Math 155 Final Report

Alfredo Gomez and Alfredo Gomez

May 5, 2021

Abstract

We describe a time series model for the total electricity consumption of the United States from 2008 through 2018. We found that that an $\text{ARIMA}(0,1,1) \times \text{ARIMA}(1,1,1)_{12}$ model best fits the data and passes the runs, Shapiro-Wilk, and Box-Ljung tests. Applying the forecast of the model to 2019 through the present, we see that the actual data falls within the 95th-percent confidence band of the forecast. We take similar steps to model the electricity consumption of the commercial sector over the same time period and reach a similar $\text{ARIMA}(0,1,1) \times \text{ARIMA}(1,1,1)_{12}$ process. However the model no longer passes the Shapiro-Wilk test and the actual electricity usage in March and April 2020 falls outside of the models forecast.

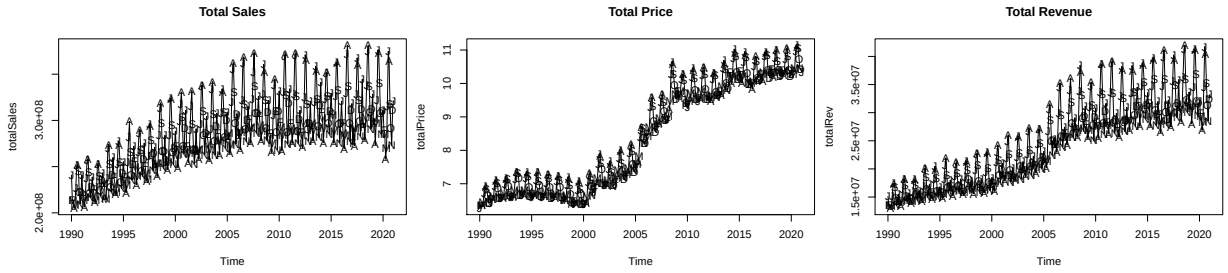
Introduction

Dataset

Our dataset includes the electricity consumption from 1990 to the present.[1] The data is broken down by state (including DC and Puerto Rico) and energy sector. The energy sectors are residential, commercial, industrial and transportation/other (the other sector is recorded from 1990 - 2002 and the transportation sector is recorded from 2003 to the present). The dataset measures the electricity consumption in four different ways: revenue in thousands of dollars, sales in MWh, price in cents/kWh, and number of customers (the number of customers is present in the dataset only after 2007).

Data Exploration

Graphs of the revenue, price, and sales of total electricity in the US are given below.

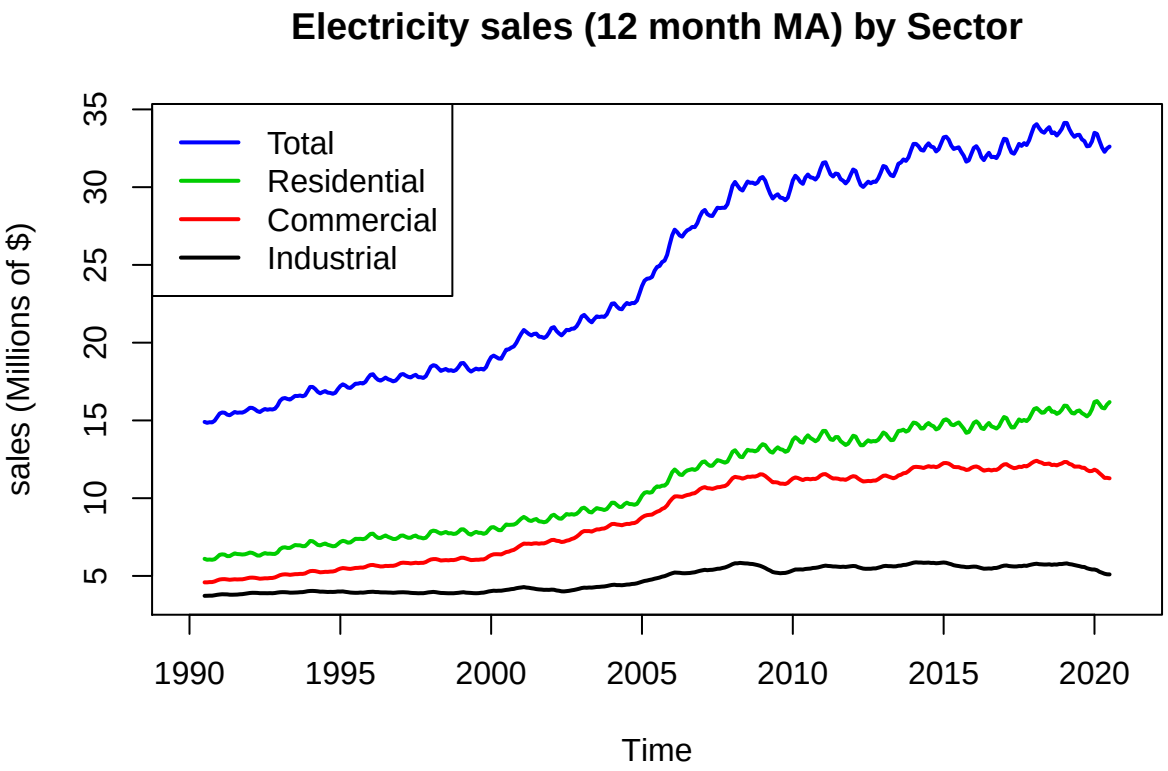


Note that all three graphs exhibit seasonal patterns but also long term trends. In particular, each decade of the data seems to behave differently. The 90s are characterized by growing usage and a cheap price of electricity. The 2000s are characterized by slower growth in the use of electricity and rising prices. Finally the 2010s are characterized by flat electricity usage and slowly rising prices. Another import trend in the sales and revenue graphs is a heteroscedastic increase in variance. Since we are interested in modeling the total revenue with predictive power, heteroscedasticity is a problem. While this can sometimes be addressed via power transformations the easiest thing in our case is to chop the dataset. We use only the data from

[1] The dataset is maintained by the US Energy Information Administration and freely available at https://www.eia.gov/electricity/data.php as form EIA-861M.

2008 onwards which we visually observed to be where the variance of the price of electricity began to stabilize. Since similar issues of heteroscedasticity appear across all of our available data, we figured this may due to some external factor changing the underlying process, and thus beleive we are justified in this decision.

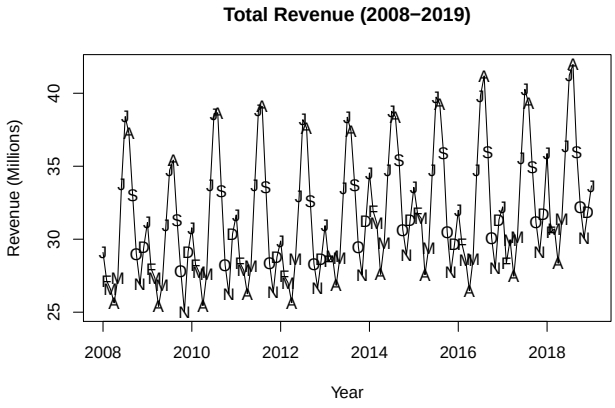
We can also see the electricity use breakdown by sector. To make the trends clearer over time, we show a 12 month moving average. Note that the residential, commercial, and industrial sectors together is about 97% of the total revenue before 2003 and more than 97% of the total revenue after 2003, the remainder falls in the other or transportation sector, not shown on this chart.



Interestingly, the 3 sectors all show their own trends. The industrial sector has little seasonal variation but is strongly impacted by the business cycle with revenues dropping significantly in great recession of 2008 - 2009 and during the lockdowns from COVID-19.

Total Revenue Data

The first series we will examine is the total revenue time series.



As noted earlier, due tot he odd behavior present throughout the data, we tightened our scope to only

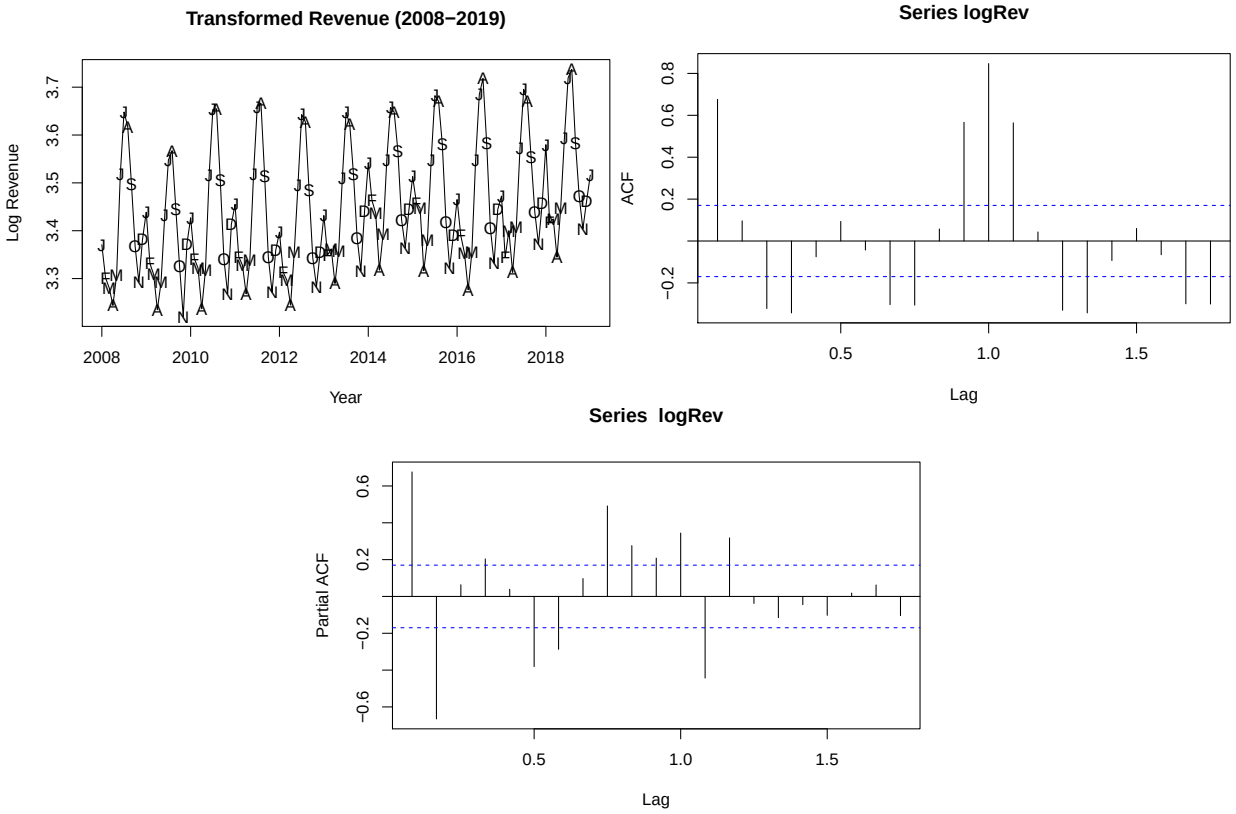
consider data past 2008. Additionally, we held out data past 2019 for the purposes of later forecasting. This resulted in a total of 132 data points. An initial observation we had was that there appeared to be a strong seasonal trend present in our data. The summer month seemed to exhibit relatively high values compared to the winter months. Additionally there appeared to be a slight upward trend in our data.

Model Specification

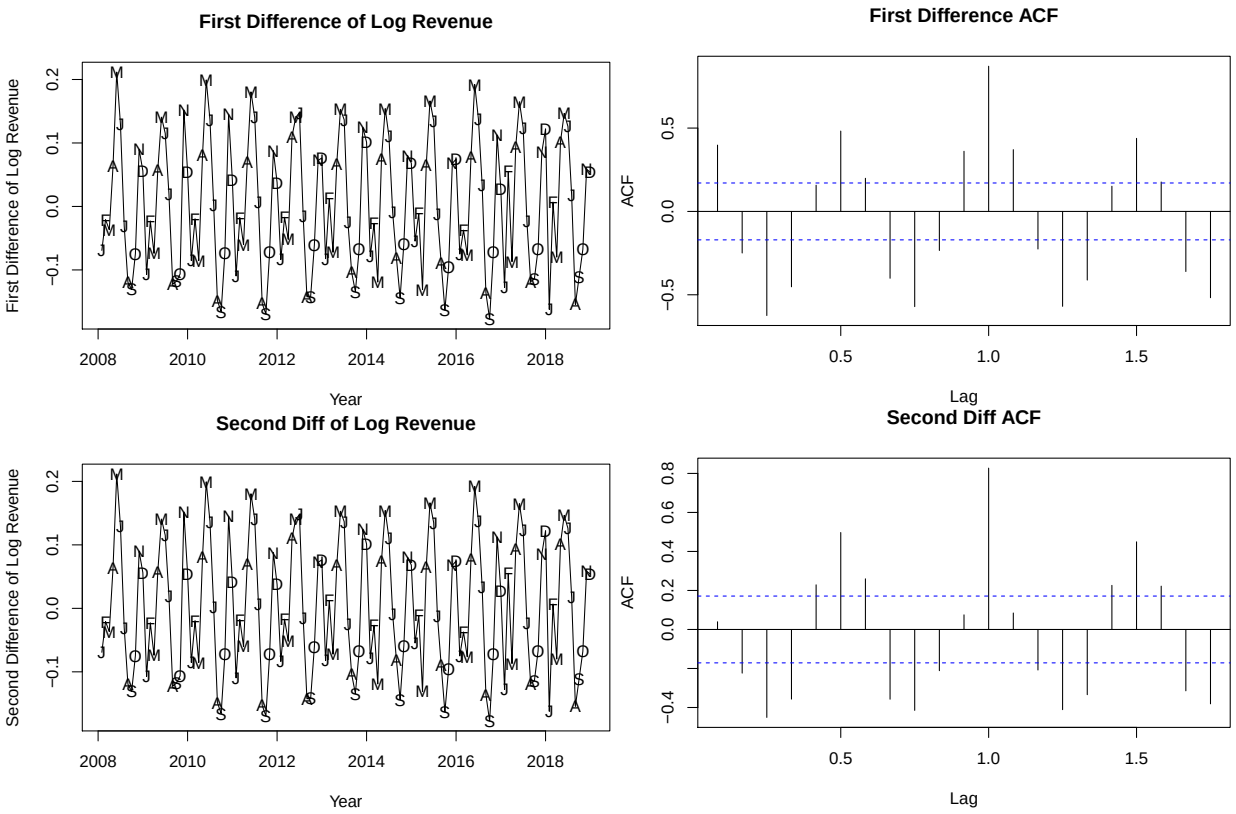
Our visual intuition is further confirmed by plotting the ACF and PACF of our data.

Plotting the ACF and PACF allows us to identify any autocorrelation present in our data, thus helping us determine an ARMA model to fit to our data. From the ACF our data is highly correlated at lag 1.0 (12 months), thus implying that our model will need a 12 month seasonal AR component. This is further exemplified by the PACF graph.

Our next step is to determine if our data needs to transformed. We do this by running a Box-Cox transformation test.

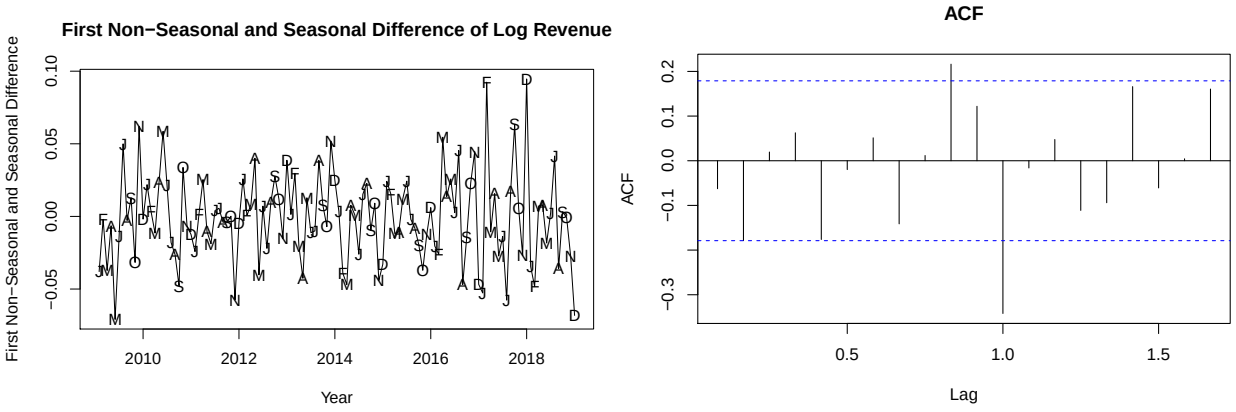


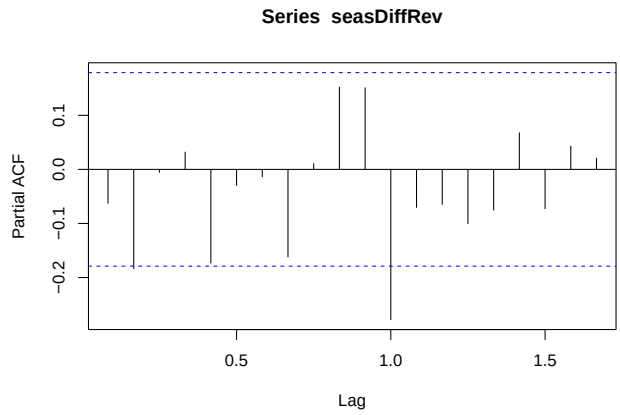
The resulting plot provides us with a possible power transformation that contains 0 and -1 within its confidence interval. However since 0 seems to be closer to the mean likelihood estimate for λ, we transformed our data by taking the logarithm. Note that the resulting time series still contains an upward trend after transforming. To account for this, we attempted to take a first and second difference of our data and evaluate them with their respective ACF's.



Visually, both of approaches were able to detrend our data and this is reflected in the ACF's by the autocorrelations being near or within our margins for most of the lags. However upon close inspection, it appears that there are a few lags resulted in higher autocorrelations within the second difference ACF, implying that this may be over-differencing the data. Hence we will proceed any further analysis using the first difference.

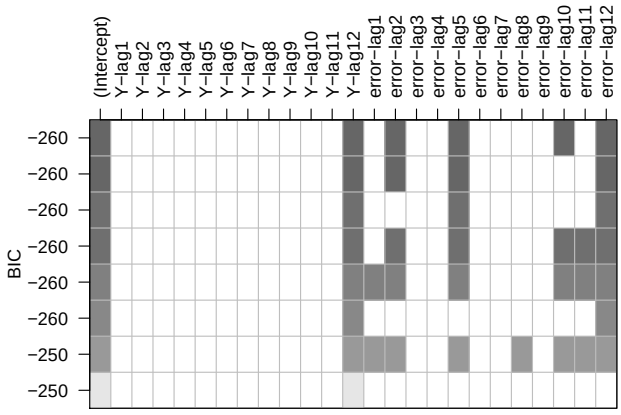
Moreover, we still observe a significant autocorrelation spike at lag 12. In order to account for this, we perform a seasonal difference to our first difference.





Now we see that that any seasonality has been largely account for. To finally determine our candidate models, we run ARMA-Subsets and EACF for the first non-seasonal and seasonal difference of revenue.

```
## AR/MA
##    0 1 2 3 4 5 6 7 8 9 10 11 12 13
## 0  o o o o o o o o o x o  x  o  o
## 1  x x o o o o o o o o o  x  o  o
## 2  o o o o o o o o o o o  x  o  o
## 3  x o x o o o o o o o o  x  o  o
## 4  o o x o o o o o o o o  x  o  o
## 5  x o x o o o o o o o o  x  o  o
## 6  x o x o o o o o o o o  x  o  o
## 7  o o x o o o x o o o o  x  o  o
## 8  o x x x o o x o o o o  x  o  o
## 9  o x x o o o x x o o o  x  o  o
## 10 x x x o o o x o o x o  x  o  o
## 11 x x x o o o x x x x x  o  o  o
## 12 o o x o x x o o o o o  x  o  o
```



Model Diagnostics

Moving forward with the four models previously identfied, we will train them on our data and evaluate their performance using a series of tests to assess randomness, normality and independnce.

```
##
## Call:
## arima(x = logRev, order = c(0, 1, 1), seasonal = list(order = c(0, 1, 2), period = 12))
##
```

```
## Coefficients:
##           ma1     sma1     sma2
##       -0.2142  -0.8220  -0.1780
## s.e.   0.1059   0.2165   0.1037
##
## sigma^2 estimated as 0.0005561:  log likelihood = 266.6,  aic = -527.2
```

```
##
## Call:
## arima(x = logRev, order = c(0, 1, 0), seasonal = list(order = c(0, 1, 1), period = 12))
##
## Coefficients:
##          sma1
##       -0.7873
## s.e.   0.1160
##
## sigma^2 estimated as 0.0006586:  log likelihood = 263.48,  aic = -524.96
```

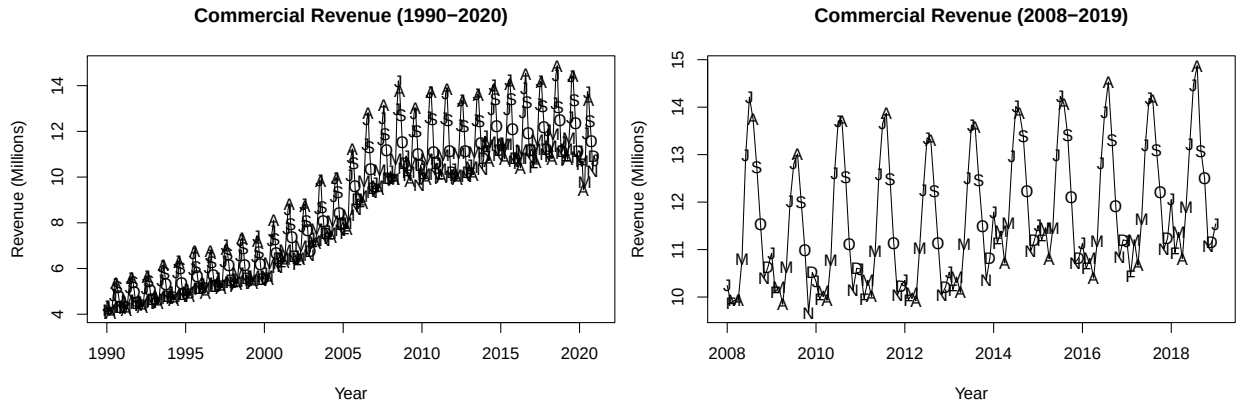
```
##
## Call:
## arima(x = logRev, order = c(0, 1, 1), seasonal = list(order = c(0, 1, 2), period = 12))
##
## Coefficients:
##           ma1     sma1     sma2
##       -0.2142  -0.8220  -0.1780
## s.e.   0.1059   0.2165   0.1037
##
## sigma^2 estimated as 0.0005561:  log likelihood = 266.6,  aic = -527.2
```

```
##
## Call:
## arima(x = logRev, order = c(0, 1, 2), seasonal = list(order = c(1, 1, 1), period = 12))
##
## Coefficients:
##           ma1      ma2    sar1     sma1
##       -0.2169  -0.1431  0.1800  -0.9997
## s.e.   0.0970   0.1036  0.1071   0.2615
##
## sigma^2 estimated as 0.0005518:  log likelihood = 267.39,  aic = -526.78
```

Forecasting

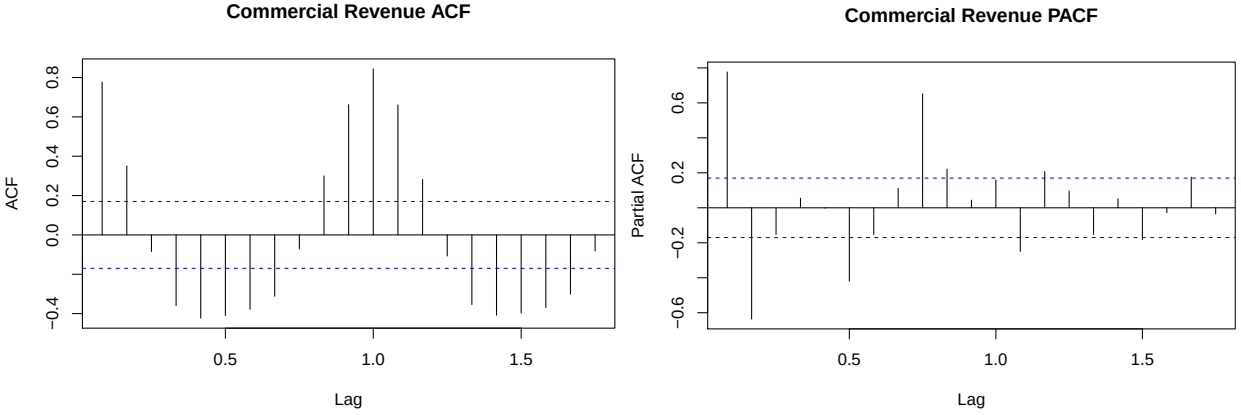
Commercial Revenue Data

The second series we will examine is the commercial revenue time series.



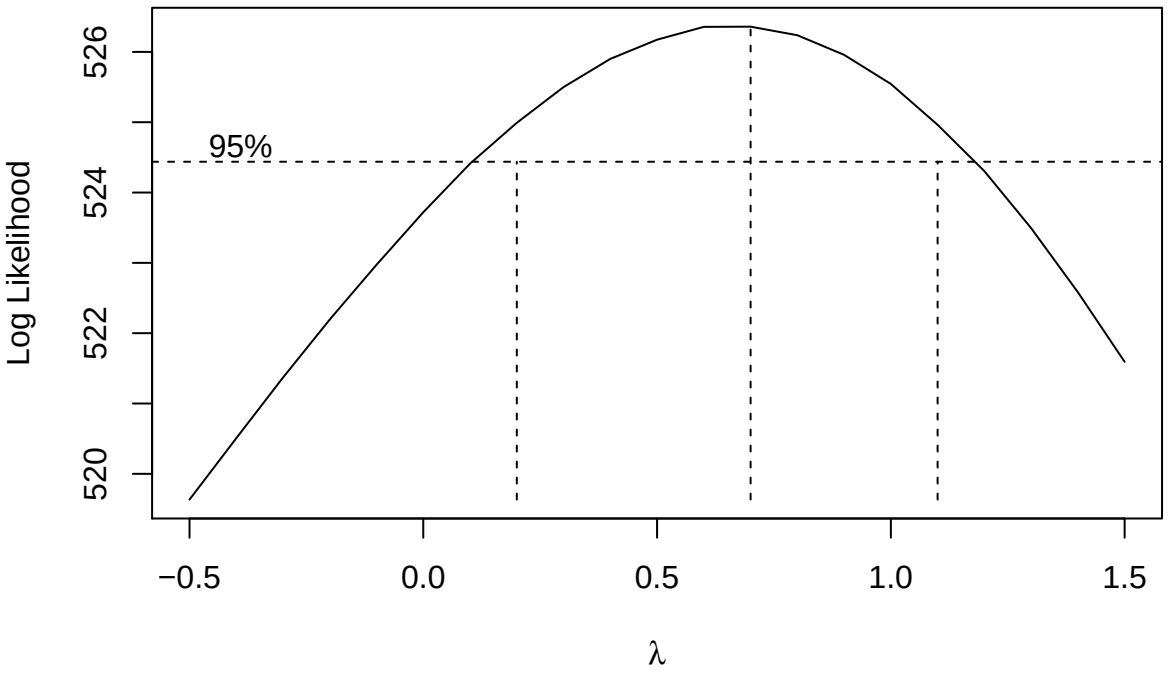
As in the total revenue series, we remove the data from the 1990s and 2000s as the underlying electricity economy was growing much faster then compared to now. We also hold out the 2019 and 2020 data for forecasting. Therefore, we model on data from January 2008 to December 2018, a total of 132 data points. Another similarity with the total revenue series is that the data has a strong seasonal component.

Looking at the plots, we also see some strong outlier seasons. For example the 2009 summer showed unusually low revenue relative to other summers while the 2014 winter was unusually high. We believe that this is due to weather effects, and these were likely a cool summer and warm winter respectively.[2]



Looking at the autocorrelation function, we can confirm that the data shows strong seasonality. The partial autocorrelation shows two significant correlations at lags 1 and 2 suggesting an AR(2) process but there are several additional spikes, most notably at lags 6 and 9.

[2]A U.S. Energy Information Administration post, explaining the seasonal variability in more detail can be found at https://www.eia.gov/todayinenergy/detail.php?id=10211



We considered taking a transformation but a Box-Cox plot reveals that none is necessary as $\lambda = 1$ (indicating the identity transformation) as within the 95% confidence interval of the plot.

Next we explore, the impact of differencing our data.

Model Specification

Model Diagnostics

Forecasting

Discussion

Conclusion

References