MSARIMA Data Analysis

Yumeng Wang

January 21, 2020

1 Exploratory Data Analysis

The dataset that I have chosen to work with is the Sales dataset, which consists of the number of sales each day for the air conditioning company Chill-E-AC from the beginning of 2015 through June 2019. We first study the dataset by plotting the raw data (Figure 1).

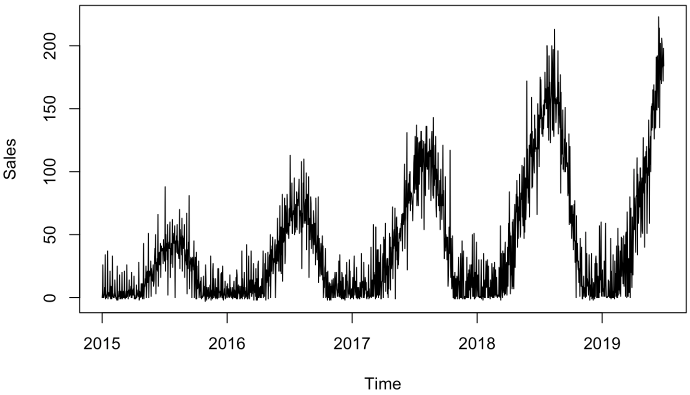

Figure 1: Sales of AC units for Chill-E-AC (from January 1, 2015 to Jun 30, 2019)

Each data point in the plot is the number of sales for a particular day, without any missing days. We observe a few negative values, which is valid as they indicate having more returns than sales for a particular day. Since the data includes a leap year (2016), I will remove the data point for February 29th in order to make the length of a year consistent for the purpose of differencing by year later on.

There are some features that we can observe from Figure 1:

- Here We can observe strong seasonality of roughly one year, with peaks around June and July each year. This is consistent with the Fourth of July holiday sale that Chill-E-AC runs every year.
- Overall, there is a general upward trend for sales through time.
- The plot also shows positive heteroscedasticity, meaning that the variability of sales changes over time with its mean.

With these features kept in mind, we can proceed to fitting models for the data.

2 Modeling a Deterministic Function of Time

The natural next step is to pursue stationarity in order to apply ARIMA models.

From the previous section, we concluded that there is heteroscedasticity with roughly linearly increasing variance, so I take the square root in order to stabilize variance. Because it can be noticed that the smallest value in all sales data is -3, I shift each value upwards by 3, so I can successfully take the square root. We will subtract by 3 after squaring back to recover the original scale. In addition, I take the difference of 365 days for seasonality. The resulting plot is shown below (Figure 2).

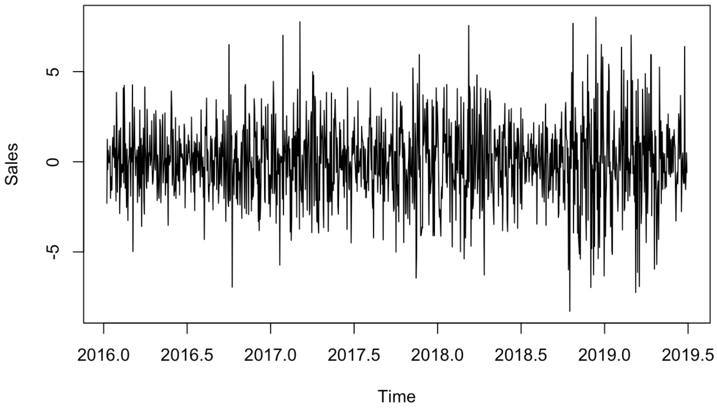

Figure 2: Differenced data of 365 days after variance stabilization

This seems to be generally stationary, which is good. This is used for Model 1 and 2 in the next section.

In addition to the previous method for pursuing stationarity, I introduced a monthly indicator model for another approach. A monthly indicator model is constructed by adding in 11 dummy features for each month to the dataset, which indicates which month each datapoint belongs in (If all indicators are 0, then the datapoint belongs in December). Then, I regress on sales and monthly indicators using a linear model to find a general trend for the data. (Figure: 3)

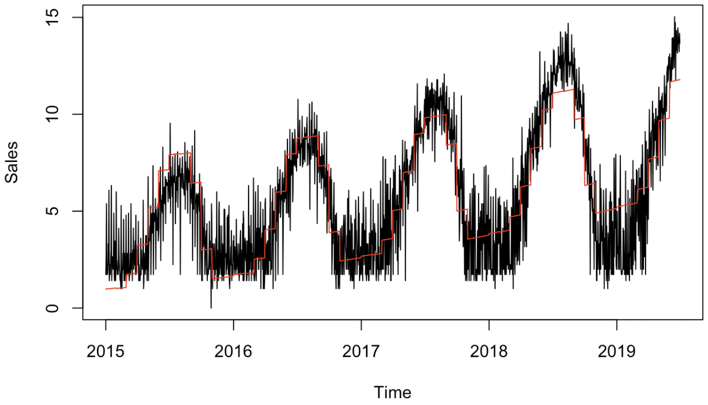

Figure 3: Variance-stabilized sales time series with a fitted linear model

The trend seems to fit quite well with the data, so I take the residuals (Figure 4). The results seems to achieve stationarity as well, so I would use it for Model 3 in the next section.

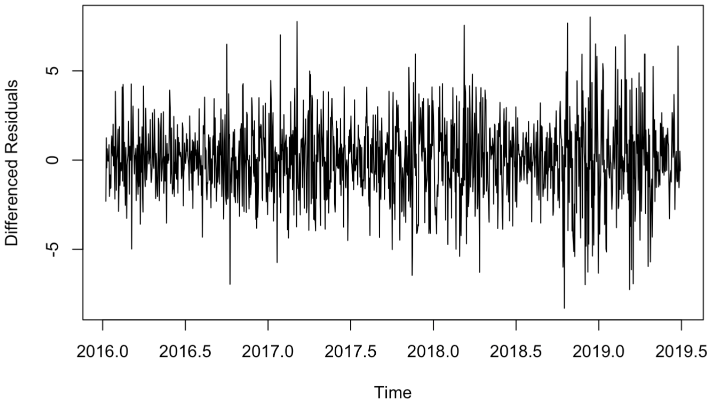

Figure 4: Residuals of the linear regression model with monthly indicator features

3 ARIMA Model Selection

In this section, I derive three SARIMA models to fit the current data in order to make future predictions. I will first introduce three models and then compare them in terms of both in-sample and out-of-sample diagnostics: Ljung-Box test results, AIC/AICc/BIC, and cross validation errors.

3.1 Model 1 Parameter Selection and Diagnostics

I first look at ACF and PACF of our first stationary data by taking the yearly difference (Figure 5)

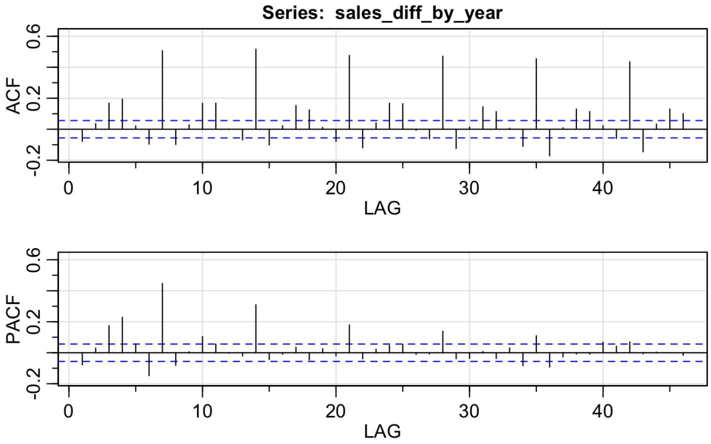

Figure 5: ACF and PACF of Sales differenced by year

From the plots generated above, I decide to use a seasonal multiplicative ARIMA model:

- In the ACF plot, there is a constant pattern within every 7 lags with a periodic spike, along with corresponding spikes in the PACF plot with exponential decay. So I take seasonality S = 7.
- The ACF spikes have same height, so I choose D = 1 and Q = 1 as seasonal ARIMA components.
- For the SAR(P) parameter, from the above analysis, I use P = 0.
- Finally, for the ARIMA part of the model, in the ACF plot, the smaller spikes before the lag 7 ends at lag 4, repeating over periods of seasonality, which is 7, so I choose to set q = 4.

In summary, I use model 1 as SARIMA(0,0,4) × $(0, 1, 1)_7$ on Sales differenced by year (Figure 6).

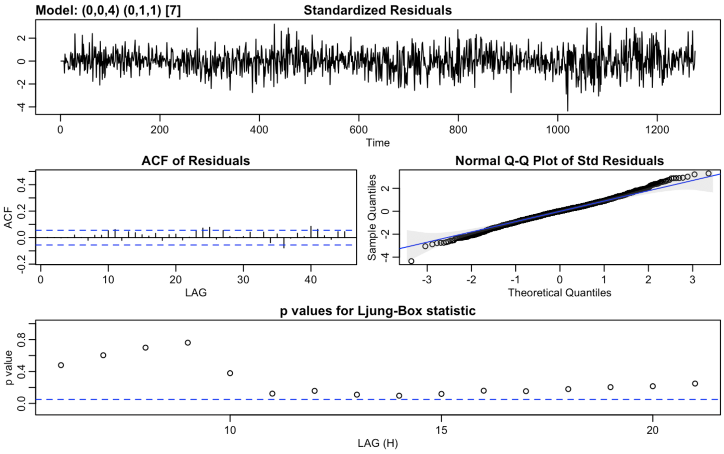

Figure 6: Diagnostics of Model 1

The diagnostic plot from this model also confirms that it is a good fit. We see stationary and normally distributed residuals, as well as high p-values for Ljung-Box test, thus showing that this model fits our data well.

3.2 Model 2 Parameter Selection and Diagnostics

Here, I take the a second difference of a day in addition to differencing by year. The ACF and PACF plots are shown below (Figure 7):

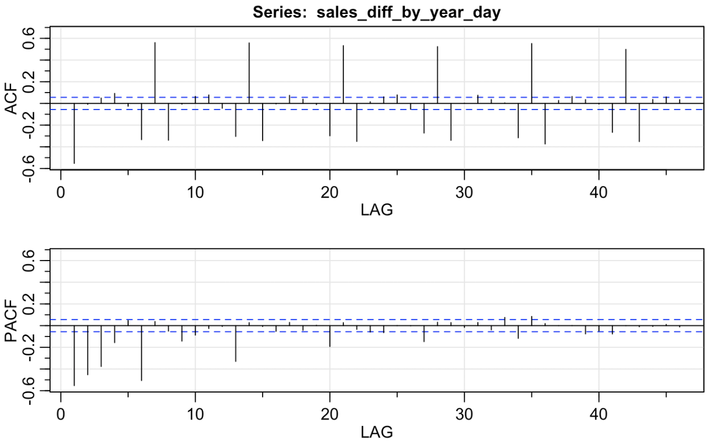

Figure 7: ACF and PACF of Twice-differenced (first by year, then by first difference)

Parameter selection:

- Since I take the first difference after differencing by year, I will naturally take d = 1 to achieve the ACF and PACF plots as above
- We see seasonality with lag 7 in the ACF plot, so we set S = 7 and first seasonal difference D = 1
- Considering the SMA(Q) component for this model, I used Q = 1 as we see at lag 7 (our seasonality), there is a high spike appearing in the ACF plot, as well as the exponential decay in seasonal spikes of lag 7 and its multiples in the PACF plot.
- As to the simple ARIMA model, again, I have discovered a pattern around the highest spike in every 7 lags in ACF plot: there are two shorter spikes at both side of the higher one, as well as a spike at lag 1, indicating that q = 1 would be reasonable.

In summary, our Model 2 is SARIMA(0,1,1)$\times(0,1,1)_7$ on Sales differenced by year and first difference.

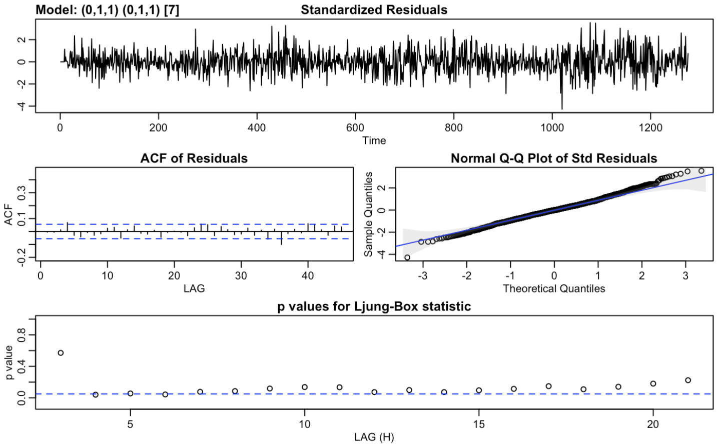

Figure 8: Diagnostics of Model 2

From diagnostics, we can see evidence that these model fits well. The residuals look stationary from the plot and ACF. Most of the ACF's fall in between the blue lines. The Q-Q plot shows that the distribution of residuals is mostly normal, except for one outlier in the data. The p-values for Ljung-Box Test Statistic are all significantly above the blue band, and thus we don't reject the null hypothesis.

3.3 Model 3 Parameter Selection and Diagnostics

For our third model, I used the ACF and PACF plots from differenced (365 days) residual for our reference of model estimation based on regression. (Figure 9):

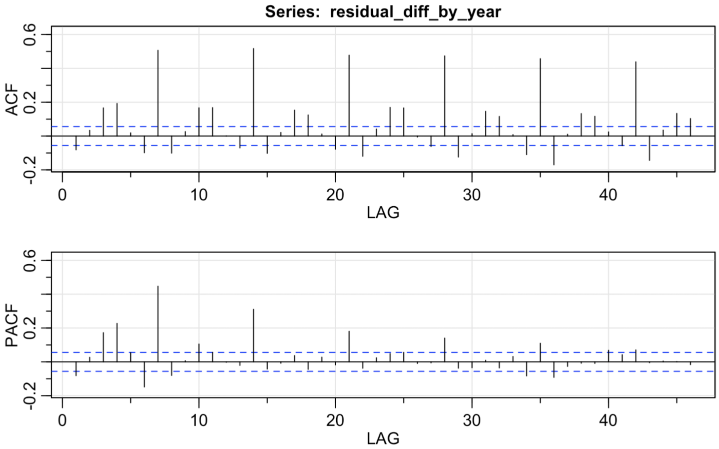

Figure 9: ACF and PACF of residuals differenced by year

Parameter Selection:

- Similar to the the ACF and PACF plot from model 1, the analysis follows.
- In the ACF plot, we see a cycles within every 7 lags, along with corresponding decreasing spikes in the PACF plot. So I take seasonality S = 7.
- The spikes in ACF are similar in height, so I choose D = 1 and Q = 1 as seasonal ARIMA components.
- For the SAR(P) parameter, from the above analysis, I use P = 0.
- For ARIMA, since the autocorrelation increases until lag 4 before lag 7 in the ACF plot, repeating over periods 7, so I choose to set q = 4.

In summary, Model 3 is SARIMA(0,0,4) $\times$ $(0, 1, 1)_7$ on the residuals.

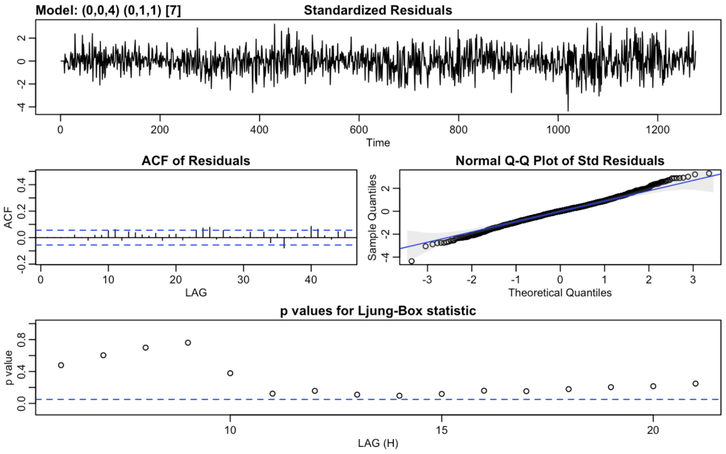

Figure 10: Diagnostics of Model 3

The diagnostic plot shows all p-values above the blue band. Residuals look stationary and normally distributed, but there's also a slight ACF spike at lag 36 similar to the previous two models.

3.4 Model Comparison

Model	AIC	AICc	BIC	CV
Model 1	3.922599	3.922651	3.950852	522.2843
Model 2	3.920373	3.92038	3.932488	572.1563
Model 3	3.922599	3.922651	3.950852	1580.0301

The IC's are very close, but model 1 has the non-trivially smallest cross validation error, along with nice p-values for Ljung-Box statistic, so I choose model 1 for predicting values in the next section.

4 Results

4.1 Estimation of model parameters

Table 1: These are our parameter estimates and corresponding standard errors for each model.

	θ_1	θ_2	θ_3	θ_4	Θ_1	μ
Estimate	0.0386	0.0393	0.0544	0.1089	-0.7889	0.0017
SE	0.0279	0.0280	0.0284	0.0279	0.0193	0.0018

From the previous section, we selected model 1 as our final model. We will use it to generate estimation of sales for the next 10 days in July. The ARIMA model is defined as follows: $(Y_t = \nabla_{365} X_t)$

$$\nabla_7(Y_t - 0.0017) = \Theta(B)\theta(B)Z_t \tag{1}$$

$$\text{Where: } \Theta(B) = 1 - 0.7889B^7 \tag{2}$$

$$\theta(B) = 1 + 0.0386B + 0.0393B^2 + 0.0544B^3 + 0.1089B^4 \tag{3}$$

4.2 Prediction

To predict sales for the next 10 days, I use the built-in function in R for SARIMA model prediction using our chosen Model 1, add back the differencing and variance stabilizing transform I did with our original data. The new values of our prediction are plotted on the original time series (Figure 11):

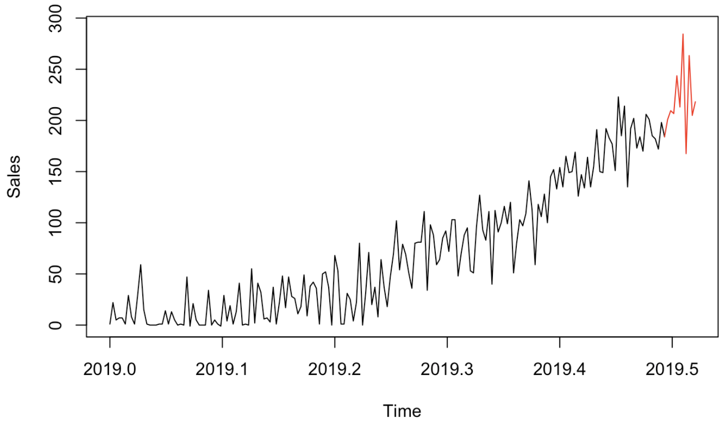

Figure 11: Predicted Sales for First 10 Days of July 2019 (days in a year is represented as fractions: e.g. Jun 30,2019 ≈ 2019.5)

The predictions are consistent with the peak summer in previous years in the original data, and we observe an increase in sales on July 4th, 2019, which is when holiday sales happens.

Although my predicted values fit the general trend that we observed exploratory analysis, there could be uncertainty due events that happen within those 10 days that are unique to the year of 2019. Also, we observe an unexpected spike on July 6th, which could be the lasting effect of the holiday season sale.