Forecasting Time Series Data Projet 1

Shalem Sumanthiran - sps9893

For this project we will be using ARIMA methods to forecast the dataset of E-commerce Retail Sales. The specifics of the dataset are as follows:

Source: U.S. Census Bureau

Release: Quarterly Retail E-Commerce Sales

Units: Millions of Dollars, Seasonally Adjusted

Frequency: Quarterly

E-commerce sales are sales of goods and services where the buyer places an order, or the price and terms of the sale are negotiated over an Internet, mobile device (M-commerce), extranet, Electronic Data Interchange (EDI) network, electronic mail, or other comparable online system. Payment may or may not be made online.

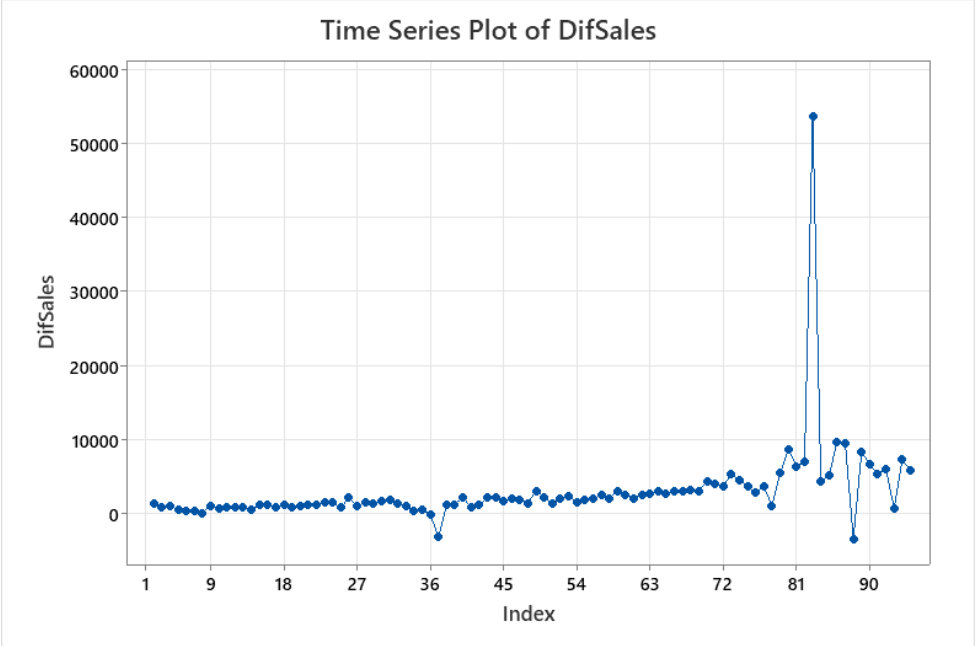
Suggested Citation:

U.S. Census Bureau, E-Commerce Retail Sales [ECOMSA], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/ECOMSA, October 31, 2023.

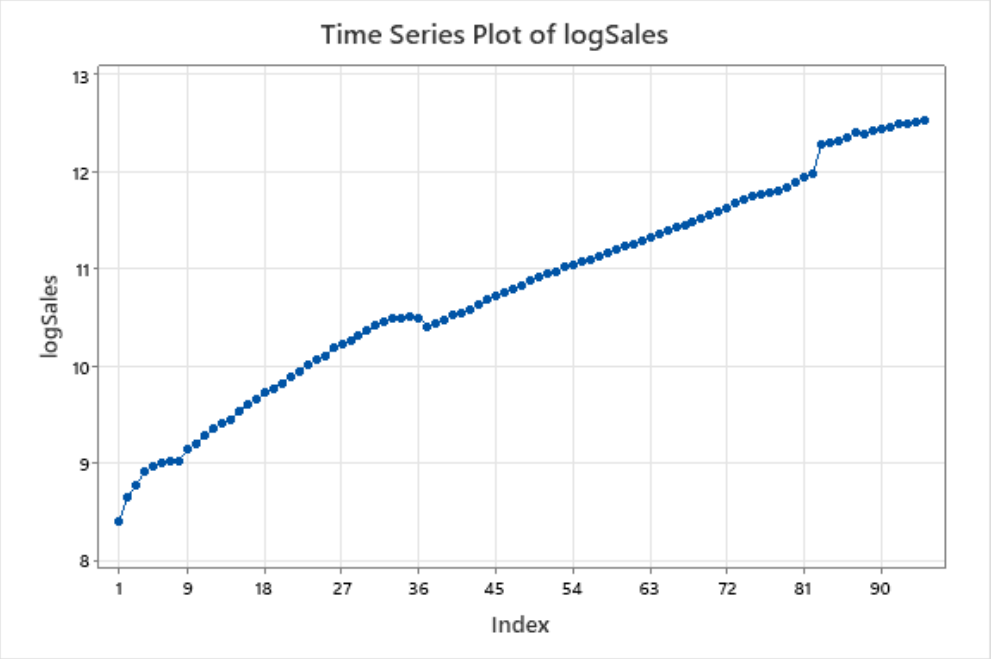
There are 95 observations, with the first observation being from the last quarter of 1999, and the most recent observation is from the first quarter of 2023.

First we plot the sales:

We also take the difference:

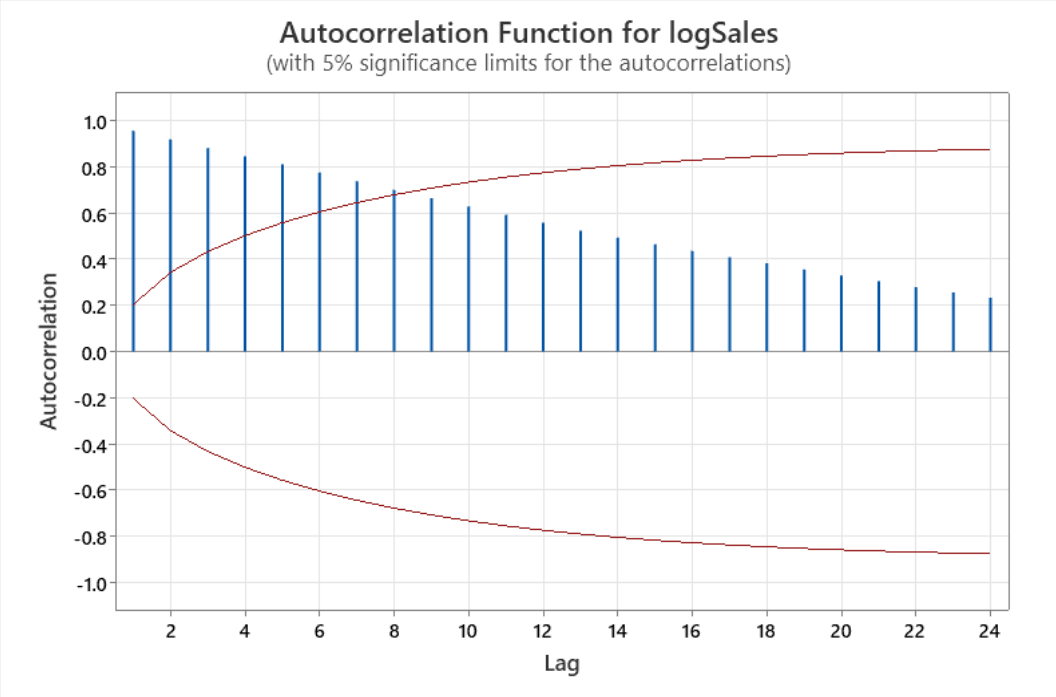


There does not seem to be level-dependent volatility. However, the values are increasing, and there may be an exponential trend. Therefore we take the log of E-commerce sales:

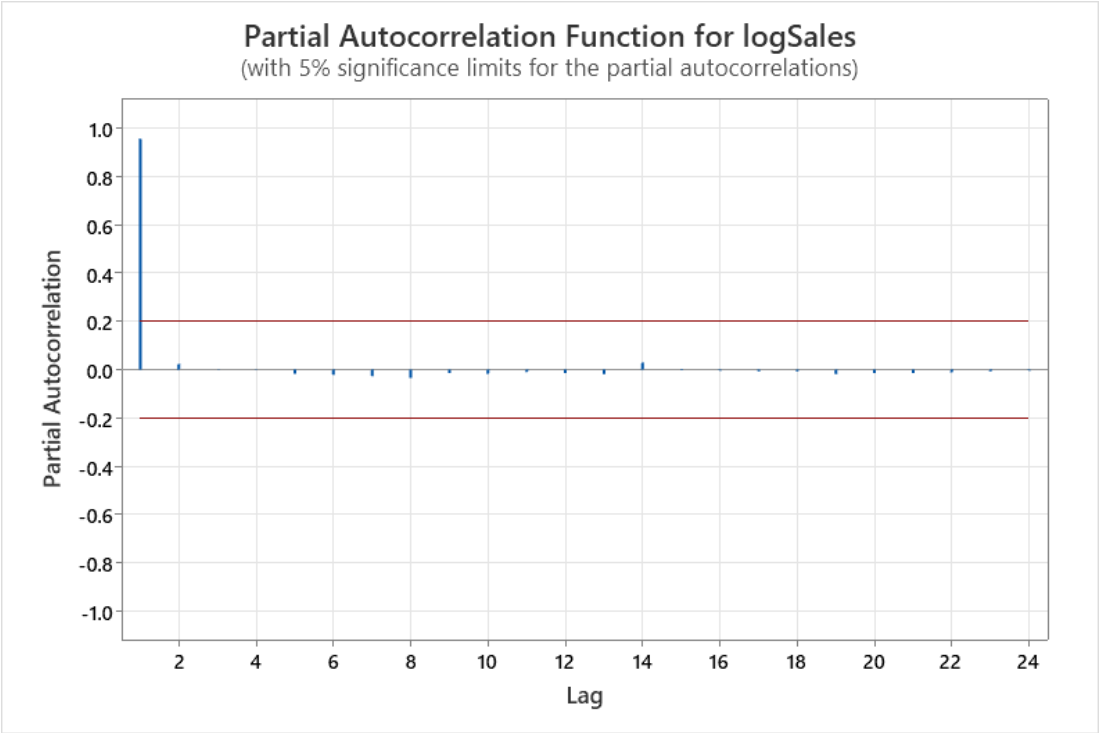


This converts the exponential trend into a linear trend, which is easier to analyze, and is roughly equivalent to the percentage change in sales, so we will work with the log of sales.

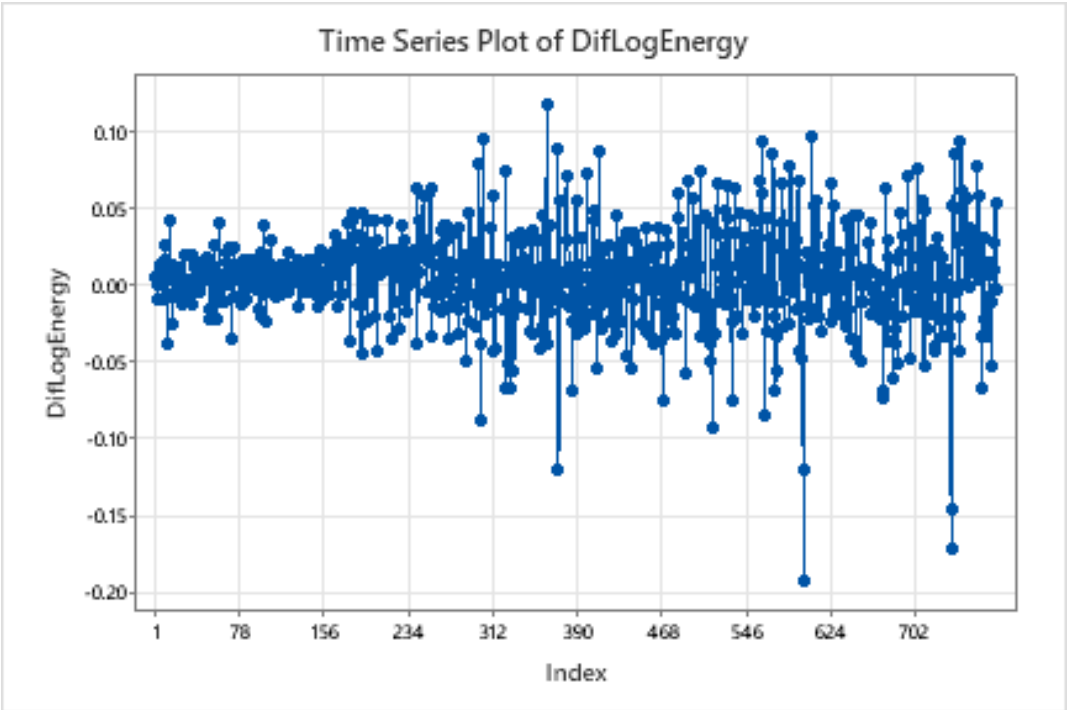
Next we take the Autocorrelation function for Log of Sales:

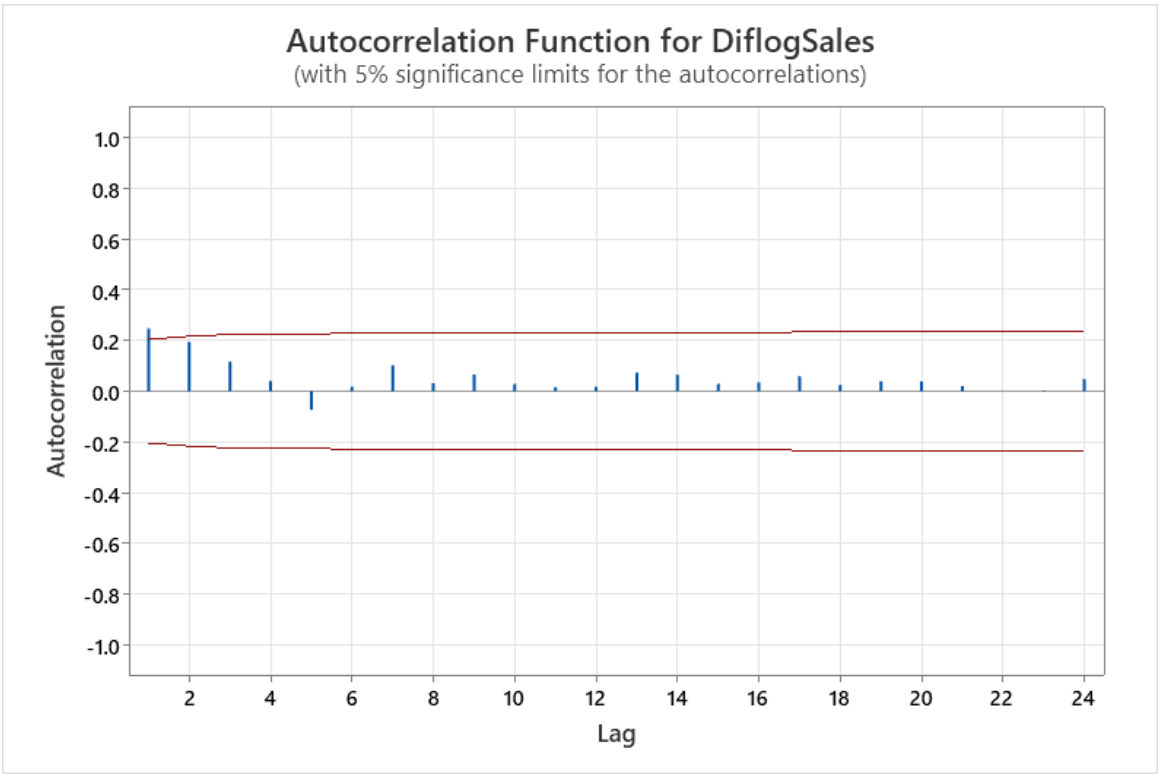


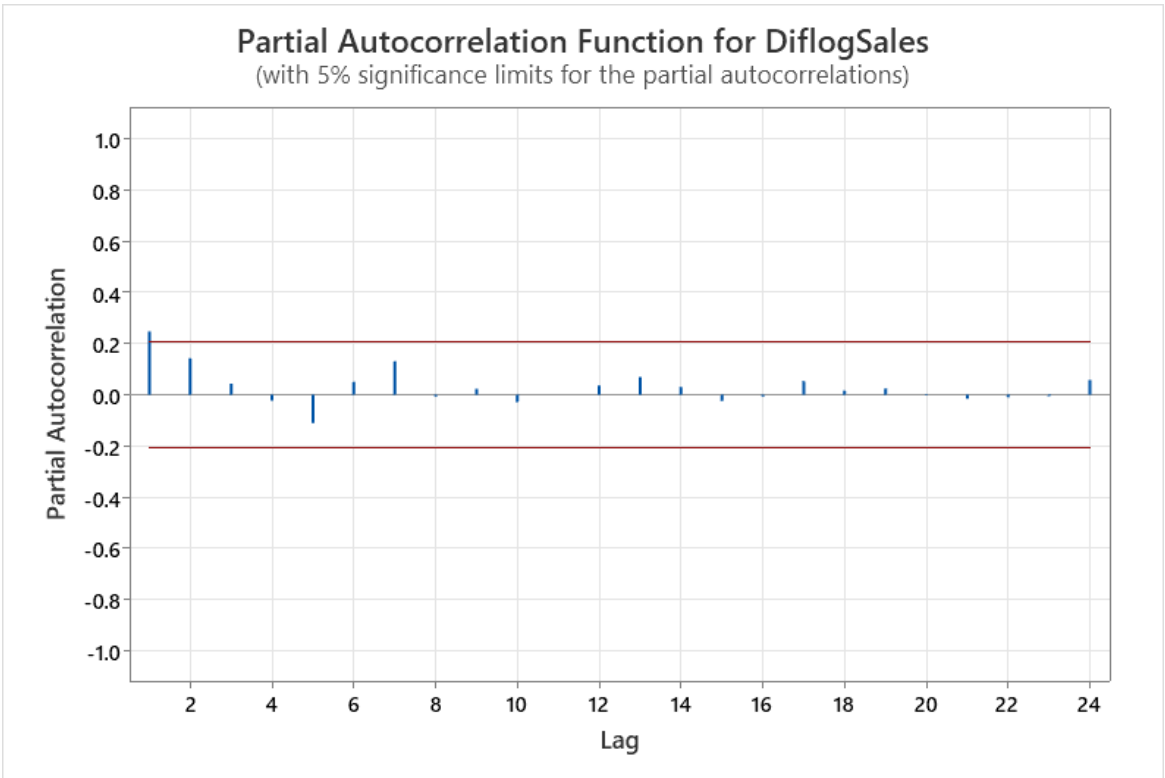
And the Partial Autocorrelation function for Log of Sales



The autocorrelation function dies down and the partial autocorrelation function cuts off after lag 1. Furthermore, there does not seem to be any seasonal component from all the graphs so far, so we do not need to adjust the data any further. We can also plot the time series, ACF and PACF for the differenced values:




Autocorrelation Function for DiflogSales
(with 5% significance limits for the autocorrelations)
Autocorrelation
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
-1.0
2
4
6
8
10
12
14
16
18
20
22
24
Lag


Partial Autocorrelation Function for DiflogSales
(with 5% significance limits for the partial autocorrelations)
Partial Autocorrelation
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
-1.0
2
4
6
8
10
12
14
16
18
20
22
24
Lag

The ACF cuts off after lag 1 and the PACF also cuts off after lag 1. There is no need for further differencing as there is no discernible pattern, and the autocorrelations are small. We also observe the AICc both with and without a constant:

With constant:

Model Selection

Model (d = 1)	LogLikelihood	AICc	AIC	BIC
p = 1, q = 1*	168.591	-328.733	-329.183	-319.009
p = 2, q = 0	168.586	-328.723	-329.173	-319.000
p = 1, q = 0	167.041	-327.815	-328.081	-320.452
p = 1, q = 2	168.980	-327.277	-327.959	-315.243
p = 2, q = 1	168.816	-326.950	-327.632	-314.916
p = 2, q = 2	169.954	-326.943	-327.909	-312.649
p = 0, q = 2	167.257	-326.064	-326.514	-316.340
p = 0, q = 1	165.898	-325.530	-325.797	-318.167
p = 0, q = 0	163.215	-322.298	-322.430	-317.343

** Best model with minimum AICc. Output for the best model follows.*

Without constant:

Model Selection

Model (d = 1)	LogLikelihood	AICc	AIC	BIC
p = 1, q = 1*	165.679	-325.091	-325.358	-317.728
p = 2, q = 1	166.679	-324.908	-325.358	-315.184
p = 1, q = 2	166.136	-323.822	-324.272	-314.099
p = 2, q = 2	167.181	-323.680	-324.362	-311.643
p = 2, q = 0	163.054	-319.841	-320.108	-312.478
p = 1, q = 0	156.188	-308.243	-308.375	-303.288
p = 0, q = 2	145.244	-282.222	-282.488	-284.858
p = 0, q = 1	143.030	-280.928	-282.060	-276.973
p = 0, q = 0	125.226	-256.408	-256.451	-253.908

Best model with minimum AICc. Output for the best model follows.

The AICc is smallest when p=1 and q=1, which suggests an ARIMA (1,1,1) with a constant term. Therefore when we fit the model with Minitab, we get the following:

Final Estimates of Parameters

Type	Coef	SE Coef	T-Value	P-Value
AR 1	0.792	0.132	5.98	0.000
MA 1	0.495	0.186	2.67	0.009
Constant	0.00975	0.00209	4.67	0.000

Differencing: 1 Regular
Number of observations after differencing: 94

All coefficients and the constant term are statistically significant at $\alpha = 0.05$.

Therefore the complete form of the fitted model is:

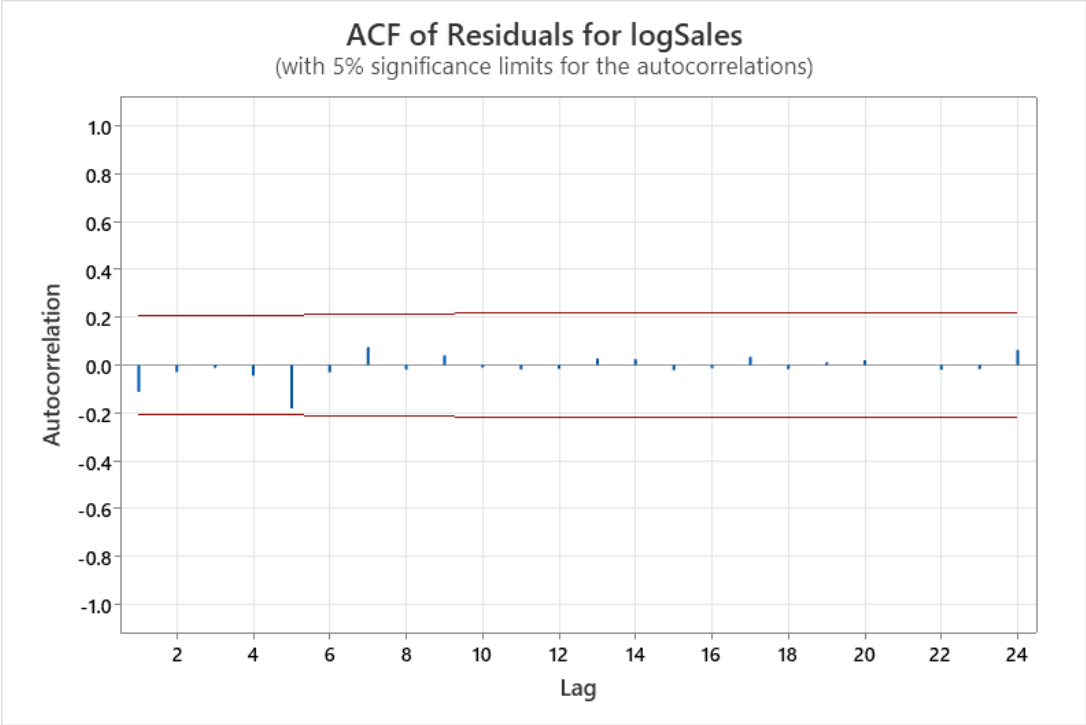
$$x_t = 0.792x_{t-1} + \varepsilon_t - 0.495\varepsilon_{t-1} + 0.00975$$

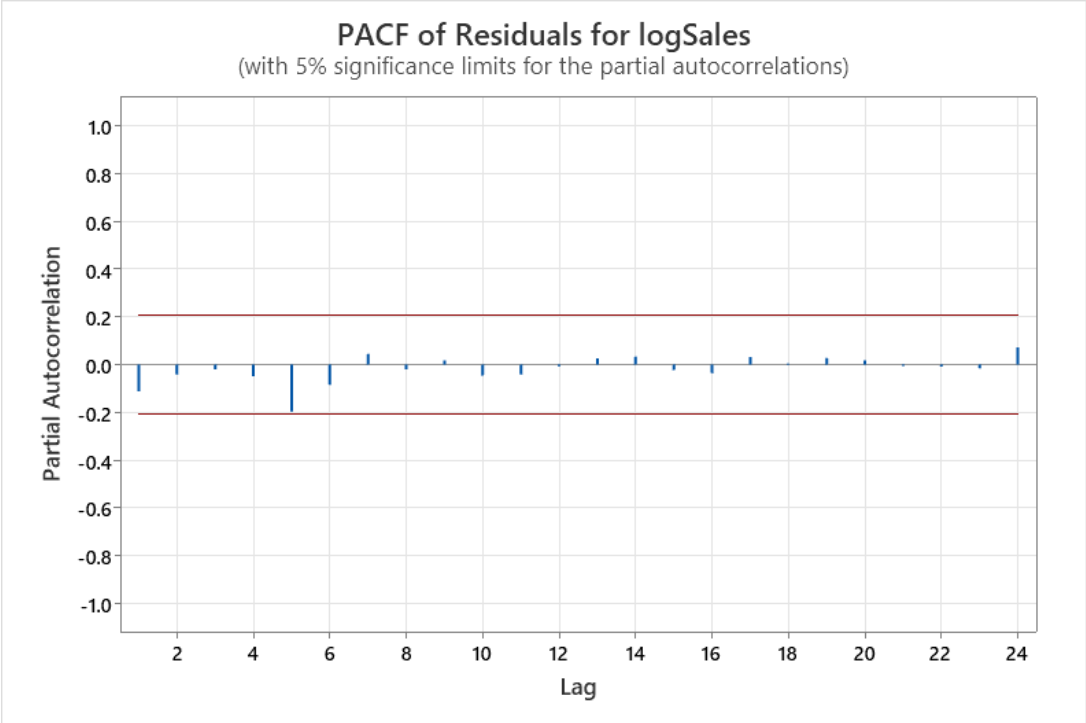
Modified Box-Pierce (Ljung-Box) Chi-Square Statistic

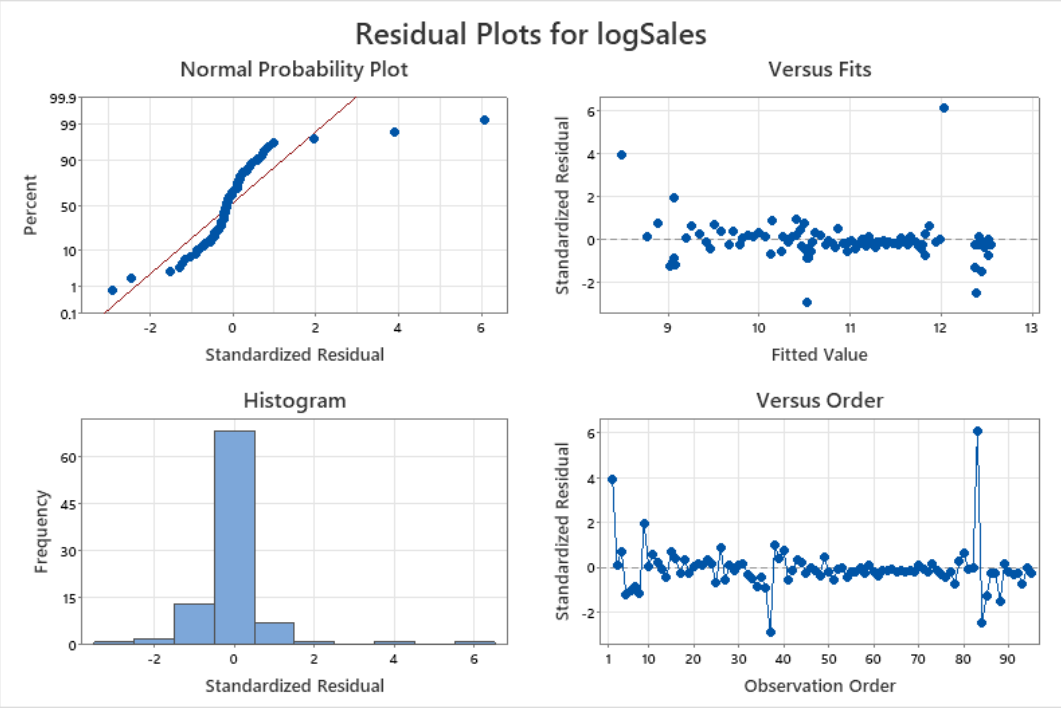
Lag	12	24	36	48
Chi-Square	5.82	6.88	13.78	26.83
DF	9	21	33	45
P-Value	0.758	0.998	0.999	0.986

The Ljung-Box statistics indicate that the p-values are greater than 0.05 at all different lags, therefore there is no evidence to suggest that the model is inadequate.

Next we plot the residuals and the ACF and PACF of the residuals,

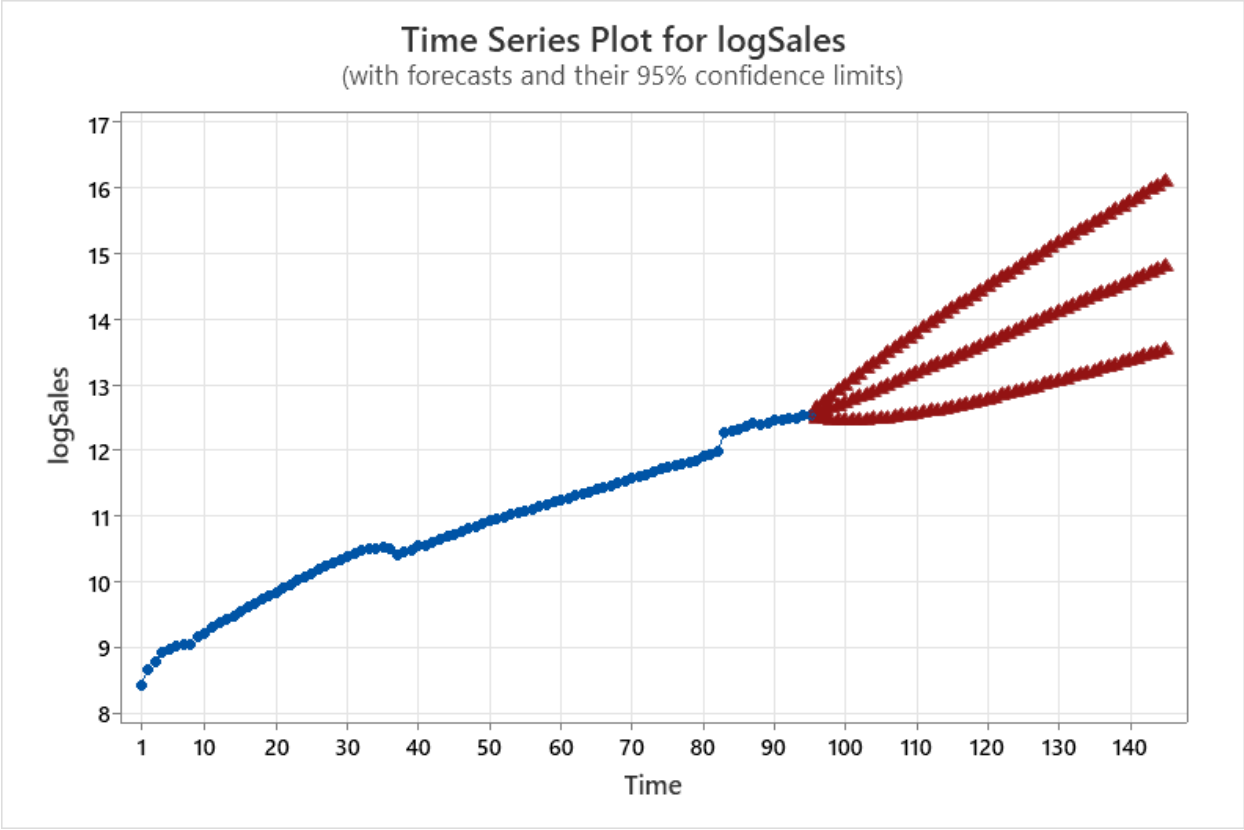

ACF of Residuals for logSales
(with 5% significance limits for the autocorrelations)
Autocorrelation
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
-1.0
2
4
6
8
10
12
14
16
18
20
22
24
Lag


PACF of Residuals for logSales
(with 5% significance limits for the partial autocorrelations)
Partial Autocorrelation
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
-1.0
2
4
6
8
10
12
14
16
18
20
22
24
Lag



There is no strong evidence to suggest that the residuals are autocorrelated, so we reject any potential inadequacies of the model.

Finally, we plot the the data, together with the forecasts, and the 95% forecast intervals.



The forecast seems reasonable as it fits the linear trend of the previous observations; with what appears to be reasonably wide forecast intervals.

To obtain the final forecast along with the original series, we calculate the exponential of the forecast values and add them to the original series: