DBA3803 Assignment 2 (E0200936)

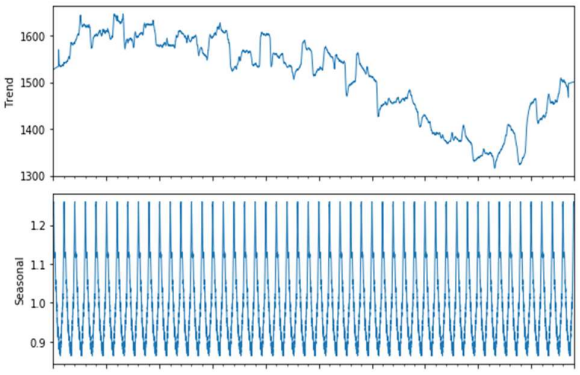
1 Introduction

This assignment report attempts to break down and analyse the trends that are present in the "England & Wales Deaths" time series data that was provided. Several methodologies that are covered include, Linear Regression, Time Series Seasonal Decomposition, as well as Exponential Smoothing, Holt's & Winters' models in order to better understand any underlying trends behind the time series data.

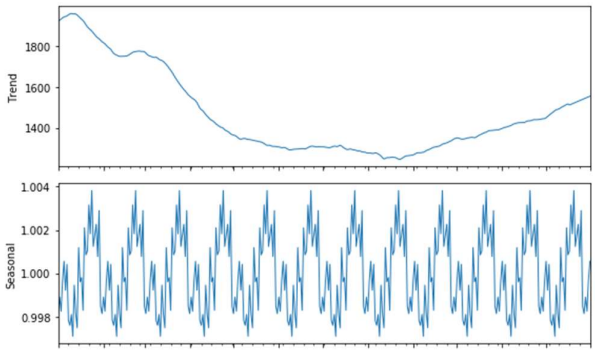
2 Linear Regression

After cleaning the data, we begin our analysis by doing a linear regression of the "England & Wales Deaths" data against the "Year" which the deaths were recorded. From a plot of the fitted regression line, we can observe a slight downwards trend over time. From the regression results summary, we gather that "Year" is a statistically significant predictor (at 5% significance level) since its p-value is less than 0.05. However, since the adjusted R-squared value is very low at 0.129, the overall regression model is not effective in predicting and explaining this downward trend. Hence, we will also have to rely on other methods to reliably establish the trend of our time series data.

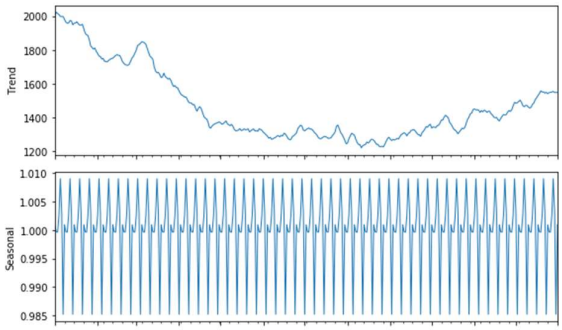
3 Time Series Decomposition



Before we begin time series decomposition, we first plot the raw data to understand the shape of our data as shown on the left. The first observation that can be made from this plot is that there is yearly seasonality in the time series data. With this in mind, we then move on to decompose the time series data based on yearly season. From the decomposition plots, we obeserve a general downward trend from 1970 to 2009 with a slight upwards trend from 2010 to 2018. This downward trend corresponds with what was observed previously in our linear regression plot, hence indicating that there might be a general downward trend in England and Wales' deaths within the time window specified by our dataset. However, the trend reversal from 2010 to 2018 might potentially suggest that this time series follows a larger cyclical pattern or a larger upward or downward trend, which extends beyond the period from 1970 to 2018. Due to insufficient data to cover a longer time window (beyond 1970 to 2018), we are unable to determine if this "trend reversal" is indeed indicative of a larger cyclical pattern, or whether it is also part of a long-term upward or downward trend.



When performing time series decomposition on a monthly basis (freq=30), taking the last 365 data points for decomposition plot gives us the graph shown on the left. We can observe that there appears to be slight seasonality across the different months in the past year. The trend plot also tells us that there is a downward trend which becomes an upward trend somewhere around Jul and Aug 2018.



Performing time series decomposition on a weekly basis (freq=7) gives us a decomposition plot of the last 365 data points as shown on the left. From the plot, we can observe that the same trend occurs with a downward trend until Jul and Aug 2018 followed by an upward trend. There also appears to be a strong weekly seasonality pattern in this case.

For all 3 of the above models, the residuals plots are all centred around 1 with an average value of 1. Considering that a multiplicative model has been used for all the above plots, this suggests that there are probably no irregularities with the above time series decompositions.

4 Exponential Smoothing, Holt's and Winters' Models

When attempting to isolate different trends from our time series data, Exponential Smoothing, Holt's and Winters' trend models were also considered in this case. All 3 models were fitted with datasets of different lengths on a weekly frequency basis, from the full dataset, to the most recent 1 year (365 data points), to the most recent 1 month (30 data points) and the most recent 1 week of the dataset (8 data points). The 3 models were also fitted to a full dataset with annual frequency to account for fitting with yearly seasonality.

	SES	Holt's (Add)	Winter's	Holt's (Mul)
1 Week (8 days)	31.022	38.776	2.227	42.721
1 Month	49.225	46.706	43.923	51.398
1 Year	60.285	60.285	58.949	59.957
All	59.978	59.978	58.568	59.975
All (freq=annual)	59.978	59.978	59.656	59.975

The fitted values of each models across different lengths of data and across both weekly and yearly seasonality were then used to compute a Root Mean Squared Error (RMSE) score to facilitate our evaluation of the best fit provided amongst the 4 models, which might highlight any possible trends in the data based on the trend and seasonality methods that are passed through the model parameters and as handled by the different models. From the summary table of the results above, we can see that Winters' Triple Exponential Smoothing Model consistently outperformed the other models across all categories of data to give the lowest RMSE amongst the 4 models. Since all of the Winters' models uses an 'additive' trend, it is likely that the upward or downward trend in our time series data is linear rather than exponential. It is also worth noting that the anomaly in Winters' model for the 1-week dataset could indicate a case of model overfitting, since there is close to no errors given by the model.

5 Further Analysis & Conclusion

Besides trend, seasonality and residual analysis, it is more interesting to understand the real world implications behind these trends. One possible explanation for the downward trend is the advancement of science and technology in healthcare, allowing human to live longer leading to less deaths over time. The annual seasonality in our dataset could also be explained by the fact that England and Wales are in the Northern Hemisphere, which experiences four seasons throughout the year. Incidentally, the peak in deaths towards the end of the year (winter season) in our dataset coincides with the yearly flu season, which tends to occur during colder months of seasonal countries since humans are more likely to come down with flu in winter. The higher incidence of flu could be part of the explanation for increased deaths towards the end of the year. In conclusion, time series data is largely limited by its inability to 'see' past trends outside of its time window and its relevance lies in its ability to relate to and explain trends and seasonality in the real world.