Alicia N Phan

BSTA 477/677 – Winter 2021

Tutorial 5 - March 28th, 2021

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.

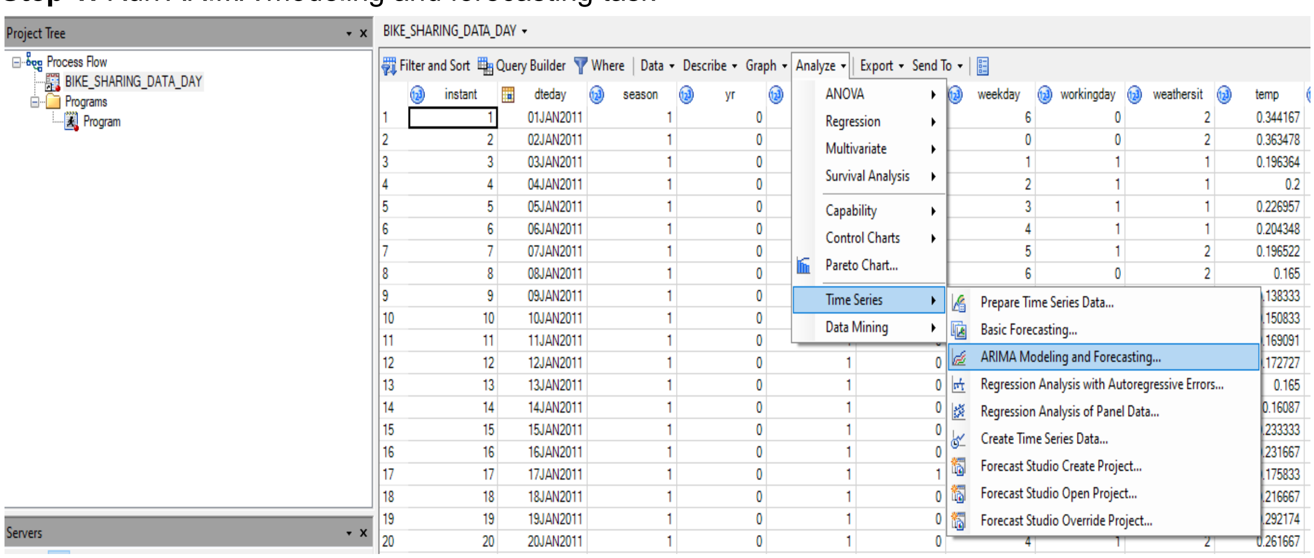
Choose ARIMA models: ACF, PACF

The following are the general steps to select ARIMA models:

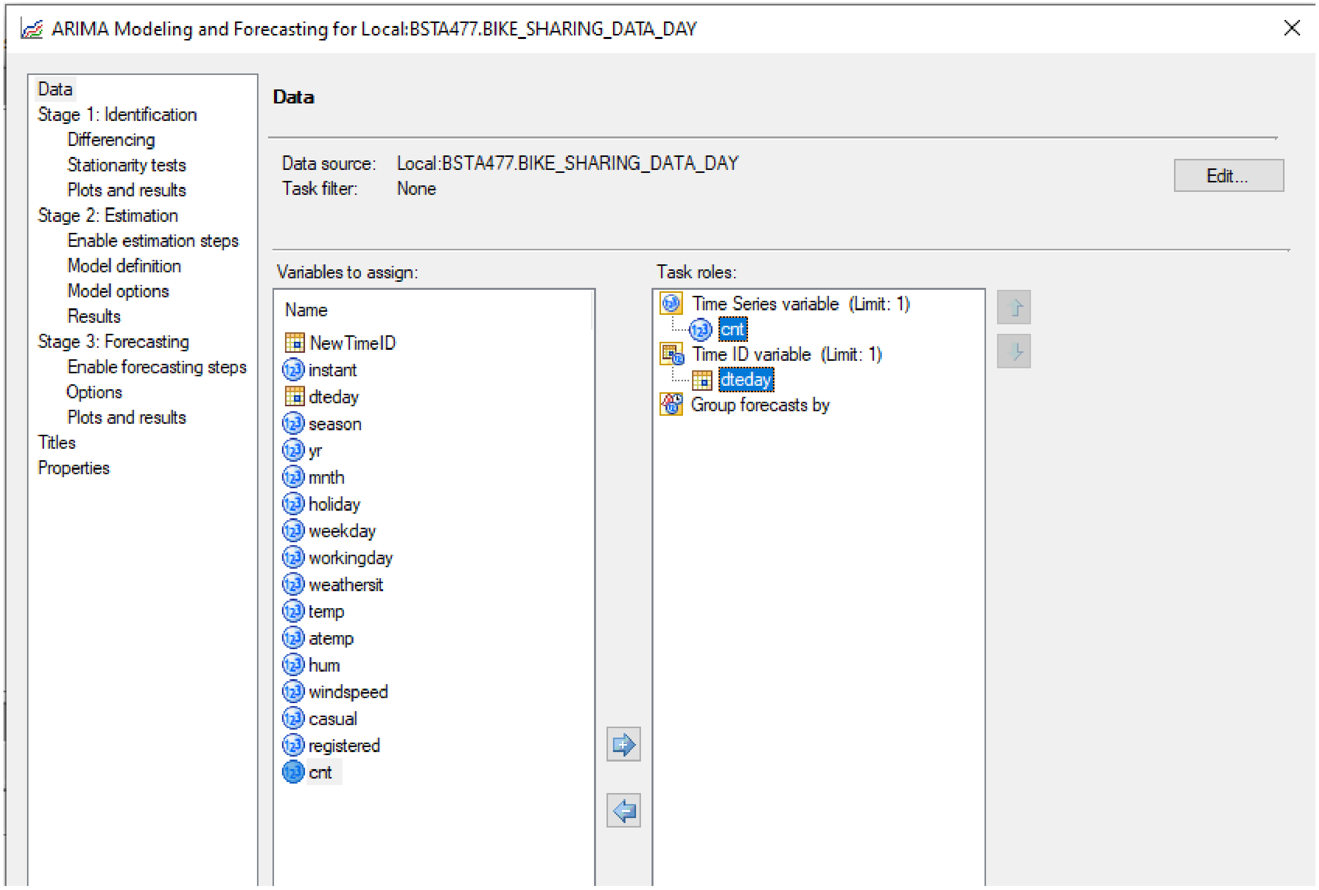
1. Run the ARIMA modeling and forecasting task on the full dataset.
2. Examine autocorrelations: ACF and PACF plots
3. Conduct differencing (if trend component is present)
4. Determine AR process, and MA process

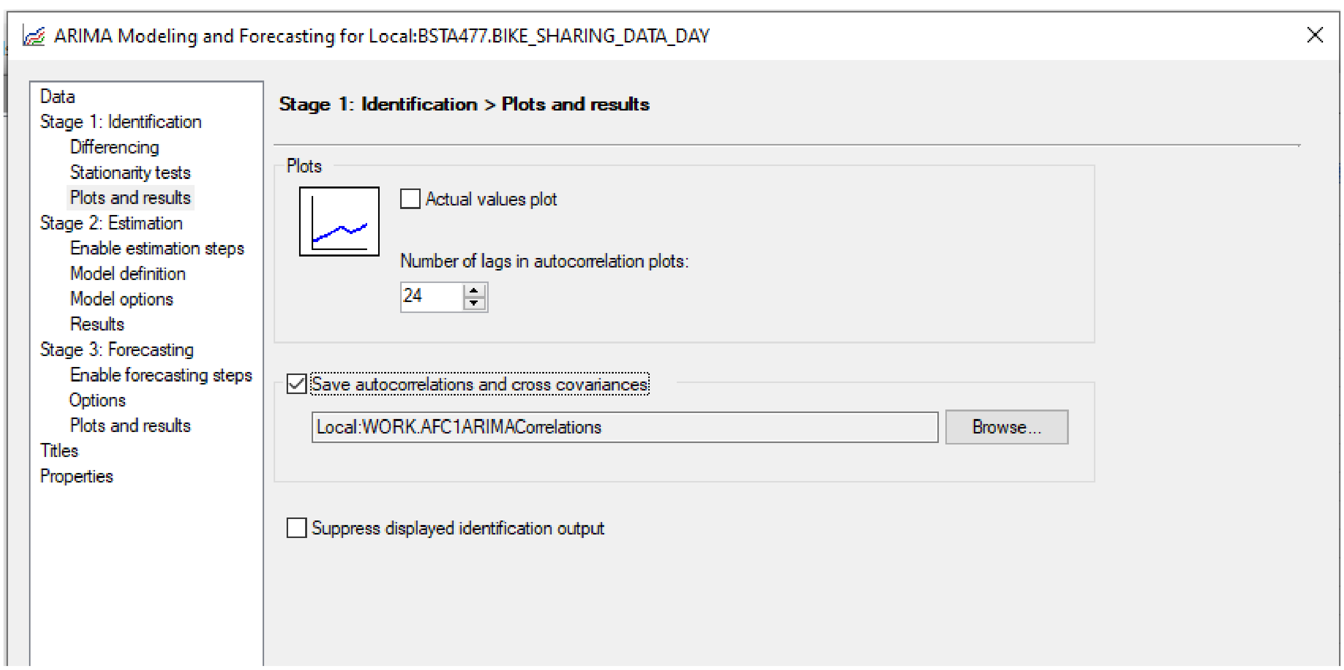
Note: For the ARIMA process, determining the model is based on trial and errors. Adjustments in the AR and MA process need to be made until there is low or no autocorrelation is present in ACF and PACF of the residuals, then this is the final model.

Step 1: Run ARIMA modeling and forecasting task



Alicia N Phan





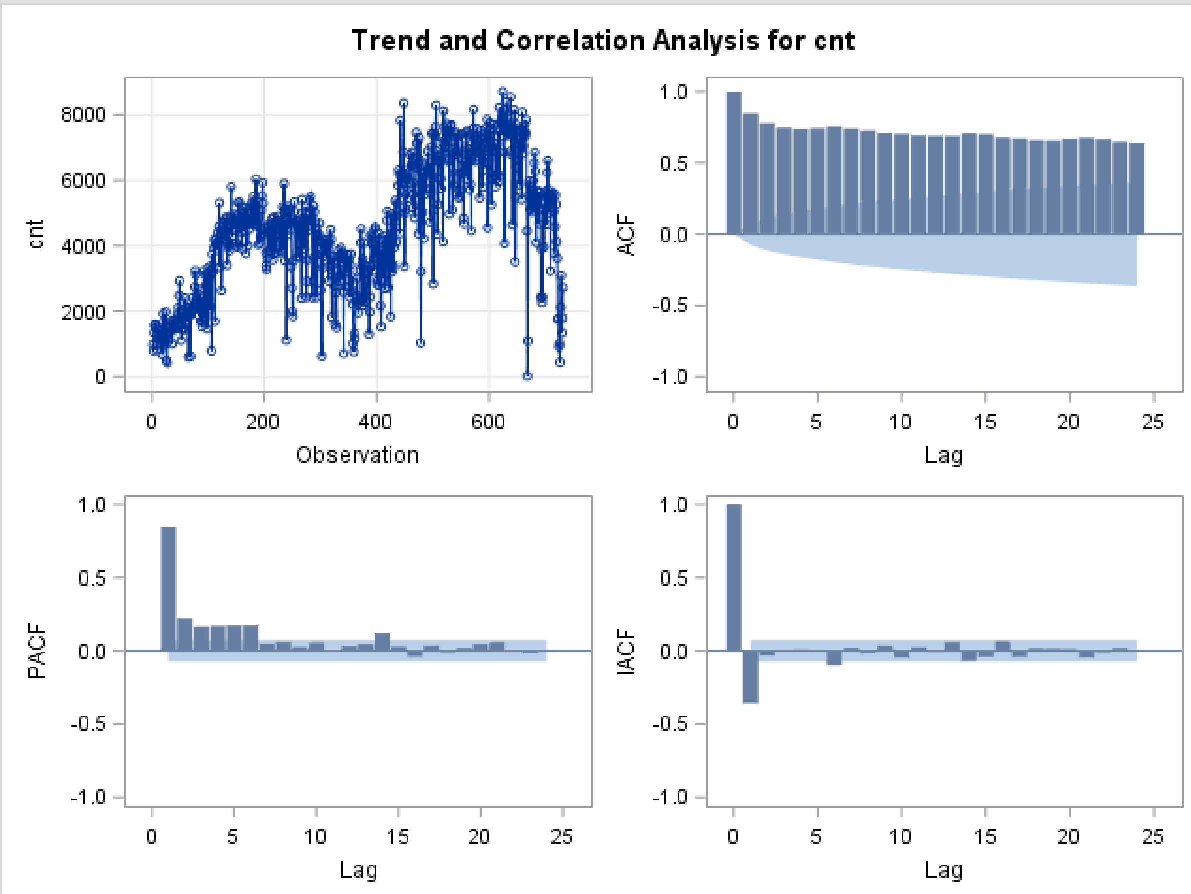
=> Click run.

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

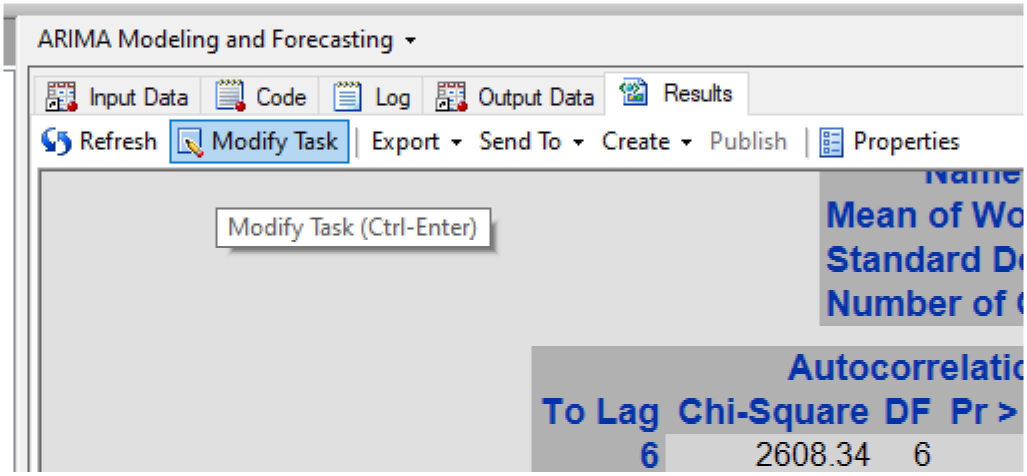


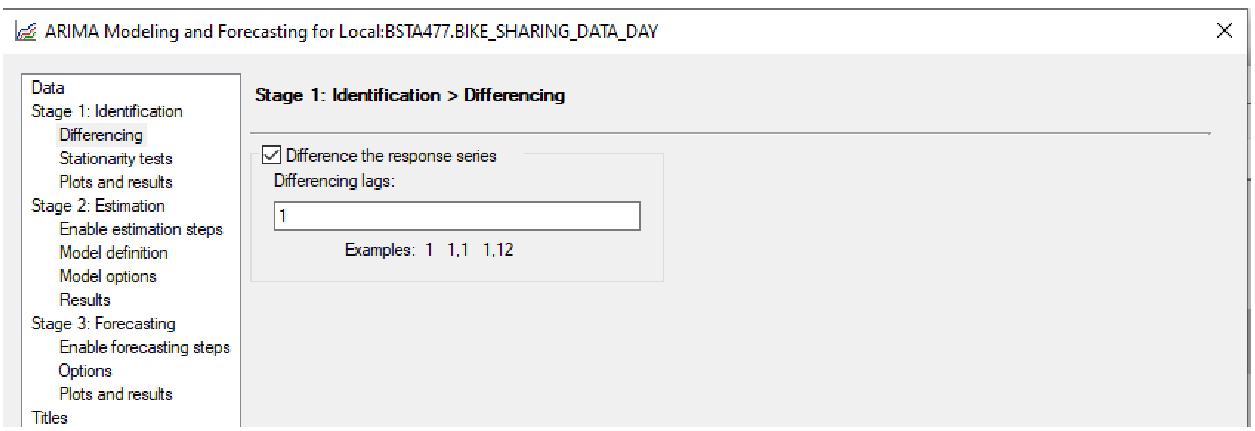
Based on the results above, we noted 3 things:

- 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





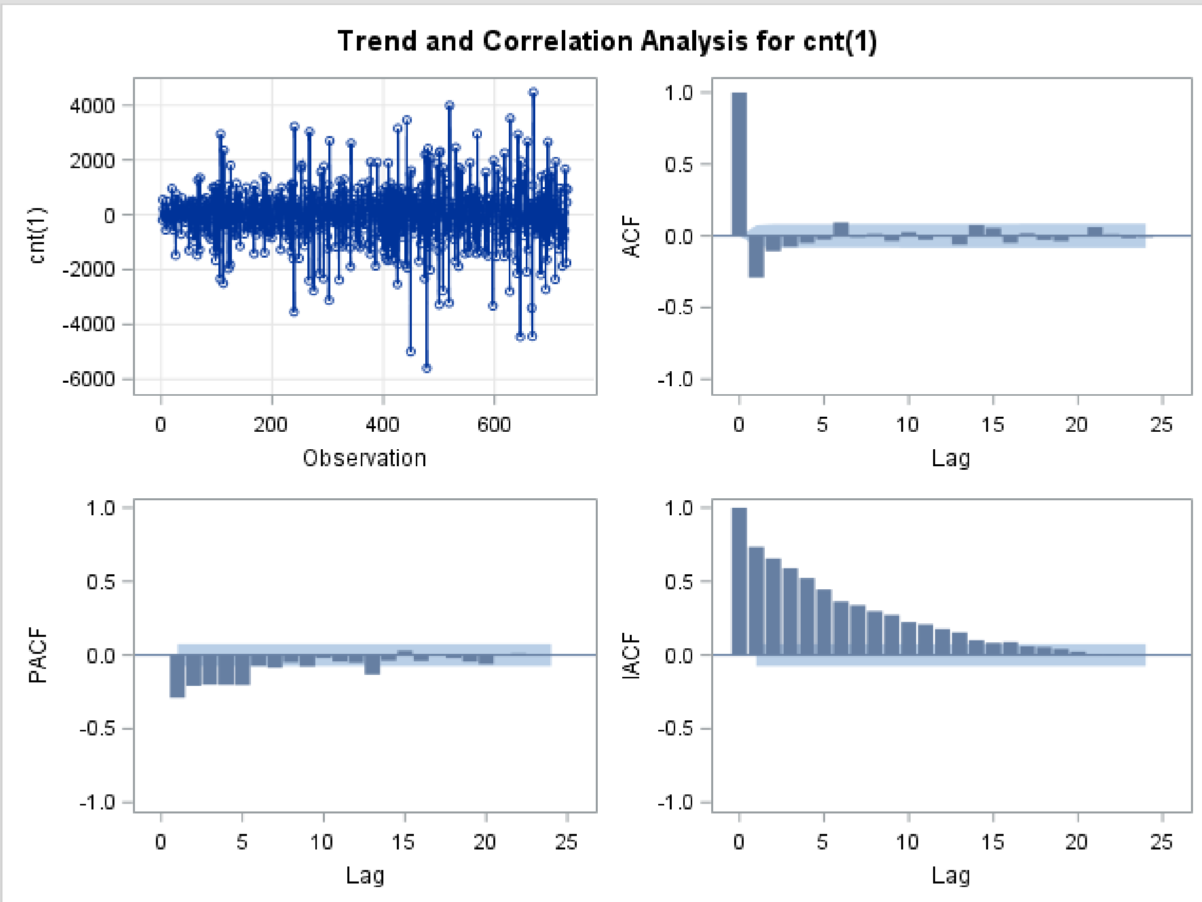
=> Click run

Step 4: Determine AR process and MA process

Alicia N Phan

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.

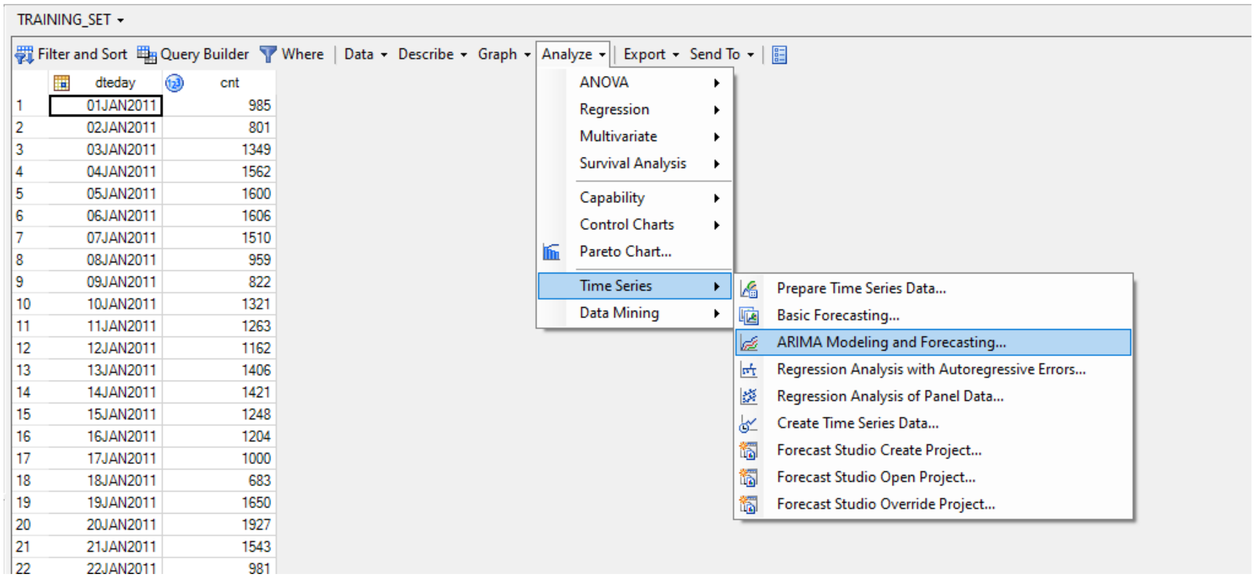
=> The model might be ARIMA(1,1,1). Let’s try to apply this on the training set and forecast the validation set.

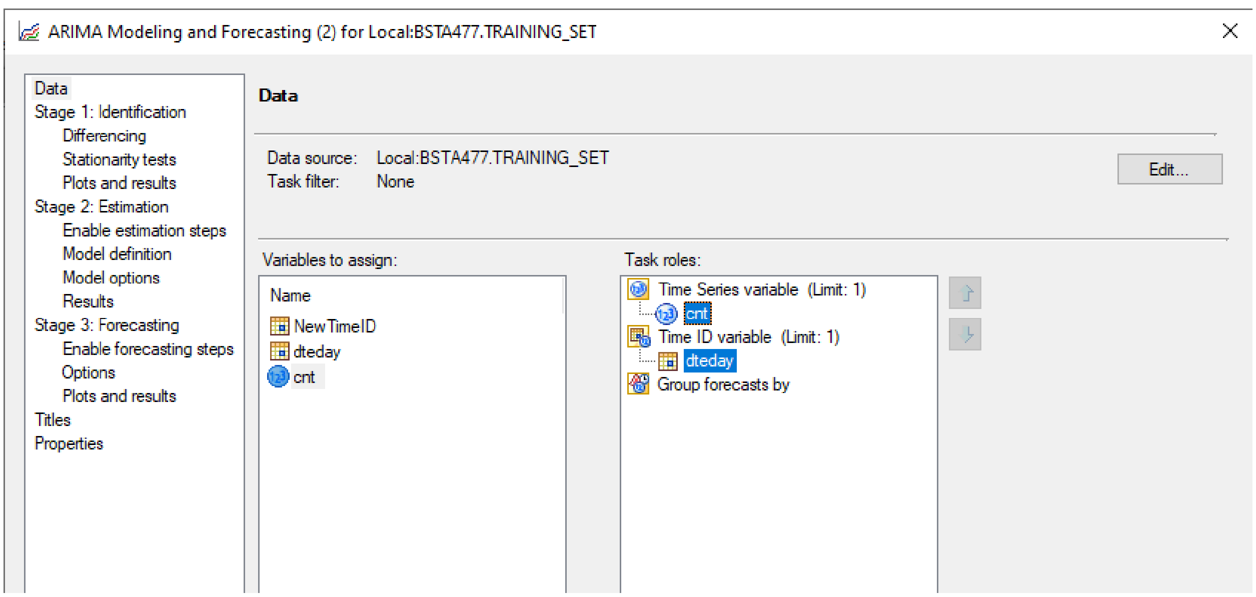
Alicia N Phan

ARIMA with SAS EG

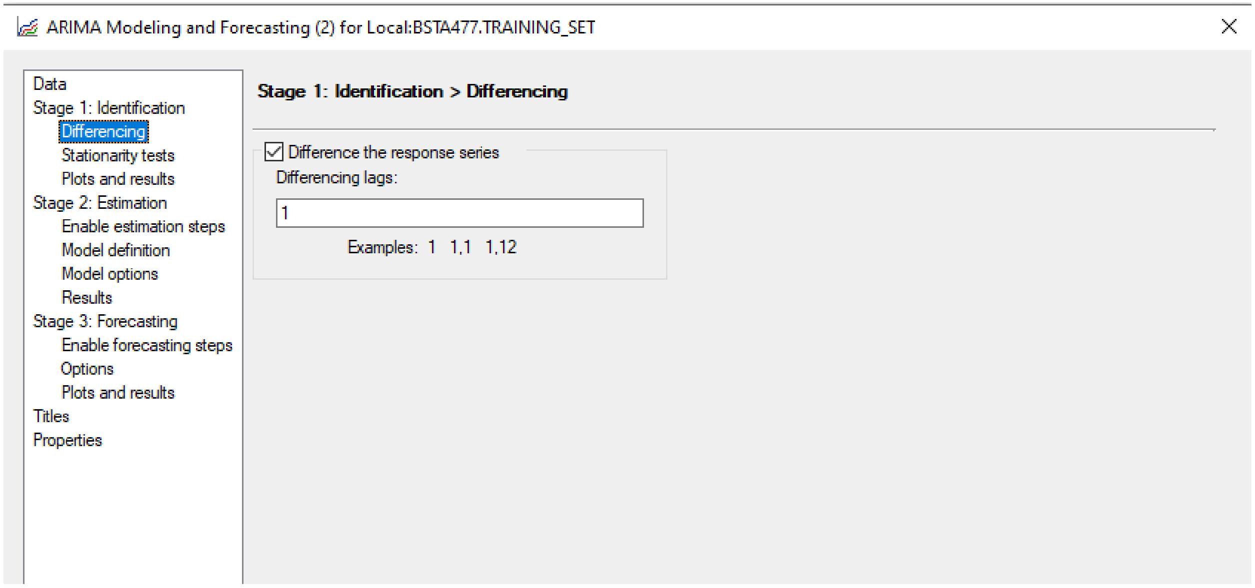
Training set

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





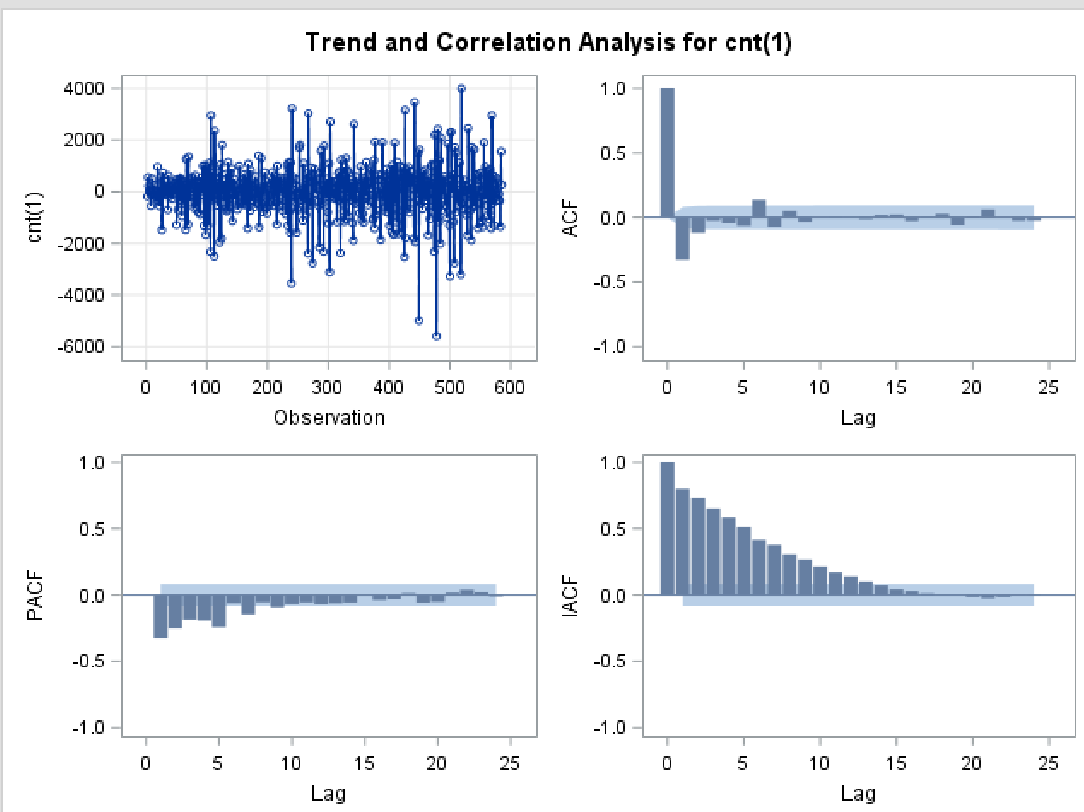
Alicia N Phan



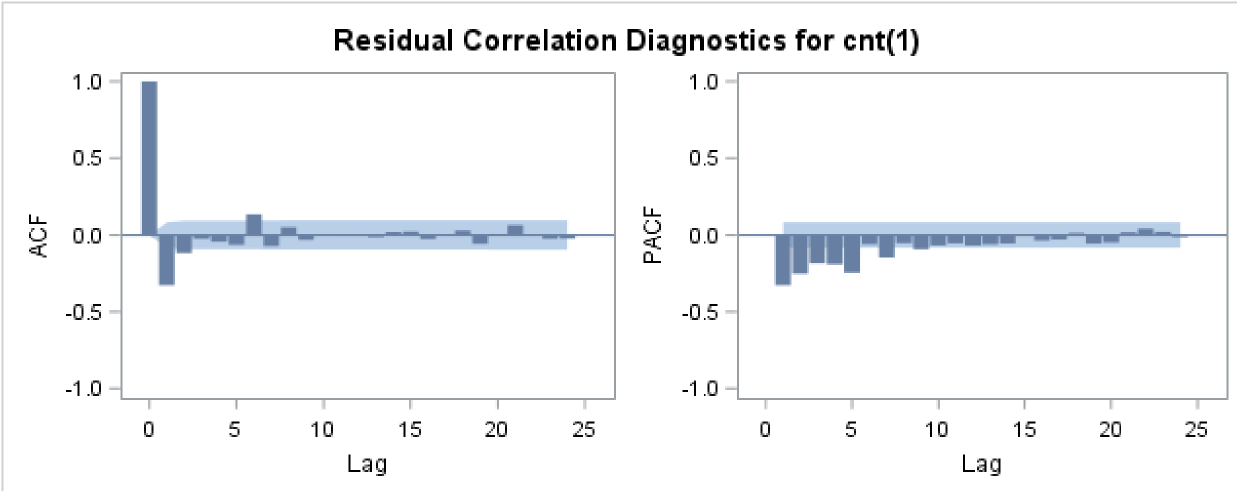
=> Click Run.

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



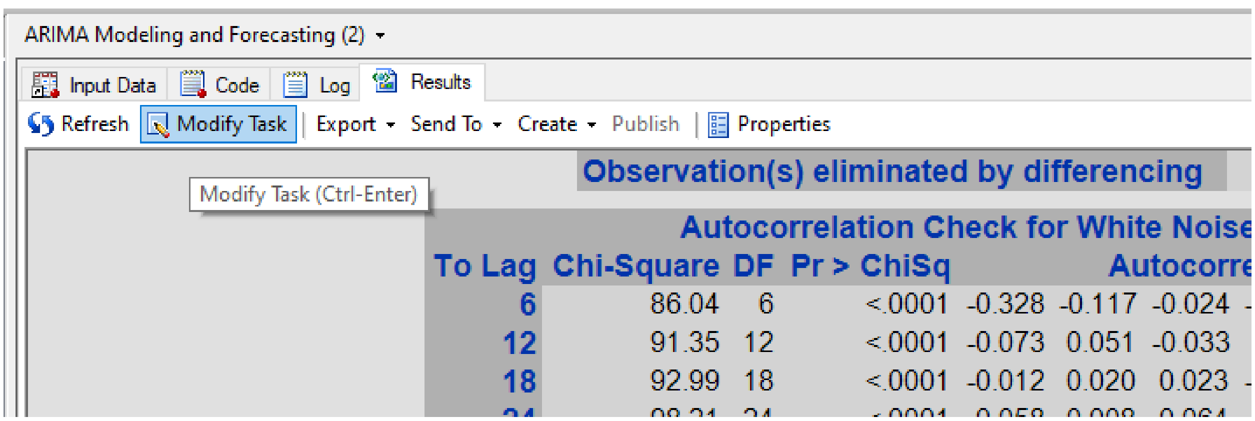
Alicia N Phan



Based on the above results, we noted:

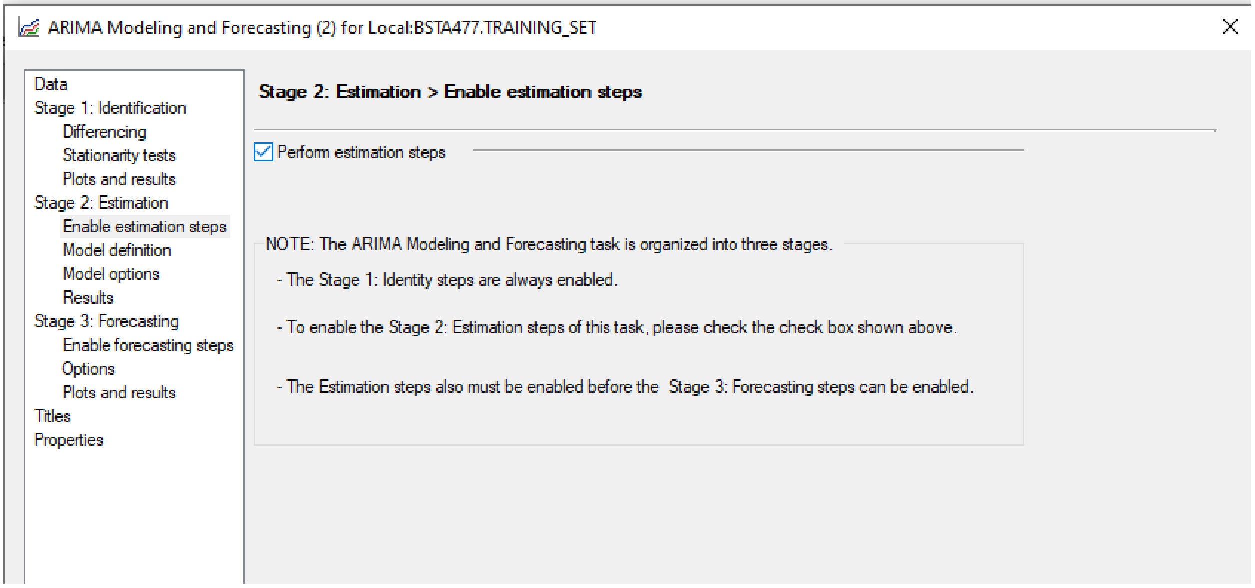
- From the ACF plot, there are moderate autocorrelations. => AR(1) might be needed.
- From the PACF plot, there are moderate negative autocorrelations. => MA(1) might be needed.

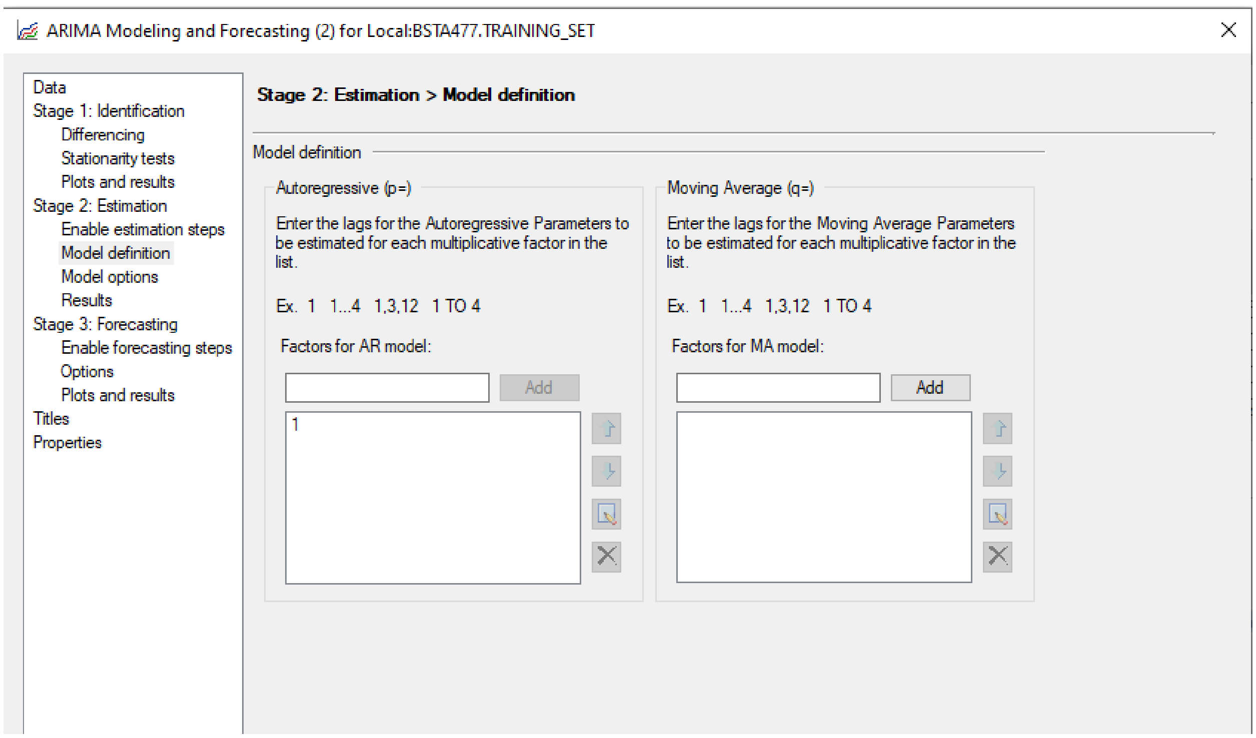
From the analysis, we try to adjust one by one. We apply AR(1) first.



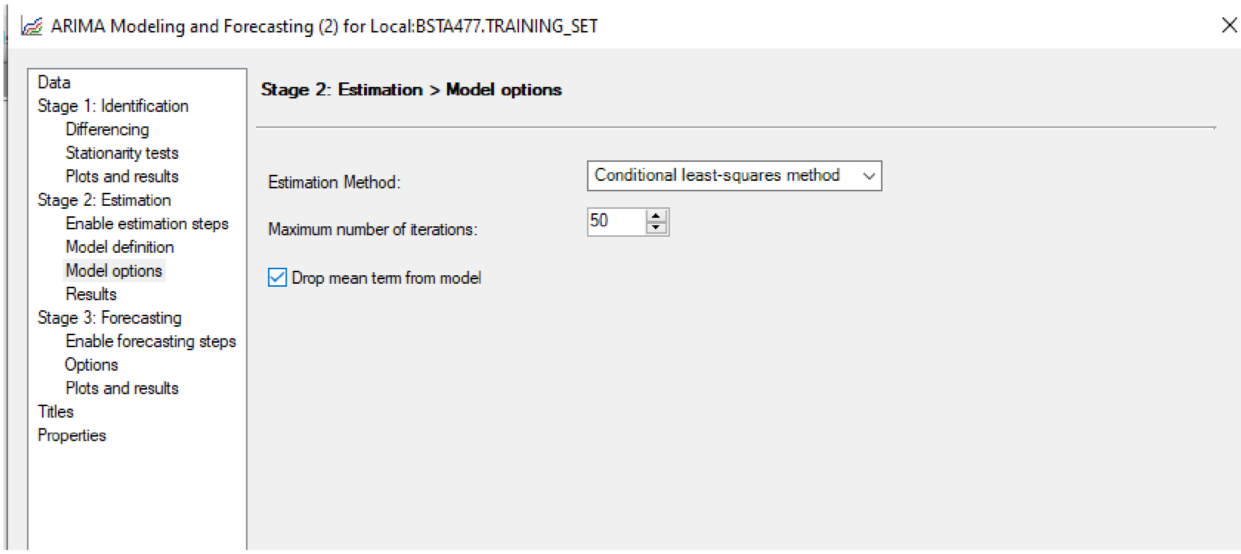
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorre		
6	86.04	6	<.0001	-0.328	-0.117	-0.024
12	91.35	12	<.0001	-0.073	0.051	-0.033
18	92.99	18	<.0001	-0.012	0.020	0.023

Alicia N Phan





Alicia N Phan



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

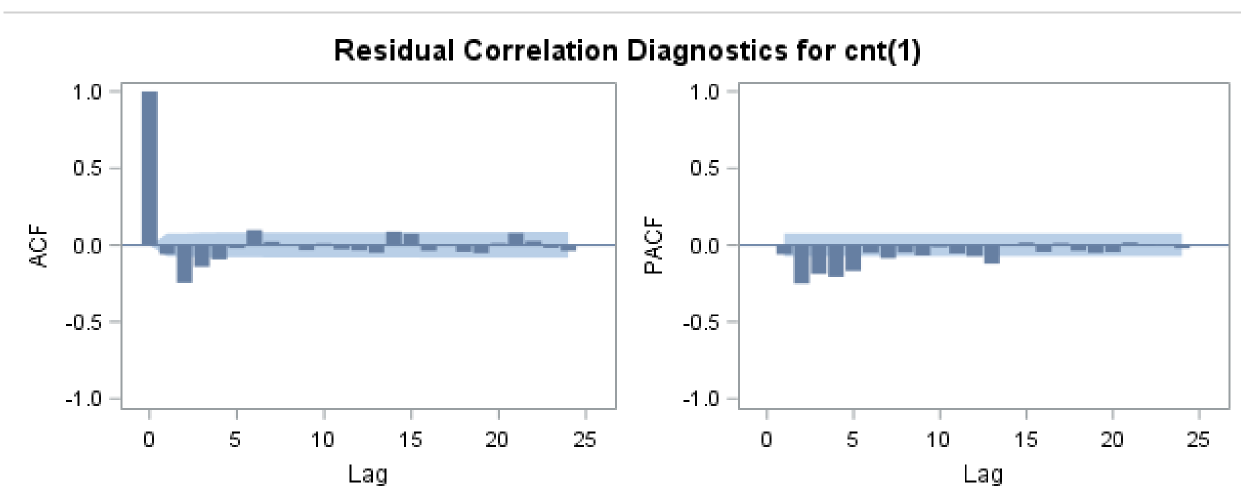
=> Click Run.

Conditional Least Squares Estimation

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

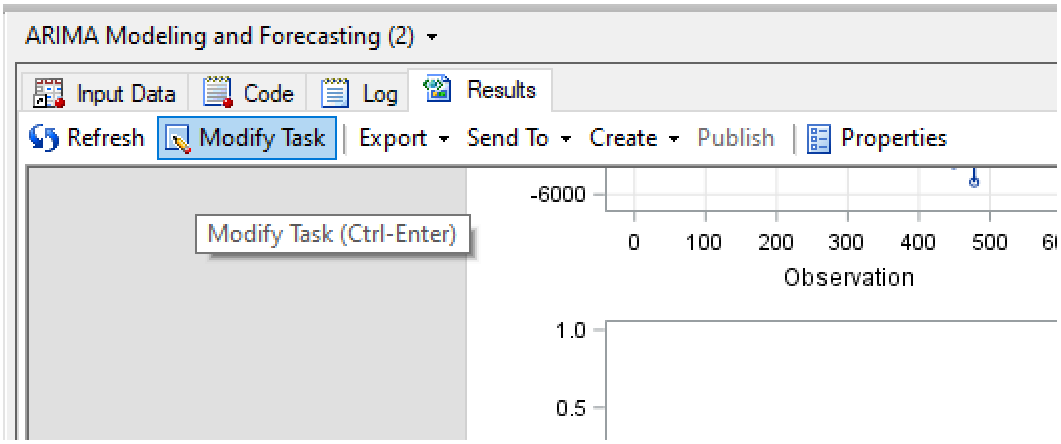
Here we see that the p value < 0.0001. If our significant level is 0.05. Then, the AR coefficient of our model is significant.

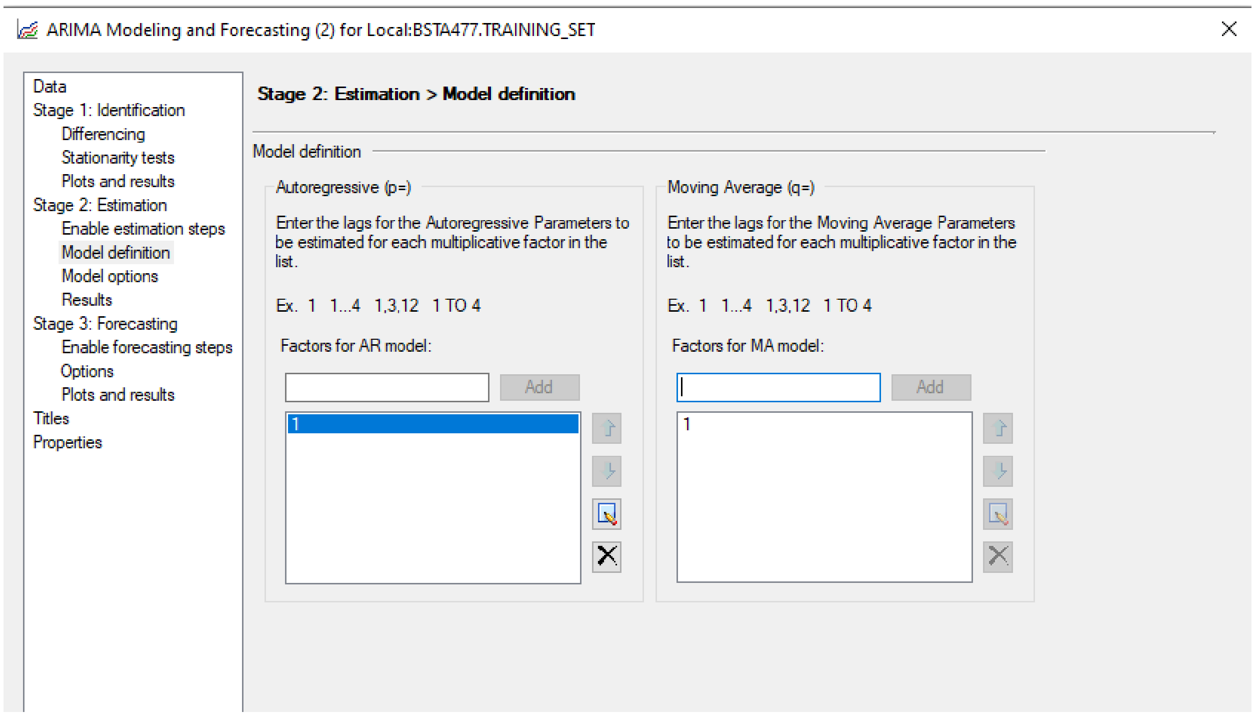
Result:

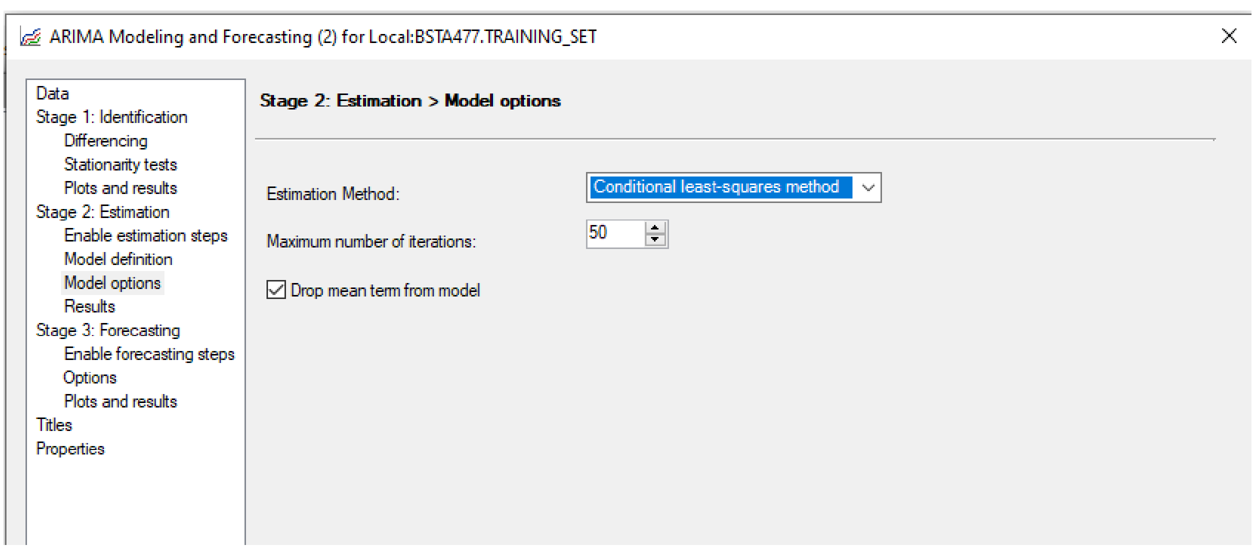


We apply MA(1) now.

Alicia N Phan





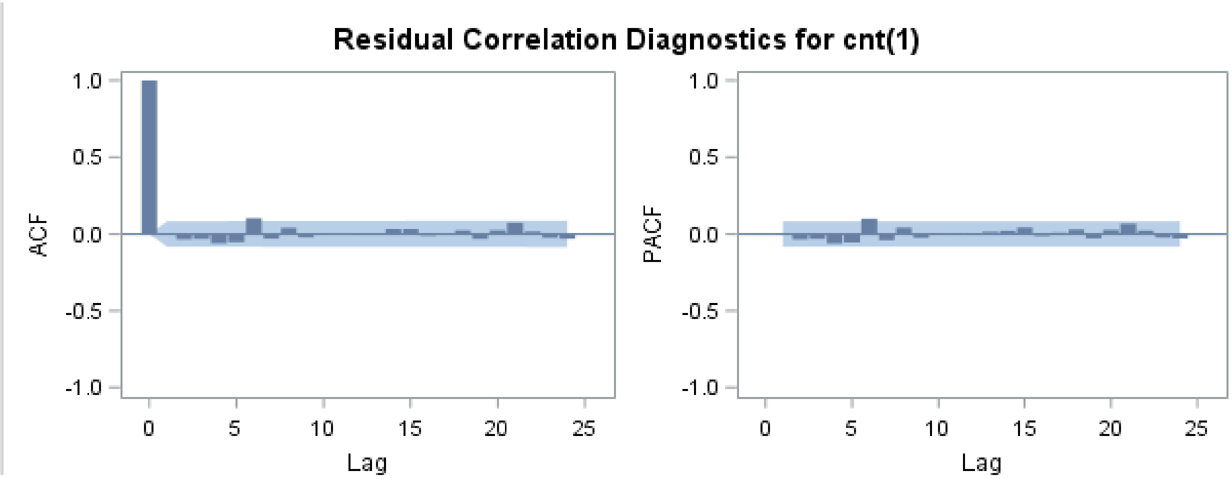


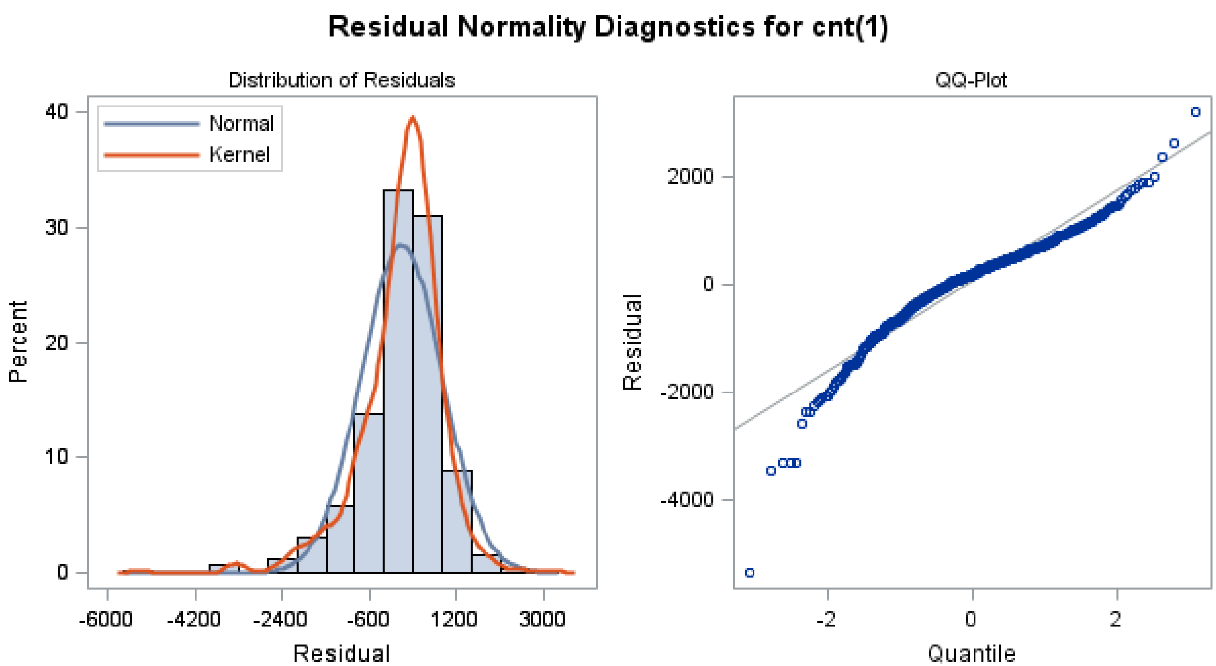
Alicia N Phan

=> 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

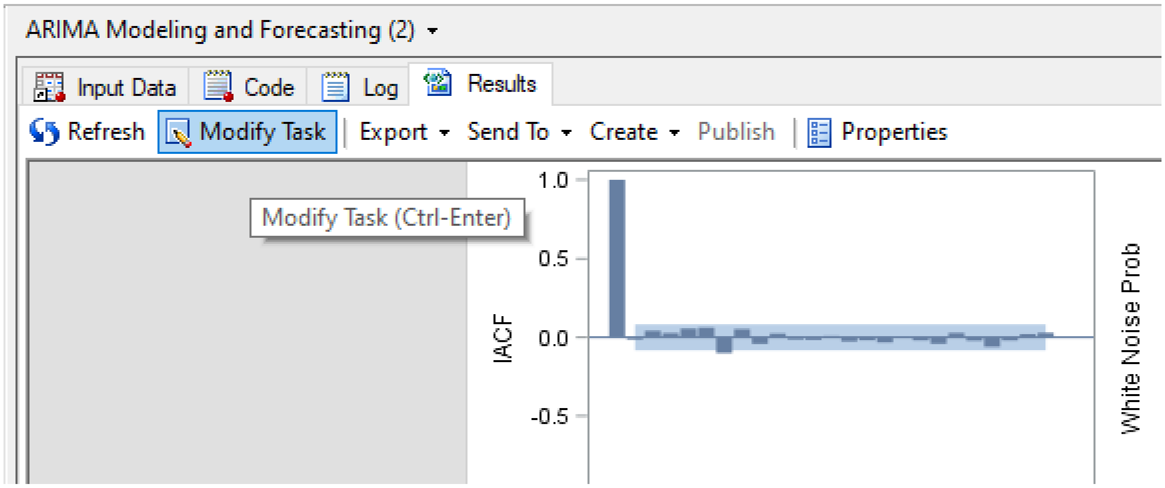
Alicia N Phan

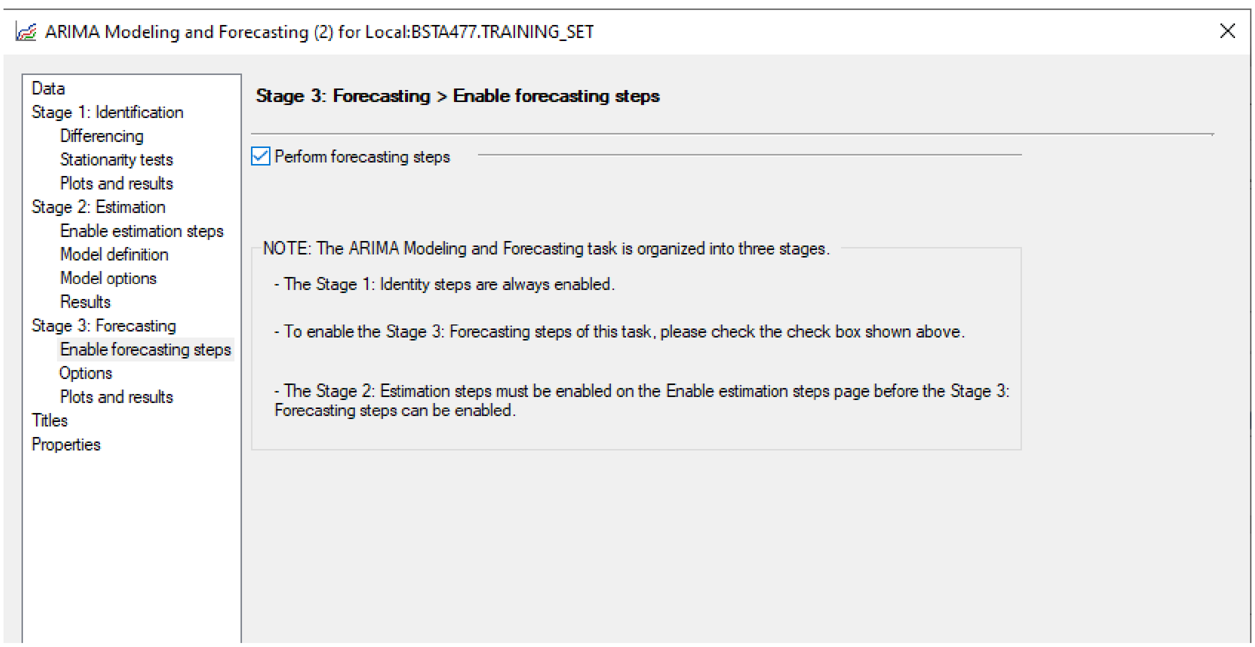
- The residuals seem to be normally distributed.

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

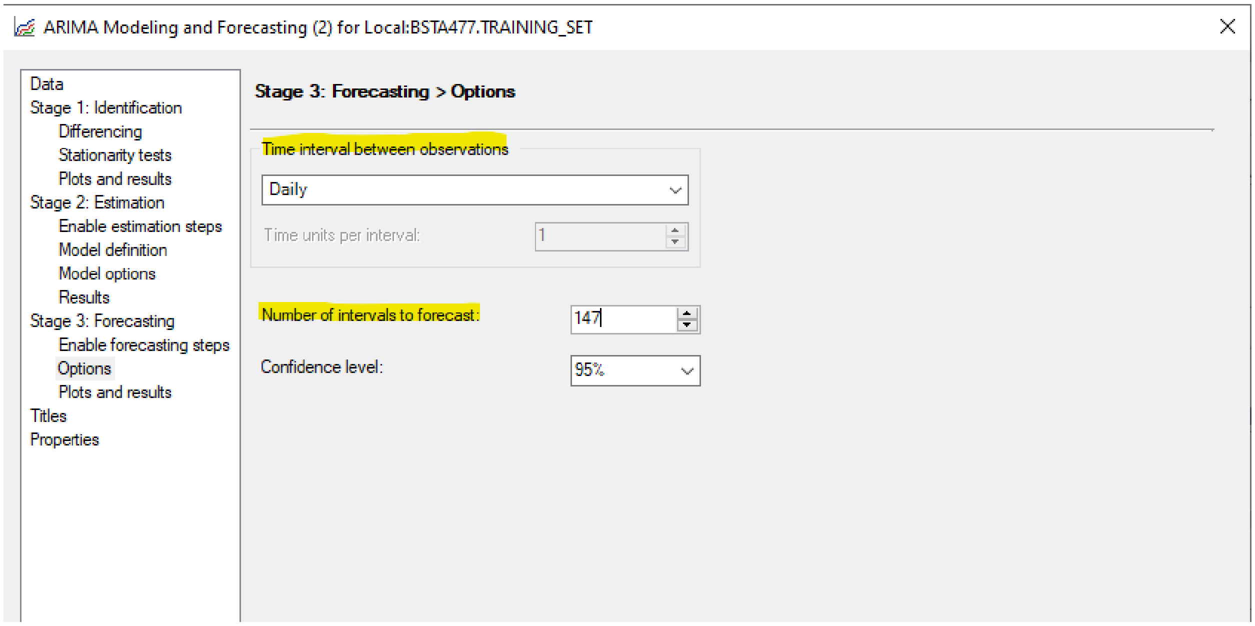
Forecast and validation

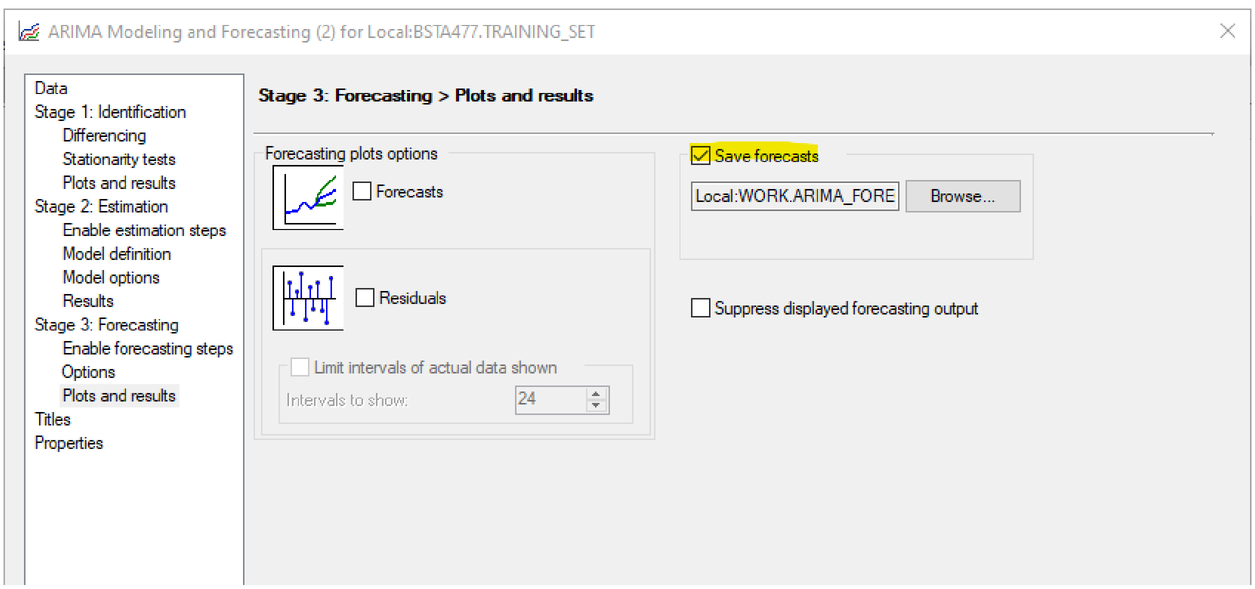
Forecast observations for validation set.





Alicia N Phan





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.arima_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

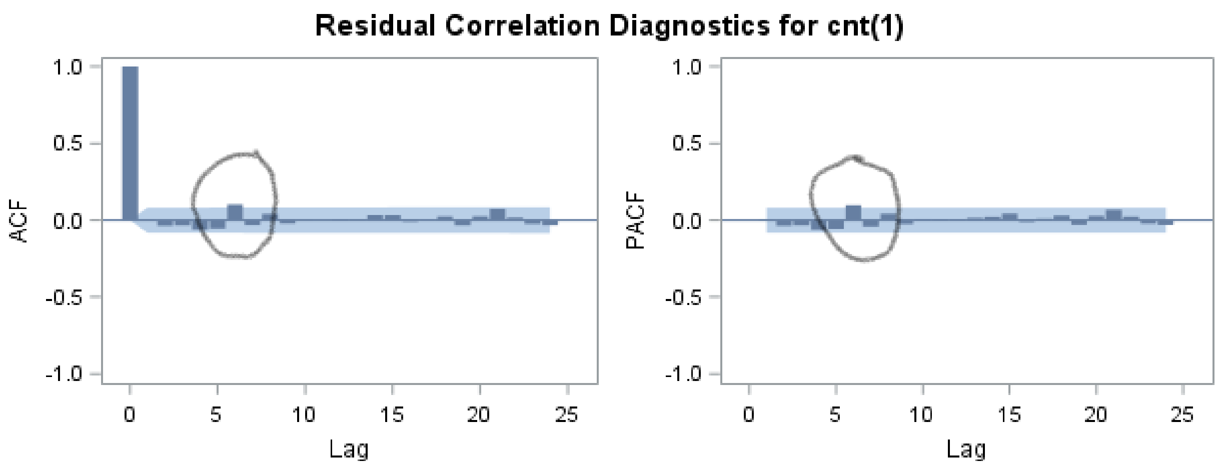
To handle ARIMA with seasonality, we need to add the seasonal component part of the model.

Alicia N Phan

$$\text{ARIMA} \quad \underbrace{(p,d,q)}_{\substack{\uparrow \\ \text{Non-seasonal part} \\ \text{of the model}}} \quad \underbrace{(P,D,Q)_m}_{\substack{\uparrow \\ \text{Seasonal part} \\ \text{of the model}}}$$

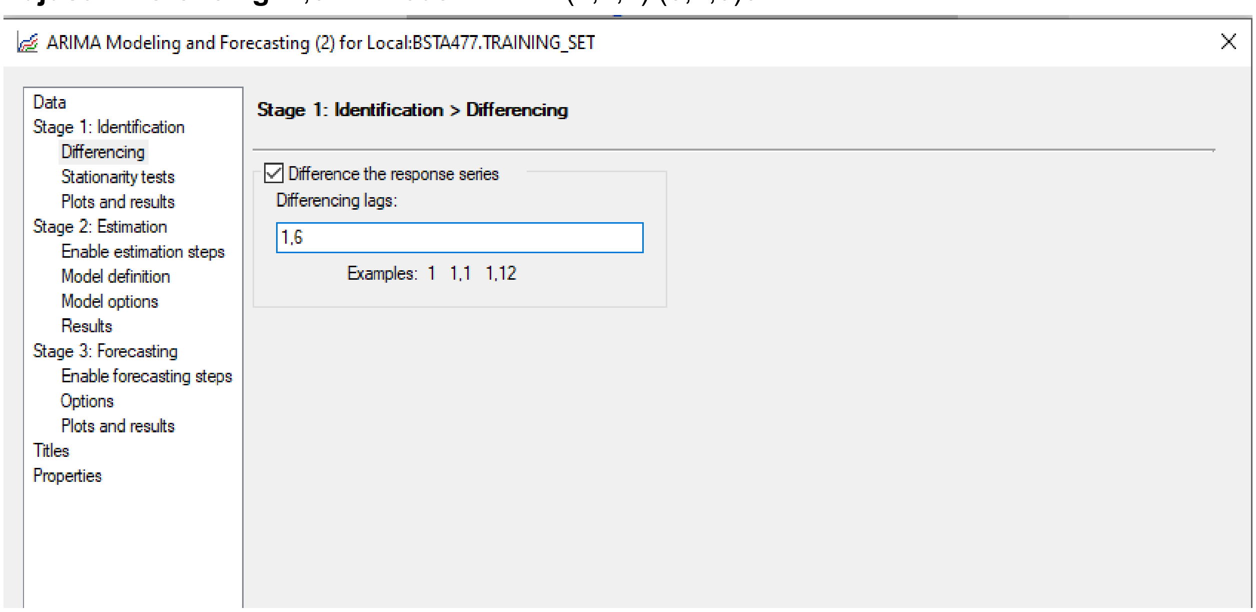
Based on the above analysis of the current dataset, we currently use the ARIMA(1,1,1) model.

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.

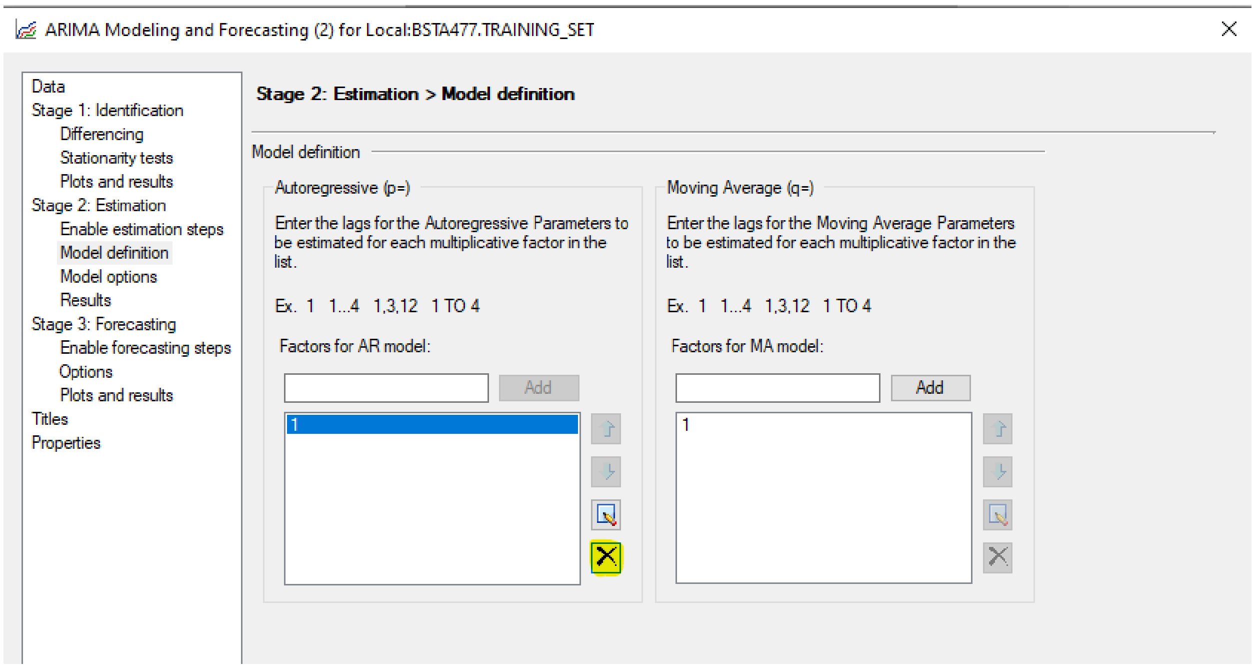
Depending on the significant lags:

- 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.

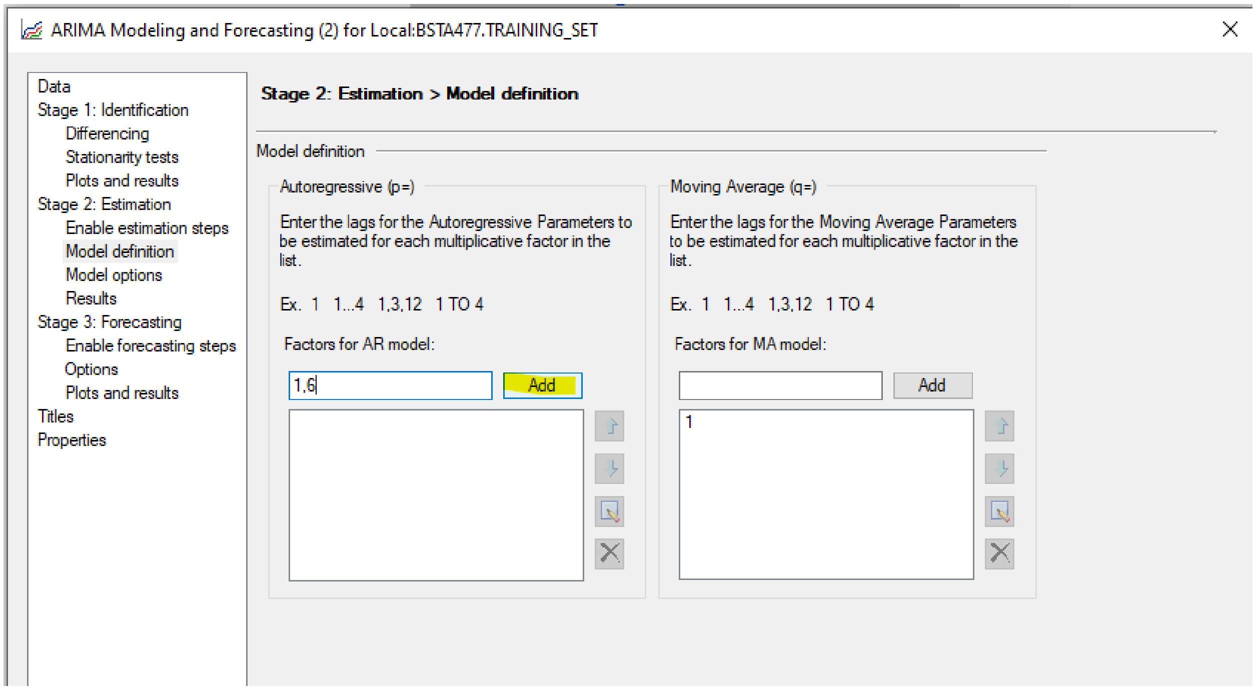
- **Add AR process**: 1,6 => Model: ARIMA (1,1,1) (1,1,0)6 (if MA process is not adjusted)

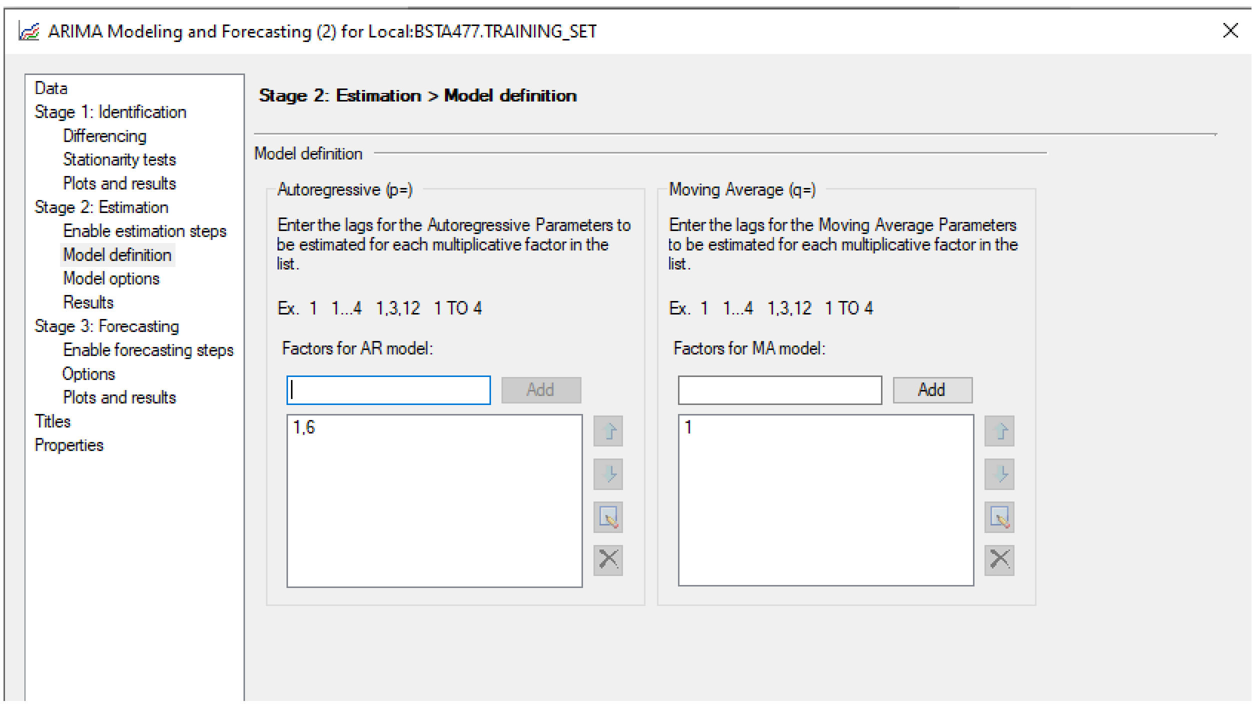
If there are still significant lags in the ACF. Then add an AR process into the seasonality.

1. Omit the previous AR process:



2. Add the new AR process





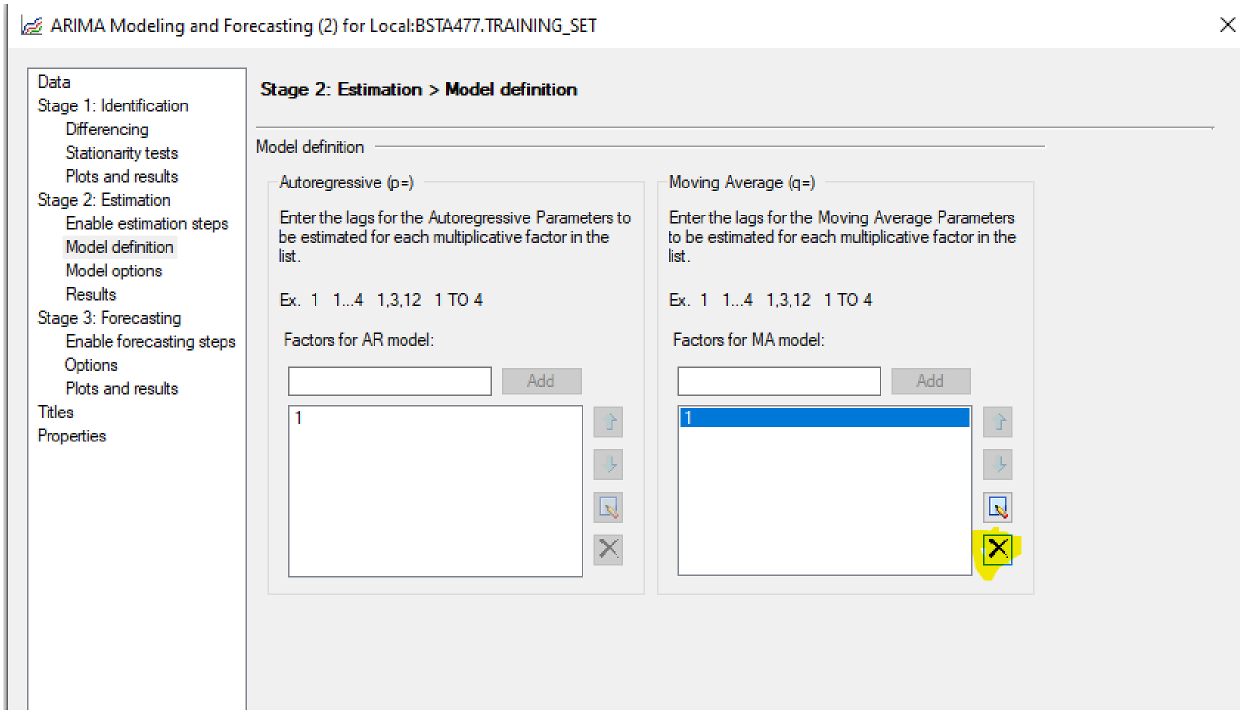
Click Run and check outputs. Pay attention to the ACF and PACF of the residuals. If there are still significant lags in PACF then add the MA process step below.

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

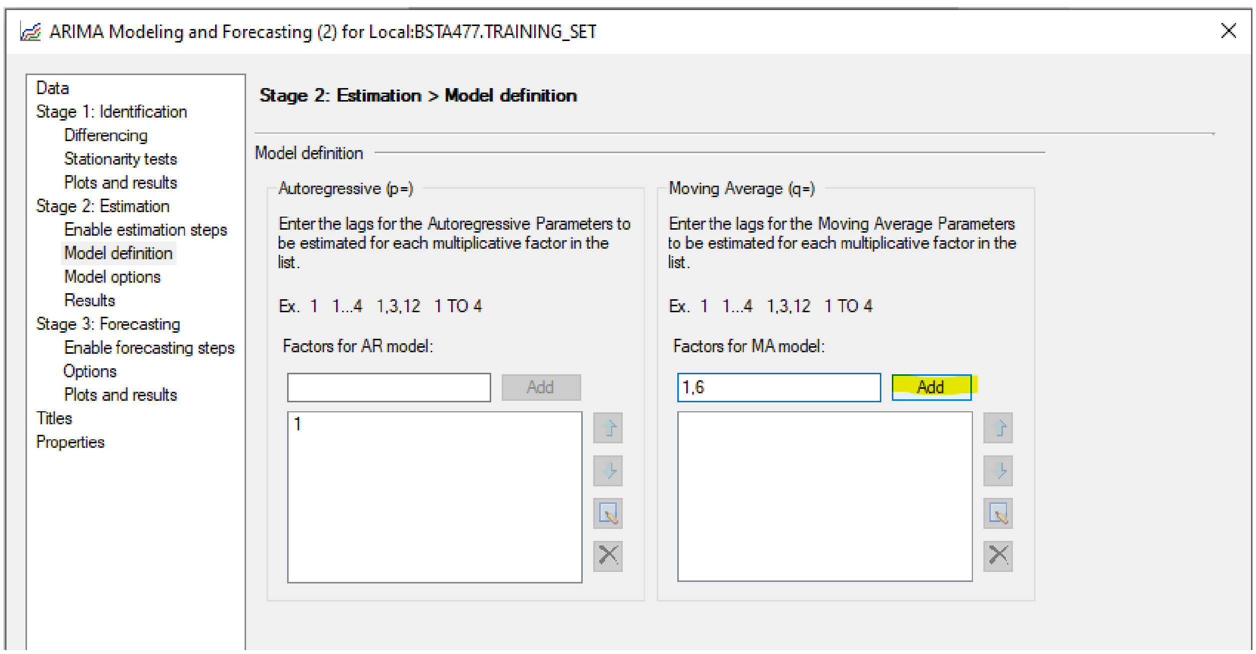
If there are only significant lags in PACF then only add an MA process using the following steps:

1. Omit the previous MA process:

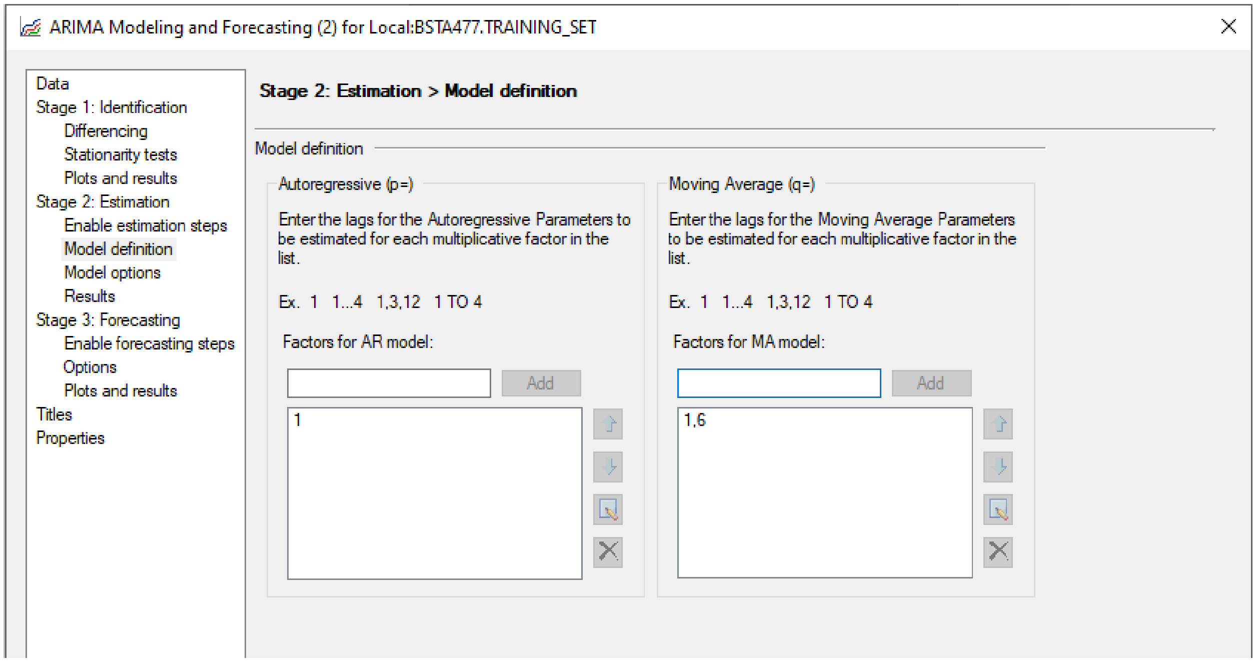
Alicia N Phan



2. Add MA process:

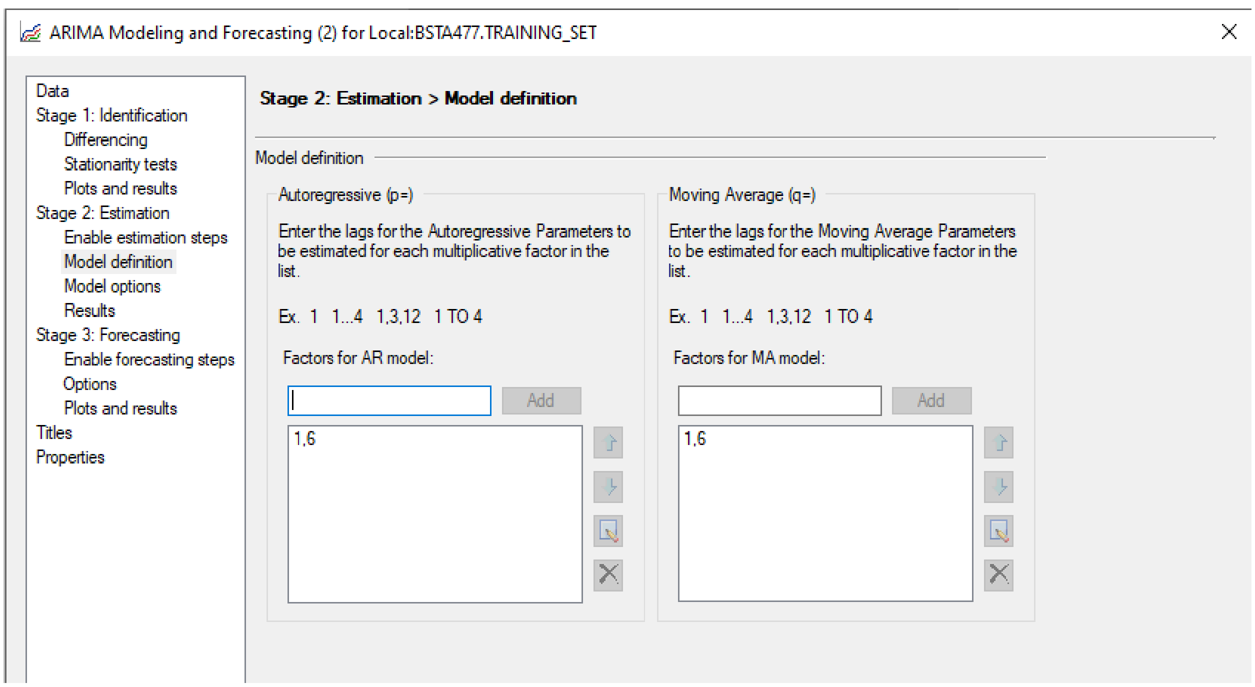


Alicia N Phan



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.