**Time Series Analysis on United Air Revenue.**

**BAN 673\_01**

Team -

Manisha Boyina

Dileep Lingamallu

Abhinay Parasa

Krushi Teja Reddy Padamati

**Executive Summary**

The goal of this time series project is to analyze the quarterly revenue data of United Air from 2000 to 2019 and develop a forecasting model to predict future revenue. The dataset includes quarterly revenue for each year from 2000 to 2019.

In the first step of the project we will visualize the data to understand the trend, seasonality, and other patterns. Next, we will use various time series techniques, such as autocorrelation to explore the data and identify the appropriate forecasting model. Different time series models such as Regression-based models, advanced exponential smoothing models and, autoregressive integrated moving average models (ARIMA) were utilized for this project.

The dataset is divided into training and validation sets, where the training set will be used to build and validate the model, and the validation set will be used to evaluate the model's performance. To achieve successful outcomes, additional regression and advanced exponential smoothing models were built. A trailing moving average for residuals and an autoregressive model for residuals were added to the regression models as needed. When necessary, the same improvements were made to the advanced exponential smoothing models. The RMSE and MAPE accuracy metrics were used as the basis for model evaluation.

Finally, the forecasting model is used to predict future quarterly revenue for United Air. The results and findings of the project will be presented in a report that will provide insights to the stakeholders to make better decisions regarding United Air's future revenue.

**Introduction**

United Airlines is one of the largest airlines in the world and has been providing passenger and cargo transportation services since 1926. The company operates more than 4,900 flights daily to 356 airports across five continents, making it a major player in the global aviation industry.

Time series analysis is an important tool for airlines because it helps them to better understand and forecast demand for their services. Airlines operate in a highly dynamic and complex environment, where demand for air travel is influenced by a variety of factors, such as economic conditions, fuel prices, exchange rates, weather patterns, and geopolitical events. These factors can have a significant impact on airline revenue and profitability, making it essential for airlines to be able to forecast demand and adjust their operations accordingly.

Time series analysis provides airlines with a range of analytical techniques that can be used to identify patterns and trends in historical data, and to forecast future demand based on these patterns. In this project, we will be utilizing regression-based models, advanced exponential smoothing models, and autoregressive integrated moving average models (ARIMA) to analyze the revenue data of United Airlines. As time series data often exhibits patterns and trends that can be difficult to identify with simple statistical analysis, these models will allow us to more accurately forecast future revenue and identify potential factors that influence revenue fluctuations over time. This analysis can provide valuable insights for decision-making and planning for United Airlines.

This, in turn, enables airlines to optimize their capacity and pricing strategies, and to make more informed decisions about route planning, fleet management, and other key operational areas. Ultimately, time series analysis is an important tool for airlines seeking to improve their revenue management and competitiveness in the highly competitive aviation industry.

**Eight Steps of Forecasting**

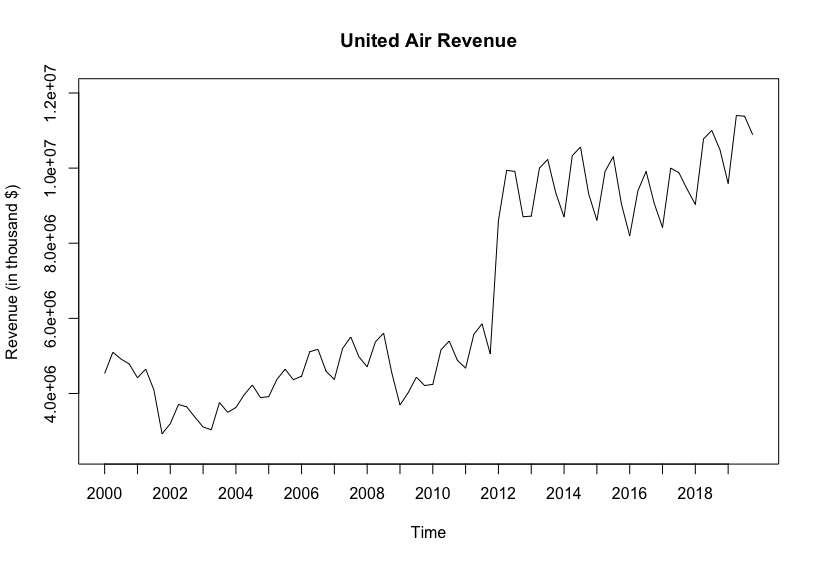
**Step 1: Define the Goal**

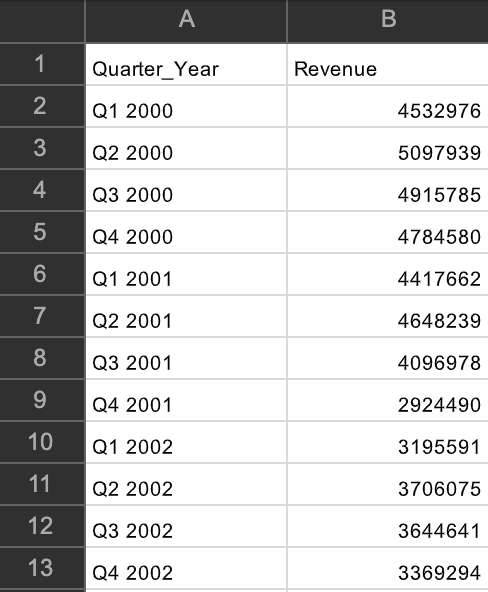
The goal of this project is to create a predictive model which will properly consider both the trend and seasonal components of the historical data and effectively forecast the desired quarters for future periods. Using the revenue data of United Airlines from 2000 to 2019 considering various time series models, the model with the highest accuracy will be considered the model of choice. This analysis can help United Airlines make data-driven decisions and plan for the future, potentially improving revenue and overall business performance. The forecasting models developed for this project were done using the R programming language.

**Step 2: Get Data**

This report will focus on the time series dataset provided by United Air representing the total quarterly revenues of United Airlines. The time period for the dataset ranges from Q1 of 2000 to Q4 of 2019. The data is measured in $1000. For example, for Q1 of 2000, the revenue of 4532976 means $4,532,976,000 or $4.532 Billion.

**Step 3: Explore and Visualize Data**

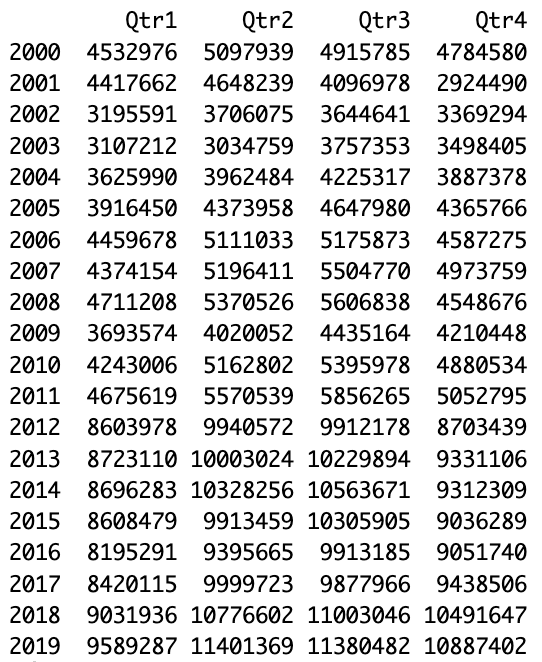




The above are the data plots of the United Air Quarterly time series data. Time series appear to have a normal up trend in the beginning years and graph drastically moved upward trend with seasonality in the year 2012 and maintained the constant trend till the end of 2019. The above is the data is used as input for the time series data set. Data set consists of quarter, year and revenue.

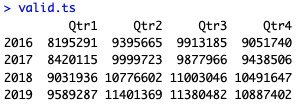
**Step 4: Data Preprocessing**

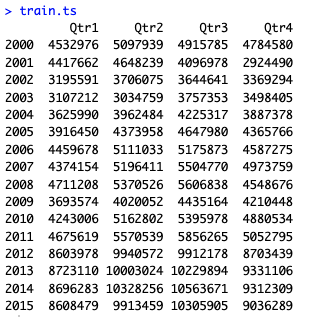
From the original data from United Air, we choose 10 years quarterly data. Doing so, only the most relevant data will be considered for the analysis. There are a total of 80 data points each year containing 4 quarters. Below is the time series data set of it.



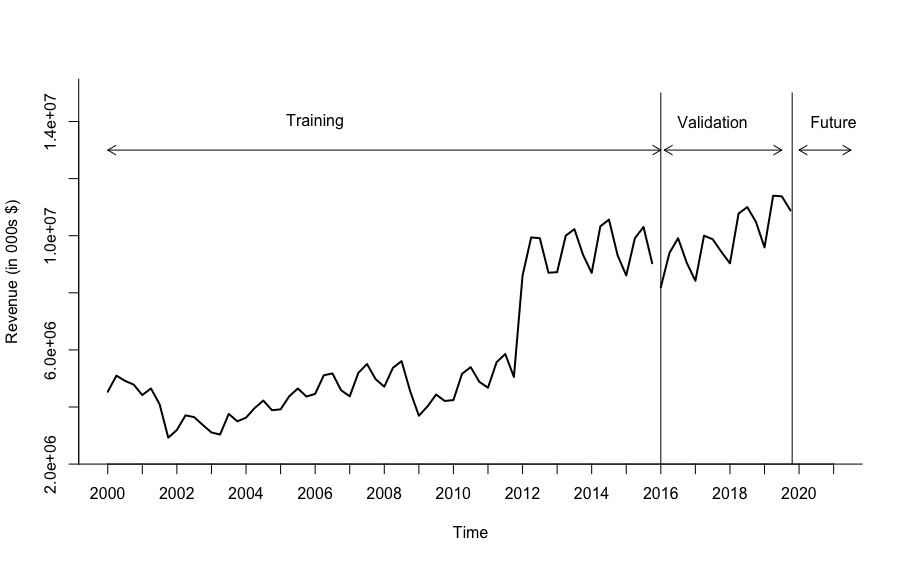
**Step 5: Partition Series**

We created a data partition of 64 records for the training period and 16 records for the validation period. These partitioned validation and training data sets are (2000-2015) and (2016-2019) respectively are below named as valid.ts and train.ts.



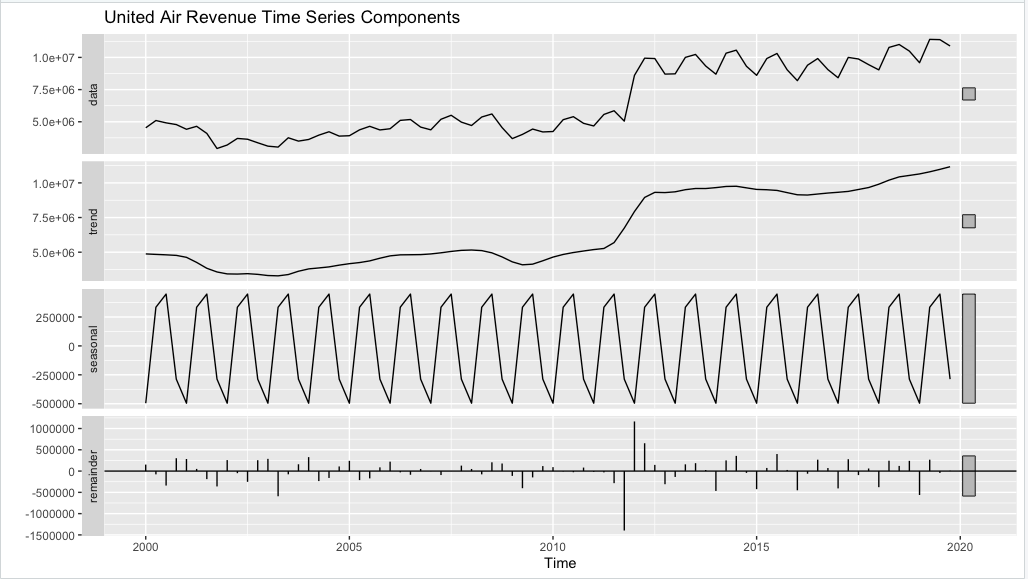


Visual representation of the training and validation partitions of the data.

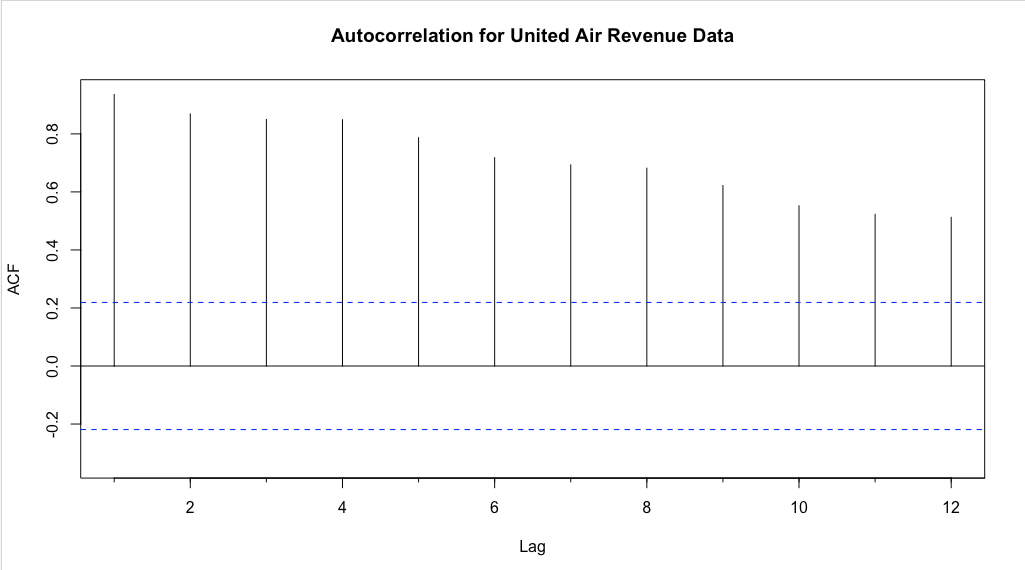


**Step 6 & 7: Apply Forecasting & Comparing Performance**

With the below time series component we can tell the United airlines revenue has an overall upward trend and also an additive seasonality.

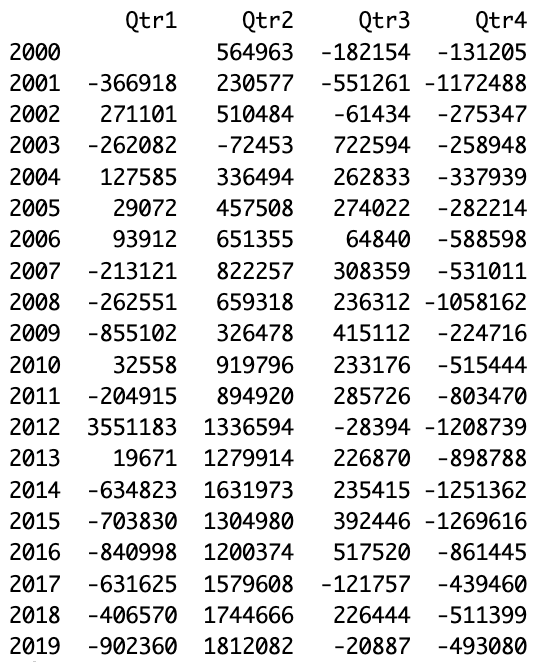


Also, the autocorrelation of the data appears to be statistically significant at all lags implying that there is strong autocorrelation in the data. For all the lags, the ACF is above the upper threshold making the data significant. Which means by further processing of the data with forecasting models we would get better results. Also at lag 1 which represents the trend the ACF is substantially higher than the other lags. And at lag 12 which represents the seasonality the ACF is lower but still significant which tells us that the data has trend and seasonality.

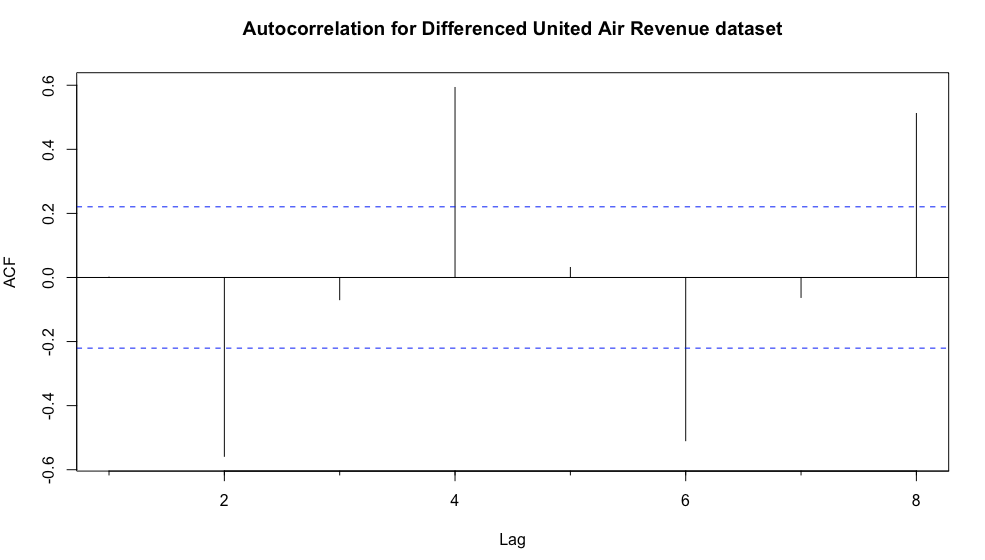


**Predictability Test for the Dataset using Lag 1 Differencing -**

Below is the data after lag 1 differencing:



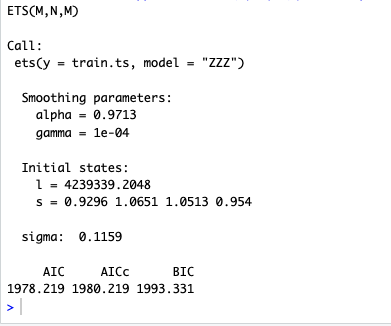
Below is the correlogram generated using the lag 1 differencing data from above and maximum lag of 8.



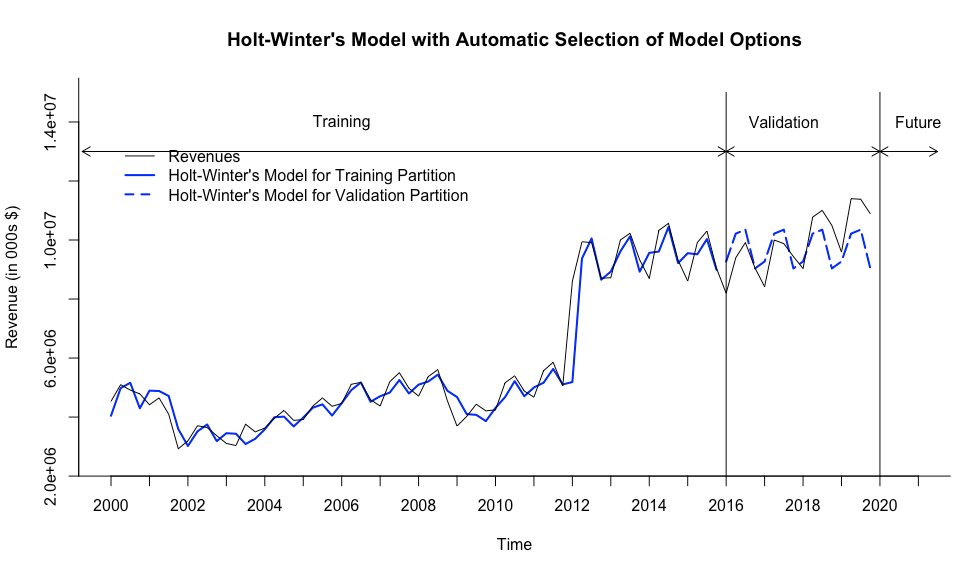
From the correlogram of Lag 1 differencing method, we can say that the data is not a random walk as the ACF at lag 2,4,6,8 are above the level of significance.

**Holts Winter Model for training set:**

Holt Winter’s Model for prediction is used for time series that contains trend and seasonality. Here we have used the automated selection of model options(Z, Z, Z) and the optimal parameters by using ETS().



* The optimal Holt - Winters model obtained is a model of (M, N, M) which represents multiplicative error, multiplicative seasonality and no trend.
* The optimal smoothing parameters for the model are:
  + Alpha (smoothing constant for exponential smoothing) = 0.9713
  + Beta (smoothing constant for trend estimate) = 0
  + Gamma (smoothing constant for seasonality estimate) = 0.0001

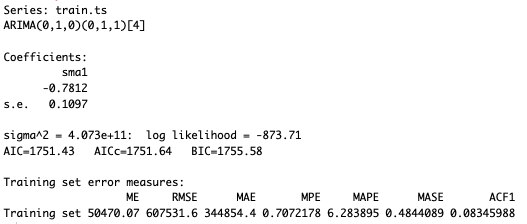


The above plot shows us that there’s an underestimate of this model.

**Auto ARIMA model for training set**

The Autoregressive Integrated Moving Average (ARIMA) model is a flexible model that can be used for forecasting on data with level, trend, and seasonal components. Since our data consists of all three, this model is appropriate to use for analysis. We generated an optimal ARIMA model with automatic selection of (p,d,q) (P,D,Q) parameters using the auto.arima() function.

Auto Arima model which uses automatic selection of the optimal parameters. The Auto ARIMA model summary is as below:



The optimal Auto ARIMA model obtained is *ARIMA (p, d, q) (P, D, Q)[m]* -

- *p = 0,* order 0 autoregressive model *AR (0)*

- *d = 1*, first differencing

- *q = 0*, order 0 moving average *MA (1)* for error lags

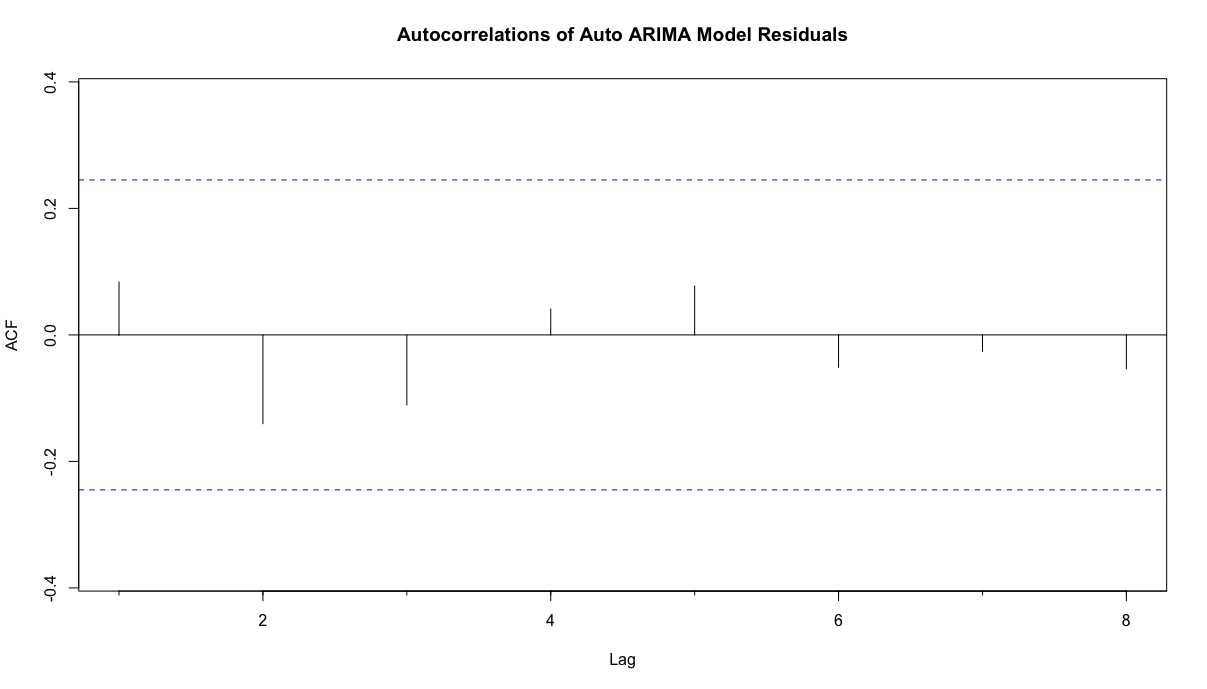
- *P = 0,* order 0 autoregressive model *AR (0)* for the seasonal part

- *D = 1*, first differencing for the seasonal part

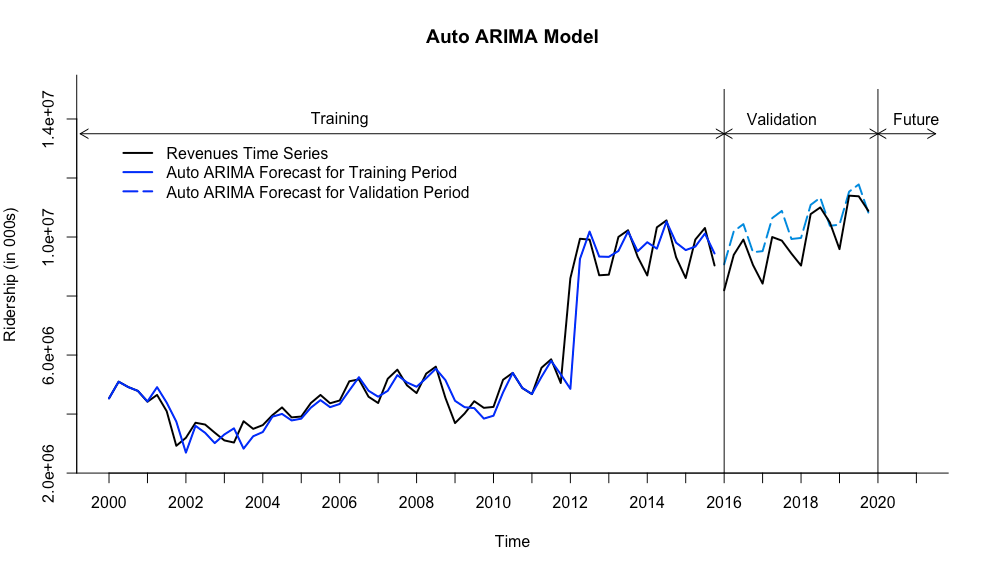
- *Q = 1*, order 1 moving average *MA (1)* for the seasonal error lags

- *m = 4*, for quarterly seasonality.

The ACF plot below has been generated using the residuals AUTO ARIMA model above.



* Based on the ACF plot we can see that the model has captured the trend, seasonality and any other patterns that existed in the original dataset and has incorporated them into the model.
* The autocorrelation at all lags now fall within the levels of significance signifying there are no more patterns in the residuals.

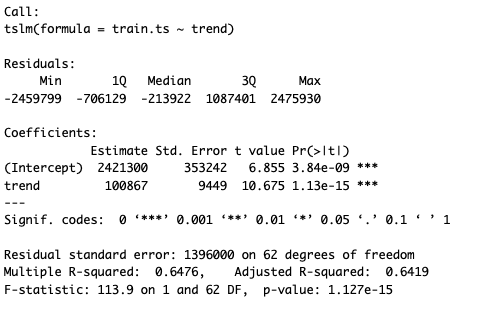


**Regression Models:**

**Regression model with linear trend**: Used to fit a global trend that is applied to the training set of time series and will apply in the forecasting period.

The equation of the below model is:

* The model has an r2 of 64.76% which can be considered as good fit.
* Also, all regression coefficients (trend, intercept) are statistically significant making this model a good fit for the dataset.

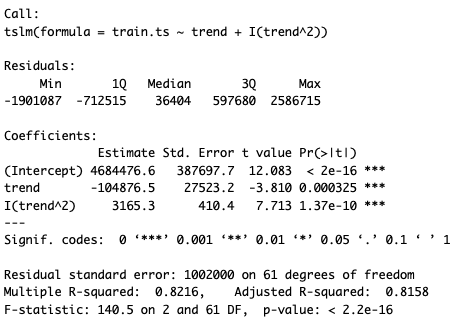


**Regression model with Quadratic trend:**

The equation of the below model is:

+ 3165 t^2

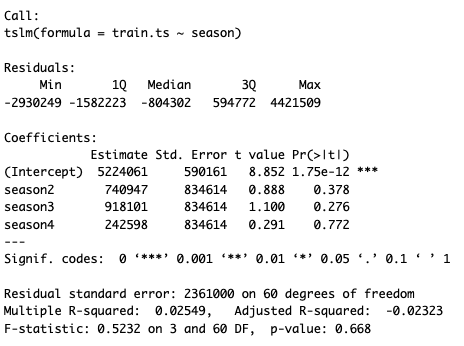
* The model has an r2 of 82.16% which can be considered as good fit.
* Also, all regression coefficients (trend, intercept) are statistically significant making this model a good fit for the dataset.



**Regression Model with Seasonality:**

The equation of the below model is:

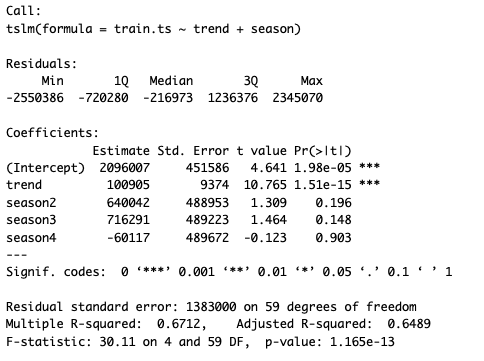
* The model has an r2 of 2.5% which is very low.
* The trend coefficient is statistically significant for this model. However, given the very low r2 we can determine that this model is not a good fit for the given dataset.



**Regression model with linear trend and seasonality:**

The model equation is as below:

* The model has an r2 of 67.12% which can be considered as good fit.
* Also, all numeric non-seasonal regression coefficients (trend, intercept) are statistically significant making this model a good fit for the dataset.

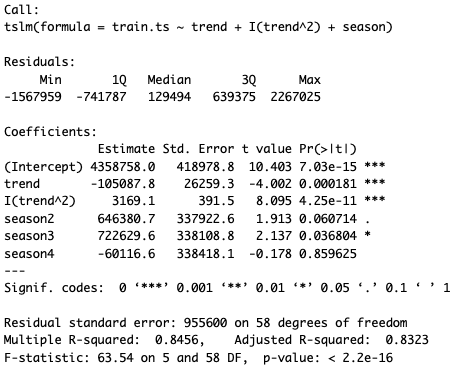


**Regression Model with quadratic trend and seasonality:**

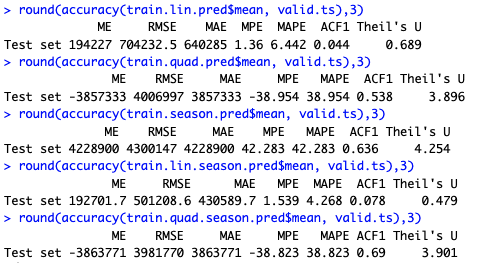
The regression model with quadratic trend and seasonality contains 5 independent variables: trend index (t), squared trend index (t2), and 3 seasonal dummy variables for Q2 (season2 – D2), Q3 (season3 – D3) and Q4 (season4 – D4).

The equation for this model is presented below:

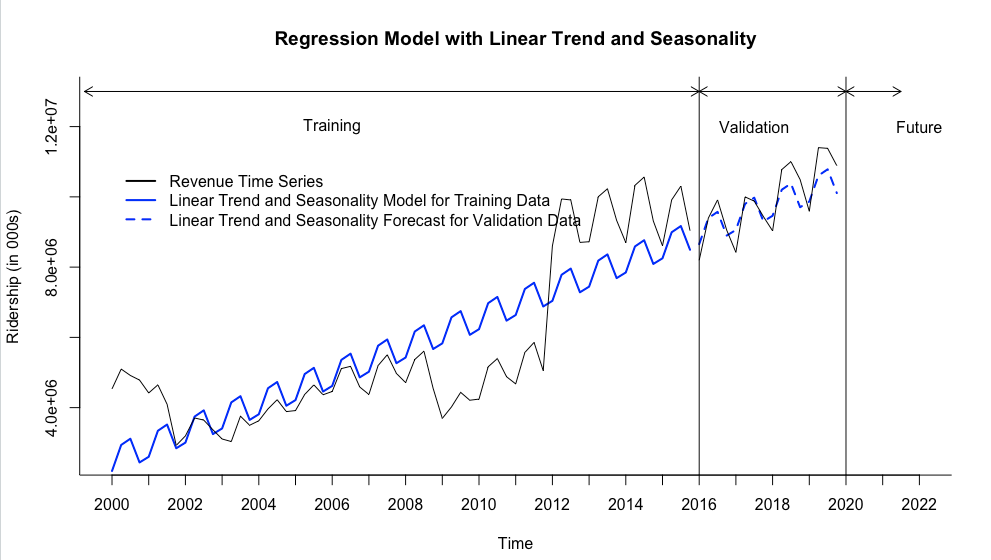
* The model has an r2 of 84.56% which can be considered as good fit.
* Also, all numeric non-seasonal regression coefficients (trend, intercept) are statistically significant making this model a good fit for the dataset.



**Accuracy measures for Regression models above:** Below are the accuracy measures for all the regression models developed on the training partition:

:

* A model with the least RMSE and least MAPE is considered to be the best model.
* Based on the above accuracy measures for the training partition, the best model is Linear trend and seasonality with the lowest MAPE of 4.2% and RMSE of 3981770.
* Below is the Plot for the same.



Below, the ACF plot of the residuals from the level 1 forecasting model(Regression model with Linear Trend and Seasonality)



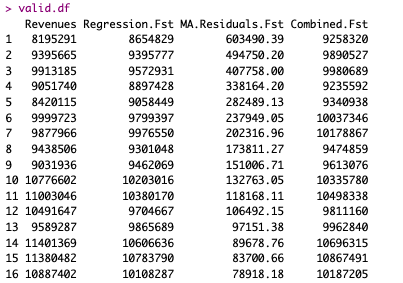
* From the above correlogram it can be seen that for most of the lags the ACF is still above the upper threshold.
* This signifies that there are still certain patterns (trend) existing in the residuals which need to be captured to further improve our forecasts.
* We intend to do this with a 2-Level forecasting model.

**2-LEVEL FORECASTING REGRESSION + MA**

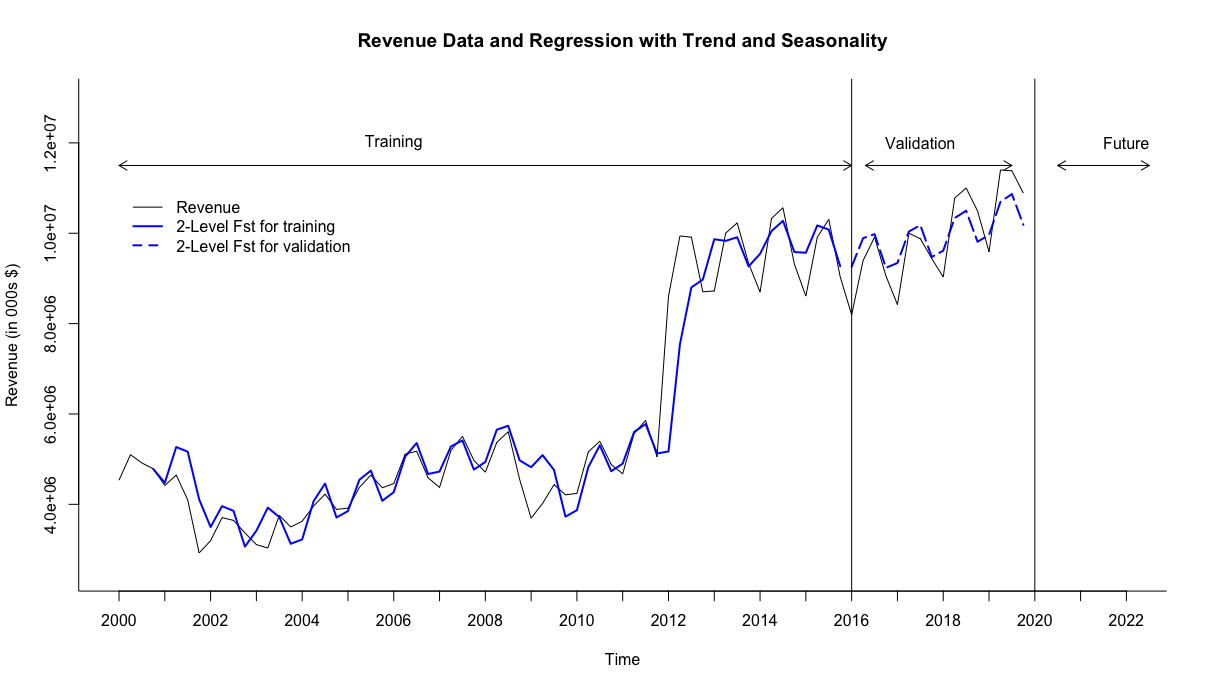
Given that the best model is Regression model with Linear trend with Seasonality, we choose the same for Level 1 forecasting of our 2-Level forecasting model.

Below are the steps involved in applying the 2-Level forecast.

* Identified regression residuals for training partition (differences between actual and regression values in the same periods).
* Display the Autocorrelation correlogram for the Revenue training residuals.
* Apply trailing MA for residuals with window width k = 4 for training partition.
* Create residuals forecast for validation period.
* Develop a two-level forecast for validation period by combining regression forecast and trailing MA forecast for residuals.
* Create and represent a table for validation period: validation data, regression forecast, trailing MA for residuals and total forecast.



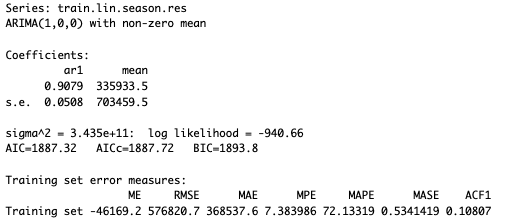
Plot for 2-Level forecasting for the validation period.



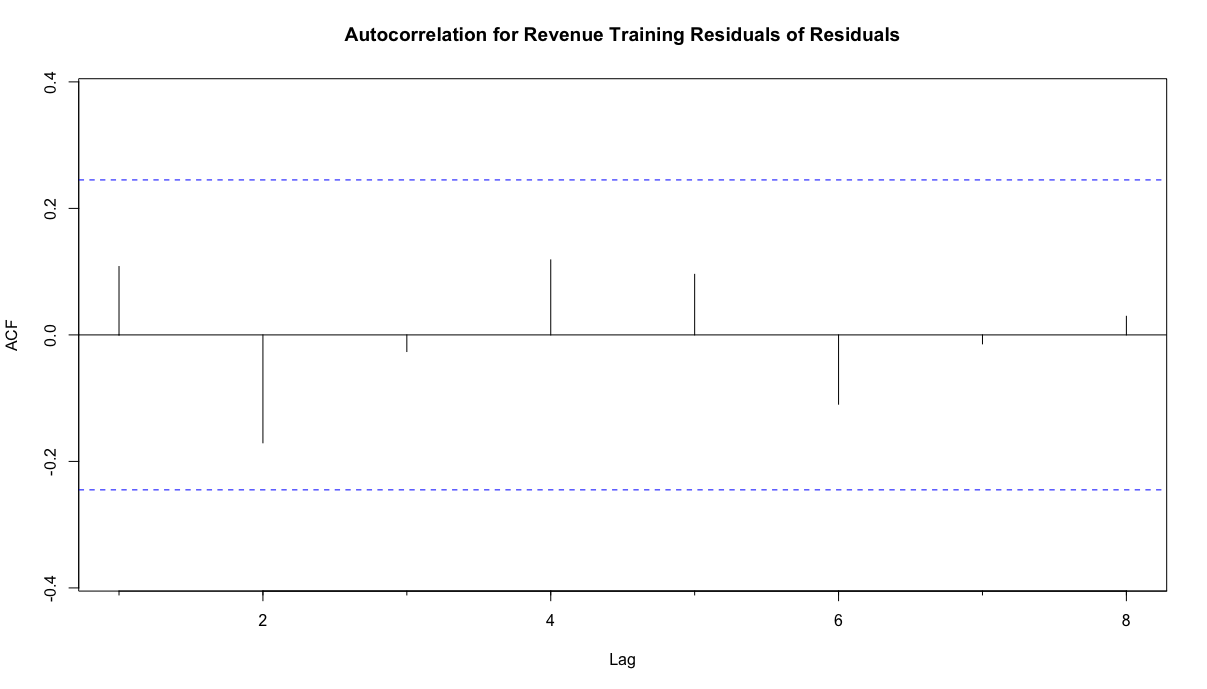
**2-LEVEL FORECASTING REGRESSION + AR model**

In this model, in addition to Level - 1 regression forecasting, we will use an AutoRegressive model of order 1 to forecast the residuals from the level-1 model and improve our forecasts.

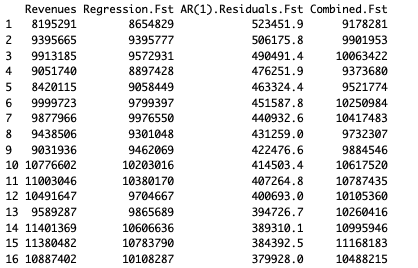
The summary of the AR (1) developed is as below:

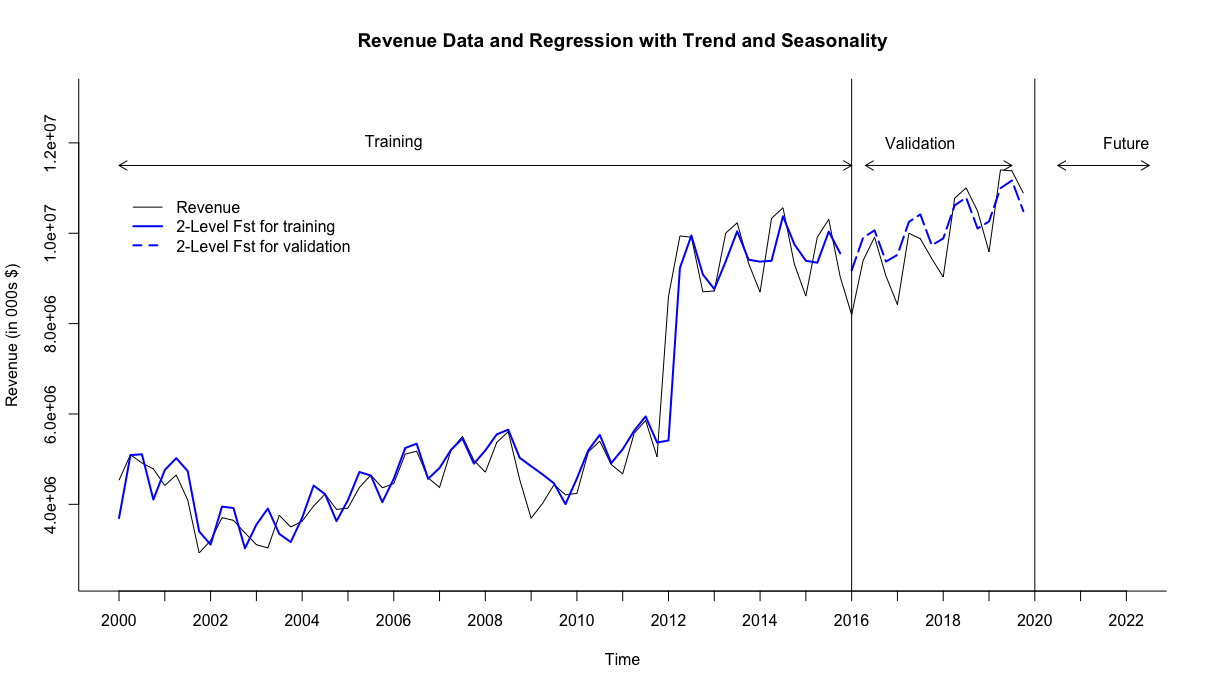


* *ARIMA (1, 0, 0)* is an autoregressive (AR) model with order 1, no differencing, and no moving average model.

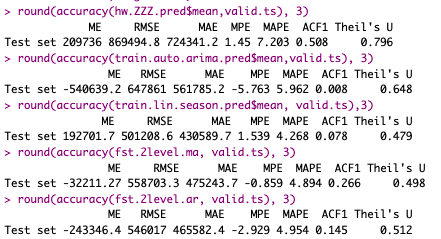


* Autocorrelations of residuals of residuals produced by the AR (1) model can be inferred from this correlogram to be random. Hence, significant autocorrelation in all lags has been observed by the AR (1) model for residuals.
* Data table with validation data, regression forecast for validation period, AR (1) residuals for validation, and two level model results.





Below are the accuracy measures for all the forecasting models considered above namely: (1) Holt Winter’s model; (2) Auto ARIMA model; (3) Regression Model with Linear Trend and Seasonality; (4) 2 Level Forecasting (Regression Model with Linear Trend and Seasonality and Moving Averages); (5)2 Level Forecasting (Regression Model with Linear Trend and Seasonality and AR (1))



* The model with the least RMSE at 501208.6 and MAPE at 4.268% is the Regression model with Linear trend and Seasonality.
* The other models which have slightly higher MAPE and RMSE but still performing good include the 2-Level forecasting models.
* Thus, we would like to consider applying these 3 models on the entire dataset.

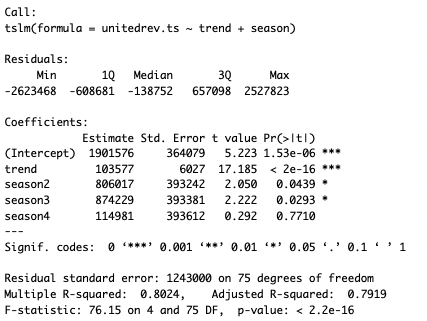
**Now fitting the optimal model on the Entire Dataset**

The chosen models to apply on the entire dataset include:

1. **Regression model with Linear Trend and Seasonality**
2. **2-Level Forecasting Regression + MA**
3. **2-Level Forecasting Regression + AR () model**

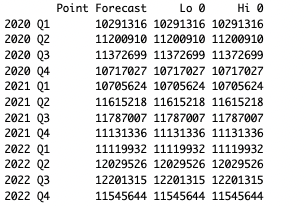
1. Regression Model with Linear trend and Seasonality for entire dataset

Below is the summary of the same.

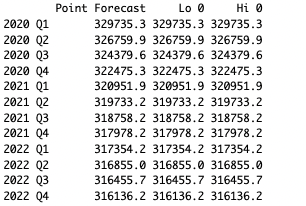


* The regression model contains 4 independent variables: trend index (t) and 3 seasonal dummy variables for Q2 (season2 – D2), Q3 (season3 – D3) and Q4 (season4 – D4).
* All numeric coefficients are below 0.05 making these coefficients significant.
* R-squared and Adj.R-squared are at 80.24% and 79.19%, which is very good.
* Thus, this model equation is significant, making it a good fit for forecasting.

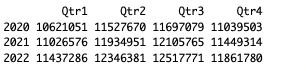
Below is the future 12 periods prediction using Regression model with Linear Trend and Seasonality.



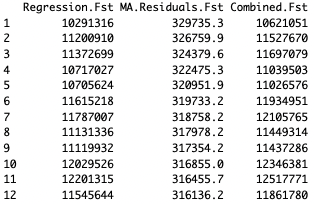
Later, using trailing MA residuals forecast for the future 12 periods.



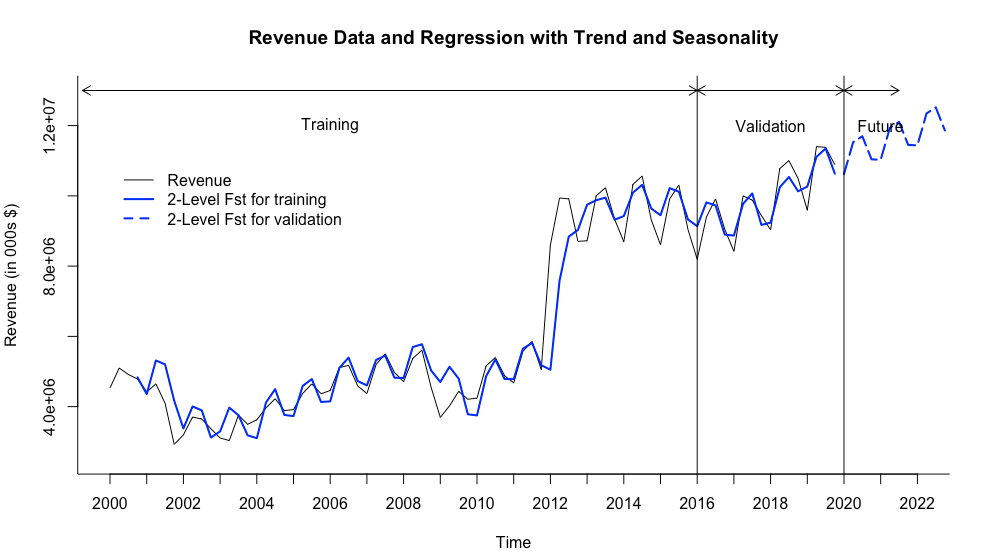
Now Developing a 2-Level forecast for the entire dataset by combining regression forecast and trailing MA forecast for residuals and presented the future 12 periods prediction below.



After that, created a table that shows the Regression Forecast, MA.Residuals Forecast, Combined Forecast for future 12 periods.

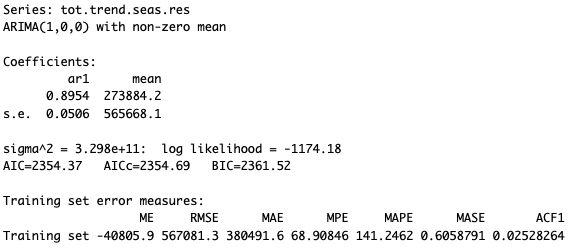


Below is the Regression with trailing MA and seasonality for the entire revenue dataset:



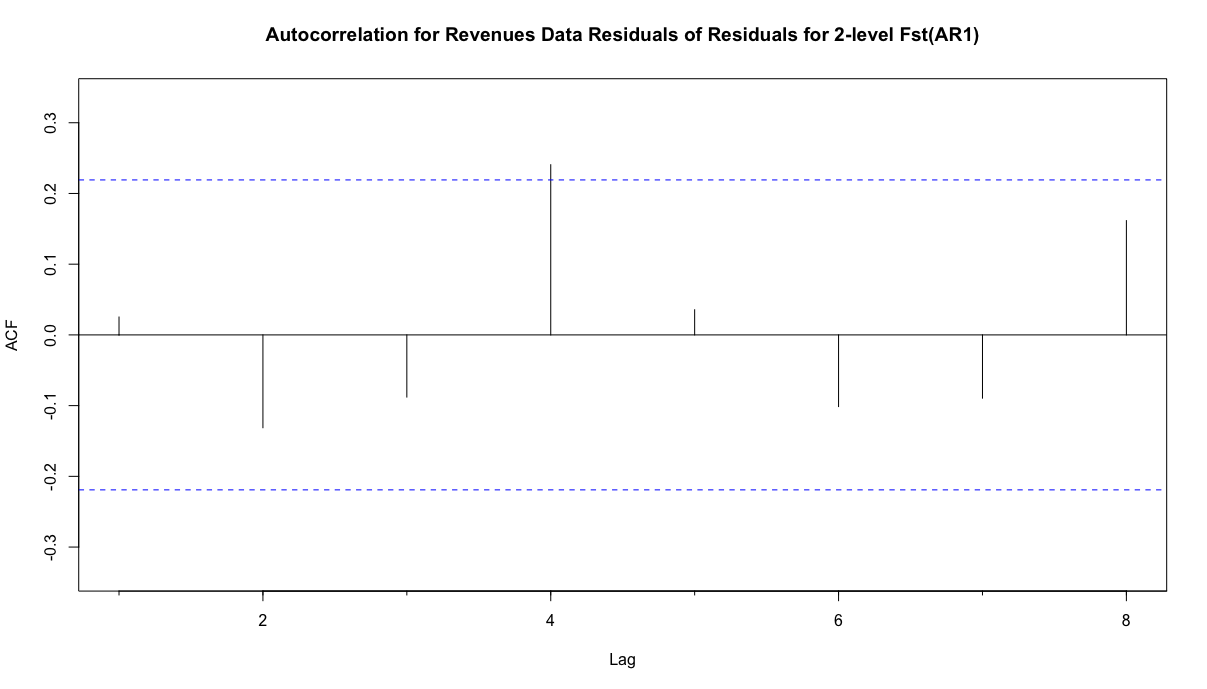
**2-Level forecasting model with Regression and AR model for entire time series dataset.**

Summary of Fitted Two Level forecasting with AR model:



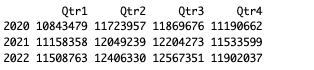
* *ARIMA (1, 0, 0)* is an autoregressive (AR) model with order 1, no differencing, and no moving average model.

Autocorrelation for the entire dataset using the 2-level Forecasting with regression and AR model.

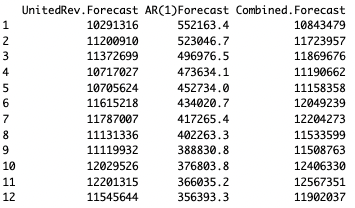


* In the autocorrelation for residuals of residuals for 2-level forecast using AR (1) model all the lags are within the significance threshold except for lag 4 which is weakly significant.
* By looking at the correlogram we can say that all the patterns of residuals are considered in this model.

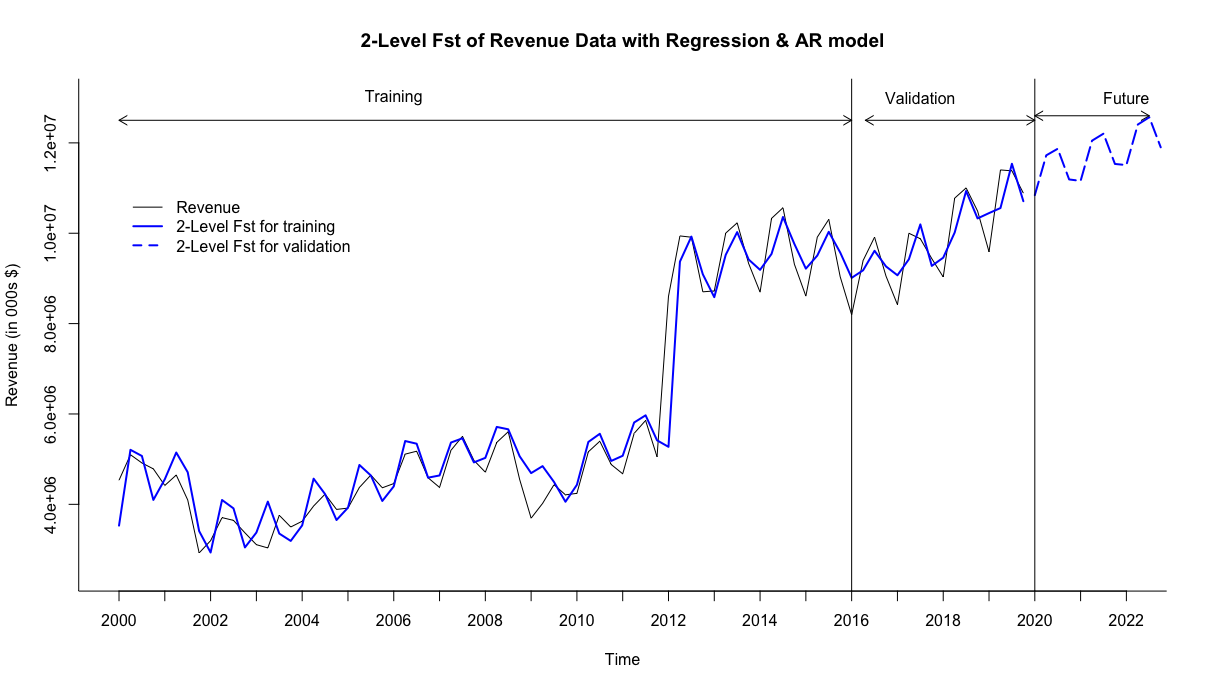
Developed a two-level model's forecast with linear trend and seasonality regression + AR (1) for residuals for future periods.



Data table with Future 12 periods data, regression forecast for Future 12 periods, AR (1) residuals for Future 12 periods, and 2-level model results.

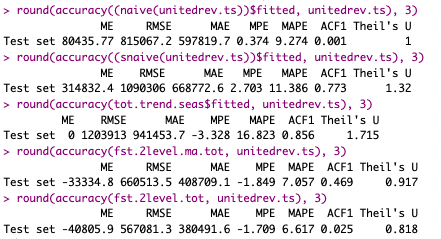


Plot for 2-Level forecast of entire revenue data with Regression and AR model.



**Step 8: Implement Forecast**

* Below are the accuracies for all the models chosen for the entire dataset and also the Naive and Seasonal Naive models.
* The least MAPE off all the models is 6.617% and the least RMSE is 567081.3
* Based on the MAPE and RMSE the best model that forecasts the future Revenue for United Airlines appears to be 2-Level Forecasting model with Regression with Linear Trend and seasonality and Auto Regressive for AR (1)

****

**Conclusion**

After considering various models for forecasting for United Airlines Revenue, the best model based on least RMSE and MAPE is 2-Level Forecasting model with Regression with Linear Trend and Seasonality and Auto Regressive for AR (1). However, the other models considered specifically the 2-Level Forecasting model with Regression with Linear Trend and seasonality and MA can also be used to forecast. Thus, it is important for the forecasting team to semiannually inspect the performances of these 2 models as data gets updated with new quarters and use the right forecasting model. This constant checkup will help the team to improve the forecasting.

**Appendix**

1. With the help of the ARIMA () function, we test the predictability of the United Airlines Revenue dataset to fit the AR (1) model. This is tested for the entire dataset with the beta coefficient and the standard error is considered with the alpha value of 0.05.

