Dow Jones Share Market Index evaluation

Submitted by: Karun Dhingra

Time Series Analysis Final Project Under Prof. Jerome Busemeyer

Abstract:

This report covers the idea for predicting the percent change in the next month price using the predictors like open price for the market, closing price and the volume of shares. In predicting stock prices you collect data over some period of time - day, week, month, etc. But you cannot take advantage of data from a time period until the next increment of the time period. For example, assume you collect data daily. When Monday is over you have all of the data for that day. However you can invest on Monday, because you don't get the data until the end of the day. You can use the data from Monday to invest on Tuesday. In this research each record (row) is data for a month. Each record also has the percentage of return that stock has in the following week (percent_change_next_month_price).

Introduction:

The two main purpose of the Time series analysis: (a) identifying the nature of the phenomenon represented by the sequence of observations, and (b) forecasting (predicting future values of the time series variable). Both of these goals require that the pattern of observed time series data is identified and more or less formally described. Once the pattern is established, we can interpret and integrate it with other data (i.e., use it in our theory of the investigated phenomenon, e.g., seasonal commodity prices.

We used the data from the Dow Jones Market Index for the listed companies and did our analysis using R. Here the data for the open and close is in $ and the volume is the quantity for the stock for a particular company. Percent change for next month price is our target variable for which we need to find any pattern or trend and see if the analysis is dependent on the predictors like the open, close and volume of these shares.

Dataset:

Dataset consist of Dow Jones index monthly record for the open, close, Volume of the 30 listed companies like Apple, Goldman Sachs, and Cisco etc.

Predictors

Open: the price of the stock at the beginning of the week

Close: the price of the stock at the end of the week

Volume: the number of shares of stock that traded hands in the week

Response Variable

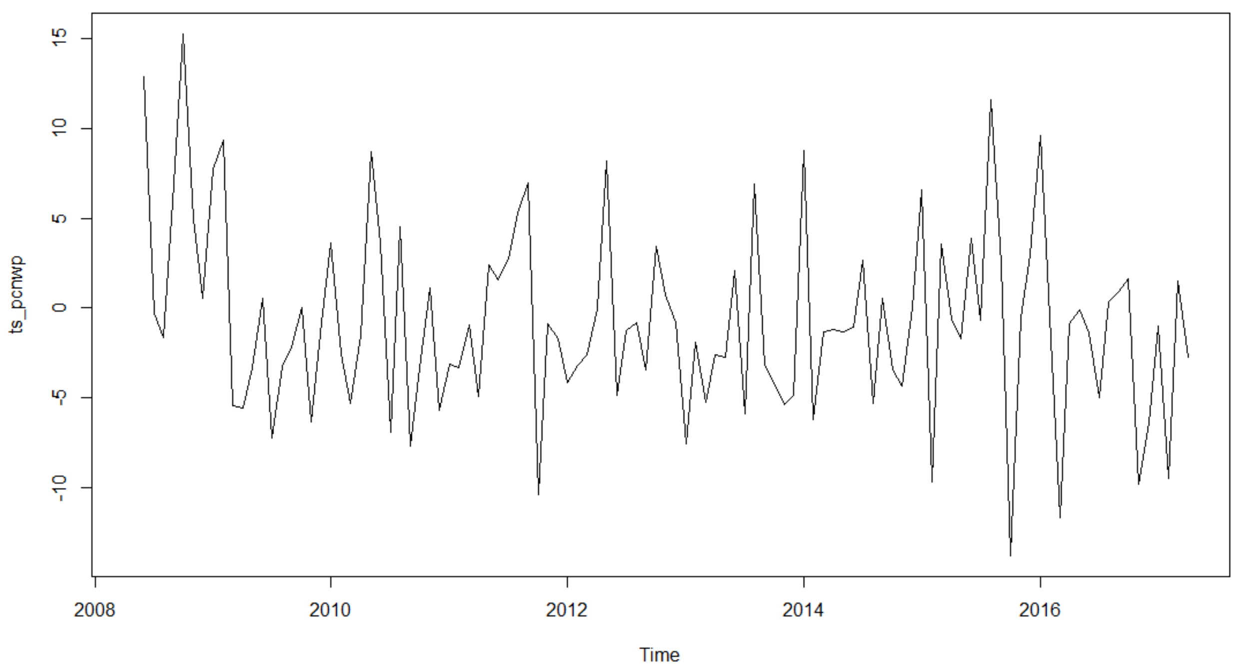
percent_change_next_weeks_price: the percentage change in price of the stock in the following week

Structure:

We will first review techniques used to identify patterns in time series data (such as smoothing and curve fitting techniques and autocorrelations), then we will introduce a general class of models that can be used to represent time series data and generate predictions (autoregressive and moving average models). Finally, we will review some simple but commonly used modeling and forecasting techniques based on linear regression.

Trends for the variables:

First we will go with the response variable on the time series plot from May'2008 to Mar'2016. We use the time series plot in R to represent the figure.

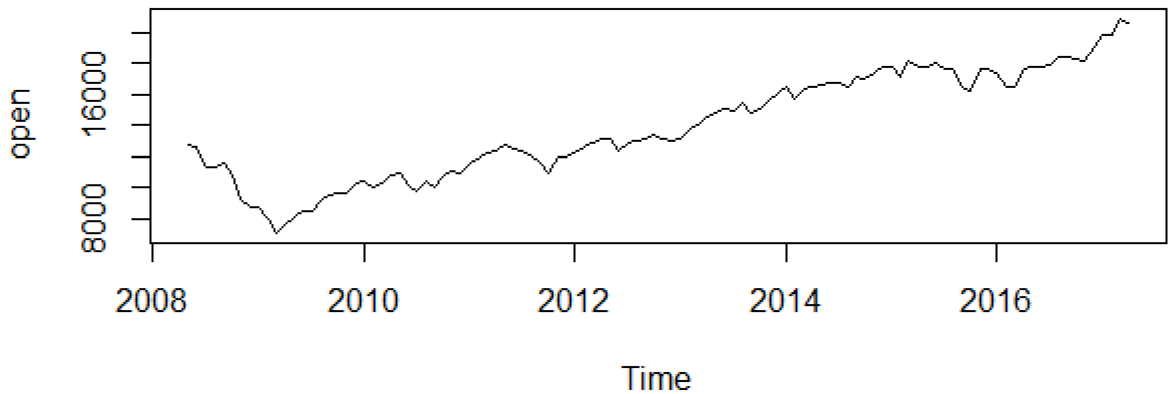


Analysis

We can see how in 2008 the recession has its own effect in the time series data. Clearly the change in the percent change in the price for the next month is on a higher scale than the rest of the years, except for the 2016 recession in the Chinese market which brought too much variation in the response variable.

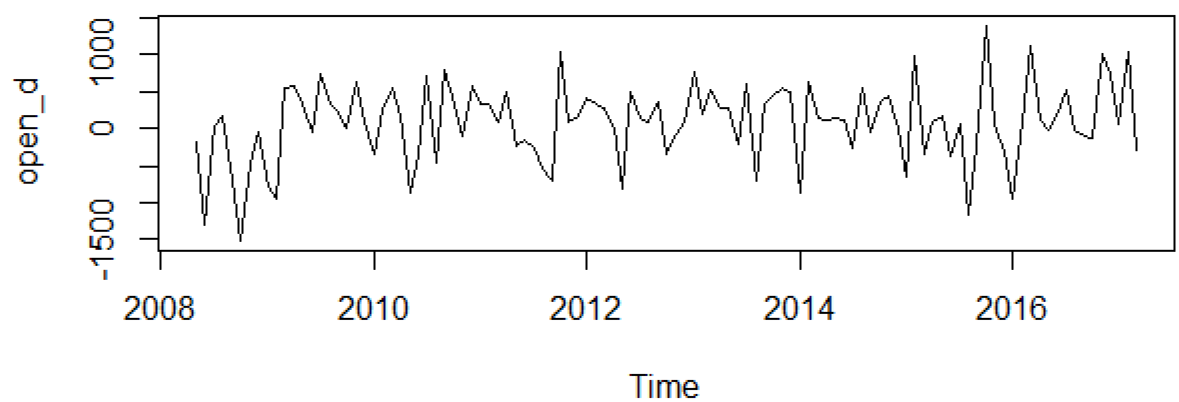
Predictors:

Opening price of the market monthly has a trend that you can see. AnaA

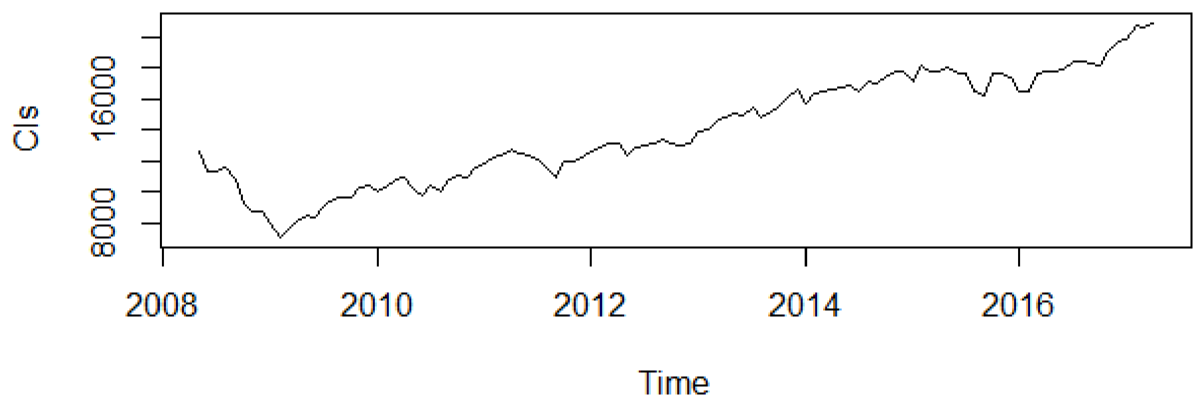


Analysis

We can there is a rising trend which is kind of linear after the 2009, ofcourse the behaviour before that has a recession impact. We need to get rid of the linear trend hence we make use of the first order difference of the predictor variable.

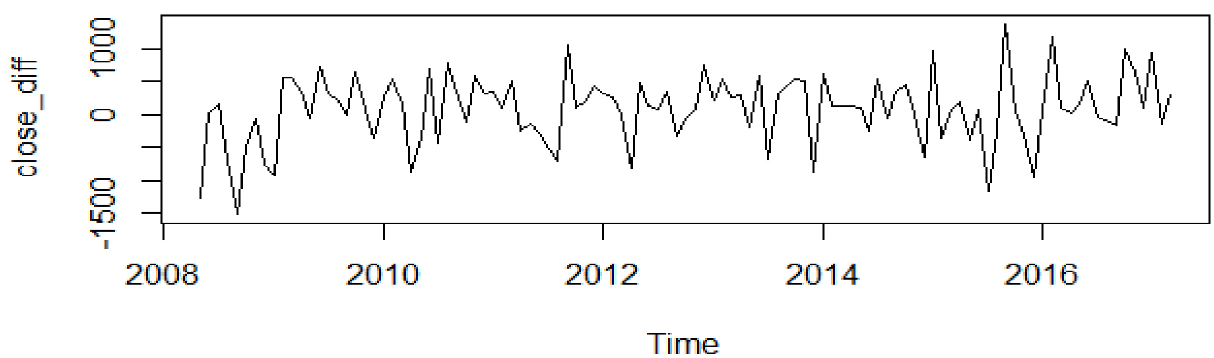


Close:

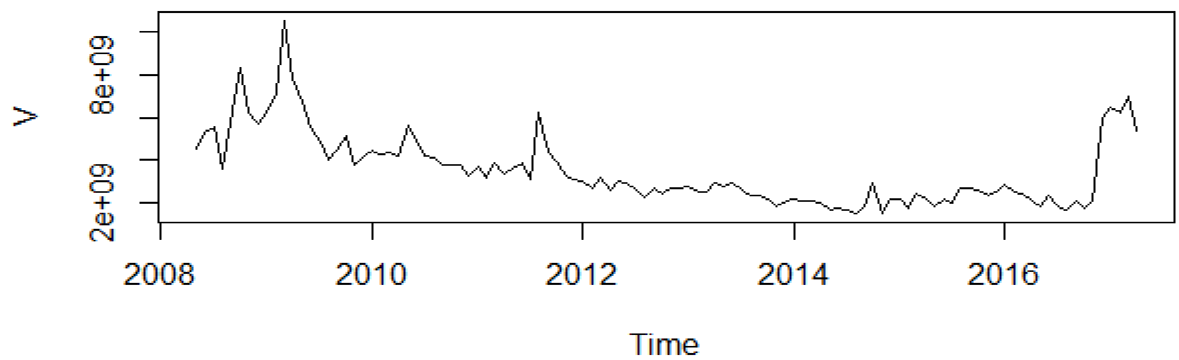


Analysis

We can there is a same rising trend which is kind of linear after the 2009, ofcourse the behaviour before that has a recession impact. Closing trend has the same behaviour as the open.

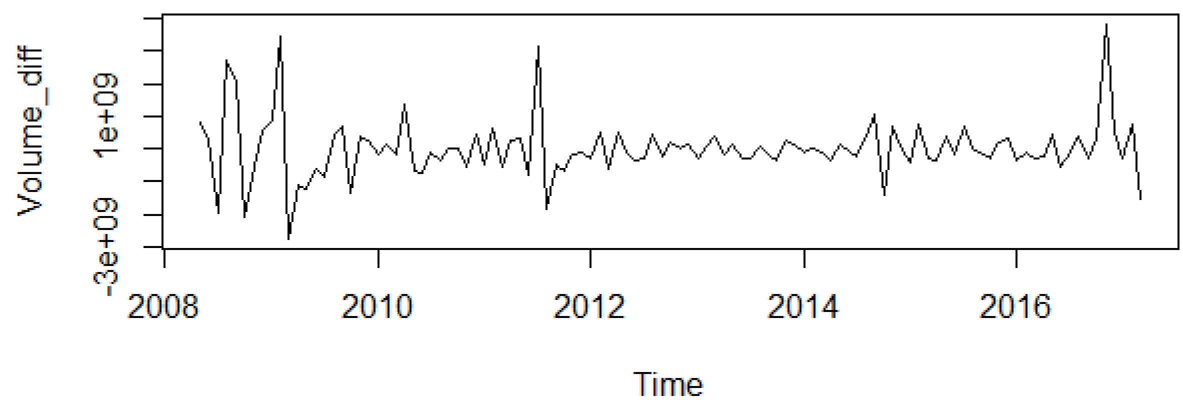


Volume:



Analysis

The trend in the volume is opposite to the one in open and close as the volume of the shares kept decreasing in a sequence, we see the upset of the trend on 2008.



On the difference of the volume of the actual volume data we see spikes in the 2017 and 2011. Couldn't find any substantial reason for the spike for these two years.

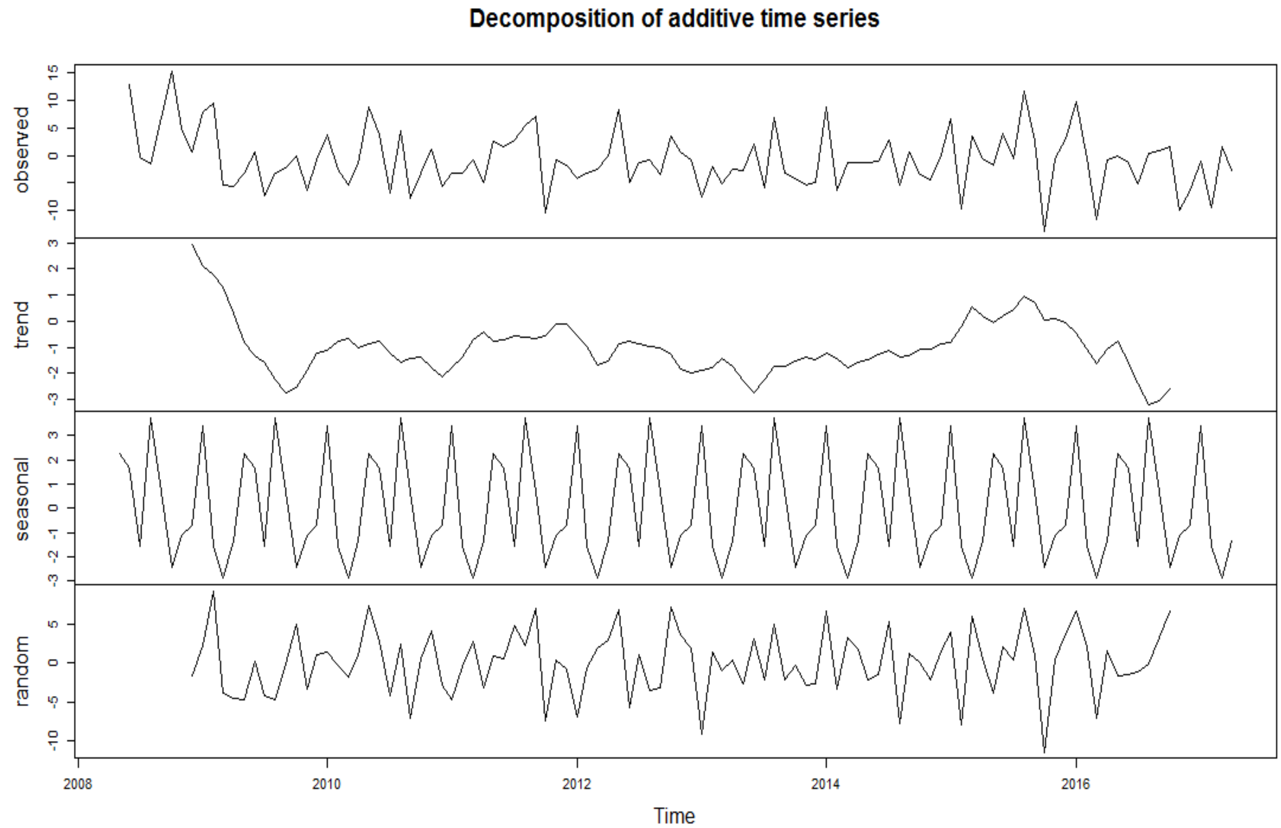
Response Variable Analysis:

For find out the seasonal trend and the regular trend in the percent change in next month price we need to find a method for the decomposition of the time series. We used the decompose function in R for the same but only after we first make the time series read in as a seasonal pattern of 12 months.

ts_pcnwp = ts(DJ_monthly$percent_change_next_weeks_price, frequency = 12, start = c(2008,5))

dec_pcnwp = decompose(ts_pcnwp, "additive")

pcnwp = ts_pcnwp - dec_pcnwp$trend - dec_pcnwp$seasonal



The random or the residual of this decomposed is the part of the time series that we want to continue our analysis with as it does have the trend(both seasonal and regular) have been removed from the time series.

We can see trend following a dip at the start of the 2008 may, and slowly follows sine wave.

Jan	Feb	Mar	Apr	May	Jun	Jul
3.3735338	-1.5532782	-2.9003245	-1.3460166	2.2340609	1.6280068	-1.5439576

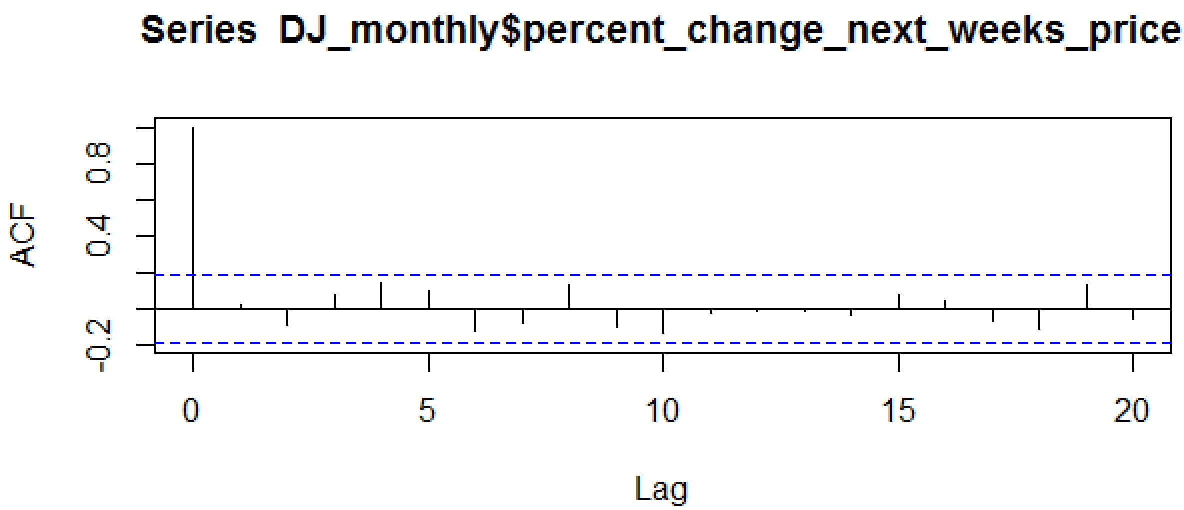
Aug	Sep	Oct	Nov	Dec
3.718247	0.6745401	-2.4331369	-1.1417507	-0.7099245

August has been even more of a headache for investors more recently. Since 1987, the Dow Jones industrial average has posted average losses of 1.1% in August, making it the worst month for equities in the past 30 years, according to Jeff Hirsch, editor of the Stock Trader's Almanac.

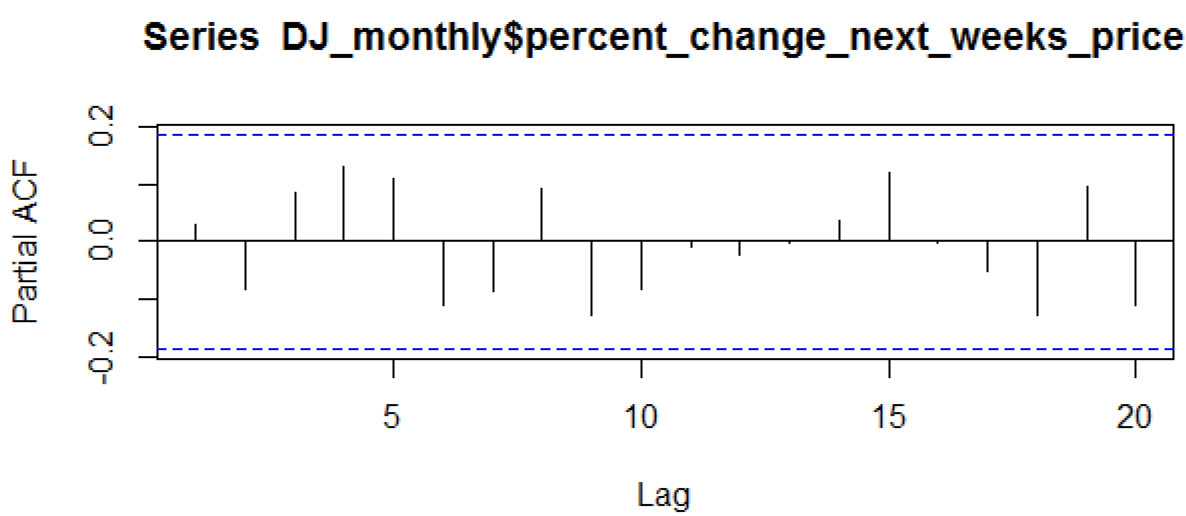
Correlation:

For time series data it is good to take note of the behavior of the correlation to see how the lag plays a role on determining good prediction for the time series.

We will go with the ACF or the autocorrelation function and PACF (Partial autocorrelation function).



The ACF plot shows the lag is pretty much under the significant level but we can see that the wave closing in at 4 going negative.



Model Selection:

First we will go with the ARIMA model for only the response variable without the predictors and see how well it can fit into the ARIMA model.

AR (autoregressive) model is usually used to model a time series which shows longer term dependencies between successive observations.

MA (moving average) model is usually used to model a time series that shows short-term dependencies between successive observations.

As the lag goes to zero after lag 4 and pacf follows the same pattern we can have a model for the time series for the percent change in next week price as ARMA(4,0,0)

ARMA(0,0,3) - If we want to see the short term dependency of the monthly impact on the next one for the percent change

We can try different model or maybe go for HoltWinter Filtering as in this case we are not assuming the regressors to take part in prediction.

Results:

ARMA(4,0,0)

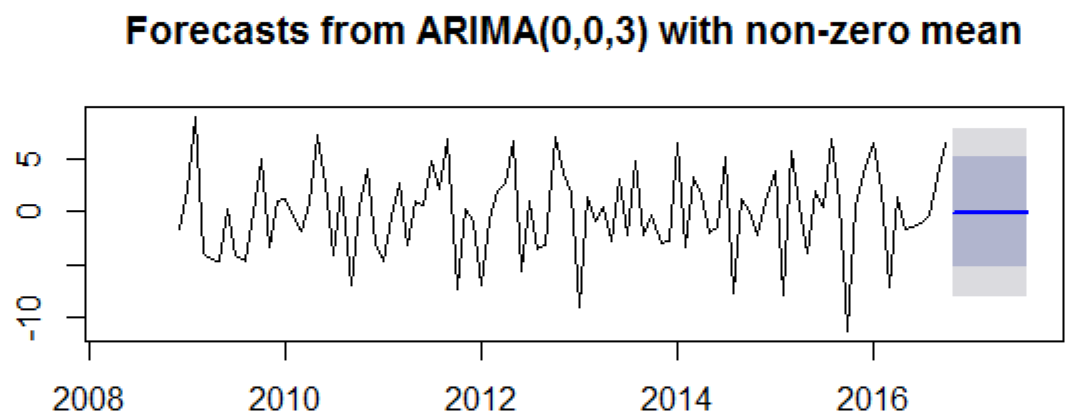
- Coefficients:
- ar1 ar2 ar3 ar4 intercept
- 0.0333 -0.0784 0.0847 0.1409 -0.7602
- log likelihood = -326.89, aic = 665.78 BIC(m1) [= 666.318

ARMA (0,0,3)

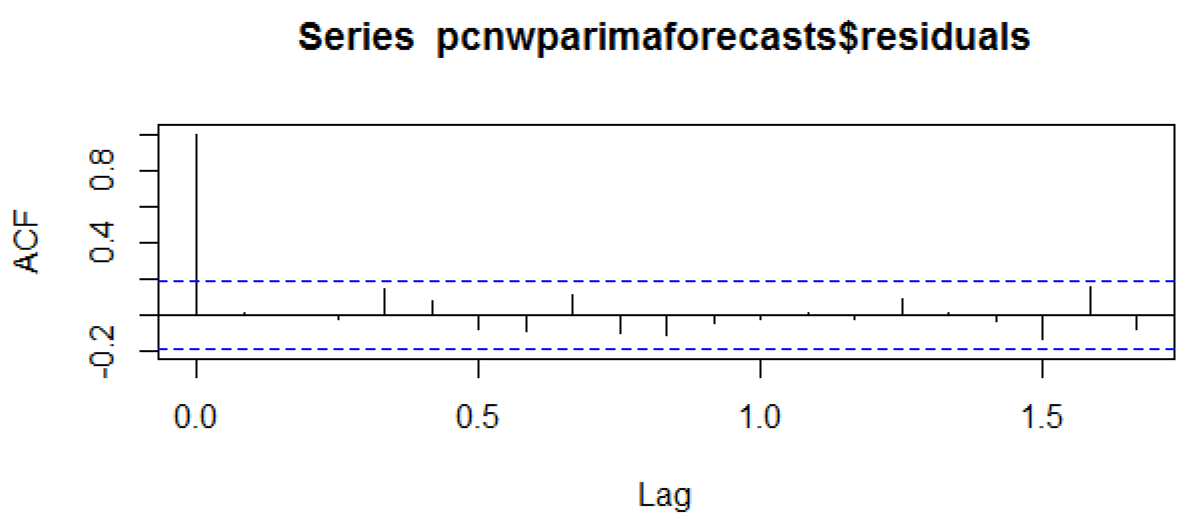
- Coefficients:
- ma1 ma2 ma3 intercept
- 0.0072 -0.0859 0.1191 -0.7806
- log likelihood = -327.79, aic = 665.58 BIC(m1) = 661.948

Forecast:

- pcnwparimaforecasts
- Point Forecast Lo 80 Hi 80 Lo 95 Hi 95
- May 2017 -2.0068523 -8.641524 4.627819 -12.15371 8.140002
- Jun 2017 -0.2208761 -6.855718 6.413966 -10.36799 9.926239
- Jul 2017 -1.0841049 -7.743380 5.575170 -11.26859 9.100378
- Aug 2017 -0.7806382 -7.486651 5.925374 -11.03660 9.475323
- Sep 2017 -0.7806382 -7.486651 5.925374 -11.03660 9.475323



Residual Plot:



ARIMA (With Predictors)

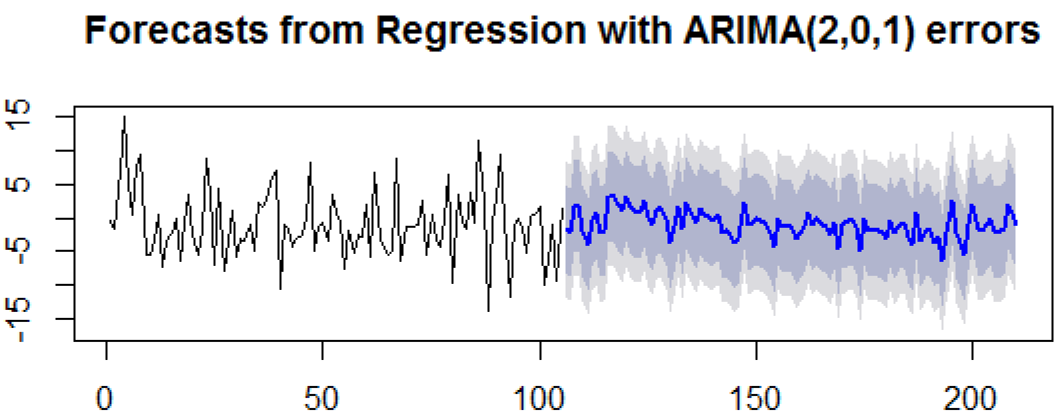
Nested model for checking the effect of the predictors, like adding and removing volume or closing stats in the percent change next week price

Trying different lag to check for the model for say a.x(t – 1) for lag 1 or adding another lag to the regressor.

Different ARMA model for taking into account the Autoregressive components and the moving average components (We don't have the difference component as the time series is stationary

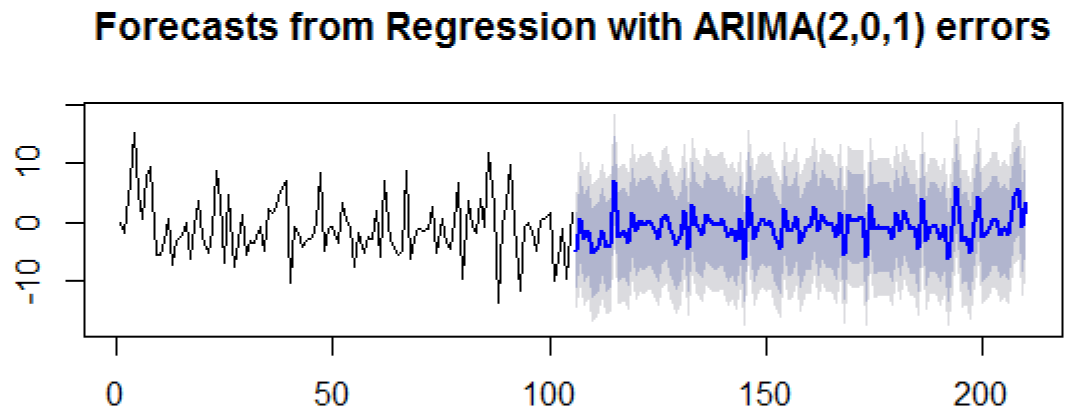
ARIMA (2,0,1) Without volume as predictor:

```
Regression with ARIMA(2,0,1) errors
•
• Coefficients:
•          ar1     ar2      ma1  intercept  open_1  close_1        T
  T2
•       0.1340  0.2293  -0.0709     2.3880  0.0035    6e-04  -1.2468  0.0
  821
• s.e.  0.3453  0.1837   0.4237     2.2409  0.0016    2e-03   0.9635  0.0
  849
•
• sigma^2 estimated as 26.63:  log likelihood=-317.2
• AIC=652.4   AICc=654.29   BIC=676.29
•
• Training set error measures:
•                        ME     RMSE      MAE      MPE     MAPE      MASE
• Training set 0.0003622429 4.960133 3.872891 101.0843 270.3931 0.6909171
•                       ACF1
• Training set -0.0003217476
```



ARIMA (2,0,1) With volume as predictor:

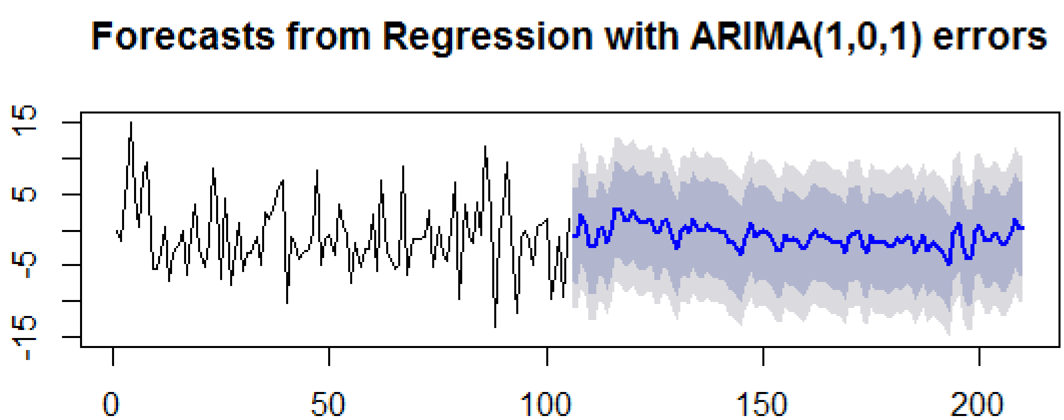
```
Regression with ARIMA(2,0,1) errors

Coefficients:
          ar1      ar2     ma1  intercept   open_1  close_1  vol_1
      -0.2651  -0.0570  0.8436    -1.1566  -0.0013   0.0050      0
s.e.   0.2369   0.3355  0.1630     0.7436   0.0031   0.0019    NaN

sigma^2 estimated as 25.6:  log likelihood=-316.06
AIC=648.11   AICc=649.61   BIC=669.34

Training set error measures:
                      ME     RMSE      MAE      MPE     MAPE      MASE
ACF1
Training set -0.02857106 4.887973 3.853627 363.4154 507.5191 0.6874805 -0.011
65455
```



Other model selections:

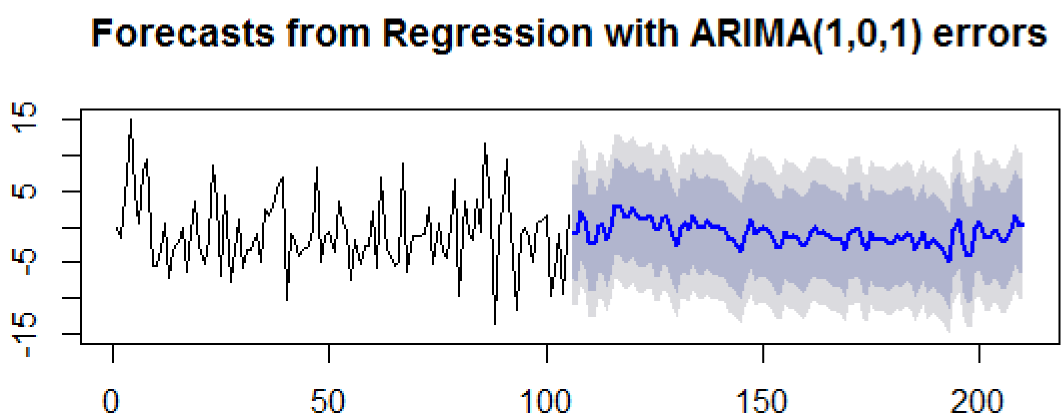
ARIMA (1,0,1):

```
Regression with ARIMA(1,0,1) errors

Coefficients:
         ar1      ma1  intercept  open_1  close_1        T      T2
      0.5498  -0.3922     2.3292  0.0021   0.0015  -1.2249  0.0817
s.e.  0.3731   0.4675     2.1610  0.0013   0.0020   0.9300  0.0810

sigma^2 estimated as 26.52:  log likelihood=-317.48
AIC=650.97   AICc=652.47   BIC=672.2

Training set error measures:
                       ME     RMSE      MAE      MPE     MAPE      MASE
ACF1
Training set -0.003391986 4.975509 3.863241 71.16593 238.3193 0.6891956 -0.00
4858677
```

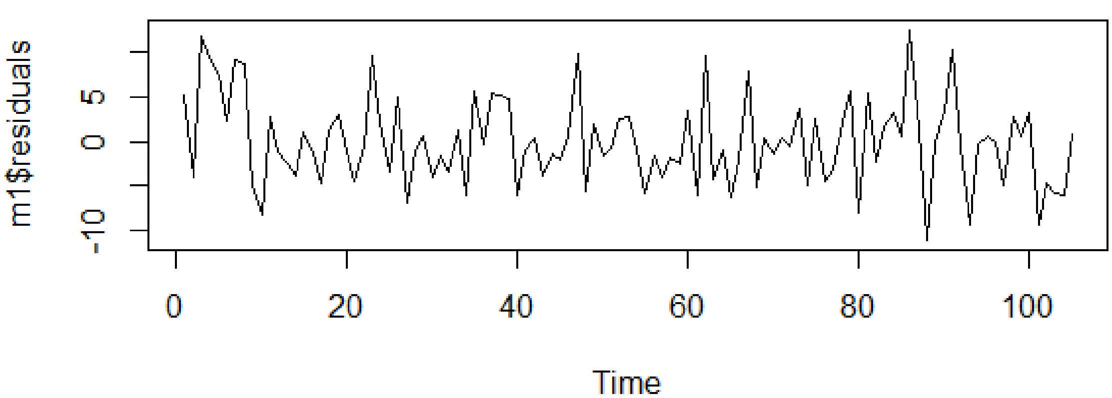


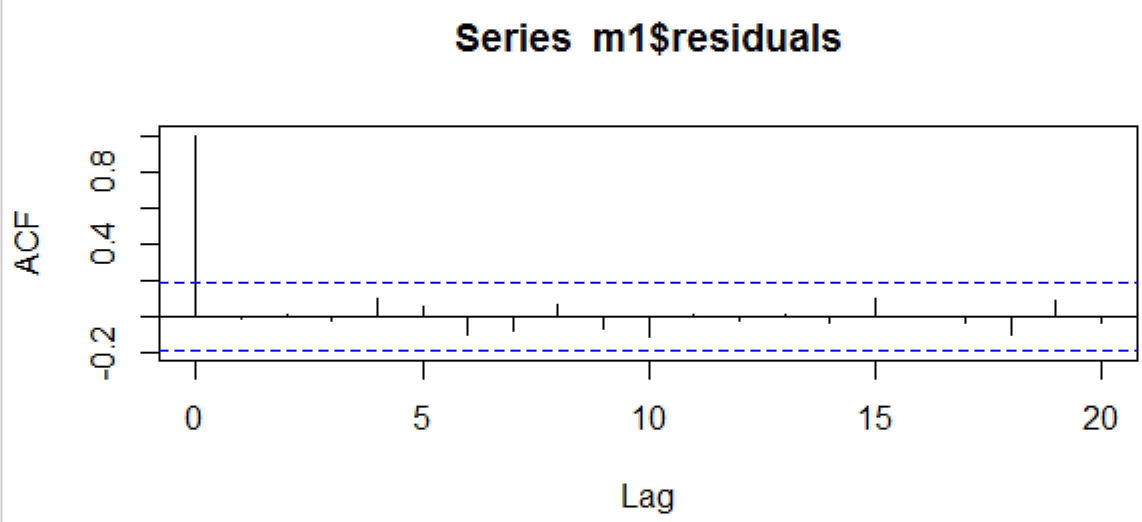
ARIMA (1,0,0):

```
Regression with ARIMA(1,0,0) errors

Coefficients:
         ar1  intercept  open_1  close_1        T      T2
      0.1816     2.2625  0.0016   0.0017  -1.2077  0.0816
s.e.  0.3257     2.1939  0.0014   0.0032   0.9584  0.0804

sigma^2 estimated as 26.35:  log likelihood=-317.67
AIC=649.33   AICc=650.49   BIC=667.91

Training set error measures:
                        ME     RMSE      MAE      MPE     MAPE      MASE
ACF1
Training set -0.0003961936 4.98432 3.881408 153.6582 306.2064 0.6924367 -0.00
2995505
```



Residual and its ACF for ARIMA (2,0,1) with Volume as predictor:



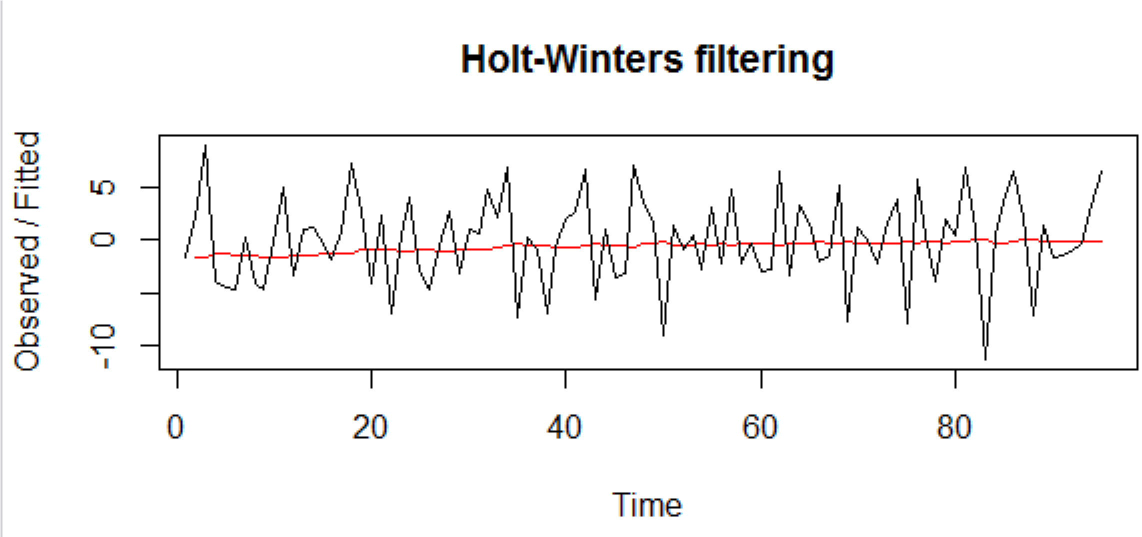


Forecasts using Exponential Smoothing

If you have a time series that can be described using an additive model with constant level and no seasonality, you can use simple exponential smoothing to make short-term forecasts.

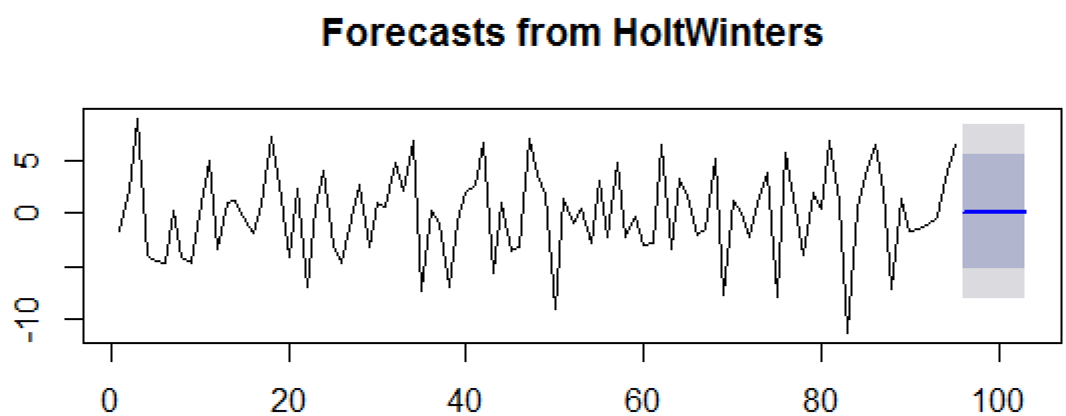
Holt-Winters exponential smoothing without trend and without seasonal component.

```
Call:
HoltWinters(x = pcnwp[8:102], beta = FALSE, gamma = FALSE)

Smoothing parameters:
 alpha: 0.03135965
 beta : FALSE
 gamma: FALSE

Coefficients:
       [,1]
a 0.1590332
```




```
> pcnmperiesforecasts$SSE
[1] 1695.59
```

This is the prediction plot for future values.



Results:

First we have used only the response variable for the analysis without using the predictors like open , close or volume. Looking at the ACF we decided to go with two different ARIMA model (4,0,0) and (0,0,3) and shows comparisons.

On the second level of analysis we tried the ARIMA model for 2 AR and 1 MA model with Volume and without Volume for comparison. We have used AIC and BIC for the goodness of fit for the model. Also 2 different models of the ARIMA has been summarized and forecasted

I have also mentioned a technique for the forecasting called exponential forecast which has some assumptions built into it.

Conclusion:

Looking at the results we can see that the ARIMA model (2, 0, 1) suits the best for the model fitting as it will not discard the assumption of long term and short term impact of price .ARIMA with volume is a better predictor than without which states the role of the volume of the stocks shift plays a major role in the percent change in price.

The ARIMA without the predictors for both 4 AR and 3 MA shows the same summary and trend which was expected of it.

To wind up the statement I would say that our time series model on the share market has performed in good confidence interval of the actual values. We can include further regressors to see if their behavior has an impact on the percent change.

References:

Relevant Papers:

- Brown, M. S., Pelosi, M. & Dirska, H. (2013). Dynamic-radius Species-conserving Genetic Algorithm for
  the Financial Forecasting of Dow Jones Index Stocks. Machine Learning and Data Mining in Pattern
  Recognition, 7988, 27-41.
- http://a-little-book-of-r-for-time-series.readthedocs.io/en/latest/src/timeseries.html
- https://www.r-bloggers.com/forecasting-stock-returns-using-arima-model/