Evaluation of Fama French five factor model using Autoregression Models

Problem 1

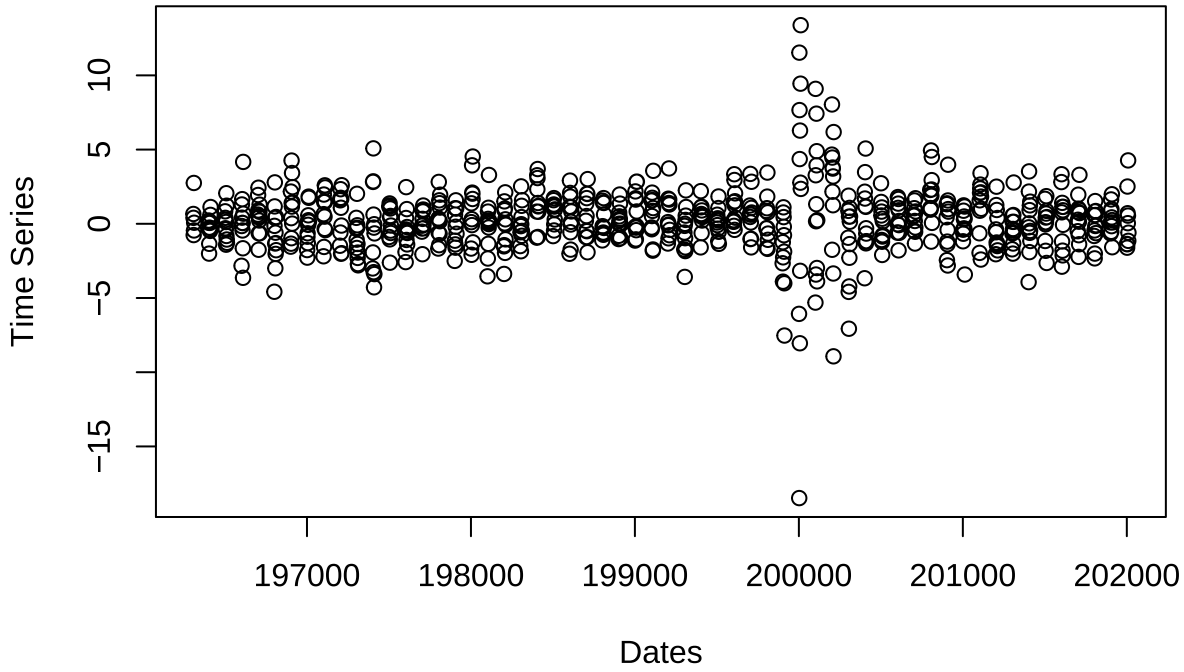
Fama and French (2015) propose a five-factor model for expected stock returns. One of the factor is based on cross-sectional sorts on firm probability. In particular, the factor portfolio is long firms with high probability (high earnings divided by book equity; high ROE) and short firms with low probability (low earnings divided by book equity; low ROE). This factor is called RMW - Robust Minus Weak.

a) Go to Ken Frenchís Data Library (google it) and download the Fama/French 5 Factors (2x3) in CSV format. Denote the time series of value-weighted monthly factor returns for the RMW factor from 1963.07-2020.10 as "rmw." Plot the time-series, give the annualized mean and standard deviation of this return series.

Solution:

```
#Pulling F-F 5 factor data
library(DataAnalytics)
library("readxl")
full <- as.data.frame(read_excel
("~/Documents/Documents/Empirical Methods/F-F_Research_Data_5_Factors_2x3.xls"))
rmw. <- as.data.frame(read_excel
("~/Documents/Documents/Empirical Methods/F-F_Research_Data_5_Factors_2x3.xls"))
```

```
#Selecting RMW Column
rmw. <- rmw.[,5]
rmw. <- as.data.frame(rmw.)

plot(full[,1],full[,5],xlab="Dates",ylab="Time Series")
```



```
#Annual Mean & SD
Annual_mean <- 12*mean(rmw.[,1])
Annual_SD <- sqrt(12)*sd(rmw.[,1])

Annual_mean
```

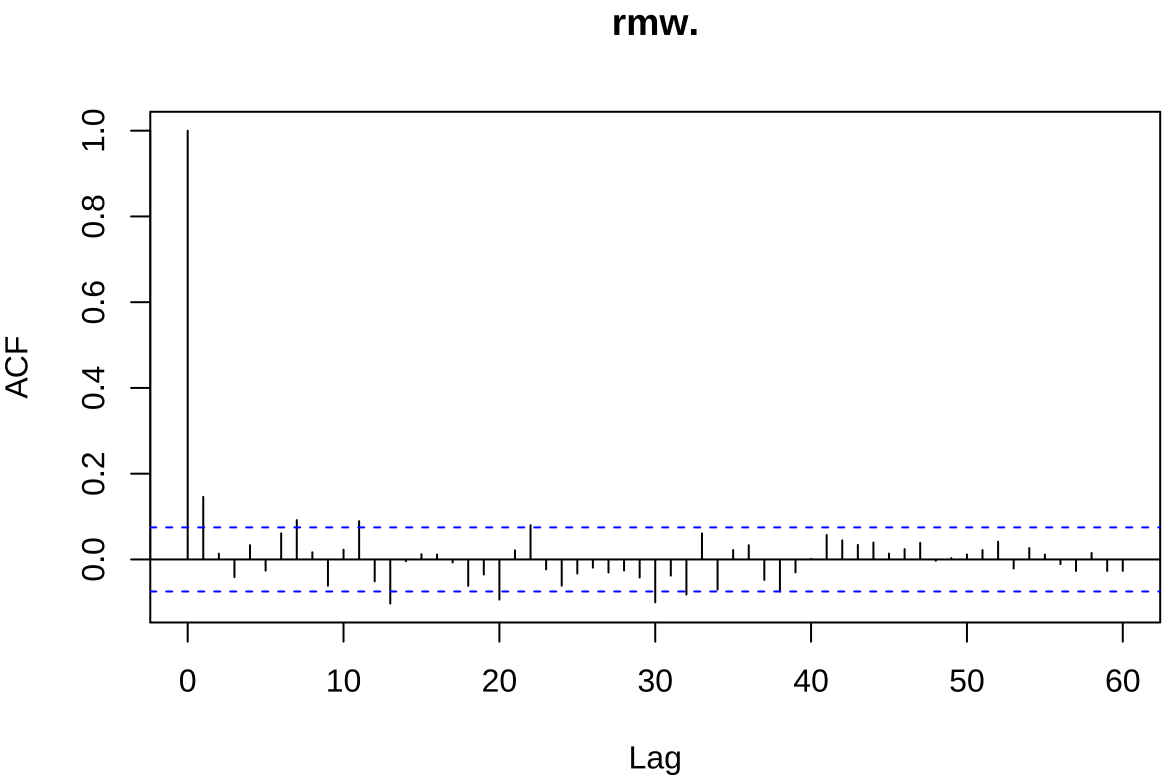
```
## [1] 3.079535
```

```
Annual_SD
```

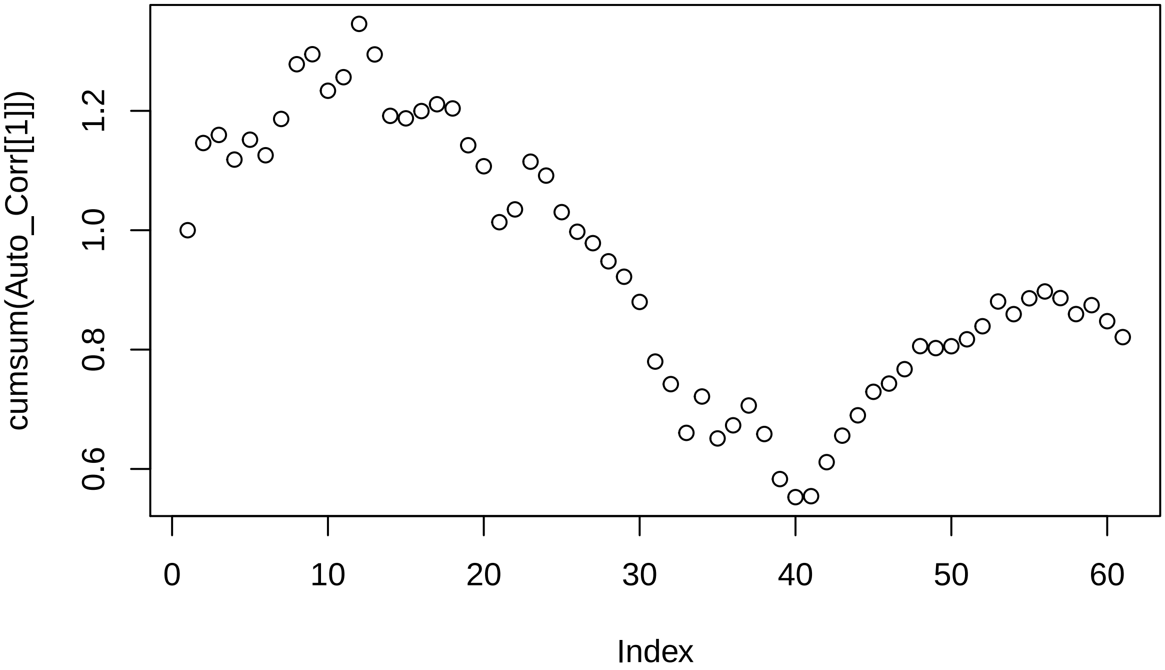
```
## [1] 7.500104
```

b) Plot the 1st to 60th order autocorrelations of rmw. Also plot the cumulative sum of these autocorrelations (that is, the 5th observation is the sum of the first 5 autocorrelations, the 11th observation is the sum of the first 11 autocorrelations, etc.). Describe these plots. In particular, do the plots hint at predictabilty of the factor returns? What are the salient patterns, if any
Solution:

```
#Autocorrelations till lag 60
Auto_Corr <- acf(rmw.,lag.max=60)
```



```
#Cumulative ACF plot
plot(cumsum(Auto_Corr[[1]]))
```



The first plot indicates the significance of autocorrelation of lag variables. From the first plot, we can see that the first lag variable is significant and after that, there are some lag variables in between that are significant (lag 7, lag 9, lag 11 etc.).

The second plot indicates presence of seasonality, like for how many lag variables, the correlation is cumulatively positive and the period after which it starts decreasing. In the second plot, we can see that the cumulative sum increases until lag variable of about 13 and then it decreases until about 40 and then again increases (salient patterns asked in the question).

These plots do hint at the predictability of factor returns, the first plot hints about how many lag variables need to be considered for regression and the second plot hints if the cumulative correlation act positively or negatively.

c) Perform a Ljung-Box test that the first 6 autocorrelations jointly are zero. Write out the form of the test and report the p-value. What do you conclude from this test?

Solution:

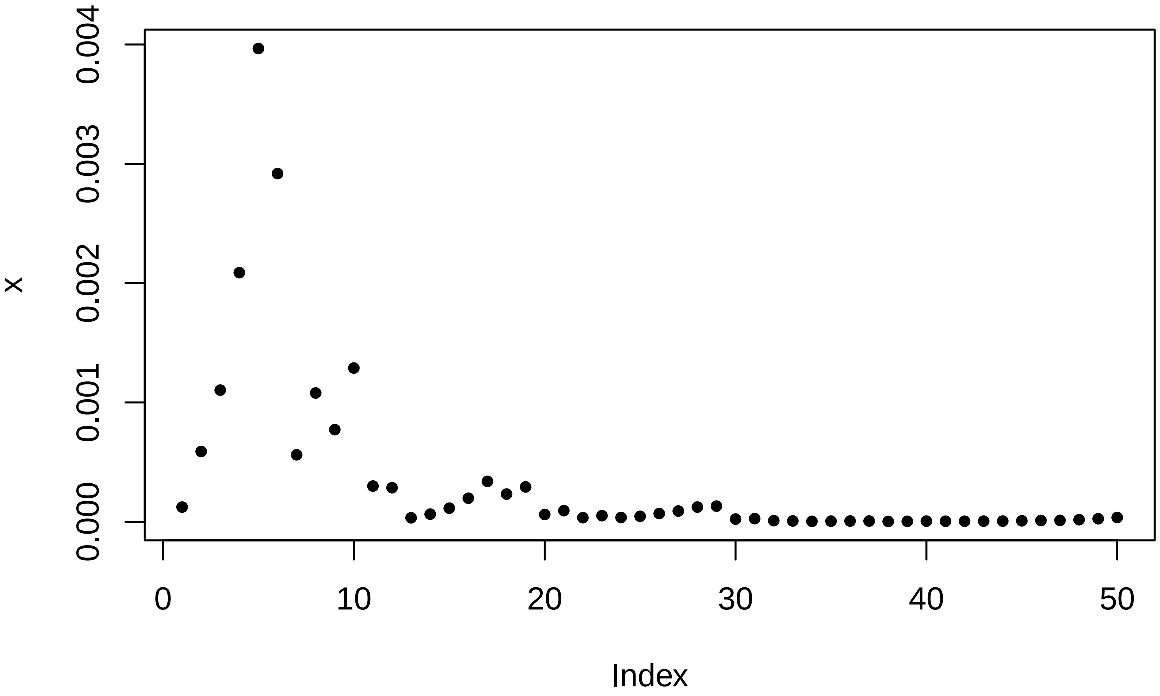
```
#Performing Box-Ljung test on 6 lagged AC
Box.test(rmw.,type="Ljung-Box",lag=6)
```

```
##
##  Box-Ljung test
##
## data:  rmw.
## X-squared = 19.872, df = 6, p-value = 0.002918
```

p-value for the Ljung-Box test above is about 0.0029 which signifies that with 99.71% confidence level, we can say that there is not much auto-correlation left in the lag variables when 6 lag variables are considered for regression.

d) Based on your observations in (2) and (3), propose a parsimonious forecasting model for rmw.
Solution:

```
x <- 0
for(i in 1:50){x[i] <- Box.test(rmw., lag = i, type = "Ljung-Box")$p.value}
plot(x, pch =20)
```



We feel taking 7 lag variables is ideal for the model. By plotting the autocorrelations graph, we see that the correlations until lag value of 6 is is insignificant and lag variable 7 has good correlation. Also, by plotting the p-value graph using Box-Ljung test, we can see that the p-value increases until lag of 5 and gradually decreases later. We see one sudden dip in p-value at lag value of 7 and then increase a little to further decrease after lag 10. Hence, after lag 10, we feel it is overfitting and hence would eliminate them. Because we see sudden dip in p-value at lag value 7 compared to its immediate neighborhood, we feel it is parsimonious to consider lag value until 7.

e) Estimate the proposed model. Report Robust (White) standard errors for β, as well as the regular OLS standard errors.
Solution:

```
library("sandwich")

#Regular OLS Standard Errors
Forecast_rmw <- lm(rmw. ~ back(rmw.) + back(rmw.,2) + back(rmw.,3) + back(rmw.,4) + back(rmw.,5) + back
lmSumm(Forecast_rmw)
```

```
## Multiple Regression Analysis:
##     8 regressors(including intercept) and 681 observations
##
## lm(formula = rmw. ~ back(rmw.) + back(rmw., 2) + back(rmw., 3) +
##     back(rmw., 4) + back(rmw., 5) + back(rmw., 6) + back(rmw.,
##     7), data = rmw.)
##
## Coefficients:
##                Estimate Std Error t value p value
## (Intercept)    0.193900   0.08518    2.28   0.023
## back(rmw.)     0.149700   0.03843    3.90   0.000
## back(rmw., 2) -0.003674   0.03882   -0.09   0.925
## back(rmw., 3) -0.051420   0.03887   -1.32   0.186
## back(rmw., 4)  0.057270   0.03882    1.48   0.141
## back(rmw., 5) -0.048800   0.03882   -1.26   0.209
## back(rmw., 6)  0.057650   0.03886    1.48   0.138
## back(rmw., 7)  0.079290   0.03852    2.06   0.040
## ---
## Standard Error of the Regression:  2.143
## Multiple R-squared:  0.038  Adjusted R-squared:  0.028
## Overall F stat: 3.82 on 7 and 673 DF, pvalue= 0
```

```
#Robust (White) Standard Errors

Robust_White <- coeftest(Forecast_rmw, vcov = vcovHC(Forecast_rmw, type="HC0"))
Robust_White
```

```
##
## t test of coefficients:
##
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.1939139  0.1072175  1.8086  0.07096 .
## back(rmw.)     0.1497340  0.1062934  1.4087  0.15939
## back(rmw., 2) -0.0036736  0.0716775 -0.0513  0.95914
## back(rmw., 3) -0.0514168  0.0750842 -0.6848  0.49371
## back(rmw., 4)  0.0572679  0.0793486  0.7217  0.47071
## back(rmw., 5) -0.0487957  0.0756406 -0.6451  0.51908
## back(rmw., 6)  0.0576519  0.0563451  1.0232  0.30658
## back(rmw., 7)  0.0792910  0.0633098  1.2524  0.21085
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Problem 2

1) Simulate T time series observations each of of the following two return series N times:

$$r_{1,t} = \mu + \sigma\epsilon_{1,t}$$
$$r_{2,t} = \mu + \sigma\epsilon_{2,t}$$

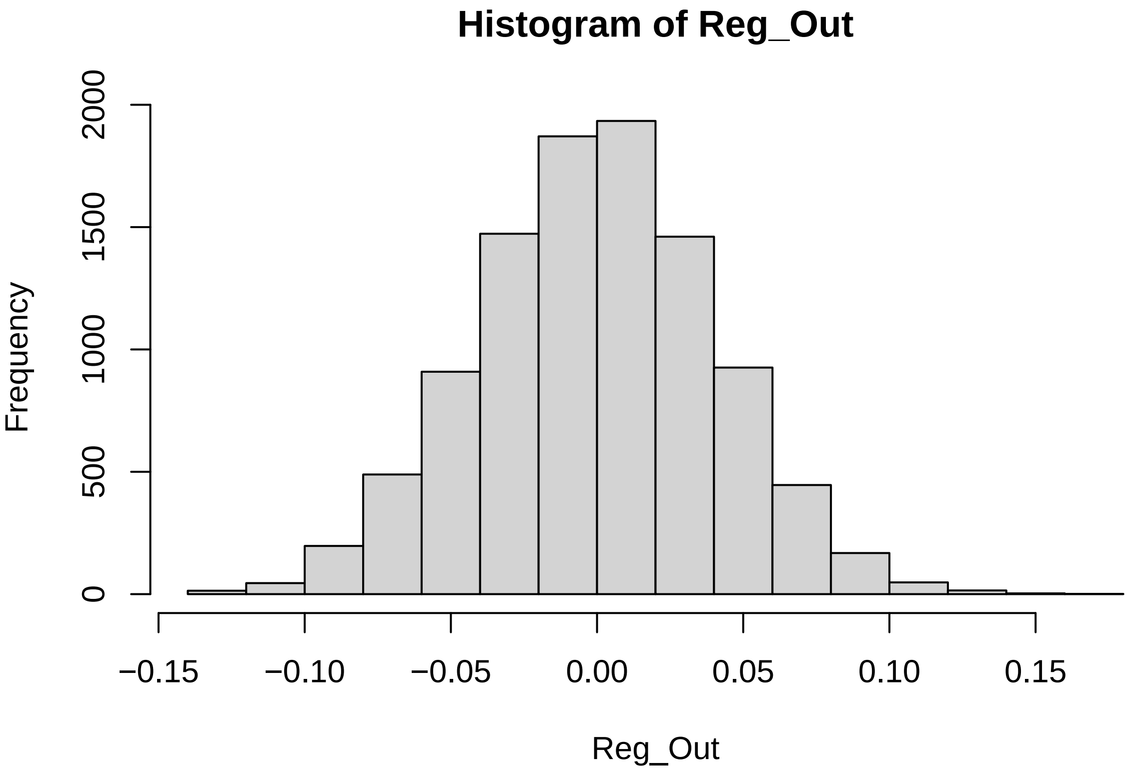
where μ = 0:5%, σ = 4%, and the residuals are uncorrelated standard Normals. Let T = 600 and N = 10000. For each of the N time-series, regress:

$$r_{1,t} = \alpha + \beta r_{2,t} + \epsilon_t$$

and save the slope coefficient as $\beta^{(n)}$, where n = 1,. . . ,N. Give the mean and standard deviation of across samples n and plot the histogram of the 10000 β's. Does this correspond to the null hypothesis = 0? Do the regress standard errors look ok?

Solution:

```
mu <- 0.005
sigma <- 0.04
Time <- 600
N <- 10000

r_1_t <- matrix(0,nrow=Time,ncol=N)
r_2_t <- matrix(0,nrow=Time,ncol=N)

eps_1_t <- double(Time)
eps_2_t <- double(Time)



for (i in 1:N) {
  eps_1_t <- rnorm(Time)
  eps_2_t <- rnorm(Time)
  r_1_t[,i] <- mu + sigma*eps_1_t
  r_2_t[,i] <- mu + sigma*eps_2_t
}

Reg_Out <- matrix(0,nrow=N,ncol=1,byrow=TRUE)

for (i in 1:N) {
  Reg_Out[i] <- (lm(r_1_t[,i] ~ r_2_t[,i]))$coefficient[2]
}

mean_beta <- mean(Reg_Out)
sd_beta <- sd(Reg_Out)

hist(Reg_Out)
```



```
#Checking whether null hypothesis beta=0 is true using t-statistic
t_beta=(mean_beta-0)/sd_beta
```

We get t-statistic as -0.008242198 whose absolute value is below the t value of 1.96 (at a 95% confidence interval). Hence, we can say that the value of beta is not statistically significant and fail to reject the null hypothesis that beta=0

By plotting the regression standard errors, we can see that it follows a normal distribution. We can see that the mean of regression standard error is about 0.04 which is equal to sigma. This is intuitive because the standard deviation in r is multiple of sigma and is the only source of deviation in the regression of r_1_t and r_2_t. Hence, we can verify that true value is almost equal to expected value. Also, we can see that Central Limit Theorem test can be applied and therefore we can verify that return process is stationary.

2) Next, construct N price sample of length T based on each return using:

$$p_{1,t} = p_{1,t-1} + r_{1,t}$$
$$p_{1,t} = p_{1,t-1} + r_{1,t}$$

using $p_{1,0} = p_{2,0} = 0$ as the initial condition. Now, repeat the regression exercise using the regression:

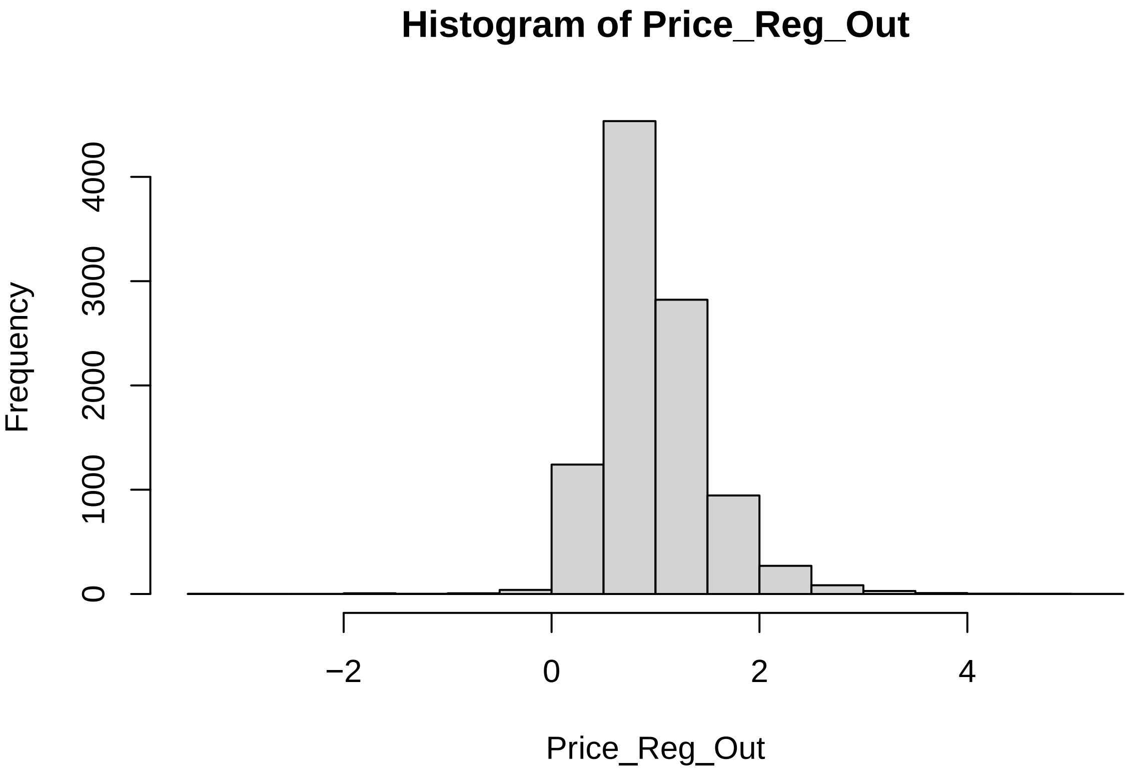
$$p_{1,t} = \alpha + \beta p_{2,t} + \epsilon_t$$

Again report the mean and standard deviation of the N estimated β's and plot the histogram. Does this correspond to the null hypothesis $\beta = 0$? Do the regression standard errors look ok? Explain what is going on here that is different from the previous return-based regressions.

Solution:

```
p_1_t <- matrix(0,nrow=Time,ncol=N)
p_2_t <- matrix(0,nrow=Time,ncol=N)
```

```r
for (i in 1:N) {
  for(j in 1:Time) {
    if(j==1) {
      p_1_t[j,i] <- 0 + r_1_t[j,i]
      p_2_t[j,i] <- 0 + r_2_t[j,i]
    }else{
      p_1_t[j,i] <- p_1_t[j-1,i] + r_1_t[j,i]
      p_2_t[j,i] <- p_2_t[j-1,i] + r_2_t[j,i]
    }

  }
}

Price_Reg_Out <- matrix(0,nrow=N,ncol=1,byrow=TRUE)

for (k in 1:N) {
  Price_Reg_Out[k] <- (lm(p_1_t[,k] ~ p_2_t[,k]))$coefficient[2]
}

hist(Price_Reg_Out)
```



```r
mean_beta_p <- mean(Price_Reg_Out)
sd_beta_p <- sd(Price_Reg_Out)

#Checking whether null hypothesis beta=0 is true using t-statistic
t_beta=(mean_beta_p-0)/sd_beta_p
```

We get t-statistic as 1.95 whose absolute value is very close to the t value of 1.96 (at a 95% confidence interval). If we compare the t-statistic with t values at lower confidence intervals (for example at 90% confidence interval t value is 1.64) it is far greater and hence, we can say that the value of beta is statistically significant at lower confidence intervals and reject the null hypothesis (that beta=0) at these levels.

In this second regression we notice that the value of p(1,t) is not completely independent of p(2,t) (since beta is not always zero) whereas in the previous return based regression, we observe that both r(1,t) and r(2,t) are independent. This could be because in the second regression, the current values of p are dependent on the previous values of p and it is likely that p(2,t) might have some values that are similar to p(1,t) and since the future values depend on the previous older values there could be some relationship between both the estimated values. In the previous returns regression, both the regressions are completely independent and depend majorly on the given mu values and hence it is very unlikely for them to have any dependence.

Also, we can see that Central Limit theorem test do not apply here as the plot β of p(1,t) and p(2,t) regression is slightly skewed. Hence, we can verify that price process is not stationary.