U.S. Clothing Retail Sales Forecast with SARIMA Model

Ziping Wei

Introduction

For project2, I continued to analysis the U.S. Clothing Retail Sales data, which was used for project 1. Unlike in project 1, where I applied the AR model to forecast sales, in this project I applied the SARIMA to forecast sales. The data used in project 1 is sales from January 1992 to August 2019 and the data shows an obvious decline in 2008 which may because of recession. In this project, I only choose the 10-year sales data from 2009 to October 2019 with 130 observations[1].

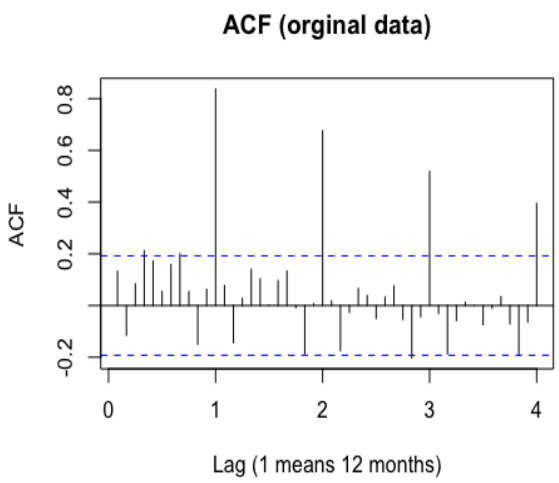
Explanatory Data Analysis

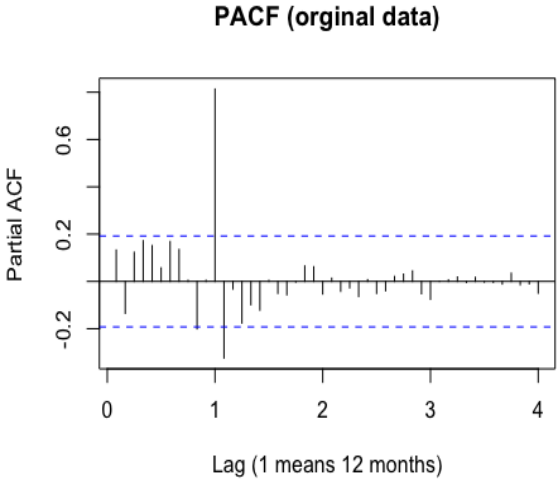
The plot below shows the monthly retail sales of clothing stores in U.S. from Jan.2009 to Oct. 2019. In general, clothing sales shows an increasing trend from 2009 to 2019. We could see that the data is not stationary with an obvious positive linear trend.

Fig.2.1 *U.S. 2009.01-2019.10 Clothing Retail Sales*

To further test my thoughts, I checked the ACF and PACF of the original data.

[1] U.S. Census Bureau, Retail Sales: Clothing Stores [MRTSSM4481USN], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/MRTSSM4481USN, December 20, 2019.

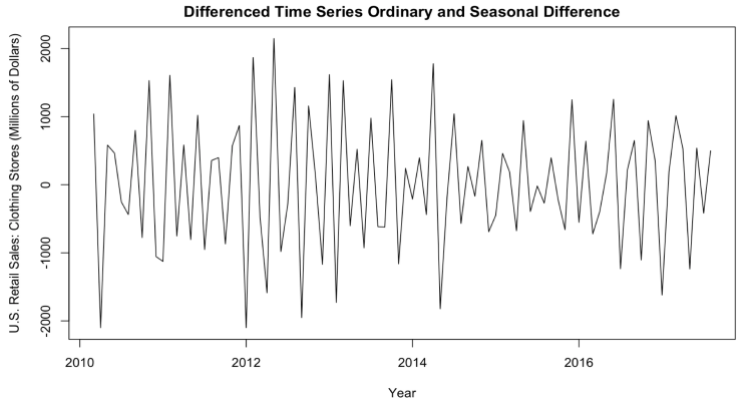


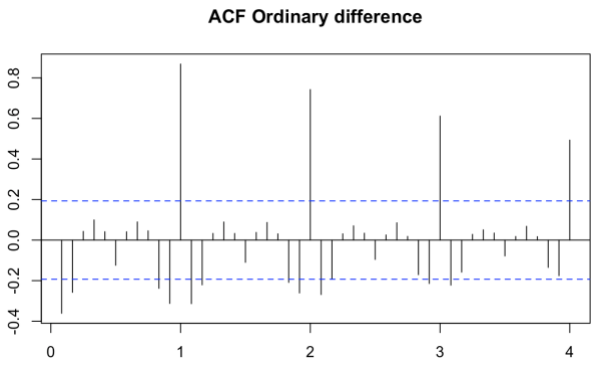


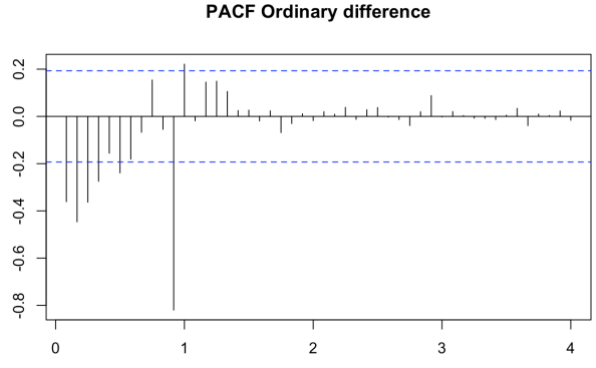
The ACF and PACF plot further prove the nonstationarity of the data, with ACF decays very slowly.

Model Selection

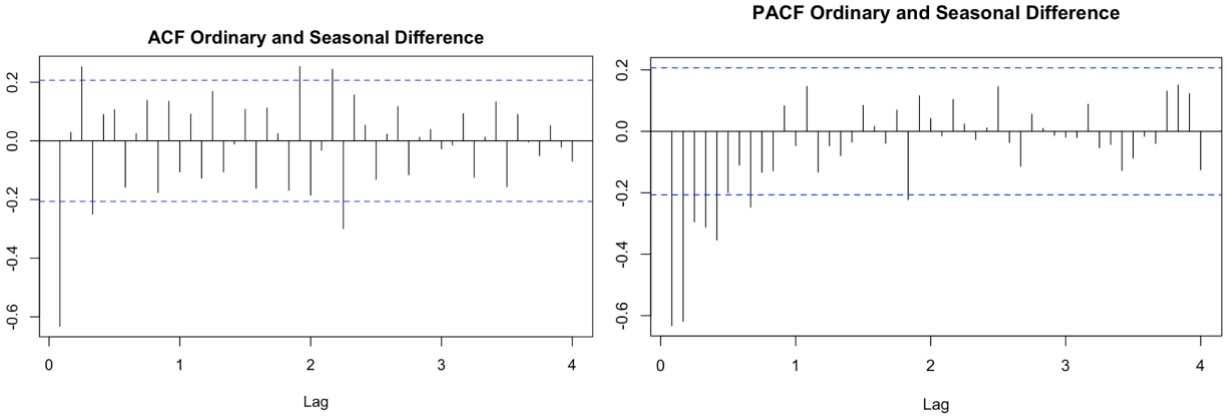
To test the accuracy of our models later, I split the data into Train and Test datasets with splitting ratio 0.8. After splitting, the train dataset has 104 observations and the test dataset has 26 observations. To remove nonstationarity of the train data, I first take one lag differencing.



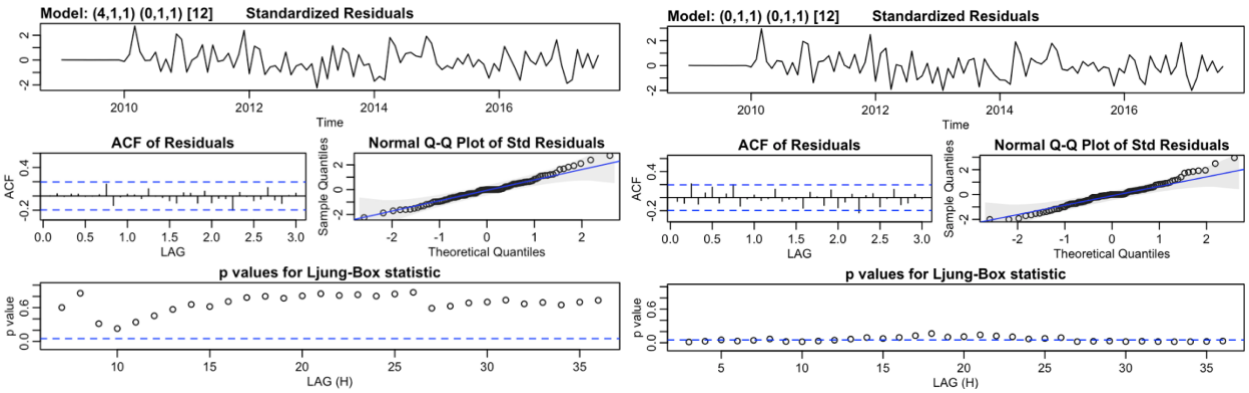




After taking one lag difference, the time series plot shows a constant mean around 0 and a stable variance, compared to the original one. Noticed that on x-axis, 1 means 12 months. Therefore, if we consider ACF cuts off at lag 2 and PACF tails off slowly, MA(2) could be a candidate model for the ordinary difference data. The first two peak of ACF are significant, which may because of the influence of the seasonal pattern. Therefore, MA(1) could be a candidate model as well. If we consider PACF cuts off at lag 4 and 6, the ARIMA models could be possible. In addition, the ACF indicates a seasonal pattern, as the ACF spikes at lag 12, 24, 36. In order to achieve stationarity, I further take a seasonal difference with period 12.

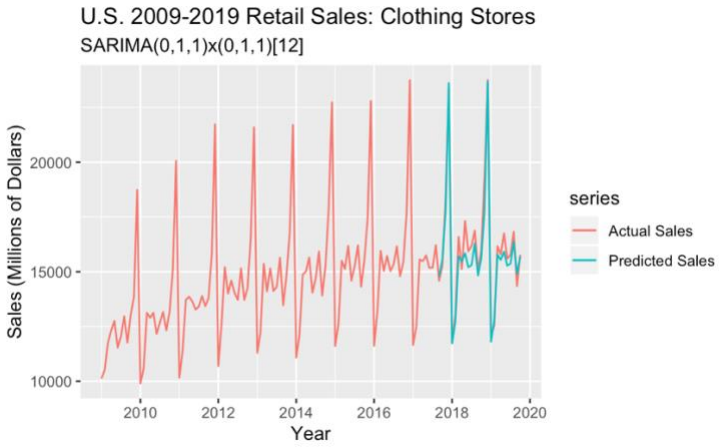


The ACF cuts off at lag 1 and 4 and PACF decays slowly. MA(1) and MA(4) could be possible models for the seasonal difference data. Combine the possible models for the ordinary difference data and the seasonal difference data, below are possible models: SARIMA(p,1,d) x (0,1,D)[12] with p could be 0, 4, 6, d could be 1, 2 and D could be 1, 4. After fitting all SARIMA models with all combinations of possible p, d and D values. SARIMA(4,1,1) x (0,1,1)[12] has the lowest AIC score and the SARIMA(0,1,1) x (0,1,1)[12] has the lowest BIC score. To compare these two models, I then checked the diagnostic plots.



The residuals of both models are fairly normal distributed. The p values of SARIMA(4,1,1) x (0,1,1) are significant in general compared to p values of SARIMA(0,1,1) x (0,1,1)[12]. However, for SARIMA(4,1,1) x (0,1,1), only ar1, ma1 and sma1 have significant coefficient. In

addition, BIC has more penalty on model complexity. Taking two reasons above into consideration, I decided to choose SARIMA(0,1,1) x (0,1,1)[12] as the final model which has the 2nd lowest AIC score and the lowest BIC score. (The scores of all possible models could be seen in the R codes). Fitting the SARIMA(0,1,1) x (0,1,1)[12] on train data, the predicted value vs. actual value of test data are shown below.



Fitting SARIMA(0,1,1) x (0,1,1)[12] Model and Forecast

Fitting the whole data set to SARIMA(0,1,1) x (0,1,1)[12] and the Final model is

$$(1-B)(1-B^{12})x_t = (1-0.8153B)(1-0.2187B^{12})w_t$$

```
Call:
arima(x = train, order = c(0, 1, 1), seasonal = list(order = c(0, 1, 1), period =
12))
Coefficients:
        ma1     sma1
      -0.8153  -0.2187
s.e.   0.0591   0.1550
sigma^2 estimated as 180646:  log likelihood = -680.73,  aic = 1365.46
```

Using the SARIMA(0,1,1) x (0,1,1)[12] model, the next 12 months values are forecasted.