Example Project

Jared Fisher

12/8/2020

1 Executive Summary

Chill-E-AC's sales of air conditioners is expected to be very high in July 2019, which is attributable to the high annual growth and strong seasonal pattern of high sales in the summer. The Fourth of July should also contribute to especially high sales. According to our parametric model with ARMA(1,1)x(1,0)[7] noise, the increasing sales seen in early summer will start to level off, but at this peak, Chill-E-AC should see some of the highest selling days in its history. [Note to students: this document contains comments to you inside of square braces [x] like this comment is, but they're not red to avoid distracting from the overall look.]

2 Exploratory Data Analysis

Air conditioner sales for Chill-E-AC (hereafter referred to simply as "sales") have been growing annually, as seen below in the left panel of Figure 1. There is a strong seasonal pattern: sales spike in the summer and are almost nonexistent in the winter. Also of note is that the variance of sales in their offseasons appears to increase over time, and may imply the same issue during their sales seasons as well.

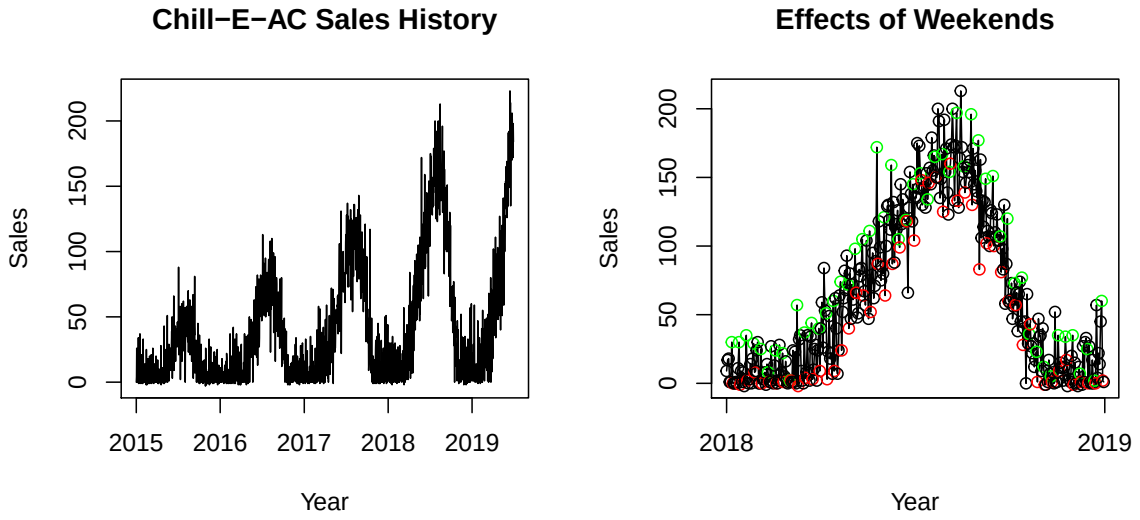

Figure 1: Daily air conditioner sales for Chill-E-AC. In the right panel, the green and red circles denote Saturdays and Sundays respectively.

When looking closer, there are also a few peculiar days in the dataset. The right panel of Figure 1 shows the effects of different days of the week. Saturdays have higher than average sales and Sundays are lower than average. Also, the company has a big discount every the Fourth of July, so sales are naturally higher on this holiday every year [note to students: this the kids of detail that is in the README file]. Lastly, 2016 is a leap year, such that the existence of February 29 will make some modeling methods more difficult to employ.

3 Models Considered

To model the natural signal in this data, both a parametric model and a differencing approach are used. Both of these models of the signal will be complimented with ARMA models for the remaining noise.

3.1 Parametric Signal Model

First, a parametric model is considered. To create a sinusoid that increases in amplitude every year but is flat during the offseason, a sinusoid with period 365.25 is interacted with time and with an indicator for the months of the sales season, April to October. Additionally, indicators for day of the week and the Fourth of July are added. This deterministic signal model is detailed in Equation (1) below, where X_t is the additive noise term.

$$\begin{aligned}\text{Sales}_t = &\beta_0 + \beta_1 t + \beta_2 I_{\text{season}_t} + \beta_3 t I_{\text{season}_t} + \beta_4 I_{\text{July4}_t} + \sum_{j=1}^{6} \beta_{4+j} I_{\text{weekday}_{jt}} + \beta_{11} I_{\text{season}_t} cos\left(\frac{2\pi t}{365.25}\right) \\ &+ \beta_{12} I_{\text{season}_t} sin\left(\frac{2\pi t}{365.25}\right) + \beta_{13} t I_{\text{season}_t} cos\left(\frac{2\pi t}{365.25}\right) + \beta_{14} t I_{\text{season}_t} sin\left(\frac{2\pi t}{365.25}\right) + X_t \end{aligned} \quad (1)$$

Figure 2 presents the fit as well as the residuals, which appear reasonably stationary. Both plots focus in on the last two years of data in order to show the fine details.

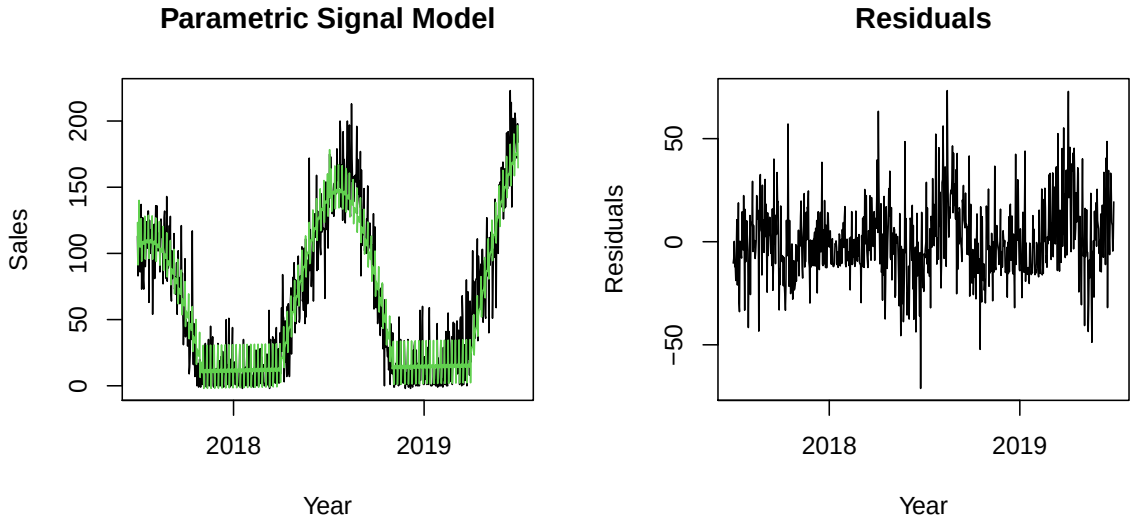

Figure 2: The parametric signal model. The left panel shows this model's fitted values in green, plotted atop the sales data in black. The right panel shows the residuals of this model.

The right panel of Figure 2 shows that there potentially is heteroscedasticity when comparing the summer sales season with the winter "offseason". [note to students: I don't address this here, as this dataset contains negative values, and we cannot take the log or sqrt. To cure the heteroscedasticity, we would either need a superior signal model or an advanced VST approach.]

3.1.1 Parametric Signal with ARMA(1,1)x(1,0)[7]

The ACF and PACF plots for the parametric model residuals are shown in Figure 3. The lags with the largest magnitude ACF values [note I didn't say "spikes" here, but formally described what we see] occur at lags 1 and 7. Furthermore, there are more significant lags in the ACF than in the PACF. These two observations lead to proposing p=P=1 as a potential fit, however trial and error showed that this shape is not well fit unless q=1 as well. Thus, ARMA(1,1)x(1,0)[7] is proposed, and this model implies the ACF and

PACF indicated by the red circles in Figure 3 which fit the general pattern of the sample autocorrelations. [Note to students: you could demonstrate fit with the Ljung Box plot, the ACF of ARMA's residuals, or even all of sarima's diagnostic plots. Up to you and the narrative you construct on your report.]

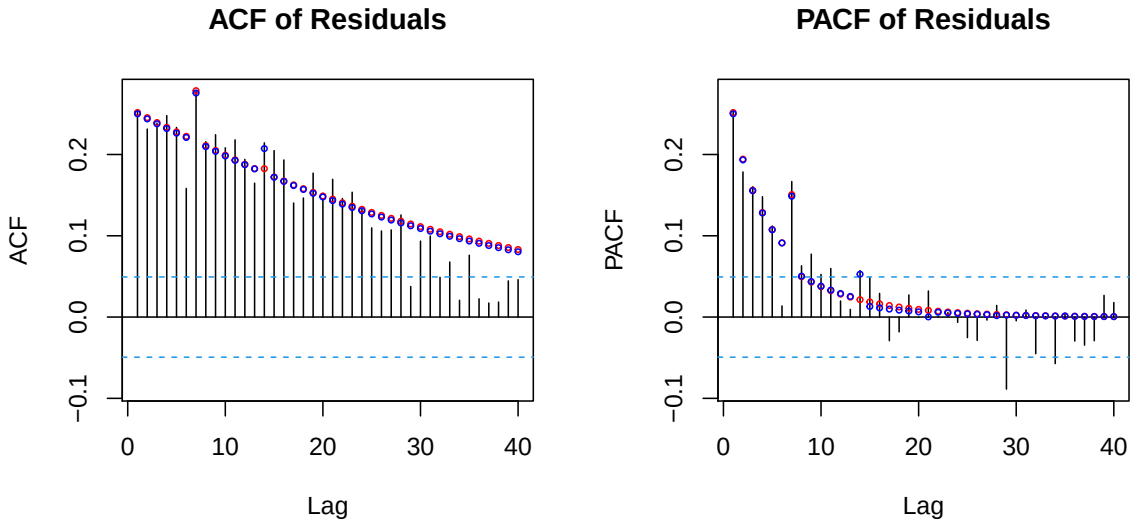

Figure 3: Autocorrelation function (ACF) and partial autocorrelation function (PACF) values for the parametric signal model's residuals. Red circles reflect the AR(1)xSAR(1)[7] model, while the blue circles reflect the ARMA(2,2) model.

3.1.2 Parametric Signal with ARMA(1,1)x(0,2)[7]

The R function auto.arima(), with the no differencing option specified, suggests ARMA(1,1)x(0,2)[7]. This seems plausible as the seasonal AR component (P=1) is replaced by a larger order of a seasonal MA component (Q=2). This model's ACF and PACF are included as blue points on Figure 3, which look like an incrementally better fit to the sample autocorrelations than the red circles from the first suggested ARMA model.

3.2 Differencing

As previously addressed, there are annual peaks in the summers, so lag-365 differencing will be helpful. Note that the leap day on 2/29/2016 will interfere with this, so everything before 3/1/2016 is trimmed off from the dataset when differencing. There are also weekly effects (high Saturday sales, low Sunday sales), so lag-7 differencing will also be beneficial, and as this is now twice differenced, any linear or quadratic trend will be accounted for. We see this is the case in the implied fitted values of this differencing approach, which are shown in Figure 4. Also shown is the time series of the differences, which appears stationary. [note to students: this is essentially the residual plot! And you could reasonably argue it's nonstationary due to heteroscedasticity, but again, we can't take the log of the raw data in this problem.]

3.2.1 Differencing with SMA(1)[7]

The sample ACF and PACF for these differences are shown in Figure 5. Significant values at every seventh lag of the PACF plot suggest that Q=1 and S=7. There are other lags that appear significant and similar to the q=Q=1 pattern, however, adding q=1 does not fit these points well, so the simpler model is chosen. The fit of this choice is shown in Figure 5 with red circles. These red circles show that this SMA(1)[7] specification fits well at each of the most-significant lags: PACF values at multiples of seven and the ACF at lag seven. [I could instead show all of, or a subset of, the sarima diagnostics...]

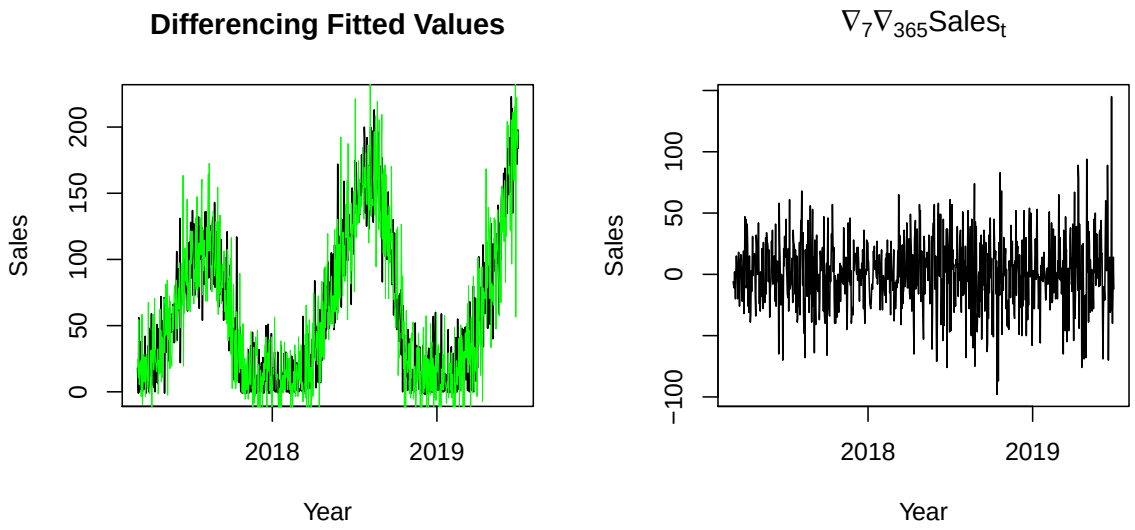

Figure 4: Diagnostics for differencing "signal model". The left panel shows the data in black and the fitted values in green The right plot shows the differences themselves, to be assessed for stationarity.

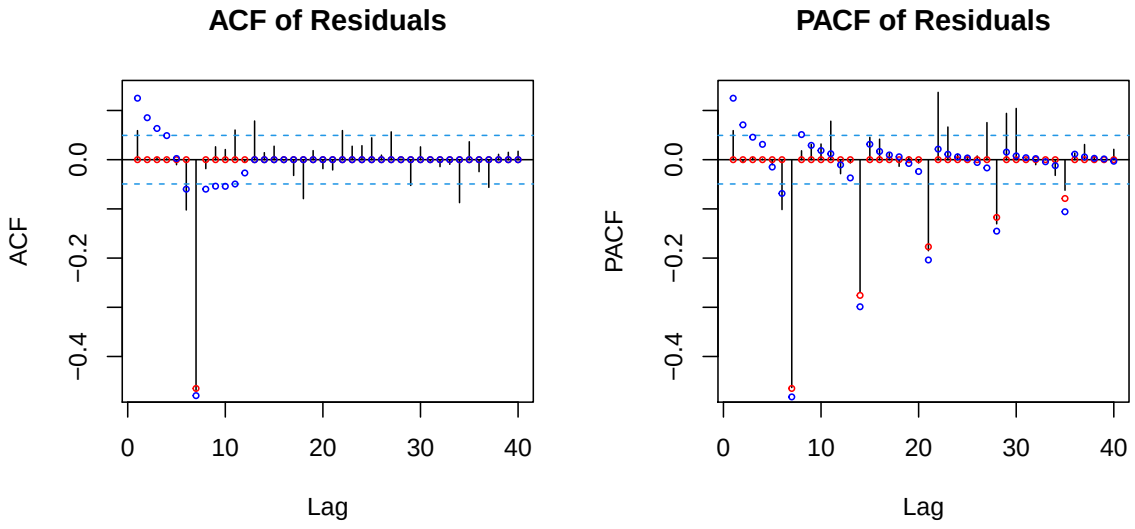

Figure 5: Autocorrelation function (ACF) and partial autocorrelation function (PACF) values for the differencing model. Red circles reflect the SMA(1)[7] model, while the blue circles reflect the ARMA(0,5)x(0,1)[7].

Table 1: Cross-validated out-of-sample root mean squared prediction error for the four models under consideration.

	RMSPE
Parametric Model + ARMA(1,1)x(1,0)[7]	21.32146
Parametric Model + ARMA(1,1)x(0,2)[7]	21.43294
Annual Differencing + Weekly Differencing + SMA(1)[7]	26.29462
Annual Differencing + Weekly Differencing + ARMA(0,5)x(0,1)[7]	26.94549

3.2.2 Differencing with ARMA(0,5)x(0,1)[7]

As with the previous signal model's second ARMA specification, this second noise model will be chosen by the R function auto.arima(), with the no differencing option specified. This automated procedure suggests ARMA(0,5)x(0,1)[7], i.e. q=5 and Q=1 for S=7, which is a more complex version of the SMA(1)[7] model in the previous subsubsection. [Students: note that auto.arima's default is to cap q at 5...] The ACF and PACF of this ARMA model are included as blue circles on Figure 3, which look similar but more sporadic compared to the same values from SMA(1)[7] in red circles. They are similar in that the PACF values at multiples of seven (and thus the ACF at lag seven) are accounted for. However, this more complex model also tries to account for some of the PACF values off of lag seven, though this seems to cause some of the ARMA ACF values (in blue) to stray away from the sample ACF values in black.

4 Model Comparison and Selection

These four model options are compared through time series cross validation. The nonoverlapping testing sets roll through the last 180 days in the data, 1/2/2019 through 6/30/2019, in 10 day segments. Thus there will be 180 forecasted points over these 10 windows. The training sets consist of all data that occur before the appropriate testing set. The models' forecasting performances will be compared through root-mean-square prediction error (RMSPE). The model with the lowest RMSPE will be chosen as the model for predicting sales in July.

Table 1 shows that the parametric model with ARMA(1,1)x(1,0)[7] has the lowest cross-validated forecast error, as measured by RMSPE, though the parametric model with ARMA(1,1)x(0,2)[7] is a close second. Thus the parametric model with ARMA(1,1)x(1,0)[7] is the chosen forecasting model. [Students: Your cross validation procedure may be different, and you may have AIC/BIC values that are comparable across models.]

5 Results

To forecast sales in July, a parametric model of time will be used. Let Sales_t be the air conditioner sales on day t with additive noise term X_t, as previously shown in Equation (1), which is restated below as part of Equation (2). X_t is a stationary process defined by ARMA(1,1)x(1,0)[7], where W_t is white noise with variance σ_W^2.

$$\begin{aligned}\text{Sales}_t =& \beta_0 + \beta_1 t + \beta_2 I_{\text{season}_t} + \beta_3 t I_{\text{season}_t} + \beta_4 I_{\text{July4}_t} + \sum_{j=1}^{6} \beta_{4+j} I_{\text{weekday}_{jt}} + \beta_{11} I_{\text{season}_t} cos\left(\frac{2\pi t}{365.25}\right) \\ &+ \beta_{12} I_{\text{season}_t} sin\left(\frac{2\pi t}{365.25}\right) + \beta_{13} t I_{\text{season}_t} cos\left(\frac{2\pi t}{365.25}\right) + \beta_{14} t I_{\text{season}_t} sin\left(\frac{2\pi t}{365.25}\right) + X_t \\ X_t =& \phi X_{t-1} + \Phi X_{t-7} - \phi\Phi X_{t-8} + W_t + \theta W_{t-1} \end{aligned} \tag{2}$$

There are several binary indicators in this model. I_{season_t} indicates if day t is in one of the months of the sales season, April to October. I_{July4_t} indicates if day t is the Fourth of July. $I_{\text{weekday}_{jt}}$ indicates if day t is the

jth day of the week. ϕ, Φ, θ, and all of the β's are coefficients that will be estimated in the next subsection. [note to students: I don't have a mean of X here, because the mean of the residuals of linear models is 0 (however, sarima estimates it anyway, so I suppose it be a bit more accurate to report it here). If you are differencing, the mean difference is probably not zero, and the same is true of smoothing's residuals.]

5.1 Estimation of model parameters

Estimates of the model parameters are given in Table 2 in Appendix 1. It is particularly interesting to note that the Fourth of July averages 34 more sales than it would as a normal day. For the days of the week, Friday is set as the baseline category, such that Saturdays average 19 more sales than Fridays, Sundays 13 fewer, and the other weekdays show little difference.

5.2 Prediction

Figure 6 shows the forecasted values of sales for the first ten days of July. The model predicts that the peak of sales season is near, as the expected sales are growing at a slower pace than in previous weeks. We do see high expected sales on both Thursday, July 4, and Saturday, July 6. Looking at the prediction intervals, either of these two days may set the company daily sales record. [Note to students: this interval contains the uncertainty of the ARMA model, not of the signal model as we assume m_t and s_t are deterministic, right?]

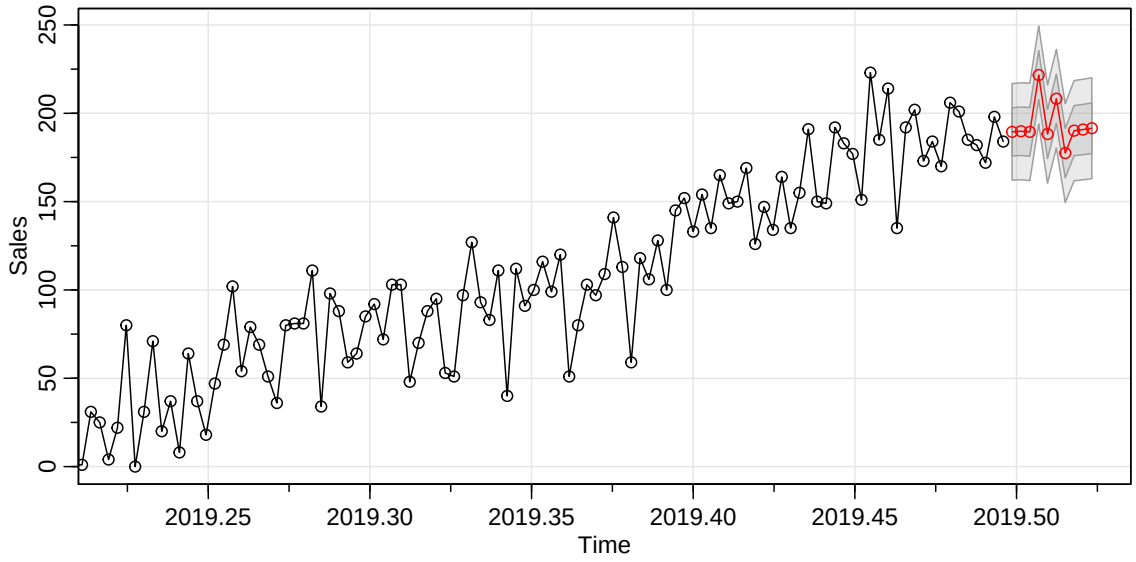

Figure 6: Forecasts of air conditioner sales for Chill-E-AC. The x-axis is time in years. The black points are the recent historical sales data. The red points are the forecasts for the first ten days in July 2019. The dark/light grey bands are the one/two standard error bands, representing 68%/95% prediction intervals, respectively. [Note to students: the x-axis would ideally be dates, but currently it is time in whole years, which is better than just integers "t". Sarima.for is proving to be a challenge to customize with their options; of course, I/you could just take their code and customize it correctly.]

6 Appendix 1 - Table of Parameter Estimates

Table 2: Estimates of the forecasting model parameters in Equation (2), with their standard errors (SE). [As a not-required bonus, I've included a brief description of what each of the coefficients are, because there are a lot of betas here... Also note that sarima() will give you the estimate of σ_W^2 (see code in my .Rmd), but I've yet to find its SE.] [Second note: if we don't assume independence, which lm regressions must, are these SE's of the β's very helpful?]

Parameter	Estimate	SE	Coefficient Description [not required]
β_0	2.738	1.450	Intercept
β_1	0.008	0.001	Time
β_2	-9.189	2.043	Sales season
β_3	0.033	0.002	Time $\times$ sales season interaction
β_4	34.387	7.620	Fourth of July
β_5	19.163	1.398	Saturday
β_6	-13.433	1.397	Sunday
β_7	-0.312	1.399	Monday
β_8	0.130	1.399	Tuesday
β_9	0.796	1.399	Wednesday
β_{10}	-1.588	1.397	Thursday
β_{11}	-17.001	2.556	Cosine $\times$ sales season interaction
β_{12}	-7.763	1.403	Sine $\times$ sales season interaction
β_{13}	-0.061	0.003	Cosine $\times$ time $\times$ sales season
β_{14}	-0.016	0.001	Sine $\times$ time $\times$ sales season interaction
ϕ	0.971	0.008	AR coefficient
θ	-0.873	0.016	MA coefficient
Φ	0.083	0.025	Seasonal AR coefficient
σ_W^2	186.521		Variance of White Noise