S&P Global Quarterly Composite Index Modeling

Adheesh Ajit Phadnis

Paridhi Rajyaguru

Nan Zhang

**Abstract**:

In this project, we collect quarterly data of S&P Index from 1936 to 2020, to build different models to predict S&P Index. First, we analyze the data and compare different plots to find out a proper response variable, DIFFINDEX (The difference of the SPINDEX between this year and last year). Before tried different models, the dataset was divided into two subsets: All data (1936-2020) and Post-World War 2 (1946-2020). The first model is autoregressive model built with first lag of DIFFINDEX for all data, and for Post-World War 2 dummy variable (after 2008) is also included. The second model is autoregressive model with first lag and second lag of DIFFINDEX. The third model is ARIMA model, combined differencing with autoregression and a moving average model. After comparison between AIC of different ARIMA models, we find that ARIMA (3,0,3) is better. In the last step, we compare the MAE and RMSE of autoregressive models and ARIMA (3,0,3) and conclude that ARIMA (3,0,3) is the best one.

1. **Introduction**

An important task of a financial analyst is to quantify costs associated with future cash flows. We consider here funds invested in a standard measure of overall market performance, the Standard and Poor’s (S&P) 500 Composite Index. The goal is to forecast the performance of the portfolio for discounting of cash flows. In particular, we examine the S&P Composite Quarterly Index for the years 1936 to 2020, inclusive.

The response variable chosen is Log returns of S&P Global composite index. Instead of forecasting the actual index, the research is targeted in forecasting the differenced logarithmic series. The reason is that the actual series has a unit root. It can be observed using Augmented Dicky Fuller test that after first order differencing the log returns the series becomes stationary. Moreover, the difference of logarithm can be interpreted as proportional changes. The independent variables for the research question are the first lag of the differenced logarithmic series. The independent variables are the first lag and the dummy variable after2008 (1 = during or after 2008 or 0 = before 2008)

1. **Data Description**

The below table gives a detailed description of the dataset.

TABLE 1

DESCRIPTION OF DATASET

|  |  |  |
| --- | --- | --- |
| File Name: Number of Number of SP500Quarterly observations: 284 variables: 5 | Number of SP500Quarterly observations: 339 | Number of  variables: 5 |
| Variable | Num of Observations Missing | Description |
| YEAR  SPINDEX  DIFFINDEX  LNSPINDEX  DIFFLNSP |  | Year  The Standard and Poor’s (S&P) 500 Composite Index  The difference of the SPINDEX between this year and last year  The natural logarithm of SPINDEX  The difference of LNSPINDEX between this year and last year |

(Source: Center for Research on Security Prices, University of Chicago and Yahoo Finance)

**Exploratory Data Analysis:**

*Comparative plot for 4 variables in data*

Chart

Description automatically generated

Figure 1: Comparative plot

From the original index values in the upper- left-hand panel, we see that the mean level and variability increase with time. This pattern clearly indicates that the series is nonstationary.

The time series plot of the differences, in upper-right-hand panel, still indicates a pattern of variability increasing with time.

An alternative transformation is to consider logarithmic values of the series. The time series plot of logged values, presented in lower-left-hand panel of Figure 1, indicates the mean level of the series increases over time and is not level. Thus, the logarithmic index is not stationary.

Another approach is to take the difference of log of index. This is especially desirable when looking at indices, or “breadbaskets”, because the difference of logarithms can be interpreted as proportional changes. From the final time series plot, in the lower-right-hand panel of Figure 1, we see that there are fewer discernible patterns in the transformed series, the difference of logs. This transformed series seems to be stationary. Hence, we consider DIFFLNSP as the response variable.

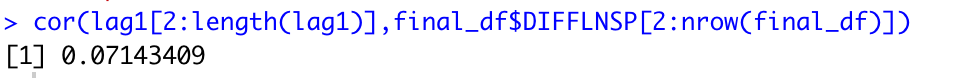
*Exploring the relationship between first lag of DIFFLNSP and DIFFLNSP*

Chart, scatter chart

Description automatically generated

Figure 2: Relationship between first lag of DIFFLNSP and DIFFLNSP

From Figure 2 we can see that there is weak linear relationship between the first lag and DIFFLNSP. The correlation coefficient between lag1 and DIFFLNSP is 0.07 which confirms the aforementioned statement.



Summary statistics:

A picture containing text

Description automatically generated

As we see the mean becomes close to zero as from SPINDEX to DIFFLNSP confirming with the observation from the graphical analysis that the process is close to white noise.

1. **Selection**

**Feature Engineering**

We tried modelling the response variable (DIFFLNSP) using various approaches. For all the models we divided the dataset into training and testing for model evaluation. In doing so we tried to train the data on two subsets viz. **All data** (1936-2020) and **Post-World War 2** (1946-2020). The below table summarizes the datasets after splitting.

TABLE 2

DIVIDED DATASET

|  |  |  |
| --- | --- | --- |
|  | **All data** | **Post-World War 2** |
| *Train* | 1936 - 2016 | 1946 - 2015 |
| *Test* | 2017 – 2020 | 2016 - 2020 |

We also created a dummy variable called after2008 (0: pre-financial crisis i.e., before 2008 and 1: after 2008).

* 1. **First Model**

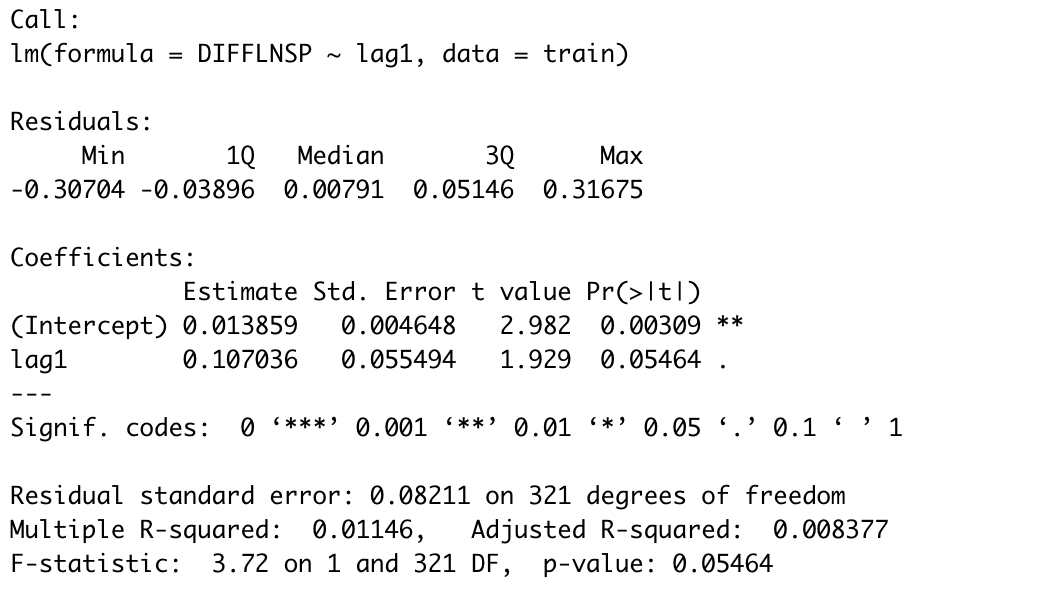
**AR (1)**

**All data:**

Here, we used the first lag of DIFFLNSP as the predictor variable. The regression line is as below:

**DIFFLNSP (predicted) = 0.013859 + 0.107036 \* DIFFLNSP**

The model output is as follows:



**Post-World War:**

Here, we used the first lag of DIFFLNSP and after 2008 as the predictor variables. The reason for including the dummy variable in this model is to experiment to see if the dummy variable provides any information gain reducing the p-value. Since the training set here contains a mix of 0 and 1 values for the dummy variable, the hypothesis was that the variation in DIFFLNSP would be explained by the dummy variable. The regression line is as below:

Table

Description automatically generated

We can see that the adjusted R squared decreases as compared to the previous model because of the addition of after2008 variable. However, the p-value of this model increased substantially.

**DIFFLNSP (t) = 0.016377 + 0.082351\* DIFFLNSP(t-1) - 0.006632 \* after2008**

* 1. **Second Model**

**AR (2)**

**All data:**

Here, we used the first lag and second lag of DIFFLNSP as the predictor variable. The regression line is as below:

**DIFFLNSP (predicted) = 0.013859 + 0.107036 \* DIFFLNSP**

The model output is as follows:

Table

Description automatically generated

If we compare the output of this model to the AR (1) model we see that the adjusted R squared increases with the additional lag variable. We also tried adding the after2008 variable but the adjusted R-squared increases further. Hence further analysis on this model is not done.

**Post-World War:**

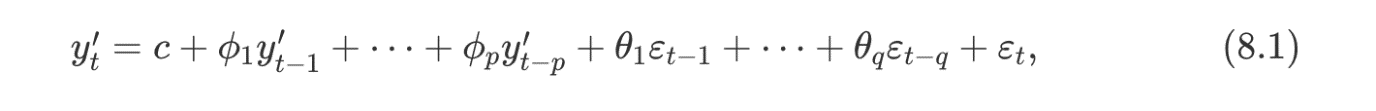
From the AR (2) all data model it was clear that lag2 should not be included in the model. Hence further analysis for Post-World war data was not done.

* 1. **Final Model**

**ARIMA model**

If we combine differencing with autoregression and a moving average model, we obtain a non-seasonal ARIMA model.

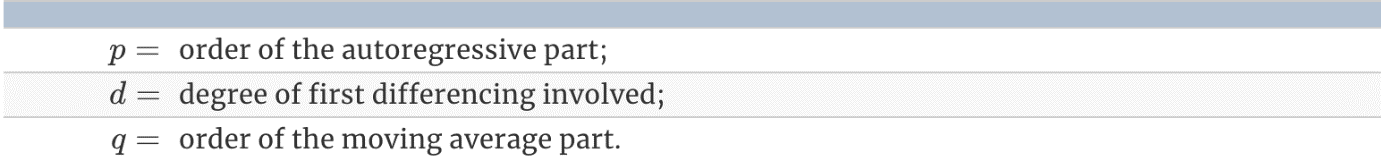
The equation for the model is



where yt′ is the differenced series (it may have been differenced more than once). The “predictors” on the right-hand side include both lagged values of yt and lagged errors. We call this an **ARIMA (**p, d, q**) model**

TABLE 3

MEANING OF P, D, Q



For this model we have used Forecast package in R. There are two types of functionalities in selecting the p, d, q for the ARIMA model. We can use the auto.arima function or we can chose our own model. Fitting auto. Arima yield the following output.

Text

Description automatically generated

The auto.arima function selected ARIMA(0,0,1) model based on the series. The d part of the model makes sense since the series is already differenced and pretty much stationary. Hence it does not require differencing. P says no lag is significant to be included in the regression. And the order of the MA part is 1 which means the first lag of the errors should be included in the regression equation.

The regression line is:

**DIFFLNSP (t) = c + 0.1174 \* ma1**

**C = 0.0154 \* (1- 0.1174) = 0.0136**

**DIFFLNSP (t) = 0.0136 + 0.1174 \* ma1**

We also tried different ARMA (p,0, q) models manually. The table below summarizes the output.

TABLE 4

OUTPUTS OF DIFFERENT ARIMA MODELS

|  |  |  |
| --- | --- | --- |
| Model | AIC | BIC |
| ARIMA (1,0,1) | -692.63 | -677.65 |
| ARIMA (2,0,2) | -691.64 | -669.24 |
| ARIMA (3,0,3) | -688.19 | -658.42 |

The model with lowest AIC value is ARIMA (3,0,3). Hence, we select this model for further predictions on test set.

Regression Output:

Text

Description automatically generated

**Predictions:**

**Text

Description automatically generated**

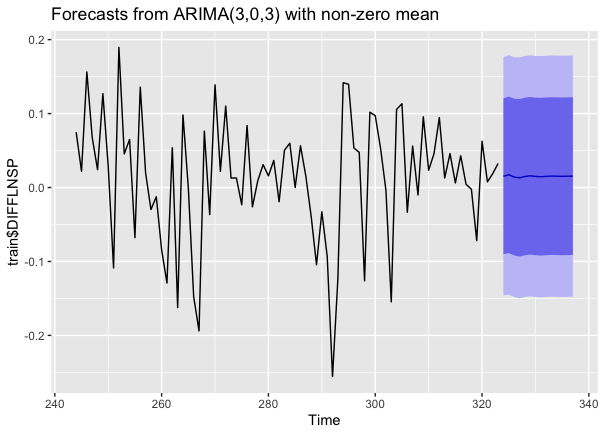
****

Figure 3: Forecast from ARIMA (3,0,3) with non-zero mean

TABLE 5

MAE AND RMSE OF DIFFERENT MODELS

|  |  |  |
| --- | --- | --- |
| Model | MAE | RMSE |
| AR (1) – All data | 0.072 | 0.102 |
| AR (1) – WW2 | 0.060 | 0.090 |
| AR (2) – All data | 0.074 | 0.106 |
| ARIMA (3,0,3) | 0.059 | 0.088 |

The best model is ARIMA (3,0,3)

1. **Final Model Interpretation**

**Assumptions of ARIMA model**

* 1. Data should be stationary – by stationary it means that the properties of the series don’t depend on the time when it is captured. A white noise series and series with cyclic behavior can also be considered as stationary series.
* 2. Data should be univariate – ARIMA works on a single variable. Auto-regression is all about regression with the past values.

**Estimated regression line**

ARIMA (3,0,3) model:

Yt = c + 0.9327yt−1− 0.6288yt−2 + 0.4819yt−3 - 0.8310εt−1+ 0.4945εt−2 - 0.4838εt−3 + εt

where c=0.0155× (1−0.9327) = 0.00104315 and εt is white noise with a standard deviation of 0.082 =√0.006753

**Interpretation of slope and coefficients**

* Intercept interpretation:

Holding all lags and shocks constant, the predicted return at time t on an average is 0.001043 = 0.1%

* Coefficients:

Holding everything else constant, with 1% increase in the previous quarter return the predicted return at time t is expected to increase by 0.9325 %

Holding everything else constant, with 1% increase in the last-to-last quarter return the predicted return at time t is expected to decrease by 0.6288 %

Holding everything else constant, with 1% increase in the last to last to last quarter return the predicted return at time t is expected to increase by 0.4819 %

Holding everything else constant, with 1% increase in the previous quarter unexpected news shock the predicted return at time t is expected to decrease by 0.8310 %

Holding everything else constant, with 1% increase in the last-to-last quarter unexpected news shock the predicted return at time t is expected to increase by 0.4845 %

Holding everything else constant, with 1% increase in the last to last to last quarter unexpected news shock the predicted return at time t is expected to decrease by 0.4838%

**AIC and BIC**

Akaike’s Information Criterion (AIC), which was useful in selecting predictors for regression, is also useful for determining the order of an ARIMA model.

AIC=−2log⁡(L)+2(p+q+k+1),

where LL is the likelihood of the data, k=1 if c≠0 and k=0 if c=0. Note that the last term in parentheses is the number of parameters in the model (including σ2σ2, the variance of the residuals).

And the Bayesian Information Criterion can be written as

BIC=AIC+[log(T)−2] (p+q+k+1).

Good models are obtained by minimizing the AIC, AIC or BIC. Our preference is to use the AIC.

Hence, we select the model ARIMA (3,0,3) yielding the lowest AIC and BIC values.

**Residual analysis**

The Ljung-Box test has p value of 0.5334.

**H0:** The data are independently distributed (i.e., the correlations in the population from which the sample is taken are 0, so that any observed correlations in the data result from randomness of the sampling process).

**Ha:** The data are not independently distributed; they exhibit serial correlation

The p – value is > 0.05 we cannot reject the null that the data are independent and identically distributed.

Text, letter

Description automatically generated

Chart

Description automatically generated

From first graph we can see that the residuals are independent and identically distributed and the process is white noise.

1. **Conclusion**

From the analysis above, we can use ARIMA model and put in the former three quarters’ data and the influence of former three quarters’ unexpected news shock to get a predicted S&P Index. But how to qualify unexpected news shock is still a problem. Taking into more economic factors into consideration may make the predict result closer to true value but it will also increase the complexity.

1. **Appendix**

**List of tables and figures:**

* [Table 1- DESCRIPTION OF DATASET](#Table1)
* [Table 2-Divided dataset](#Table2)
* [Table 3-Meaning of p, d, q](#Table3)
* [Table 4-Outputs of different ARIMA models](#Table4)
* [Table 5-MAE and RMSE of different models](#Table5)
* [Figure 1-Comparative plot](#Figure1)
* [Figure 2-Relationship between first lag of DIFFLNSP and DIFFLNSP](#Figure2)
* [Figure 3-Forecast from ARIMA (3,0,3) with non-zero mean](#Figure3)

**References:**

* <https://otexts.com/fpp2/arima-estimation.html>
* <https://datascienceplus.com/time-series-analysis-using-arima-model-in-r/#:~:text=Assumptions%20of%20ARIMA%20model&text=A%20white%20noise%20series%20and,regression%20with%20the%20past%20values>.
* <https://en.wikipedia.org/wiki/Ljung%E2%80%93Box_test>
* <https://nwfsc-timeseries.github.io/atsa-labs/sec-tslab-autoregressive-moving-average-arma-models.html>