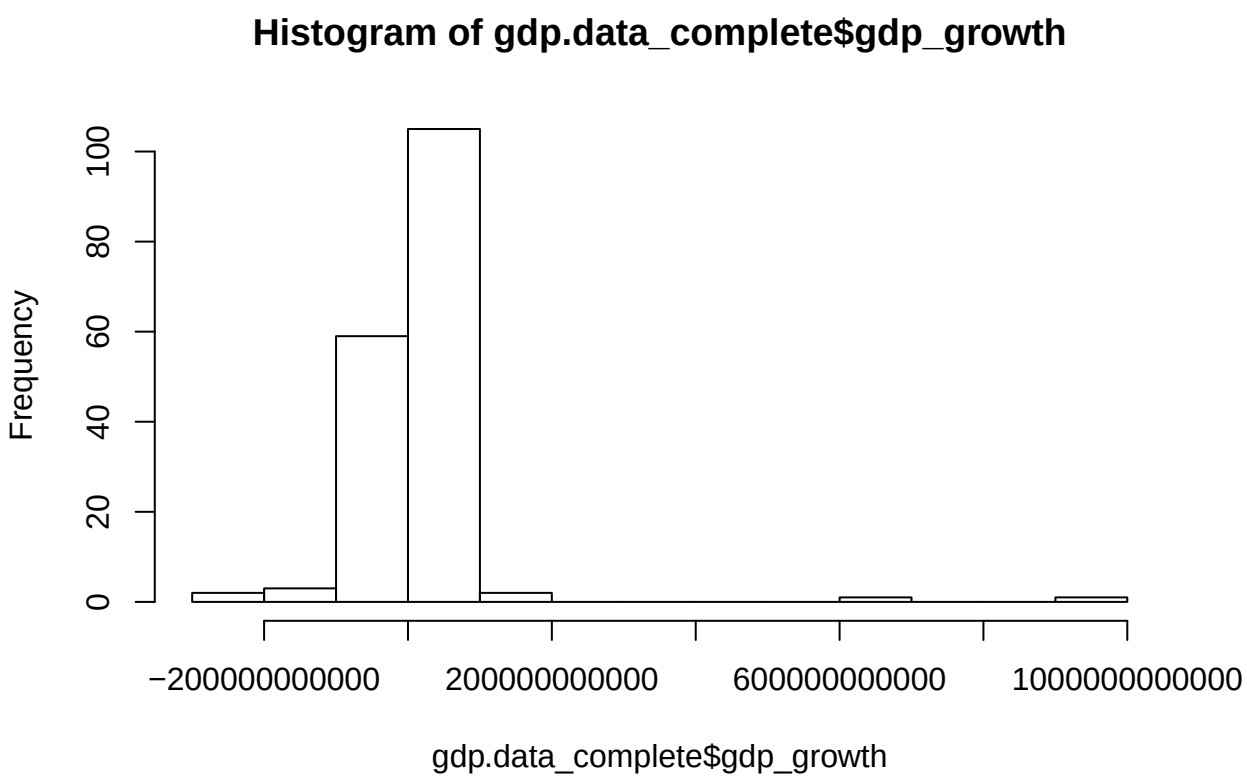
```r
gdp.data_complete <- data.complete(gdp.data, c('gdp2011', 'gdp2012'))

#create new nominal variable based on gdp increase
gdp.data_complete$gdp_growth <- gdp.data_complete$gdp2012 - gdp.data_complete$gdp2011

#create and store mean of new variable
mx <- mean(gdp.data_complete$gdp_growth)
paste("mean is: ", mx)
```

```
## [1] "mean is:  7172376796.32376"
```

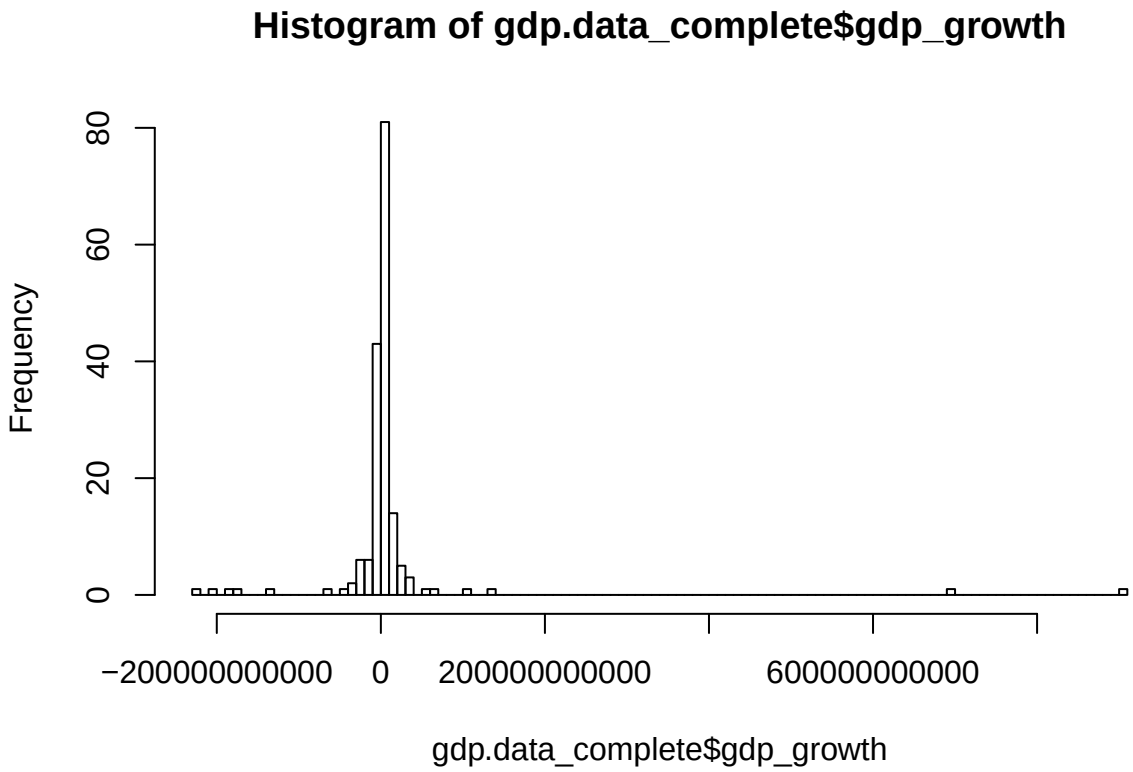
RESPONSE: As indicated above in the block of code, the mean (controlling for NAs) is 7172376796

11. Create a histogram of your new variable from question 10. Is it normally distributed? Describe its shape in terms of the distribution properties from the class.

```r
#view the histogram and observe its shape
hist(gdp.data_complete$gdp_growth)
```




```r
#change number of breaks to observe slightly different version
hist(gdp.data_complete$gdp_growth, breaks = 100)
```



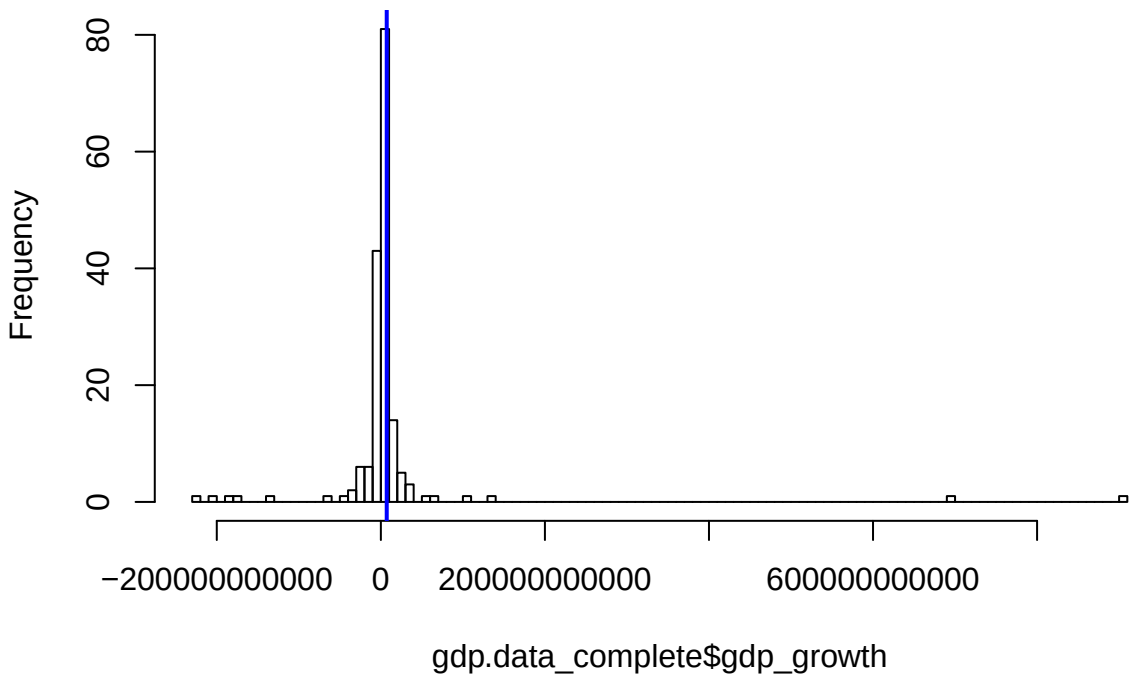
RESPONSE: The graph above is not normally distributed but rather exhibits leptokurtic positive skew given the right tail is longer and the mass of the distribution is concentrated on the left of the figure.

12. Create a new Boolean variable, high_growth, that equals TRUE if a country's gdp growth is higher than the mean. How many countries have above average growth, and how many have below average growth? Explain this result in terms of the shape of the gdp_growth distribution.

```
# Question 12
#create new high_growth variable based on comparison of gdp_growth to population mean
gdp.data_complete$high_growth <- gdp.data_complete$gdp_growth > mx

#count TRUE / FALSE values
summary(gdp.data_complete$high_growth)
```

```
##    Mode   FALSE    TRUE    NA's
## logical     142      31       0
```

```
hist(gdp.data_complete$gdp_growth, breaks = 100)
#add mean line to histogram to explain TRUE/FALSE breakdown
abline(v = mx, col = "blue", lwd = 2)
```



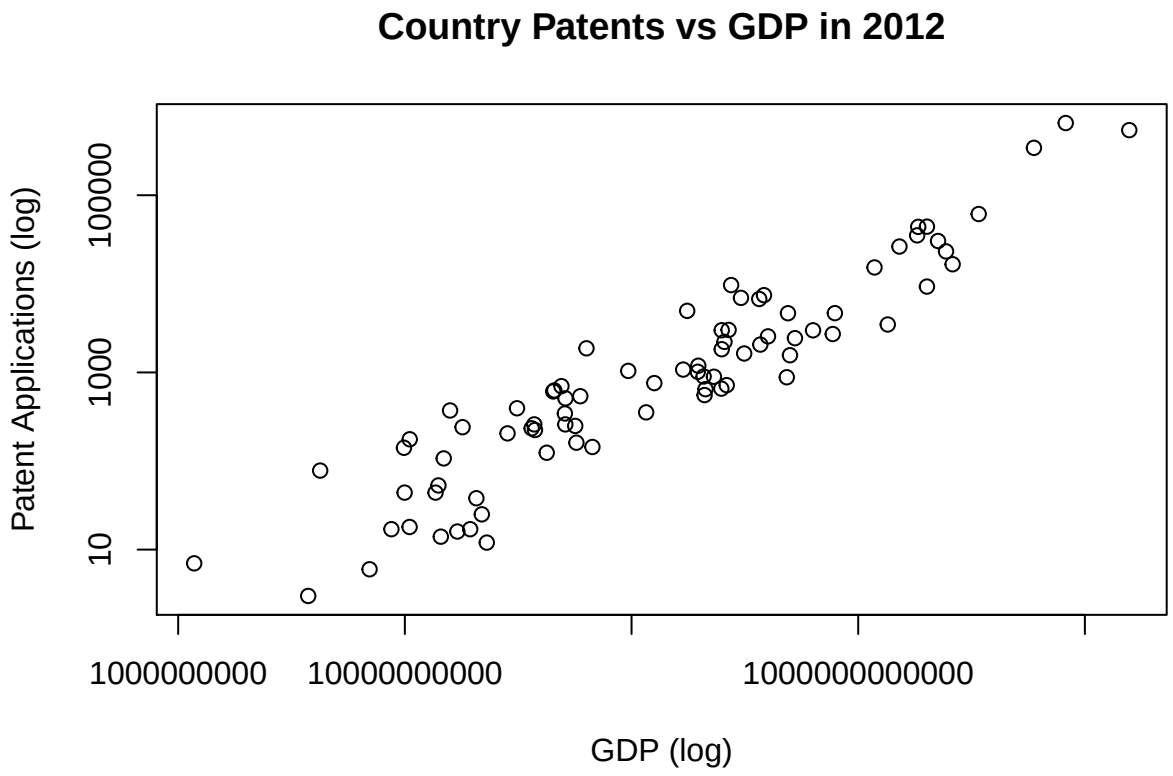
RESPONSE: Based on the printed summary of T/F values, there are 31 countries with above average growth between 2011-2012 and 142 below. The reason there are so many countries below is due to the number of countries with negative growth. This is further explained by the positive skew of the gdp growth distribution and viewing the histogram with more breaks. I've shown a mean line on the graph to demonstrate this further

Part 2b. Data Import (25 points)

13. Find one new metric country-level variable from some public source, and merge it into your dataset ... Once you have successfully merged your dataframes, create one more graph that you think is interesting that involves your new variable. Write a few sentences about what the graph shows.

```
#Data Sourced from World Intellectual Property Organization
#http://ipstats.wipo.int/ipstatv2/index.htm?tab=patent
#Report Parmaters:
#Intellectual Property Right: Patent
#Year Range: 2011 - 2012
#Reporting Type: Total Count by Filing Office
#Indicator: 1 - Total patent applications

#did minor data cleaning in text editor to remove extraneous header lines
patent.data <- read.csv("patent_applications_2011_2012_cleaned.csv", sep=",", header = TRUE)
patent.data_complete <- patent.data[c("Office","X2011","X2012")]
names(patent.data_complete) <- c("Country","PatentApplications2011","PatentApplications2012")

#create growth variable
patent.data_complete$patent_growth <- patent.data_complete$PatentApplications2012 - patent.data_complete

#merge data together -- keep only those observations that match
```

```r
gdp_patent.data_complete <- merge(gdp.data_complete,patent.data_complete,by="Country")
gdp_patent.data_complete <- data.complete(gdp_patent.data_complete,c('gdp_growth','patent_growth'))

#plot just the nominal values - patent apps vs gdp in 2012
plot(gdp_patent.data_complete$gdp2012
  , gdp_patent.data_complete$PatentApplications2012
  , log = "xy"
  , xlab="GDP (log)"
  , ylab="Patent Applications (log)"
  , title("Country Patents vs GDP in 2012"))
```

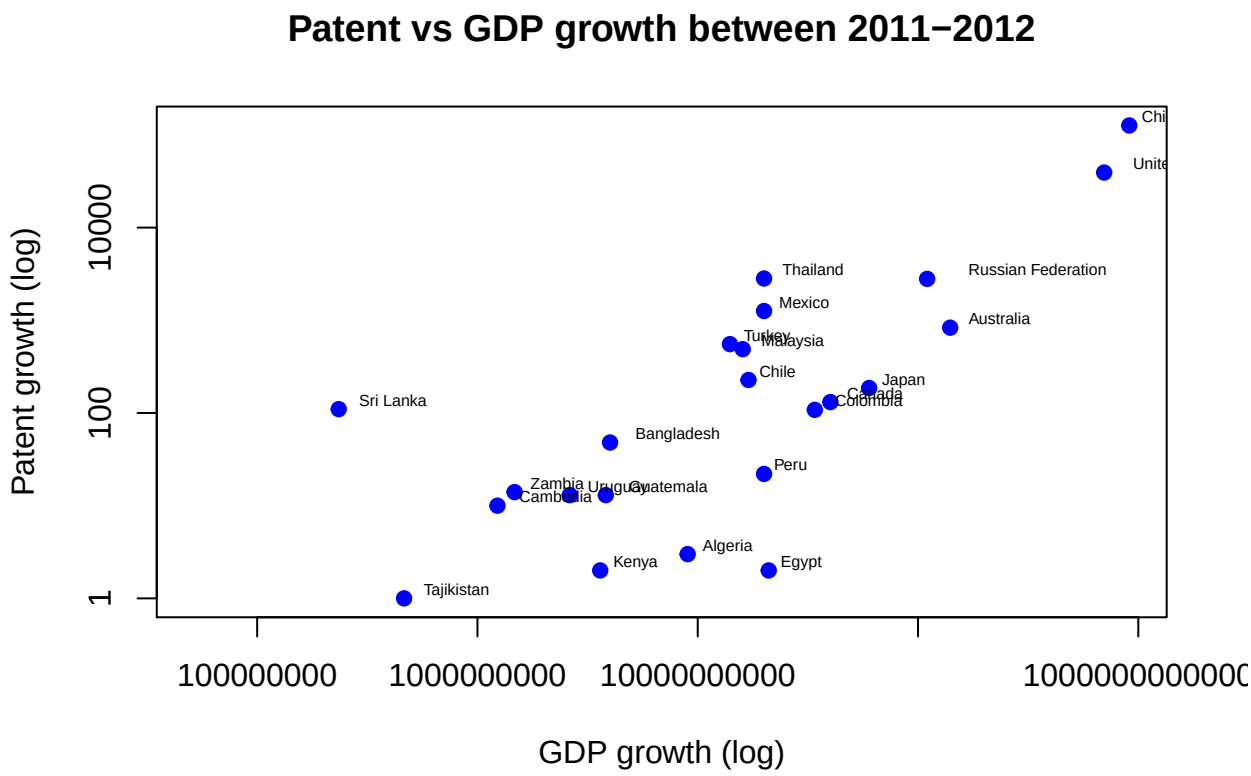



```r
##final plot
##this plots the positive growth only, omits negative growth due to log transform
plot(gdp_patent.data_complete$gdp_growth
     , gdp_patent.data_complete$patent_growth
     , log = "xy"
     , xlab="GDP growth (log)"
     , ylab="Patent growth (log)"
     , title("Patent vs GDP growth between 2011-2012")
     , col= "blue", pch = 19, cex = 1, lty = "solid")
```

```
## Warning in xy.coords(x, y, xlabel, ylabel, log): 41 x values <= 0 omitted
## from logarithmic plot
```

```
## Warning in xy.coords(x, y, xlabel, ylabel, log): 38 y values <= 0 omitted
## from logarithmic plot
```

```
#add country labels to graph
textxy(gdp_patent.data_complete$gdp_growth, gdp_patent.data_complete$patent_growth, gdp_patent.data_comp
```



RESPONSE: For those countries with positive GDP growth, the final graph of interest entitled [Patent vs GDP growth between 2011-2012], demonstrates a positive correlation between growth in patent applications and gdp from 2011-2012. China's lead is unsuprising to me but I was suprised Mexico and Thailand where with such patent growth. Other countries such as Kenya, Algeria and Egypt exhibit small patent growth with fairly sizeable gdp growth. Perhaps characteristic of their economic outputs being less tech focused.