I. Objective

The aim of this project is to delve into regression and univariate analysis, demonstrating understanding of basic statistics and R code.

Data Source: JP Morgan stock historical prices from Yahoo Finance

Price considered in the analysis: Close price adjusted for dividends and splits

3.1.1 Basic Statistics

1. Calculate in R:
   1.1. Average stock value
   1.2. Stock volatility
   1.3. Daily stock return

3.1.2 Linear Regression

1. Implement a two-variable regression in R
   Explained variable: JP Morgan stock (adjusted close price)

3.1.3 Univariate Time Series Analysis

Forecast S&P/Case-Shiller U.S. National Home Price Index using an ARMA model.

Data source: https://fred.stlouisfed.org/series/CSUSHPINSA Period considered in the analysis

1. Implement the Augmented Dickey-Fuller Test for checking the existence of a unit root in Case-Shiller Index series
2. Implement an ARIMA(p,d,q) model. Determine p, d, q using Information Criterion or Box-Jenkins methodology. Comment results
3. Forecast the future evolution of Case-Shiller Index using the ARMA model. Test model using in sample forecasts

II. Introduction

Linear Regression- In statistics, linear regression is a linear approach to modeling the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables). The case of one explanatory variable is called simple **linear regression**.

Moving Average Trendline- A moving average trendline smoothes out fluctuations in data to show a pattern or trend more clearly. A moving average trendline uses a specific number of data points (set by the Period option), averages them, and uses the average value as a point in the trendline.

Explained Variable- Dependent Variable or response variable or outcome variable

Explanatory Variable- Independent Variable or regressor or predictor variable

Analysis ToolPak- The Analysis ToolPak is an Excel add-in program that provides data analysis tools for financial, statistical and engineering data analysis.

LINEST Function- The LINEST function uses the least squares regression method to calculate a straight line that best explains the relationship between the variables and returns an array describing that line.
LINEST function syntax

The syntax of the Excel LINEST function is as follows:

LINEST(known_y's , [known_x's], [CONST], [STATS])

Where:

- ***known_y's*** (required) is a range of the dependent y-values in the regression equation. Usually, it is a single column or a single row.
- ***known_x's*** (optional) is a range of the independent x-values. If omitted, it is assumed to be the array {1,2,3,...} of the same size as known_y's.
- ***const*** (optional) - a logical value that determines how the intercept (constant a) should be treated:
  If TRUE or omitted, the constant a is calculated normally.
  If FALSE, the constant a is forced to 0 and the slope (b coefficient) is calculated to fit y=bx.
- ***stats*** (optional) is a logical value that determines whether to output additional statistics or not:
  If TRUE, the LINEST function returns an array with additional regression statistics.
  If FALSE or omitted, LINEST only returns the intercept constant and slope coefficient(s).

Since LINEST returns an array of values, it must be entered as an array formula by pressing the Ctrl + Shift + Enter shortcut. If it is entered as a regular formula, only the first slope coefficient is returned

ARMA Model- In the statistical analysis of time series, autoregressive–moving-average (ARMA) models provide a parsimonious description of a (weakly) stationary stochastic process in terms of two polynomials, one for the autoregression (AR) and the second for the moving average (MA).

Given a time series of data Xt , the ARMA model is a tool for understanding and, perhaps, predicting future values in this series. The AR part involves regressing the variable on its own lagged (i.e., past) values. The MA part involves modeling the error term as a linear combination of error terms occurring contemporaneously and at various times in the past. The model is usually referred to as the ARMA(p,q) model where p is the order of the AR part and q is the order of the MA part.

ARIMA Model- ARIMA stands for autoregressive integrated moving average model. ARIMA (p, d, q) has both the AR and MA components in the equation.
• p – the order of the autoregressive model.
• d – the order required to make the variable stationary.
• q – the order of the moving average model.
It is a generalization of an autoregressive moving average model. Both of these models are fitted to time series data either to better understand the data or to predict future points in the series

Box Jenkins Method- Box - Jenkins Analysis refers to a systematic method of identifying, fitting, checking, and using integrated autoregressive, moving average (ARIMA) time series models.
The method is appropriate for time series of medium to long length (at least 50 observations)

Multiple R.-It is the ***Correlation Coefficient*** that measures the strength of a linear relationship between two variables. The correlation coefficient can be any value between -1 and 1, and its absolute value indicates the relationship strength. The larger the absolute value, the stronger the relationship:

- 1 means a strong positive relationship
- -1 means a strong negative relationship
- 0 means no relationship at all

***R Square*.-**It is the Coefficient of Determination, which is used as an indicator of the goodness of fit. It shows how many points fall on the regression line. The R2 value is calculated from the total sum of squares, more precisely, it is the sum of the squared deviations of the original data from the mean.

Adjusted R Square-It is the *R square* adjusted for the number of independent variable in the model. We will use this value instead of *R square* for multiple regression analysis.

Standard Error- It is another goodness-of-fit measure that shows the precision of your regression analysis - the smaller the number, the more certain we can be about your regression equation. While R square represents the percentage of the dependent variables variance that is explained by the model, Standard Error is an absolute measure that shows the average distance that the data points fall from the regression line.

Observations- It is simply the number of observations in the model.

df is the number of the degrees of freedom associated with the sources of variance
SS is the sum of squares. The smaller the Residual SS compared with the Total SS, the better the model fits the data
MS is the mean square
F is the F statistic, or F-test for the null hypothesis. It is used to test the overall significance of the model.
Significance F is the P-value of F

The **Significance F** value gives an idea of how reliable (statistically significant) the results are. If Significance F is less than 0.05 (5%), your model is OK. If it is greater than 0.05, we should probably better choose another independent variable.

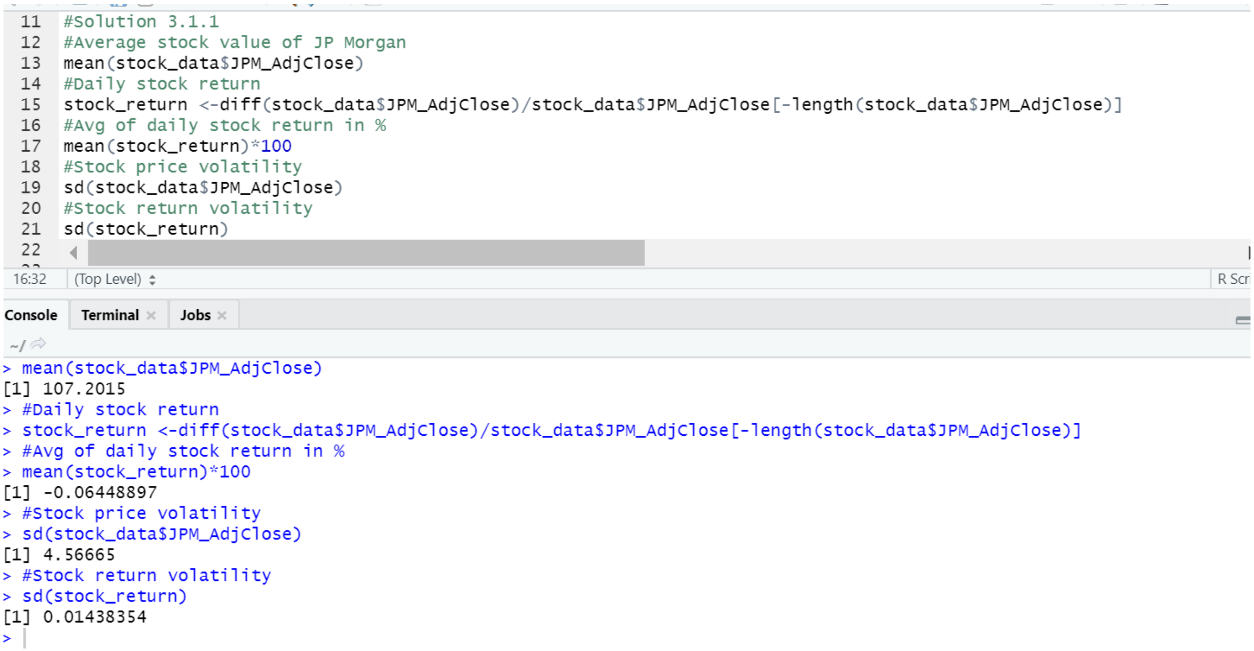
III. Results

3.1.1 Basic Stats Computation

R Results

Description	Command	Computed Value
Average stock value of JP Morgan	mean(stock_data$JPM_AdjClose)	107.2015
Daily Stock Return	diff(stock_data$JPM_AdjClose)/stock_data$JPM_AdjClose[-length(stock_data$JPM_AdjClose)]	
Average of Daily Stock Return	mean(stock_return)*100	-0.06448897
Stock Price volatility	sd(stock_data$JPM_AdjClose)	4.56665
Stock Return Volatility	sd(stock_return)	0.01438354

Result Snapshot



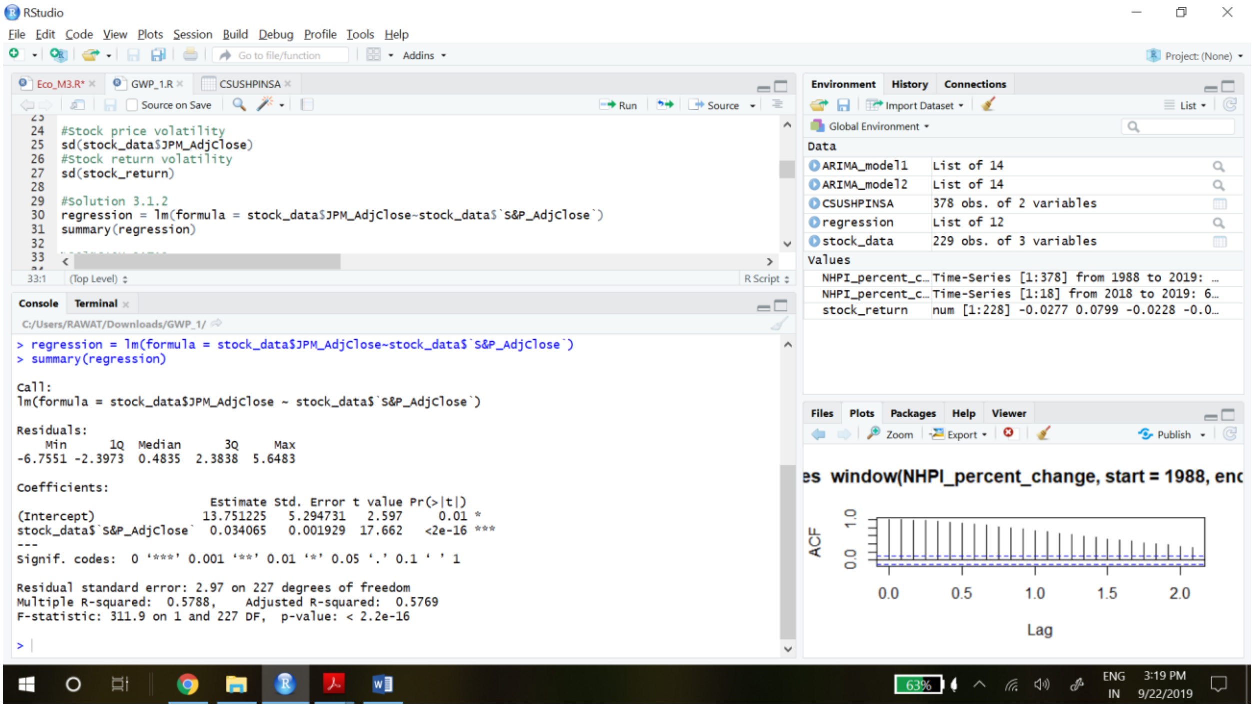
```
#Solution 3.1.1
#Average stock value of JP Morgan
mean(stock_data$JPM_AdjClose)
#Daily stock return
stock_return <-diff(stock_data$JPM_AdjClose)/stock_data$JPM_AdjClose[-length(stock_data$JPM_AdjClose)]
#Avg of daily stock return in %
mean(stock_return)*100
#Stock price volatility
sd(stock_data$JPM_AdjClose)
#Stock return volatility
sd(stock_return)
```

```
> mean(stock_data$JPM_AdjClose)
[1] 107.2015
> #Daily stock return
> stock_return <-diff(stock_data$JPM_AdjClose)/stock_data$JPM_AdjClose[-length(stock_data$JPM_AdjClose)]
> #Avg of daily stock return in %
> mean(stock_return)*100
[1] -0.06448897
> #Stock price volatility
> sd(stock_data$JPM_AdjClose)
[1] 4.56665
> #Stock return volatility
> sd(stock_return)
[1] 0.01438354
> 
```

Basic Stats computed from R

3.1.2 Linear Regression

1. Implement a two-variable regression in R.



```
#Stock price volatility
sd(stock_data$JPM_AdjClose)
#Stock return volatility
sd(stock_return)

#Solution 3.1.2
regression = lm(formula = stock_data$JPM_AdjClose~stock_data$`S&P_AdjClose`)
summary(regression)
```

```
> regression = lm(formula = stock_data$JPM_AdjClose~stock_data$`S&P_AdjClose`)
> summary(regression)

Call:
lm(formula = stock_data$JPM_AdjClose ~ stock_data$`S&P_AdjClose`)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.7551 -2.3973  0.4835  2.3838  5.6483 

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)    
(Intercept)               13.751225   5.294731   2.597     0.01 *  
stock_data$`S&P_AdjClose`  0.034065   0.001929  17.662   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.97 on 227 degrees of freedom
Multiple R-squared:  0.5788,	Adjusted R-squared:  0.5769 
F-statistic: 311.9 on 1 and 227 DF,  p-value: < 2.2e-16

> 
```

```
Data
ARIMA_model1       List of 14
ARIMA_model2       List of 14
CSUSHPINSA         378 obs. of 2 variables
regression         List of 12
stock_data         229 obs. of 3 variables
Values
NHPI_percent_c…    Time-Series [1:378] from 1988 to 2019: …
NHPI_percent_c…    Time-Series [1:18] from 2018 to 2019: 6…
stock_return       num [1:228] -0.0277 0.0799 -0.0228 -0.0…
```

Regression analysis output: Summary Output

In our case, R2 is 0.57 (rounded to 2 digits), which is not good. It means that only 57% of our values fit the regression analysis model. In other words, 57% of the dependent variables (y-values ie JP Morgan stock (adjusted close price)) are explained by the independent variables (x-values i.e. SP500). Generally, R Squared of 95% or more is considered a good fit.

Regression analysis output: ANOVA

Basically, it splits the sum of squares into individual components that give information about the levels of variability within the regression model:

Regression analysis output: coefficients

This section provides specific information about the components of the analysis:

- The most useful component in this section is Coefficients. It enables you to build a linear regression equation in Excel***: y=bx + a***
- For our data set, where y is the number of umbrellas sold and x is an average monthly rainfall, our linear regression formula goes as follows:
- ***Y(JP Morgan_Adj Close) = SP500_ Adj Close* x + Intercept***
- Equipped with a and b values rounded to three decimal places, it turns into: ***Y=0.034*x+13.751***

Regression analysis output: residuals

If we compare the estimated and actual ***JP Morgan_Adj Close*** prices corresponding to the SP 500 Adj Close Price of **2821.97998**, we will see that these numbers are slightly different:

- Estimated: 109.883064149983 (calculated in Excel)
- Actual: 112.262558 (row 2 of the source data)

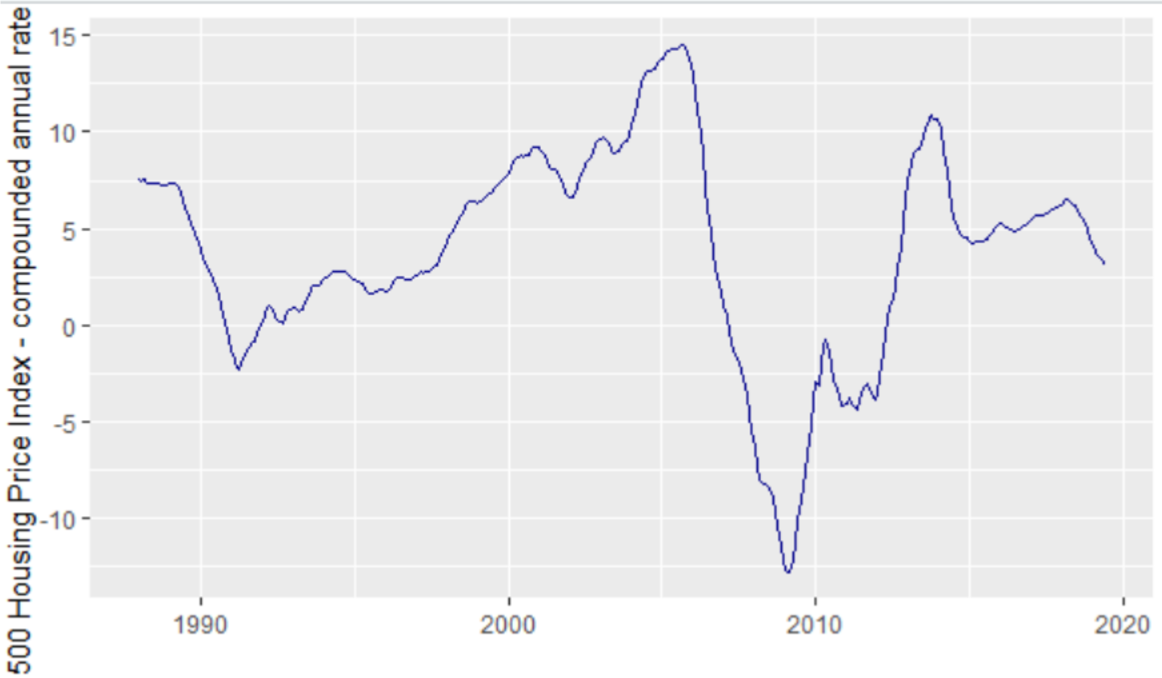
The difference is because independent variables are never perfect predictors of the dependent variables. And the residuals can help us understand how far away the actual values are from the predicted values:

For the first data point (SP500 Adj Close Price of **2821.97998**), the residual is approximately -**2.37949385001745**. So, we add this number to the predicted value, and get the actual value: **109.883064149983 + 2.37949385001745 = 112.262558**

3.1.3 Univariate Time Series Analysis

Forecast S&P/Case-Shiller U.S. National Home Price Index using an ARMA model.

We would first Visualise the series to identify type of time series to be considered.

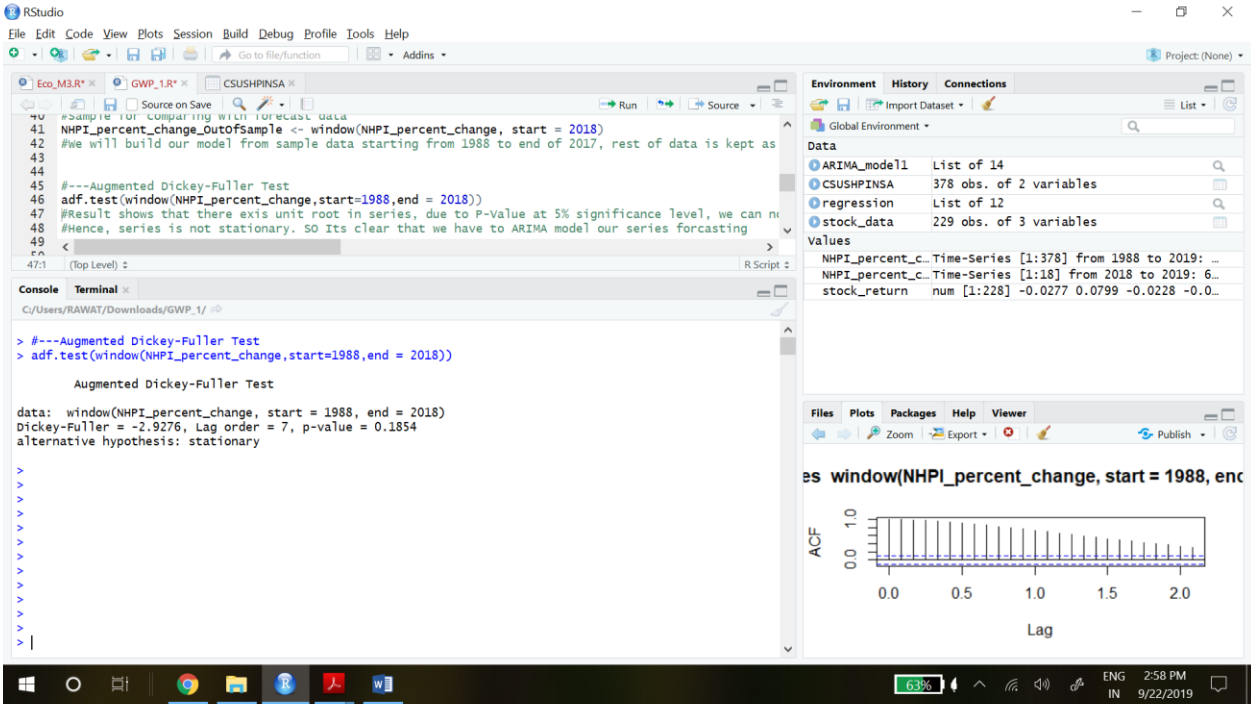


This series, viewed as a whole, appears quite non-stationary. There is no obvious mean over the whole sample. If we apply real world knowledge of what we intend to model we know in 2008 Global fiancial crisis hit US and housing prices tanked. Graph depicts the same.

In 2008 it has a huge dip in the housing prices and post it got recovered.

Implement the Augmented Dickey-Fuller Test for checking the existence of a unit root in Case-Shiller Index series

We will build our model from sample data starting from 1988 to end of 2017, rest of data is kept as Out of sample (i.e. start of 2018 to June 2019)



```
#Sample for comparing with forecast data
NHPI_percent_change_OutOfSample <- window(NHPI_percent_change, start = 2018)
#We will build our model from sample data starting from 1988 to end of 2017, rest of data is kept as


#---Augmented Dickey-Fuller Test
adf.test(window(NHPI_percent_change,start=1988,end = 2018))
#Result shows that there exis unit root in series, due to P-Value at 5% significance level, we can n
#Hence, series is not stationary. SO Its clear that we have to ARIMA model our series forcasting
```

C:/Users/RAWAT/Downloads/GWP_1/

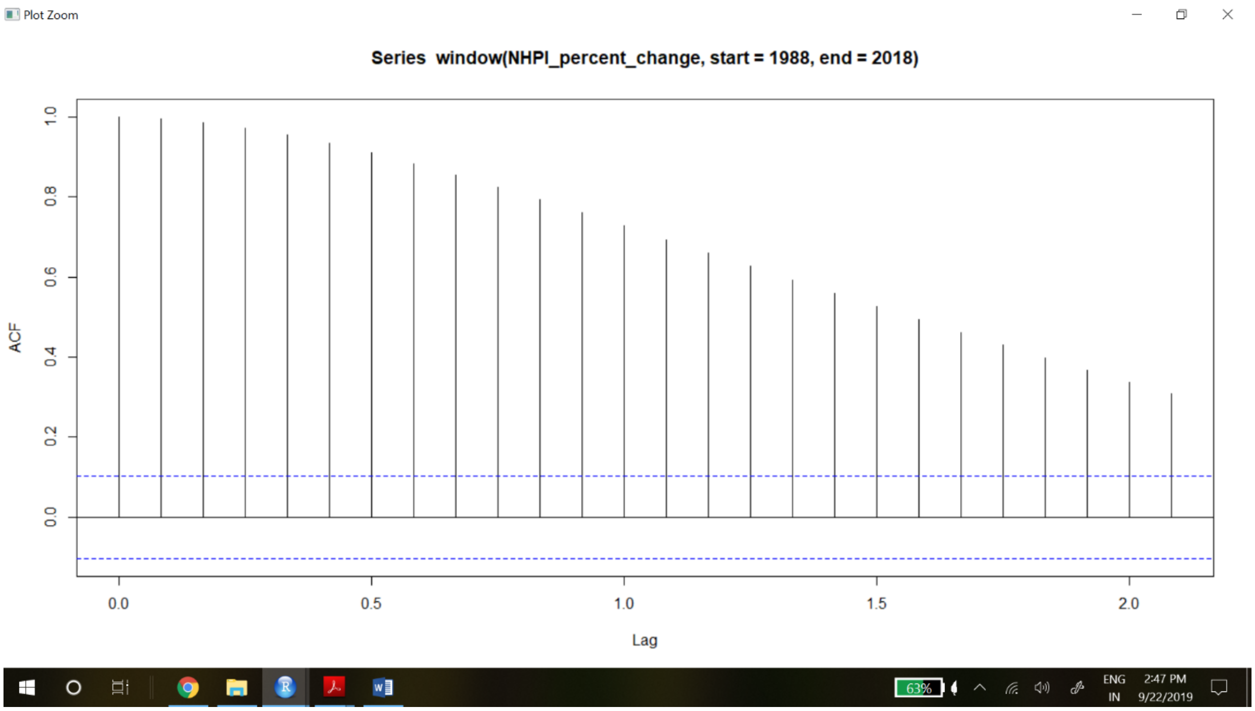
```
> #---Augmented Dickey-Fuller Test
> adf.test(window(NHPI_percent_change,start=1988,end = 2018))

	Augmented Dickey-Fuller Test

data:  window(NHPI_percent_change, start = 1988, end = 2018)
Dickey-Fuller = -2.9276, Lag order = 7, p-value = 0.1854
alternative hypothesis: stationary

>
>
>
>
>
>
>
>
>
>
>
>
>
> |
```

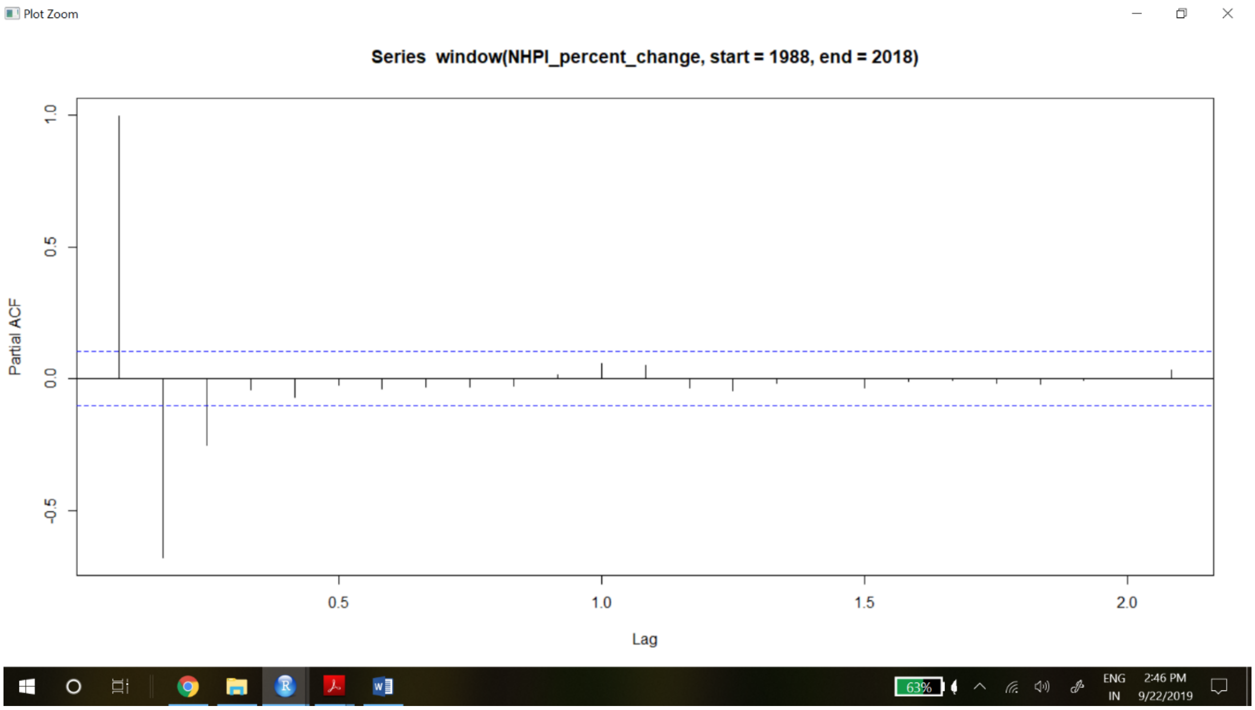
```
Data
ARIMA_model1          List of 14
CSUSHPINSA            378 obs. of 2 variables
regression            List of 12
stock_data            229 obs. of 3 variables
Values
NHPI_percent_c…       Time-Series [1:378] from 1988 to 2019: …
NHPI_percent_c…       Time-Series [1:18] from 2018 to 2019: 6…
stock_return          num [1:228] -0.0277 0.0799 -0.0228 -0.0…
```

Above result shows that there exist unit root in series. Due to P-Value at 5% significance level, we can not reject Null hypothesis, hence, series is not stationary.

Implement an ARIMA(p,d,q) model. Determine p, d, q using Information Criterion or Box-Jenkins methodology.

We will first analyse ACF and PACF chart of series. Below are the ACF and PACF charts of our sample data.



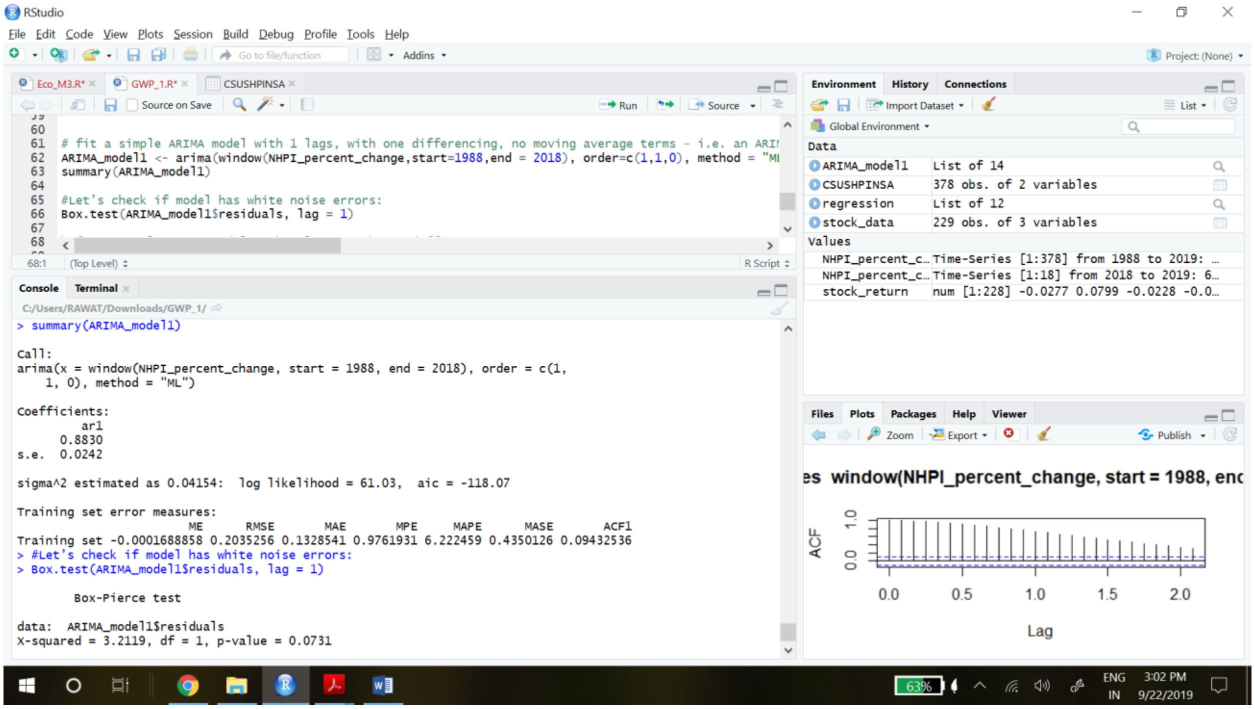


The ACF of the full-time path of the index variable is typical of a non-stationary process: auto-correlations are very close to 1 and slowly fade. PACF showing there has strong impact of the first lag – notice that pacf() by default drops the zero-*th* lag.
According to pattern of ACF and PACF, its shows sign that series has AR signature, we can take ARIMA (1,1,0) for our starting point (fitting and evaluating process)

Fitting & Evaluating the Model

Here we fit a number of models to the Price Index process from 1988 to 2018 to allow some room to explore its forecasting performance. We saw in the ACF and PACF in that there is significant autocorrelation for many lags.

Let’s Estimate with ARIMA(1,1,0)



```
# fit a simple ARIMA model with 1 lags, with one differencing, no moving average terms - i.e. an ARI
ARIMA_model1 <- arima(window(NHPI_percent_change,start=1988,end = 2018), order=c(1,1,0), method = "MI
summary(ARIMA_model1)

#Let's check if model has white noise errors:
Box.test(ARIMA_model1$residuals, lag = 1)
```

```
> summary(ARIMA_model1)

Call:
arima(x = window(NHPI_percent_change, start = 1988, end = 2018), order = c(1,
    1, 0), method = "ML")

Coefficients:
         ar1
      0.8830
s.e.  0.0242

sigma^2 estimated as 0.04154:  log likelihood = 61.03,  aic = -118.07

Training set error measures:
                        ME      RMSE       MAE       MPE     MAPE      MASE       ACF1
Training set -0.0001688858 0.2035256 0.1328541 0.9761931 6.222459 0.4350126 0.09432536
> #Let's check if model has white noise errors:
> Box.test(ARIMA_model1$residuals, lag = 1)

        Box-Pierce test

data:  ARIMA_model1$residuals
X-squared = 3.2119, df = 1, p-value = 0.0731
```

Data

ARIMA_model1	List of 14
CSUSHPINSA	378 obs. of 2 variables
regression	List of 12
stock_data	229 obs. of 3 variables

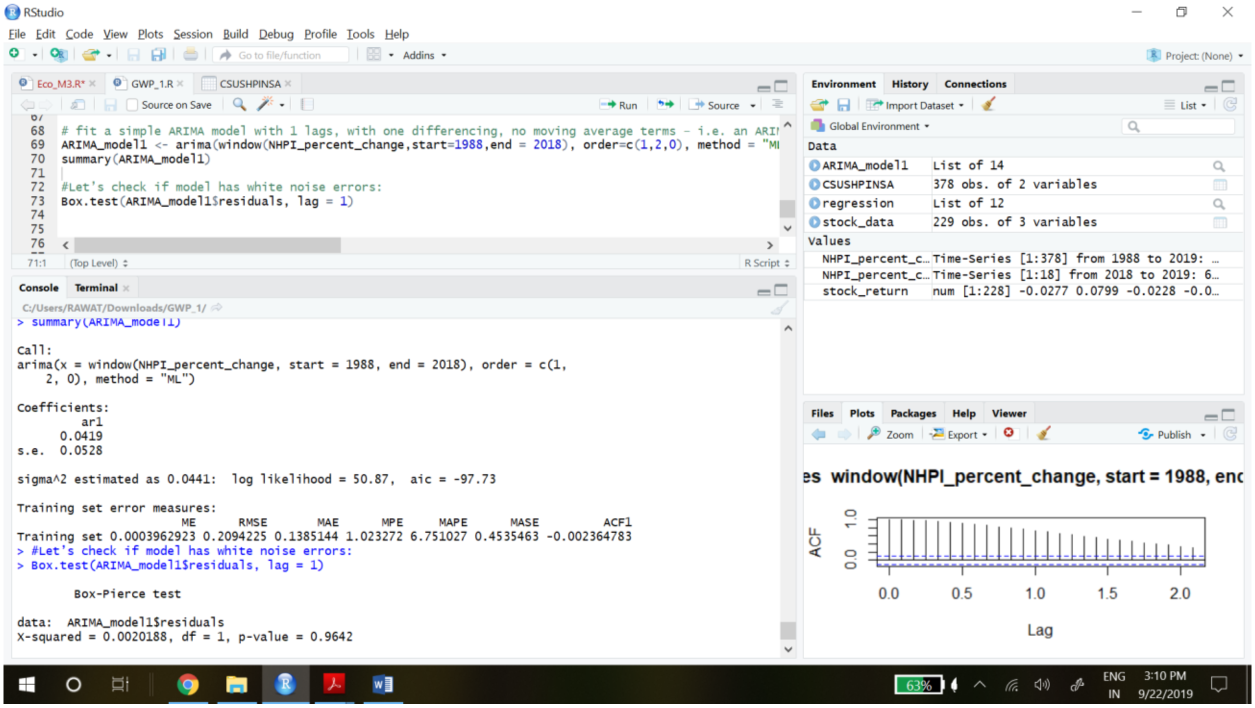
Values

NHPI_percent_c…	Time-Series [1:378] from 1988 to 2019: …
NHPI_percent_c…	Time-Series [1:18] from 2018 to 2019: 6…
stock_return	num [1:228] -0.0277 0.0799 -0.0228 -0.0…

Result shows that coeffiecient is not statistical significange, as difference between the estimated variable and model variable is more than two times of standard deviation (i.e. at 5% significance level or at 95% confidence level). So this is not a parsimonious model.

We have also checked if model has still white noise errors after One level differencing through Box-Pierce test. The Box-Pierce test is one of a range of tests for serial correlation with Null hypothesis that there is no auto correlation between residuals. In this case, the probability value is more that 5% significance value . Thus, we clearly accept the hypothesis that the errors are white noise.

But still we have to increase our fit of the model , So we turned to ARIMA (1,2,0) model to see if more differencing will improve the fit of the model:



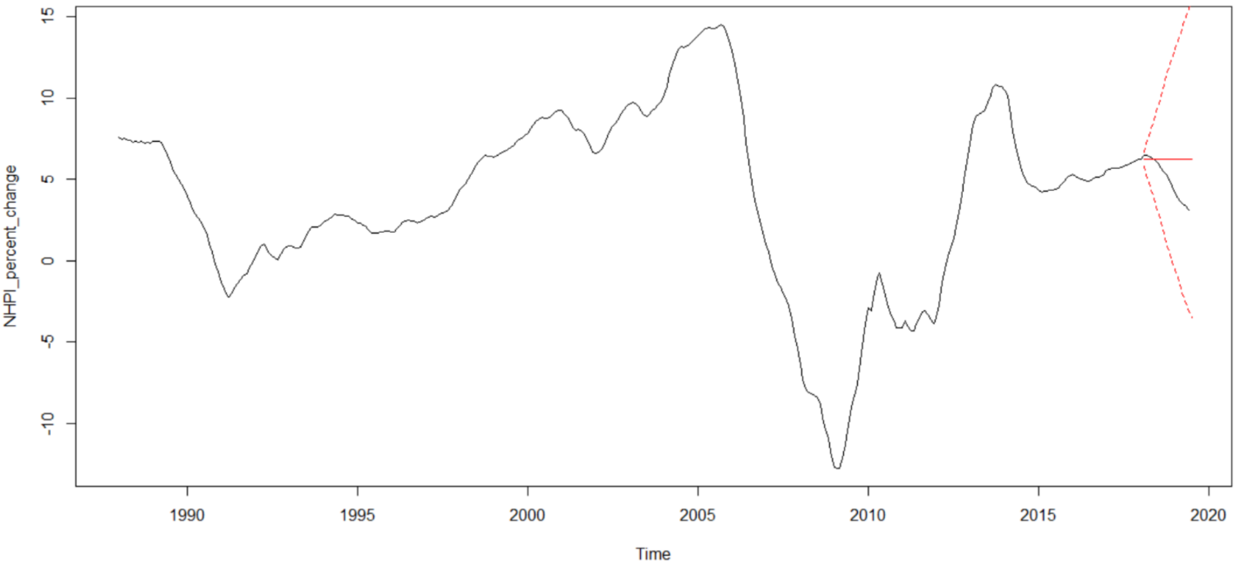
```
# fit a simple ARIMA model with 1 lags, with one differencing, no moving average terms - i.e. an ARI
ARIMA_model1 <- arima(window(NHPI_percent_change,start=1988,end = 2018), order=c(1,2,0), method = "MI
summary(ARIMA_model1)

#Let's check if model has white noise errors:
Box.test(ARIMA_model1$residuals, lag = 1)
```

```
> summary(ARIMA_model1)

Call:
arima(x = window(NHPI_percent_change, start = 1988, end = 2018), order = c(1,
    2, 0), method = "ML")

Coefficients:
         ar1
      0.0419
s.e.  0.0528

sigma^2 estimated as 0.0441:  log likelihood = 50.87,  aic = -97.73

Training set error measures:
                       ME      RMSE       MAE      MPE     MAPE      MASE         ACF1
Training set 0.0003962923 0.2094225 0.1385144 1.023272 6.751027 0.4535463 -0.002364783
> #Let's check if model has white noise errors:
> Box.test(ARIMA_model1$residuals, lag = 1)

        Box-Pierce test

data:  ARIMA_model1$residuals
X-squared = 0.0020188, df = 1, p-value = 0.9642
```

```
Data
ARIMA_model1        List of 14
CSUSHPINSA          378 obs. of 2 variables
regression          List of 12
stock_data          229 obs. of 3 variables
Values
NHPI_percent_c…     Time-Series [1:378] from 1988 to 2019: …
NHPI_percent_c…     Time-Series [1:18] from 2018 to 2019: 6…
stock_return        num [1:228] -0.0277 0.0799 -0.0228 -0.0…
```

Above result shows that now the coefficients are statistically significant. The residuals tests as white noise at a 5% level. So we get our better model for forecast , there might be a chance that there would be some model which can be more effiecent from this.

Forecast the future evolution of Case-Shiller Index using the ARMA model. Test model using in-sample forecasts



IV. Bibliography/References

- https://www.diva-portal.org/smash/get/diva2:725091/FULLTEXT01.pdf
- https://www.fanniemae.com/resources/file/research/housingsurvey/pdf/2015nabemennisaward.pdf
- https://pdfs.semanticscholar.org/5ad0/afbbafb06ec4cf618dcb66fbc64401957e76.pdf
- https://www.investopedia.com/ask/answers/correlation-inflation-houses.asp
- http://www.ugr.es/~scarbo/Mortgage.pdf