Time Series Final Project- Prediction of Future Sales Data

Chuching Ho

- Abstract

In this project, I try to use three different time series data: sales of Retail Trade and Food Services, sales of Manufacture's shipment, inventory and orders, and sales of Motor Vehicle or Auto parts to predict the future trends of these three series. I came up with an ARIMA model for each series, and then I combined the three series with a multivariate autoregressive model. The results showed that those three series can significantly predict each other.

- Introduction

Gross Domestic Product(GDP) is the total value of final goods and services produced within a country over a period of time. The manufacture data can be a valuable indicator for the macroeconomic market. The sales of retail trade and food services can represent the personal consumption which plays a significant role in the health of economy. For instance, if the retail and food services growth is slowing or stalled, this indicates that consumers are not spending at previous levels. In addition, the sales of motor vehicle or auto parts can also be representative of the personal consumption. Thus, I choose to analyze these series from 2000 to 2018. One is sales of Retail Trade and Food Services, another is monthly sales of Manufacturer's Shipment, Inventory and Orders, the other is sales of Motor Vehicle or Auto parts. The monetary unit of the three series are in millions of US dollars. The three time series data are from the United States Census Bureau with each month a data point from 2000 to 2018.

- Statistical Methods

Before fitting the model, exploratory data analyses (EDA) were used to check the time series if it is more likely to be ARMA. The time series plots were plots for every series respectively to see the general trend and the stationarity. To see if a time series is stationary, the mean function needs to be constant and independent of time t and the autocorrelation function depends only through the time difference of two time points, that is the absolute value of s-t.

The autocorrelation function (ACF) is defined as $$\rho(s,t) = \frac{\gamma(s,t)}{\sqrt{\gamma(s,s)\gamma(t,t)}}.$$

The Autocorrelation Function measures the linear predictiability of the series at time t using the value at time s. Partial Autocorrelation Function measures the correlation between X and Y with the linear effect of Z removed. I then plot the Autocorrleation Function and Partial Autocorrelation Function to check the time series. For the MA(q), the ACF will be zero for lags greater than q. The ACF will not be zero before and at lag q. For AR(p), the PACF will be zero for lags greater than p. The PACF will not be zero before and at lag q.

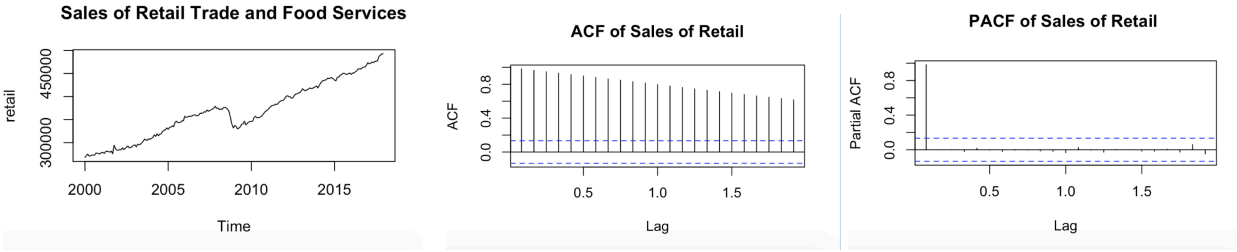
I choose to use ARIMA model to fit the data respectively. In addition, split the data into two parts, one is from 2000 to 2017 as the analysis part of the time series. The other is from 2018. I use the data from 2000 to 2017 to predict the data for the first ten months of 2018. I compare the predicted value with the true value.

Based on the series I chose, there might exist some correlation between the series. Thus, I use the Vector Autoregressive Model (VAR) to analyze three-dimensional series together. Vector Autoregressive Model (VAR) can capture the linear interdependencies among multiple time series.

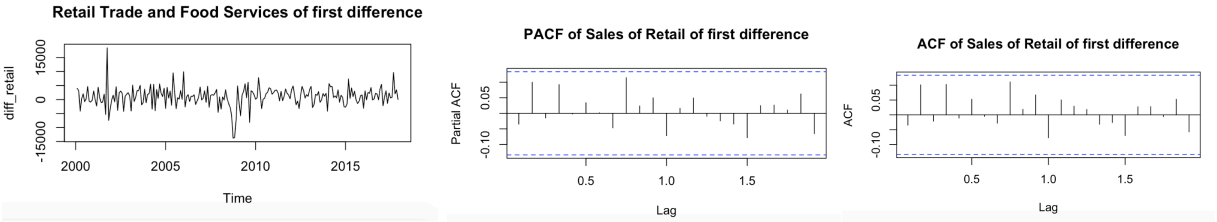
- Results

1. Sales of Retail Trade

   The following are the time plot (left), ACF (middle), PACF (right) of the Sales of Retail Trade and Food Services series from 2000 to 2017.



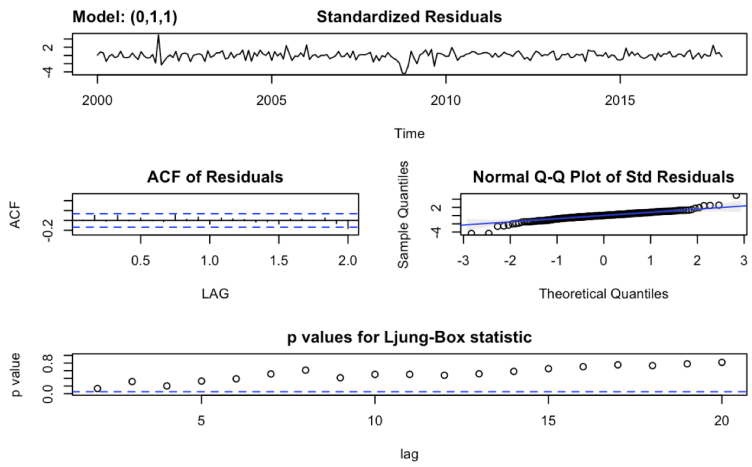
   From the ACF plot, there exists autocorrelation. The Sales of Retail Trade and Food Services is not stationary. Thus, I take first difference of the series.



   By plotting the ACF of the differenced time series, it appears that the differenced process shows minimal autocorrelation. I then plotted the EACF to identify the potential models.

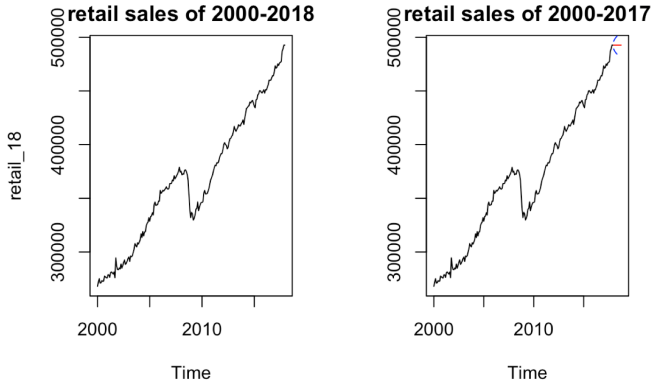
```
AR/MA
   0 1 2 3 4 5 6 7 8 9 10 11 12 13
0  o o o o o o o o o o o  o  o  o
1  x o o o o o o o o o o  o  o  o
2  x x o o o o o o o o o  o  o  o
3  x x o o o o o o o o o  o  o  o
4  o x o x o o o o o o o  o  o  o
5  o x o o o o o o o o o  o  o  o
6  o x o o x o o o o o o  o  o  o
7  o x x o x x x o o o o  o  o  o
```

   I fit the model using ARIMA(1,1,2). The following is the residual diagnostic of a ARIMA(0,1,1)



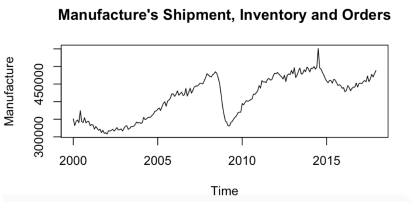
From the multivariate Ljung-box test, the test statistic is 4.6198, p-value= 0.464. The Box-Ljung test failed to reject the null hypothesis of white noise. We conclude that the model is a good fit. From the Normal QQ plot, we can see that there are more extreme values than expected. Other than the extreme values, the sample quantiles are aligned with the theoretical quantiles.

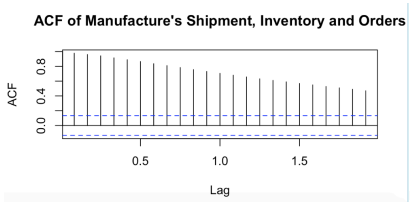
The time plot are true retail sales from 2000 to 2018(left), true retail sales from 2000-2017 and the predicted values of sales for 2018(right).

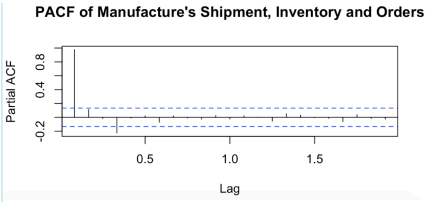


2. Manufacture's shipment, inventory and orders

The following are the time plot, ACF and PACF of the sales of the manufacture's shipment, inventory and orders from 2000 to 2017.





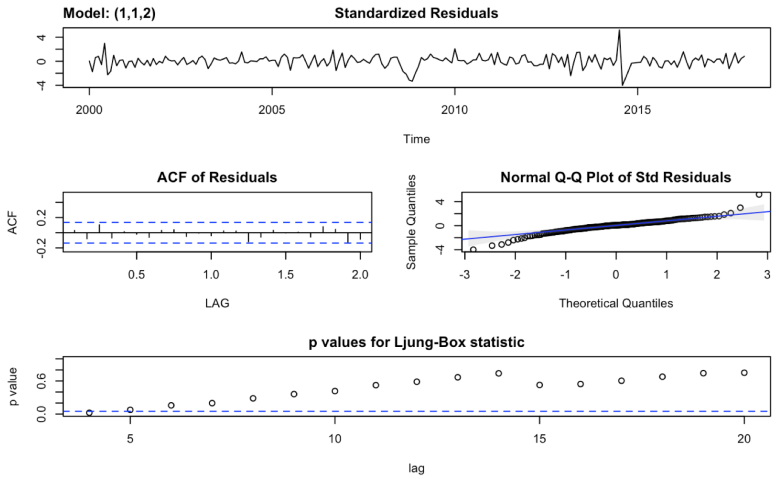


From the ACF plot, there exists autocorrelation. The sales of Manufacture's Shipment, Inventory and Orders is not stationary. Thus, I take first difference of the series. After differencing, the time plot looks more like a stationary series. By plotting the ACF of the differenced time series, it appears that the differenced process shows minimal autocorrelation.

```
> eacf(diff_Manufacture)
AR/MA
  0 1 2 3 4 5 6 7 8 9 10 11 12 13
0 x o x o o o o o o o o  o  o  o
1 x o x o o o o o o o o  o  o  o
2 x x x o x o o o o o o  o  o  o
3 x o x x x o o o o o o  o  o  o
4 x o o x x o o o o o o  o  o  o
5 x x o o x o o o o o o  o  o  o
6 x x o o o o o o o o o  o  o  o
7 x x o o o x x x o o o  o  o  o
```

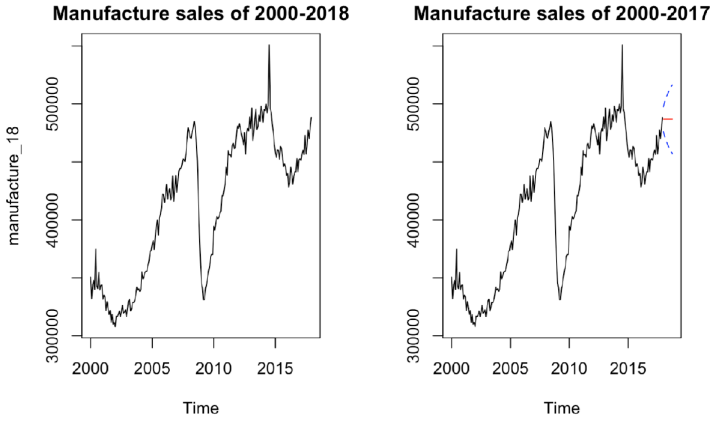
I first chose to use ARIMA(0,1,1) and ARIMA(1,1,1) to fit the model; however, the Box-Ljung test rejected the null hypothesis of white noise. I conclude that these models are not a good fit.

I fit the model using ARIMA(1,1,2). The following is the residual diagnostic of a ARIMA(1,1,2).



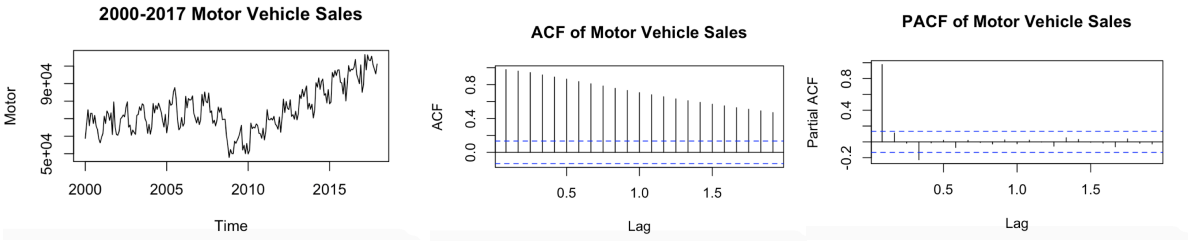
From the multivariate Ljung-box test, the test statistic is 5.1264, p-value= 0.4006. The Box-Ljung test failed to reject the null hypothesis of white noise. We conclude that the model is a good fit. From the Normal QQ plot, we can see that there are more extreme values than expected. Other than the extreme values, the sample quantiles are aligned with the theoretical quantiles.

I forecast the Sales of Manufacture data for 2018.



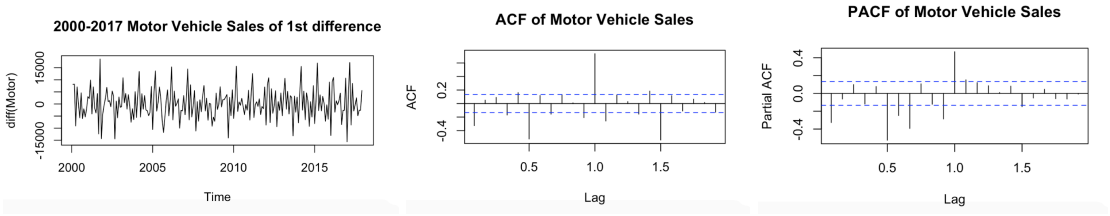
3. Motor Vehicle Sales

The following are the time plot(left), ACF(middle) and PACF(right) of the sales of the manufacture's shipment, inventory and orders from 2000 to 2017.

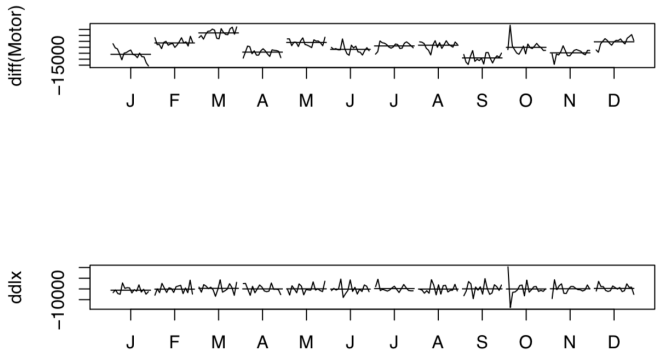


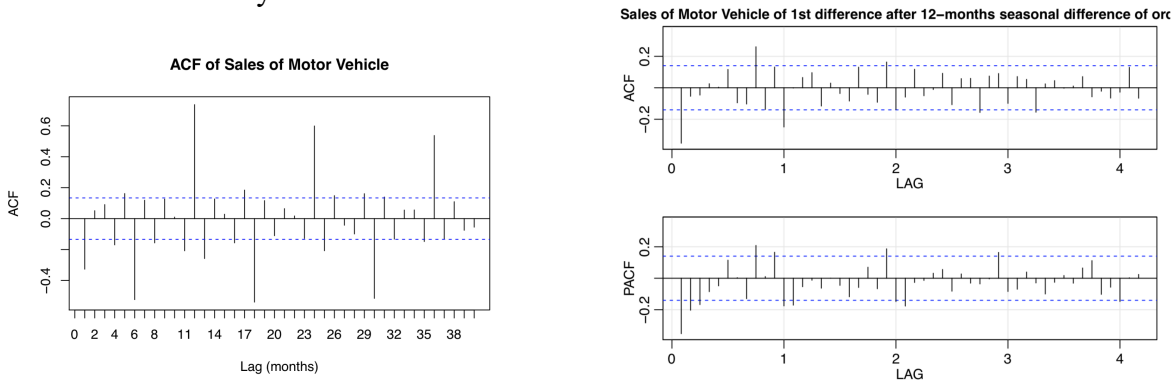
From the ACF plot, there exists autocorrelation. The series of sales of Motor Vehicle is not stationary. Thus, I take first difference of the series. After differencing, the time plot looks

more like a stationary series. By plotting the ACF of the differenced time series, it appears that the differenced process shows minimal autocorrelation.



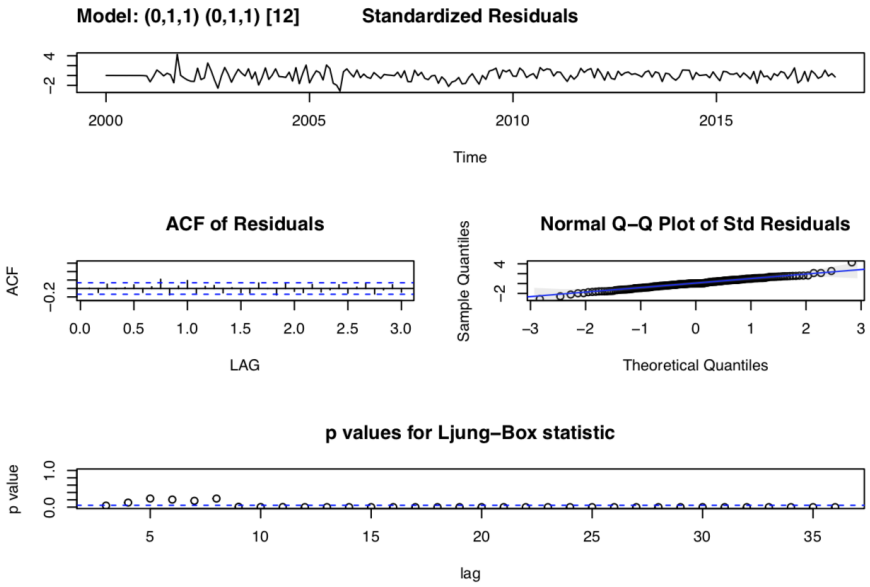
To inspect seasonal nonstationarity, I inspected the two plots (right). The month plot(top) indicates seasonal nonstationarity. After considering a 12-months seasonal difference of order D=1, the month plot(bottom) indicate that the nonstationarity is removed.



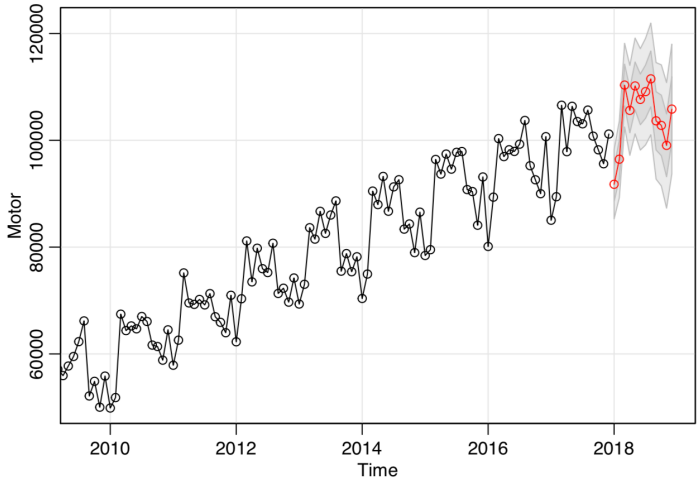


We can see that every six months there is a spike in the ACF. The spikes are slowly decreasing every 12 months. That behavior points at a seasonal nonstationarity. Hence, we can try to remove this seasonal nonstationarity by considering a 12-months seasonal difference of order D=1

Consider the ACF and PACF of Sales of Motor Vehicle series after 12-months seasonal difference of order D=1

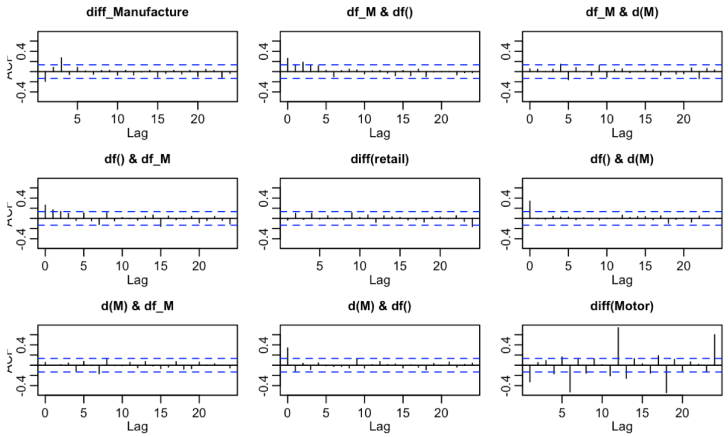


I forecast the Sales of Motor Vehicle data for 2018.

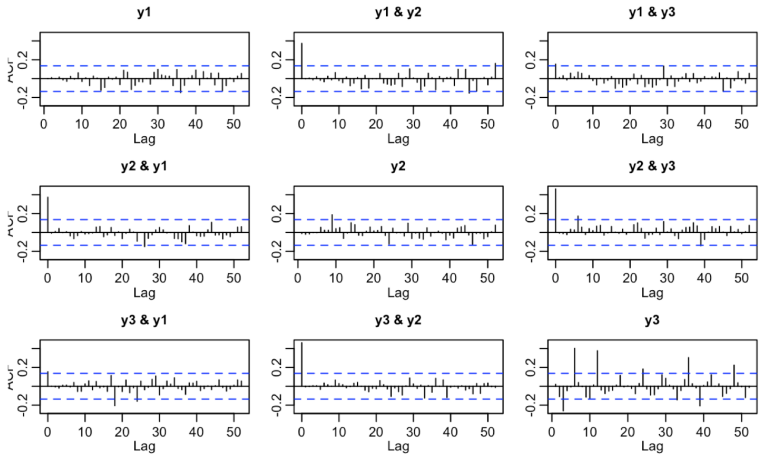


4. Multivariate Autoregressive model (VAR)

VAR: I use multivariate Autoregressive model to estimate the three-dimensional time series. I plot the ACF(right) to check the autocorrelations of the three-dimensional series.



The plot below is the residual ACF after fitting the autoregressive model.



The cross-correlations of the residuals shows the ACFs of the individual residual series along the diagonal. The off diagonals display the CCFs between pairs of residual series. We notice that most of the correlations in the residual series are negligible, however, the zero-order correlation of retail data with manufacture residuals is about 0.37.

This means that the AR model is not capturing the concurrent effect of retail(y2) on manufacture(y1). The zero-order correlation of manufacture data with retail residuals is about 0.37. The zero-order correlation of manufacture data with motor residuals is about 0.43. This means that the AR model is not capturing the concurrent effect of motor(y3) and

manufacture(y1) on retail(y2). The zero-order correlation of retail data with motor residuals is about 0.43. This means that the AR model is not capturing the concurrent effect of retail(y2) on motor(y3).

Note that from the VARselect function, BIC picks the order p = 1 model while AIC, Hannan-Quinn and FPE pick an order p = 8 model.

The following regressions are the ones with significant variable.

y1 = -.280*y1.lag1 + .679*y2.lag1 +.820*y2.lag2 -.530 y2.lag6

y2 = .122*y1.lag1 -.218* y2.lag1 +.114*y1.lag2 +.068*y1.lag3 -.206*y2.lag3 -.060*y1.Iag7+1230*constant

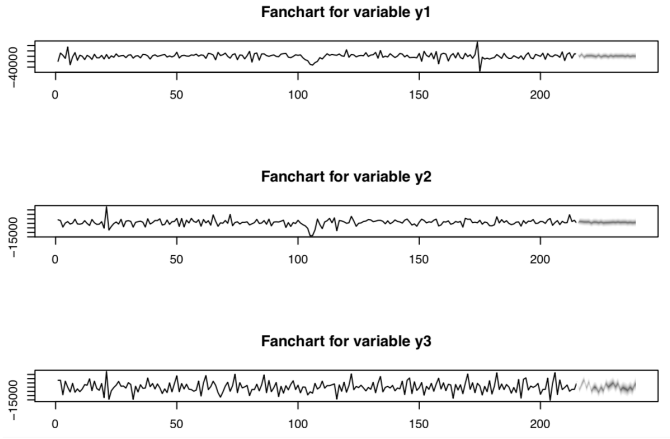
y3= .617*y3.lag1 -.507*y3.lag2 -.129*y3.lag4 -.114* y3.lag5 -.089*y1.lag6 +.620*y2.lag6 - -.810*y3.lag6 -.148*y1.lag7 -.316*y2.lag7 -.561*y3.lag7 +.240 y2.lag8 -.527y3.lag8

To examine the residuals, I plot the ACF of the residuals and examine the multivariate Ljung Box plot. From the multivariate Ljung-box test, the null hypothesis of white noise is rejected. We conclude that the model is not a good fit.

```
mLjungBox(xdata, lag = 5)
```

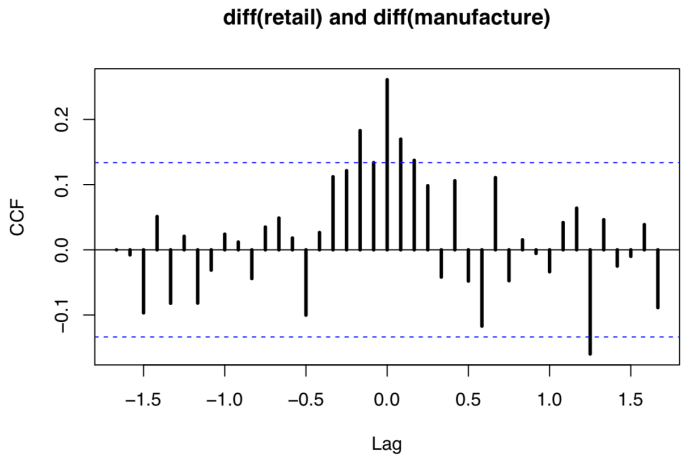
```
##   K   Q(K) d.f. p-value
## 1 5 134.71   45       0
```

The Fanchart for variable y1 represents predictions and errors of the sales Manufacture's Shipment, Inventory and Orders. The Fanchart for variable y2 represents the predictions and errosr of sales of Retail Trade and Food Services. The Fanchart for variable y3 represents the predictions and errors sales of Motor Vehicle.



The cross-correlation function between changes in the of retail and changes in the manufacture is shown below. The largest absolute cross-correlations are around zero lags and these correlations are positive.

This means that an above-average change in sale of retail predicts a future change in manufacture that is above average.



- Discussion

For the sales of Retail and Food Services series and the Manufacture series, I predicted the sales for 2018 and compared them with the true values. I found that the predicted values are quite different from the true values. Although the models are good fits, it is still difficult to accurately predict the future trend given the historical data. Future studies are required to identify new

methods that can more accurately predict the future economic trend.

According to the analysis of the sales of motor vehicle, we predict the sales of the motor vehicle for 2018. We captured the macroeconomic trend and the seasonality of the sales of motor vehicle. Since sales of the motor vehicle are one part of the consumption, we therefore can know that one part of the GDP trend and seasonality.

For the multivariate AR model, it can account for the correlation of these series that may be left behind when doing the ARIMA modeling respectively. The residual diagnostics of the fitted model shows that there exists some concurrent effect that still did not explained by the model.

- Reference

R.Shumway, Time Series Analysis and its Applications with R examples