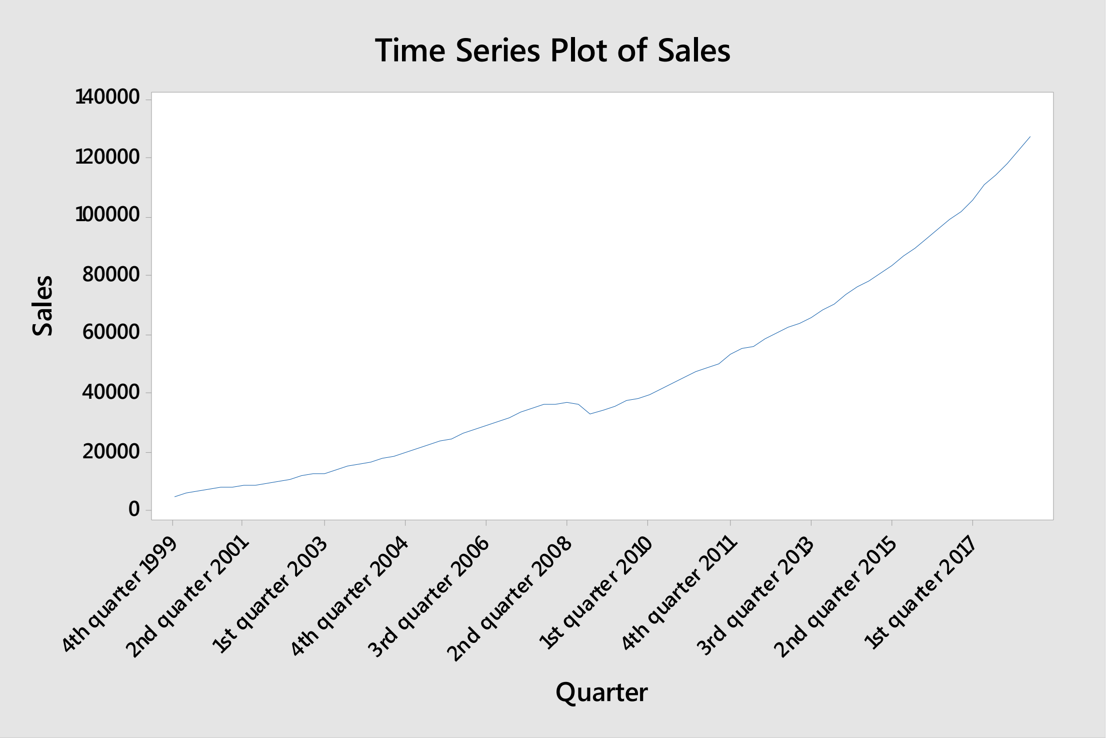
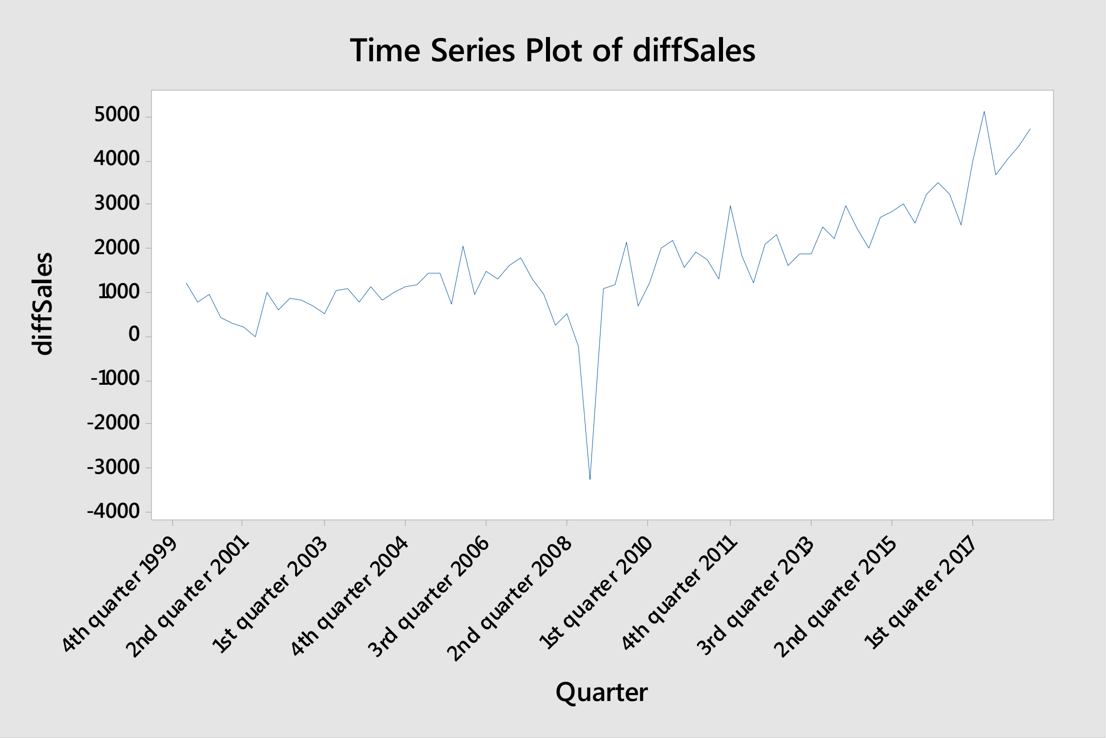
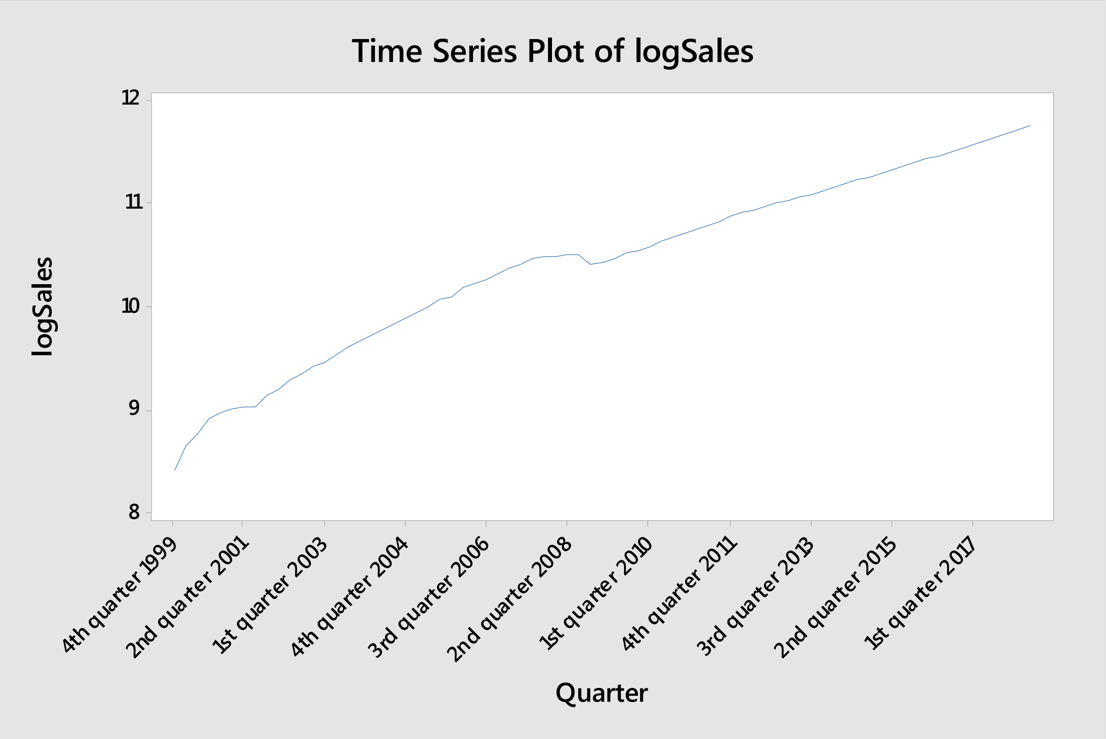
Forecasting Time Series Data Project **Tianze Lin tl2718**

In this project, E-commerce retail sales of the United States from 4th quarter, 1999 to 2nd quarter, 2018 will be analysed. The number of observations in this dataset is 75. The dataset is available on United States Census Bureau website, https://www.census.gov/retail/index.html#ecommerce.

First, plot the sales, the first difference of sales and log sales.

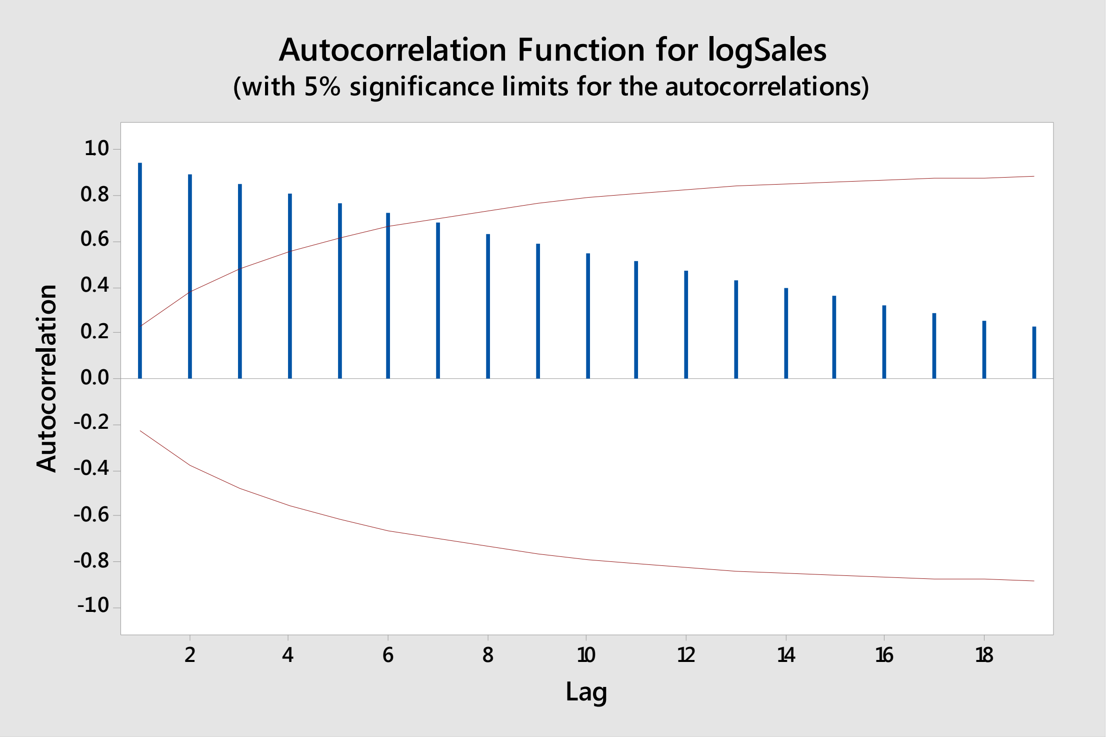


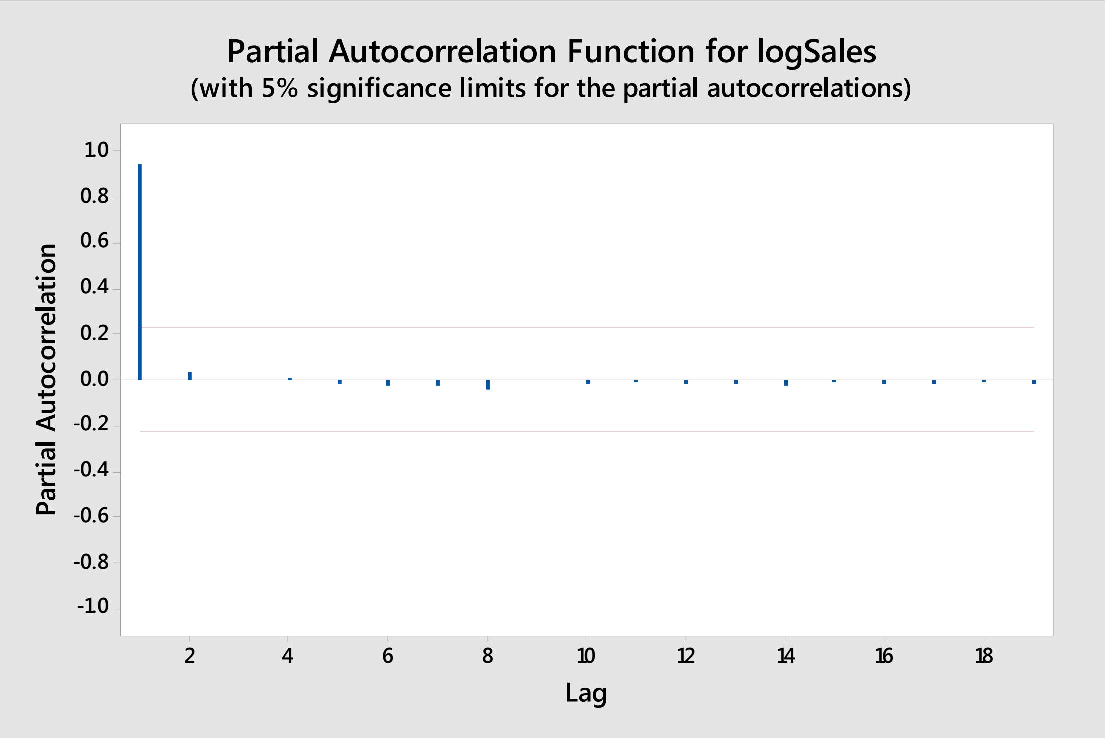




From time series plots above, there is no strong evidence that the volatility of sales is dependent on the level of sales. However, we can see that the sales might have an exponential trend, so log transformation can convert the exponential trend into a linear tread, which is easier to analyse. Furthermore, the change in natural logs is approximately equal to the percentage change of sales. Therefore, we should work with log sales.

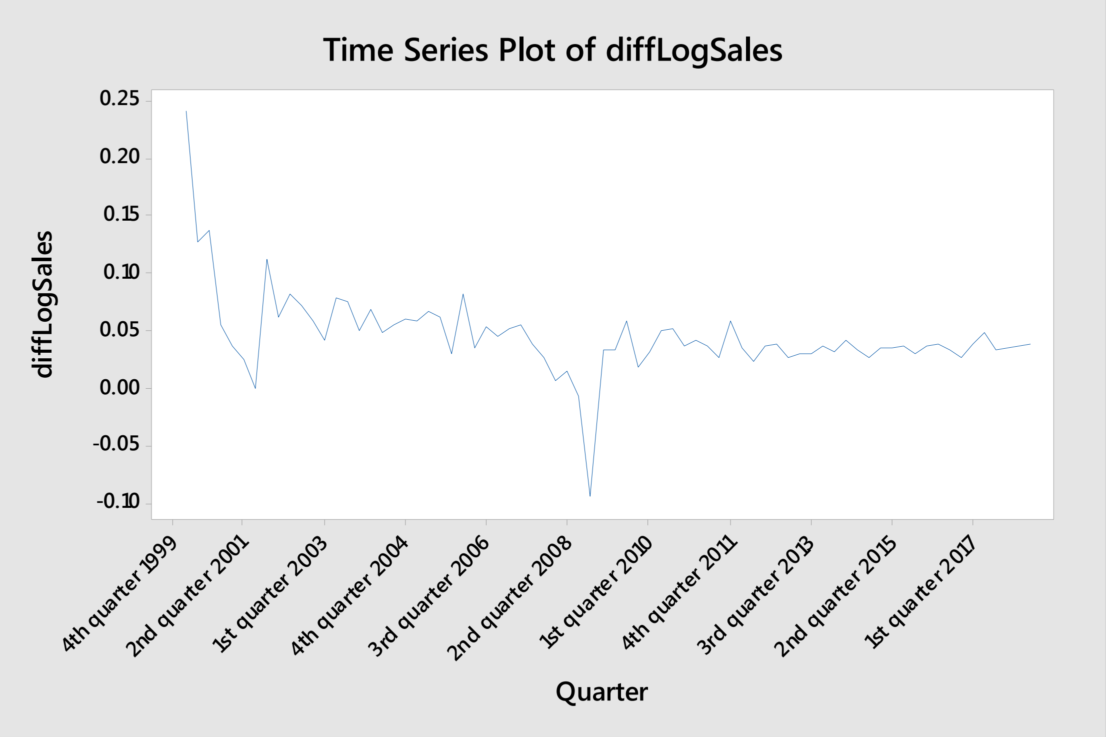
Next, plot ACF and PACF of log sales.

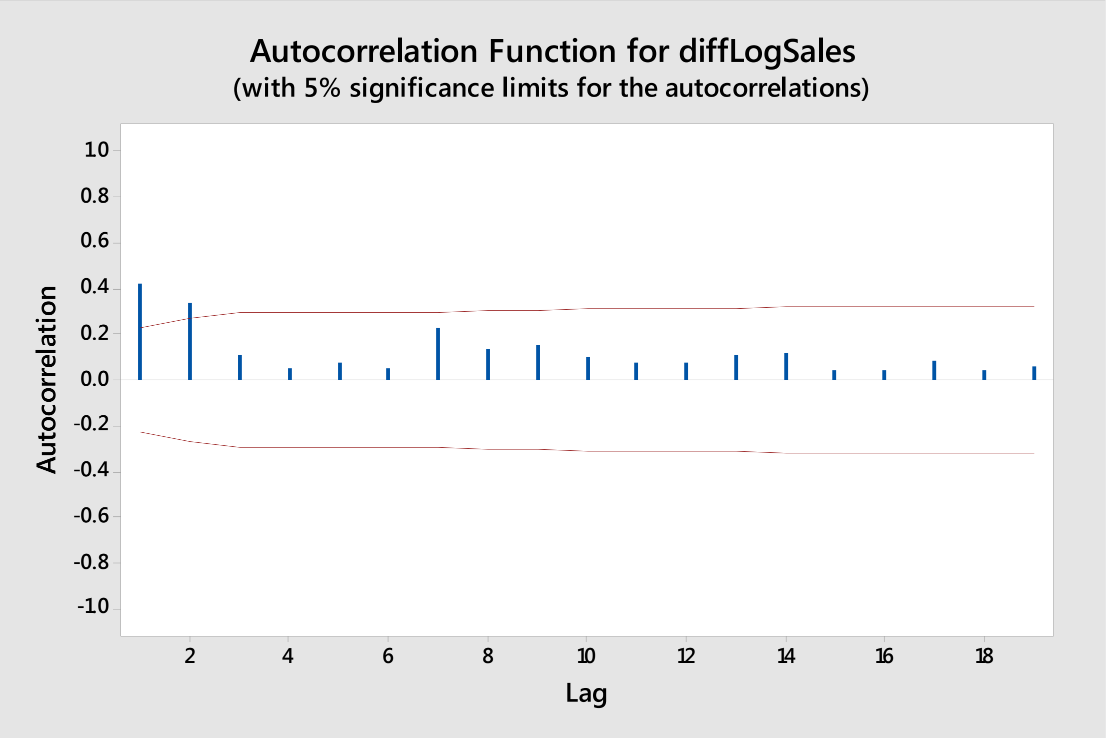


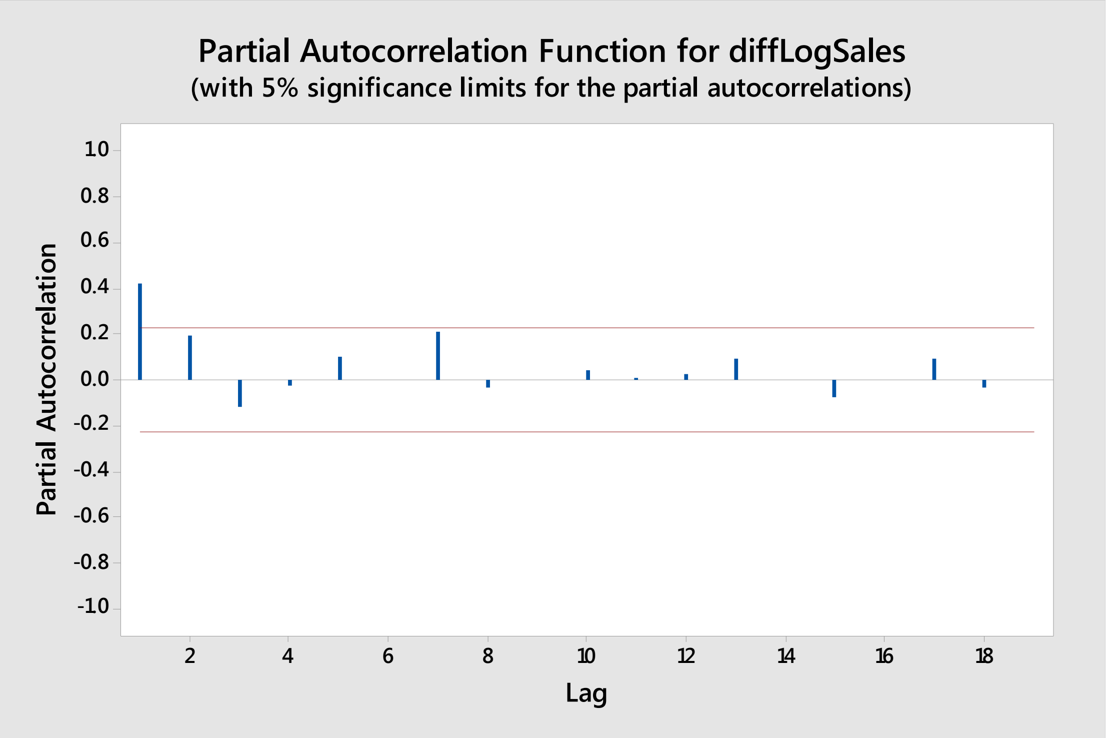


ACF decreases slowly and PACF cuts off beyond lag 1, and this suggests differencing the log sales. From time series plots, ACF and PACF, there is no strong seasonal pattern, so the seasonal component may not exist.

Next, plot the time series, ACF and PACF of differenced log sales.

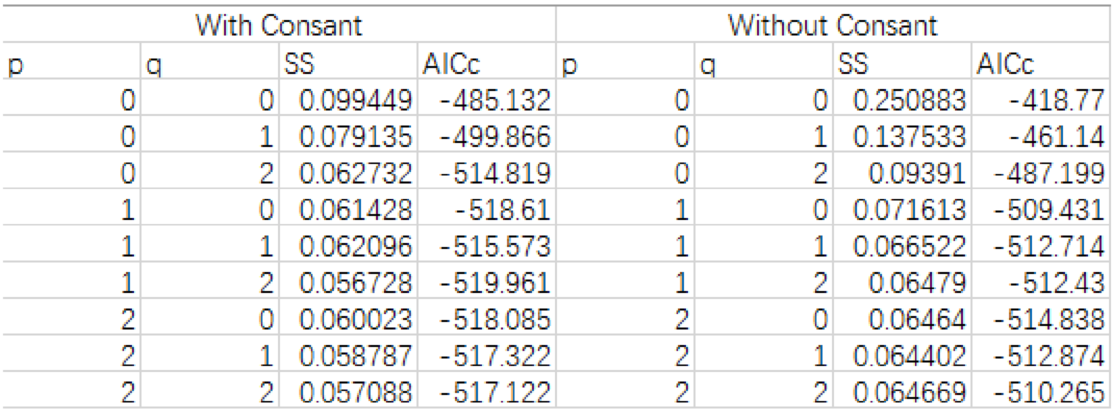






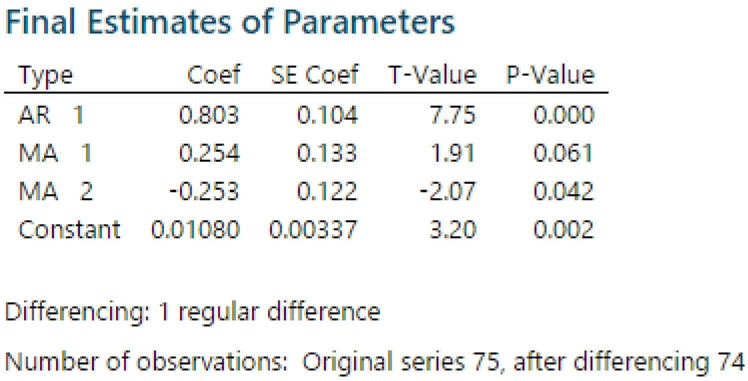
ACF cuts off beyond lag 2 and PACF cuts off beyond lag 1. The autocorrelations are small and pattern less, so higher order of differencing is not needed.

ACF and PACF don’t provide enough information on selecting p and q of ARIMA model, so AICc should be used to select p and q.



With Consant				Without Consant			
p	q	SS	AICc	p	q	SS	AICc
0	0	0.099449	-485.132	0	0	0.250883	-418.77
0	1	0.079135	-499.866	0	1	0.137533	-461.14
0	2	0.062732	-514.819	0	2	0.09391	-487.199
1	0	0.061428	-518.61	1	0	0.071613	-509.431
1	1	0.062096	-515.573	1	1	0.066522	-512.714
1	2	0.056728	-519.961	1	2	0.06479	-512.43
2	0	0.060023	-518.085	2	0	0.06464	-514.838
2	1	0.058787	-517.322	2	1	0.064402	-512.874
2	2	0.057088	-517.122	2	2	0.064669	-510.265

AICc is smallest when p=1 and q=2 with constant, so it suggests an ARIMA (1,1,2) model with a constant term. Next, fit the model with Minitab.



Final Estimates of Parameters

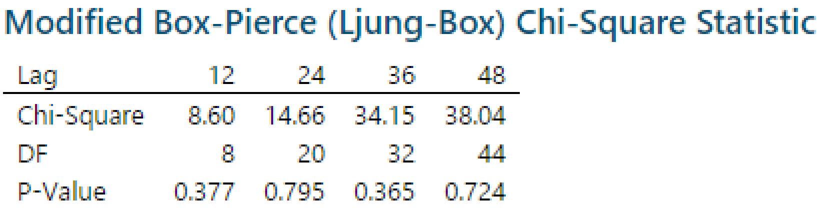
Type	Coef	SE Coef	T-Value	P-Value
AR 1	0.803	0.104	7.75	0.000
MA 1	0.254	0.133	1.91	0.061
MA 2	-0.253	0.122	-2.07	0.042
Constant	0.01080	0.00337	3.20	0.002

Differencing: 1 regular difference

Number of observations: Original series 75, after differencing 74

The p-value of MA 1 coefficient is greater than 0.05, so there is no strong evidence that the MA 1 coefficient is statistically significant at α=0.05. Other coefficients and constant term are statistically significant.

Let x_t be the differenced log sales. Then the complete form of our fitted model is
$x_t = 0.803x_{t-1} + \varepsilon_t - 0.254\varepsilon_{t-1} + 0.253\varepsilon_{t-2} + 0.01080$

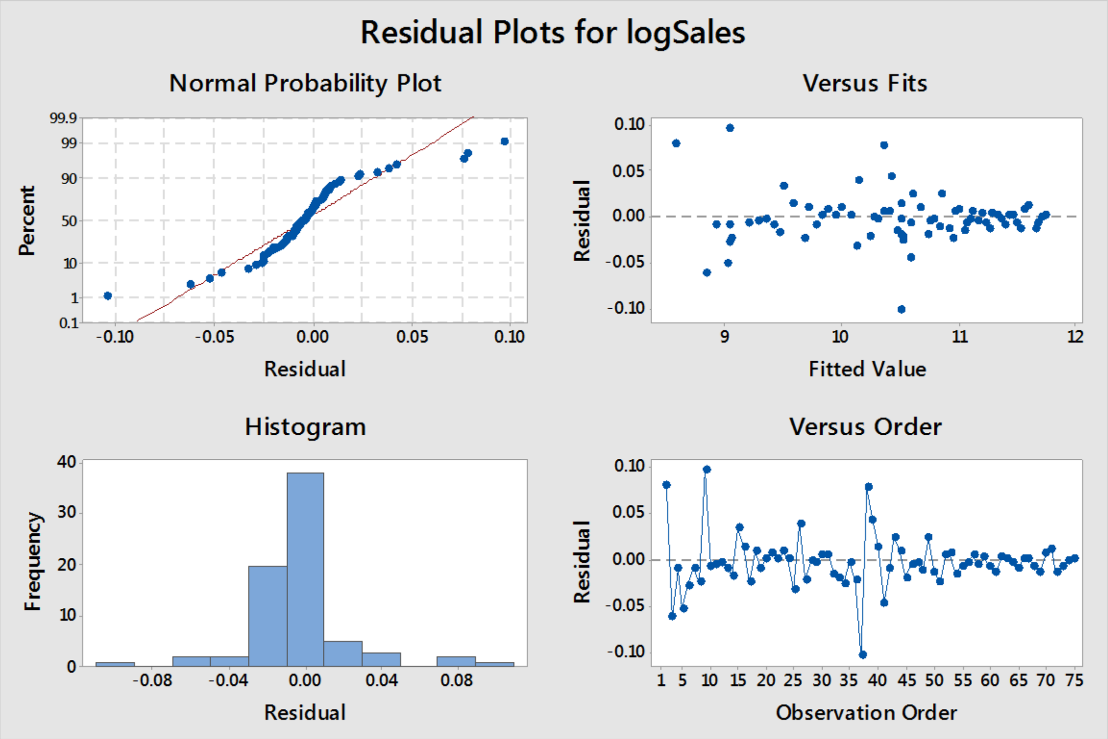


Modified Box-Pierce (Ljung-Box) Chi-Square Statistic

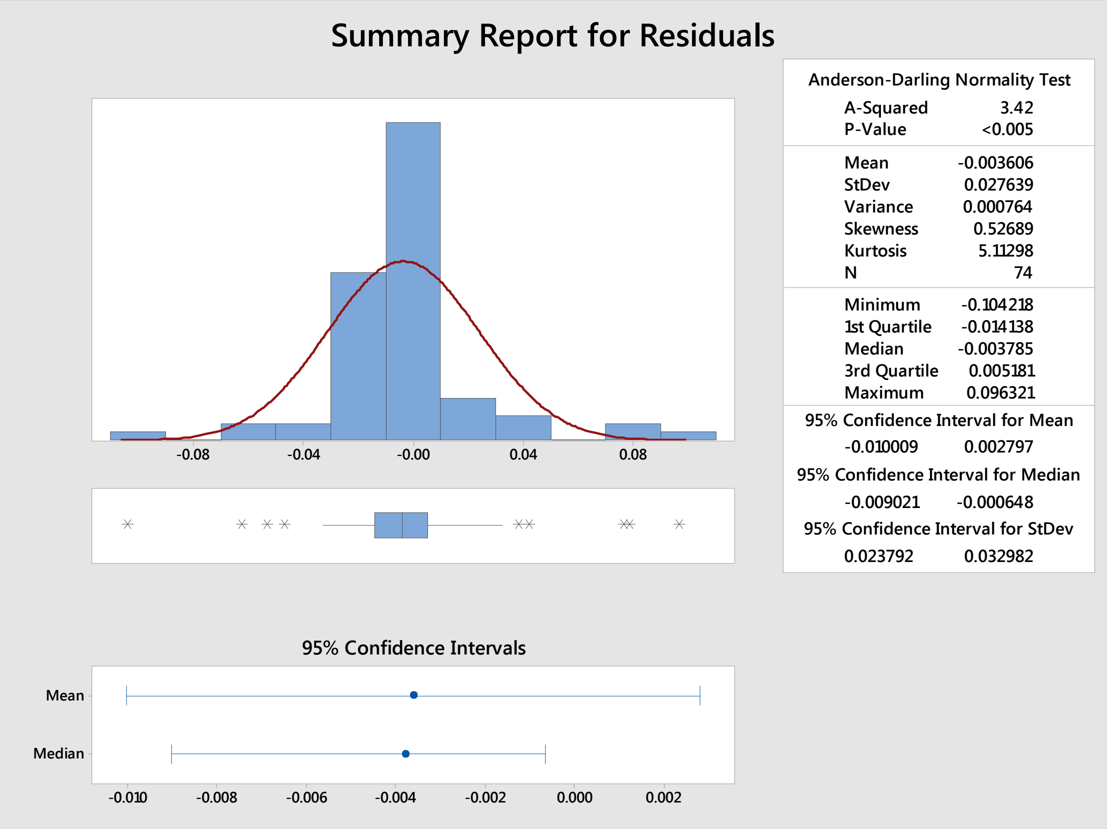
Lag	12	24	36	48
Chi-Square	8.60	14.66	34.15	38.04
DF	8	20	32	44
P-Value	0.377	0.795	0.365	0.724

The Box-Pierce Chi-Square Statistic shows that all the p-values at different lags are higher than 0.05, so there is no strong evidence that our model is inadequate at the significance level of 0.05.

Next, let’s plot the residuals from fitted model and the ACF and PACF of residuals.



Residual Plots for logSales



Summary Report for Residuals

Anderson-Darling Normality Test

A-Squared	3.42
P-Value	<0.005
Mean	-0.003606
StDev	0.027639
Variance	0.000764
Skewness	0.52689
Kurtosis	5.11298
N	74
Minimum	-0.104218
1st Quartile	-0.014138
Median	-0.003785
3rd Quartile	0.005181
Maximum	0.096321

95% Confidence Interval for Mean

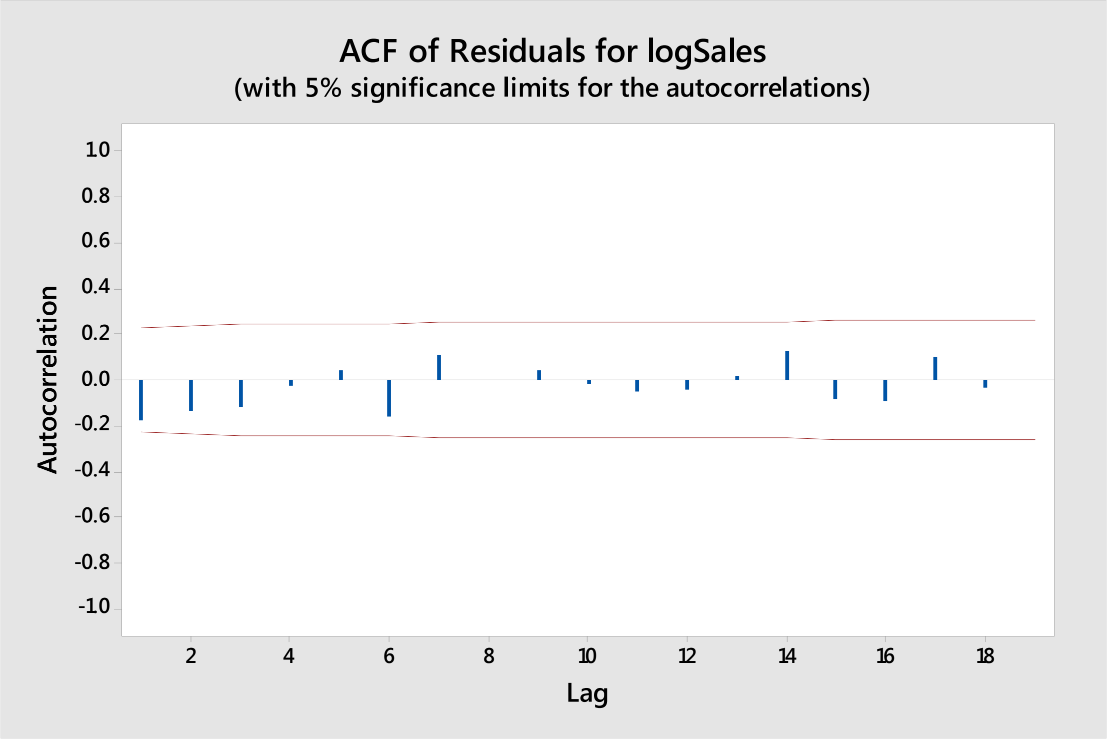
-0.010009 0.002797

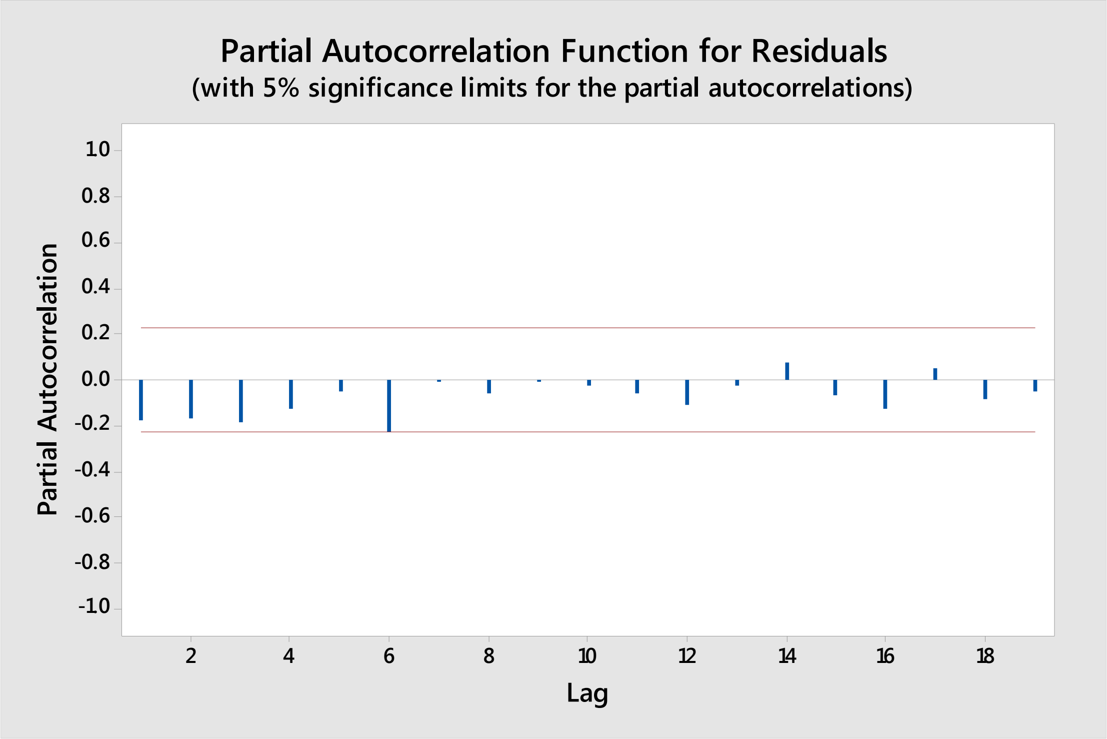
95% Confidence Interval for Median

-0.009021 -0.000648

95% Confidence Interval for StDev

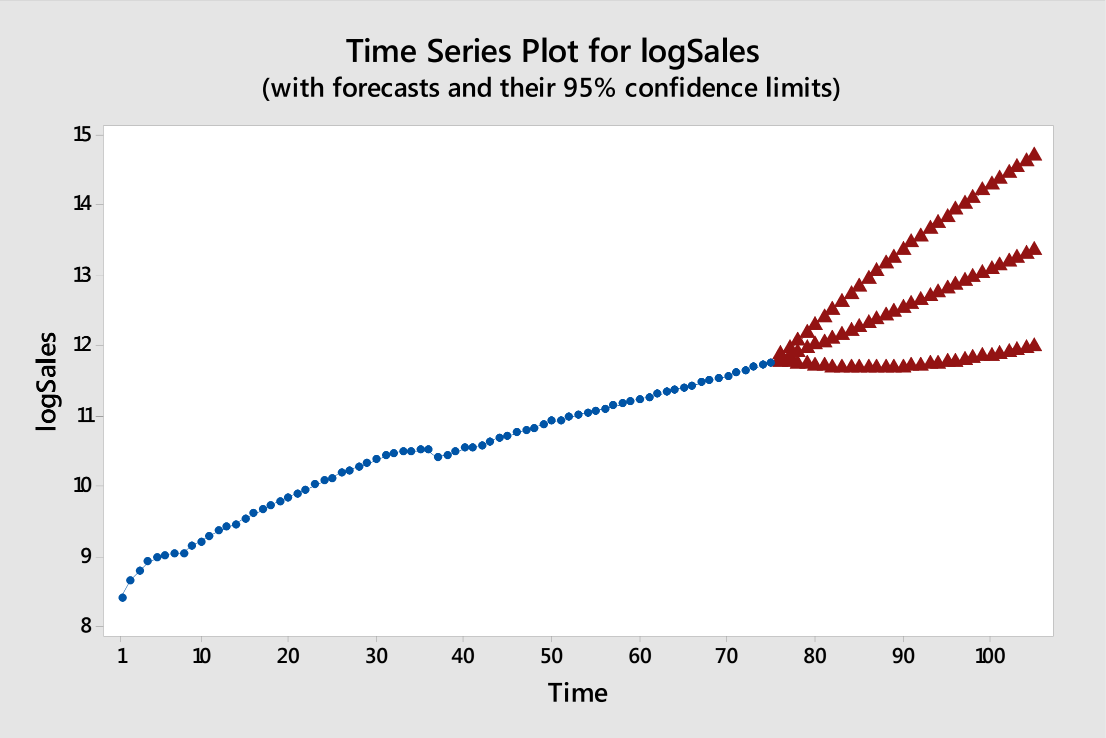
0.023792 0.032982



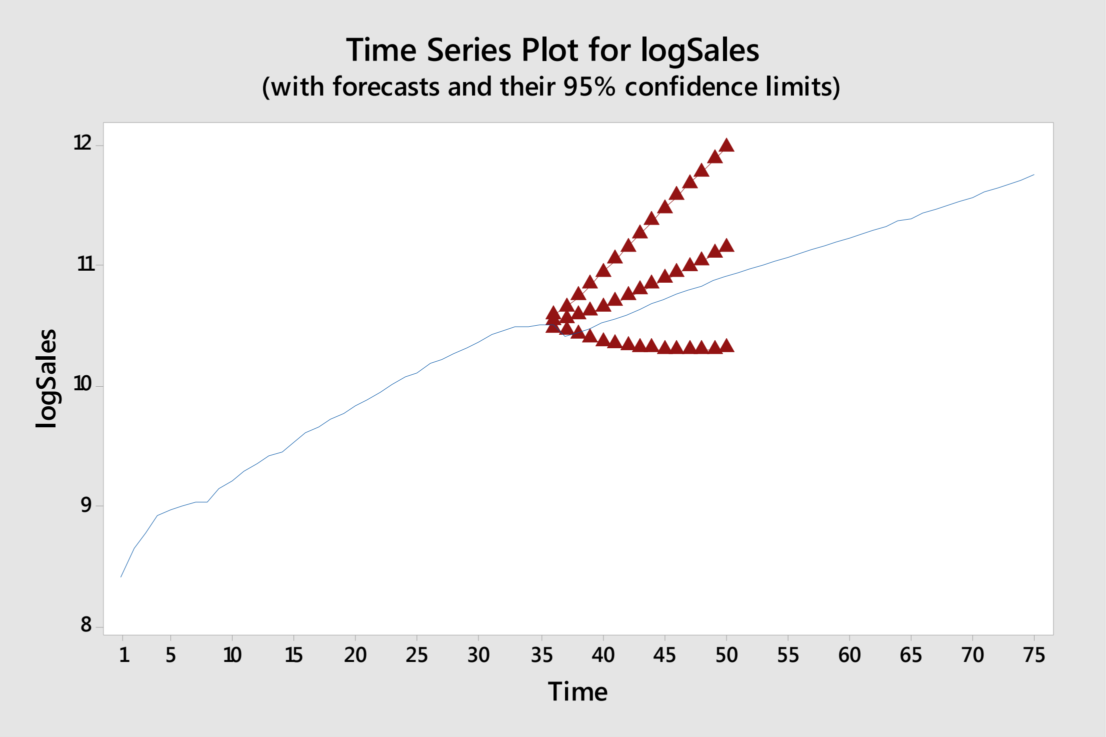


There is no strong evidence that residuals are autocorrelated. This doesn’t indicate any inadequacies of our model.

Next, obtain forecast and 95% forecast intervals for lead times 1-30 and plot them.

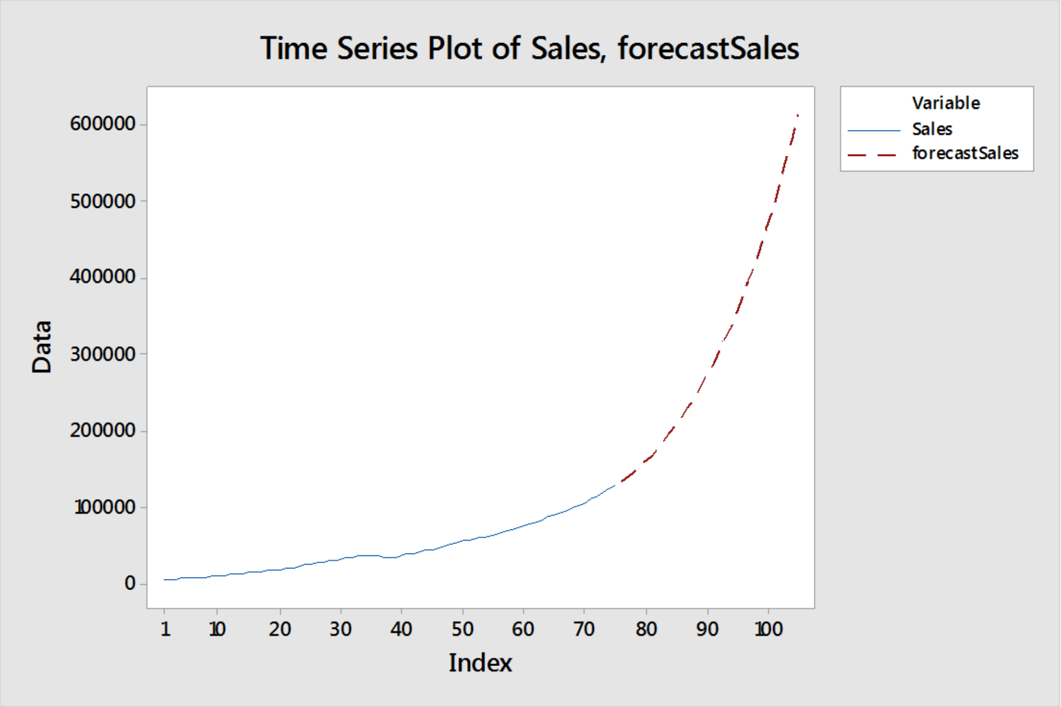


These forecasts seem reasonable since they fit the linear trend with previous observations, and these forecast intervals seem too wide, but actually these intervals are narrow. We can verify this by fitting ARIMA (1,1,2) with constant from observations 1 to 35 and obtain 95% forecast intervals for lead times 1 to 15.



From the plot, the observations 36 and 37 almost exceed the forecast intervals. This indicates the forecast intervals are narrow. This might be caused by the heavy-tailedness of shocks in our model.

Finally, calculate the exponential of forecast log sales to obtain forecast sales and plot them with sales.


Time Series Plot of Sales, forecastSales
Data
Index
Variable
Sales
forecastSales