**Time Series Analytics Case Study**

**Forecasting Walmart’s Revenue**

1. **Plot the data and visualize time series components.**

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

revenue.ts <- ts(revenue.data$Revenue,

start = c(2005, 1), end = c(2020, 2), freq = 4)

revenue.ts

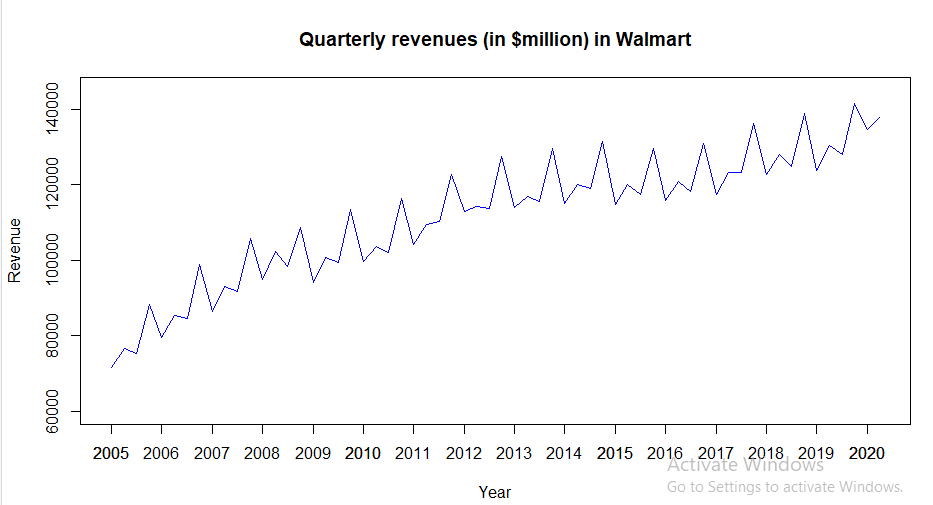
Here, we have loaded the dataset in R by using the read.csv() function followed by creating a time series of the revenue data using the ts() function for years from 2005 through 2020 with a frequency of 4 to capture quarterly data.

**1.b Apply the plot() function to create a data plot with the historical data**

plot(revenue.ts, xlab = "Year", ylab = "Revenue", ylim = c(60000, 145000),

main = "Quarterly revenues (in $million) in Walmart", col = "blue")

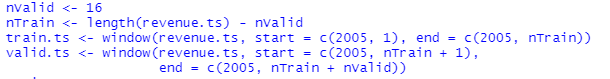
axis(1, at = seq(2005, 2020))

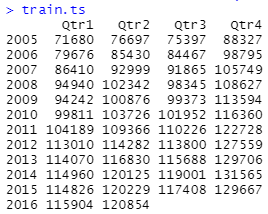
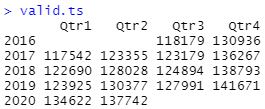


From the above plot of Walmart’s revenue, it shows that there’s an upward trend in the first and third quarters while the even quarters are shown by a downward trend in each of the years from the time series data. This mean that there is a seasonality pattern with highest revenue in the beginning of each year and low revenue towards the end of the year. There likely exists a seasonal non- linear component with multiplicative trend in this revenue times series data plot, with certain flatness from 2013 – 2017.

**2 Apply five regression models using data partition**

**2.a Develop data partition with the validation partition of 16 periods and the rest for the training partition.**





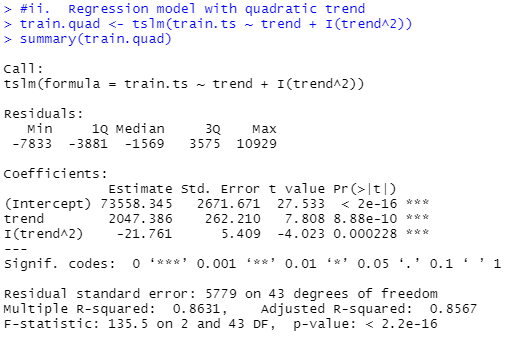
The data is partitioned with the validation of 16 periods and the rest for the training partition as above

**2b. Use the tslm() function for the training partition to develop each of the 5 regression models from the above list. Apply the summary() function to identify the model structure and parameters for each regression model, show them in your report, and also present the respective model equation. Use each model to forecast revenues for the validation period using the forecast() function.**

For the asked question, based on the patterns that are reflected in data plot for Q1, it is reasonable to consider a regression model that incorporates a non-linear trend and seasonality. Hence, following are the models for forecasting:

* Regression model with quadratic trend
* Regression model with seasonality
* Regression model with quadratic trend and seasonality
* Regression model with linear trend and seasonality (Since seasonality component is observant)

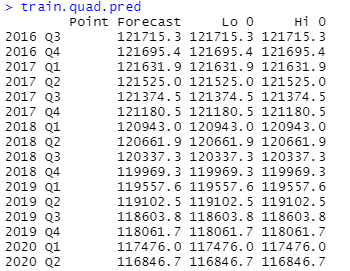
**Regression mode with quadratic trend**



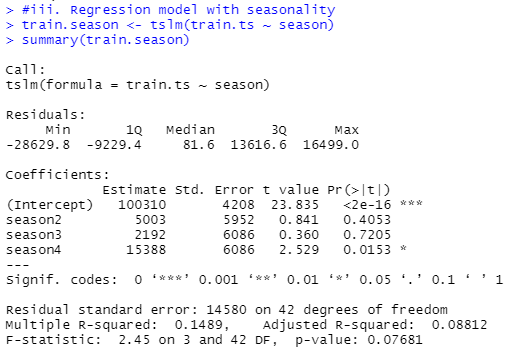
Based on the summary of regression model with quadratic trend, it contains two variables that are of independent nature: period index (t) and squared period index (t2). The model’s equation is:

*yt = 73558.345+2047.39 t – 21.76 t2*

It can be observed that this model is statistically significant as the R squared is high with 0.8631. But the coefficient for t and t2 variables are insignificant for the reason p-values are above 0. Along with that F-statistic is also statistically significant (p-value is less than 0.01). Hence, we can conclude that, this regression equation may be included for further consideration as it would be appropriate for time series forecasting. Following is the forecast for the validation period:



**Regression model with seasonality**

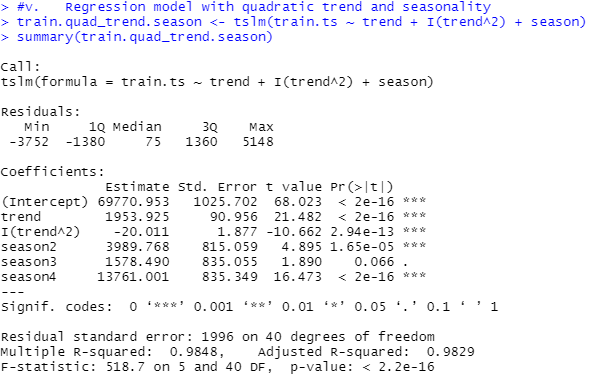


Based on the summary of regression model with seasonality, it contains three variables that are of independent seasonal and dummy in nature for Q2 (season2 – D2), Q3 (season3 – D3) and Q4 (season4 – D4). The model’s equation is:

*yt = 100310 + 5003 D2 + 2192 D3 + 15388 D4*

It can be observed that this model is not statistically significant as the R squared is low with 0.1489, and

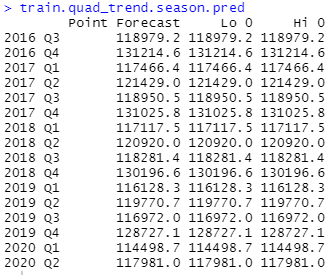
F-statistic (p-value is substantially greater than 0.01), intercept and two regression coefficients for D2 andD3 being statistically insignificant (p-values for both coefficients are above 0.01). The p-value for D4 coefficient is slightly above 0.01, and hence it is statistically insignificant. It we can conclude that, this regression equation may be excluded from further consideration as it would not be appropriate for time series forecasting.

**Regression model with quadratic trend and seasonality**

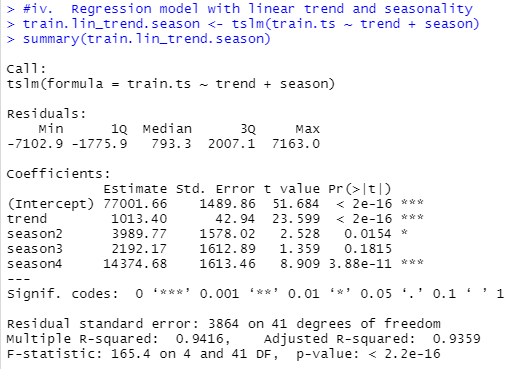
Based on the summary of regression model with quadratic trend & seasonality, it contains five variables that are of independent while there exists 3 dummy variables in nature for Q2 (season2 – D2), Q3 (season3 – D3) and Q4 (season4 – D4) along with period index (t) and squared period index (t2). The model’s equation is:

*yt = 69770.95 + 1953.92 t – 20.01 + 3989.76 D2 + 1578.49 D3 + 13761 D4*

It can be observed that this model is statistically significant as the R squared is high with 0.9848, and statistically significant F-statistic (p-value is substantially lower than 0.01). It can also be observed that most regression coefficients including intercept being statistically significant (for p-value <0.01) except for regression coefficient for D3 (which is statistically significant for p-value <0.05). we can conclude that, this regression equation may be included for further consideration as it would be appropriate for time series forecasting. Following is the forecast for the validation period:



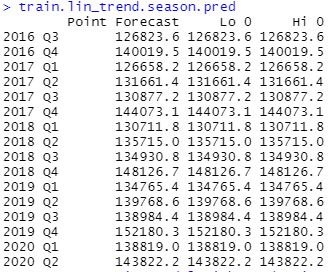
**Regression model with linear trend and seasonality (This can be considered as optional, since seasonality)**



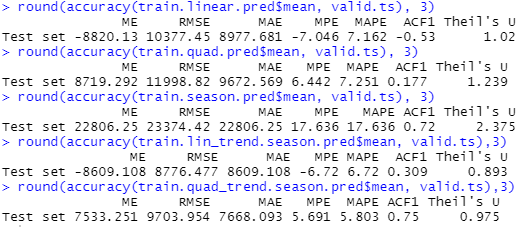
Based on the summary of regression model with linear trend & seasonality, it contains four variables that are of independent while there exists 3 dummy variables in nature for Q2 (season2 – D2), Q3 (season3 – D3) and Q4 (season4 – D4). period index (t) and squared period index (t2). The model’s equation is:

*yt =77001.666– 1013.40 t + 3989.77 D2 + 2192.17 D3 - 14374.68 D4*

It can be observed that this model is statistically significant as the R squared is high with 0.9416, and statistically significant F-statistic (p-value is substantially lower than 0.01). It can also be observed that most regression coefficients including intercept being statistically significant (for p-value <0.01) except for regression coefficient for D2 (which is statistically significant for p-value <0.05). It we can conclude that, this regression equation may be included for further consideration as it would be appropriate for time series forecasting. Following is the forecast for the validation period:



**2c Apply the accuracy() function to compare performance measure of the 5 forecasts you developed in 2b. Present the accuracy measures in your report, compare them, and, using MAPE and RMSE, identify the two most accurate regression models for forecasting.**

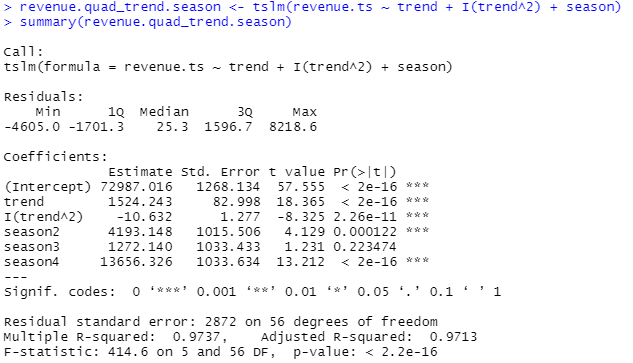


From the above accuracy figures, it shows that the MAPE and RMSE accuracy measures for both the validation period (as well as for the training period), out of the three considered forecasts as significant for further considerations from 2b, the *regression model with quadratic trend and seasonality* provides the best (lowest) accuracy measures.

1. **Employ the entire data set to make time series forecast**

**3a. Apply the two most accurate regression models identified in question to make the forecast for the last two quarters of 2020 and first two quarters of 2021. For that, use the entire data set to develop the regression model using the tslm() function. Apply the summary() function to identify the model structure and parameters, show them in your report, and also present the respective model equation. Use each model to forecast Walmart’s revenue in the 4 quarters of 2020 and 2021 using the forecast() function, and present this forecast in your report.**

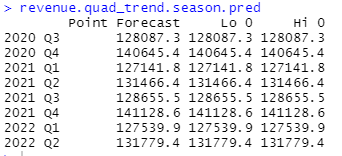
**First One:**



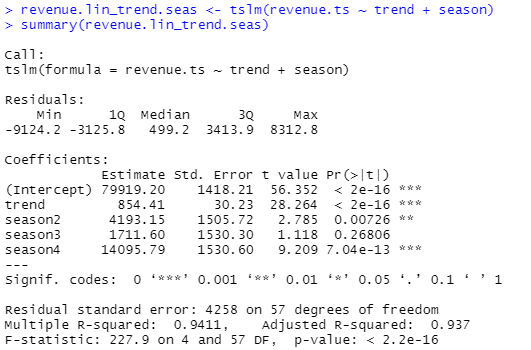
This *regression model with quadratic trend and seasonality* for the whole data set shows that it contains 5 independent variables: trend index (t), squared trend index (t2), and 3 seasonal dummy variables for Q2 (season2 – D2), Q3 (season3 – D3) and Q4 (season4 – D4). The regression equation is:

*yt = 72987.01 + 1524.243 t – 10.63 + 4193.14 D2 + 1272.14 D3 + 13656.32 D4*

Based on the model’s summary, we can observe that it shows a high R-squared of 0.9737, statistically significant F-statistic (p-value is substantially lower than 0.01), and all regression coefficients including intercept being statistically significant (p-value <0.01) except for D3. Hence, we can conclude that overall, this regression model is statistically significant and can be used for forecasting revenue for next years. Following is the forecast made based on the model of quadratic trend and seasonality:



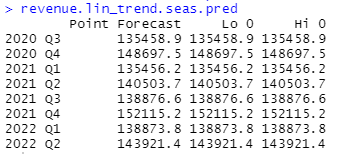
**The second one:**



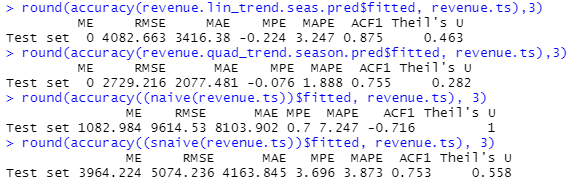
This regression model with linear trend and seasonality for the whole data set shows that it contains 4 independent variables: trend index (t), and 3 seasonal dummy variables for Q2 (season2 – D2), Q3 (season3 – D3) and Q4 (season4 – D4). The regression equation is:

*yt = 79919.2 + 854.41 t + 4193.15 D2 + 1711.6 D3 + 14095.7 D4*

Based on the model’s summary, we can observe that it shows a high R-squared of 0.9411, statistically significant F-statistic (p-value is substantially lower than 0.01), and all regression coefficients including intercept being statistically significant (p-value <0.01) except for D3. Hence, we can conclude that overall, this regression model is statistically significant and can be used for forecasting revenue for next years. Following is the forecast made based on the model linear trend and seasonality:



**3b. Apply the accuracy() function to compare the performance measures of the regression models developed in 3a with those for naïve and seasonal naïve forecasts. Present the accuracy measures in your report, compare them, and identify, using MAPE and RMSE, which forecast is most accurate to forecast Walmart’s quarterly revenue in 2020 and 2021.**



From the above accuracy figures for the models forecasted in 3a, it shows that the MAPE and RMSE measure, the regression model with quadratic trend and seasonality is evidentially more accurate than both the seasonal naïve forecast and more accurate than the naïve.

It can also be learned from the above figures the linear trend and seasonality is evidentially more accurate than both the seasonal naïve forecast and more accurate than the naïve.

But when both compared. The regression model with quadratic trend and seasonality is more accurate than the linear trend and seasonality model. Therefore, the regression model with quadratic trend and seasonality is the more accurate model to use in forecasting.