Empirical Methods Homework 1

YiTao Hu, Jin (Jane) Huangfu, Charles Rambo, Junyu (Kevin) Wu

1/14/2020

Problem 1

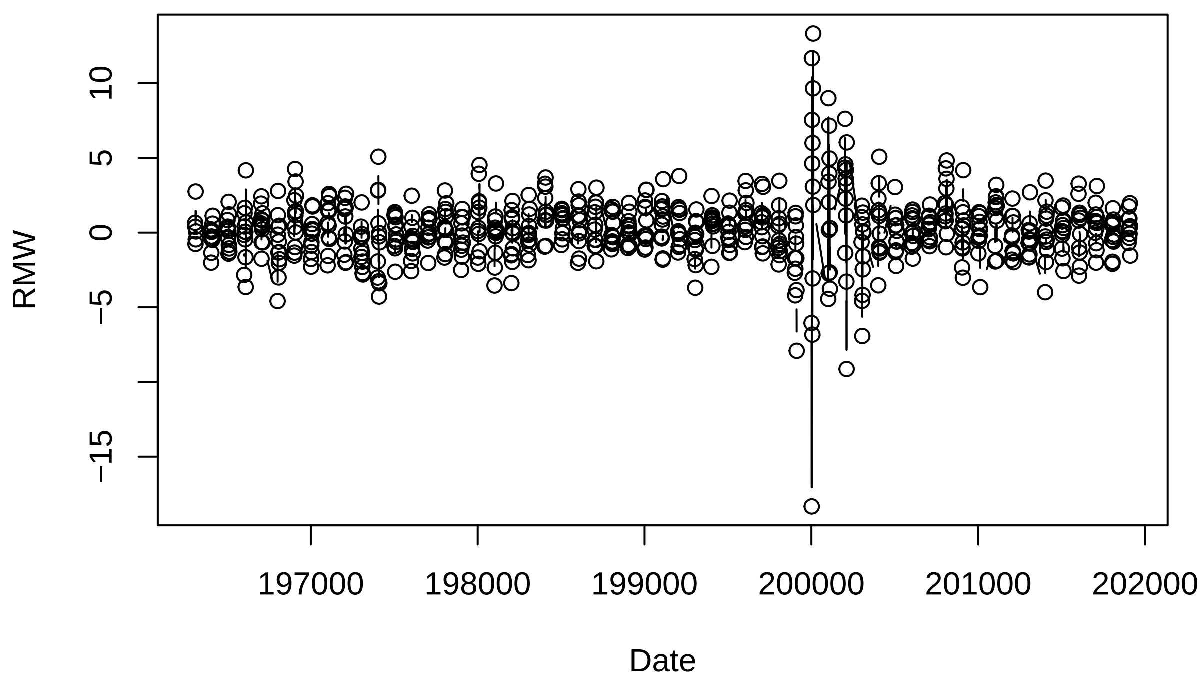
Question 1

We obtained the Fama/French 5 Factors (2x3) CSV file from

`https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html`.

For simplicity, we added an eighth column which was a date with a hyphen between the year and month in Excel, so R would be able to recognize the date. We then replaced the first column with our modification. The annualized arithmetic mean was 3.111669%, annualized geometric mean was 2.868361%, and a standard deviation was 7.472613%.

```
data <- read.csv(file=file.choose(), header=TRUE, sep=",", skip = 3, nrows = 677)
names(data) <- c("Date", "Mkt.RF", "SMB", "HML", "RMW", "CMA", "RF")
rmw <- data$RMW
plot(x = data$Date, y = rmw, type="b", xlab = "Date", ylab = "RMW")
```




```
arith_average_retun <- 12*mean(rmw)
geo_average_return <- 100*(prod(1 + rmw/100)^(12/length(rmw)) - 1)
standard_deviation <- sqrt(12)*sd(rmw)
arith_average_retun
```

```
## [1] 3.111669
```

```r
geo_average_return
```

```
## [1] 2.868361
```

```r
standard_deviation
```

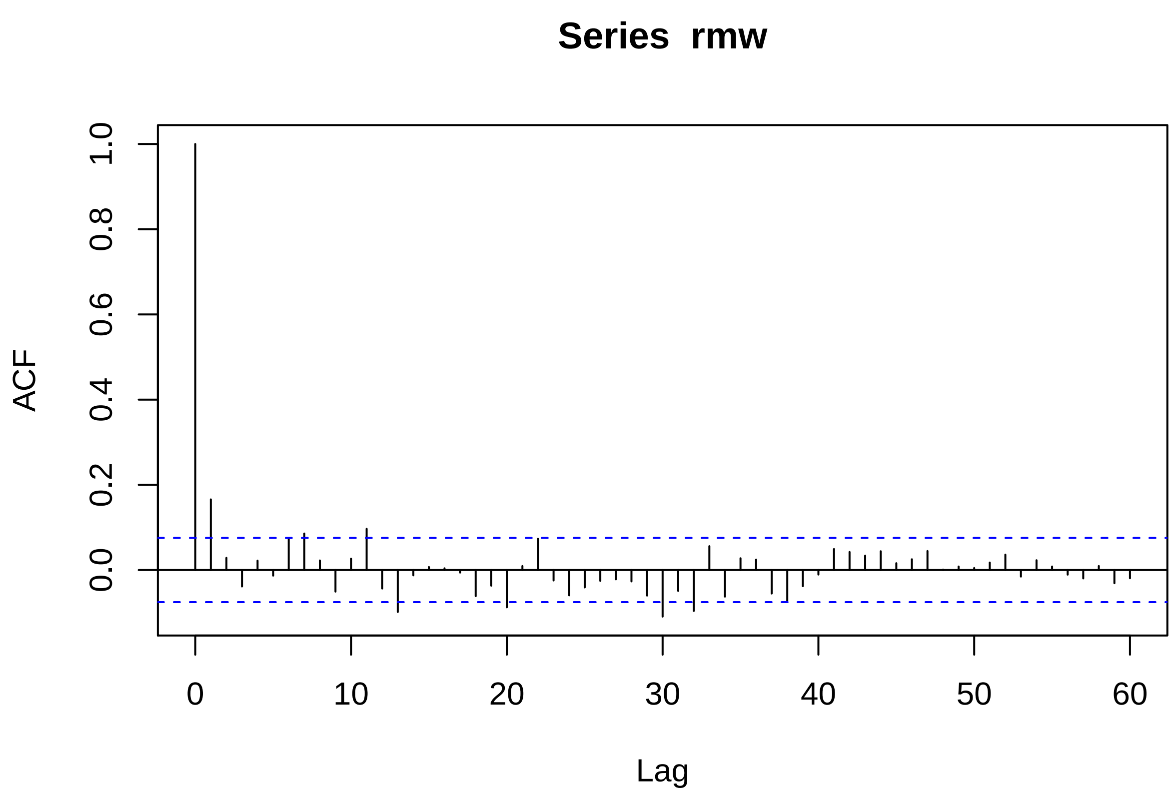
```
## [1] 7.472613
```

Question 2

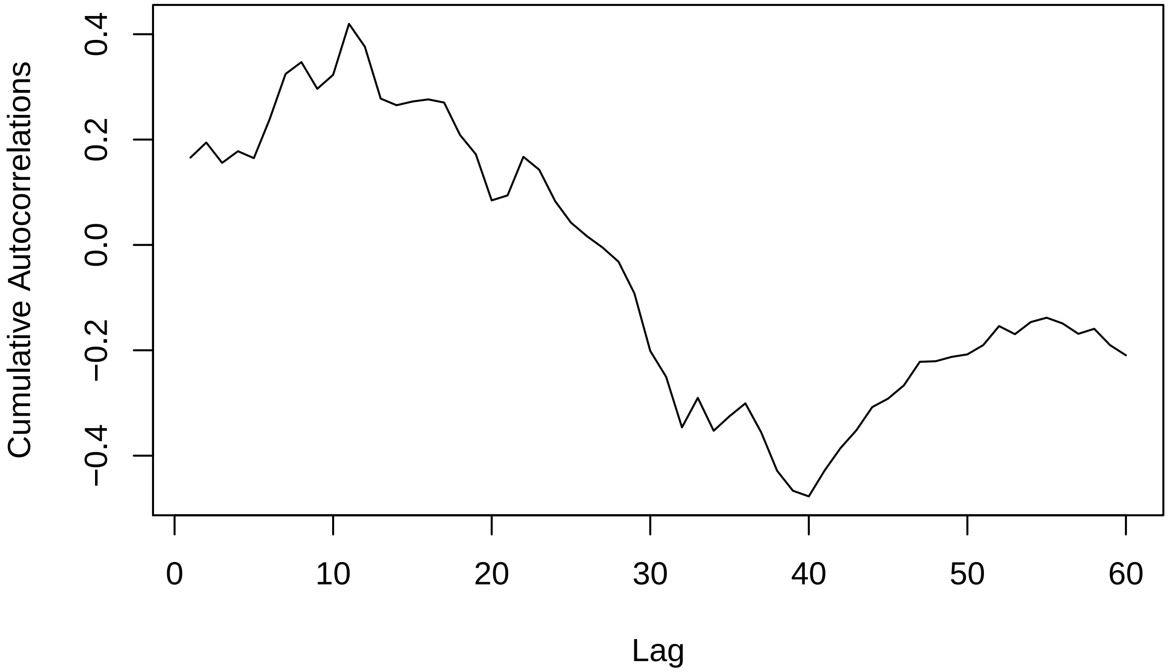
The autocorrelations are, on balance, positive until the 11th, and they are net negaitve until the 41st, though there are a few positive autocorrelations so the cumulative sum is not decreasing monotonically, e.g. the 22nd lag appears to be positive.

From the graph of the acf function, we see that the only statistically significant lags are the 1st, 7th, 11th, 13th, 20th, 30th, and 32nd.

```r
acf(rmw, lag.max = 60)
autocorrelations <- as.vector(acf(rmw, lag.max = 60)$acf)
```




```r
cumulative_autocorrelations <-  cumsum(autocorrelations[-1])
plot(x = 1:60, y = cumulative_autocorrelations, type = "l", xlab = "Lag",
     ylab = "Cumulative Autocorrelations" )
```



Question 3

We will perform a Box-Ljung test on the data. The hypothesis test is of the form

$$H_0 : \rho_1 = \rho_2 = \ldots = \rho_6 = 0 \quad H_a : \rho_i \neq 0 \text{ for some } i$$

The p-value of our test is 0.0004246266. Hence, we reject the null hypothesis that the first six autocorrelations are all zero at a significance level of 5%, i.e. at least one is non-zero.

```
library(stats)
Box.Ljung <- Box.test(rmw, lag = 6, type = "Ljung-Box")
p_value <- Box.Ljung$p.value
p_value
```

```
## [1] 0.0004246266
```

Question 4

Via examination of the acf plot, it is clear that $AR(1)$ is

$$rmw_{t+1} = \alpha + \beta' rmw_t + \epsilon_{t+1}.$$

Note that we replaced the NA in the lagged RMW variable with the frist entry of RMW to avoid errors in our calculations later.

```
library(Hmisc)
```

```
## Warning: package 'Hmisc' was built under R version 3.6.2
```

```
## Loading required package: lattice

## Loading required package: survival

## Loading required package: Formula

## Loading required package: ggplot2

##
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
lag_rmw <- Lag(rmw, shift = 1)
lag_rmw[1] <- rmw[1]
model <- lm(rmw ~ lag_rmw)
summary(model)
```

```
##
## Call:
## lm(formula = rmw ~ lag_rmw)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.5411  -1.0978   0.0138   0.9929  14.5081
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.21571    0.08242   2.617  0.00907 **
## lag_rmw      0.16606    0.03797   4.373 1.42e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.129 on 675 degrees of freedom
## Multiple R-squared:  0.02755,    Adjusted R-squared:  0.02611
## F-statistic: 19.12 on 1 and 675 DF,  p-value: 1.42e-05
```

We preformed a Box-Ljung test on the residuals to affirm that we have obtained all meaningul autocoreelations and got a p-value of approximitely 48%. As a result, we do not reject the null hypothesis that the first six autocorrelations are all zero.

```
Box.Ljung.residuals <- Box.test(model$residuals, lag = log(length(model$residuals)), type = "Ljung-Box")
Box.Ljung.residuals$p.value
```

```
## [1] 0.4801619
```

The fact that an $AR(1)$ is well suited for this data is somewhat intuitive because we would expect investors' expectation of ROE on the returns of stocks to be relatively stable overtime but we do not expect the relationship to be strong or long lasting because, if it were, more investors would simply utilize the strategy and inflate the stock price and therefore reduce future returns.

Question 5

We will calculate the variances by means of matracies. Note that

$$\begin{aligned}
\hat{\beta} &= (X'X)^{-1}X'y \\
&= (X'X)^{-1}X'(X\beta + \epsilon) \\
&= (X'X)^{-1}X'X\beta + (X'X)^{-1}X'\epsilon \\
&= \beta + (X'X)^{-1}X'\epsilon
\end{aligned}$$

Thefore, we have

$$\begin{aligned}
Var(\hat{\beta}) &= E\left[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'\right] \\
&= E\left[\left((X'X)^{-1}X'\epsilon\right)\left((X'X)^{-1}X'\epsilon\right)'\right] \\
&= E\left[(X'X)^{-1}X'\epsilon\epsilon'X(X'X)^{-1}\right] \\
&= (X'X)^{-1}X'E\left[\epsilon\epsilon'\right]X(X'X)^{-1}
\end{aligned}$$

When we consider Var^{OLS}, we suppose

$$E\left[\epsilon\epsilon'\right] = \sigma^2 I,$$

which makes are equation

$$Var^{OLS}(\hat{\beta}) = \sigma^2(X'X)^{-1}.$$

For the White estimator, we suppose that

$$E\left[\epsilon\epsilon'\right] = ee' =: \Lambda,$$

where $e = y - X\hat{\beta}$. This gives,

$$Var^{White}(\hat{\beta}) = (X'X)^{-1}X'\Lambda X(X'X)^{-1}.$$

In our particular case, we are interested in the lower right entries. Hence, for the OLS variance, we obtain 0.001439951 and for the White variance we obtain 0.01261432. Taking square roots leads us to conclude that the OLS and White standard errors are 0.03794669 and and 0.1123135, respectively.

```r
X <- matrix(1,  nrow = length(lag_rmw), ncol = 2)
X[, 2] <- lag_rmw
sigma_squared_ols <- var(model$residuals)
var_ols <- sigma_squared_ols*solve(t(X) %*% X)
var_ols[2, 2]
```

```
## [1] 0.001439951
```

```r
SE_ols <- sqrt(var_ols[2, 2])
SE_ols
```

```
## [1] 0.03794669
```

```r
Lambda <- diag(model$residuals^2)
var_white <- solve(t(X) %*% X) %*% t(X) %*% Lambda %*% X %*% solve(t(X) %*% X)
var_white[2, 2]
```

```
## [1] 0.01261432
```

```r
SE_white <- sqrt(var_white[2, 2])
SE_white
```

```
## [1] 0.1123135
```

Problem 2

We provide the code for a our simulation below.

```
set.seed(42)
mu <- 0.005
sigma <- 0.04
N <- 10000
T <- 600

betas_r <- c()
betas_p <- c()
p_1 <- c(0)
p_2 <- c(0)

for(k in 1:N){
  epsilon_1 <- rnorm(T)
  epsilon_2 <- rnorm(T)
  r_1 <- mu + sigma*epsilon_1
  r_2 <- mu + sigma*epsilon_2
  betas_r[k] <- lm(r_1 ~ r_2)$coef[2]
  for(l in 2:(T+1)){
    p_1[l] <- p_1[l - 1] + r_1[l - 1]
    p_2[l] <- p_2[l - 1] + r_2[l - 1]
  }
  betas_p[k] <- lm(p_1 ~ p_2)$coef[2]
}
```

Question 1

The mean value of β^* in question 1 is 0.0001563299 and the standard deviation is 0.04082026. The 95% confidence interval is $(-0.07975113, 0.08026071)$. Hence, at a 5% significance level, we fail to reject the null hypothesis that $\beta = 0$.

```
mean_beta_r <- mean(betas_r)
sd_beta_r <- sd(betas_r)
mean_beta_r
```

```
## [1] 0.0001563299
```

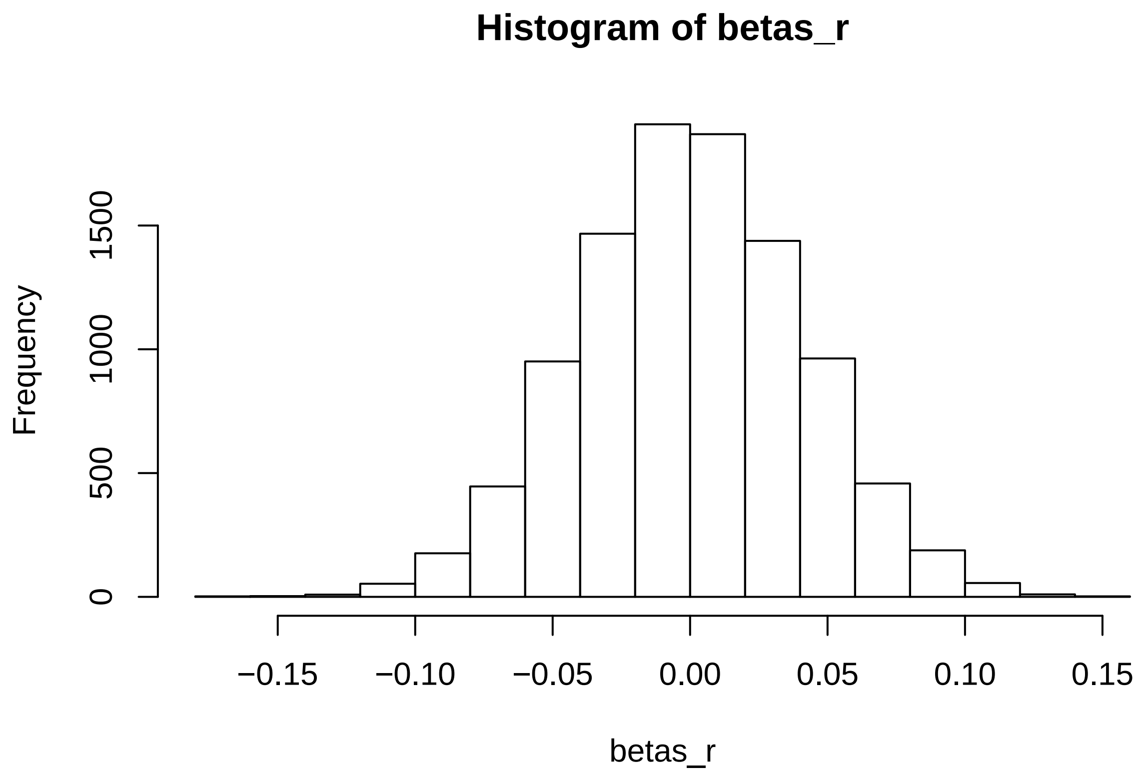
```
sd_beta_r
```

```
## [1] 0.04082026
```

```
quantile(betas_r, c(0.025, 0.975))
```

```
##        2.5%       97.5%
## -0.07975113  0.08026071
```

```
hist(betas_r)
```



Question 2

For question 2, the mean value of β^* is -0.01283888 and the standard deviation is 0.02587436. The 95% confidence interval is $(0.2373364, 2.1884223)$. Therefore, at a 95% confidence level, we reject the null hypothesis that $\beta = 0$.

```
mean_beta_p <- mean(betas_p)
sd_beta_p <- sd(betas_p)
mean_beta_p
```

```
## [1] 0.9857455
```

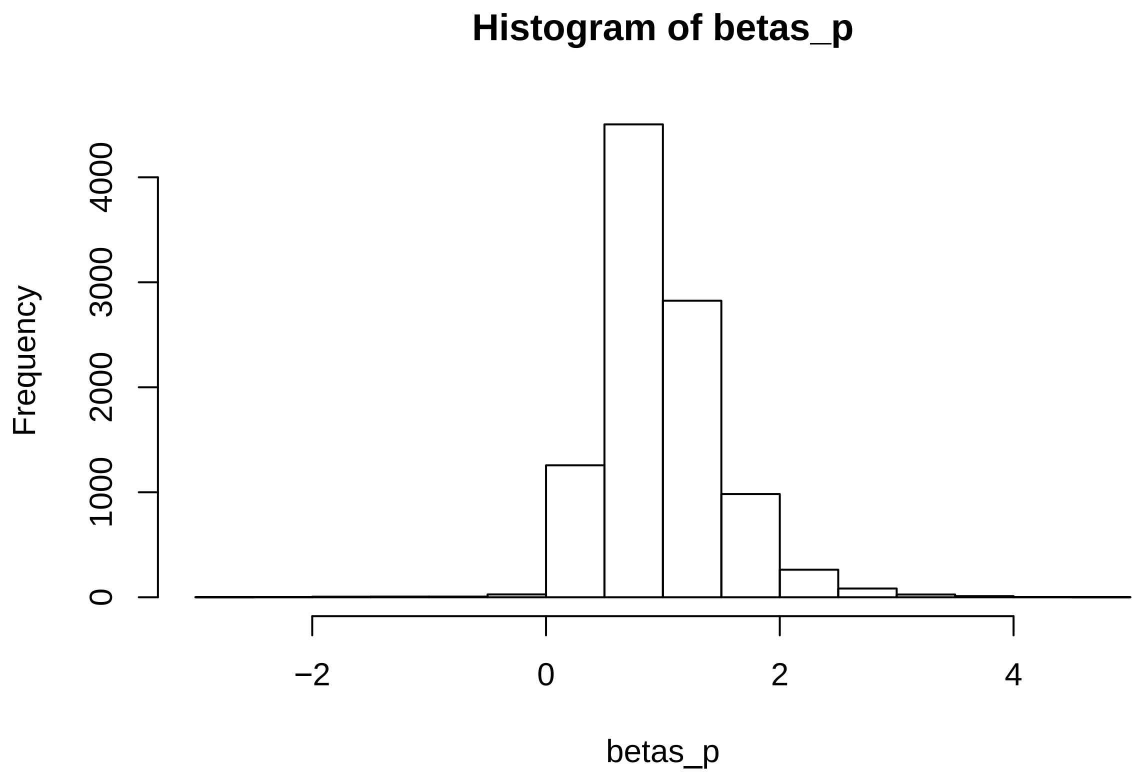
```
sd_beta_p
```

```
## [1] 0.5083531
```

```
quantile(betas_p, c(0.025, 0.975))
```

```
##      2.5%     97.5%
## 0.2373364 2.1884223
```

```
hist(betas_p)
```



Histogram of betas_p
Frequency
0
1000
2000
3000
4000
−2
0
2
4
betas_p