ECO374 Essay - Analyzing and Forecasting Unknown Time Series Data

Introduction

In this short essay, we will explore an unknown monthly time series data and formulate a 12-step ahead forecast. We will begin by analyzing the original data, and necessarily transforming it to perform relevant statistical analysis. This will involve detrending the series using an optimal linear model and extracting its residuals. Once we have successfully formulated predictions for the deterministic component and our residual series, we will generate the forecast for the original series. Ultimately, we hope to obtain predictions that are consistent with the data we observe and result in a low forecasting error, measured by the sum of squared errors (SSE).

Setup and Data

The first priority when dealing with any time series would be visualizing the data. In order to fit a time series model, we need to make sure that our data is stationary. The existence of trends in a series will violate the assumption stationarity. Plotting the original series can therefore help us detect potential trends and patterns in the data. If the data is indeed stationary, we would expect our variable to be randomly oscillating values across time. Figure 1 illustrates our original series, which has a total of 180 monthly periods. We can observe a very distinct pattern that seems to be occurring every year (12 periods).

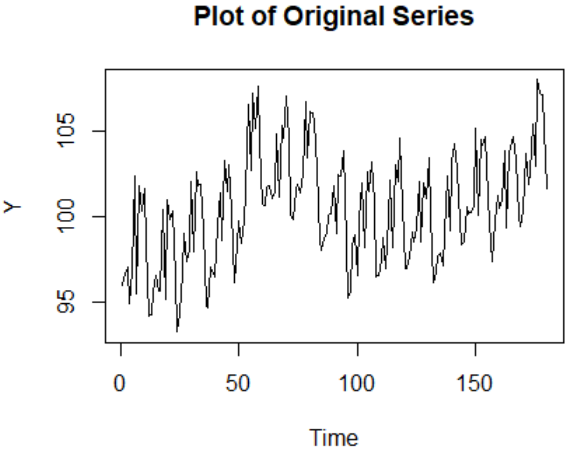


Figure 1: Plot of the original time series

This indicates that we may be dealing with some sort of seasonal pattern in the data. Furthermore, we also observe that the average of the series has been slowly increasing, and thus we can infer the existence of a deterministic trend. From a quick visual inspection, the original data cannot be stationary due to the reasoning above. We can confirm our belief by running the Augmented Dickey-Fuller (ADF) test, which tests the null hypothesis that a unit root is present in our time series sample (i.e. series is not stationary) against the null hypothesis that the series is stationary. The result of ADF tells us that the data is indeed non-stationary, given the large p-values for all lags up to 12. We can further have confirmation with the sample auto-correlation plot of the series (Figure 2). Below, we can observe peaks at annual lags, i.e. every twelve lags. This directly confirms our belief that seasonal effect is present in the series, and that the series is not stationary.

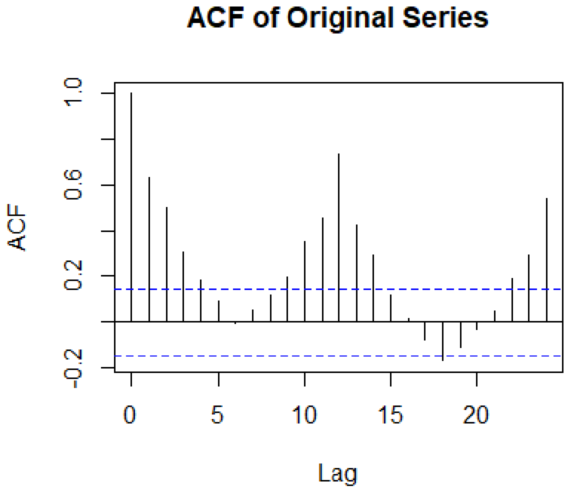


Figure 2: Plot of the sample auto-correlation function

We checked that our data is non-stationary, and now our task now is to detrend the original series by regressing time polynomial of order k as dummy variables and extract the fitted residuals. As mentioned above, we are observing a mixture of a deterministic trend and seasonal variations. To account for the seasonality, we must decide whether the seasonality is deterministic or stochastic. Deterministic seasonality would be characterized by constant magnitude of fluctuations at each time t at the annual level, while stochastic seasonality would exhibit non-constant magnitudes of the fluctuation. This is particularly difficult to assess visually with our series. We found that incorporating deterministic seasonality, as opposed to its stochastic variation, resulted in a doubtful, explosive forecast that seemed too inconsistent with the pattern in our data. Therefore, we choose to incorporate stochastic seasonality in detrending our original series; and as we will see in later section, this led to a far more believable forecast.

We must also consider the order of time polynomials we will be using to regress the original data. In general, we want the linear model fitted to have a low AIC value, as it reflects the amount of variation captured by the deterministic component of the original series. But it is also possible for the AIC value to increase again beyond some order k. To see this, we formulate ten linear models incorporating stochastic seasonality with different polynomial orders (Figure 3). The general form of these models can succinctly be written as:

$$y_t \sim \sum_{k=1}^{k} t^k + \underbrace{y_{t-12}}_{\text{Stochastic Seasonality}} + \epsilon_t \qquad (1)$$

From Figure 3 we deduce that the marginal gains from choosing a model with $k > 6$ is negligible. Therefore, we will choose the model from (1) with time polynomial of order $k = 6$. From Figure 4, we can see that our fitted values closely follow the dynamics of the original

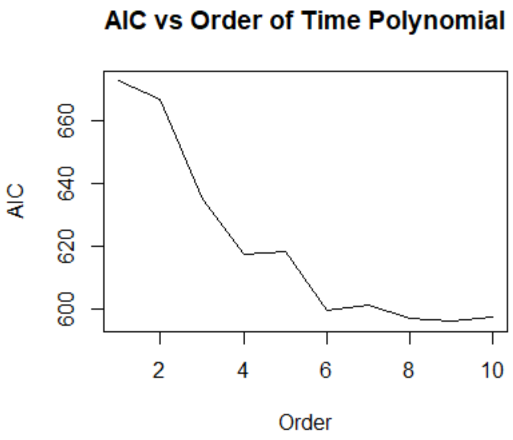


Figure 3: AIC fitted against order of the time polynomial

series. Note that because we are incorporating a 12-period lag into our linear model, we 'lose out' on the first twelve observations of the original data.

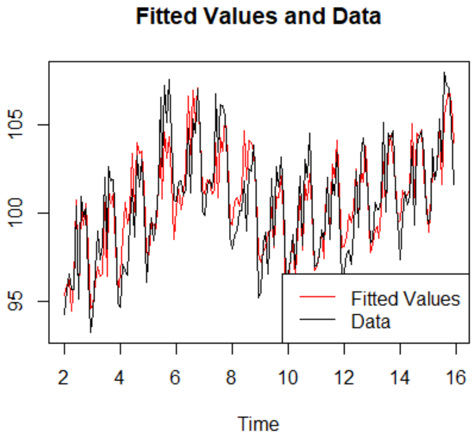


Figure 4: Plot of the original time series and fitted values from (1), $k = 6$

Now we can extract the residuals from the fitted model and use those residuals as a new series to fit a time series model. The ACF of the residuals is plotted on the left of Figure 5. There are still some seasonality effects that were not taken care of, and to remedy this, we first difference our residuals, i.e. $\delta e_t = e_t - e_{t-1}$. The right of Figure 5 shows us that any seasonal variations within the data has successfully been controlled for. Under the ADF we obtain p-values lower than 0.01 for all relevant lags, so we will believe that the data (differenced residuals) is stationary. Furthermore, under the Ljung-Box test, we find there exists non-zero auto-correlations within the series. A time series model will be needed.

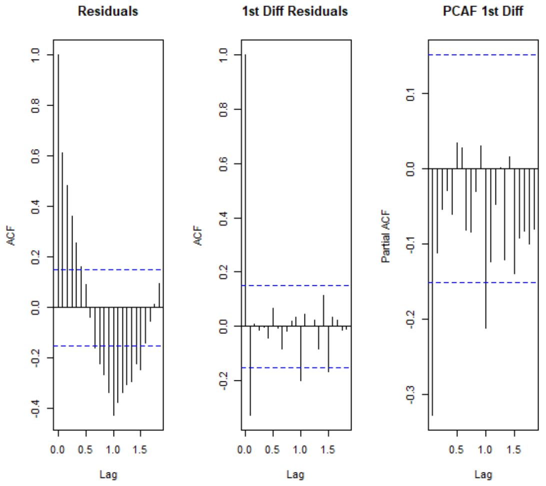


Figure 5: ACF plots of residual and its first difference

Modelling the Residuals

We have deterended the original series, extracted its residuals, and first differenced it to make sure the series is stationary and clear of any seasonal variations. Now that we have a stationary, auto-correlated time series, we can proceed to fit a time series model using ‘auto.arima()‘ function. From Figure 5 we saw that ACF cuts off after lag 1, and lags after 15 periods (months) is no longer significant. Our PACF plot also cuts immediately right off, and lags after 10 periods are no longer significant. So we will input $max.p = 15, \;\; max.q = 10$ as our inputs to the auto.arima function. We obtain a ARIMA(0,0,1) x seasonal ARIMA(0,0,1) model.

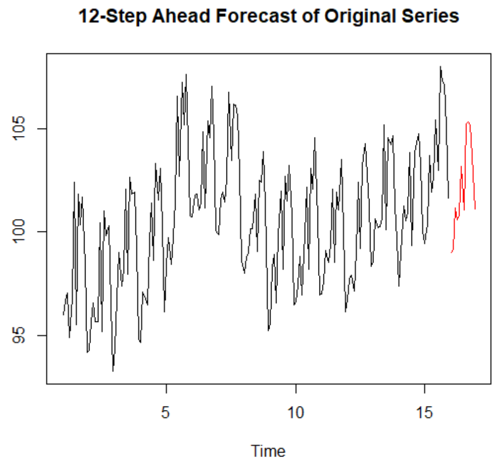
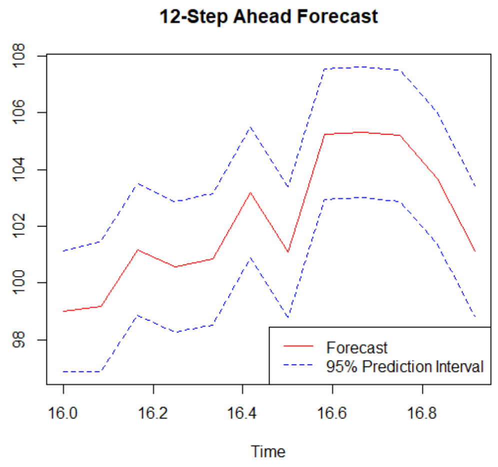


Figure 6: ACF plots of residual and its first difference



Forecast

In order to make a forecast, we must sum the forecast of the deterministic trend and the forecast of the residual series. For the latter, we must also first integrate the predictions because they were initially differenced. Figure 7 shows the 12-step ahead forecast for the residual along with the original data. Figure 8 shows specifically those 12 forecasts and their prediction intervals. The prediction intervals were built on the standard errors of the predictions and using critical values for the 95% interval. In this process, we assumed that the standard errors are normally distributed. As we can see, our forecast very closely follows the historical data. Our forecast was able to capture the the seasonal volatility present in the data.

Conclusion

In this short essay, we have analyzed, deconstructed, and devised a 12-steps ahead forecast of an unidentified time series. Our initial assessment of the data showed that the data itself was not stationary with evident effects of seasonality. We subsequently detrended the series using regression and also controlling for stochastic seasonality. Further differencing of the residual series obtained from the regression led to a non-seasonal, stationary series. Then, we predicted the differenced residuals using an ARIMA(0,0,1)xSARIMA(0,0,1) model, and constructed a prediction band based on the assumption of normally distributed errors. The forecast seems reasonable given the historical pattern of the data, and we hope our forecast results in a small prediction error.

References