174 Final Project

Vance Sine

3/7/2022

Abstract

This Dataset was created by Makridakis, Wheelwright, and Hyndman in 1998. This contains monthly number of shales of shampoo over a three year period. This dataset provides us with insight on public perception on hygeine and month-to-month differences in consumer sales. Our goal for this project is to find a suitable model to predict the next ten years of shampoo sales, based off of our current figures. Analyzing the increase and decrease in overall sales can provide shampoo manufacturers with valuable insight on consumer trends and help alleviate any supply constraints/overstocking.

In order to make this prediction, I started by plotting the time series in order to gain an initial understanding of my dataset. I noticed an increasing trend, seasonality, and a non-stable variance. In hopes of minimizing the variance, I applied a box-cox transformation. I could see from a plot that our observed trend and seasonality witheld, while effectively minimizing variance. I also tried a log-transform as this gives us a symmetric histogram and more even variance. I see that, the log transformation minimizes variance the most. After this, I look at the ACF and PACF of our transformed dataset to select models to test. Following this, I perform diagnostic checks and compaired AIC to select the best-fit. I then plot the forecasted future values and check our confidence interval.

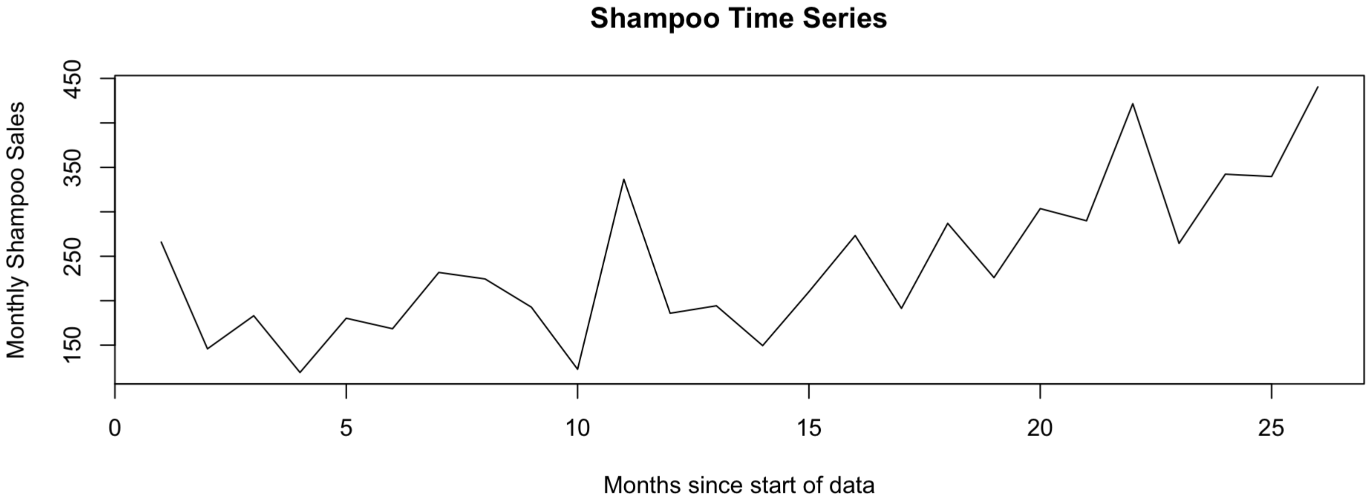
Introduction

People have different hygeine needs during different times of the year. Users are likely to need shampoo more often during seasons that have extreme weather and require more showering(etc. heat wave during summer). During this time of year, people are more active and need to clean themselves more. This dataset contains shampoo sales for each month, for three continuous years. With an increasing trend in sales, our goal is to forecast future shampoo sales in order to help manufacturers effectively deal with an increased demand. To create a forecast, I start by transforming our data to minimize variance. In order to account for a seasonal component, I apply lags 1 and 12 in order to remove any linear trend and seasonality. I then identified models: ARIMA(0,1,1) ARIMA(0,1,2), ARIMA(0,1,3), ARIMA(1,1,1), ARIMA(1,1,2), and ARIMA(1,1,3) as potential fits for our dataset. After performing diagnostic testing to see which prediction fits best. I then analyze the residuals to make sure that they follow white-noise. I then test this model to ensure an approximately normal distribution, along with ensuring independence. I then use this model to forecast the next 10 years and verify their accuracy via the use of a 95% confidence interval.

Loading Data

Our Dataset has monthly shampoo sales for 36 months, I will use this to conduct our time series analysis.

Plotting Time Series

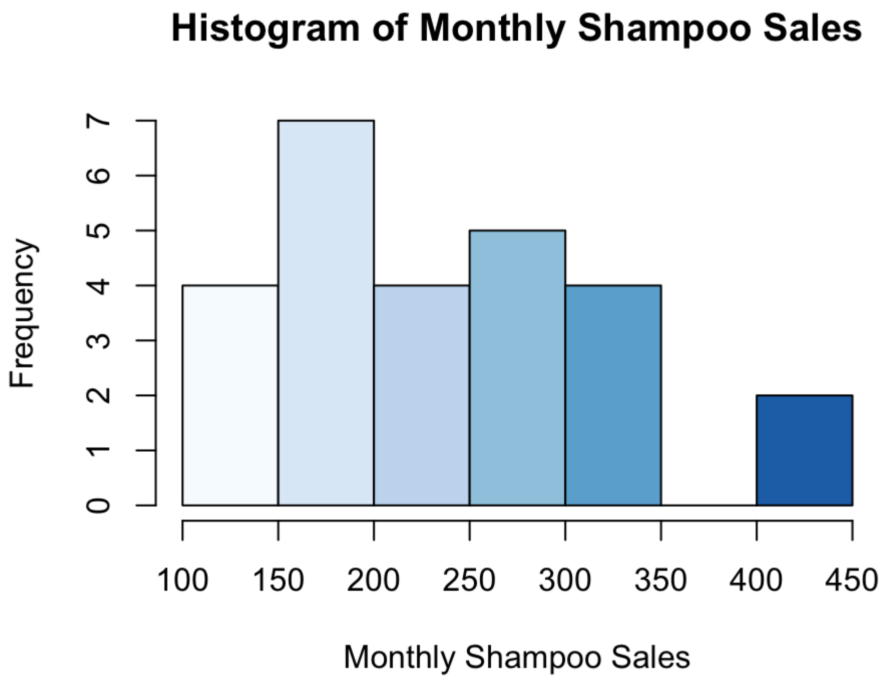


i)Trend Analysis: I see an increasing trend in sales. As the number of months increase, the monthly sales increase.

ii)Seasonal Analysis: I see peaks toward the end of each year, with a sharp drop in the following month.

iii)Sharp Changes in Behavior: Seeing as there is a seasonal component and a changing variance, this data is not normally distributed.

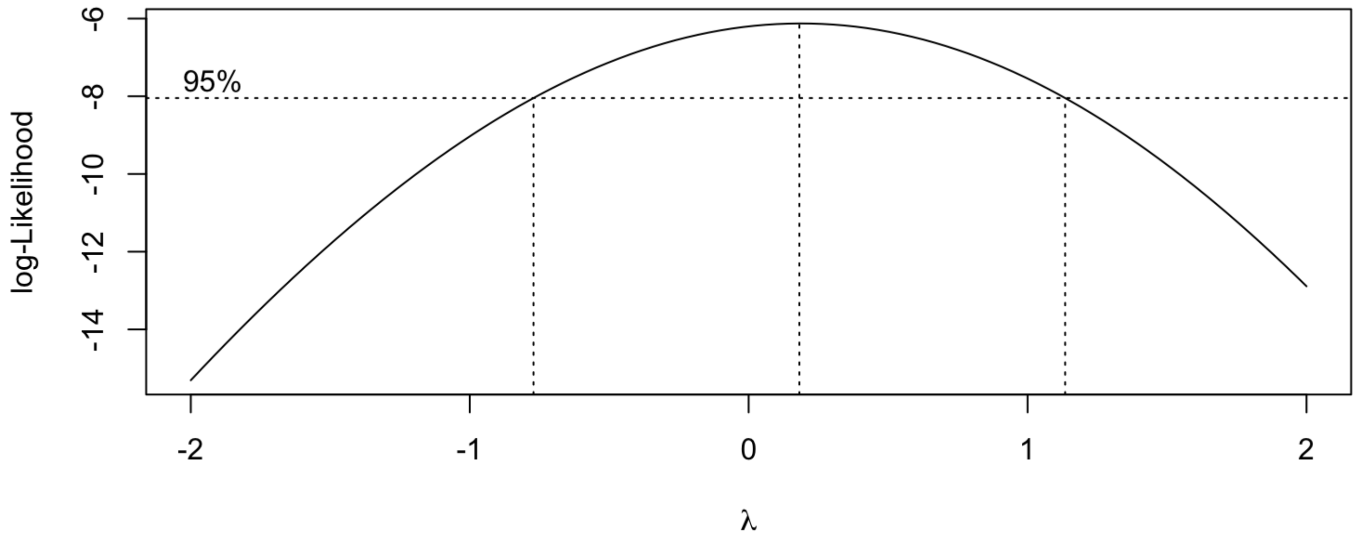
I see that seasonality appears to be present. Our data set covers monthly data collected over three years. I see three distinct peaks toward the latter months of each year. Additionally, there is a noticable increasing trend throughout.



I can see that this histogram for our data is skewed to the right. I want to normalize this via transformation. Let’s start with box-cox and then compare the variance to a log-transformation.

Implementing Box-Cox to stabilize the variance

In order to make our data resemble a normal distribution, I implement a box cox transformation in hopes of stabilizing the variance.

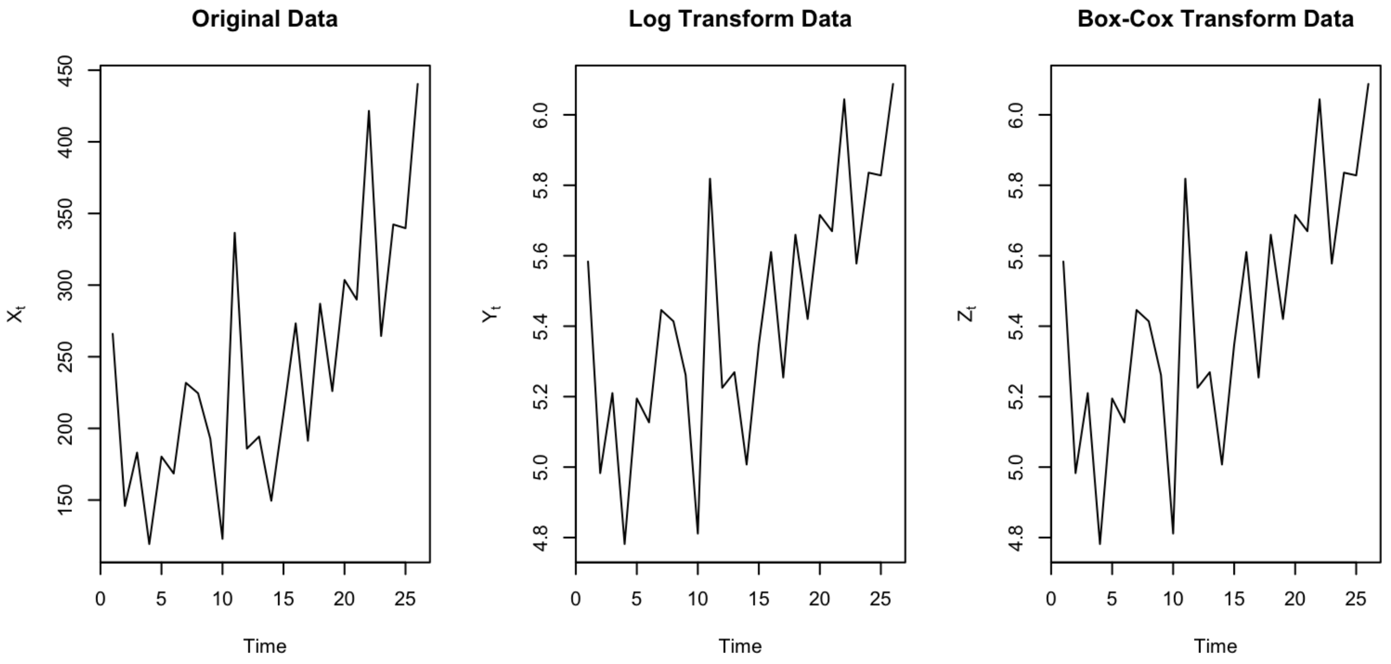


Finding optimal Lambda value

```
## [1] 0.1818182
```

It appears that our optimal lambda value is 0.1414141 to maximize our-log likelihood.

Plot original and transformed data Time Series



The variance is much lower after our log transformation. Additionally, I can see that the graph is much less volatile. However, our variance is actually minimized when I use a log.

Compare Variances

Actual Variance

```
## [1] 7285.102
```

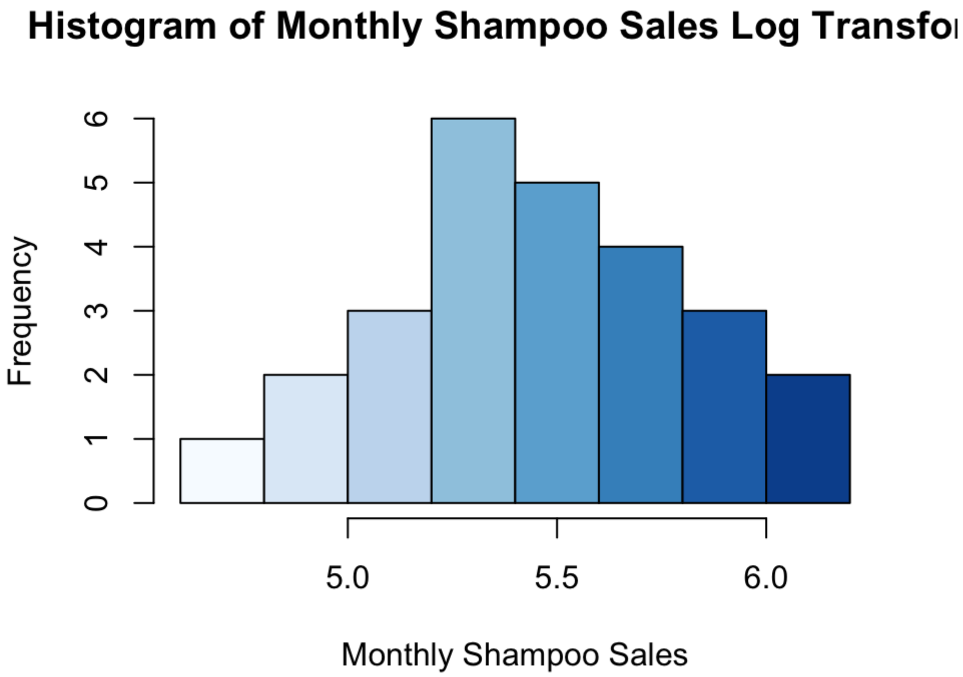
Box-Cox Variance

```
## [1] 0.8877589
```

Log Variance

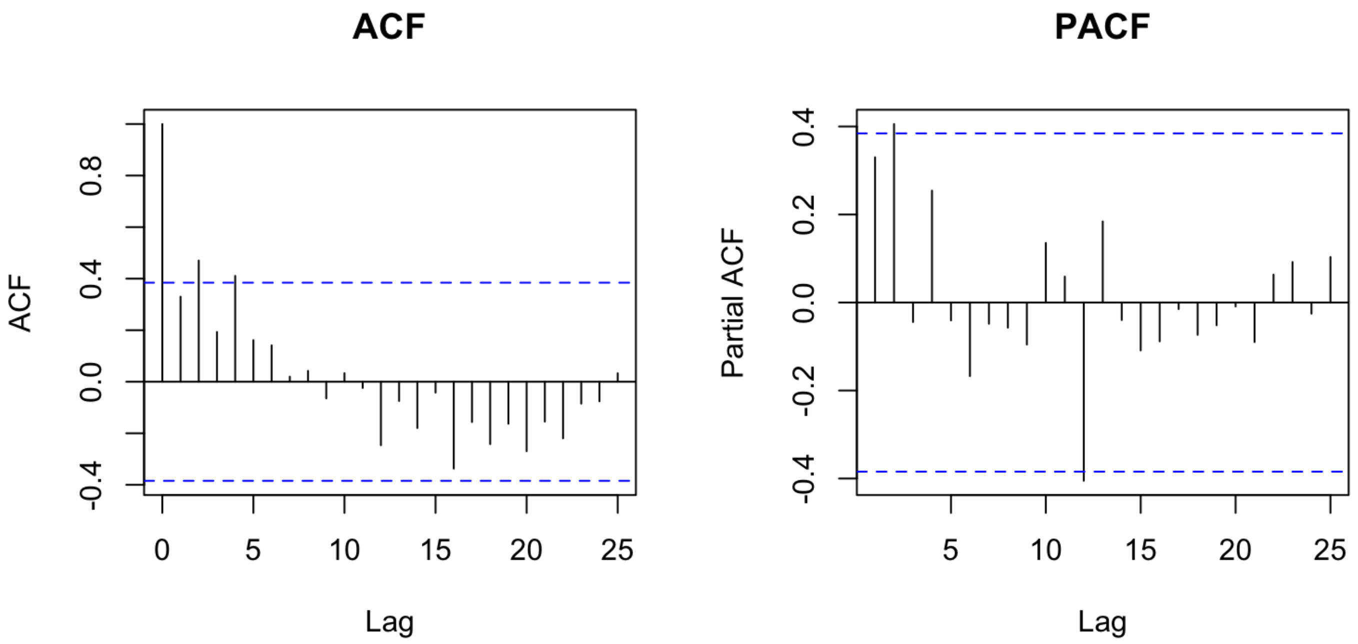
```
## [1] 0.1225972
```

Above are the variances for our data, our data under a box cox transformation, and our data under a log transform respectively. I see that each transformation improved our variance, as expected. However, the log transformation produces the most desireable results.

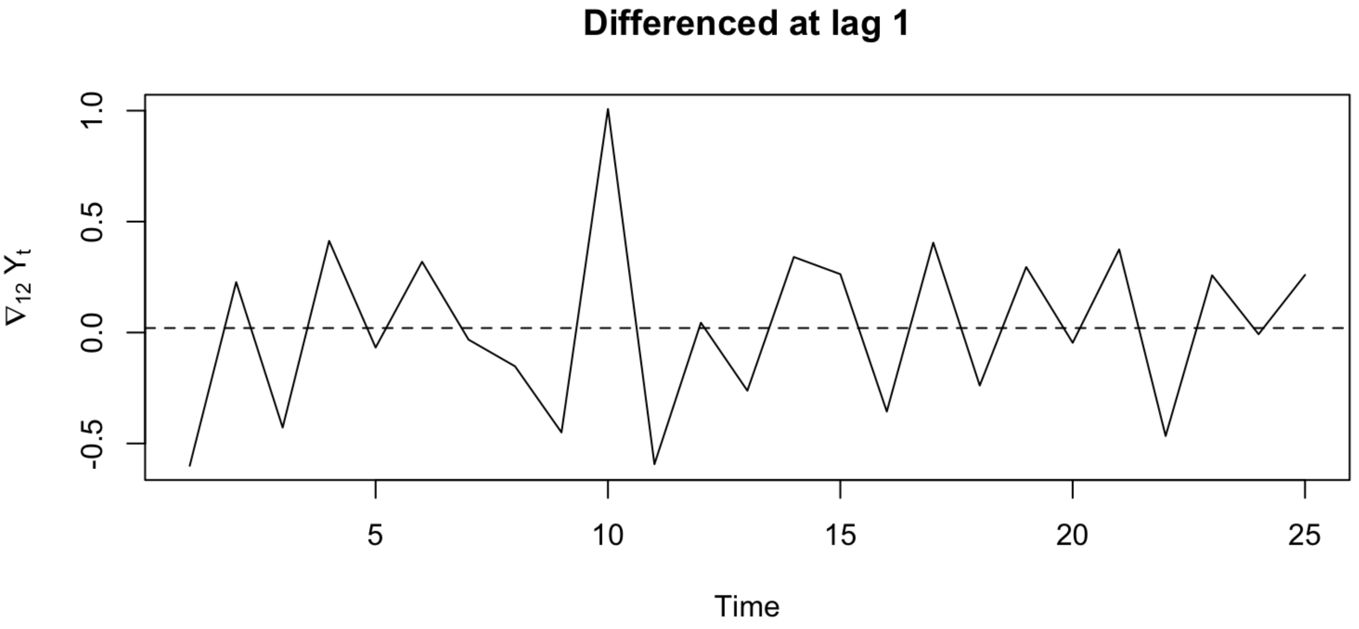


This appears to have a more normal distribution when compared to our initial histogram. This was made after performing the log transformation; which produces a much more symmetric histogram than originally plotted.

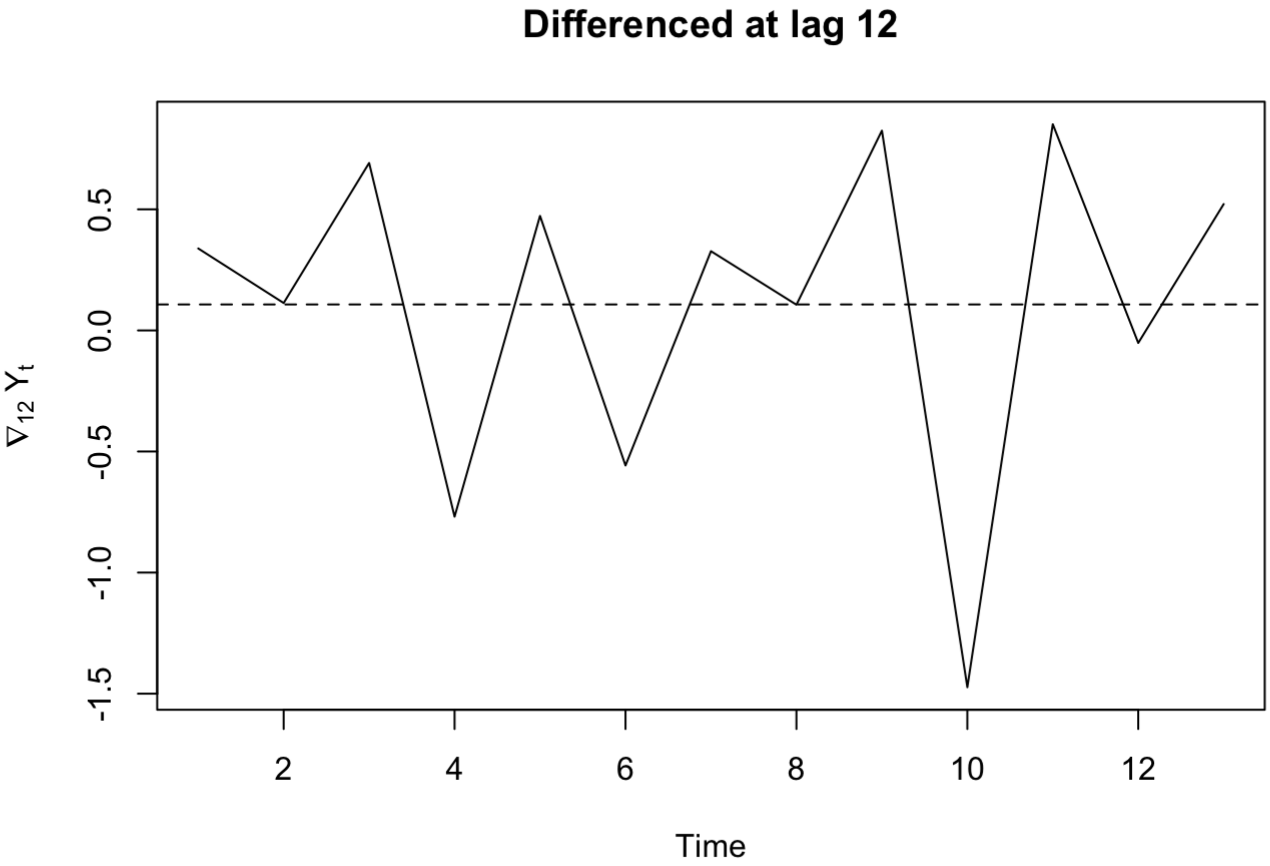
Plotting our transformed data



Difference to remove the trend



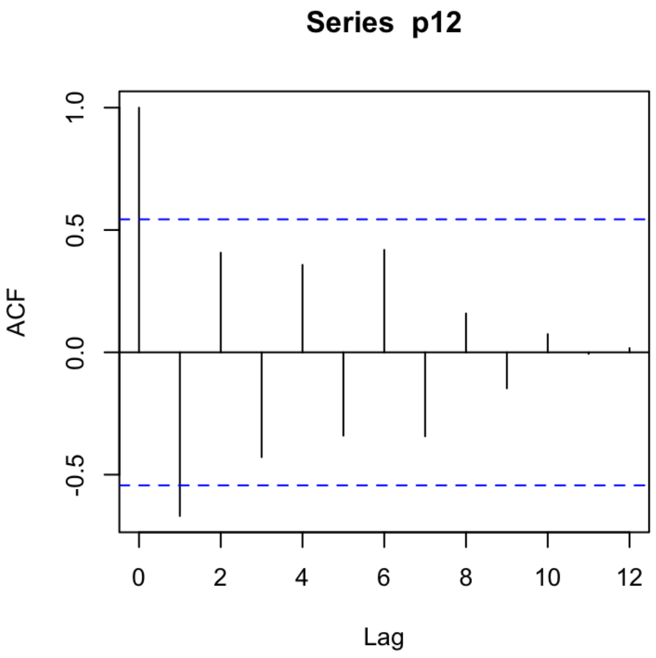
Differencing at lag 1 removes the linear trend.

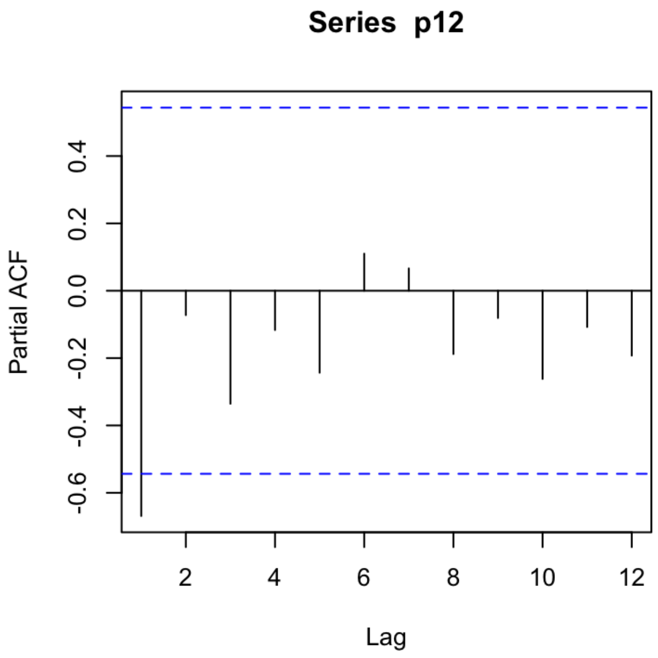


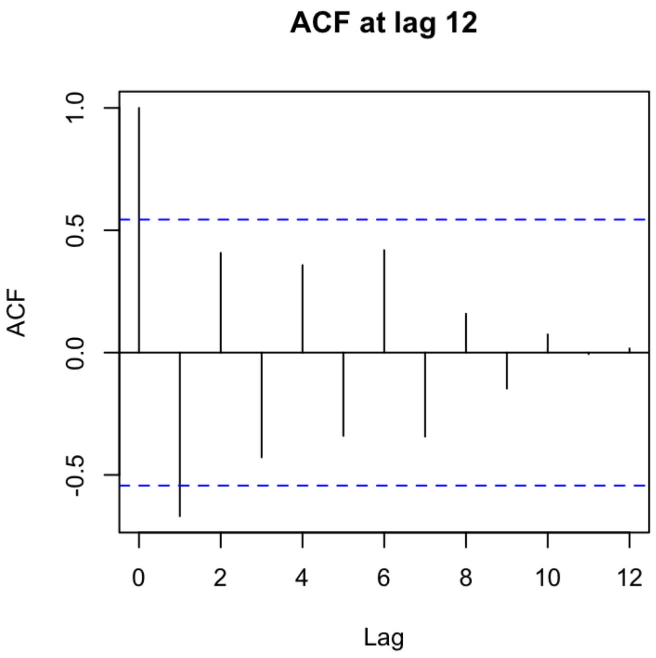
Differencing at lag 12 has removed seasonality, as desired.

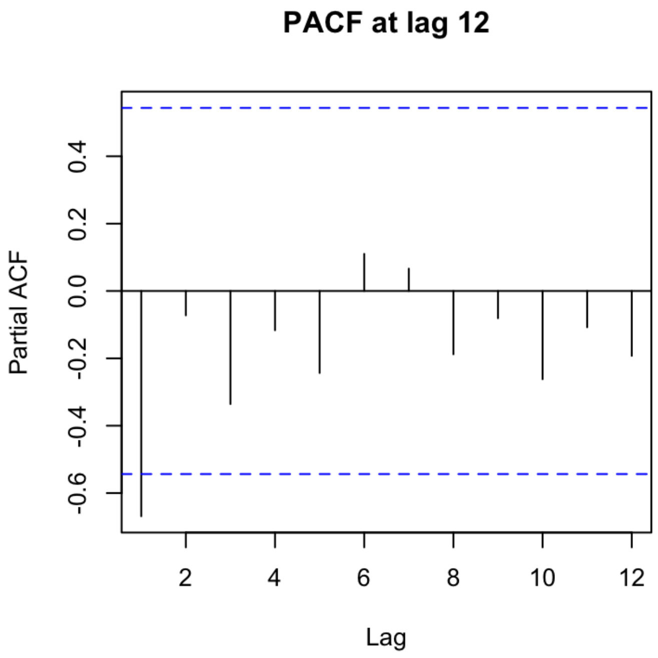
Plotting the Differenced Data at different lags

Now, I plot the ACF at different lags and base my model predictions off of the following output.









Looking at our ACF plot, there appears to be:

It is differenced at lag 1 so d = 1.

On the ACF chart up to lag 50, we can see significant lags at 1, 2, and 3. So good guesses for q are 1, 2, or 3. On the full PACF chart, we can see a significant lag at 1. Meaning, that it could also be a pure moving average model. I guess are 0 and 1 for p. This gives us candidate models: ARIMA(0,1,1) ARIMA(0,1,2), ARIMA(0,1,3), ARIMA(1,1,1), ARIMA(1,1,2), ARIMA(1,1,3). These were obtained by looking at the plots of the acf and pacf.

Model Estimation

After gathering our predictions from the ACF/PACF above, we want to compare them to see which works best with our dataset.

Test to see a model prediction

```
## Series: shampoo.log
## ARIMA(1,1,0)
##
## Coefficients:
##           ar1
##       -0.6584
## s.e.   0.1539
##
## sigma^2 = 0.08848:  log likelihood = -4.94
## AIC=13.87   AICc=14.42   BIC=16.31
```

The auto.arima function provides us with an estimation ARIMA(1,1,0). Now I will run diagnostics to see if there exists a better model.

```
##
## Call:
## arima(x = shampoo.log, order = c(0, 1, 1), seasonal = list(order = c(0, 0, 0),
##     period = 12))
##
## Coefficients:
##           ma1
##       -0.6803
## s.e.   0.1205
##
## sigma^2 estimated as 0.08653:  log likelihood = -5.19,  aic = 14.39
```

```
##
## Call:
## arima(x = shampoo.log, order = c(1, 1, 0), seasonal = list(order = c(0, 0, 0),
##     period = 12))
##
## Coefficients:
##           ar1
##       -0.6584
## s.e.   0.1539
##
## sigma^2 estimated as 0.08494:  log likelihood = -4.94,  aic = 13.87
```

```
##
## Call:
## arima(x = shampoo.log, order = c(0, 1, 2), seasonal = list(order = c(0, 0, 0),
##     period = 12))
##
## Coefficients:
##           ma1     ma2
##       -0.9660  0.3280
## s.e.   0.2809  0.2679
##
## sigma^2 estimated as 0.07858:  log likelihood = -4.17,  aic = 14.34
```

```
##
## Call:
## arima(x = shampoo.log, order = c(0, 1, 3), seasonal = list(order = c(0, 0, 0),
##     period = 12))
##
## Coefficients:
##           ma1     ma2     ma3
##       -0.9967  0.3037  0.0844
## s.e.   0.2268  0.2451  0.1704
##
## sigma^2 estimated as 0.07735:  log likelihood = -4.05,  aic = 16.11
```

```
##
## Call:
## arima(x = shampoo.log, order = c(1, 1, 1), seasonal = list(order = c(0, 0, 0),
##     period = 12))
##
## Coefficients:
##           ar1      ma1
##       -0.3892  -0.5026
## s.e.   0.2492   0.2262
##
## sigma^2 estimated as 0.079:  log likelihood = -4.15,  aic = 14.3
```

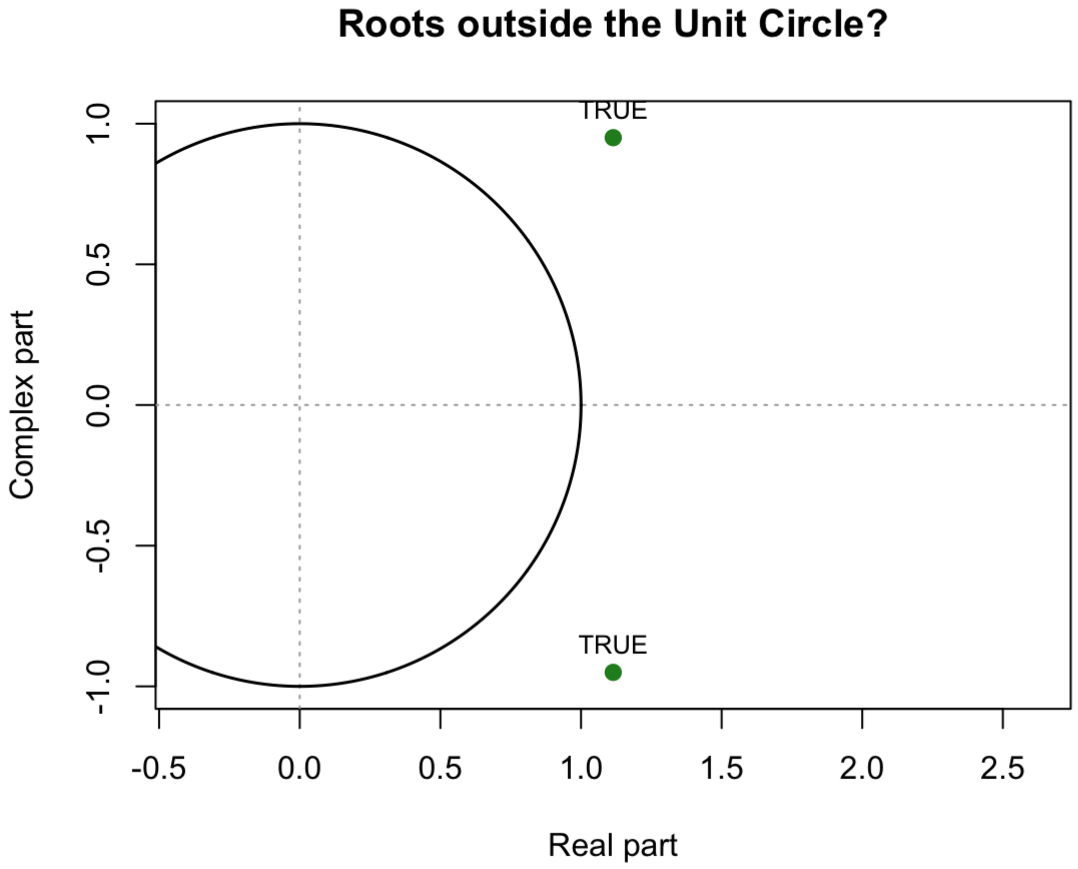
```
##
## Call:
## arima(x = shampoo.log, order = c(1, 1, 2), seasonal = list(order = c(0, 0, 0),
##     period = 12))
##
## Coefficients:
##          ar1      ma1     ma2
##       0.4242  -1.3697  0.6166
## s.e.  0.8558   0.6869  0.4000
##
## sigma^2 estimated as 0.07642:  log likelihood = -3.9,  aic = 15.81
```

```
##
## Call:
## arima(x = shampoo.log, order = c(1, 1, 3), seasonal = list(order = c(0, 0, 0),
##     period = 12))
##
## Coefficients:
##          ar1      ma1     ma2      ma3
##       0.9392  -1.9583  1.1693  -0.1469
## s.e.  0.0991   0.2833  0.4132   0.2558
##
## sigma^2 estimated as 0.06866:  log likelihood = -3.37,  aic = 16.74
```

Use Aikake Information Curve to select optimal values From the analysis above, it appears that our AIC is minimized at ARIMA(0,1,2). This also has only 2 coefficients, therefore this is the preferred model. This contradicts our auto.arima prediction of ARIMA(1,1,1) and presents us with an improved model.

Use unit circle test to check casual and invertible

```
##       real   complex outside
## 1 1.114436  0.950281    TRUE
## 2 1.114436 -0.950281    TRUE
## *Results are rounded to 6 digits.
```



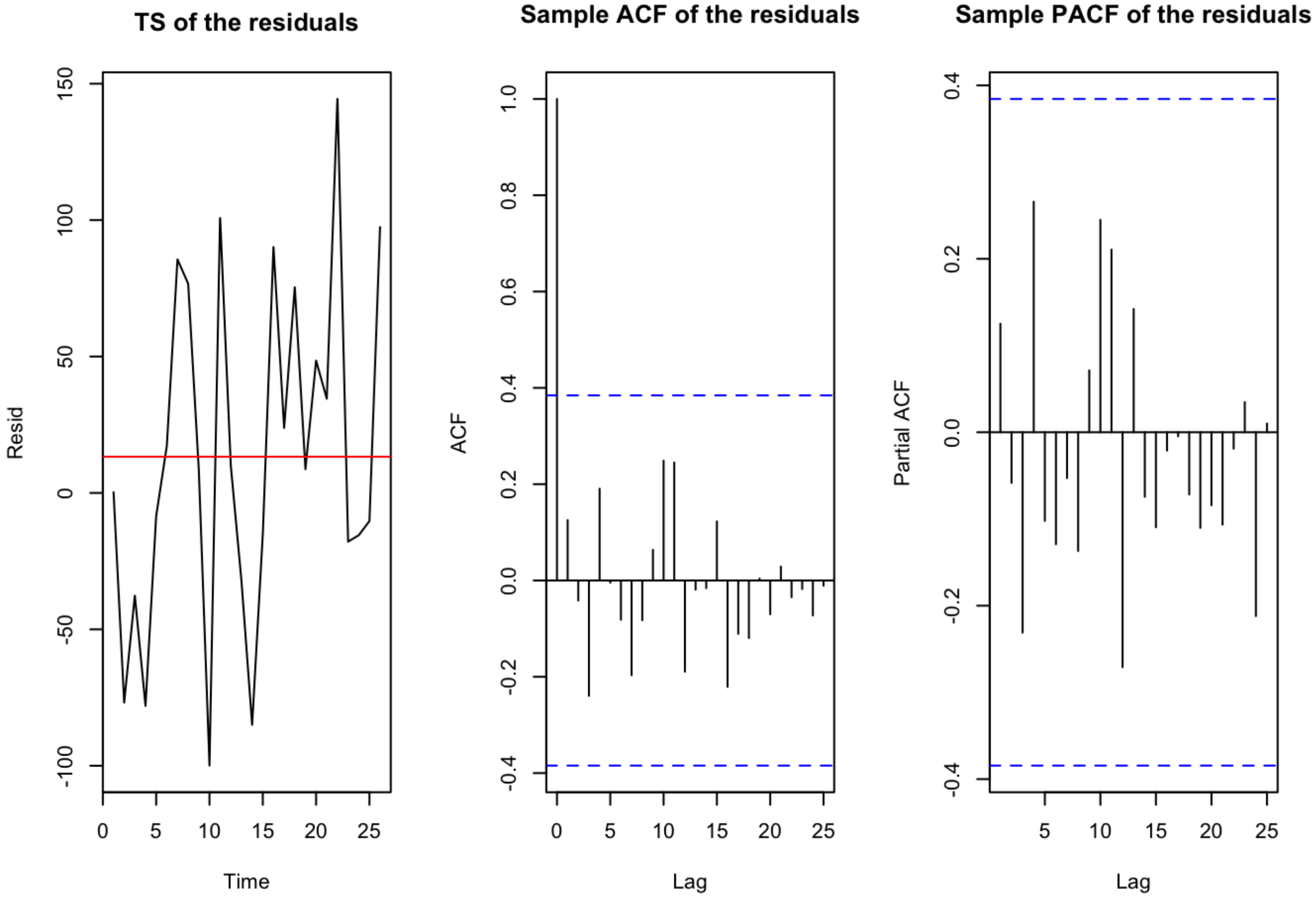
Using the unit circle test, we are able to conclude that our model is stationary and invertible. Our roots are outside of the unit circle.

Residual Analysis

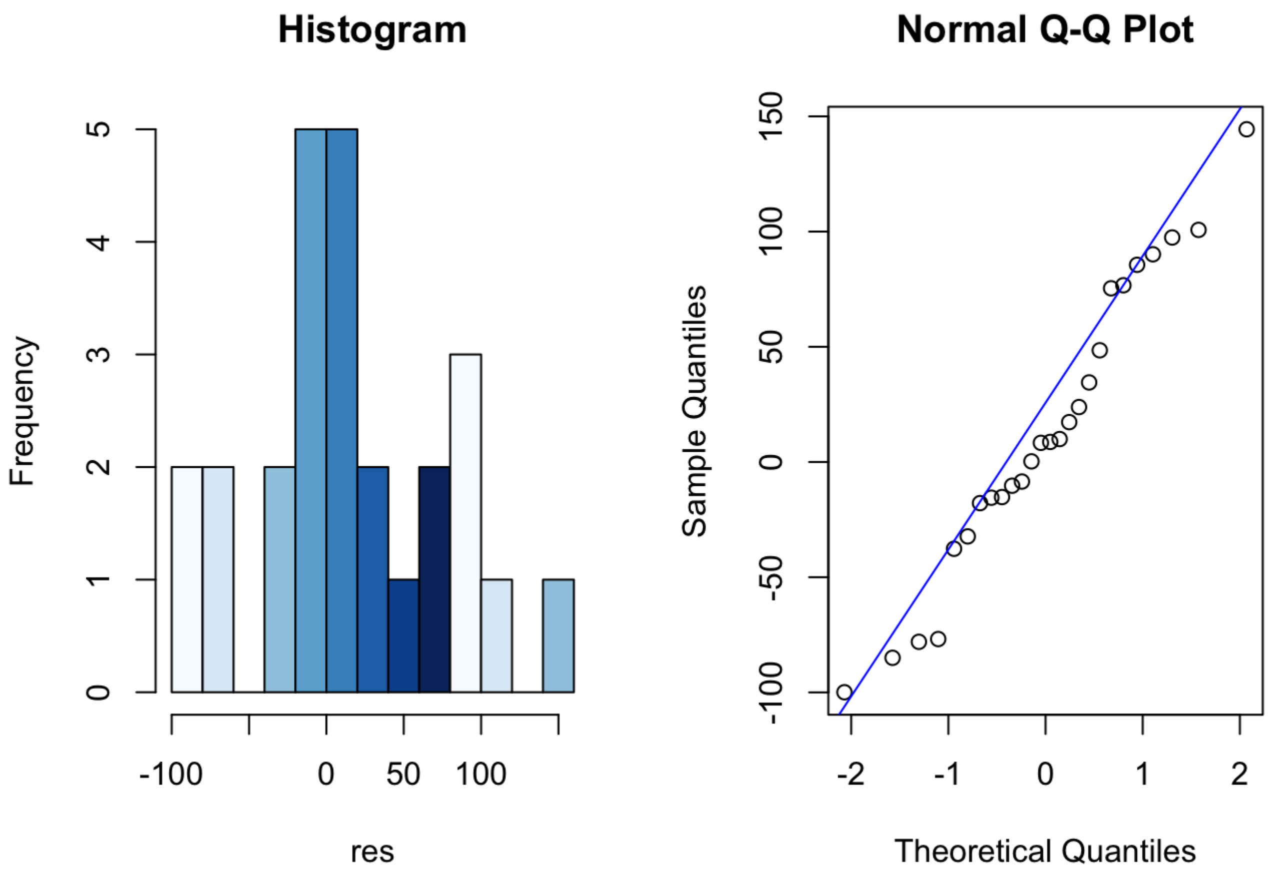
Mean and Variance of our Residuals

```
## [1] 13.25553
```

```
## [1] 4009.403
```



This appears to resemble white noise. There are no lags that break our confidence interval.



Our histogram looks approximately normal. This is corroborated by our qq-plot that appears to follow that of a normal distribution.

Now I run a series of diagnostic tests on our selected model to ensure that each test passes.

```
## 
##  Shapiro-Wilk normality test
## 
## data:  res
## W = 0.9699, p-value = 0.6208
```

Our p value is greater than 0.05. Based on the parameters of the Shapiro-Wilk normality test; we fail to reject the hypothesis that this is normally distributed.

Testing Residuals

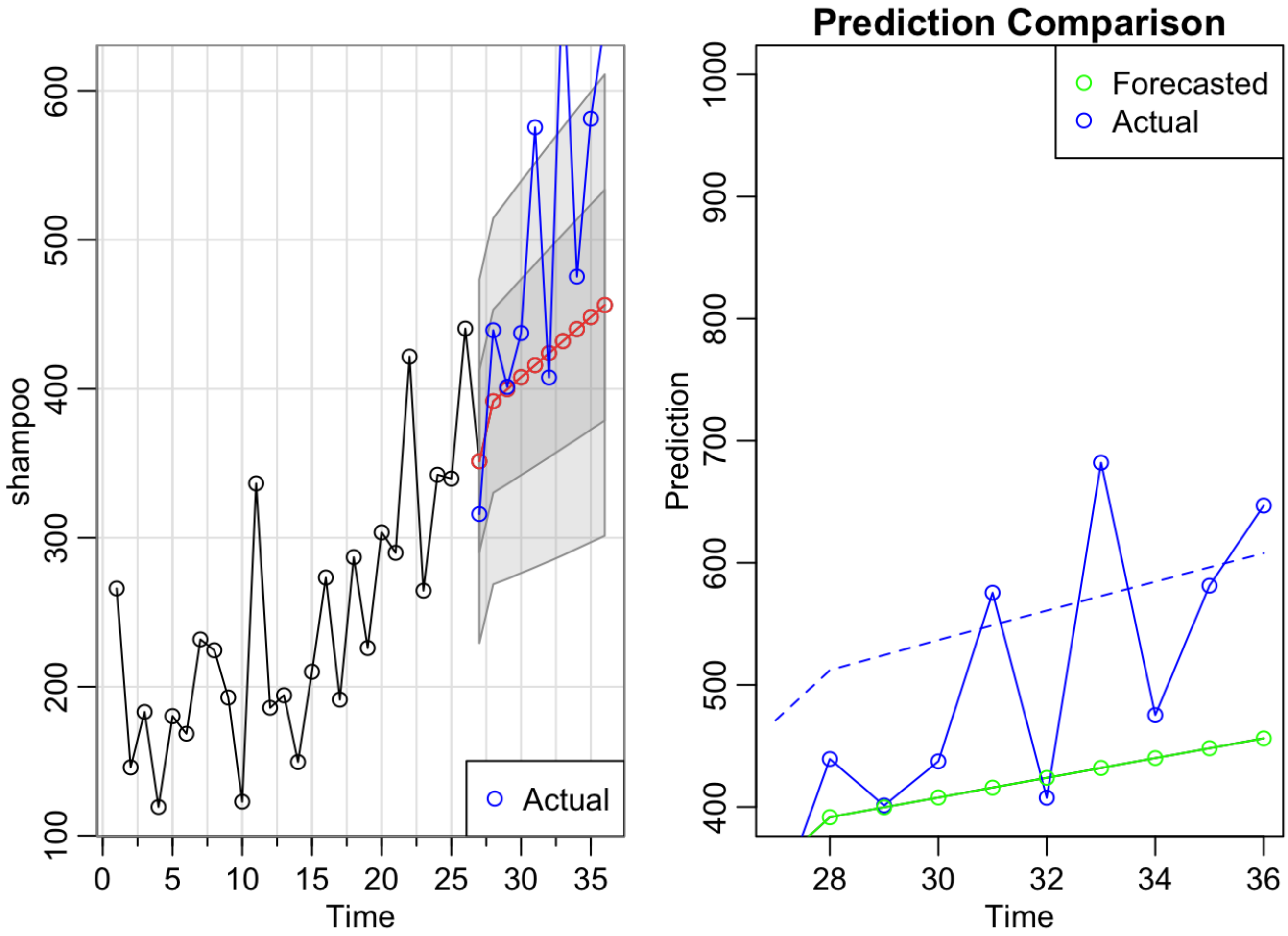
```
## 
##  Box-Pierce test
## 
## data:  res
## X-squared = 2.8897, df = 3, p-value = 0.4089
```

```
##
##  Box-Ljung test
##
## data:  res
## X-squared = 3.5282, df = 3, p-value = 0.3171
```

```
##
##  Box-Ljung test
##
## data:  res^2
## X-squared = 8.7359, df = 5, p-value = 0.1201
```

We know that our h is approximately 5 and that that 2 degrees of freedom are present for our 2 coefficients. Using this, we employ the portmanteau tests to further analyze our residuals. From the tests above, it appears that all of the p-values are sufficient and pass. Therefore, we can say that the residuals are independently distributed; meaning that they follow a white-noise model.

Predictions



As we can see, our forecasted values differentiate from the actual values. Our model definitely has some shortcomings. The first being being the variance of our actual values. The second would be the large spikes where our forecast leaves the 95% confidence intervals. It appears as if the true values overestimate the next 10 months, when compared to our forecast. My model does a good job forecasting the trend, however lacks some precision.

Conclusion

The goal of this project is to construct an accurate model to predict the next ten months of Shampoo sales based on a time series data set. In order to do this I stabilized the variance via a log-transformation and removed any seasonality by differencing our data. I then analyzed the ACF/PACF chart that was produced to develop a few predictions to test. Following this, we found the parameters and checked the residuals to ensure independence and normality. After I concluded that the residuals follow white noise, I determined that found that the model $(1 - 1.0391B + 0.4662B^2)Z_t$ is the best model out of our predictions. When we forecast the values, I see that the actual values have much more variance and tend to go above our 0.95 confidence interval.

While this project has not been easy, I have enjoyed the opportunity to learn a new way to analyze data. I've enjoyed learning about such an applicable

Sources 1) Pstat 174 Lab and Lecture Material 2) Shampoo Sales Dataset by Makridakis, Wheelwright, and Hyndman (1998). https://www.kaggle.com/redwankarimsony/shampoo-saled-dataset (https://www.kaggle.com/redwankarimsony/shampoo-saled-dataset)

Code Appendix

```r
knitr::opts_chunk$set(echo = TRUE)

library(MASS)
library(forecast)
library(tsdl)
library(MuMIn)
library(astsa)

shampoo.load <- read.csv('/Users/vancesine/Downloads/shampoo_sales.csv')

shampoo <- shampoo.load$Sales[1:26]
test <- shampoo.load$Sales[27:36]


class(shampoo)

#We make a test set of the final 10 observations to be used in our forecasting

#plot the data and note any important visual queues
ts.plot(shampoo, main = "Shampoo Time Series", xlab = "Months since start of data", ylab
= "Monthly Shampoo Sales" )
#plot histogram
hist(shampoo, main = "Histogram of Monthly Shampoo Sales", xlab= "Monthly Shampoo Sales"
, c = blues9)
library(MASS)
t <- 1:length(shampoo)
fit <- lm(shampoo ~ t)
bcTransform <- boxcox(shampoo ~ t, plotit = TRUE)
lambda <- bcTransform$x[which(bcTransform$y == max(bcTransform$y))]
lambda
#find lambda for box-cox transform
shampoo.bc <- (shampoo^lambda-1)/lambda
shampoo.log <- log(shampoo)
par(mfrow=c(1,3))
#plot the original data and compare with transformed data
ts.plot(shampoo, main = "Original Data", ylab = expression(X[t]))
ts.plot(shampoo.log, main = "Log Transform Data", ylab = expression(Y[t]))
ts.plot(shampoo.log, main = "Box-Cox Transform Data", ylab = expression(Z[t]))
#compare variances of transformations
var(shampoo)

var(shampoo.bc)
var(shampoo.log)
#plot histogram of transformation to ensure more normality
hist(shampoo.log, main = "Histogram of Monthly Shampoo Sales Log Transform", xlab= "Mont
hly Shampoo Sales", c = blues9)
par(mfrow = c(1,2))
#plot and compare the acf and pacf for our transformed data
acf(shampoo.log, lag.max = 36, main = "ACF")
pacf(shampoo.log, lag.max = 36, main = "PACF")
par(mfrow = c(1,1))
#difference at lag one to remove linear trend
```

```
p1 <- diff(shampoo.log,1)
ts.plot(p1, main = "Differenced at lag 1",
     ylab = expression(nabla[12]~Y[t]))
abline(h = mean(p1), lty = 2)

#difference at 12 to remove seasonality
par(mfrow = c(1,1))
p12 <- diff(p1, 12)
ts.plot(p12, main = "Differenced at lag 12",
        ylab = expression(nabla[12]~Y[t]))
abline(h = mean(p12), lty = 2)
par(mfrow = c(2,2))
#plot acf at different lags
acf(p12, lag.max = 50, )
pacf(p12, lag.max = 50,)
acf(p12, lag.max = 12, main = "ACF at lag 12")
pacf(p12, lag.max = 12,main = "PACF at lag 12" )
library(forecast)
auto.arima(shampoo.log)
(arima(shampoo.log,order=c(0,1,1),seasonal=list(order=c(0,0,0), period=12)))
(arima(shampoo.log,order=c(1,1,0),seasonal=list(order=c(0,0,0), period=12)))
(arima(shampoo.log,order=c(0,1,2),seasonal=list(order=c(0,0,0), period=12)))
(arima(shampoo.log,order=c(0,1,3),seasonal=list(order=c(0,0,0), period=12)))
(arima(shampoo.log,order=c(1,1,1),seasonal=list(order=c(0,0,0), period=12)))
(arima(shampoo.log,order=c(1,1,2),seasonal=list(order=c(0,0,0), period=12)))
(arima(shampoo.log,order=c(1,1,3),seasonal=list(order=c(0,0,0), period=12)))
library(UnitCircle)
uc.check(pol_ = c(1,-1.0391, 0.4662), plot_output = TRUE)
fit <- arima(shampoo, order = c(0,1,2), seasonal = list(order =c(0,0,0), period =12), me
thod = "ML")
res <- residuals(fit)
mean(res);var(res)
res <- residuals(fit)
par(mfrow=c(1,3))
plot(res, xlab='Time',ylab='Resid',main="TS of the residuals"); abline(h = mean(res), co
l = "red")
acf(res, lag.max = 25, main = "Sample ACF of the residuals")
pacf(res, lag.max = 25, main = "Sample PACF of the residuals")

par(mfrow=c(1,2))
hist(res,main = "Histogram",breaks = 10, c= blues9)
qqnorm(res)
qqline(res,col ="blue")
shapiro.test(res)
Box.test(res, lag = 5, type = c("Box-Pierce"), fitdf = 2)
Box.test(res, lag = 5, type = c("Ljung-Box"), fitdf = 2)
Box.test(res**2, lag = 5, type = c("Ljung-Box"), fitdf = 0)
par(mfrow=c(1,2))
pred.tr <- sarima.for(shampoo, n.ahead = 10, plot.all = F, p = 0, d = 1, q = 2, P = 0, D
= 0, Q = 0, S = 0)
lines(27:36, test, col="blue")
points(27:36, test, col="blue")
```

```
legend("bottomright", pch=1, col=c("blue"),
legend=("Actual"))
plot(pred.tr$pred, main = "Prediction Comparison", ylim = c(400, 1000), xlim = c(27, 36
), ylab = "Prediction")
lines(27:36, pred.tr$pred+1.96*pred.tr$se,lty=2, col = "blue")
lines(27:36, pred.tr$pred-1.96*pred.tr$se,lty=2, col = "blue")
lines(27:36, pred.tr$pred, col="green")
lines(27:36, test, col="blue")
points(27:36, test, col="blue")
points(27:36, pred.tr$pred, col = "green")
legend("topright", pch=1, col=c("green", "blue"),
legend=c("Forecasted", "Actual"))
```