Time Series Analysis of Ice Cream Sales

Greg DeVore

August 19th, 2017

Overview

The purpose of this assignment is to explore the properties and methods associated with time series analysis in statistics. A time series is any quantity that changes with respect to time, such as sales figures for a company, signal data from a measurement device, or the price of a company stock. Often with time series, the ultimate goal is to make predictions for future values. For example, to forecast energy demands for a power company, or to know how much of a certain product to make to ensure there is enough on hand to meet demand, but not so much that product is going to waste.

In this assignment, we'll be analyzing ice cream sales over nearly 20 years. Our objectives are to first understand the behavior of the time series, and then to forecast demand over the next year. Specifically, we'll attempt to answer the following questions:

1. Is the time series stationary? That is, do the mean and variance of the data change over time?
2. Is there a significant seasonal component to the time series? That is, do sales change in a predictable manner at some regular interval?
3. After the time series has been decomposed into seasonal and trend components, what order ARMA model best fits the residual?
4. Once we create a forecast for ice cream demand, how do the predicted values and associated confidence intervals behave over time?

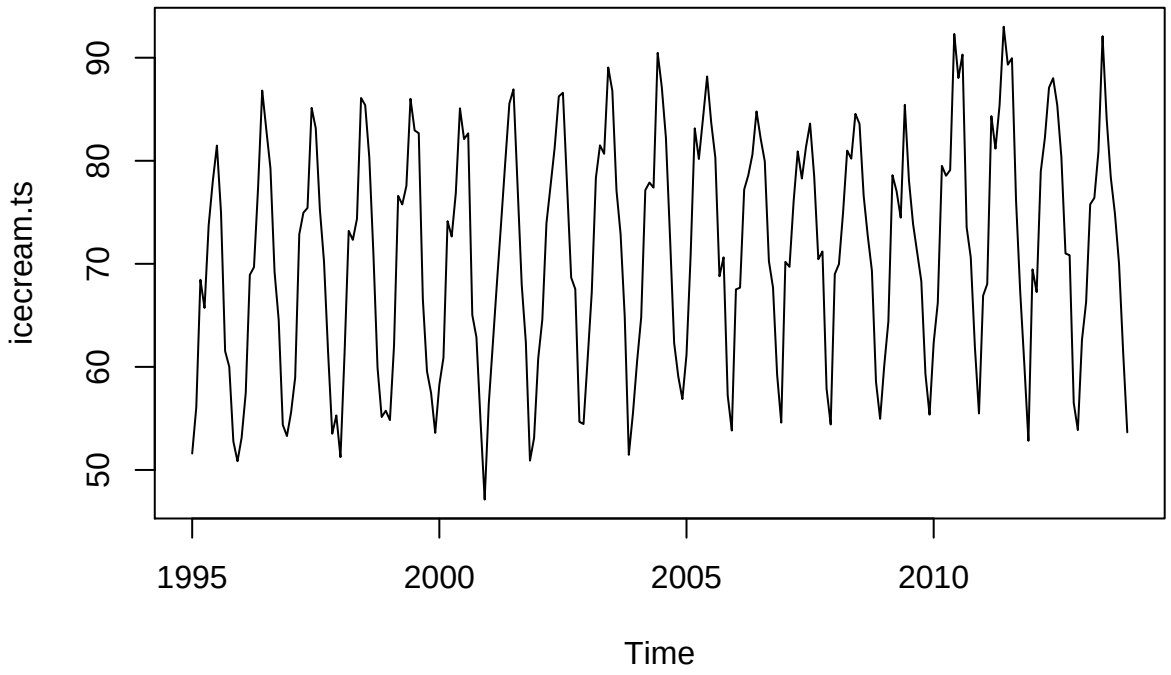
After completing our analysis and forecasting of the ice cream time series, it will be found that:

1. The ice cream sales data is not stationary, and in fact it exhibits strong periodic behavior.
2. The ice cream sales data shows a strong seasonal component.
3. The best fit to the residual data is an ARMA(3,1) model. That is, the order of the autoregressive (AR) model is 3, and the order of the moving average (MA) model is 1.
4. The confidence intervals for the forecast means are reasonably small, and don't appear to vary much over time.

Ice Cream Sales

Our data contains sales numbers from 1995-2013, with values reported every month. To begin, let's create a time series object from our ice cream sales data and plot it. Note that ice cream sales appear to be quite periodic, which makes sense given that more ice cream is probably sold in the summer than the winter. Also, sales appear to be generally increasing over time.

```
# Create time series object from ice cream sales
icecream.ts <- ts(dairy$Icecream.Prod, start = dairy$Year[1], frequency = 12)
# Plot time series
plot(icecream.ts)
```



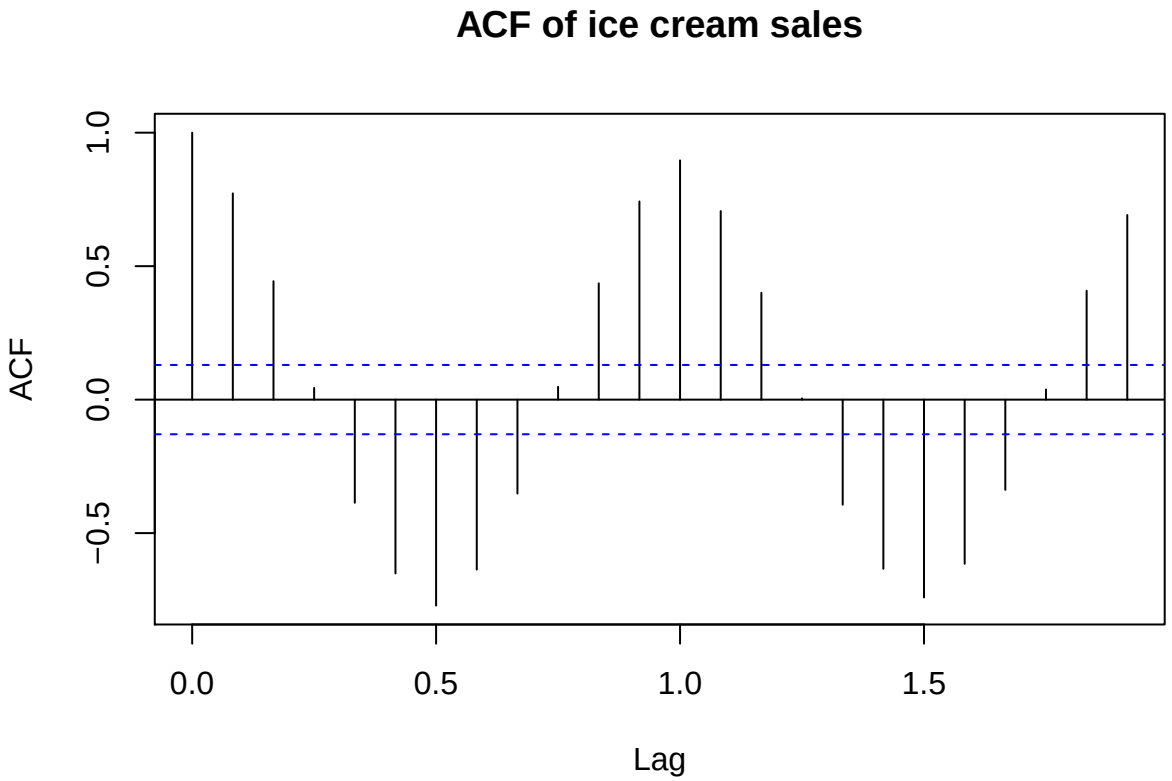
Stationarity

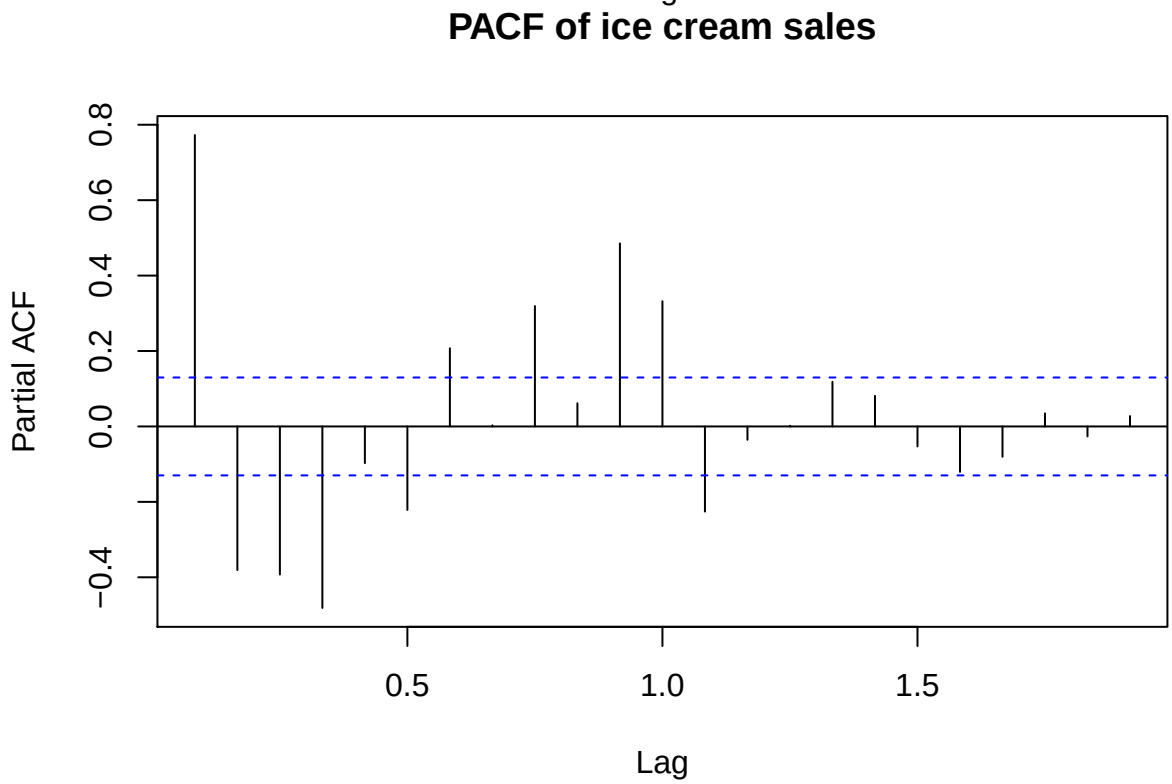
Once of the most important properties of a time series is whether or not it is stationary. Stationary refers to the fact that the mean, variance, and other properties of the time series are constant with time. We are going to forecast ice cream sales using an ARIMA model, which stands for Autoregressive Integrated Moving Average. Those terms will be defined later on, but the first thing to know about ARIMA models is that they require a time series to be stationary. The I term, which stands for Integrated, can help stabilize a time series when it is not stationary by differencing consecutive terms. Let's see if our ice cream data is in fact stationary.

Auto Correlation and Partial Auto Correlation

One of the main ways to determine whether or not a time series is stationary is through the use of auto correlation (ACF) and partial auto correlation (PACF) plots. Auto correlation measures the correlation of a time series with a lagged (offset) version of itself, and partial auto correlation measures correlation of a time series and its lags after removing the effects of correlations at smaller lag values. Let's take a look at the ACF and PACF plots for ice cream sales. In both plots, the blue dotted lines represent confidence intervals, and values that lie outside of them are considered to be significant.

```
# Plot ACF and PACF
plot.acf(icecream.ts, col = 'ice cream sales', is.df = FALSE)
```



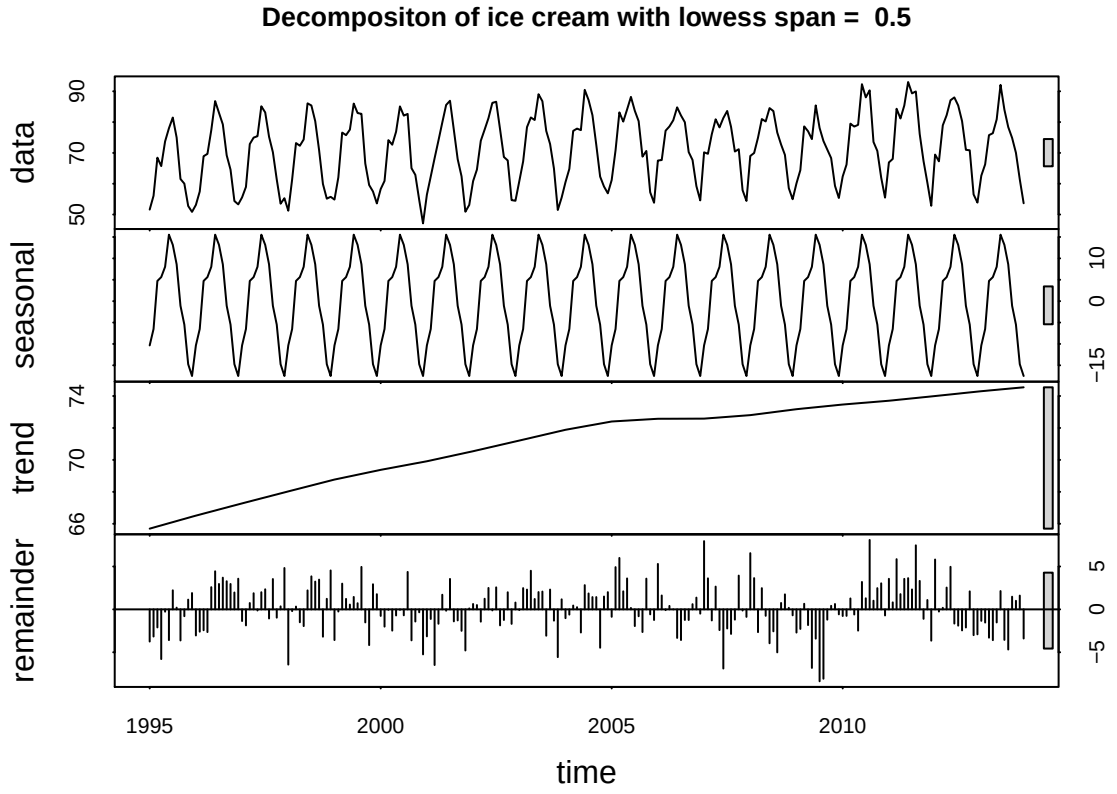


If the time series was stationary, we would expect a significant value in the ACF plot only when the lag is zero (since the series perfectly correlates with itself at zero lag). Here, not only are there significant values at many lags, but the behavior is strongly periodic, and does not decay at larger lags. For the PACF, we expect no significant values for a stationary time series, however there are quite a few in the plot above until the values decay at larger lags. Both of these plots indicate that the ice cream time series is not stationary.

Seasonality

After stationarity, the next area of exploration for our time series data is seasonality. Given that we are exploring ice cream sales, we would expect some seasonality in our data, as more ice cream is probably sold during the summer than the winter. To explore the seasonality, we can decompose our time series into three components using a seasonal, trend, and residual model (STL). Specifically, this decomposition will remove the trend (increase or decrease over time) using a LOESS regression model, and remove the seasonal component using a regression on periodic components. The remainder after both of these has been removed is the residual. Note that there are two primary types of STL models, additive and multiplicative. This refers to the relationship between the seasonal, trend, and residual components of the time series (whether they add together to create the series, or multiply). For our data, since the magnitude of the seasonality appears to be steady over time, we'll assume an additive relationship. A plot of all three components for our ice cream data is shown below.

```
icecream.stl <- ts.decomp(icecream.ts, Mult = FALSE, is.df = FALSE, span = 0.5)
```



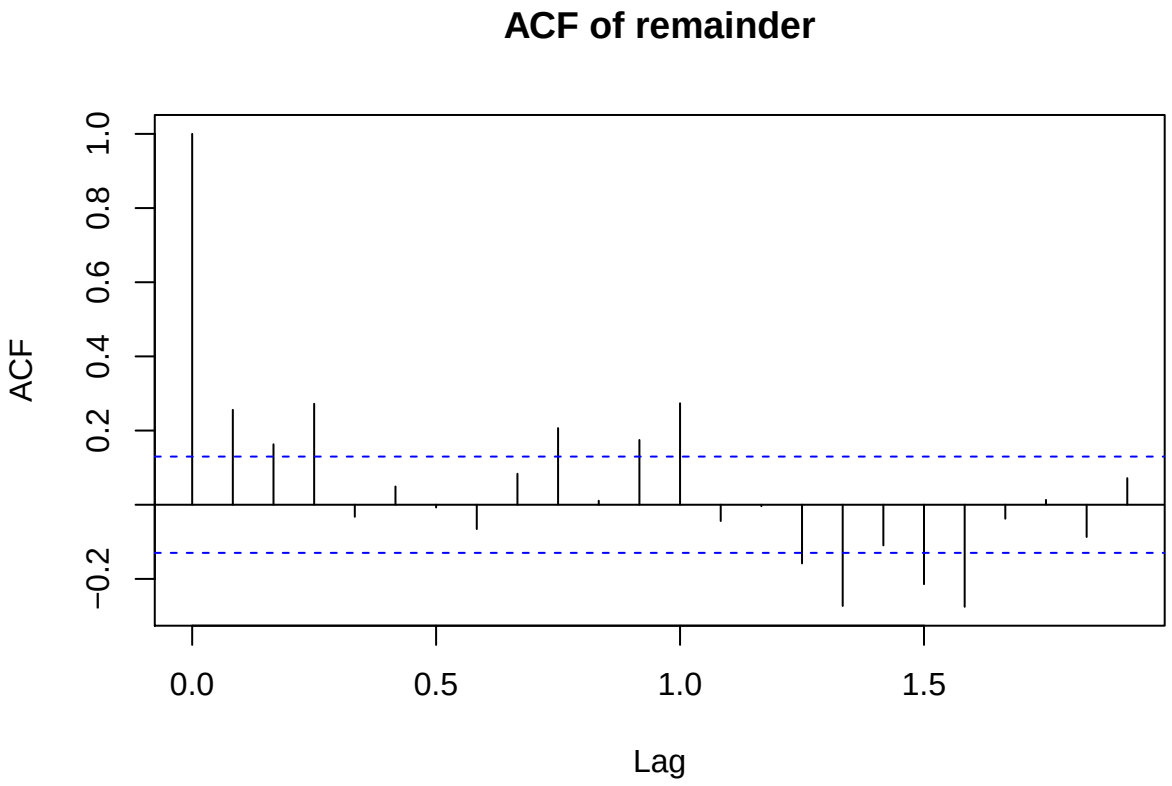
Some observations about the plots:

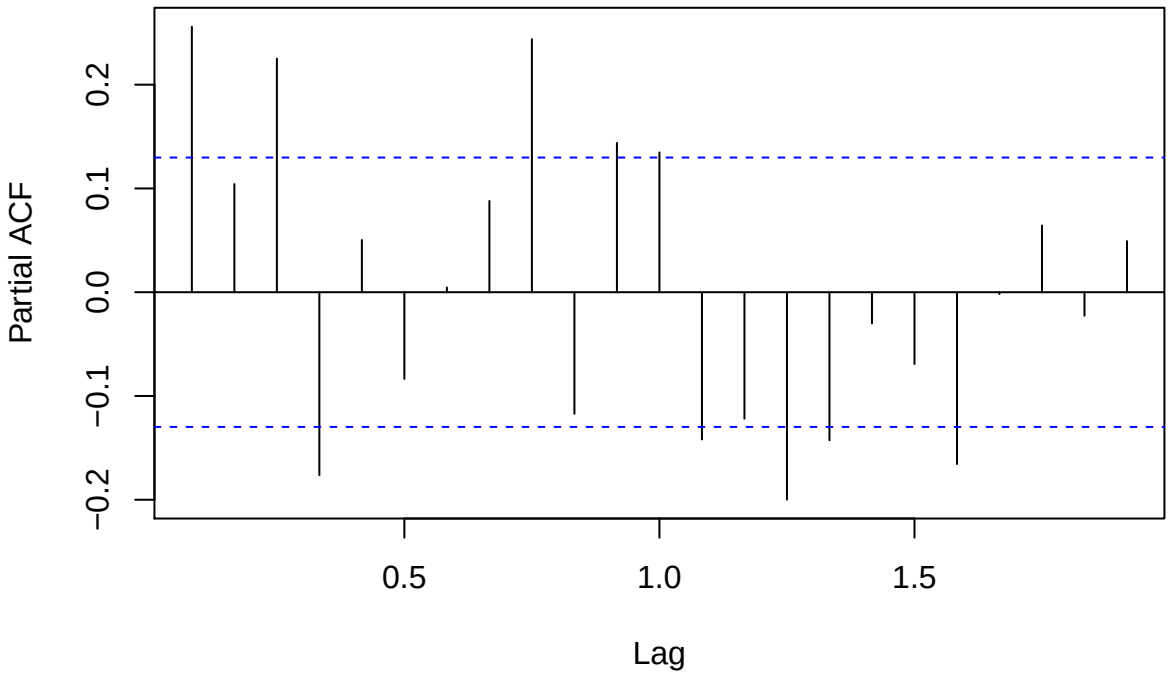
- The first plot shows the raw ice cream sales data.
- The second plot shows the seasonal component of the ice cream data. Note the sinusoidal like behavior showing peaks every summer (the data starts in January, where sales are lowest, and they peak in June/July).
- The third plot shows the trend component of the ice cream data. Note that the overall sales increase for the first 10 years, then flatten out for a bit before increasing again. If you look closely at the raw data, you can see this behavior.
- The fourth plot shows the residual data. Note that there is a non-trivial amount of data remaining, so some behavior of our time series was not captured in the seasonal and trend components. The remainder looks to be relatively steady in terms of amplitude, so perhaps the trend was fully captured by the decomposition. However, there still appears to be some seasonality to the remainder.

Fitting an ARMA model to the residual

Let's take a closer look at the residual left over after the trend and seasonal components were removed. Specifically, let's plot the ACF and PACF, which are shown below. For the ACF plot, note that the periodic behavior is greatly reduced when compared to the raw time series, but a bit remains. For the PACF plot, nearly all of the periodic behavior has been eliminated.

```
# Plot ACF and PACF of residual
plot.acf(icecream.stl, col = 'remainder', is.df = TRUE)
```





We can fit an ARMA(p,q) model (Autoregressive Moving Average) to the residual data in an attempt to capture the behavior. The AR stands for an autoregressive time series, which assumes that the current value in a time series is a linear combination of past values, with different weights applied to each previous value. The MA stands for a moving average time series, which assumes that the current value in a time series is a linear combination of past error terms, with different weights applied to each previous value. The p and q terms refer to the order (number of previous value or error terms used) of the AR, and MA models, respectively. Both AR and MA models only work for stationary time series. Since the residual data is still exhibiting some periodic behavior, it may not be possible to capture all of the behavior using an ARMA model.

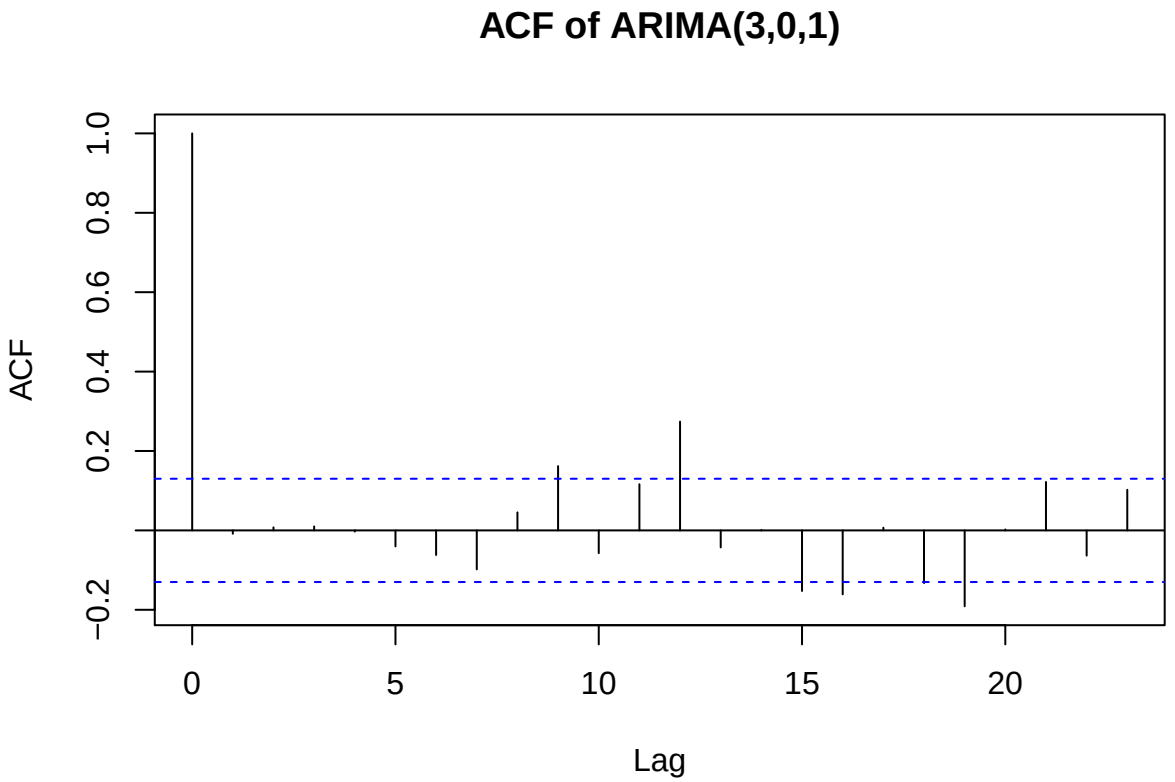
To determine the order (p and q) of the best ARMA model, we will look at the model which minimizes the Aikake Information Criterion (AIC). It turns out the best model to fit the residual of the STL decomposition is an ARMA(3,1) model.

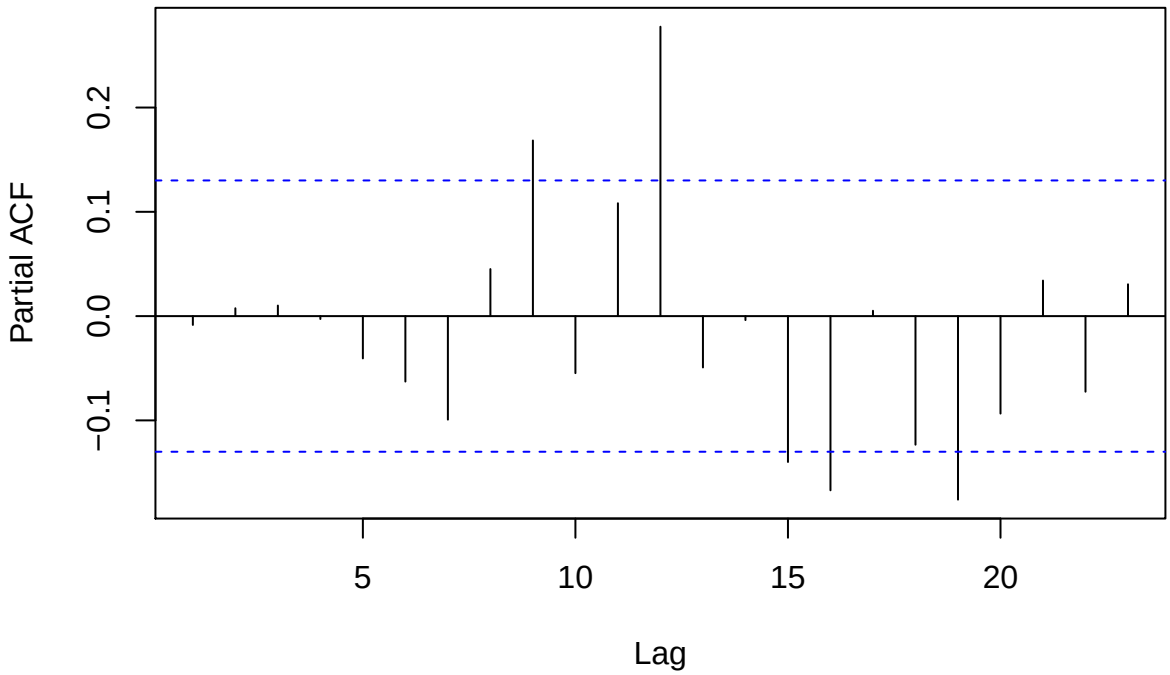
```
# AR = 3, MA = 1 are best for ARMA model, produces lowest AIC (1106)
# All coefficients are significant
icecream.ARMA <- ts.model(icecream.stl[, 3], order = c(3,0,1))
```

```
##
## Call:
## arima(x = ts, order = order, include.mean = FALSE)
##
## Coefficients:
##           ar1     ar2     ar3     ma1
##       -0.4617  0.2082  0.2890  0.7385
## s.e.   0.1182  0.0759  0.0671  0.1135
##
## sigma^2 estimated as 7.15:  log likelihood = -548.02,  aic = 1106.04
```

Note that all coefficients are significant, and the AIC is 1106. Let's plot the ACF and PACF of the residual from this model, to see if all of the behavior has been captured.

```r
# Plot ACF and PACF of residual
plot.acf(icecream.ARMA$resid[-1], col = 'ARIMA(3,0,1)', is.df = F)
```





The ACF has fewer significant values than the residual from the STL decomposition, and more closely resembles white noise than the former. The ARMA(3,1) model has done a good job of capturing some of the

rest of the behavior of the residual, despite the limitations of the ARMA model only working for stationary time series.

ARIMA Forecast

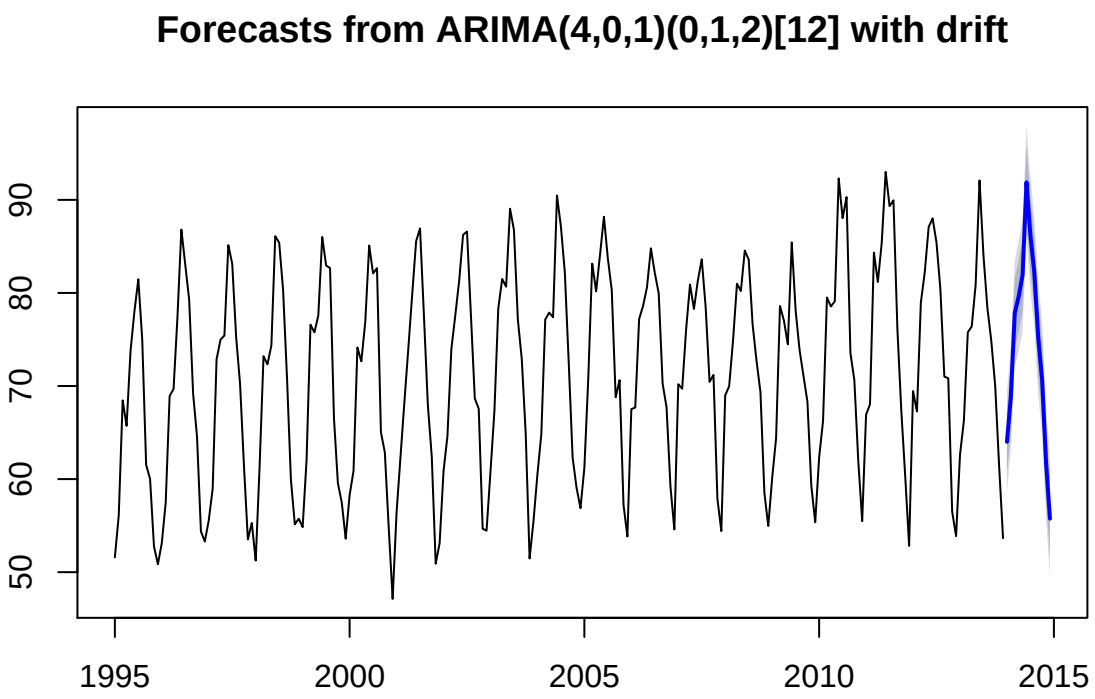
The final step in our analysis is to forecast ice cream production for the next 12 months. To start, we'll fit an ARIMA model to the base ice cream time series. As stated previously, the I in ARIMA stands for Integrated. It is the only piece that can handle a non-stationary time series, and is used to model the random walk component. Since our data is not stationary, we'll most likely need it to fully capture the behavior. To create our ARIMA model, let's use the *auto.arima* function in R. It will determine the optimum coefficients for our model for us. In addition to the AR and MA coefficients, the auto arima model adds additional seasonal terms to model seasonal trends and behavior. The full model is an ARIMA(p,d,q)(P,D,Q), where the second set of parenthesis contain the additional seasonal terms (seasonal autoregressive, differencing, and moving average). Let's create the full model, summarize the results, and plot the forecasted values for the ice cream sales.

```
# Create ARIMA fit
icecream.fit <- auto.arima(icecream.ts, max.p = 5, max.q = 5, max.P = 2, max.Q = 2,
                           max.order = 5, max.d = 2, max.D = 1, start.p = 0,
                           start.q = 0, start.P = 0, start.Q = 0)
# Forecast
icecream.forecast = forecast(icecream.fit, h=12)
summary(icecream.forecast)
```

```
##
## Forecast method: ARIMA(4,0,1)(0,1,2)[12] with drift
##
## Model Information:
## Series: icecream.ts
## ARIMA(4,0,1)(0,1,2)[12] with drift
##
## Coefficients:
##           ar1     ar2     ar3     ar4     ma1     sma1     sma2   drift
##       -0.6703  0.2777  0.3759  0.1641  0.9305  -0.5005  -0.2550  0.0379
## s.e.   0.0876  0.0773  0.0777  0.0770  0.0548   0.0725   0.0683  0.0112
##
## sigma^2 estimated as 7.63:  log likelihood=-526.93
## AIC=1071.87   AICc=1072.74   BIC=1102.25
##
## Error measures:
##                       ME     RMSE      MAE        MPE     MAPE      MASE
## Training set 0.08607263 2.638315 2.017612 0.04295444 2.845819 0.7309931
##                      ACF1
## Training set 0.009999409
##
## Forecasts:
##          Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
## Jan 2014       64.00784 60.46765 67.54803 58.59359 69.42210
## Feb 2014       68.82175 65.16363 72.47987 63.22713 74.41637
## Mar 2014       77.87455 74.19823 81.55088 72.25210 83.49701
## Apr 2014       79.62275 75.70930 83.53620 73.63764 85.60785
## May 2014       81.93482 78.01923 85.85040 75.94644 87.92319
## Jun 2014       91.88244 87.92496 95.83992 85.82999 97.93488
```

```
## Jul 2014       86.17298 82.20965 90.13630 80.11159 92.23436
## Aug 2014       82.19863 78.22515 86.17211 76.12171 88.27555
## Sep 2014       75.17528 71.20037 79.15019 69.09618 81.25438
## Oct 2014       70.60148 66.62239 74.58057 64.51598 76.68697
## Nov 2014       61.68990 57.71050 65.66929 55.60394 67.77585
## Dec 2014       55.74677 51.76601 59.72752 49.65873 61.83481
```

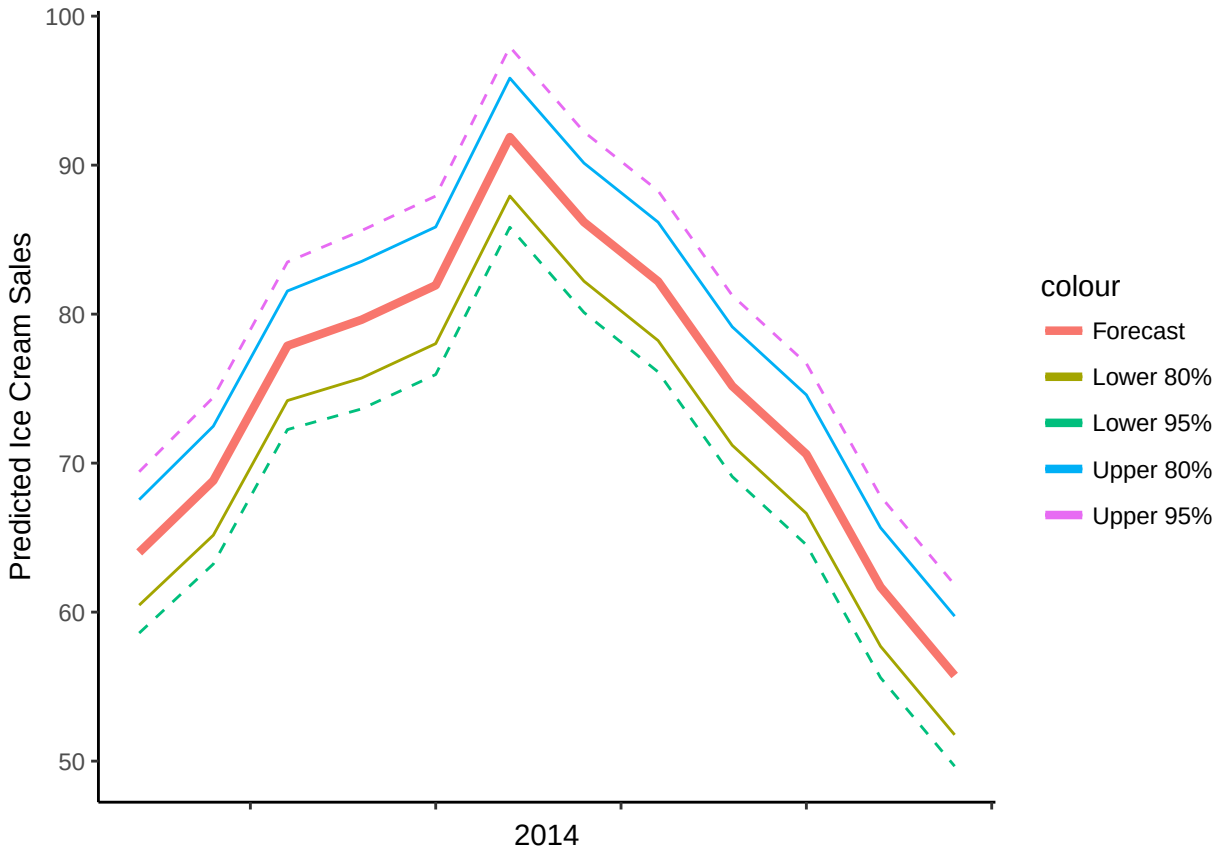
```
plot(icecream.forecast)
```



Note that all of the coefficients of the ARIMA model are significant, and that the AIC is 1071. The forecast looks reasonable given the prior years sales numbers, and from the numeric data for the forecast, the confidence intervals (shown in grey on the plot) are relatively tightly bound to the forecast means. Let's just plot the forecast means and confidence intervals to get a closer look.

```
# Compile CI as data frame for plotting
icecream.forecast.CI <- data.frame(predict = icecream.forecast$mean,
                                   L80 = icecream.forecast$lower[,1],
                                   L95 = icecream.forecast$lower[,2],
                                   U80 = icecream.forecast$upper[,1],
                                   U95 = icecream.forecast$upper[,2])
# Plot forecast + CI
ggplot(icecream.forecast.CI, aes(x = 1:12, y = predict, color = 'Forecast')) +
  geom_line(size = 1.5) +
  geom_line(aes(x = 1:12, y = L80, color = 'Lower 80%')) +
  geom_line(aes(x = 1:12, y = L95, color = 'Lower 95%'), linetype = 2) +
  geom_line(aes(x = 1:12, y = U80, color = 'Upper 80%')) +
  geom_line(aes(x = 1:12, y = U95, color = 'Upper 95%'), linetype = 2) +
  theme_classic() + theme(axis.text.x=element_blank())  +
  xlab('2014') + ylab('Predicted Ice Cream Sales')
```

```
## Don't know how to automatically pick scale for object of type ts. Defaulting to continuous.
```



By plotting the forecast means separately from the rest of the sales data, we can get a closer look and confirm that the confidence intervals are reasonably small compared to the forecast means. Also, the confidence intervals do not seem to drift towards or away from the forecast as time evolves. This confirms what we saw in the numeric data. Overall, the forecast appears to do a good job in predicting future values for ice cream sales.

Summary and Conclusion

The purpose of this assignment was to explore the properties and methods associated with time series analysis in statistics. We did this by looking at ice cream sales data, and our first goal was to understand the behavior of the time series. We explored the stationarity and seasonality of the data, and decomposed the time series using an STL decomposition. After that, we attempted to fit an ARMA model to the residual from the STL decomposition, before moving on to our final goal of forecasting demand. Once we created a forecast for ice cream demand, we looked at the forecast means and associated confidence intervals, and how they behaved over time.

To summarize our results, after completing our analysis and forecasting of the time series, we found that:

1. The ice cream sales data was not stationary, and in fact exhibited strong periodic behavior.
2. In addition, the ice cream sales data showed a strong seasonal component.
3. The best fit to the residual data from the STL decomposition was an ARMA(3,1) model. It produced the lowest AIC, and all of the coefficients were significant.
4. After fitting a seasonal ARIMA model to the data in order to forecast, we found that the confidence intervals for the forecast means were reasonably small, and didn't appear to vary much over time.