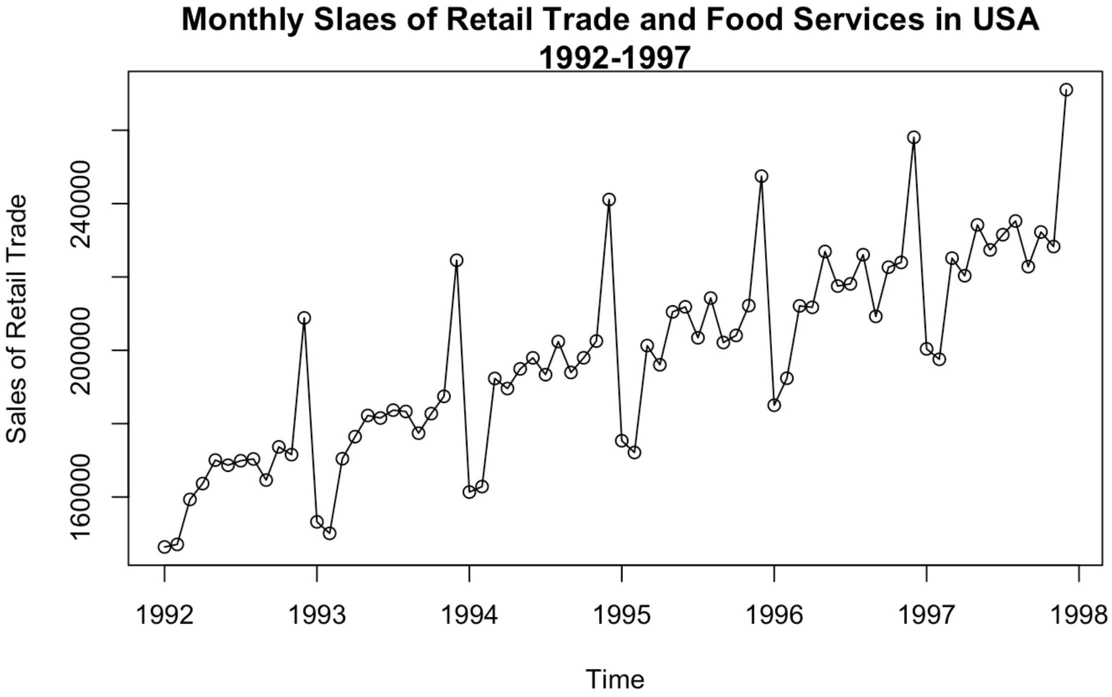
I. **Introduction:**

The selected dataset shows the monthly sales of retail trade and food services in USA from January 1992 to January 2020. All numbers are shown in millions of dollar. Since the overall trend and seasonal pattern are similar throughout the data, I decide to only include the time frame from January 1992 to December 1997.



According to the monthly graph above, we can see that overall trend of the monthly sales of foods were increasing steadily with seasonality within each year. For every single year in the dataset, the amount of sales increased from January and had an intensively surge in December. After that, sales drop back to the lowest point of a year and repeat the pattern again. To be noticed that the fluctuation of the sales from February to November became more and more strong for some unknown reason.

In this case, I decide to use the seasonal means model to remove the seasonality and the smoothing method to estimate the overall trend. After removing those non-stationary component, I think the overall trend will be smooth and flat. The fluctuation between month and month will also be flatten.

Source: https://www.kaggle.com/datasets/landlord/usa-monthly-retail-trade.

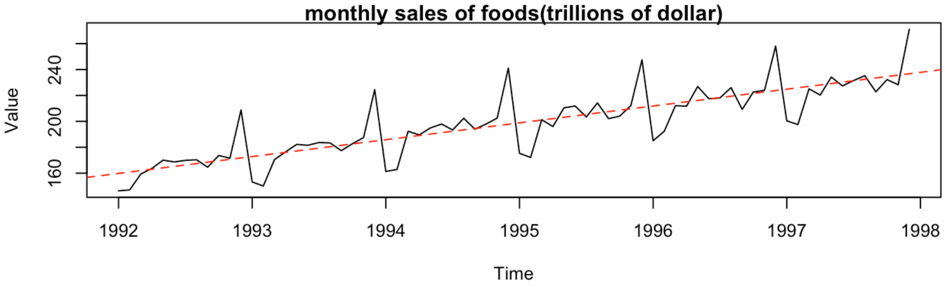
II. **Models for Trend and Seasonality:**

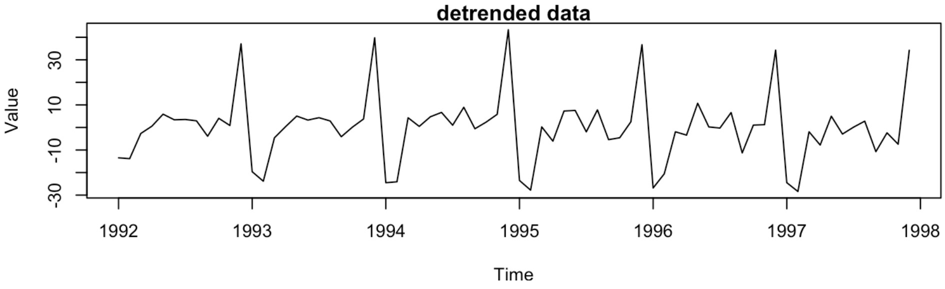
We will use both the linear regression model and method of seasonal means to analysis the overall trend and the seasonality mentioned in part I. Since the numbers shown in millions are too large to interpretive, the rest of the numbers and graph in the project will be all in billions of dollars. The linear regression model:

$x_t = -25749.68 + 13.01\beta_1 t + w_t$, where $w_t \sim wn(0,\sigma^2)$

was fit to estimate the trend. The estimated trend and the detrended data are shown below.

The estimated trend looks appropriate but did not eliminate the seasonal effect within each year.

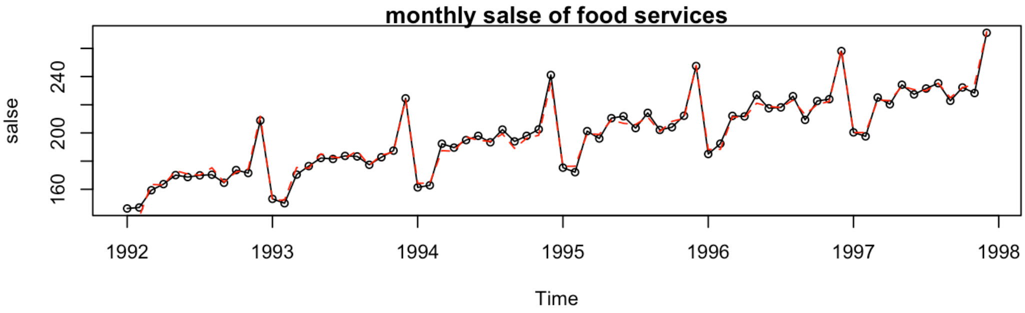


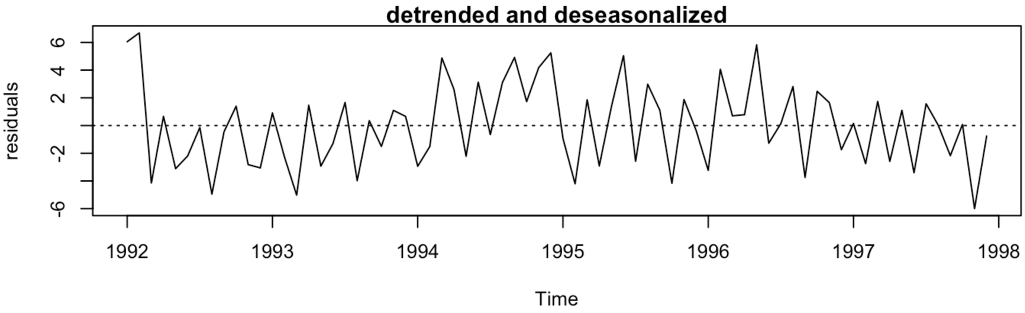


Hence we need to further consider fitting a seasonal means model to remove both overall trend and seasonality: $y_t = 11.98t + \hat{s}_t + x_t$. The value of $\hat{s}_t$ is in the following table, which is the estimated seasonal effect:

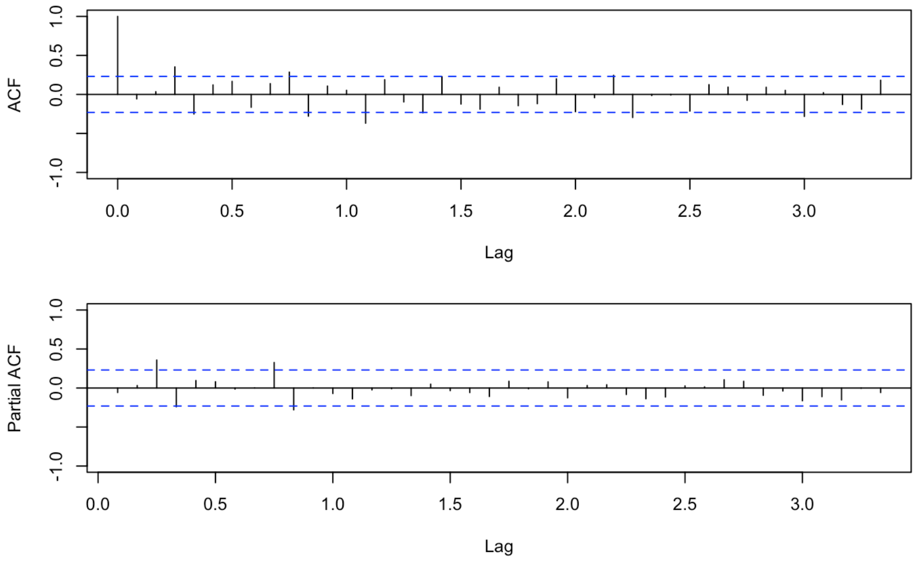
Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
$\hat{s}_1$	$\hat{s}_2$	$\hat{s}_3$	$\hat{s}_4$	$\hat{s}_5$	$\hat{s}_6$	$\hat{s}_7$	$\hat{s}_8$	$\hat{s}_9$	$\hat{s}_{10}$	$\hat{s}_{11}$	$\hat{s}_{12}$
-23728.	-23729.	-23707.	-23709.	-23699.	-23703.	-23705.	-23700.	-23711.	-23705.	-23704.	-23668.

The following figures are the the estimated trend and seasonality of the data, and the data after detrending and deseasonalizing.





We can see that the second plot shows a long-range curved pattern, which means it is pretty close to stationarity after removing the trend and seasonal effect. According to the following graphs of ACF and PACF value, we can further conclude that the detrended and deseasonalized data are stationary.



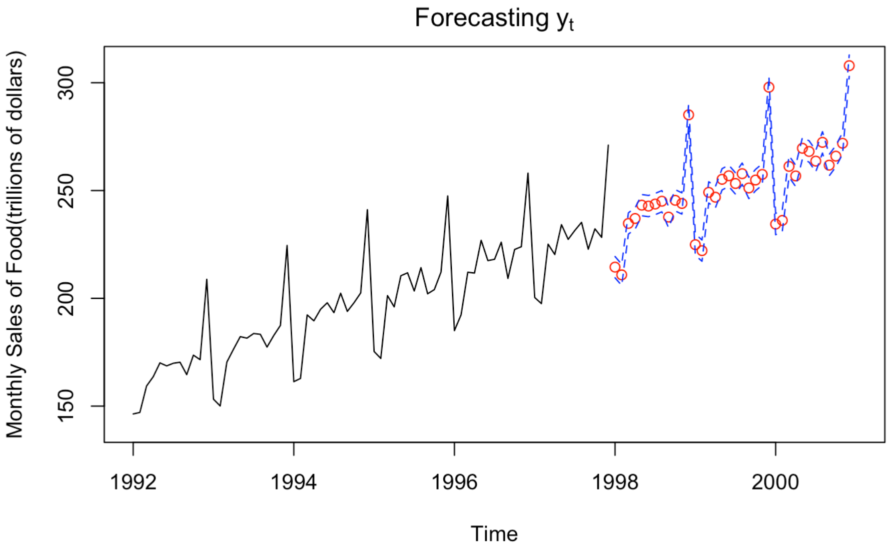
Additionally, since both of ACF and PACF gradually tails off as the number of lags increasing, an ARMA model with p>0 and q>0 could be applicable for this dataset.
A couple of models are tried including ARMA(1, 1), ARMA(2, 1) , ARMA(1, 2), ARMA(2, 2). The summaries of these three models are included in the following table:

Model	AIC
ARMA(1, 1)	359
ARMA(2, 1)	357.43
ARMA(1, 2)	364.54
ARMA(2, 2)	347.94

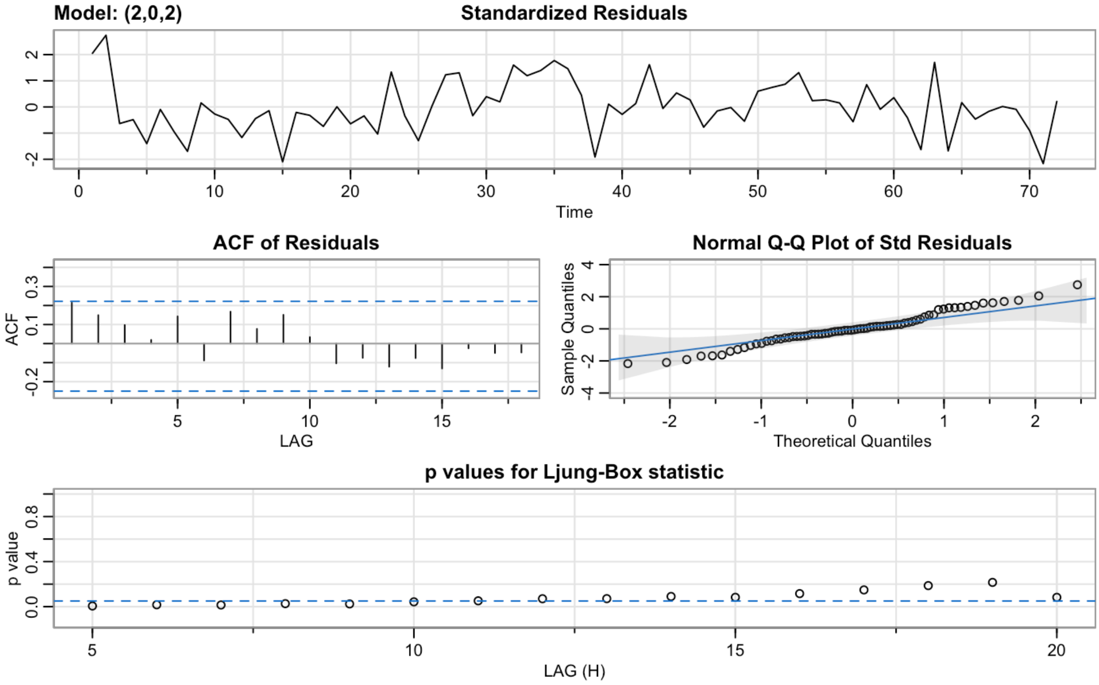
The value of AIC of the model ARMA(2, 2) is the smallest. Hence we can say that the dataset can be best explained by it. Therefore, the final fitted model for this data is a seasonal means model with a linear trend plus an ARMA(2, 2) stationary process:

$$y_t = 11.98t + \hat{s}_t + x_t, \quad x_t = -0.99x_{t-2} - 1.17x_{t-1} + w_t + 1.15w_{t-1} + 0.99w_{t-2}, \quad w_t \sim iidN(0,5.907)$$

The following figure forecast the original time series(containing trend and seasonality) out to December 2000.

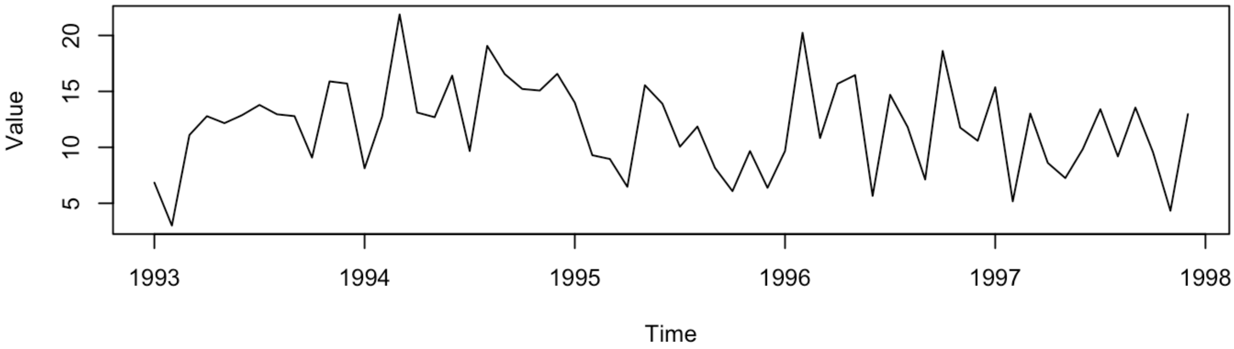


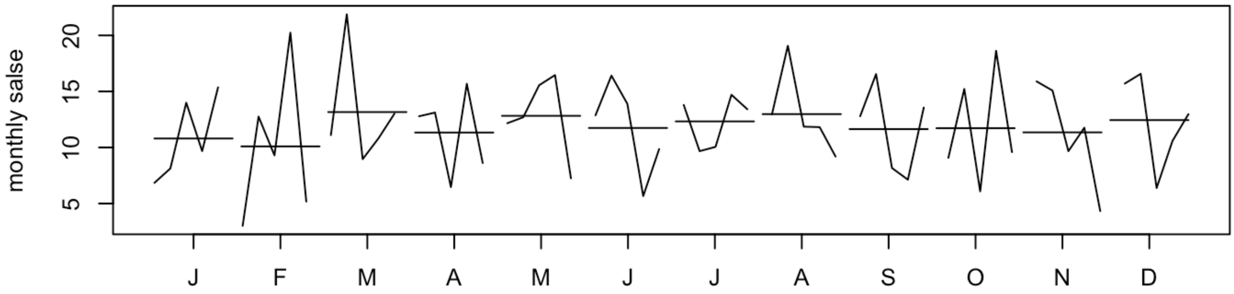
However, the p-value of the Ljung-Box test is pretty small, which means there are no enough evidences to support that iid noise. Hence, even though the value of AIC for this model is relative small. We cannot say it is perfectly fit to the dataset.



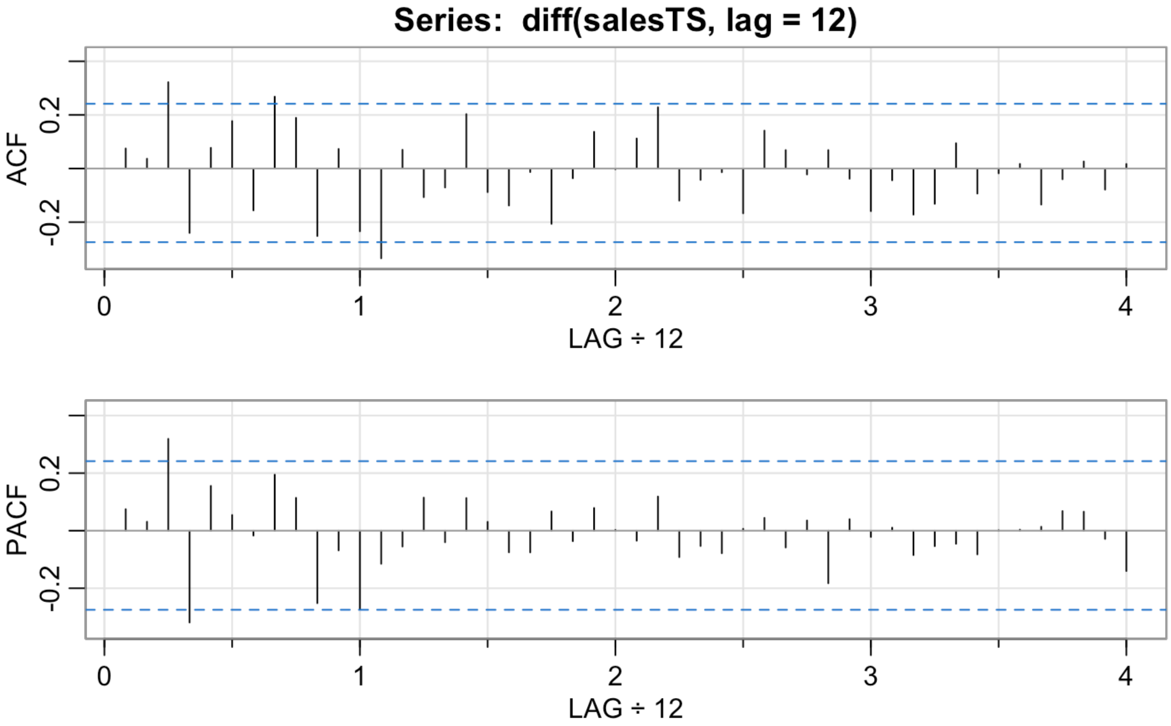
III. SARIMA Modelling:

By using the method of differencing with lag=12, the overall trend of the dataset has been removed. Hence no more differencing is needed. The detrended and deasonalized data and month plot are shown below.



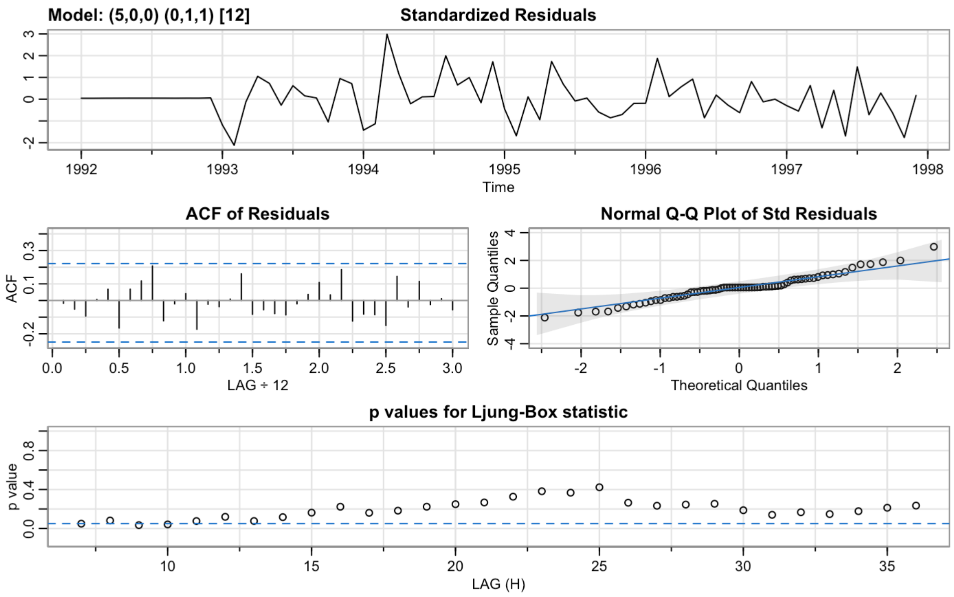


The following figures are ACF and PACF of the differenced data, both of them tails of at around the lag 12.



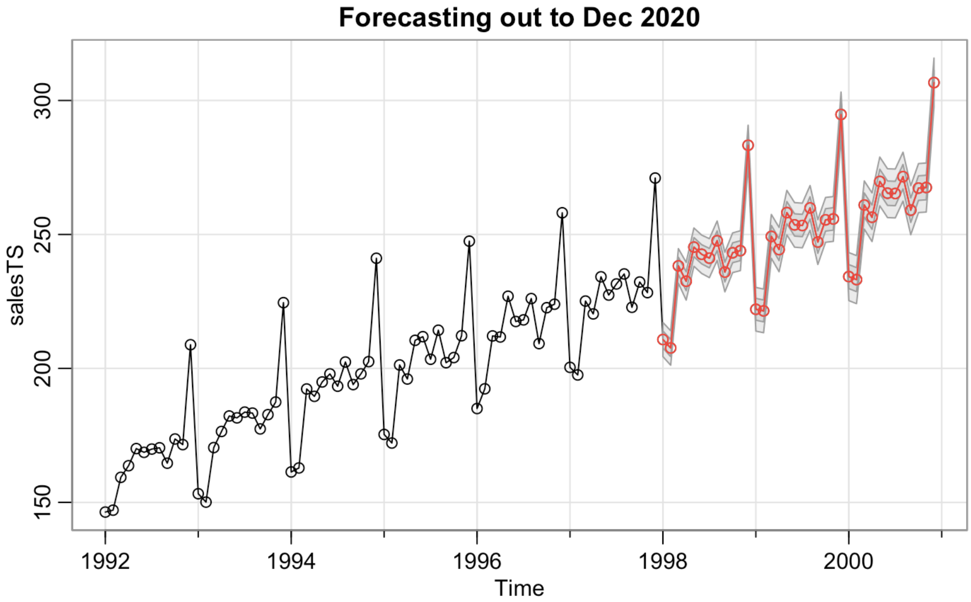
I fit the model with $ARIMA(5{,}0{,}0)*(0{,}1{,}1)_{12}$, $ARIMA(1{,}0{,}0)*(0{,}1{,}1)_{12}$ and $ARIMA(1{,}0{,}1)*(0{,}1{,}1)_{12}$. The AIC summaries and related plots are all shown below. We can see that the value of AIC of model ARIMA(5, 0, 0)*(0, 1, 1)[12] is much smaller than which of the other models. Additionally, the p-values for Ljung-Box statistic of the first model is large, which indicates no evidence against iid noise. And its sample ACF displays no large autocorrelations. Compare to ARIMA(5, 0, 0)*(0, 1, 1)[12], although the residuals of ARIMA(1, 0, 0)*(0, 1, 1)[12] appear to be normally distributed, the sample ACF of it is relatively significant and the Ljung-Box test also indicates there is no enough evidence to support iid noise. Hence in this case, the first model ARIMA(5, 0, 0)*(0, 1, 1)[12] is more appropriate, and the final model is:

$$(1-0.10B-0.44B^3+0.30B^4)(1-B^{12})x_t = (1-0.48B^{12})w_t, \quad w_t \sim idd\ N(0,10.35)$$



model	AIC
ARIMA(5, 0,0)*(0, 1, 1)_12	329.68
ARIMA(1, 0, 0)*(0, 1, 1)_12	338.86
ARIMA(1, 0, 1)*(0, 1, 1)_12	339.74

The following graph shows the forecasting value by using the model above.



IV. Model Comparison:

Both of this two models predict the seasonality and the overall trend well. The result of forecasting for the three years(1998-2000) in the future are increasing with similar seasonality within each year. However, the first-year prediction of ARMA(2, 2) doesn't perform the fluctuation between every month very well, and its value of AIC is relatively larger than which of the ARIMA(5, 0, 0)*(0, 1, 1)[12]. The Ljung-Box test also indicates that ARMA(2, 2) is not white noise, while the p-value of Ljung-Box test for ARIMA(5, 0, 0)*(0, 1, 1)[12] are all significantly large, which means it is white noise.

V. Conclusion:

According to all the sections above, we can see that the the model ARIMA(5, 0, 0)*(0, 1, 1)[12] is a appropriate model for our dataset sales of food services. It not only has relatively small AIC(329.68) but also predict the fluctuation within each year more precisely.