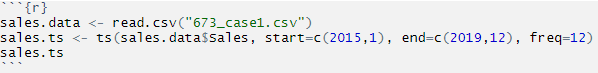
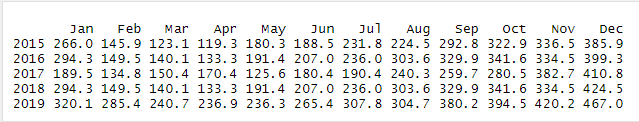
**Case Study #1: Sales Forecasting**

In this case, consider the data on worldwide monthly sales from a large chain of grocery stores (673\_case1.csv). The monthly sales are given for a period of 2015-2019 and measured in millions of dollars. The goal is to identify the best forecasting model to predict monthly sales in 12 months of 2020.

**1. Identify time series components and plot the data.**

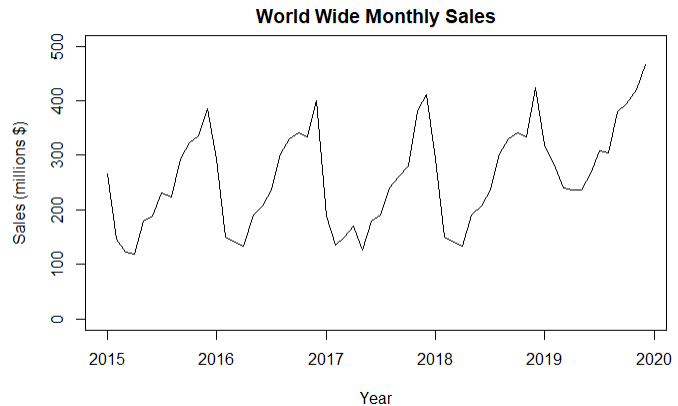
1. Create time series data set in R using the ts() function.





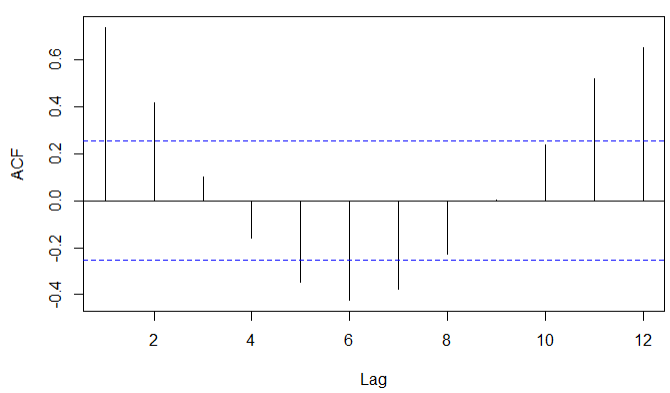
* Above shows codes using ts() for the historic sales data and the resulting output.

1. Employ the plot() function to create a data plot with the historical data, provide it in your report, and explain what data patterns can be visualized in this plot.



* At a glance, the historical sales data seems to have a repeating cyclical pattern with nearly equal sized amplitudes during each yearly interval. The trend also seem to be quite constant with the exception of a somewhat larger jump at around 2019. Overall, it would appear that the historical data has a constant linear trend with additive seasonality.

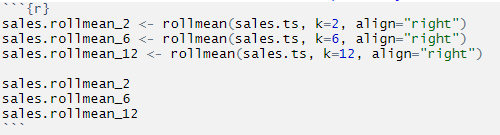
1. Apply the Acf() function to identify possible time series components. Provide in the report the autocorrelation chart and explain the time series components existing in the historical data.



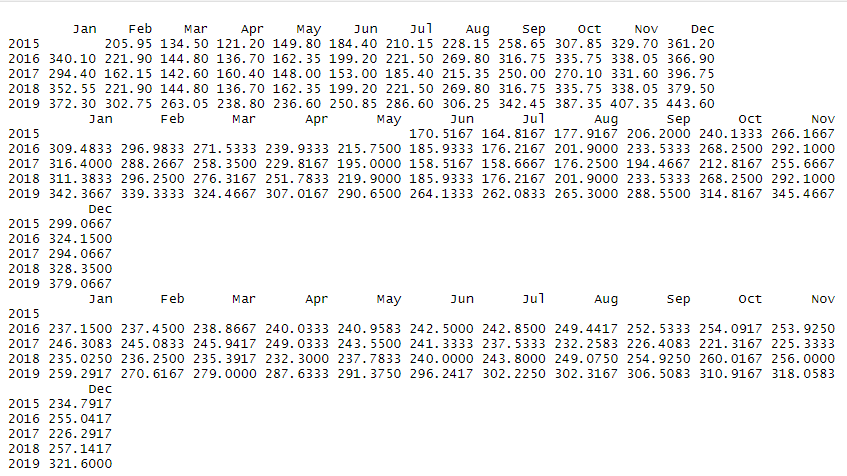
* A lag time of 12 was used because the historic sales data are monthly. As shown by the chart above, there are nonrandom correlations with the sales data at time lags of 1, 2, 5, 6, 7, 11, and 12. A relatively large positive ACF at time lag 1 in the chart above indicates that the sales data have a positive trend. Time lag 12 also has a relatively high ACF, which indicates that the sales data are seasonal.

**2. Use trailing MA for forecasting time series.**

1. Use the rollmean() function to develop three trailing MAs (apply the entire data set with no partitioning) for the window width of 2, 6, and 12, respectively. Present the R code for these MAs in your report.

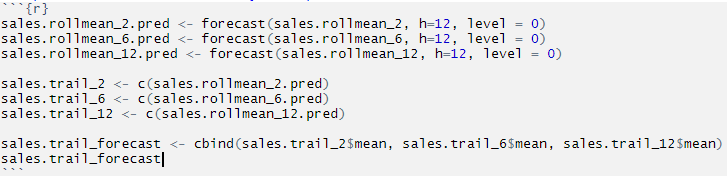


* Above shows codes for the three MAs for historic sales with window widths of 2, 6, and 12

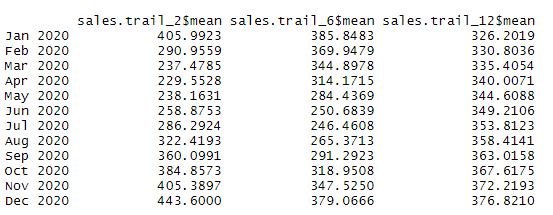


* Above shows trailing MA results for window widths of 2, 6, and 12, respectively.

1. Use the forecast() function to create a trailing MA forecast for each window width in 12 months of 2020, and present these forecasts in your report.

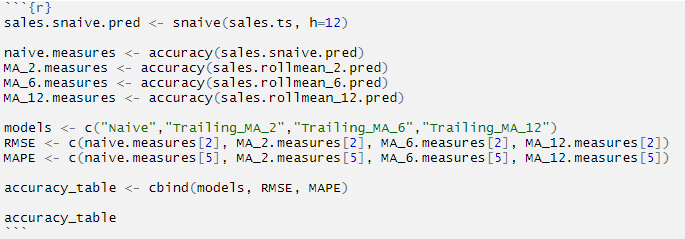


* Codes above calculates a 12 month forecast for the year 2020 for each of the trailing MAs. The forecasts are then put in a table where the results are presented.

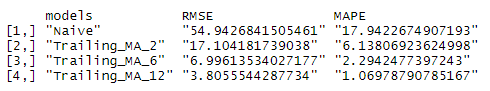


* Above shows the forecast results for the three trailing MAs with window widths 2, 6, and 12, respectively.

1. Develop a seasonal naïve forecast for the entire historical data set, and apply the accuracy() function to compare accuracy of the four models: seasonal naïve forecast and trailing MAs with window width of 2, 6, and 12, respectively. Present the accuracy measures in your report, compare MAPE and RMSE of these forecasts, and identify the best forecasting model.



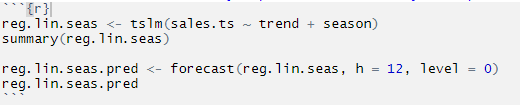
* Above are the codes used to assess the accuracy of the models.



* Above lists the accuracy measurements for the naïve forecast of the entire dataset and the trailing MA forecasts with window widths 2, 6, and 12. Each measured forecast predicts 12 months of data points. Looking at the RMSE and MAPE for each model, it is clear that all the trailing MA models performed better than the naïve model. Overall, the trailing MA with a window width of 12 has the lowest RMSE and MAPE, making it the best forecasting model out of the four models.

**3. Apply the two-level forecast with regression and trailing MA for residuals.**

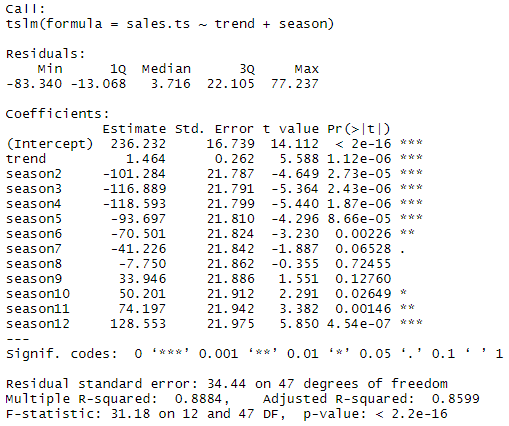
1. To de-trend and de-seasonalize the historical data for the entire data set, develop using the tslm() function a regression model with linear trend and seasonality and forecast sales in 2020 with the forecast() function. Present and briefly explain the model in your report.



* Above codes create a regression model with linear trend and seasonality using the historic sales data, and then uses the model to forecast 12 months of sales data in 2020.

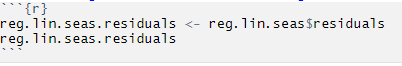


* Above shows the 12 months forecasted sales data in 2020 using the regression model.

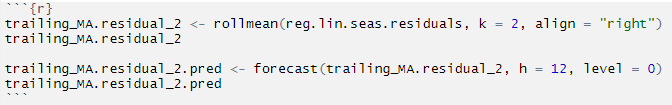


* Above is a summary of the linear model. The ~trend + season specification indicates that the model will be linear. This model was trained using the entire sales dataset. The “Residuals” portion shows the residuals in the model during fitting. This model has a high R-squared and adjusted R-squared value, as well as a good F-statistic value and a significantly low p-value; this suggests that the model fits decently well with the data. The model also contains a number of statistically significant coefficients and intercept, with the exception of the season8 and season9 coefficients.

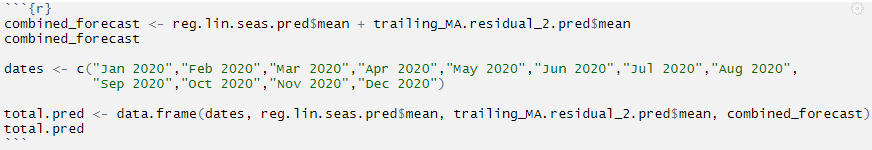
1. Identify regression residuals, apply a trailing MA (window width of 2) for these residuals using the rollmean() function, and forecast worldwide monthly sales in 12 months of 2020 (use the forecast() function). Combine the regression and trailing MA residuals’ forecast for 2020, and present in your report a table that contains regression forecast, trailing MA forecast for residuals, and total (combined) forecast in 2020.



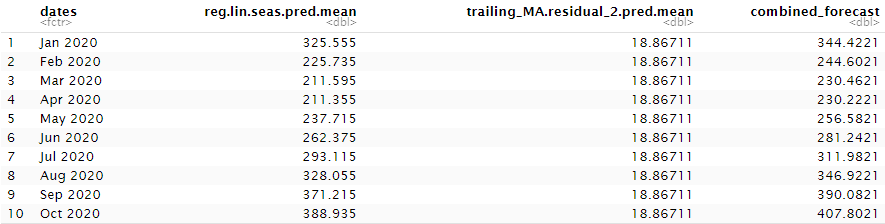
* Codes above retrieves the residuals from the regression model created in 3a.



* Codes above creates a trailing MA with window width of 2 using the regression model residuals. Then, it predicts the residuals in the next 12 months in 2020.



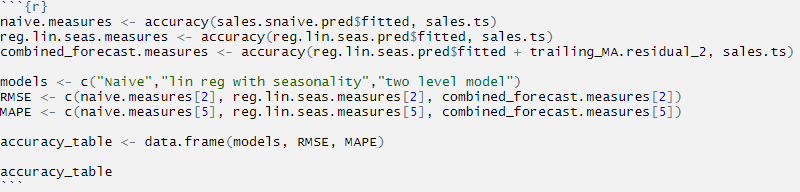
* Codes above combines the 12 month predicted residuals in 2020 with the 12 month predicted sales calculated in 3a. It then outputs a table summarizing all predicted data.



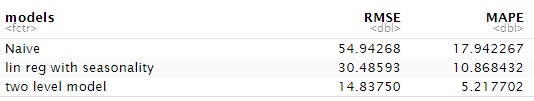


* Above output table shows the predicted sales, predicted residuals, and combined sales and residuals for the 12 months of 2020.

1. Apply the accuracy() function to compare accuracy of the three forecasting models: seasonal naïve forecast (applied in question 2c), regression model with trend and seasonality, and two-level (combined) model with regression and trailing MA for residuals. Present the accuracy measures in your report, compare MAPE and RMSE of these forecasts, and identify the best forecasting model.

****

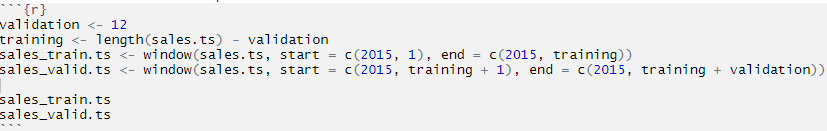
* Above codes are for measuring the accuracy of the three models: the naïve seasonal model, the linear regression with seasonality model, and the two level regression and trailing MA for residuals model.

****

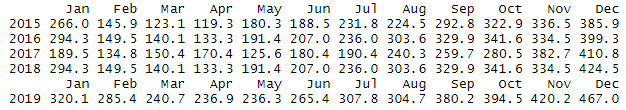
* The seasonal linear regression model and the two level combined model both outperforms the Naïve model. In this case, the two level model outperforms the rest in terms of having minimal RMSE and MAPE. Hence, the two level combined model is the best forecasting model of the three.

**4. Use advanced exponential smoothing methods.**

1. Develop data partition with the validation partition of 12 historical periods and training partition of the rest of the historical periods. Present in your report the data associated with each partition.

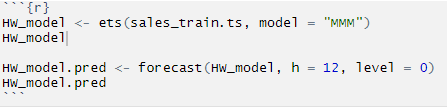


* Above codes are for setting aside 12 historical sales periods as validation data, and the rest as training data.

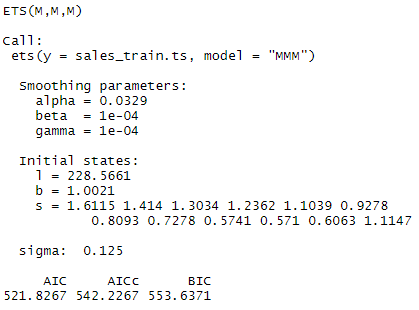


* Above shows the result of the partitioning. This output includes the training (top set) and validation (bottom set) data.

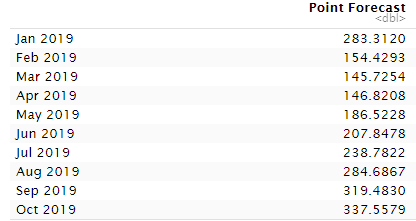
1. For the training partition, use the ets() function to develop a Holt-Winter’s model with multiplicative error, multiplicative trend, and multiplicative seasonality options, and automated selection of smoothing parameters for the training partition. Present and explain the model in your report. Use the model to forecast worldwide sales for the validation period using the forecast() function.



* Above codes are for creating the Holt-Winter’s model, and then forecasting for the validation period.



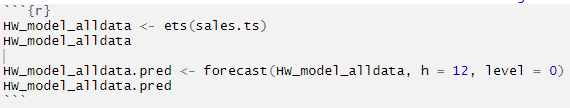
* The “MMM” for the model indicates that the model has the multiplicative error, multiplicative trend, and multiplicative seasonality specification type. This model will train using the training data, which includes all sales data from the beginning of 2015 to the end of 2018. It contains three smoothing parameters, alpha (0.0329), beta (1e-04), and gamma (1e-04), which were auto-estimated by the model.



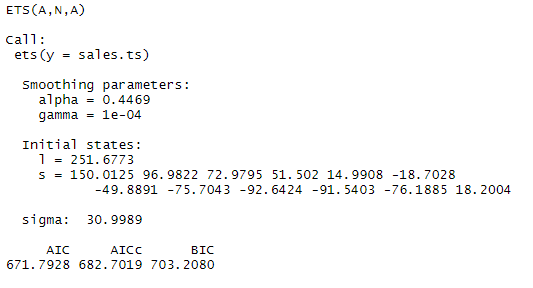


* Above are the 12 months predicted sales in 2019 using the Holt-Winter’s model that was trained with the training partition.

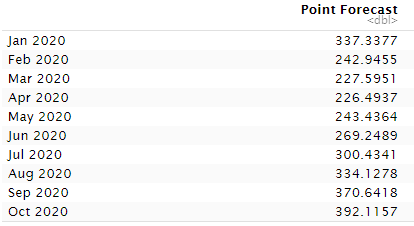
1. To make a forecast for the 12 months of 2020, use the entire data set (no partitioning) to develop the Holt-Winter’s model using the ets() function with the automated selection of error, trend, and seasonality options, and automated selection of smoothing parameters. Present and explain the model in your report. Use the model to forecast worldwide sales for the 12 months of 2020 using the forecast() function.



* Above are the codes used to create a Holt-Winter’s with automatic selection for all paramaters



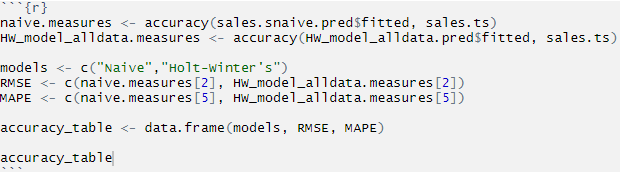
* This model was trained with all historic sales data. Because no parameters were specified during the initiation of this model, all parameters were auto-estimated. This model specification is “ANA”, which initiates to the additive error, no trend, and additive seasonality specification type. This model only consists of two smoothing parameters: alpha (0.4469) and gamma (1e-04).



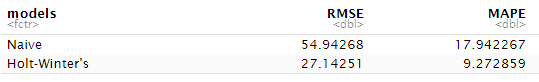


* Above are the 12 months predicted sales in 2020 using the Holt-Winter’s model that was trained with the entire historical sales dataset.

1. Apply the accuracy() function to compare the two forecasting models: seasonal naïve forecast (applied in question 2c) and Holt-Winter’s model developed in question 4c. Present the accuracy measures in your report, compare MAPE and RMSE of these forecasts, and identify the best forecasting model.

****

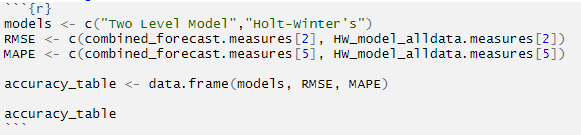
* Above codes are for measuring the accuracy of the two models: the naïve model and the Holt-Winter’s model.

****

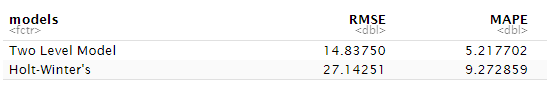
* The accuracy results clearly shows that the Holt-Winter’s model has lower RMSE and MAPE. Going off of these two metrics, the Holt-Winter’s model is the better forecasting model.

1. Compare the best forecasts identified in questions 3c and 4c. Explain what your final choice of the forecasting model in this case will be.

* The two level model was the best performing model in 3c, and the Holt-Winter’s model was the best performing model in 4c.



* Above are the codes to create a table to compare the two level model and the Holt-Winter’s model.



* Both the RMSE and MAPE are lower in the two level model when compared to the Holt-Winter’s model. Using only these two metrics, the two level model is clearly the better performing model.