**MODEL 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.

|  |  |  |
| --- | --- | --- |
|  | **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).

**Models**

**Autoregressive**

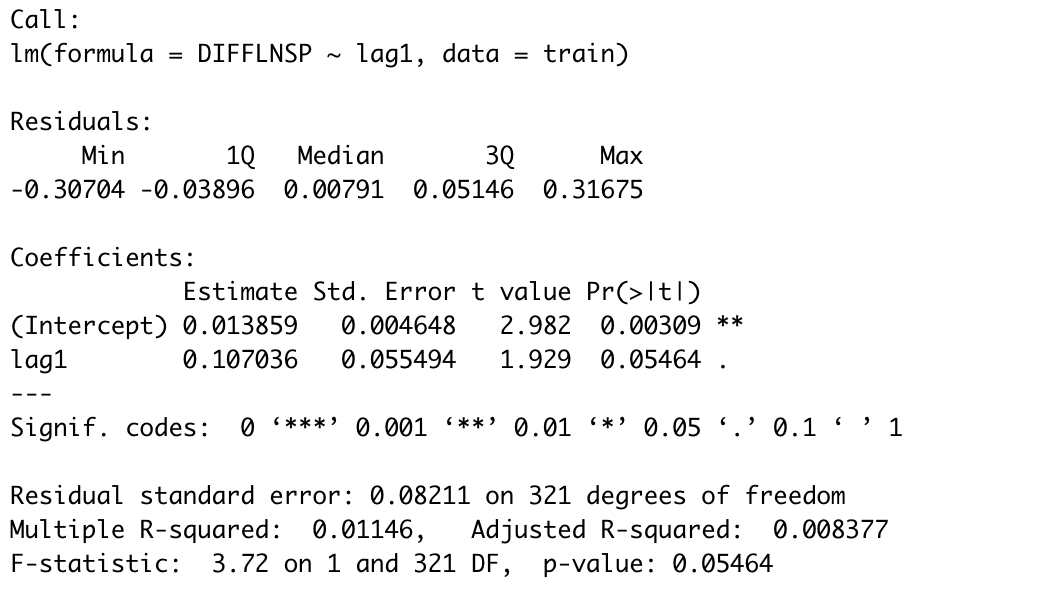
**AR (1)**

**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**

**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.

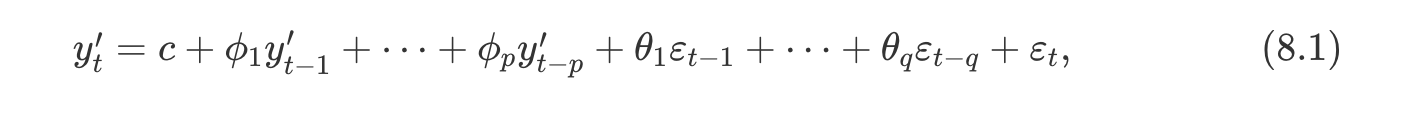
**Post-World War:**

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.

**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 ytyt and lagged errors. We call this an **ARIMA(**p,d,q**) model**



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.

|  |  |  |
| --- | --- | --- |
| 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:

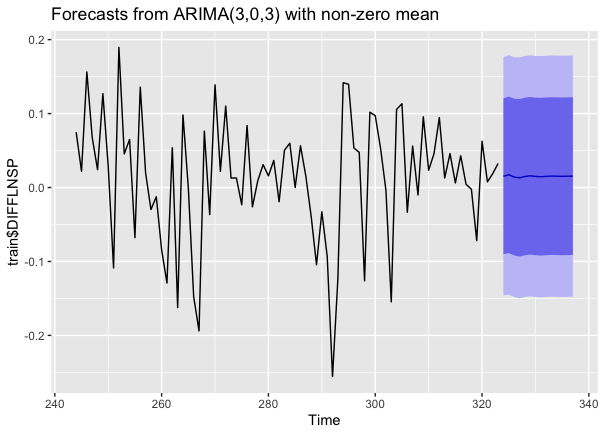
Text

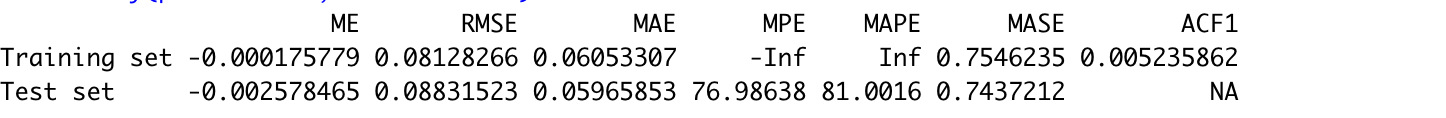
Description automatically generated

**Predictions:**

**Text

Description automatically generated**

****



Model comparison

|  |  |  |
| --- | --- | --- |
| 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)

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=1k=1 if c≠0c≠0 and k=0k=0 if c=0c=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 iid.

Text, letter

Description automatically generated

Chart

Description automatically generated

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