Case Study #3: Forecasting with AR and ARIMA Models

Case Solutions

Consider the quarterly data on Walmart revenues (in $million) from the first quarter of 2005 through the first quarter of 2022 (*673_case2.csv*). The goal is to forecast Walmart's quarterly revenue in the four quarters (Q1-Q4) of 2023 and 2024.

As you did in *case study #2*, start this case with the following:
Create time series data set in R using the *ts()* function (this part will not be graded in case #3).

```
> revenue.ts
       Qtr1   Qtr2   Qtr3   Qtr4
2005  71680  76697  75397  88327
2006  79676  85430  84467  98795
2007  86410  92999  91865 105749
2008  94940 102342  98345 108627
2009  94242 100876  99373 113594
2010  99811 103726 101952 116360
2011 104189 109366 110226 122728
2012 113010 114282 113800 127559
2013 114070 116830 115688 129706
2014 114960 120125 119001 131565
2015 114826 120229 117408 129667
2016 115904 120854 118179 130936
2017 117542 123355 123179 136267
2018 122690 128028 124894 138793
2019 123925 130377 127991 141671
2020 134622 137742 134708 152079
2021 138310 141048 140525 152871
2022 141569 152859 152813 164048
```

Develop data partition with the validation partition of 20 periods and the rest for the training partition (this part will not be graded in case #3).

```
> train.ts
       Qtr1   Qtr2   Qtr3   Qtr4
2005  71680  76697  75397  88327
2006  79676  85430  84467  98795
2007  86410  92999  91865 105749
2008  94940 102342  98345 108627
2009  94242 100876  99373 113594
2010  99811 103726 101952 116360
2011 104189 109366 110226 122728
2012 113010 114282 113800 127559
2013 114070 116830 115688 129706
2014 114960 120125 119001 131565
2015 114826 120229 117408 129667
2016 115904 120854 118179 130936
2017 117542 123355 123179 136267
2018 122690 128028 124894 138793
```

```
> valid.ts
       Qtr1   Qtr2   Qtr3   Qtr4
2019 123925 130377 127991 141671
2020 134622 137742 134708 152079
2021 138310 141048 140525 152871
2022 141569 152859 152813 164048
```

1. Identify time series predictability.

1a. Using the *AR(1)* model for the historical data, Provide and explain the *AR(1)* model summary in your report. Explain if the Walmart revenue is predictable.

The output of the *AR(1)* model for *revenue.ts* time series data is presented below. *ARIMA(1, 0, 0)* is an autoregressive (AR) model with order 1, no differencing, and no moving average model.

```
Series: revenue.ts
ARIMA(1,0,0) with non-zero mean

Coefficients:
         ar1       mean
      0.9269  117007.26
s.e.  0.0525   13308.28

sigma^2 = 94816130:  log likelihood = -763.36
AIC=1532.71   AICc=1533.07   BIC=1539.54

Training set error measures:
                   ME     RMSE      MAE       MPE     MAPE     MASE       ACF1
Training set 972.5455 9601.164 8070.539 0.2146869 7.057306 1.765948 -0.6390037
```

The model's equation is:

$$Y_t = 117007.26 + 0.9269\ Y_{t\text{-}1}$$

The coefficient of the *ar1* ($Y_{t\text{-}1}$) variable, β1 = 0.9269, and standard error of estimate, s.e. = 0.0525. We will use these two parameters for hypothesis testing about the value of the AR(1) regression coefficient.

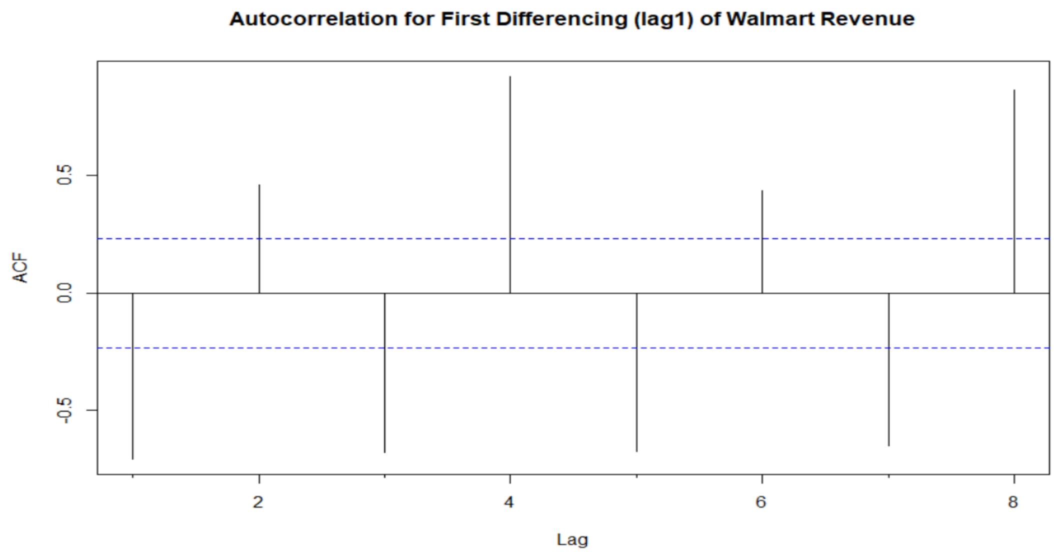
Hypothesis Testing: Z- Test
Null hypothesis Ho: $\beta1 = 1$
Alternative hypothesis H1: $\beta1 \neq 1$

z-statistic = (β1 - 1)/(s.e.) = (0.9269 - 1)/0.0525 = -1.392
p-value for z-statistic = 0.0819

Based on the p-value of 0.0819, which is greater than 0.05, we cannot reject (need to accept) the null hypothesis that β1 = 1. Therefore, the time series data for Walmart revenue, *revenue.ts*, according to this test, is not predictable and is considered random walk.

1b. Using the first differencing (lag-1) of the historical data and *Acf()* function, provide in the report the autocorrelation plot of the first differencing (lag-1) with the maximum of 8 lags and explain if Walmart revenue is predictable.

The autocorrelation plot of the first differencing for the *revenue.ts* data is presented below.



All autocorrelation coefficients of the first differenced data are statistically significant, in particular, in lag-1 for trend and lag-4 for quarterly seasonality. Therefore, using the first differencing, we can confirm that *revenue.ts* is not random walk and is predictable. Because the results of the two predictability tests in *1b* and *1c* are opposite, we will continue to utilize the data set as predictable in forecasting Walmart revenue in Q1-Q4 of 2023 and 2024.

2. Apply the two-level forecast with regression model and AR model for residuals.

2a. For the training data set, use the *tslm()* function to develop a *regression model with quadratic trend and seasonality*. Forecast Walmart's revenue with the *forecast()* function (use the associated R code from case #2). No explanation is required in your report.

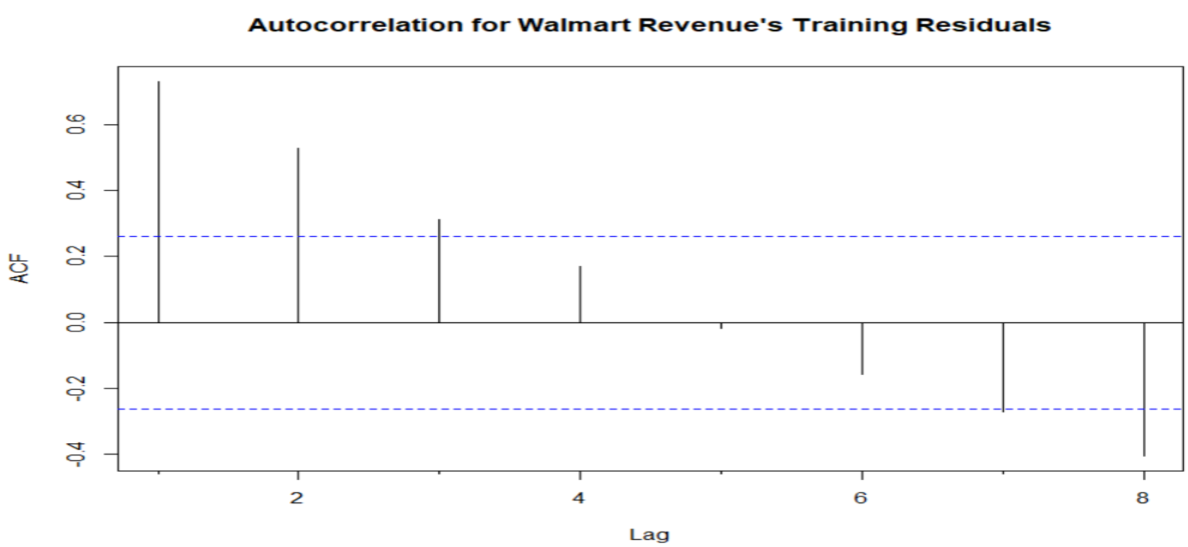
The output for the regression model with quadratic trend and seasonality for the training period and forecast for the validation period are shown below (<u>not graded in this case; were used in case #2</u>).

```
Call:
tslm(formula = train.ts ~ trend + I(trend^2) + season)

Residuals:
    Min      1Q  Median      3Q     Max 
-3583.3 -1950.1   232.7  1443.6  5664.9 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 71042.26    1042.04  68.176  < 2e-16 ***
trend        1745.66      74.67  23.379  < 2e-16 ***
I(trend^2)    -15.20       1.27 -11.974 2.67e-16 ***
season2      4175.43     838.90   4.977 8.04e-06 ***
season3      1770.27     839.51   2.109     0.04 *  
season4     14128.66     840.51  16.810  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2219 on 50 degrees of freedom
Multiple R-squared:  0.9829, Adjusted R-squared:  0.9812 
F-statistic: 575.7 on 5 and 50 DF,  p-value: < 2.2e-16
```

```
> train.trend.season.pred
        Point Forecast     Lo 0     Hi 0
2019 Q1       121150.3 121150.3 121150.3
2019 Q2       125323.1 125323.1 125323.1
2019 Q3       122884.8 122884.8 122884.8
2019 Q4       135179.7 135179.7 135179.7
2020 Q1       120957.2 120957.2 120957.2
2020 Q2       125008.3 125008.3 125008.3
2020 Q3       122448.4 122448.4 122448.4
2020 Q4       134621.7 134621.7 134621.7
2021 Q1       120277.5 120277.5 120277.5
2021 Q2       124207.0 124207.0 124207.0
2021 Q3       121525.5 121525.5 121525.5
2021 Q4       133577.1 133577.1 133577.1
2022 Q1       119111.3 119111.3 119111.3
2022 Q2       122919.2 122919.2 122919.2
2022 Q3       120116.1 120116.1 120116.1
2022 Q4       132046.1 132046.1 132046.1
```

2b. Identify the regression model's residuals for the training period and use the *Asf()* function to identify autocorrelation for these residuals. Provide the autocorrelation plot in your report and explain why it would be a good idea to add to your forecast an AR model for residuals.

The autocorrelation chart (correlogram) of the residuals from the regression model with quadratic trend and seasonality (question 2a) is provided below.



The chart shows significant autocorrelation of residuals in lags 1-3, as well as in lag 8, which means that these autocorrelations (relationships) between residuals are not incorporated into the regression model. Thus, modeling these residual autocorrelations with an *AR* model and developing a two-level model may, overall, improve the forecast.

2c. Develop an *AR(1)* model for the regression residuals, present and explain the model and its equation in your report. Use the *Acf()* function for the residuals of the *AR(1)* model (residuals of residuals), present the autocorrelation chart, and explain it in your report.

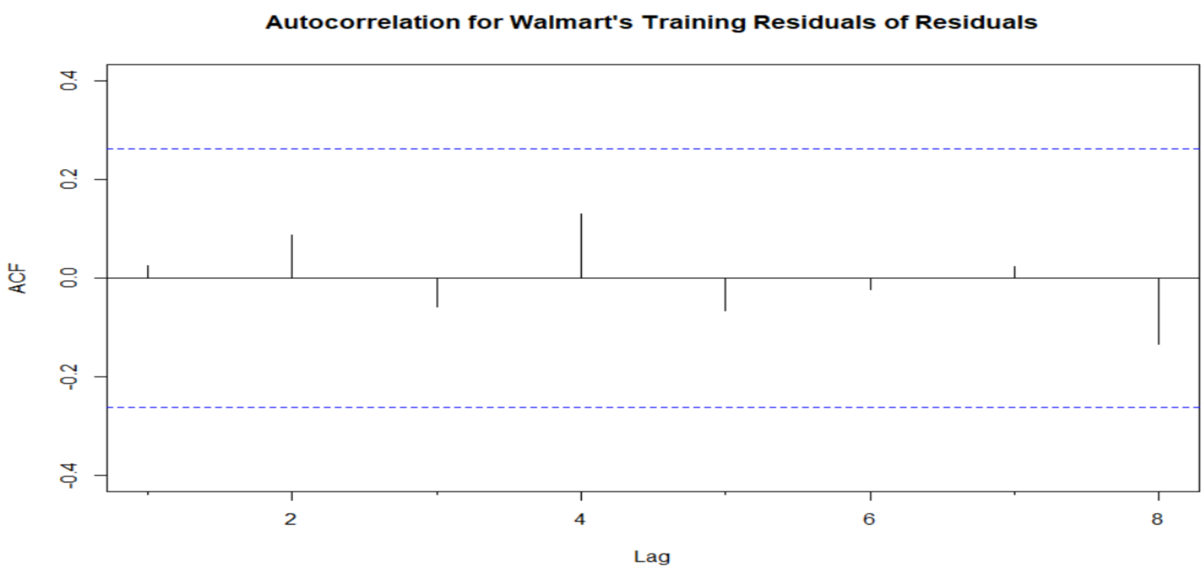
The output of the *AR(1)* model for regression residuals is presented below. *ARIMA(1, 0, 0)* is an autoregressive (AR) model with order 1, no differencing, and no moving average model.

```
Series: train.trend.season$residuals
ARIMA(1,0,0) with non-zero mean

Coefficients:
         ar1      mean
      0.7585  123.4899
s.e.  0.0876  728.1704

sigma^2 = 1987234:  log likelihood = -484.93
AIC=975.87   AICc=976.33   BIC=981.94

Training set error measures:
                   ME     RMSE      MAE      MPE     MAPE      MASE       ACF1
Training set 24.03388 1384.291 1089.501 52.83824 121.3348 0.5191297 0.02529155
```

The *AR(1)* model's equation is:

$$e_t = 123.490 + 0.759\, e_{t-1}$$

An autocorrelation chart for the *AR(1)* model's residuals (residuals of residuals) is presented below.



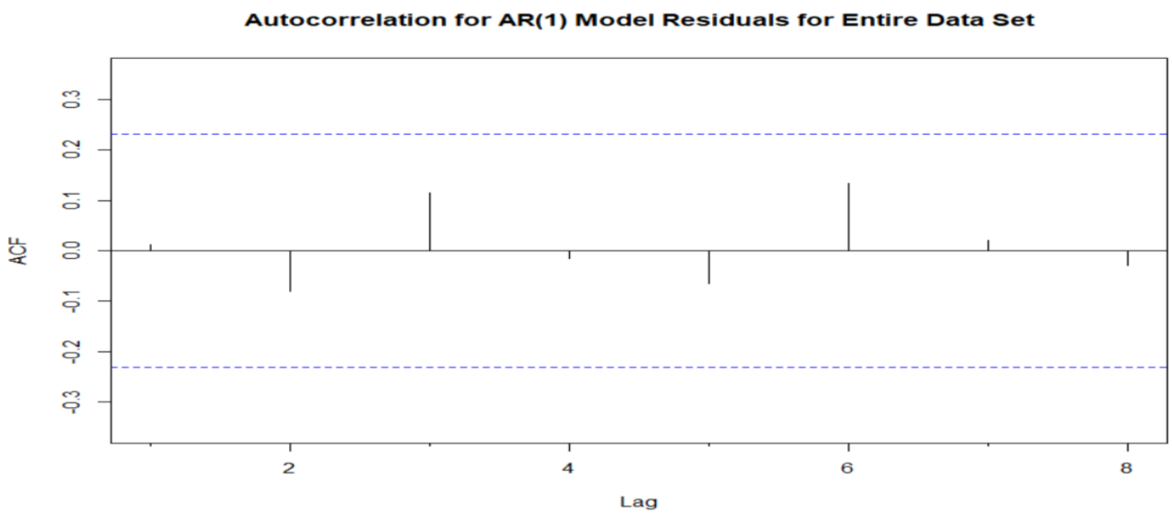
As can be seen from the autocorrelation chart (correlogram), all autocorrelations of residuals of residuals created by *AR(1)* model are random. Thus, the *AR(1)* model for residuals has absorbed significant autocorrelation in all lags. Therefore, the *AR(1)* model for residuals can be combined with the regression model in question 2a to improve the time series forecast with the two-level forecasting model.

2d. Create a two-level forecasting model (regression model with quadratic trend and seasonality + *AR(1)* model for residuals) for the validation period. Show in your report a table with the validation data, regression forecast for the validation data, *AR(1)* forecast for the validation data, and combined forecast for the validation period.

The table below describes the revenue data and forecasts in the validation partition of 16 quarters in 2019-2022 (*Valid.Revenue*), regression model's forecast in the validation period (*Reg.Forecast*), *AR(1)* model's forecast of the regression residuals in the validation period (*AR(1)Forecast*), and combined forecast (*Combined.Forecast*) as a sum of the regression and *AR(1)* models' forecasts.

	Valid.Revenue	Reg.Forecast	AR(1)Forecast	Combined.Forecast
1	123925	121150.3	2716.1151	123866.4
2	130377	125323.1	2089.9065	127413.0
3	127991	122884.8	1614.9489	124499.8
4	141671	135179.7	1254.7100	136434.4
5	134622	120957.2	981.4812	121938.6
6	137742	125008.3	774.2467	125782.5
7	134708	122448.4	617.0664	123065.5
8	152079	134621.7	497.8506	135119.5
9	138310	120277.5	407.4296	120684.9
10	141048	124207.0	338.8483	124545.8
11	140525	121525.5	286.8318	121812.3
12	152871	133577.1	247.3791	133824.5
13	141569	119111.3	217.4556	119328.8
14	152859	122919.2	194.7596	123114.0
15	152813	120116.1	177.5455	120293.6
16	164048	132046.1	164.4892	132210.6

2e. Develop a two-level forecast (regression model with *quadratic trend and seasonality* and *AR(1)* model for residuals) for the entire data set. Provide in your report the autocorrelation chart for the *AR(1)* model's residuals and explain it. Also, provide a data table with the models' forecasts for Walmart revenue in Q1-Q4 of 2023 and 2024 (regression model, *AR(1)* for residuals, and two-level combined forecast).



The autocorrelation chart above of the *AR(1)* model residuals (residuals of residuals) shows that all autocorrelations are random (within horizontal thresholds), which means that the *AR(1)* model absorbed significant autocorrelations in the residuals.

The table below provides 8 forecasts for Q1-Q4 of 2023-2024, that are associated with: the regression model with quadratic trend and seasonality (*Reg.Forecast*), *AR(1)* model for the regression residuals (*AR(1)Forecast*), and two-level combined forecast (*Combined.Forecast*) as a sum of the regression and *AR(1)* models' forecasts.

	Reg.Forecast	AR(1)Forecast	Combined.Forecast
1	141878.9	7799.626	149678.5
2	146951.0	7146.464	154097.4
3	145237.3	6550.160	151787.4
4	158350.5	6005.766	164356.3
5	144676.1	5508.761	150184.9
6	149727.7	5055.021	154782.7
7	147993.6	4640.779	152634.4
8	161086.5	4262.598	165349.1

3. Use ARIMA Model and Compare Various Methods.

3a. Use *Arima()* function to fit *ARIMA(1,1,1)(1,1,1)* model for the *training data set*. Insert in your report the summary of this ARIMA model, present and briefly explain the *ARIMA* model and its equation in your report. Using this model, forecast revenue for the *validation period* and present it in your report.

The output from the *ARIMA(1,1,1)(1,1,1)* model for the training partition is presented below.

```
Series: train.ts
ARIMA(1,1,1)(1,1,1)[4]

Coefficients:
          ar1     ma1    sar1     sma1
      -0.7265  0.6765  0.2647  -0.8859
s.e.   0.4345  0.4439  0.2159   0.2393

sigma^2 = 2793497:  log likelihood = -450.8
AIC=911.61   AICc=912.94   BIC=921.27

Training set error measures:
                    ME    RMSE      MAE        MPE      MAPE      MASE        ACF1
Training set -332.0514 1531.19 1072.693 -0.3146559 0.9838007 0.2607983 -0.02207348
```

This is a seasonal ARIMA model, *ARIMA(p, d, q)(P, D, Q)$_m$*, where:

- *p = 1,* order 1 autoregressive model *AR(1)*
- *d = 1*, first differencing
- *q = 1*, order 1 moving average *MA(1)* for error lags
- *P = 1,* order 1 autoregressive model *AR(1)* for the seasonal part
- *D = 1*, first differencing for the seasonal part
- *Q = 1*, order 1 moving average *MA(1)* for the seasonal error lags
- *m = 4*, for quarterly seasonality.

The model's equation is:

$$y_t - y_{t-1} = - 0.7265(y_{t-1} - y_{t-2}) + 0.6765e_{t-1} + 0.2647y_{t-1} - y_{t-5}) - 0.8859\rho_{t-1}$$

Using the model's equation, see below the forecast for the validation period:

```
        Point Forecast     Lo 0     Hi 0
2019 Q1       125432.2 125432.2 125432.2
2019 Q2       130637.3 130637.3 130637.3
2019 Q3       128513.3 128513.3 128513.3
2019 Q4       141960.2 141960.2 141960.2
2020 Q1       128726.5 128726.5 128726.5
2020 Q2       133845.6 133845.6 133845.6
2020 Q3       132025.9 132025.9 132025.9
2020 Q4       145326.3 145326.3 145326.3
2021 Q1       132145.8 132145.8 132145.8
2021 Q2       137228.0 137228.0 137228.0
2021 Q3       135499.1 135499.1 135499.1
2021 Q4       148753.3 148753.3 148753.3
2022 Q1       135592.2 135592.2 135592.2
2022 Q2       140660.7 140660.7 140660.7
2022 Q3       138958.7 138958.7 138958.7
2022 Q4       152198.6 152198.6 152198.6
```

3b. Use the *auto.arima()* function to develop an *ARIMA* model using the *training data set*. Insert in your report the summary of this *ARIMA* model, present and explain the *ARIMA* model and its equation in your report. Use this model to forecast revenue in the *validation period* and present this forecast in your report.

The output from using the *auto.arima()* function for the training partition is presented below.

```
Series: train.ts
ARIMA(0,1,0)(1,1,1)[4]

Coefficients:
        sar1     sma1
      0.2992  -0.9236
s.e.  0.2004   0.3077

sigma^2 = 2639340:  log likelihood = -450.91
AIC=907.81   AICc=908.32   BIC=913.61

Training set error measures:
                    ME     RMSE      MAE        MPE     MAPE      MASE        ACF1
Training set -323.7627 1519.678 1058.007 -0.3067335 0.971084 0.2572277 -0.07034954
```

This is a seasonal ARIMA model, *(0,1,0)(0,1,1)₄*, with the following parameters:

- *p = 0*, no autoregressive model
- *d = 1*, first differencing
- *q = 0*, no moving average model for error lags
- *P = 1*, no autoregressive model for the seasonal part
- *D = 1*, first differencing for the seasonal part
- *Q = 1* order 1 moving average model for the seasonal part's error lags
- m = 4, for the quarterly seasonality.

The ARIMA model's equation is:

$$y_t - y_{t-1} = 0.2992(y_{t-1} - y_{t-5}) - 0.9236\rho_{t-1}$$

This ARIMA model's forecast in the validation period is the following:

```
        Point Forecast    Lo 0     Hi 0
2019 Q1       125652.2 125652.2 125652.2
2019 Q2       130798.6 130798.6 130798.6
2019 Q3       128720.9 128720.9 128720.9
2019 Q4       142147.7 142147.7 142147.7
2020 Q1       129137.4 129137.4 129137.4
2020 Q2       134226.4 134226.4 134226.4
2020 Q3       132464.8 132464.8 132464.8
2020 Q4       145750.3 145750.3 145750.3
2021 Q1       132779.0 132779.0 132779.0
2021 Q2       137850.9 137850.9 137850.9
2021 Q3       136183.8 136183.8 136183.8
2021 Q4       149427.1 149427.1 149427.1
2022 Q1       136467.5 136467.5 136467.5
2022 Q2       141534.3 141534.3 141534.3
2022 Q3       139895.5 139895.5 139895.5
2022 Q4       153126.1 153126.1 153126.1
```

3c. Apply the *accuracy()* function to compare performance measures of the two *ARIMA* models in 3a and 3b. Present the accuracy measures in your report, compare them and identify, using MAPE and RMSE, the best *ARIMA* model to apply.

ARIMA Model (1,1,1)(1,1,1)₄

```
               ME     RMSE      MAE   MPE MAPE  ACF1 Theil's U
Test set 4978.392 6694.686 5300.759 3.346  3.6 0.675     0.698
```

Auto ARIMA (0,1,0)(1,1,1)₄

```
               ME     RMSE      MAE   MPE MAPE ACF1 Theil's U
Test set 4437.231 6141.002 4856.641 2.973  3.3 0.66      0.64
```

Based on the *MAPE* and *RMSE* accuracy measures, the best model is the auto ARIMA model, *ARIMA* $(0,1,0)(1,1,1)_4$, which has the lowest values of *MAPE* (2.97% rounded) and *RMSE* (6141.0 rounded) vs. the respective measures for the ARIMA model *ARIMA* $(1,1,1)(1,1,1)_4$,

3d. Use two *ARIMA* models from 3a and 3b for the entire data set. Present models' summaries in your report. Use these *ARIMA* models to forecast Walmart revenue in Q1-Q4 of 2023-2024 and present these forecasts in your report.

ARIMA Model $(1,1,1)(1,1,1)_4$

The output for this ARIMA model for the entire data set is shown below.

```
Series: revenue.ts 
ARIMA(1,1,1)(1,1,1)[4] 

Coefficients:
         ar1      ma1    sar1     sma1
      0.3224  -0.3978  0.0788  -1.0000
s.e.  0.7116   0.6857  0.1482   0.1121

sigma^2 = 3866094:  log likelihood = -606.59
AIC=1223.19   AICc=1224.17   BIC=1234.21

Training set error measures:
                    ME     RMSE      MAE        MPE     MAPE      MASE          ACF1
Training set -208.3464 1839.248 1279.033 -0.2302129 1.068301 0.2798705 -0.006678155
```

$$y_t - y_{t-1} = 0.3224(y_{t-1} - y_{t-2}) - 0.3978e_{t-1} + 0.0788(y_{t-1} - y_{t-5}) - 1.0\rho_{t-1}$$

The model's forecast for the 8 future quarters is the following:

```
        Point Forecast     Lo 0     Hi 0
2023 Q1       151585.0 151585.0 151585.0
2023 Q2       157361.3 157361.3 157361.3
2023 Q3       155968.4 155968.4 155968.4
2023 Q4       169102.9 169102.9 169102.9
2024 Q1       156518.7 156518.7 156518.7
2024 Q2       161850.7 161850.7 161850.7
2024 Q3       160348.6 160348.6 160348.6
2024 Q4       173631.8 173631.8 173631.8
```

Auto ARIMA Model

The output for this auto ARIMA model for the entire data set is shown below.

```
Series: revenue.ts 
ARIMA(1,0,0)(2,1,0)[4] with drift 

Coefficients:
         ar1     sar1     sar2      drift
      0.8771  -0.5464  -0.2607  1196.3907
s.e.  0.0677   0.1416   0.1525   287.2862

sigma^2 = 4921015:  log likelihood = -619.42
AIC=1248.83   AICc=1249.8   BIC=1259.93

Training set error measures:
                    ME     RMSE      MAE         MPE     MAPE      MASE       ACF1
Training set -44.65316 2091.467 1547.659 -0.04586725 1.313318 0.3386497 -0.0473287
```

This auto ARIMA model's equation is:

The equation of this model is:

$$Y_t = 1196.391 + 0.8771Y_{t-1} - 0.546(Y_{t-1} - Y_{t-5}) - 0.261(Y_{t-2} - Y_{t-6})$$

The model's forecast for the 8 future quarters is the following:

```
        Point Forecast    Lo 0     Hi 0
2023 Q1       152452.5 152452.5 152452.5
2023 Q2       158557.3 158557.3 158557.3
2023 Q3       157059.3 157059.3 157059.3
2023 Q4       169740.8 169740.8 169740.8
2024 Q1       157249.6 157249.6 157249.6
2024 Q2       163595.8 163595.8 163595.8
2024 Q3       162449.2 162449.2 162449.2
2024 Q4       174351.5 174351.5 174351.5
```

3e. Apply the *accuracy()* function to compare performance measures of the following forecasting models for the *entire data set*: (1) regression model with *quadratic trend and seasonality*; (2) *two-level* model (with *AR(1)* model for residuals); (3) *ARIMA(1,1,1)(1,1,1)* model; (4) *auto ARIMA* model; and (5) *seasonal naïve* forecast for the entire data set. Present the accuracy measures in your report, compare them, and identify, using *MAPE* and *RMSE*, the best model to use for forecasting Walmart's revenue in Q1-Q4 of 2023 and 2024.

The accuracy measures for the 5 specified models (for the entire data set) are presented below.

Regression model with linear trend and seasonality

```
         ME    RMSE      MAE    MPE  MAPE  ACF1 Theil's U
Test set  0 4050.66 3358.498 -0.144 2.935 0.846     0.417
```

Two-level model (with *AR(1)* model for residuals)

```
              ME     RMSE      MAE   MPE  MAPE  ACF1 Theil's U
Test set 129.193 1868.012 1341.747 0.087 1.169 0.011     0.182
```

ARIMA model (1,1,1)(1,1,1)₄

```
               ME     RMSE      MAE   MPE  MAPE   ACF1 Theil's U
Test set -208.346 1839.248 1279.033 -0.23 1.068 -0.007     0.176
```

Auto ARIMA model (0,1,0)(1,1,0)₄

```
              ME     RMSE      MAE    MPE  MAPE   ACF1 Theil's U
Test set -44.653 2091.467 1547.659 -0.046 1.313 -0.047     0.203
```

Seasonal naïve forecast

```
               ME     RMSE      MAE   MPE  MAPE ACF1 Theil's U
Test set 4399.824 5599.183 4570.088 3.834 3.985  0.7     0.583
```

According to the accuracy measures, the lowest MAPE of 1.07% is for the *ARIMA* $(1,1,1)(1,1,1)_4$ model, which also has the lowest RMSE of 1839.25. Based on the superiority of MAPE and RMSE, we should select the *ARIMA* $(1,1,1)(1,1,1)_4$ model as the best model for forecasting in the 4 quarters of 2023-2024.