

BSTA 477/677 – Winter 2021

Tutorial 5 - March 28th, 2021

[Choosing ARIMA models](#)

[Note about data set](#)

[Choose ARIMA models: ACF, PACF](#)

[ARIMA with SAS EG](#)

[Training set](#)

[Forecast and validation](#)

[Error terms](#)

[Training set](#)

[Validation set](#)

[ARIMA with seasonality](#)

Choosing ARIMA models

Note about data set

Ensure that your dataset has the following:

- Dependent variable
- Time variable that has the correct Time data type in SAS (Check variable type).
- Saved in a permanent SAS library or specific place.

When choosing the ARIMA model, we first look at the ACF and PACF for the **full dataset**. Once we have a sense of which ARIMA model to use, we apply to the training set. Adjust the ARIMA model further based on the training set outputs.

Choose ARIMA models: ACF, PACF

The following are the general steps to select ARIMA models:

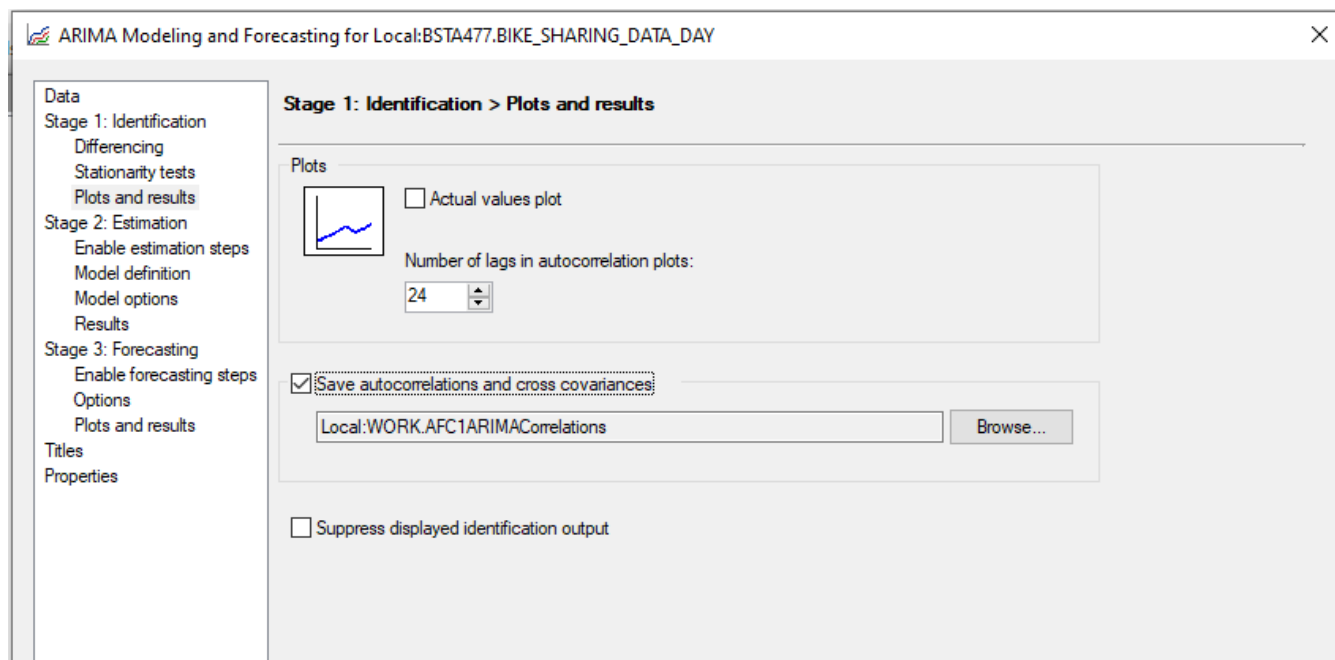
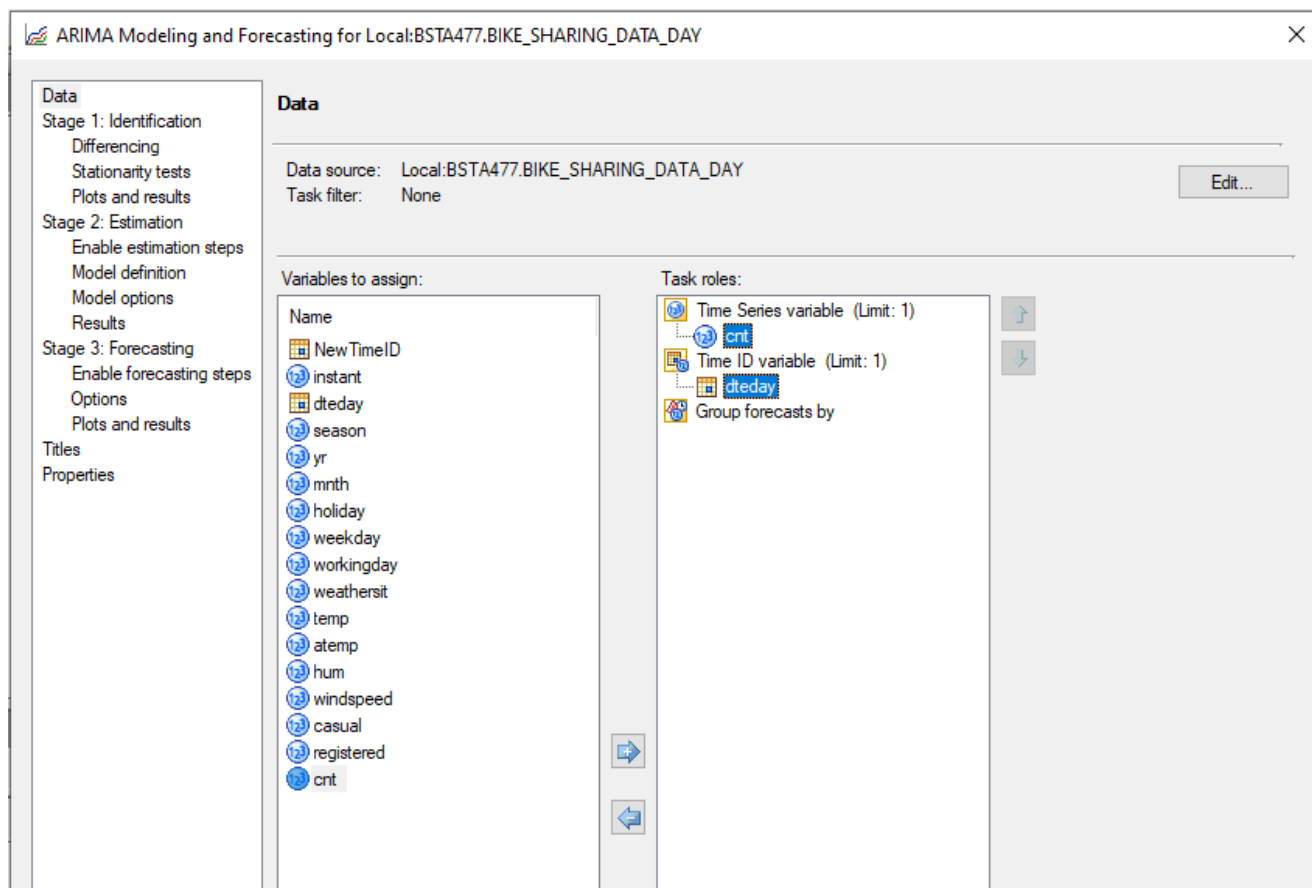
1. Run the ARIMA modeling and forecasting task on the full dataset.
2. Examine autocorrelations: ACF and PACF plots
3. Conduct differencing (if trend component is present)
4. Determine AR process, and MA process

Note: For the ARIMA process, determining the model is based on trial and errors. Adjustments in the AR and MA process need to be made until there is low or no autocorrelation is present in ACF and PACF of the residuals, then this is the final model.

Step 1: Run ARIMA modeling and forecasting task

The screenshot shows the SAS Enterprise Guide interface. On the left, the Project Tree displays a project named 'BIKE_SHARING_DATA_DAY' with a 'Program' sub-project. The main window shows a data table with columns: instant, dteday, season, yr, weekday, workingday, weathersit, and temp. The 'Analyze' menu is open, and the 'Time Series' option is selected, leading to a submenu where 'ARIMA Modeling and Forecasting...' is highlighted. Other options in the Time Series submenu include 'Prepare Time Series Data...', 'Basic Forecasting...', 'Regression Analysis with Autoregressive Errors...', 'Regression Analysis of Panel Data...', 'Create Time Series Data...', 'Forecast Studio Create Project...', 'Forecast Studio Open Project...', and 'Forecast Studio Override Project...'.

instant	dteday	season	yr	weekday	workingday	weathersit	temp
1	01JAN2011	1	0	6	0	2	0.344167
2	02JAN2011	1	0	0	0	2	0.363478
3	03JAN2011	1	0	1	1	1	0.196364
4	04JAN2011	1	0	2	1	1	0.2
5	05JAN2011	1	0	3	1	1	0.226957
6	06JAN2011	1	0	4	1	1	0.204348
7	07JAN2011	1	0	5	1	2	0.196522
8	08JAN2011	1	0	6	0	2	0.165
9	09JAN2011	1	0				0.138333
10	10JAN2011	1	0				0.150833
11	11JAN2011	1	0				0.169091
12	12JAN2011	1	0				0.172727
13	13JAN2011	1	0				0.165
14	14JAN2011	1	0				0.16087
15	15JAN2011	1	0				0.233333
16	16JAN2011	1	0				0.231667
17	17JAN2011	1	0				0.175833
18	18JAN2011	1	0				0.216667
19	19JAN2011	1	0				0.292174
20	20JAN2011	1	0				0.261667

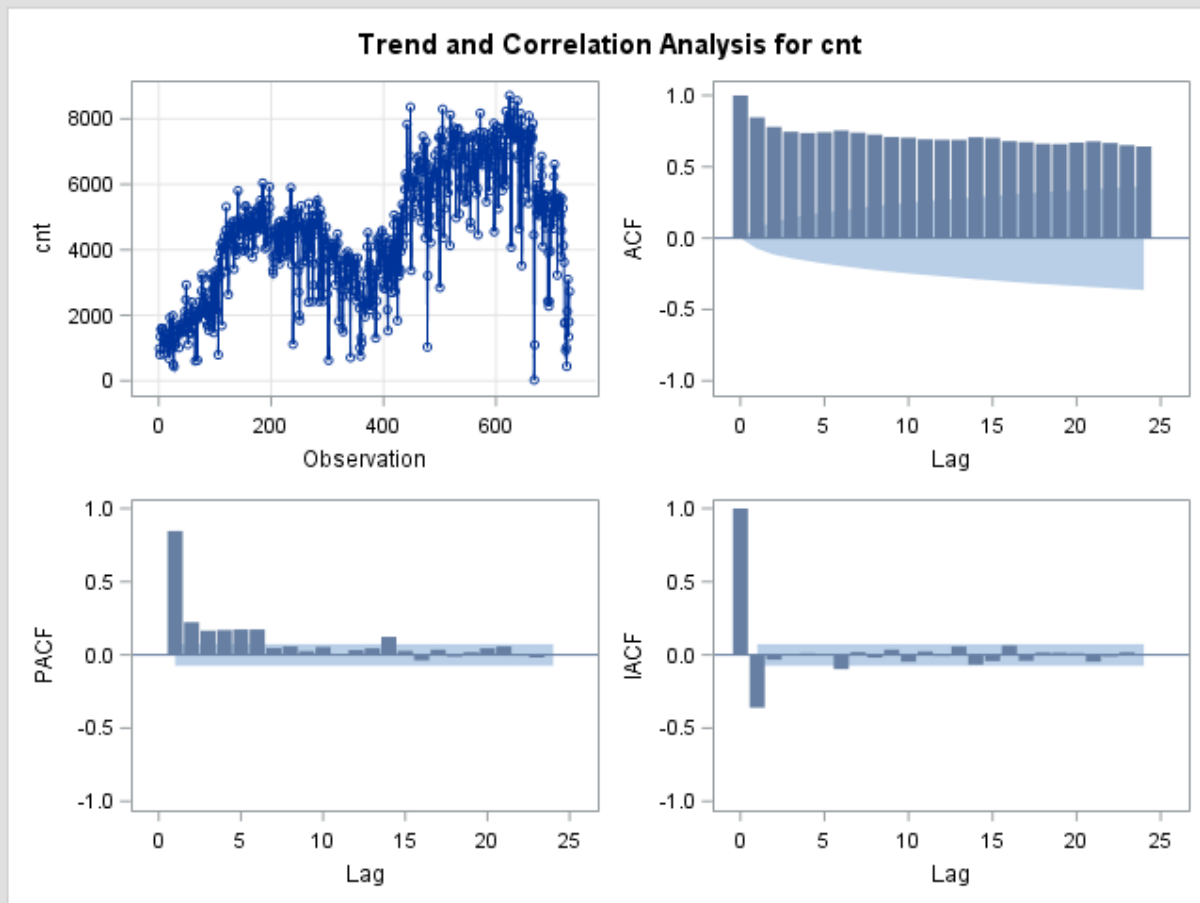


=> Click run.

Note: No other options (differencing, stationarity, etc.) were selected in this step.

Step 2: Examine results and autocorrelations

Autocorrelation Check for White Noise									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	2608.34	6	<.0001	0.846	0.779	0.746	0.737	0.742	0.755
12	4855.42	12	<.0001	0.739	0.726	0.709	0.704	0.692	0.689
18	6966.52	18	<.0001	0.689	0.707	0.702	0.680	0.673	0.661
24	8948.32	24	<.0001	0.659	0.670	0.678	0.667	0.651	0.642

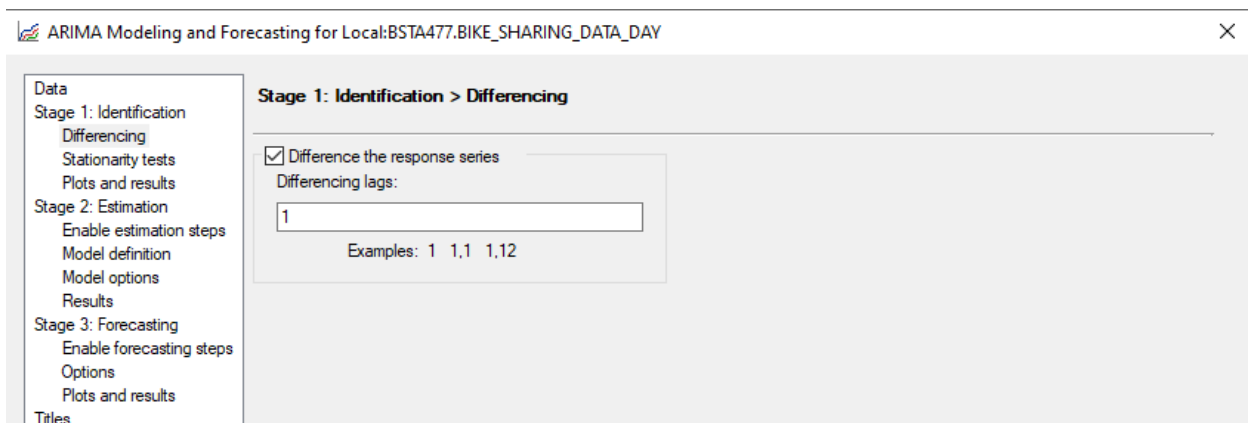
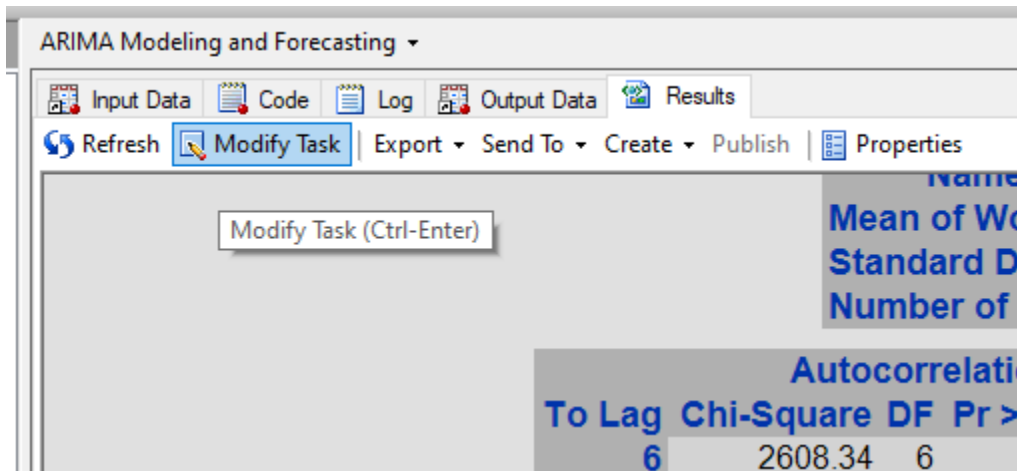


Based on the results above, we noted 3 things:

- From the Autocorrelation check for white noise table (Ljung box test - Tutorial 2): If our significant level is 0.05, then the p value < 0.05 for all the lags, this proves that the time series has significant autocorrelation. (We can also see this in the ACF plot).
- From the ACF plot:
 - The lags are high above the blue area (95% confidence interval) at lag 1 and slowly reducing. This indicates the trend element in the time series. => Differencing needed

- The lags are high above the blue area (95% confidence interval). This indicates high autocorrelation between observations. => AR process might be needed (have to difference the time series first before deciding).
- From the PACF plot: There are significant positive autocorrelations because the lags are above the blue area. => MA process might be needed.

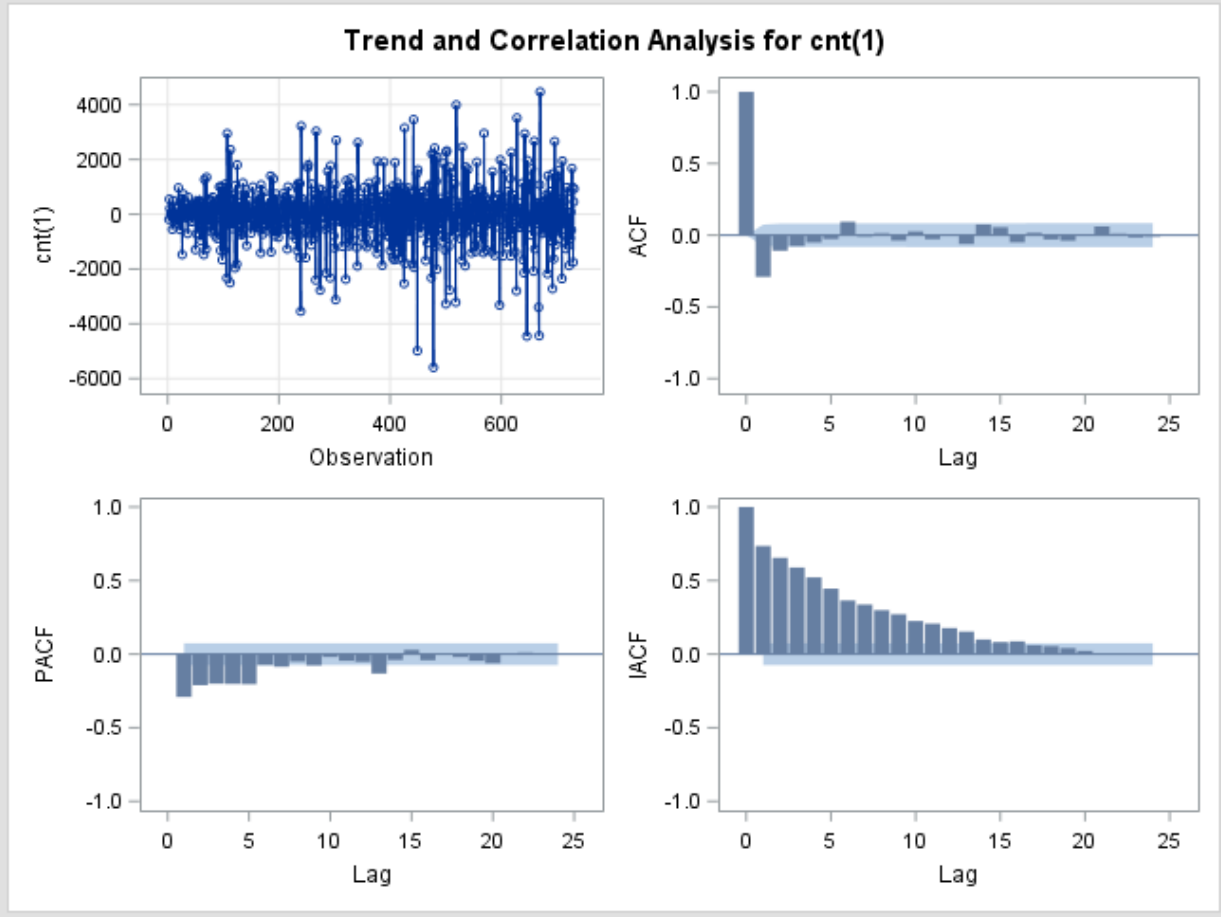
Step 3: Conduct differencing



=> Click run

Step 4: Determine AR process and MA process

Autocorrelation Check for White Noise									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	83.90	6	<.0001	-0.291	-0.109	-0.076	-0.051	-0.027	0.094
12	86.22	12	<.0001	-0.010	0.013	-0.037	0.026	-0.028	-0.004
18	98.28	18	<.0001	-0.062	0.075	0.055	-0.050	0.017	-0.030
24	102.56	24	<.0001	-0.039	0.005	0.061	0.011	-0.015	-0.010



Based on the results above, we try to determine AR and MA process:

- From the ACF plot, there are moderate autocorrelations. => AR(1) process
- From the PACF plot, there are moderate negative autocorrelations => MA(1) process.

=> The model might be ARIMA(1,1,1). Let's try to apply this on the training set and forecast the validation set.

ARIMA with SAS EG

Training set

Based on the above analysis, we difference the time series first then decide the AR and MA process again.

The screenshot displays the SAS Enterprise Guide interface. The top pane shows a data table named 'TRAINING_SET' with columns 'dteday' and 'cnt'. The 'Analyze' menu is open, and 'Time Series' is selected, leading to 'ARIMA Modeling and Forecasting...'. The bottom pane shows the 'ARIMA Modeling and Forecasting (2) for Local:BSTA477.TRAINING_SET' workflow.

TRAINING_SET Data:

	dteday	cnt
1	01JAN2011	985
2	02JAN2011	801
3	03JAN2011	1349
4	04JAN2011	1562
5	05JAN2011	1600
6	06JAN2011	1606
7	07JAN2011	1510
8	08JAN2011	959
9	09JAN2011	822
10	10JAN2011	1321
11	11JAN2011	1263
12	12JAN2011	1162
13	13JAN2011	1406
14	14JAN2011	1421
15	15JAN2011	1248
16	16JAN2011	1204
17	17JAN2011	1000
18	18JAN2011	683
19	19JAN2011	1650
20	20JAN2011	1927
21	21JAN2011	1543
22	22JAN2011	981

ARIMA Modeling and Forecasting (2) for Local:BSTA477.TRAINING_SET

Data

Data source: Local:BSTA477.TRAINING_SET
Task filter: None

Variables to assign:

Name
NewTimeID
dteday
cnt

Task roles:

Task roles
Time Series variable (Limit: 1)
Time ID variable (Limit: 1)
Group forecasts by

Data

Stage 1: Identification

Differencing

Stationarity tests

Plots and results

Stage 2: Estimation

Enable estimation steps

Model definition

Model options

Results

Stage 3: Forecasting

Enable forecasting steps

Options

Plots and results

Titles

Properties

Stage 1: Identification > Differencing

☒ Difference the response series

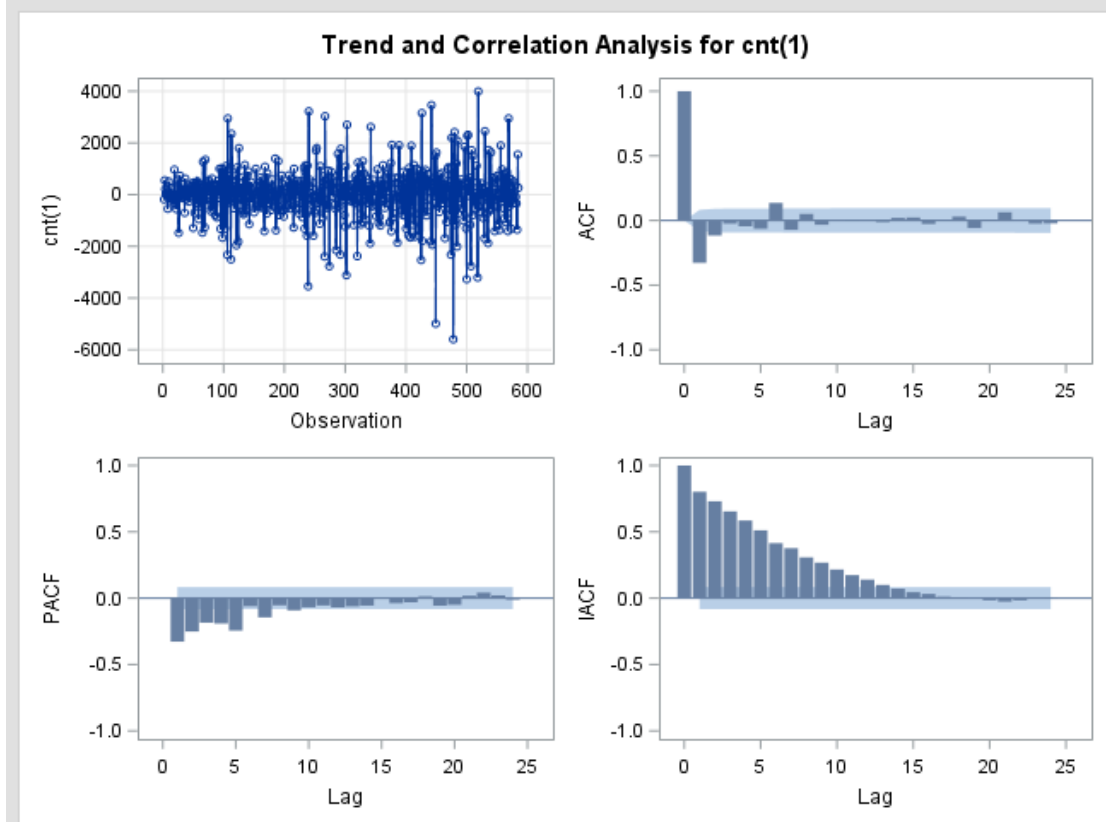
Differencing lags:

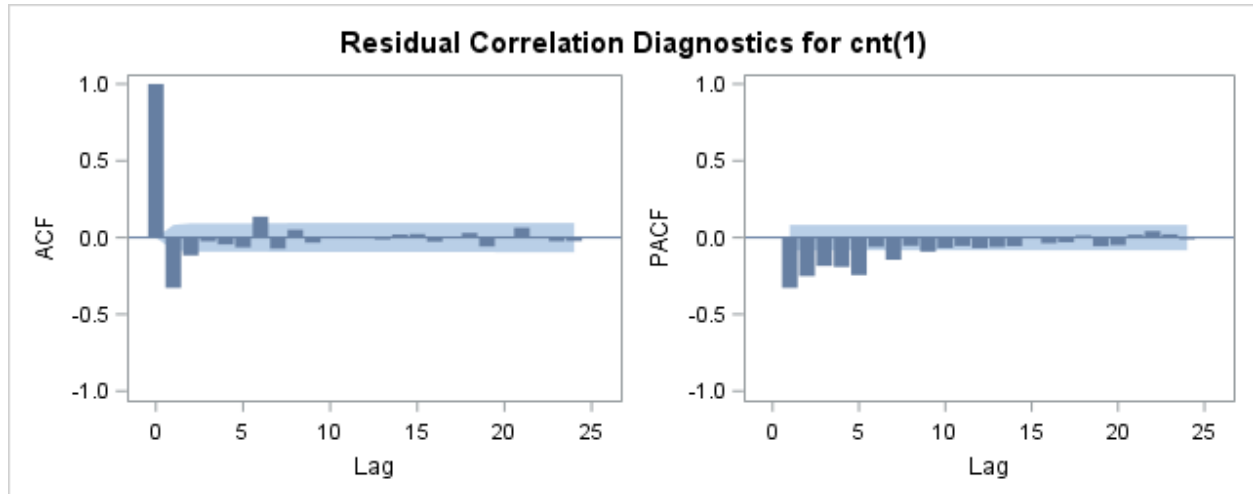
1

Examples: 1 1,1 1,12

=> Click Run.

Autocorrelation Check for White Noise									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	86.04	6	<.0001	-0.328	-0.117	-0.024	-0.045	-0.064	0.136
12	91.35	12	<.0001	-0.073	0.051	-0.033	0.000	-0.002	0.001
18	92.99	18	<.0001	-0.012	0.020	0.023	-0.027	-0.004	0.030
24	98.21	24	<.0001	-0.058	0.008	0.064	0.001	-0.024	-0.023





Based on the above results, we noted:

- From the ACF plot, there are moderate autocorrelations. => AR(1) might be needed.
- From the PACF plot, there are moderate negative autocorrelations. => MA(1) might be needed.

From the analysis, we try to adjust one by one. We apply AR(1) first.

ARIMA Modeling and Forecasting (2) ▾

Input Data | Code | Log | Results

Refresh | **Modify Task** | Export ▾ | Send To ▾ | Create ▾ | Publish | Properties

Modify Task (Ctrl-Enter)

Observation(s) eliminated by differencing

Autocorrelation Check for White Noise

To Lag	Chi-Square	DF	Pr > ChiSq	Autocorre
6	86.04	6	<.0001	-0.328 -0.117 -0.024
12	91.35	12	<.0001	-0.073 0.051 -0.033
18	92.99	18	<.0001	-0.012 0.020 0.023
24	98.24	24	<.0001	0.058 0.008 0.064

ARIMA Modeling and Forecasting (2) for Local:BSTA477.TRAINING_SET

Data

Stage 1: Identification

Differencing

Stationarity tests

Plots and results

Stage 2: Estimation

Enable estimation steps

Model definition

Model options

Results

Stage 3: Forecasting

Enable forecasting steps

Options

Plots and results

Titles

Properties

Stage 2: Estimation > Enable estimation steps

☒ Perform estimation steps

NOTE: The ARIMA Modeling and Forecasting task is organized into three stages.

- The Stage 1: Identity steps are always enabled.
- To enable the Stage 2: Estimation steps of this task, please check the check box shown above.
- The Estimation steps also must be enabled before the Stage 3: Forecasting steps can be enabled.

ARIMA Modeling and Forecasting (2) for Local:BSTA477.TRAINING_SET

Stage 2: Estimation > Model definition

Model definition

Autoregressive (p=)

Enter the lags for the Autoregressive Parameters to be estimated for each multiplicative factor in the list.

Ex. 1 1...4 1,3,12 1 TO 4

Factors for AR model:

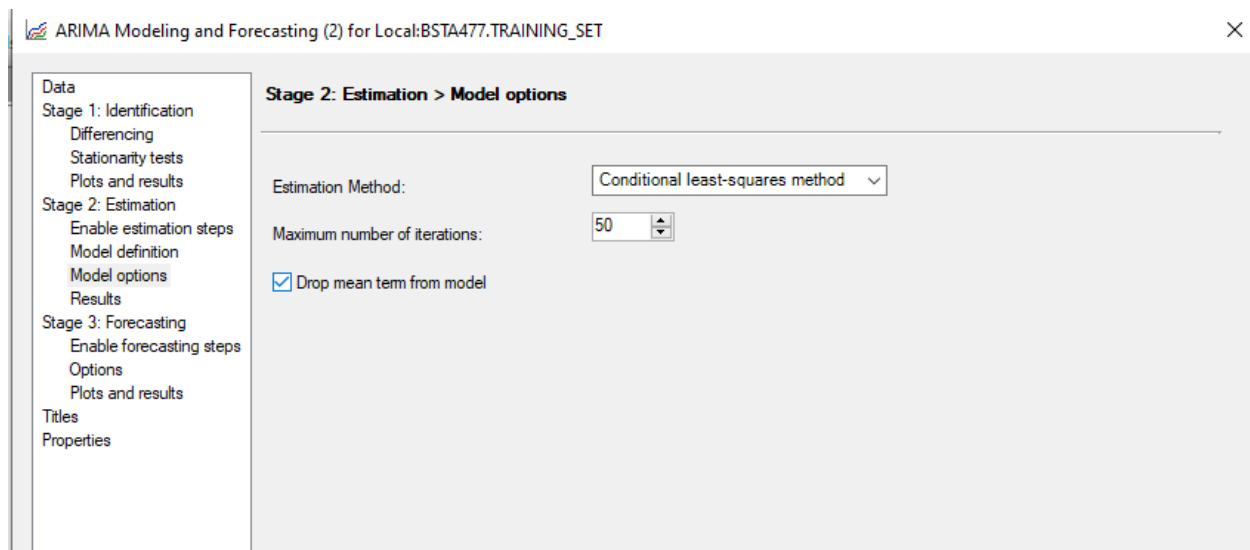
1

Moving Average (q=)

Enter the lags for the Moving Average Parameters to be estimated for each multiplicative factor in the list.

Ex. 1 1...4 1,3,12 1 TO 4

Factors for MA model:



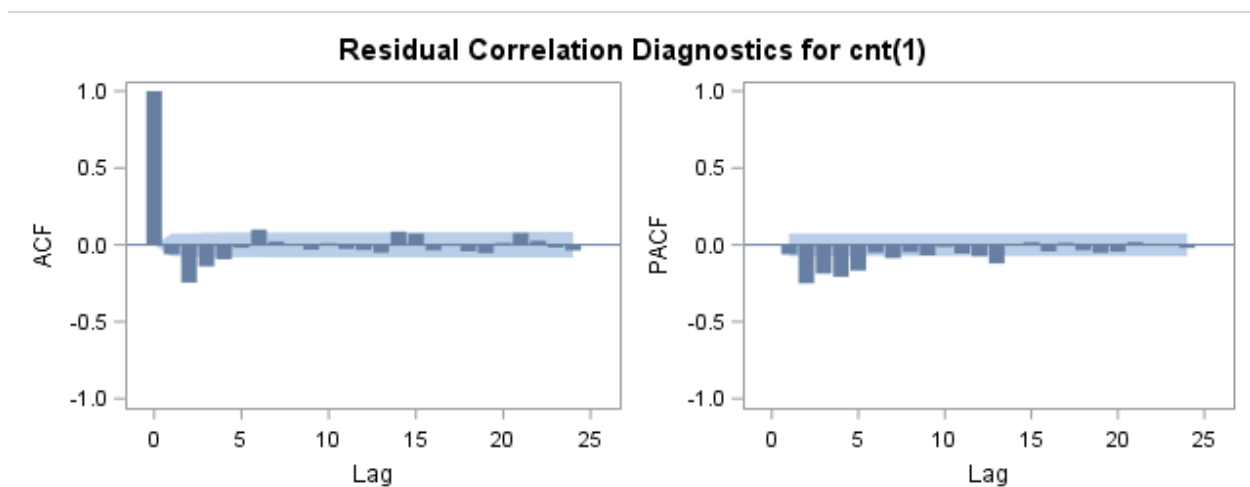
Note: We click Drop mean from the model only if differencing was applied to our time series. Otherwise, it's not necessary.

=> Click Run.

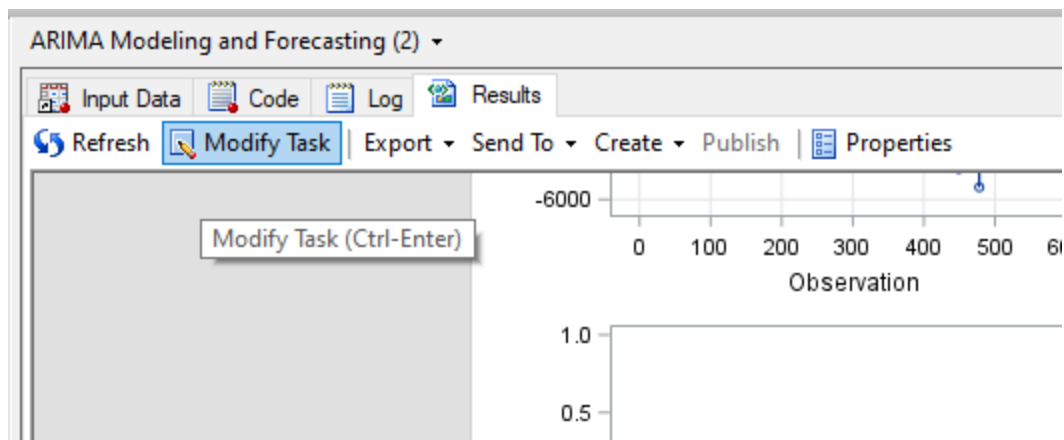
Conditional Least Squares Estimation					
Parameter	Estimate	Standard Error	t Value	Approx Pr > t	Lag
AR1,1	-0.32781	0.03913	-8.38	<.0001	1

Here we see that the p value < 0.0001. If our significant level is 0.05. Then, the AR coefficient of our model is significant.

Result:



We apply MA(1) now.



ARIMA Modeling and Forecasting (2) for Local:BSTA477.TRAINING_SET

Stage 2: Estimation > Model definition

Model definition

Autoregressive (p=)
Enter the lags for the Autoregressive Parameters to be estimated for each multiplicative factor in the list.
Ex. 1 1...4 1,3,12 1 TO 4
Factors for AR model:
1
Add

Moving Average (q=)
Enter the lags for the Moving Average Parameters to be estimated for each multiplicative factor in the list.
Ex. 1 1...4 1,3,12 1 TO 4
Factors for MA model:
1
Add

ARIMA Modeling and Forecasting (2) for Local:BSTA477.TRAINING_SET

Stage 2: Estimation > Model options

Estimation Method: Conditional least-squares method ▾

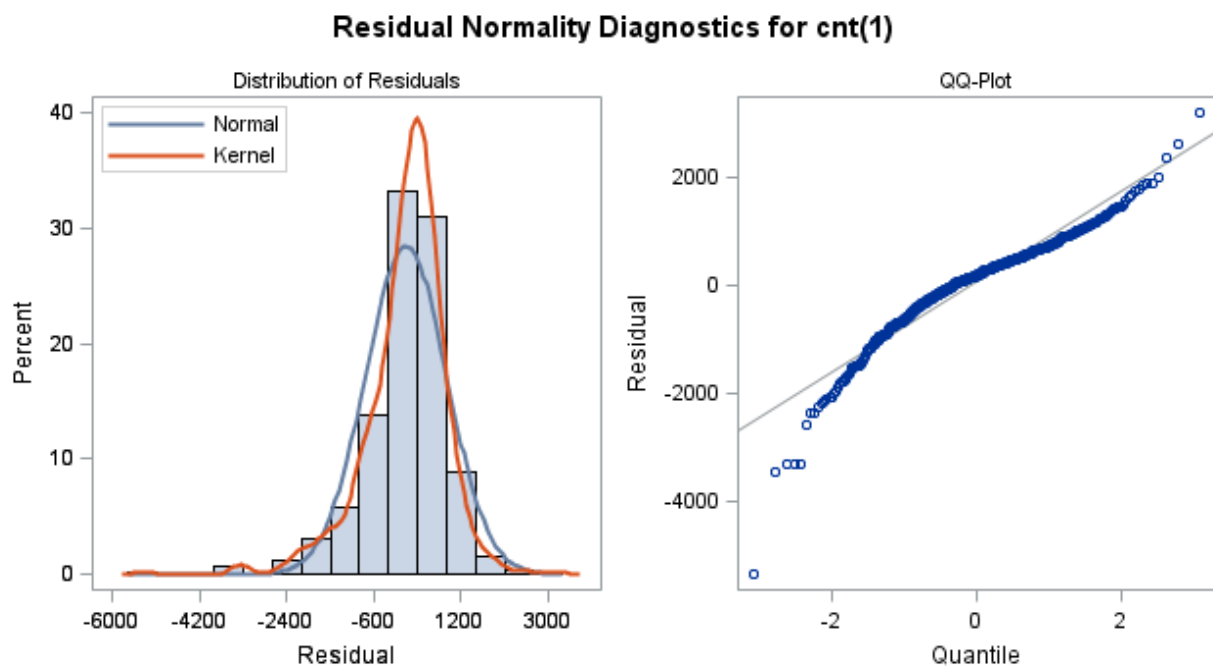
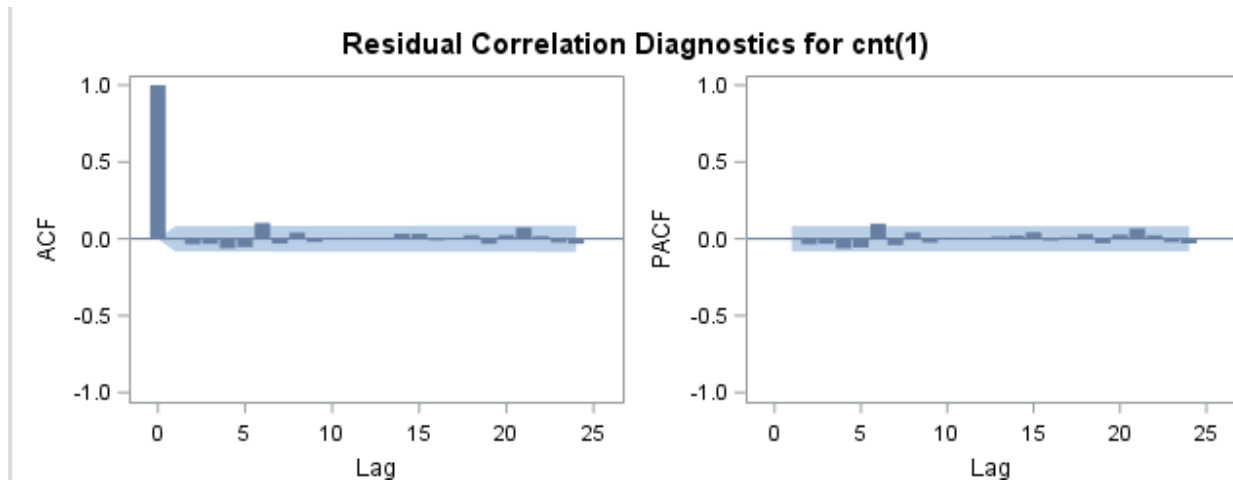
Maximum number of iterations: 50

☒ Drop mean term from model

=> Click Run.

Result:

Conditional Least Squares Estimation					
Parameter	Estimate	Standard Error	t Value	Approx Pr > t	Lag
MA1,1	0.88923	0.02321	38.31	<.0001	1
AR1,1	0.27884	0.04896	5.70	<.0001	1



To evaluate the model:

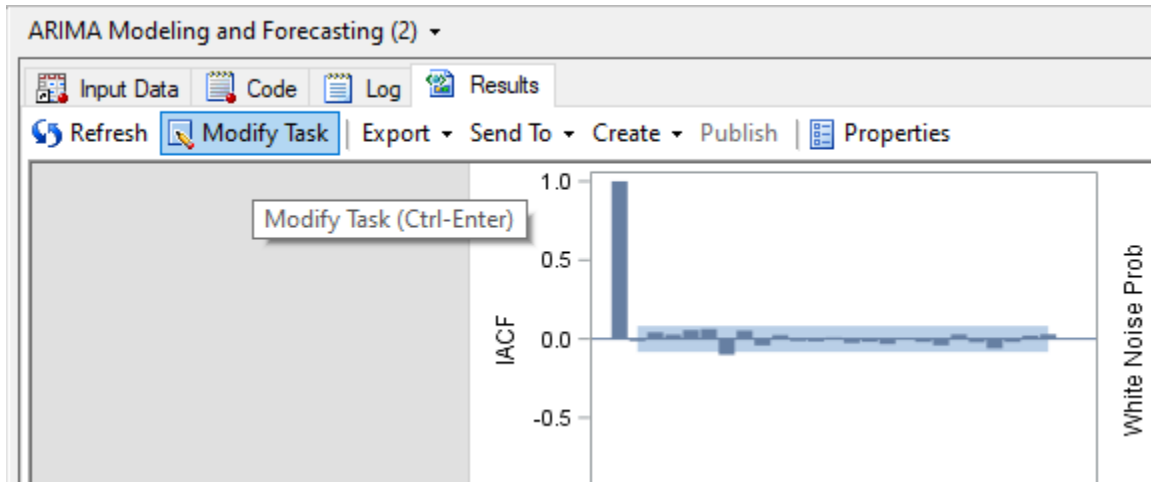
- Evaluate AR and MA process: Both AR coefficient and MA coefficient are significant.
- Evaluate the residual Correlation Diagnostics: There seems to be no strong autocorrelation

- The residuals seem to be normally distributed.

=> The model is adequate. Let's forecast observations for the validation set.

Forecast and validation

Forecast observations for validation set.



The screenshot shows the 'ARIMA Modeling and Forecasting (2) for Local:BSTA477.TRAINING_SET' window. The left sidebar lists the task stages: Data, Stage 1: Identification, Stage 2: Estimation, Stage 3: Forecasting, Titles, and Properties. The 'Stage 3: Forecasting' section is expanded, showing 'Enable forecasting steps' as the selected option. The main area displays the 'Stage 3: Forecasting > Enable forecasting steps' configuration page. It includes a checkbox labeled 'Perform forecasting steps' which is checked. Below this, a note states: 'NOTE: The ARIMA Modeling and Forecasting task is organized into three stages.' followed by three bullet points: '- The Stage 1: Identity steps are always enabled.', '- To enable the Stage 3: Forecasting steps of this task, please check the check box shown above.', and '- The Stage 2: Estimation steps must be enabled on the Enable estimation steps page before the Stage 3: Forecasting steps can be enabled.'

ARIMA Modeling and Forecasting (2) for Local:BSTA477.TRAINING_SET

Stage 3: Forecasting > Options

Time interval between observations

Daily

Time units per interval: 1

Number of intervals to forecast: 147

Confidence level: 95%

Left sidebar:

- Data
- Stage 1: Identification
 - Differencing
 - Stationarity tests
 - Plots and results
- Stage 2: Estimation
 - Enable estimation steps
 - Model definition
 - Model options
 - Results
- Stage 3: Forecasting
 - Enable forecasting steps
 - Options
 - Plots and results
- Titles
- Properties

ARIMA Modeling and Forecasting (2) for Local:BSTA477.TRAINING_SET

Stage 3: Forecasting > Plots and results

Forecasting plots options

☐ Forecasts

☐ Residuals

☐ Limit intervals of actual data shown

Intervals to show: 24

☒ Save forecasts

Local:WORK.ARIMA_FORE Browse...

☐ Suppress displayed forecasting output

Left sidebar:

- Data
- Stage 1: Identification
 - Differencing
 - Stationarity tests
 - Plots and results
- Stage 2: Estimation
 - Enable estimation steps
 - Model definition
 - Model options
 - Results
- Stage 3: Forecasting
 - Enable forecasting steps
 - Options
 - Plots and results
- Titles
- Properties

Note: Remember to save the forecast into a permanent SAS library as we would need this to calculate error terms for the training set and validation set.

=> Click Run. (The forecast output here is work.arma_forecast)

Error terms

Training set

```
data arima_training_error;
  set work.arima_forecast;
  where dteday < '08AUG2012'd;
  abs = abs(residual);
  square = residual**2;
  proportion = residual/cnt;
  abs_proportion = abs/cnt;
run;
```

```
proc means data=work.arima_training_error;
  var abs square proportion abs_proportion;
  output out=work.arima_training_eva
  mean(abs) = MAE
  mean(square) = MSE
  mean(proportion) = MPE
  mean(abs_proportion) = MAPE;
run;
```

Note:

- The “where dteday < ‘08AUG2012’d” statement is to filter out the observations of the training set.

Validation set

```
data arima_validation_forecast;
  set work.arima_forecast;
  where dteday >= '08AUG2012'd;
  rename cnt=empty_col;
run;
```

```
data arima_validation_error;
  merge work.arima_validation_forecast bsta477.validation_set;
  by dteday;
  residual = forecast - cnt;
  abs = abs(residual);
  square = residual**2;
  proportion = residual/cnt;
  abs_proportion = abs/cnt;
run;
```

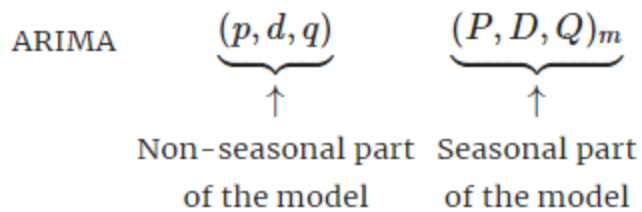
```
proc means data=work.arima_validation_error;
  var abs square proportion abs_proportion;
  output out=work.arima_validation_eva
  mean(abs) = MAE
  mean(square) = MSE
  mean(proportion) = MPE
  mean(abs_proportion) = MAPE;
run;
```

Note:

- As there are no actual observations in the work.arima_forecast data set, we have to filter out the validation forecasts and merge the new data set to validation set to calculate residuals.
 - Filter out validation forecast: where dteday >= "08AUG2012"d
 - Rename cnt because this column in the arima_forecast for validation set is empty.
 - Merge forecast with validation set: merge work.arima_validation_forecast bsta477.validation_set
- Be careful when using proc means. Make sure to name the output dataset (output out = new_data_set_name) is different from the input dataset. If the same name is used in the output dataset, the results are null.

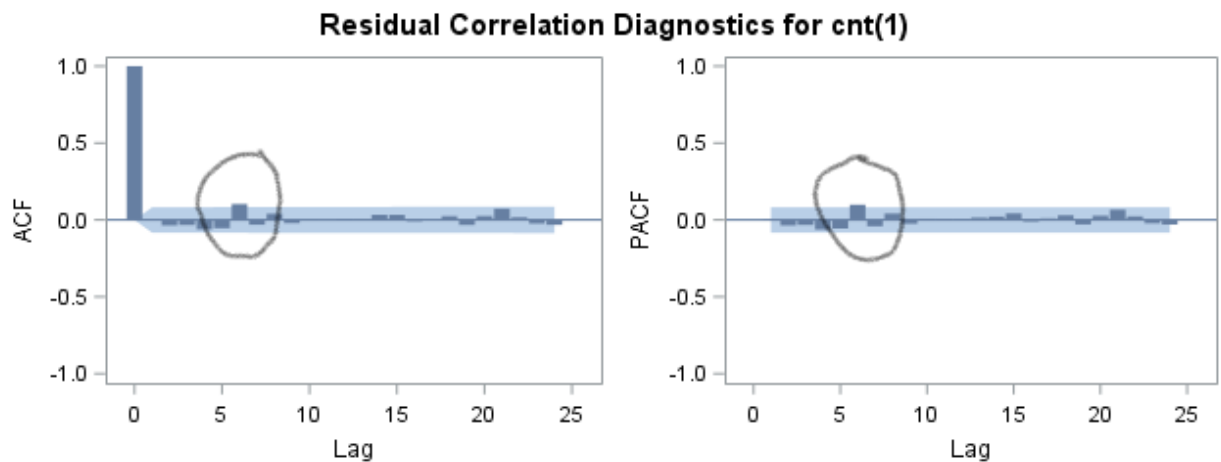
ARIMA with seasonality

To handle ARIMA with seasonality, we need to add the seasonal component part of the model.



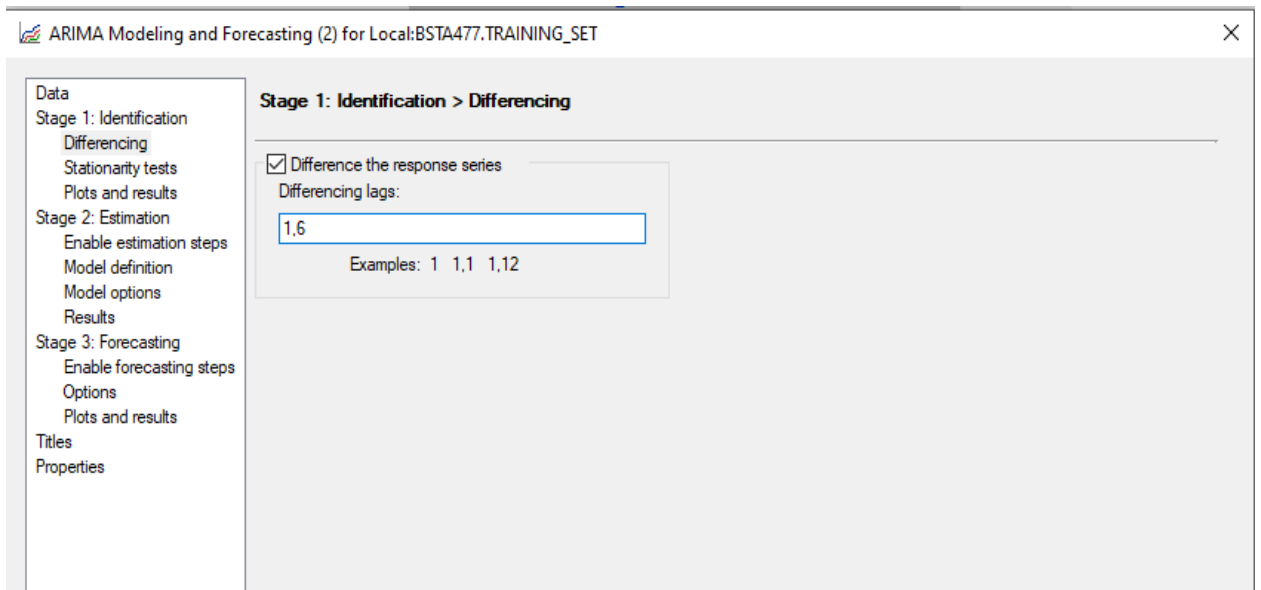
Based on the above analysis of the current dataset, we currently use the ARIMA(1,1,1) model.

However, for example, if we still see significant lag at lag 6 on both ACF and PACF, imagine lag 6 is above the blue area for the ACF and PACF below:



Then, we want to add seasonality factor, the following steps can be done to add seasonality at lag 6 in ARIMA model:

- **Adjust Differencing:** 1,6 => Model: ARIMA(1,1,1) (0,1,0)₆



Then click Run. If there are still significant lags in ACF or PACF at seasonality lags then the below steps should be added.

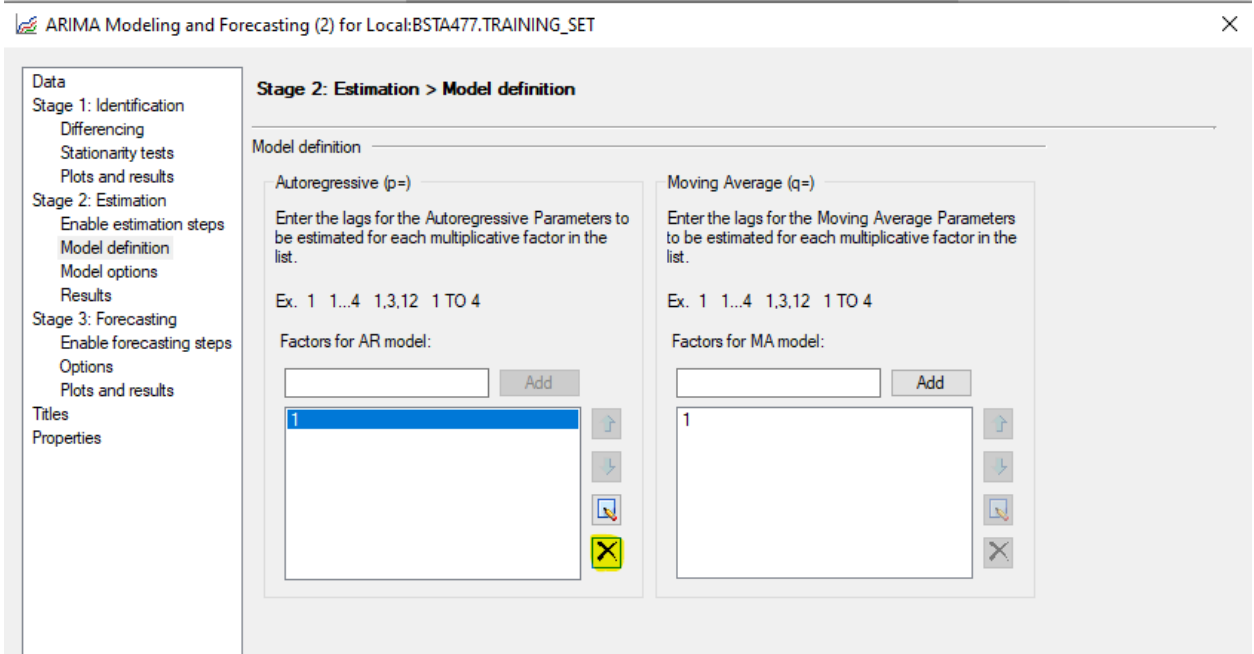
Depending on the significant lags:

- If there are only significant lags in the ACF plot then add AR process. Then check the plot again.
- If there are only significant lags in PACF plot then add the MA process. Then check the plot again.
- If there are significant lags in both ACF and PACF then add AR process first, then check the plots and add MA process.

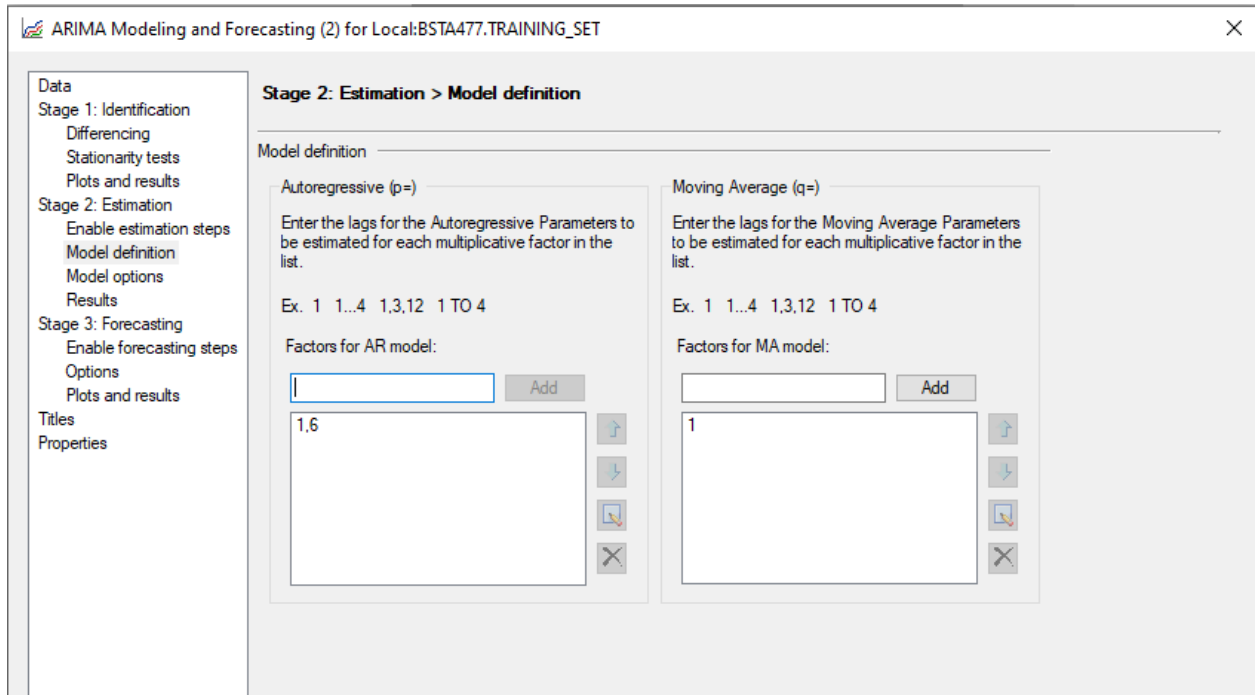
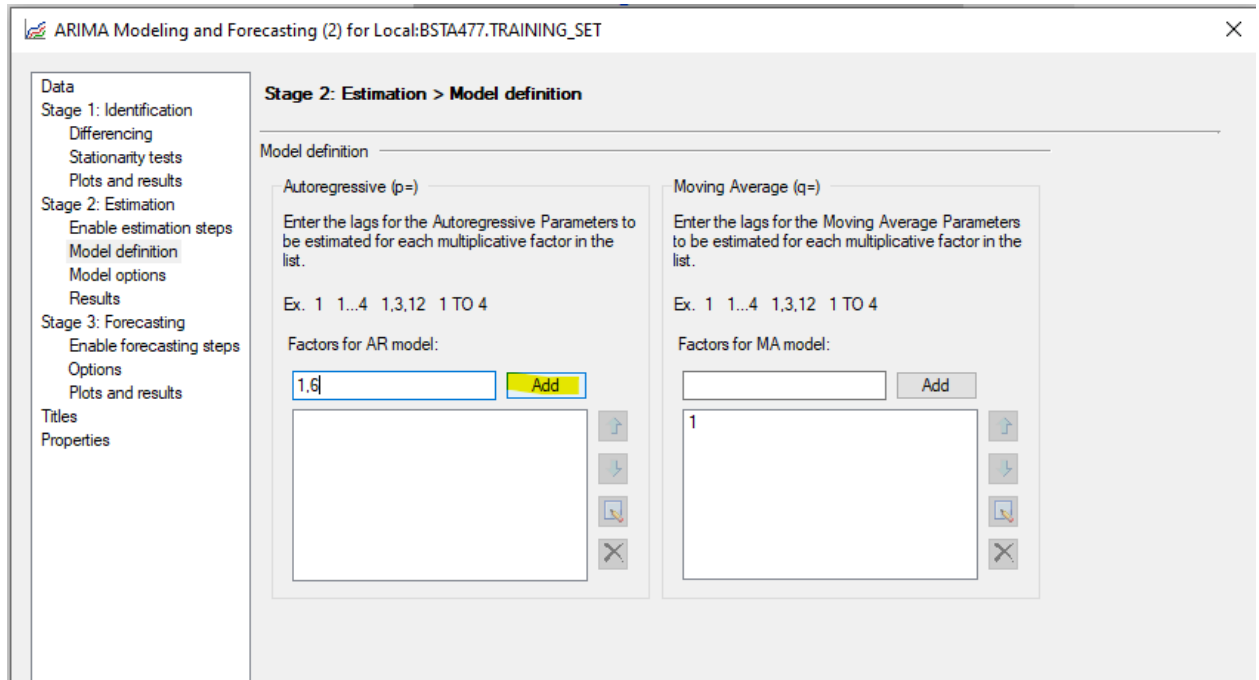
- **Add AR process:** 1,6 => Model: ARIMA (1,1,1) (1,1,0)₆ (if MA process is not adjusted)

If there are still significant lags in the ACF. Then add an AR process into the seasonality.

1. Omit the previous AR process:



2. Add the new AR process



Click Run and check outputs. Pay attention to the ACF and PACF of the residuals. If there are still significant lags in PACF then add the MA process step below.

- Adjust MA process: 1,6 => Model: ARIMA (1,1,1) (0,1,1)6 (if AR process is not adjusted)

If there are only significant lags in PACF then only add an MA process using the following steps:

1. Omit the previous MA process:

ARIMA Modeling and Forecasting (2) for Local:BSTA477.TRAINING_SET

Stage 2: Estimation > Model definition

Model definition

Autoregressive (p=)
Enter the lags for the Autoregressive Parameters to be estimated for each multiplicative factor in the list.
Ex. 1 1...4 1,3,12 1 TO 4
Factors for AR model:

1

Moving Average (q=)
Enter the lags for the Moving Average Parameters to be estimated for each multiplicative factor in the list.
Ex. 1 1...4 1,3,12 1 TO 4
Factors for MA model:

1

2. Add MA process:

ARIMA Modeling and Forecasting (2) for Local:BSTA477.TRAINING_SET

Stage 2: Estimation > Model definition

Model definition

Autoregressive (p=)
Enter the lags for the Autoregressive Parameters to be estimated for each multiplicative factor in the list.
Ex. 1 1...4 1,3,12 1 TO 4
Factors for AR model:

1

Moving Average (q=)
Enter the lags for the Moving Average Parameters to be estimated for each multiplicative factor in the list.
Ex. 1 1...4 1,3,12 1 TO 4
Factors for MA model:

Then click Run and check the outputs.

- Adjust both AR and MA => Model: ARIMA (1,1,1) (1,1,1)₆

We adjust the model until the ACF and PACF of the residuals do not exhibit seasonality factors and the residuals are white noise.