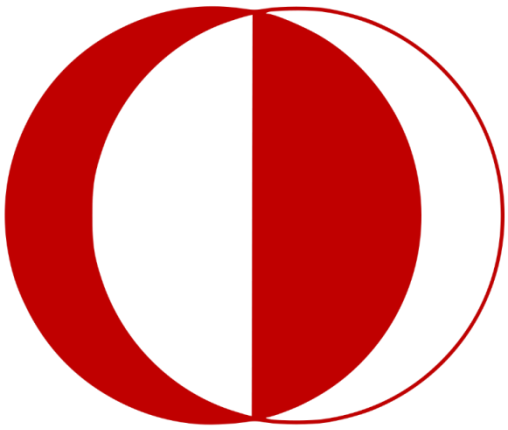


MIDDLE EAST TECHNICAL UNIVERSITY

2021

Department of Economics

TERM PROJECT

Part1

Time Series Analysis on CO2 Dataset

ECON453: BUSINESS FORECASTING

SPRING 2021

SUPERVISOR : NAZIM KADRİ EKİNCİ

Prepared by:

ZEYNEPSU KESİM – 2292340

The project is about choosing the best forecast model for the given dataset. The dataset is a monthly time series data between the years 1959 and 1998 about the Mauna Loa Atmospheric CO2 Concentration. Three main models are considered and analyzed. The last 24 observations are used for the test set while the other observations are considered as the training set to train the model. The analysis is made by using R with the packages of "baseR" and "astsa".

From the data, it can be seen a strong upward trend and a highly seasonal pattern. The seasonal pattern looks pretty much the same meaning could have almost the same variance. Therefore, even though firstly converting the log form seems redundant because of the constant pattern of the seasonality, after the comparison I made on the models (linear regression models), I decided to use the log form of the data since the adjusted R-Squared is better on the log form. Without the log, the adjusted R-Squared is 0.987 while with the log form it is 0.99. It does not differ too much, but better model is better forecasting. Therefore, throughout this paper, I use the log form of the data for estimation and forecasting.

Linear Model

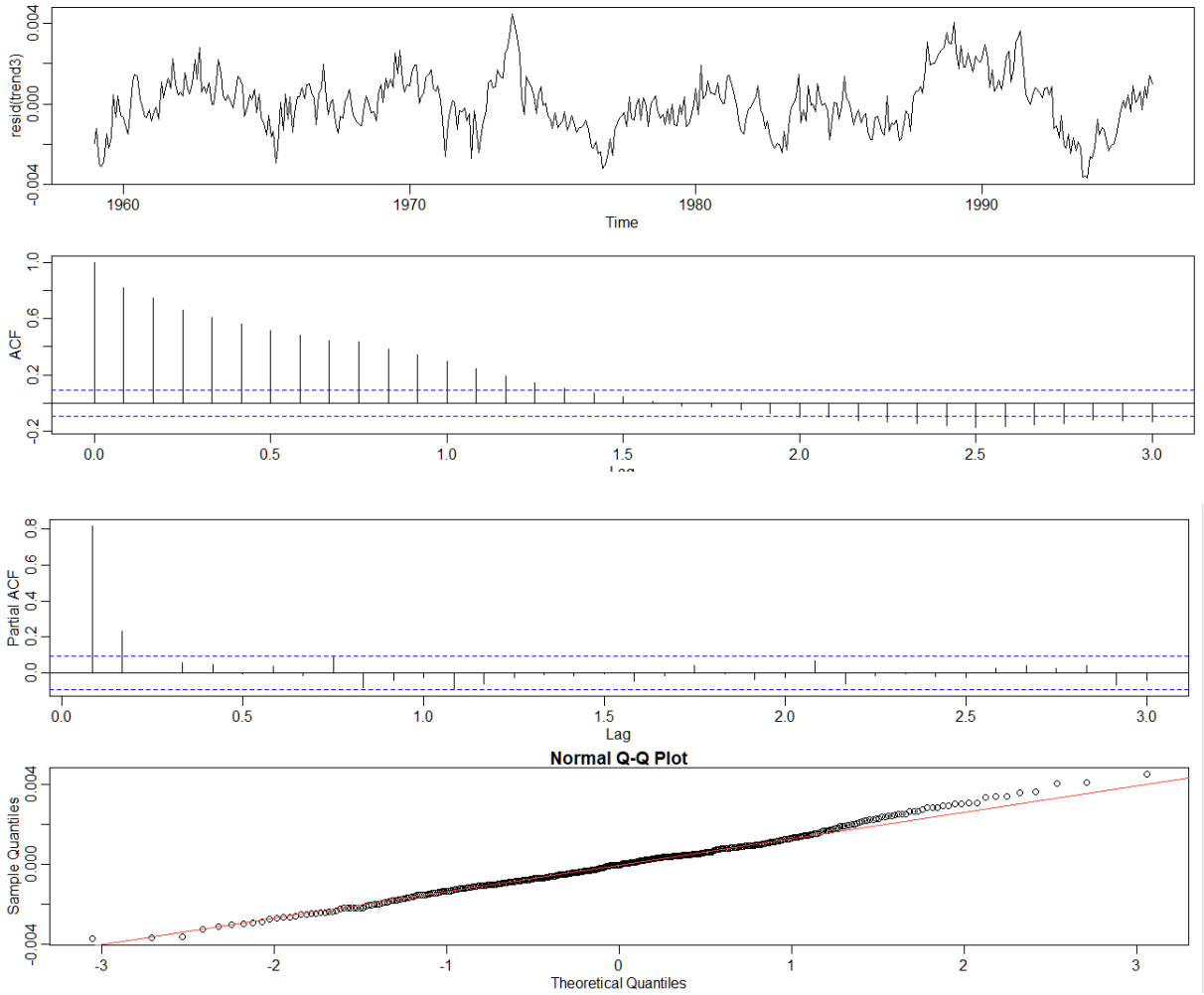
I use dummy variables for months to handle with the seasonal component, and I tried different time trends to handle the best with the upward trend. Mainly I tried 3 different time trends namely linear trend, quadratic trend and cubic trend.

The first model to be considered is the linear trend with dummy variable. The r-squared is 0.987. When we plot the estimated model with the real values of the training set, we can see that at the beginning and at the end of the data, we underestimated while in the middle parts we over estimated. Therefore, it is not the best model. Moreover, residuals do not look like a white noise since, we can observe a downward trend until the late 70s, and then an upward trend. By looking at the ACF and PACF, we can say that residuals can follow an AR(1) model since it tails off in ACF, and cuts off after lag 1 in PACF. Moreover, they are not normally distributed.

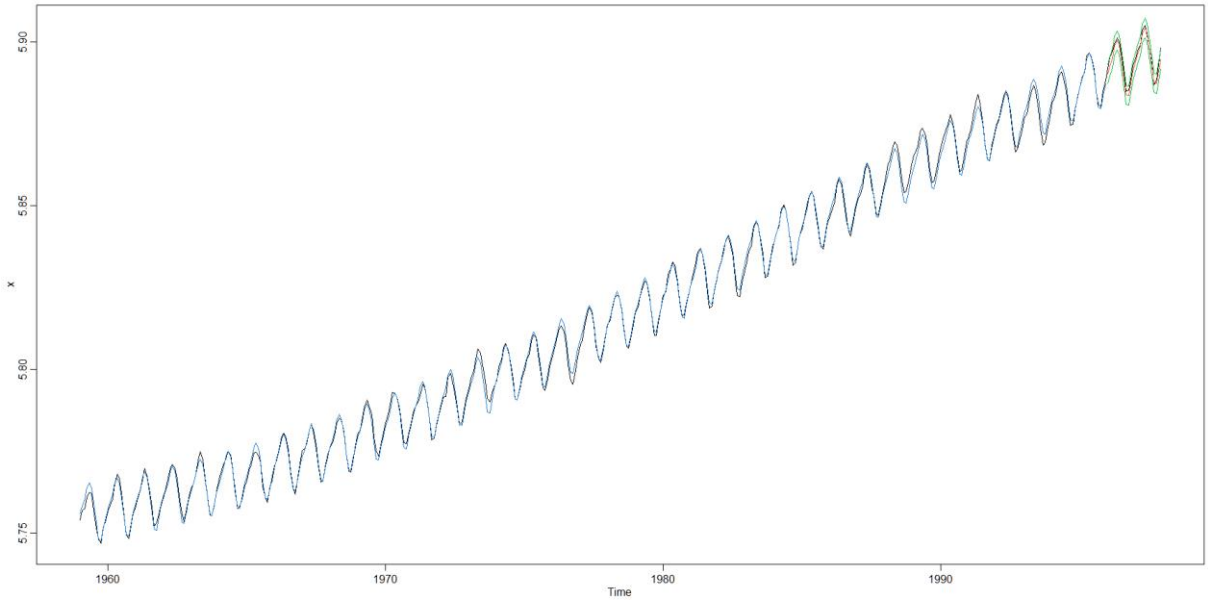
The second model is with the linear and quadratic trend. Adjusted r-squared is better, 0.9975. The coefficients of t and t^2 is both significant (by looking at the t values), therefore we include both of them. Moreover, by looking at F-value, we can say that all variables in the model jointly significant to. Thus, it is a better model than the first one. When we plot it, we the under estimation at the beginning and over estimation in the middle is almost gone. After 1990, there are little misestimations, but they can be small enough to ignore. The residuals are better, and there exist evidence that still supports that the residuals might follow AR(1) process. Furthermore, residuals are distributed precisely according to normal distribution but still not that good.

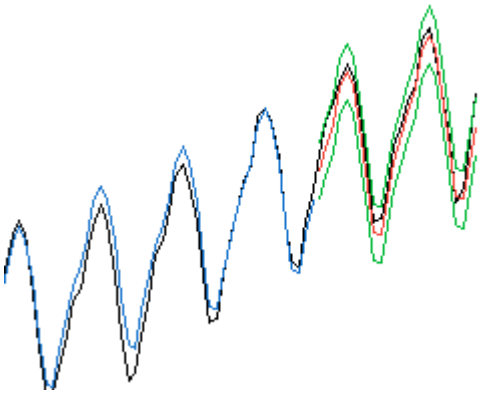
The third model is with the linear, quadratic and cubic time trend. Adjusted r-squared is better, 0.9988. Thee coefficients of t and t2 and t3 are both significant (by looking at the t values), therefore we include both of them. By looking at F-value, we can say that all variables in the model jointly significant to. ACF and PACF of the residuals are better than the other two models. Residuals look more like a white noise than the other two. Jigglier residual graph we have, there are still some little trends but not that much dominant than the other two. It looks

like the best estimation is with the cubic trend. The below graphs contains information about the residuals.



When we compare the forecasts of the above three models, in the first model our fitted values for the forecast are underestimated compared with the test set values while in the second model, they are overestimated. However, in both of the models, the test set values are in the prediction intervals according to our forecast. The first model has RMSE=0.005122 and the second model has RMSE=0.004434. As the trend increases, the RMSE falls. In the third model, which is with the cubic trend, we almost have the same fitted values with the real values in the test set. Our model captures the pattern well, with RMSE=0.001510, very close to 0.





Above graph shows the forecast for linear regression with cubic trend and dummy variable model. The black line is the real values in the data set; the blue line is our estimation according to the training set; the red line is the forecasted values for the tests set; the green lines are the prediction intervals.

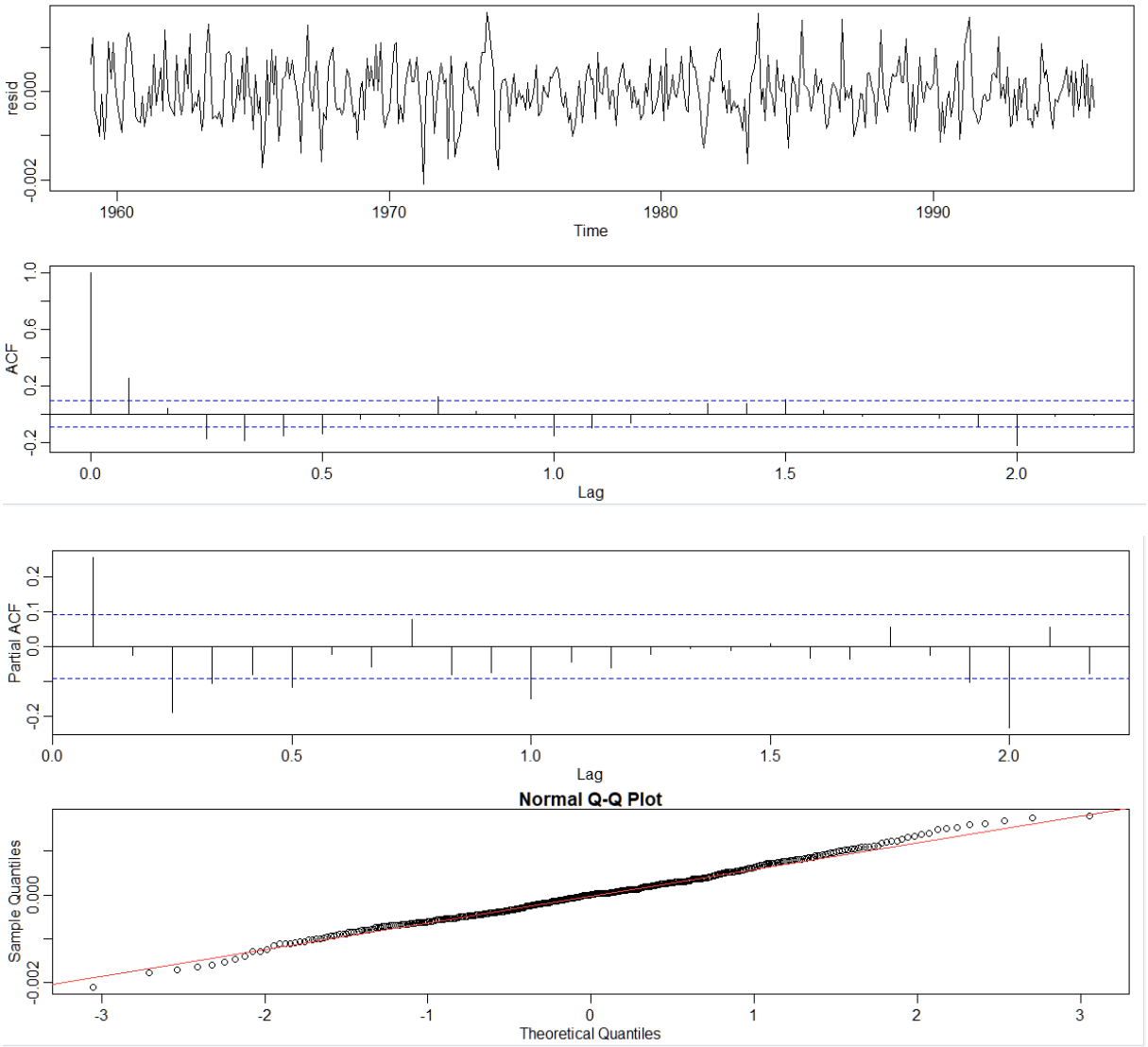
Therefore, as we can see from the above graph, the forecast is too close to the original values, and since it has the best RMSE, the best model is the third model which is with the cubic trend and dummy variables.

Model with STL

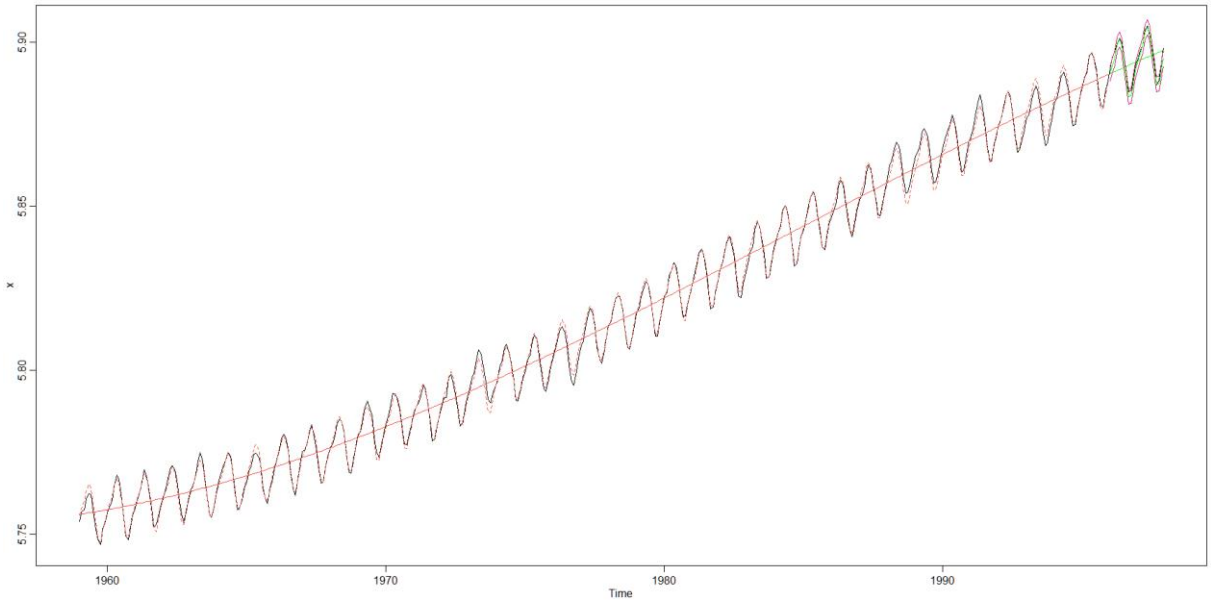
By using STL function, I decompose the training set. I use the s.window = 13 since the seasonality is pretty much the same. We use s.window=7 to capture the differences better if we know that seasonality increases or decreases over time. Nevertheless, the seasonality is almost the same here. Moreover, since we have the monthly data, it is better to use 13 since there are 12 months. In STL function, the seasonal part is not same for all the years-in the regression case we assume the seasonal pattern is the same for all years. This helps us to make our forecasts better.

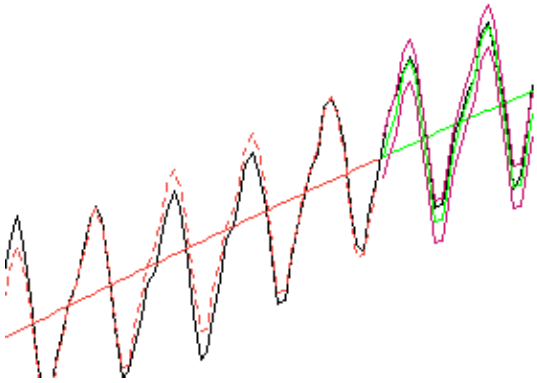
From the below graph, information about the residuals can be found. Residuals might be following a seasonal ARIMA model since in the PACF, we see that at lag 1 and 2, it cuts off. Moreover since it tails off, there can be an MA process inside the ARIMA. The plot of the

residuals is better, looking jiggly while the distribution is almost the same with the regression case.



Because from the linear regression case, we found out that the model with the cubic trend is better, to re-estimate the trend, I again use linear, quadratic, and cubic trend together (I know all of them are significant according to their p-values). After I re-estimate the trend, I add the seasonal component, and plot the estimated values with the real values of the training set. The estimation is quite good (you can find the graph below). In some years, we still see overestimation and underestimation. However, as a whole, the model fits well. To forecast, I used last two years' seasonal patterns. Since the seasonal patterns is not highly volatile over time, the mean values can also be used. However, the model with the last two years' values fitted better. After the forecast, we see that the forecast is almost the same with the actual data in the test set. The RMSE=0.001445 which is better than the best model in regression case.





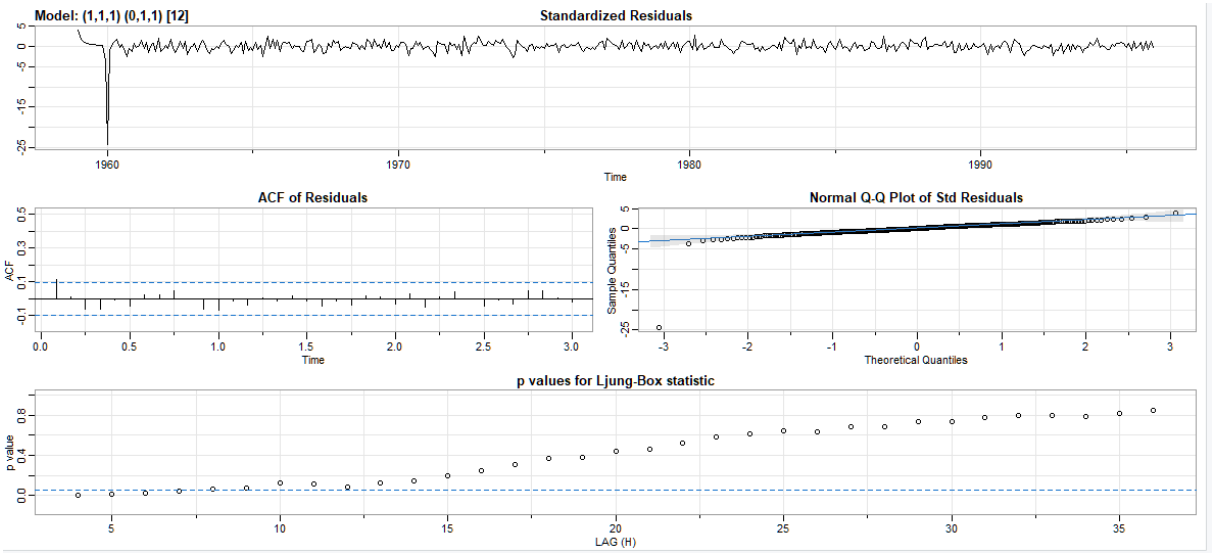
(To see graphs better, the code I supply can be used.)

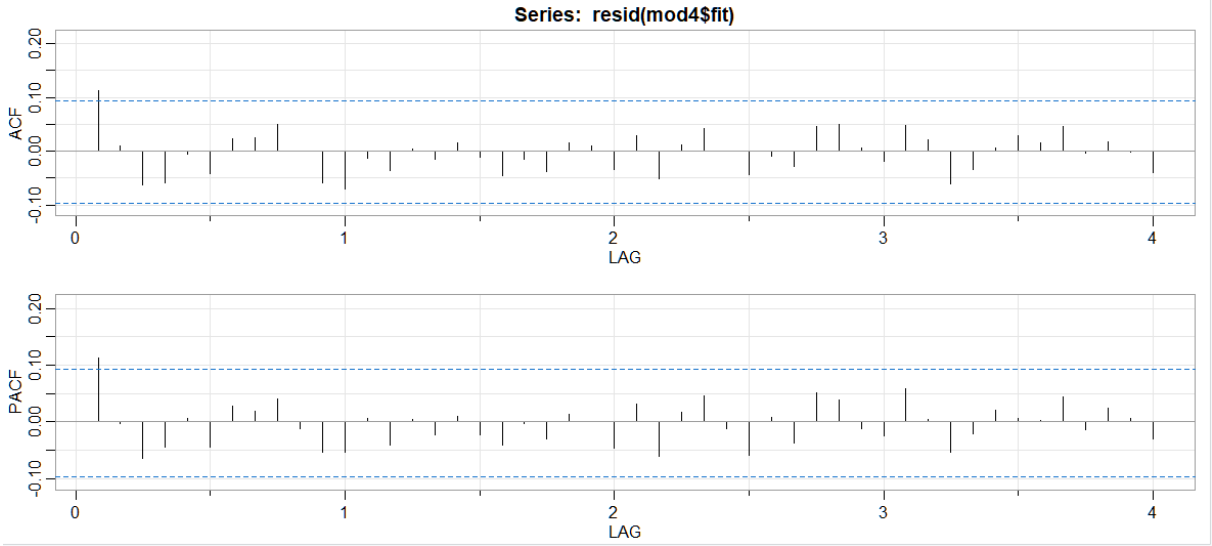
The red lines are our estimates according to the training set, and the green lines are our forecasts; pink lines are the prediction intervals, and the black line represents the real values. The straight lines represents the estimated and forecasted trends(red and green, respectively)

ARIMA Model

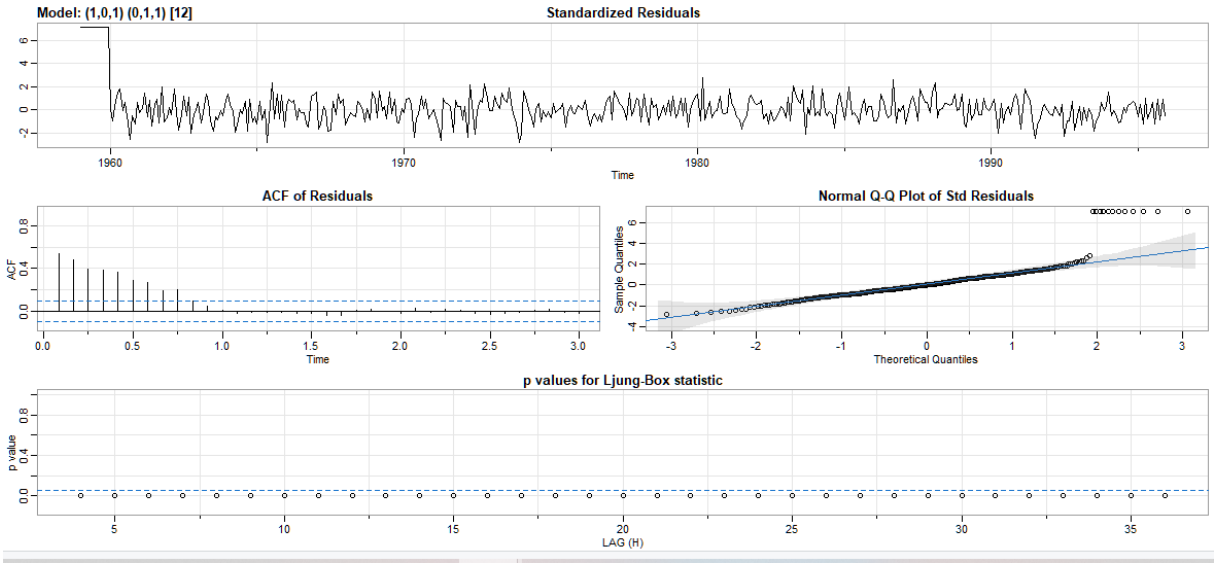
Before I started to try models, I analyzed the data set again to find out which models can be suitable. The data set consists of monthly data, and strong seasonality. By looking at the ACF of the training set, we can say that there is a strong trend since strong autocorrelations are seen. These strong autocorrelations could be because of the strong trend element. Besides the trend, there could be an AR model. By looking at the PACF, at lag1, there is a high autocorrelation so it can be an AR(1) model. In the second lag again PACF cuts the dashed line, therefore it is significant, thus it could be an AR(2) model. After lag 2, the function still looks like it slightly tails off, thus, there could be an MA process, too. Since there is seasonality and trend, we should have detrending element which is the differencing (d=1), and for seasonality, we should have seasonal differencing (D=1).

I tried many models according to my assumptions that I discuss above. The detailed information and all the models I tried can be found in the code. I compare AICcs, ACF-PACF functions, Ljung-Box statistics, and after my analysis, my best results are seen in the ARIMA(1,1,1)(0,1,1) with AICc= -11.00773, and ARIMA(2,1,1)(0,1,1) with AICc= -10.82013. Since the first model has a lower AICc, I choose it as my best.

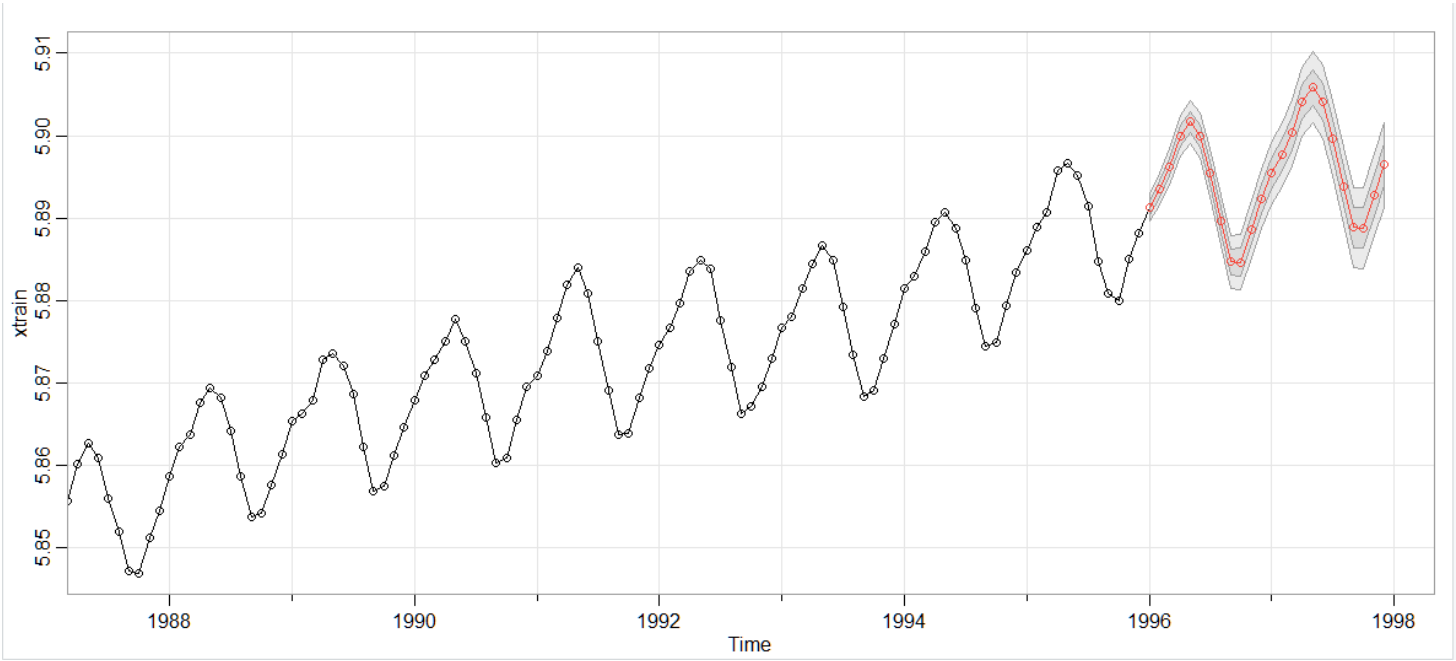




Then I tried ARIMA models with external regressors. I tried two external regressors one with quadratic term and one with cubic term. Since coefficient of the cubic term is insignificant, I decided to use the quadratic term. After trying various models (again could be found in the code), I chose the best model as ARIMA(1,0,1)(0,1,1) with external regressor (quadratic trend). However, the model without the external regressor is far better. We can say that even by looking at the residual graphs. The best model that I can found has very bad Ljung-Box statistics and worse distribution.



I forecasted, two of the above models, the forecast for the ARIMA model without the external regressor is very well. The red line below represents the forecasted values while the



grey region is the prediction intervals. This is the ARIMA(1,1,1)(0,1,1) model with RMSE=0.00101131, which is the best result comparing with the other two models (linear regression and STL).

The forecast for the external regressor was not bad, too. It was the model ARIMA(1,0,1)(0,1,1) with quadratic trend, having RMSE= 0.002229901. The RMSE is still quite good, but we have better models.

In conclusion, we can say that between these three models, the best model for forecasting this dataset is ARIMA(1,1,1)(0,1,1) after analyzing many components of the models.