# Math 155 Final Report

Steven Litvack-Winkler 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 ARIMA$(0,1,1) \times$ 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 slightly different ARIMA$(2,0,0)$ process, accounting for seasonality with monthly means and time since the start of the series as regressors. While the model does not pass pass the Shapiro-Wilk test, performs fairly well under other metrics. The final model is also able to predict most of the data fairly well, with the exception of March and April 2020, which falls outside of the models forecast. However, this behavior is expected as a result of US COVID-19 lockdown restrictions during this time.

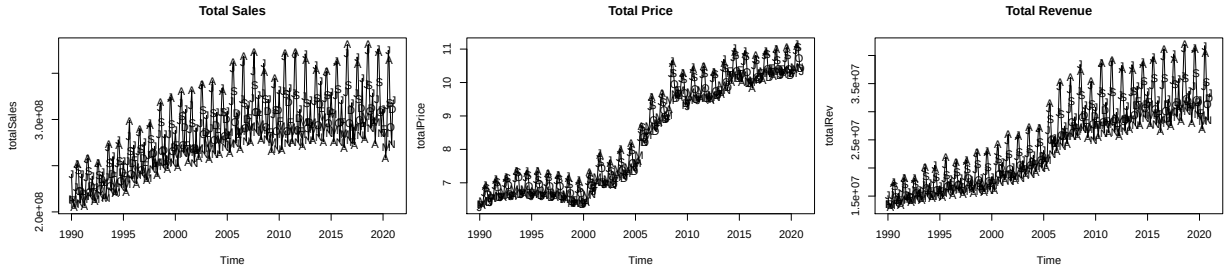
# Introduction

## 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). For this paper we focus on modeling only the national electricity revenue and we look at the total revenue and the commercial sector only. To reduce the scale of the data, the revenue is measured in millions of dollars.

## Data Exploration

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

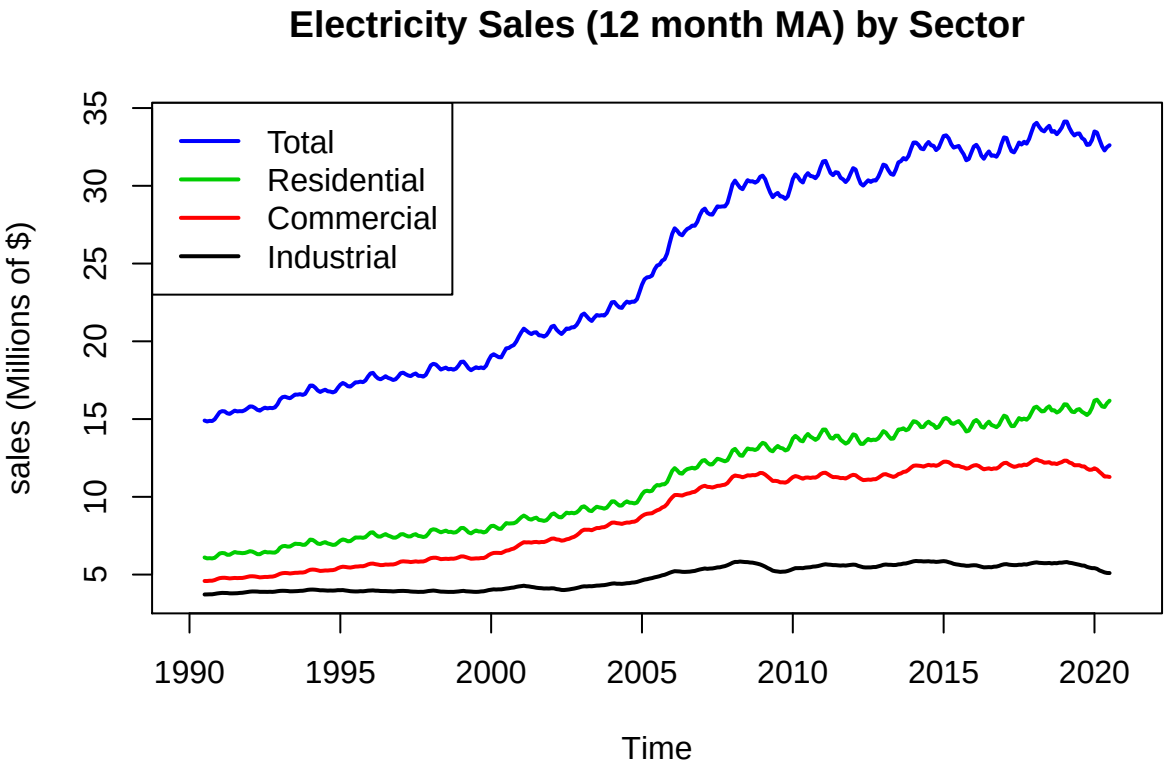



Note that all three graphs exhibit seasonal patterns but also long term trends. In particular, each decade of the data seems to behave differently. The 90s are characterized by growing usage and a cheap price of electricity. The 2000s are characterized by slower growth in the use of electricity and rising prices. Finally the 2010s are characterized by flat electricity usage and slowly rising prices. Another import trend in the

[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.

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 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.

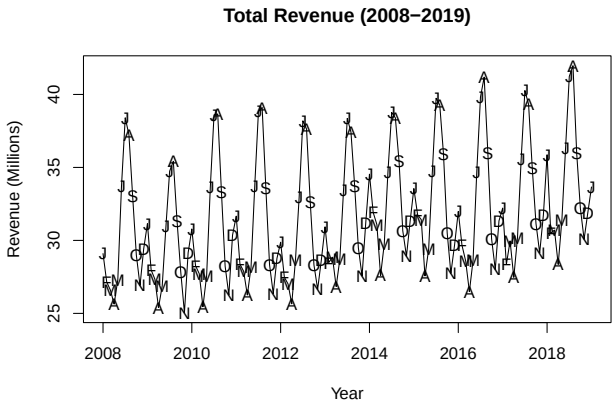
We can also see the electricity use breakdown by sector. To make the trends clearer over time, we show a 12 month moving average. Note that the residential, commercial, and industrial sectors together is about 97% of the total revenue before 2003 and more than 97% of the total revenue after 2003, the 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. The residential and commercial revenue both show noticeable seasonal trends although the residential revenues month to month variation appears to be higher. Additionally, the residential revenue has grown more over time than the commercial revenue.

## Total Revenue Data

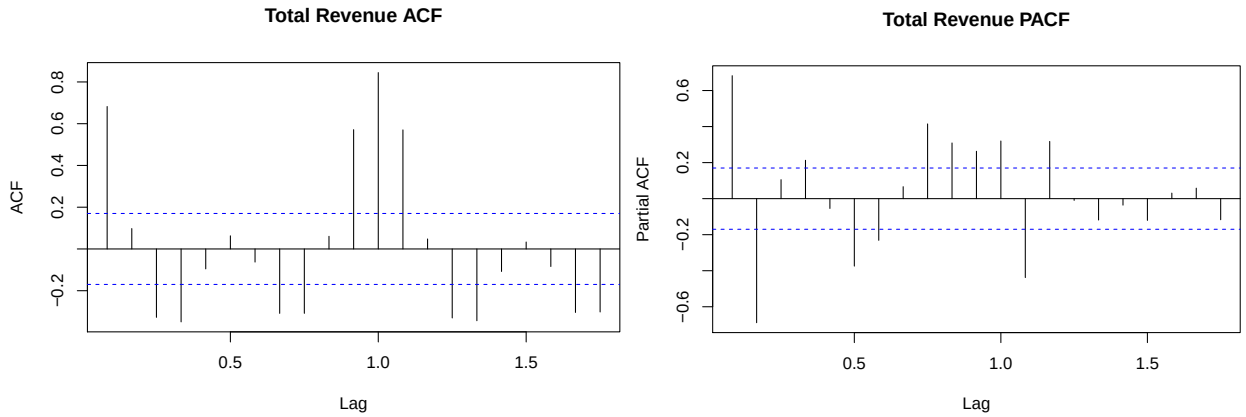
The first series we will examine is the total revenue time series.




As noted earlier, due to the 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.

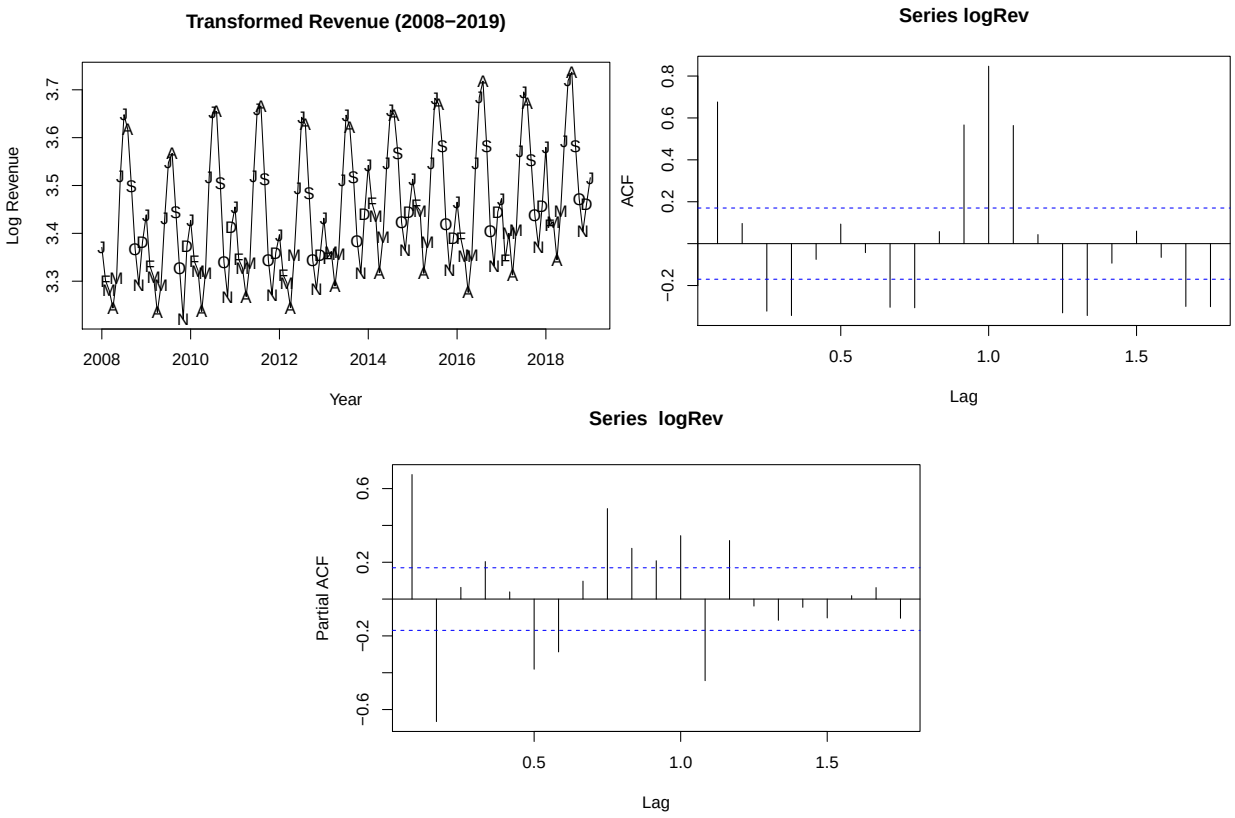
## Model Specification

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

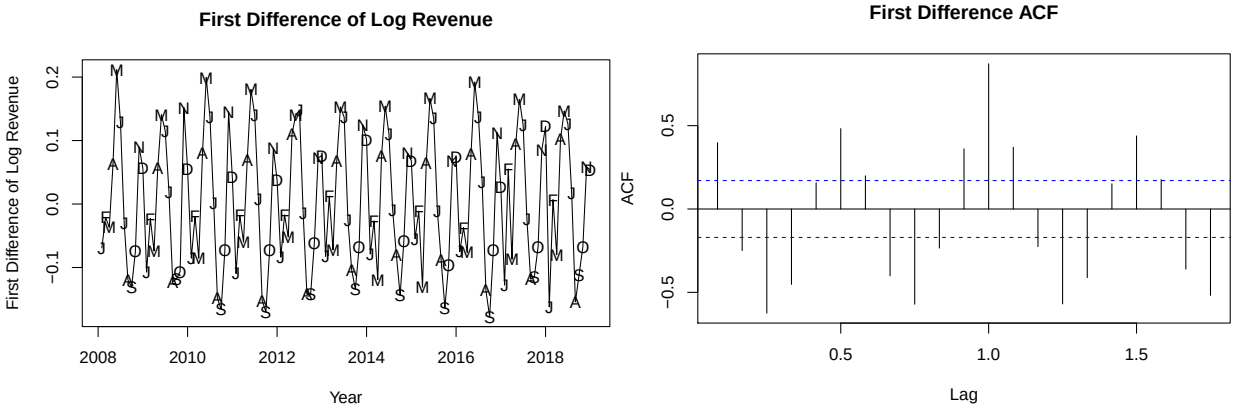


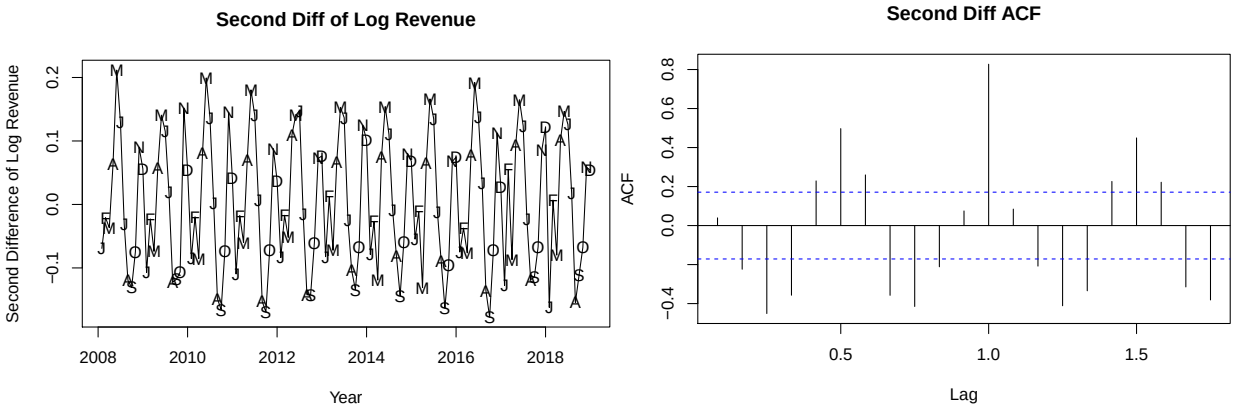

Plotting the ACF and PACF allows us to identify any autocorrelation present in our data, thus helping us determine an ARMA model to fit to our data. From the ACF our data is highly correlated at lag 1.0 (12 months), thus implying that our model will need a 12 month seasonal AR component. This is further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



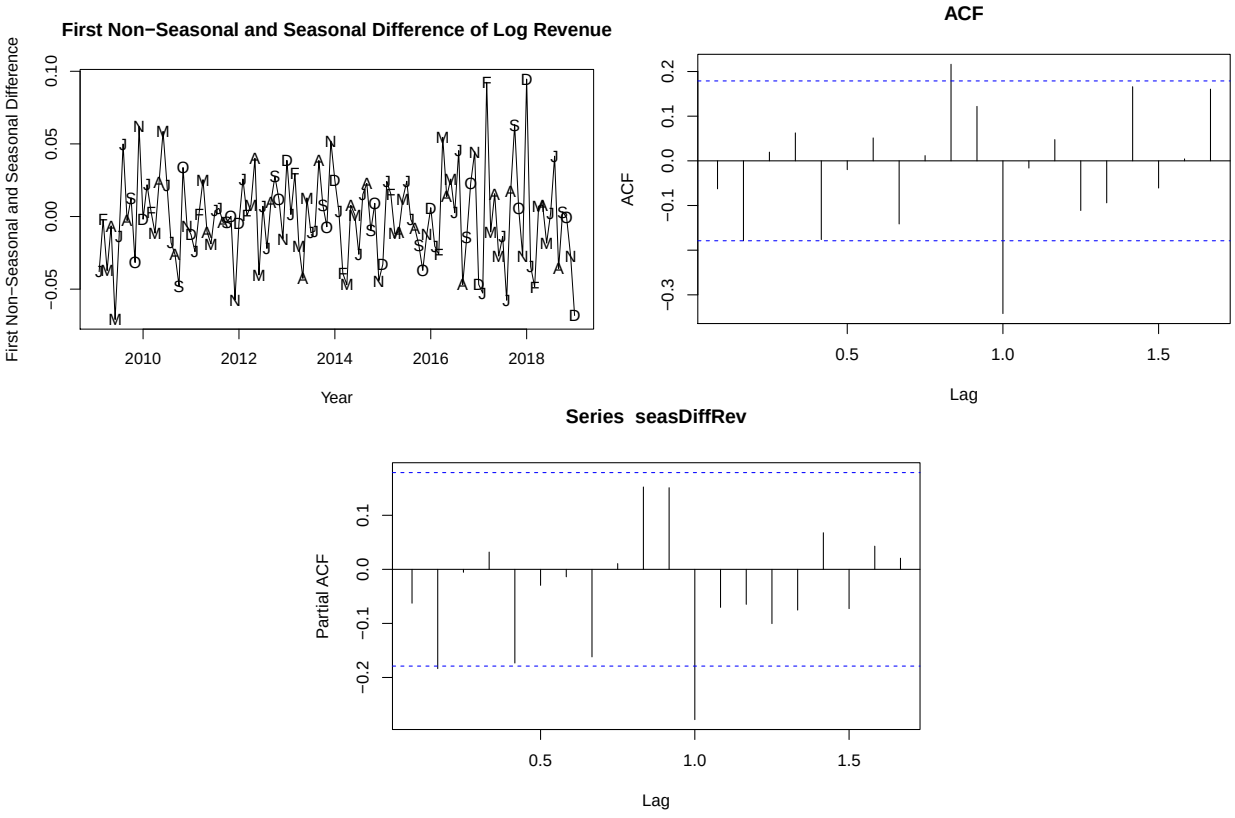

The resulting plot provides us with a possible power transformation that contains 0 and -1 within its confidence interval. However since 0 seems to be closer to the mean likelihood estimate for $\lambda$, we transformed our data by taking the logarithm. Note that the resulting time series still contains 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.

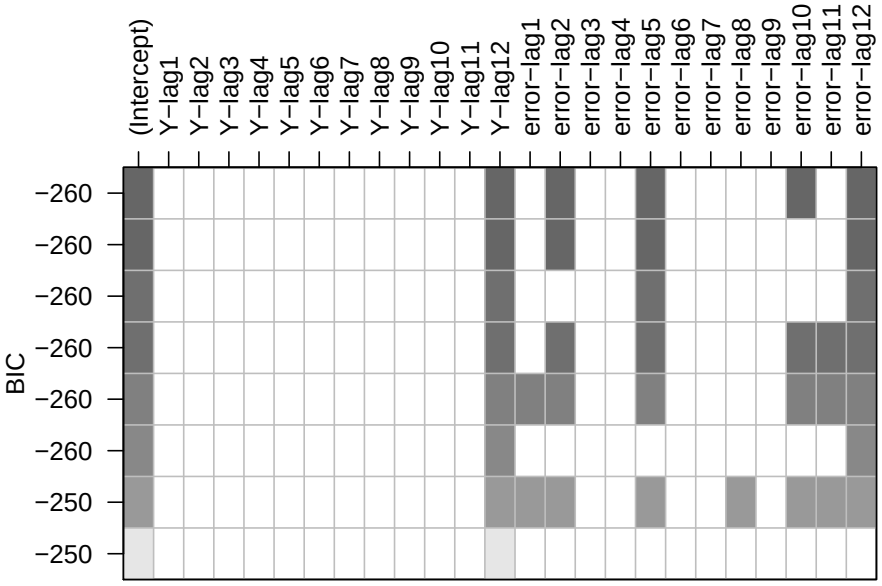
Moreover, we still observe a significant autocorrelation spike at lag 12 after the first difference. In order to account for this, we perform a seasonal difference to our first difference.




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. Firstly, the ARMA subset for the first non-seasonal and seasonal difference suggested a model with no non-seasonal AR terms, but possibly a seasonal AR term. Additionally it suggests a seasonal MA term in addition to a few non-seasonal AR terms. Thus some of the seasonal components were $(1,1,2)_{12}$, $(1,1,0)_{12}$ and $(0,1,2)_{12}$. Looking at the EACF, it seemed fairly possible that our model could contain up to two non-seasonal MA terms and at most one AR component depending on which values we count as false positives within the plot.

```
## [1] "Total Revenue EACF"
```

```
## 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
```




Using this information we reduced our search to the following models: ARIMA$(0,1,1) \times (0,1,2)_{12}$, ARIMA$(0,1,0) \times (0,1,1)_{12}$, ARIMA$(0,1,1) \times (0,1,2)_{12}$, ARIMA$(0,1,2) \times (1,1,1)_{12}$.

## Model Diagnostics

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

| Model 1 | MA1 | SMA1 | SMA2 |
|---|---|---|---|
| Coeff. | -0.2142 | -0.8220 | -0.1780 |
| Standard Error | 0.1059 | 0.2165 | 0.1037 |
| $\sigma^2 = 0.0005561$ | Log-likelihood = 266.6 | AIC =-527.2 | |

| Model 2 | SMA1 |
|---|---|
| Coeff. | -0.7873 |
| Standard Error | 0.1160 |

| Model 2 | SMA1 |
|---|---|
| $\sigma^2 = 0.0006586$ | Log-likelihood = 263.48 AIC =-524.96 |

| Model 3 | MA1 | SMA1 | SMA2 |
|---|---|---|---|
| Coeff. | -0.2142 | -0.8220 | -0.1780 |
| Standard Error | 0.1059 | 0.2165 | 0.1037 |
| $\sigma^2 = 0.0005561$ | Log-likelihood = 266.6 | AIC =-527.2 | |

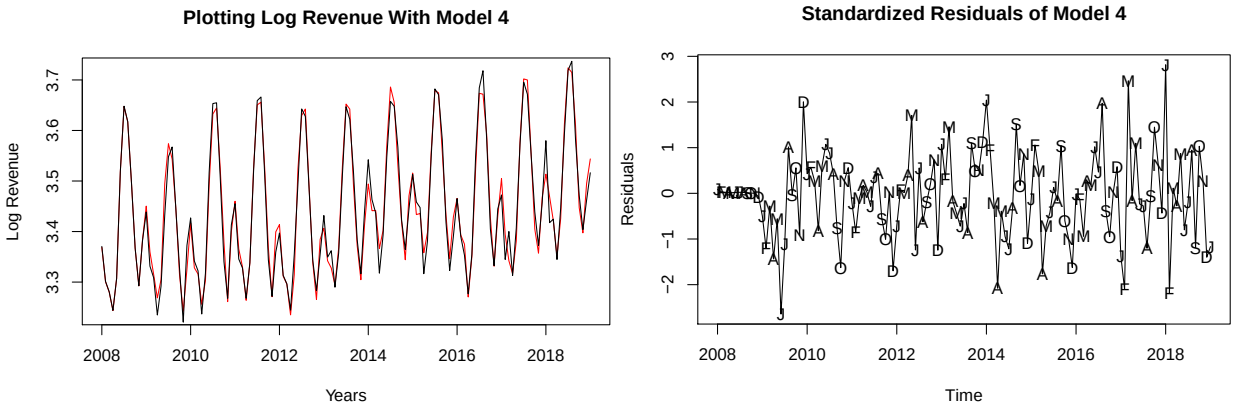
| Model 4 | MA1 | MA2 | SAR1 | SAR2 |
|---|---|---|---|---|
| Coeff. | -0.2169 | -0.1431 | 0.1800 | -0.9997 |
| Standard Error | 0.0970 | 0.1036 0.1071 | 0.2615 | |
| $\sigma^2 = 0.0005518$ | Log-likelihood = 267.39 | AIC = -526.78 | | |

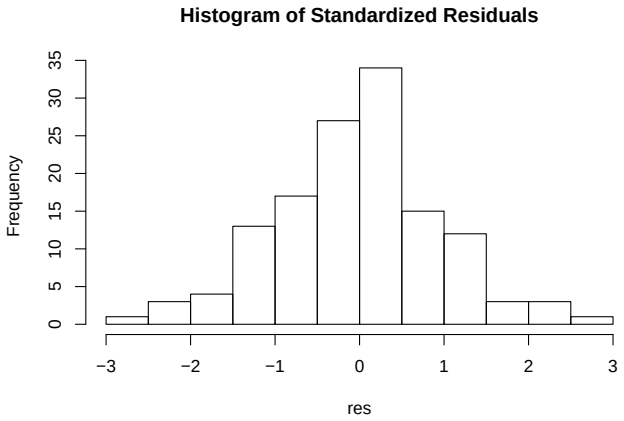
From the models themselves it is worth noting that Models 1, 3, and 4 had similar scores for AIC, whereas Model 2 appeared to score a few points lower. Additionally most, if not all, coefficients appear to be statistically significant since zero is more that a standard error away from the derived value.

| | Model 1 | Model 2 | Model 3 | Model 4 |
|---|---|---|---|---|
| Runs Test | 0.013 × | 0.506 ✓ | 0.0137 × | 0.0676 ✓ |
| Shapiro-Wilk Test | 0.6402 ✓ | 0.1992 ✓ | 0.6402 ✓ | 0.6032 ✓ |
| Box-Ljung Test | 0.09032 ✓ | 0.2217 ✓ | 0.09032 ✓ | 0.1331 ✓ |
| QQ Plot | Normal ✓ | Normal ✓ | Normal ✓ | Normal ✓ |
| Histogram | Normal ✓ | Normal ✓ | Normal ✓ | Normal ✓ |

From the summary table of diagnostics, we note that Model 1 and Model 3 both fail the runs the test, thus we must reject the hypothesis that the residuals are independent. However for examples, there is insufficient evidence to reject the hypotheses that the residuals are normally distributed. Additionally, the Box-Ljung test fails to reject the hypotheses that the models do not show lack of fit. In the end, the lack of independence within the residuals was sufficient for us to reject these two models and more closely consider the remaining.

Both Model 2 and Model 4 passed all tests. However, we went with Model 4 since it had a higher p-value under the Shapiro Test and due to its lower AIC score. For reference, we have included a plot of the historical data (black) along side the model (red) as well as the residuals resulting from our model choice.

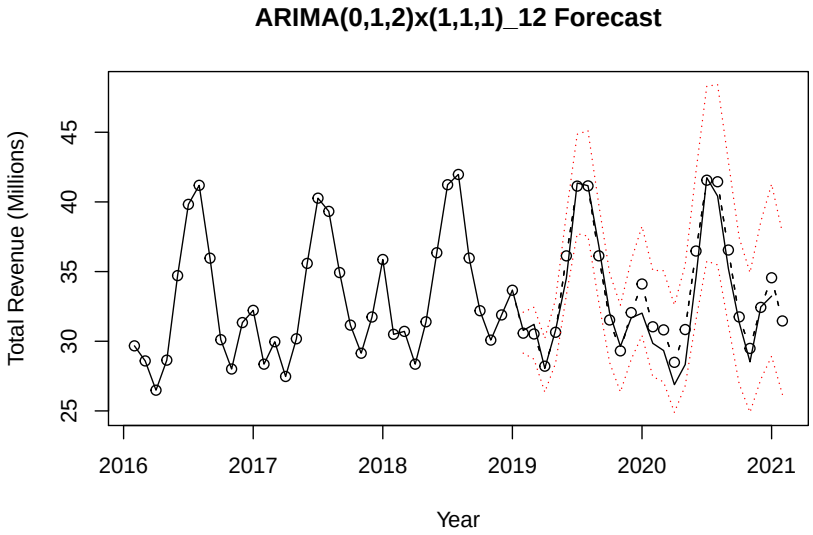




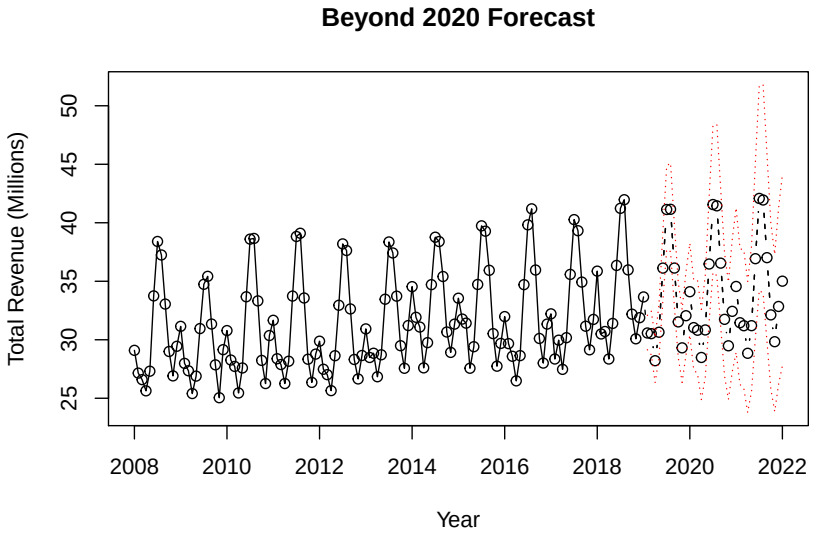

## Forecasting

To forecast future observations, we use the forecast and predict functions to forecast our Model 2 using our held out data from 2019 and 2020 at the 95% confidence level.

We first plot our model forecasts with the historical data. The solid black line represents the historical data while the dotted line represent the predicted values from our model. The red outline corresponds to our 95% confidence interval. From the plot, we observe that our model is able to account for the seasonality of our data and give us pretty accurate predictions, with all of our observations being within the 95% forecast limits. However, one thing we noticed is that our predictions are consistently over the historical value, while this is not statiscally significant given the confidence interval, it is worth bringing up this positive bias as it seems more pronounce in the Commercial revenue, as will be discussed in the next section.

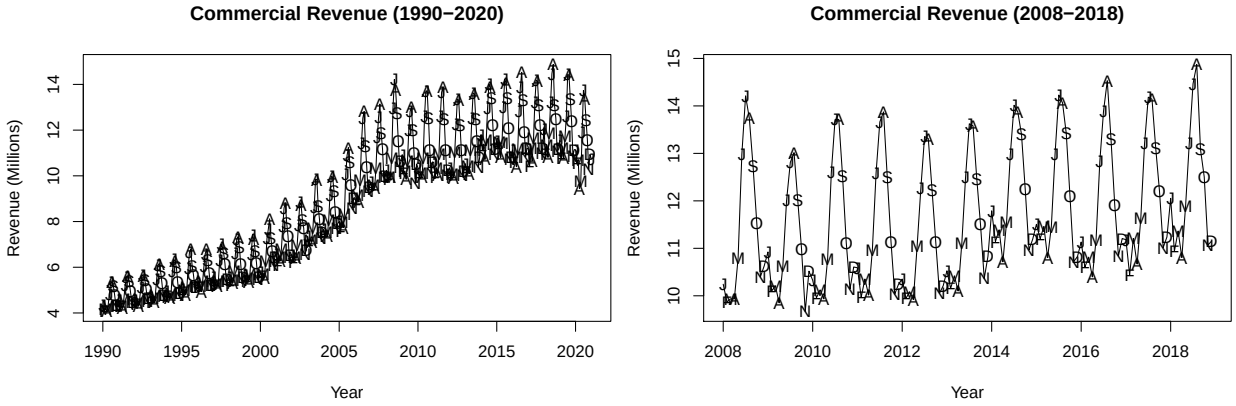



We also plot the model's predictions further into the future. Overall we do see a slight positive trend moving towards the future according to our model.




# Commercial Revenue Data

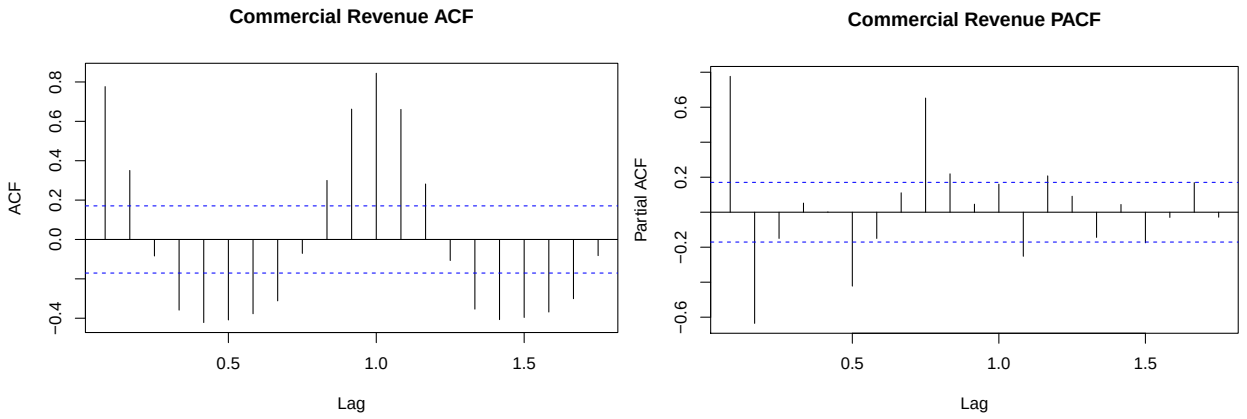
The second series we will examine is the commercial revenue time series.



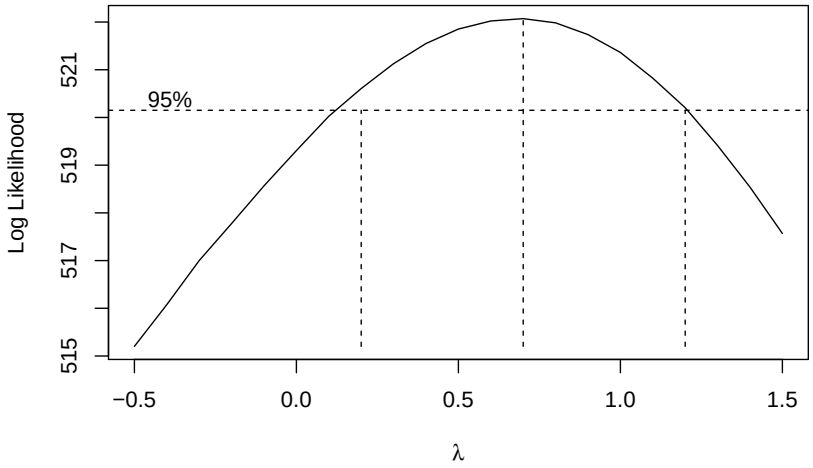

As in the total revenue series, we remove the data from the 1990s and 2000s as the underlying electricity economy was growing much faster then compared to now. We also hold out the 2019 and 2020 data for forecasting. Therefore, we model on data from January 2008 to December 2018, a total of 132 data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

[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

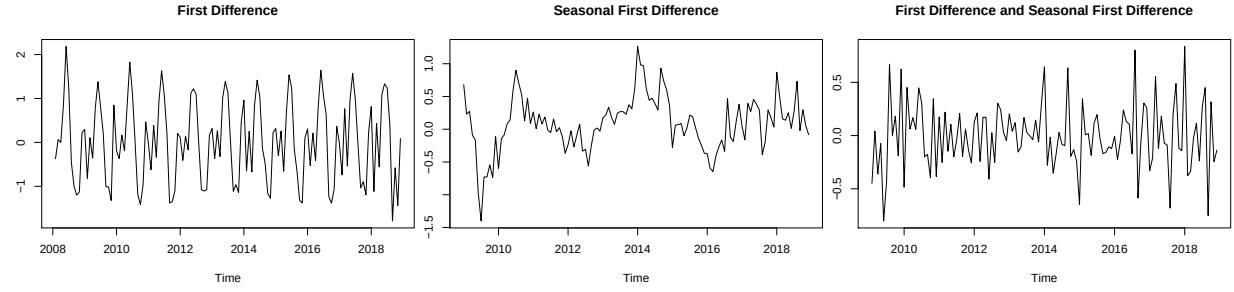



Looking at the autocorrelation function, we can confirm that the data shows strong seasonality. The partial autocorrelation shows two significant correlations at lags 1 and 2 suggesting an AR(2) process but there are several additional spikes, most notably at lags 6 and 9.




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.

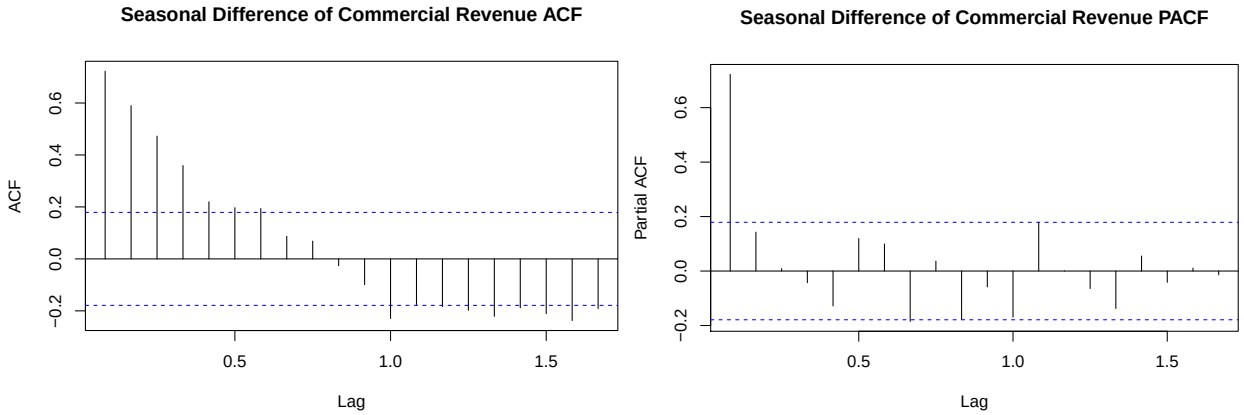
Next we explore, the impact of differencing our data. Since there are strong seasonal affects these need to be removed either through a seasonal regressor or a seasonal difference.



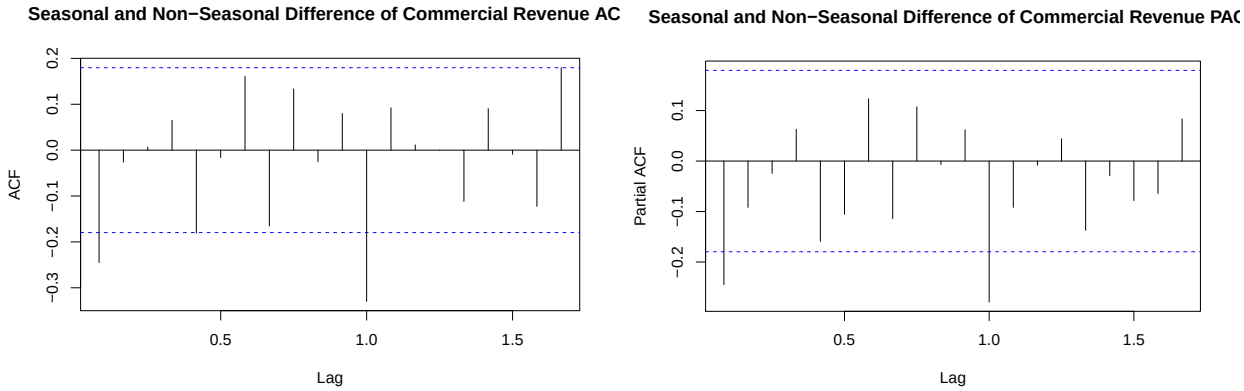

Indeed, the first difference alone is clearly leaves the seasonal patterns intact. The first seasonal difference and the combined first seasonal and non-seasonal difference both succeed in removing the seasonal patterns so we we will proceed to consider ARIMA models with one seasonal difference and either zero or one non-seasonal differences.

## Model Specification

To better understand the correlations in our series we recompute the ACF and PACF of our differenced data.




The partial autocorrelation function has a large spike at lag 1 that strongly suggests including an AR(1) term in the model. Since the autocorrelation function looks somewhat like a decaying exponential that an AR(1) process produces we can consider modeling the series as an ARIMA$(1,0,0) \times (0,1,0)_{12}$ process. We'll use this as model 1. Next we consider the series with both a seasonal and non difference applied.




This set of graphs is more difficult to interpret although both the ACF and the PACF have spikes at lag 12. The graphs also have spikes at lag 1 although they are both close to the significance threshold. Since there is no clear model suggested by the ACF and PACF we now seek additonal candidate models using the armasubsets tool in R.

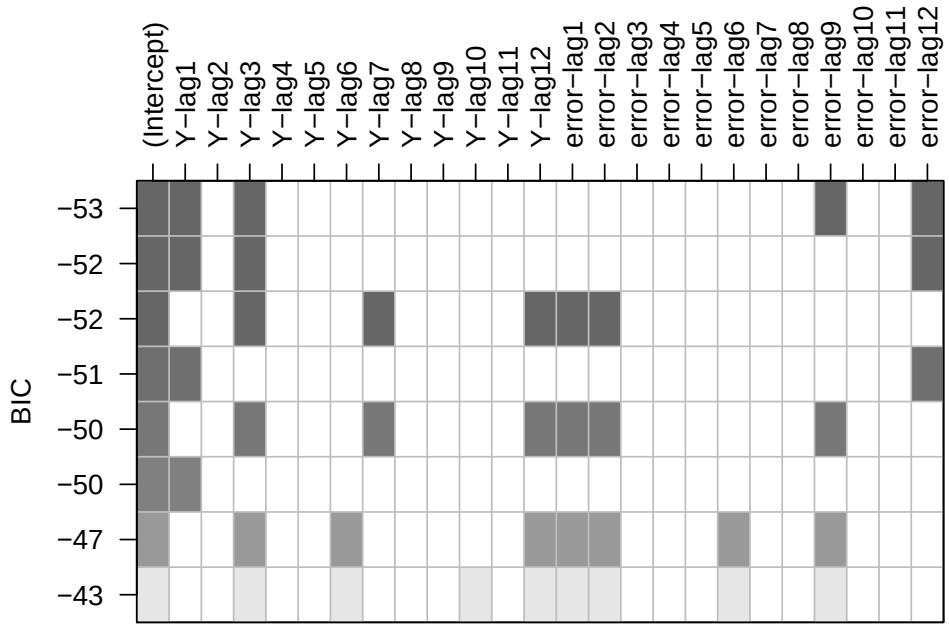



Figure 1: Arma subsets of Seasonally Differenced Commercial Revenue

Beginning with the seasonally differenced data, the second and third rows of the plot suggest two further strong candidate models (models 2 and 3), an ARIMA$(2,0,0) \times (1,1,0)_{12}$ and ARIMA$(1,0,0) \times (0,1,1)_{12}$. Note that model 3 is similar to the model 1 we guessed at from the ACF and PACF plots but also has a seasonal MA term. We can also see that our model 1 shows up as the third to last row on the armasubsets plot, so we can infer that it may be a substantially worse model.

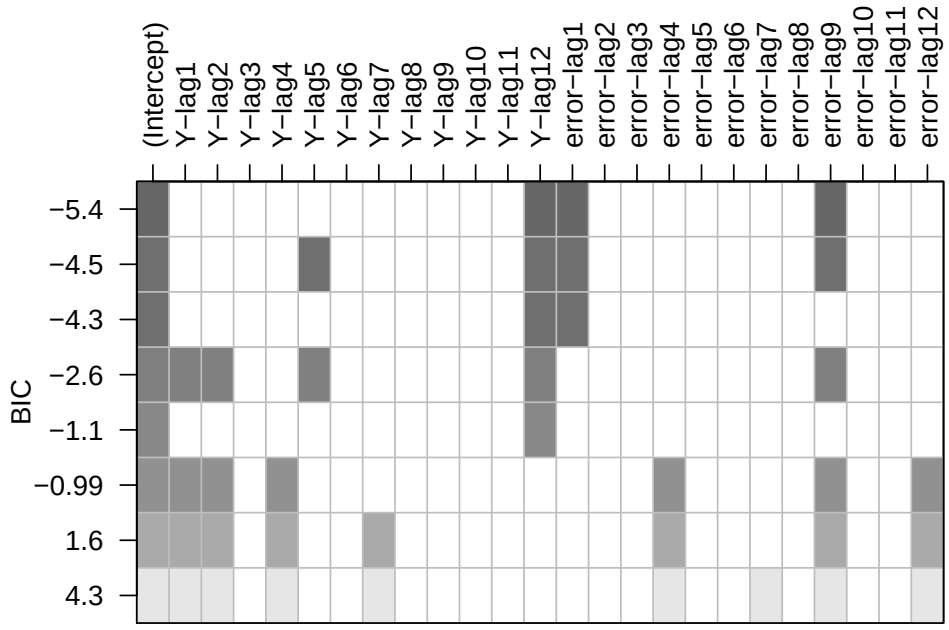



Figure 2: Arma subsets of Seasonally and Non-Seasonally Differenced Commercial Revenue

Continuing with the seasonally and non-seasonally first differenced data, the top 3 models all suggest a seasonal autoregressive term and a non-seasonal moving average term. The top 2 models also include a

moving average term of lag 9, although we choose to ignore this as it doesn't appear in the autocorrelation plot and is not parsimonious with a periodicity of 12 that shows up elsewhere in the data. Therefore we will take our 4th model to be an ARIMA$(0,1,1) \times (1,1,0)_{12}$ process.

Finally, we consider addressing the seasonal pattern using a monthly means regression instead of a seasonal difference. Since the model also has a small increasing trend over time, we include the time as well. The armasubsets function on the residuals of a linear model on the seasonal means and time of the data is

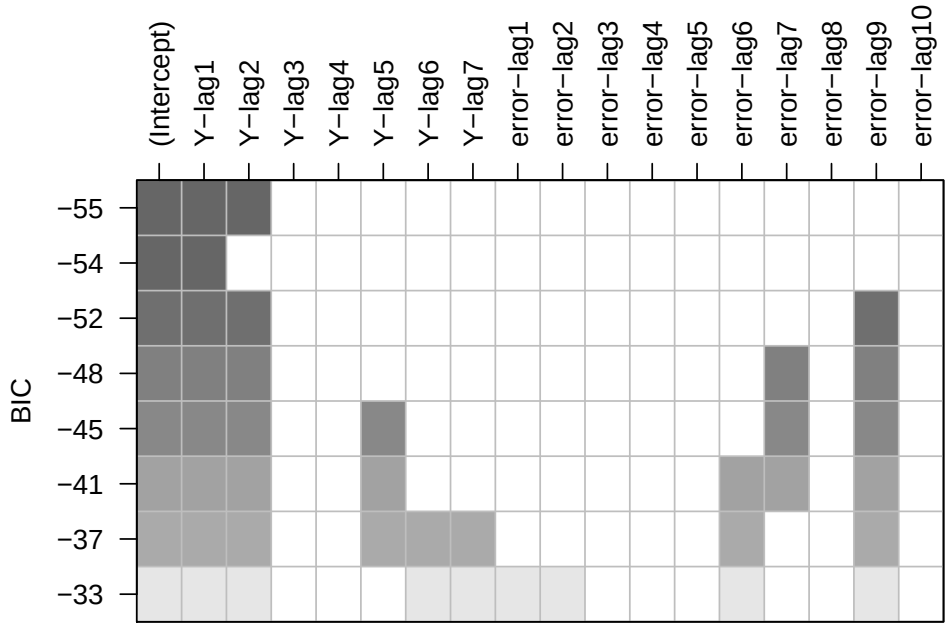



Figure 3: Arma Subsets of Regression on Month and Time Residuals

This suggests our final model is an ARIMA$(2,0,0)$ on the regression residuals.

## Model Diagnostics

We now define our 5 models and evaluate them.

The five models are defined below. The fifth model is described in two tables since it has more parameters.

| Model 1 | AR1 | $\sigma^2$ |
|---|---|---|
| Coeff. | 0.7417 | 0.0814 |

| Model 2 | AR1 | AR2 | SAR1 | $\sigma^2$ |
|---|---|---|---|---|
| Coeff. | 0.6650 | 0.1464 | -0.4407 | 0.06706 |
| Std. Error | 0.0906 | 0.0904 | 0.0958 | |

| Model 3 | AR1 | SAR1 | $\sigma^2$ |
|---|---|---|---|
| Coeff. | 0.8001 | -0.5996 | 0.06376 |
| Std. Error | 0.0591 | 0.1066 | |

| Model 4 | MA1 | SAR1 | $\sigma^2$ |
|---|---|---|---|
| Coeff. | -0.2667 | -0.4496 | 0.07225 |
| Std. Error | 0.0927 | 0.0951 | |

| Model 5 | AR1 | AR2 | Intercept | Time | Jan | Feb | Mar |
|---|---|---|---|---|---|---|---|
| Coeff. | 0.5517 | 0.2015 | 10.1166 | 0.1119 | 0.2803 | -0.3008 | -0.0326 |
| Std. Error | 0.0852 | 0.0851 | 0.1514 | 0.0215 | 0.0721 | 0.0820 | 0.0911 |

| Model 5 | Apr | May | Jun | Jul | Aug | Sep | Oct | Nov | $\sigma^2$ |
|---|---|---|---|---|---|---|---|---|---|
| Coeff. | -0.4250 | 0.4665 | 2.0153 | 3.0949 | 3.1378 | 2.042 | 0.9052 | -0.2968 | 0.0442 |
| Std. Error | 0.0911 | 0.0962 | 0.0991 | 0.0998 | 0.0987 | 0.0955 | 0.090 | 0.0803 | 0.0700 |

The first test we will perform is the runs test which checks if the residuals have remaining correlations. The runs test p-values on the models are:

```
## [1] 0.3100 0.2790 0.5770 0.0313 0.5680
```

We see that models 1, 2, 3, and 5 all have p-values above 0.05 indicating that they pass the runs test. However, models 3 and 5 have a significantly larger p-value than the first two indicating they more comfortably pass the test. Model 4 does not pass the test, and a more detailed look at the output shows that it has only 54 observed runs compared to 67.3 expected runs. This indicates that the residuals from model 4 still are positively correlated.

Moving on to the Shapiro-Wilk test, this tests the models residuals for normality. The p-values of the 5 models are

```
## [1] 0.0025461393 0.0017741788 0.0001322545 0.0170986423 0.0002853906
```

In fact, we see that all of the models fail the Shapiro-Wilk test so this won't help us much in deciding which is best. This likely indicates that the dataset contains some strong outliers that all of the models struggle to fit.

We can consider the Ljung-Box test which also checks if the correlations between the residuals are significant. Since our dataset is seasonal with lag 12 we use the default value of including 12 autocorrelations in the Ljung-Box test. The p-values from the Ljung-Box test are

```
## [1] 0.001194265 0.563645300 0.223496981 0.681920276 0.974868971
```

We see that model 1 fails the Ljung-Box test while the rest pass it.

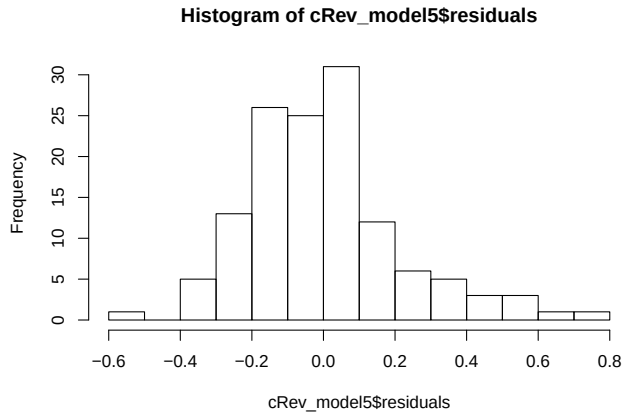
Finally we can compare the Akaike Information Criterion (AIC) and Bayesian Information Criteria (BIC) of each model. The AIC and BIC values are

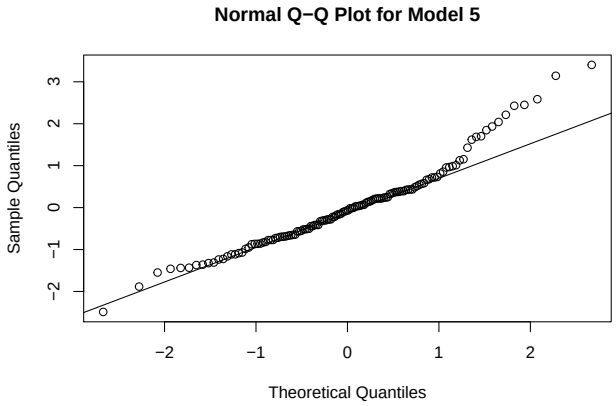
```
## [1] 44.332803 27.739551 22.511076 33.798884 -4.392709
```

```
## [1] 49.90779 38.88952 30.87355 42.13625 41.73212
```

Since models 1 and 4 failed the runs test and Ljung-Box test respectively we are mainly choosing between models 2, 3, and 5. Note that models 1 and 4 also have the worst AIC and BIC values, so the AIC and BIC largely agree with the previous tests. Model 5 has by far the best AIC but also the worst BIC of these three models. Nevertheless, we will select model 5 as the best model since most of the parameters in the model are significant and we think BIC is excessively penalizing its higher number of parameters.

Since we proceed with model 5 we will also check it's QQ-plot and histogram for normality. We note though that since the model failed the Shapiro-Wilk test, we do not expect these to look normal.



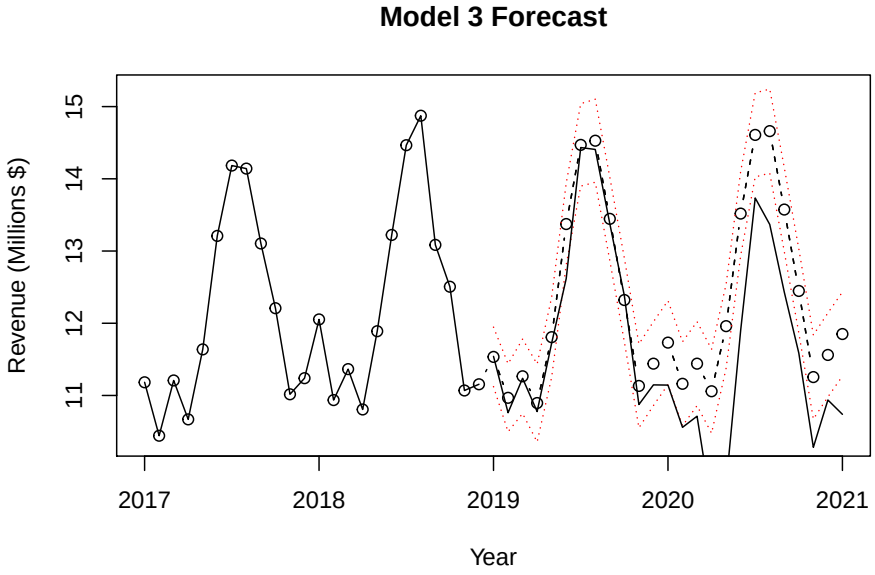

Indeed, the histogram shows substantial leftward skew and the QQ plot clearly deviates from it's expected line at positive quantiles above 1.

The full summary of test results is given in the table below:

| | Model 1 | Model 2 | Model 3 | Model 4 | Model 5 |
|---|---|---|---|---|---|
| Runs Test | 0.3100 ✓ | 0.2790 ✓ | 0.5770 ✓ | 0.0313 × | 0.5680 ✓ |
| Shapiro-Wilk Test | 2.5461e-03 × | 1.7742e-03 × | 1.3225e-04 × | 1.7099e-02 × | 2.8539e-04 × |
| Box-Ljung Test | 0.0012 × | 0.5636 ✓ | 0.2234 ✓ | 0.6819 ✓ | 0.9748 ✓ |
| AIC | 44.332803 | 27.739551 | 22.511076 | 33.798884 | -4.392709 |
| BIC | 49.90779 | 38.88952 | 30.87355 | 42.13625 | 41.73212 |

## Forecasting

To further test our model we will check its predictions on the data from 2019 and 2020. As before we plot the true data in a black line, the predicted points as blue circles and the forecasts 95% confidence band in red.




We see that the forecast performed quite well through most of 2019 although it did overestimate the observed value in June 2019. However, beginning in February 2021, the observed data lies entirely below the forecasts 95% confidence interval. This implies that the electricity revenue for 2021 was very unusual from a historical perspective. However, given the circumstances of the COVID-19 pandemic and the lockdowns that followed

it we think it is likely this impacted the commercial revenue data and explains the models overpredictions. Commercial users of electricity include schools, offices, and malls and these buildings were largely operating at far below their normal capacity. Further evidence for this theory is given by the fact that our models prediction was most off in April at the beginning of lockdowns in many states.[3]

# Discussion & Conclusion

In conclusion, this project explores time series data for U.S. electricity revenue from 2008-Present. Through trying various transformations and differencing approaches, we arrive at a handful of d models. These models are tested for independence, randomness, normality, and "good-ness of fit", from which we select a single model. The selected model is then used to analyze held-out data as well as forecast future observations.

In the case of total revenue, our final model was an $\text{ARIMA}(0,1,1) \times (1,1,1)_{12}$. We found this model captures the seasonality of our data fairly well, as for to capture all of our held-out data within a 95% confidence interval. However, we note that our model seemed to have a slight postive bias, overpredicting most points. We hypothesize that this is due to the global influence of the novel COVID-19 virus causing a measurable decrease due to lockdowns across the country.

For the commercial revenue series we used an $\text{ARIMA}(2,0,0)$ model with monthly means and time since the start of the series as regressors. This model fit the data very well with low errors although it failed the Shapiro-Wilk test indicating that the residuals of the model are not normally distributed. We believe this is due to underlying outliers in the dataset. When predicting with the model, it performed quite well in 2019 but struggled in 2020. This is likely because of the effect of COVID-19 that are model could not have anticipated since no comparable event occurred in its training data. The model for commercial revenue missed even more than our prior model for the total revenue. We believe this is because while commercial and industrial electricity use dropped during COVID, the residential sector use of electricity increased. Therefore, the total electricity revenue series had a weaker decline in 2020 than the commercial electricity revenue series.

# References

Cryer, Jonathan D., and Kung-sik Chan. Time Series Analysis with Applications in R. Springer, 2011.

https://www.eia.gov/electricity/data.php#sales

Source Code found in Appendix.

[3]California was the first state to issue a formal lockdown on March 20, 2020. https://www.ksla.com/2020/03/20/california-becomes-first-state-order-lockdown/