Xueqing Wu

E-Commerce Retail Sales
Time Series Analysis and Forecast

Abstract:
The goal of this project is to fit a model for the E-commerce retail sales to predict the future sales. The data is a quarterly data from Quarter 4 of 1999 to Quarter 3 of 2019. The dataset is divided into two parts, the training dataset with 17 seasons and the validation dataset with 3 seasons. For the ARIMA forecasting, I tried Seasonal ARIMA, subset ARMA model, and auto ARIMA approach, and the model diagnostic shows that all of the models are adequate. Besides, I applied a non-ARIMA forecasting, Holt-Winters forecast. By comparing the RMSE, MAE, and MAPE, the auto ARIMA approach fits the best model with (2, 1, 0)*(0, 1, 1)[4]. The model fitted by SARIMA and subset ARMA model approach may be overfitting due to the excessive differencing.

Introduction:

With the development of Internet, more and more people shop online instead of in physical stores. The goal of this project is to fit a model for the E-commerce retail sales to predict the future sales. The dataset is found on Kaggle (https://www.kaggle.com/census/e-commerce-retail-sales-series-data-collection#ECOMSA.csv), and you can find more about the dataset here (https://fred.stlouisfed.org/series/ECOMNSA).

The data records the quarterly total E-Commerce retail sales in the United States from Quarter 4 of 1999 to Quarter 3 of 2019. The dataset contains 80 observations, and there is no missing value. There are only two variables in the dataset: time (quarter) and the value of total sales. The unit of total sales is transformed from millions of dollars to billions of dollars by dividing 1,000 to the value variable.

Data Description and Transformation:

The dataset is divided into parts: training dataset, with 17 seasons from 1999 Quarter 4 to 2017 Quarter 3, and validation dataset, with 3 seasons from 2017 Quarter 4 to 2019 Quarter 3. The plot of training dataset (figure 1) and decomposition of the data (figure 2) are shown below.

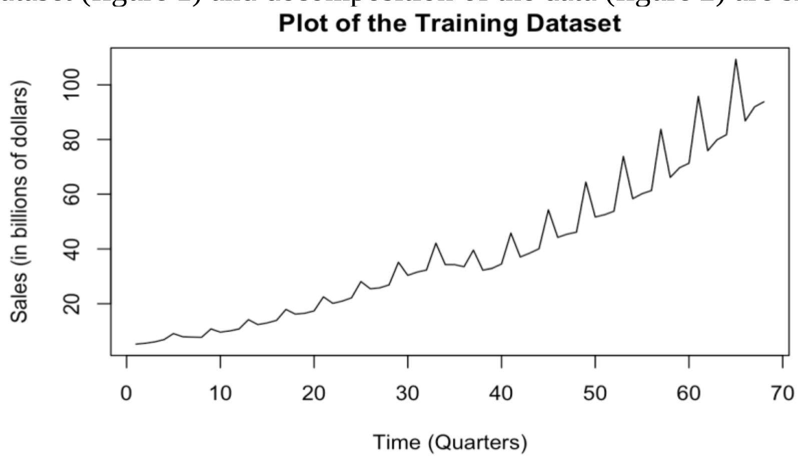


Figure 1

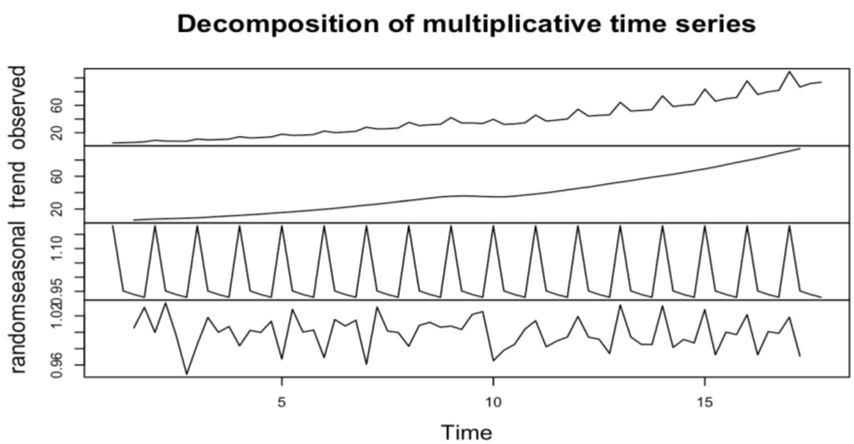


Figure 2

According to figure 1, there are no outliers in the dataset. Both figures show that the process has an evident increasing trend and seasonal pattern. The increasing rate is slow in the middle due to the 2008 financial crisis. As figure 1 manifests, the variance is not constant. Thus, I applied a log transformation to stabilize the variance.

<u>Model Identification and Parameter Estimation</u>

<u>SARIMA Model</u>

To test if the process has a unit root, I conducted an augmented Dickey-Fuller test. The p-value of the test is 0.99, so we fail to reject the null hypothesis that the process is not stationary. Therefore, the time series have a unit root, and it is not stationary. Also, figure 1 and figure 2 manifests an increasing trend in the process, so I conducted a non-seasonal differencing to the process.

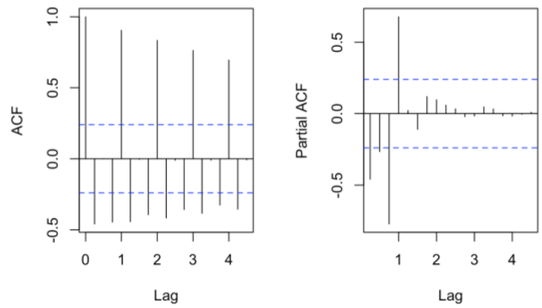
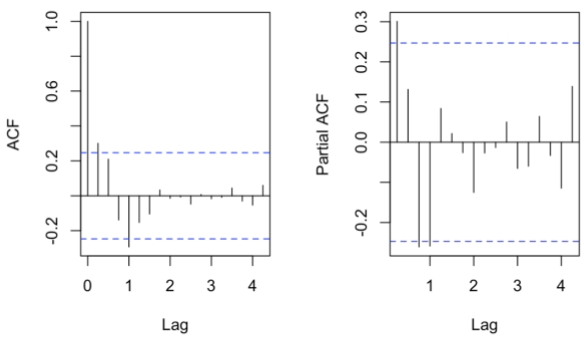


Figure 3

The ACF and PACF after the non-seasonal differencing is shown in the figure 3. Obviously, the seasonal lags are more significant than other lags. Also, the decomposition of the time series manifests a clear seasonal pattern. Therefore, I conducted a seasonal differencing.



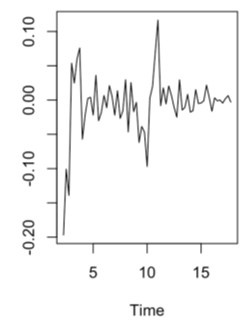
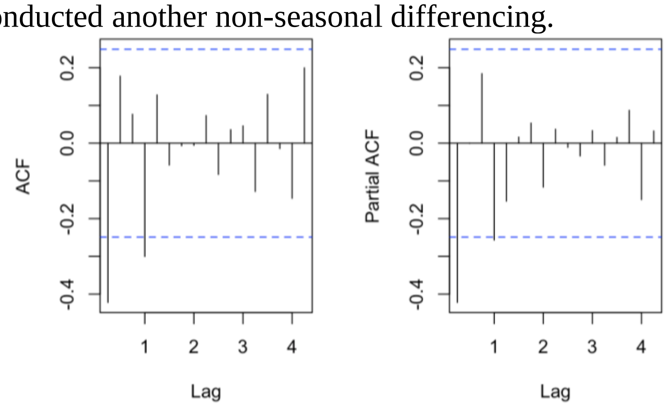


Figure 4 Figure 5

After a non-differencing and a seasonal differencing, the ACF and PACF is shown in figure 4, and the time series after differencing is shown in figure 5. Although the ACF and PACF seems stationary, the time series process shows the process is not stationary as the mean is not constant. Thus, I conducted another non-seasonal differencing.



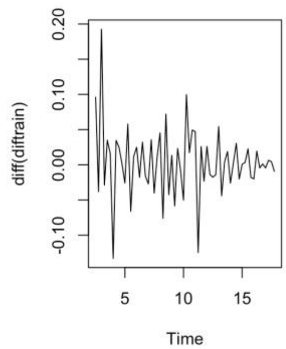


Figure 6 Figure 7

After 2 non-seasonal differencing and 1 seasonal differencing, ACF and PACF in figure 6 and the graph of time series process in figure 7 all indicate that the process is stationary. Therefore, the final differencing for this time series process is d=2 and D=1.

Nevertheless, the seasonal lags are relatively more significant than other lags in figure 7. Hence, I decide to fit SARIMA models instead of ARIMA models. According to figure 7, both ACF and PACF cuts off at lag 1, so p, q=0 or 1. By observing the seasonal lags, I found that both ACF and PACF are significant at 1S, so I decide to try P, Q=0 or 1. Therefore, I have 16 candidate models in total. By excluding the models that either p, q and P, Q are both 0, there are 9 models in total, and I order them by AICc from the lowest to the highest.

ARMA(p, q)	ARMA(P, Q)	AICc
(1, 2, 0)	(0, 1, 1)	-207.83
(1, 2, 0)	(1, 1, 0)	-206.9
(1, 2, 0)	(1, 1, 1)	-205.63
(1, 2, 1)	(0, 1, 1)	-205.58
(0, 2, 1)	(0, 1, 1)	-205.51
(0, 2, 1)	(1, 1, 0)	-204.77
(1, 2, 1)	(1, 1, 0)	-204.62
(0, 2, 1)	(1, 1, 1)	-203.51
(1, 2, 1)	(1, 1, 1)	-201.02

According to AICc, the model (1, 2, 0)*(0, 1, 1)[4] has the lowest AICc, and it is the best fitted model in the SARIMA model selection approach.

<u>Subset ARMA Model</u>
In addition to the SARIMA model, I also tried subset ARMA model. According to the previous differencing, I decide that d=2, and D=1, same as the previous. I set the maximum lag to be 15 for both AR process and MA process. The result of fitting a subset ARMA model is shown below in figure 8.

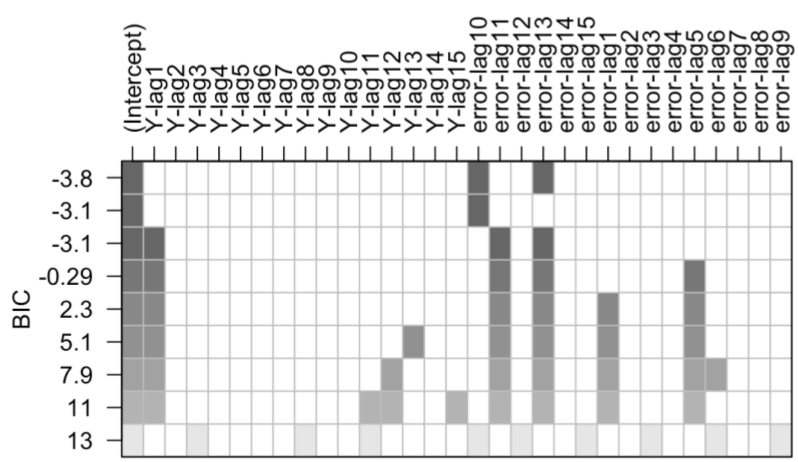


Figure 8

According to figure 8, the top row has the lowest BIC. However, the second row has similar BIC but with a cleaner model. Thus, I chose the MA(10) model on the second row. By fitting the subset of MA(10) model, the coefficient of lag 10 is -0.0907, and the model can be written as *Xt=et-0.097e(t-10)*.

Auto ARIMA
Besides the SARIMA model selection and the subset of ARMA model selection, I also applied auto ARIMA approach. I used the auto.arima function in r to fit the original training dataset without any differencing or log transformation. The result of this model selection method is shown in figure 11.

```
Series: train_s
ARIMA(2,1,0)(0,1,1)[4]
Box Cox transformation: lambda= 0

Coefficients:
         ar1     ar2     sma1
      0.3129  0.3100  -0.5005
s.e.  0.1444  0.1496   0.1223

sigma^2 estimated as 0.001548:  log likelihood=115.36
AIC=-222.72   AICc=-222.03   BIC=-214.15
```

Figure 11

The auto ARIMA selection model suggests that the best model for the training dataset is SARIMA(2, 1, 0)*(0, 1, 1)[4] with a log transformation. This result is different from the model I selected using the SARIMA selection, especially the differencing part. The auto ARIMA only apply 1 non-seasonal differencing while I apply twice.

Model Diagnostic
To ensure each model is valid, I conducted model diagnostic for each model.

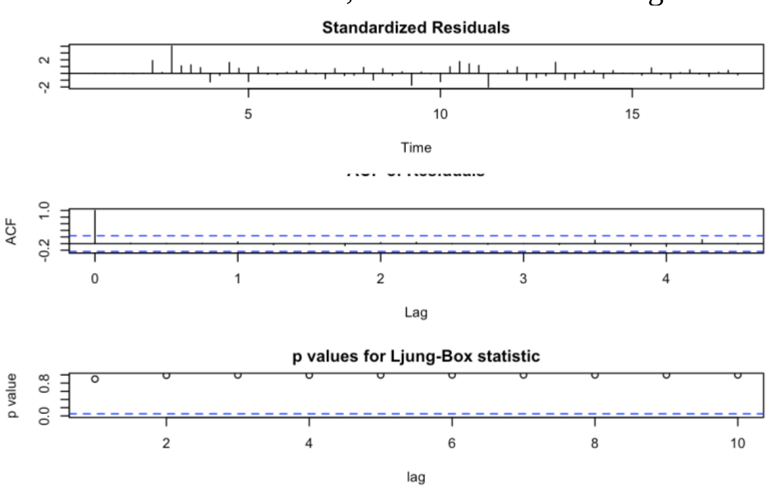


Figure 12

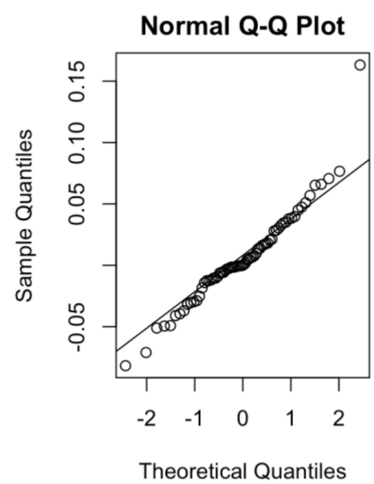


Figure 13

Figure 12 and figure 13 are the model diagnostic for SARIMA model. According to figure 12, the residuals are randomly distributed without any pattern. The ACF is not significant after lag 1, so it implies ACF=0. The p values for Ljung-Box statistics are all greater than 0.05, so we fail to reject the hypothesis that all ACFs equal to 0. Hence, I can conclude that the residuals are independent. According to the Normal QQ plot by figure 13, although the beginning and the end deviates from the benchmark and there is an outlier at the right top corner, the majority of the dots align with the benchmark. Hence, the model satisfies the normality assumption.

I repeated the same process for the model fitted with subset ARIMA and auto ARIMA. Both models have random and independent residuals, and they all satisfy the normality assumption. Thus, they are all adequate models.

Forecasting

Besides the ARIMA forecasting, I also conducted multiplicative Holt-Winters forecast. To select the best fit model, I implement out-sample forecast by comparing the forecast value with the test value and compared the errors for each model. I used MAE, RMSE, and MAPE as standard to compare, and the result is shown below.

Approach	Model	RMSE	MAE	MAPE
SARIMA	(1, 2, 0)*(0, 1, 1)[4]	4.532487	3.240306	2.360991
Auto ARIMA	(2, 1, 0)*(0, 1, 1)[4]	2.699877	1.820499	1.838883
Subset ARIMA	(0, 2, 10)*(0,1,0)[4]	3.482916	2.790716	2.101486
Holt-Winters Forecast	NA	6.845874	5.485342	4.138144

By comparing all four models, the auto ARIMA approach has the lowest MSE, MAE, and MAPE which gives the best fitted model. Figure 14 shows the forecast of the (2, 1, 0)*(0, 1, 1)[4] with the auto ARIMA method.

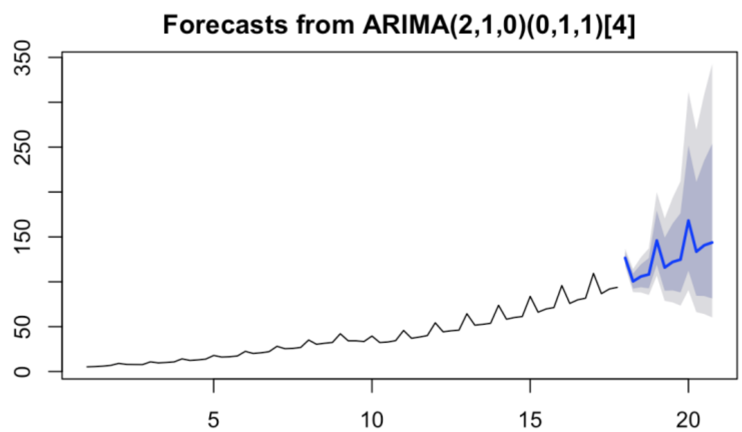


Figure 14

Conclusion:

The best model is ARIMA (2, 1, 0)*(0, 1, 1)[4] fitted by the auto ARIMA method. The reason that other two ARIMA (SARIMA and Subset ARIMA) models do not do well is because I applied 2 non-seasonal differencing which may not be necessary. I misinterpreted the graph of time series after d=1 and D=1 (figure 5) as nonstationary and conducted another non-seasonal differencing whereas the non-constant mean should be solved by the intervention analysis. The excessive differencing leads to a higher likelihood of overfitting. Therefore, I should try SARIMA model with only d=1 and D=1 next time.

This model can help predict the total retail sales of E-commerce and even predict the econmonic trend of the United States.