U.S. Car Sales 1960-1968

Michael Cambaliza

6/1/2021

Executive Summary

For this project I will be looking into U.S. Car Sales (in millions) from 1960 to 1968 which is split up into months and you can find the data set at Kaggle.com. I will create a model that can forecast future data points based off the raw data we are given. I will be splitting my data set into a training set and test set so I can compare how well my model fits once I reach the forecasting step. I first checked to see if my data is skewed and applied necessary transformations and in this case none were needed. I then checked for seasonality and trend and applied necessary differencing to make my data stationary and viewed the variance as I went along to account for overdifferencing. Once I made my data stationary I then looked at the ACF and PACF to preliminary identify possible models. Once I found possible models I then chose the two best models based off of their AICc and used diagnostic checking on each of them to choose the most adequate model and then proceeded to forecast future data points. Finally, I was then able to compare my predicted data points that I obtained from my model to my test set and checked how accurate my model was.

From my results I was able to determine that the model I chose which was a $SARIMA(0,1,1)\text{x}(0,1,1)s_{12}$, was accurate in forecasting because the entirety of my test set was contained in my predictive intervals. I predicted the next 12 data points which is the year 1968 and my model was able to accurately predict the next year.

Introduction

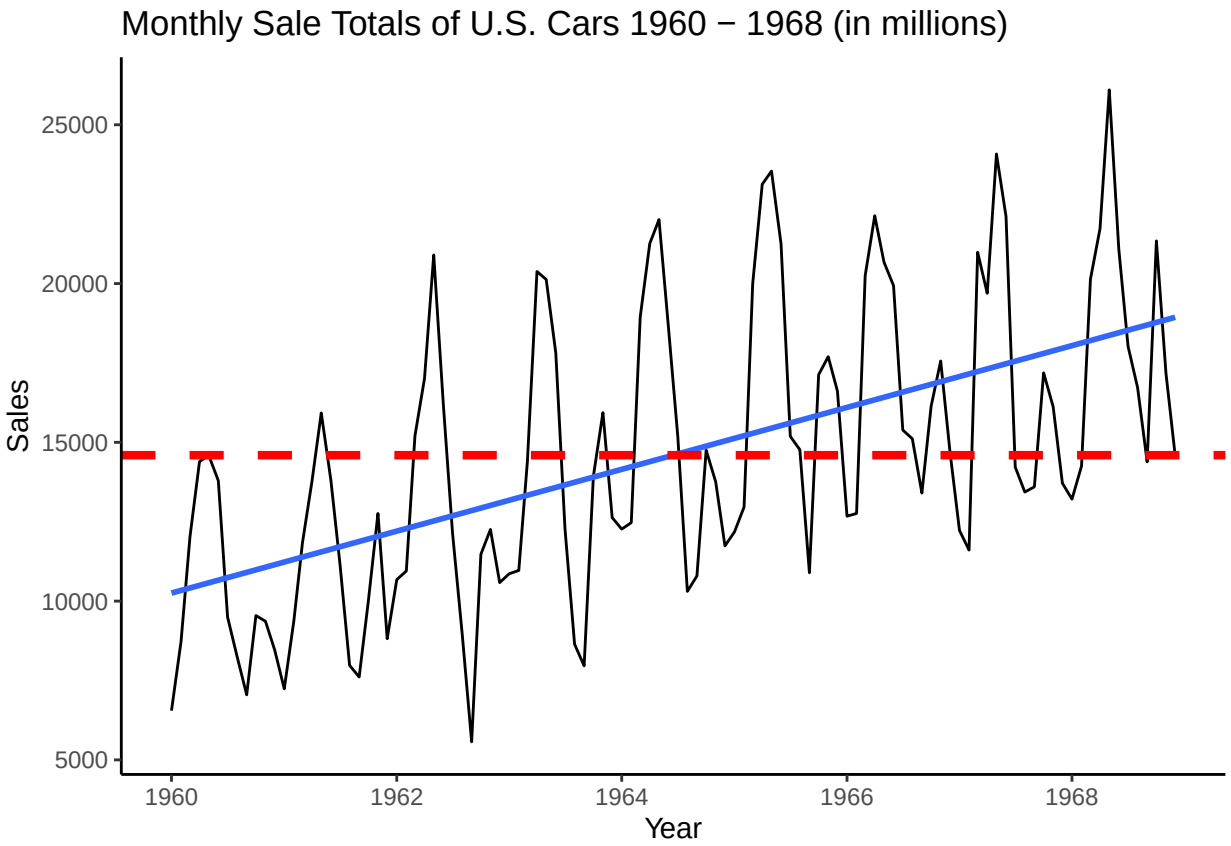
The data set I will be using is U.S. Car Sales (in millions) from 1960-1968 from Kaggle.com. I chose this dataset because I wanted to work with financial data and I have yet to do any kind of analysis that involves forecasting future sales and I plan on working in an industry that uses forecasting for costs, profits, and sales. I felt like this data set would introduce me to how sales data can act and how to tackle this kind of dataset in the future. I will use this dataset to forecast future U.S. car sales. I will also be splitting this data up into a training set and test set to compare later during forecasting.

Firstly, I wanted to check if my data was skewed. To handle the problem of skewed data I plotted histograms to check if my data was skewed or not. From my results it looked like a transformation was not necessary because the raw data already looked gaussian. I then checked for trend and seasonality by plotting my data in a line plot. It is was obvious that my data contained trend and seasonality and my assumptions were clarified after looking at the decomposition plot of my data. I proceeded to handle the trend issue by differencing at lag 1 and then differencing at lag 12 to handle the seasonality issue. After each difference I checked the variance to make sure it was decreasing to make sure I was not overdifferencing. Once I was able to make my data stationary I then looked at my ACF and PACF plots to view possible model candidates. I settled on choosing two models, my first was a $SARIMA(0,1,1)\text{x}(0,1,1)s_{12}$ and the other followed a $SARIMA(0,1,3)\text{x}(0,1,1)s_{12}$. I then compared these two models by their AICc which were extremely close to one another so I decided to do diagnostic checking on each of them to help me choose the best model for my data. Each model passed all my tests so this became an issue for me because I was not sure which to

choose. I decided to stick with my first model,$SARIMA(0,1,1)\text{x}(0,1,1)s_{12}$, because it did have the lower AICc even if it was a small difference. I then used this model to forecast the next 12 points(1 year) and create predictive intervals off of what I obtained. I finally was able to include my test set and compare it to the intervals I had obtained from my model. To my delight all the points were within my predictive intervals and this proved that my model was adequate for forecasting my data. This analysis was performed in R with compatible packages to help create some of the plots shown.

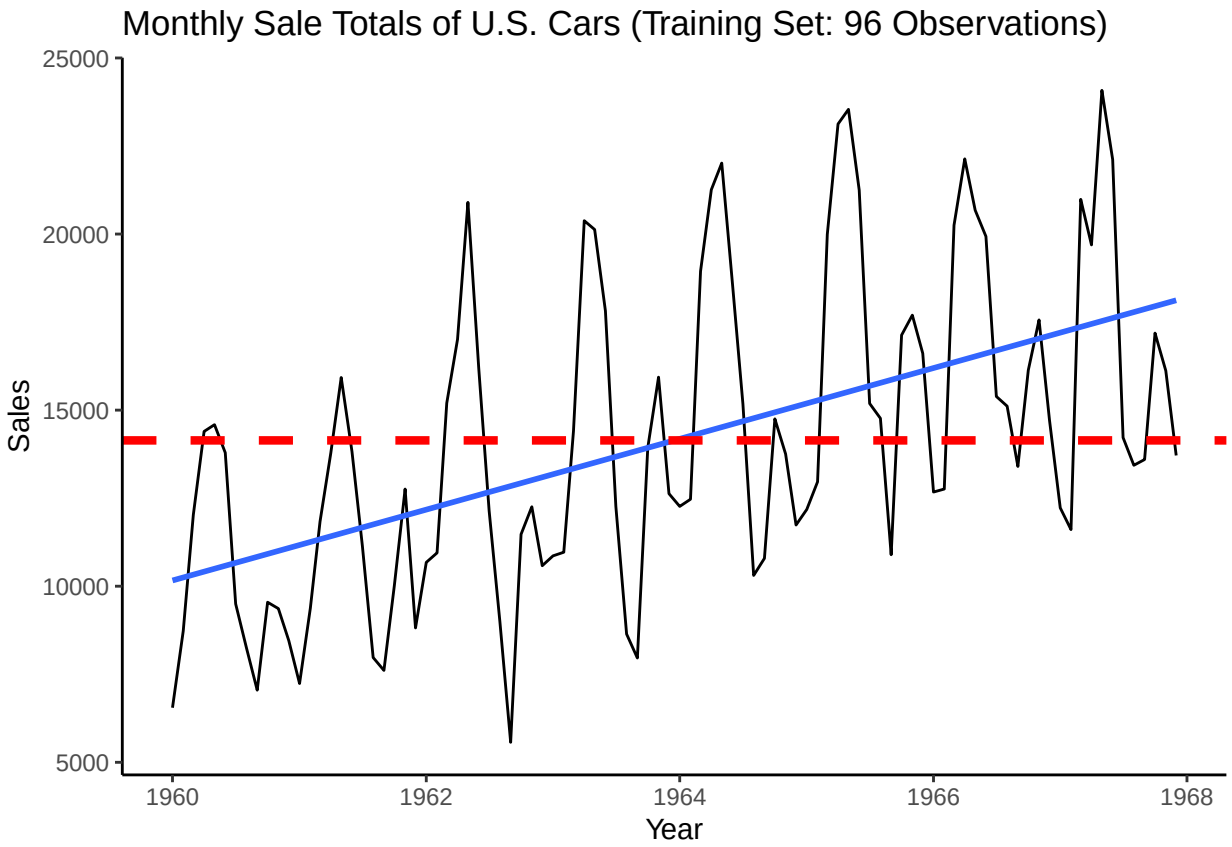
Process

I will begin by plotting my data to see any characteristics I should be aware of such as trend, seasonality, or any sharp changes in behavior.

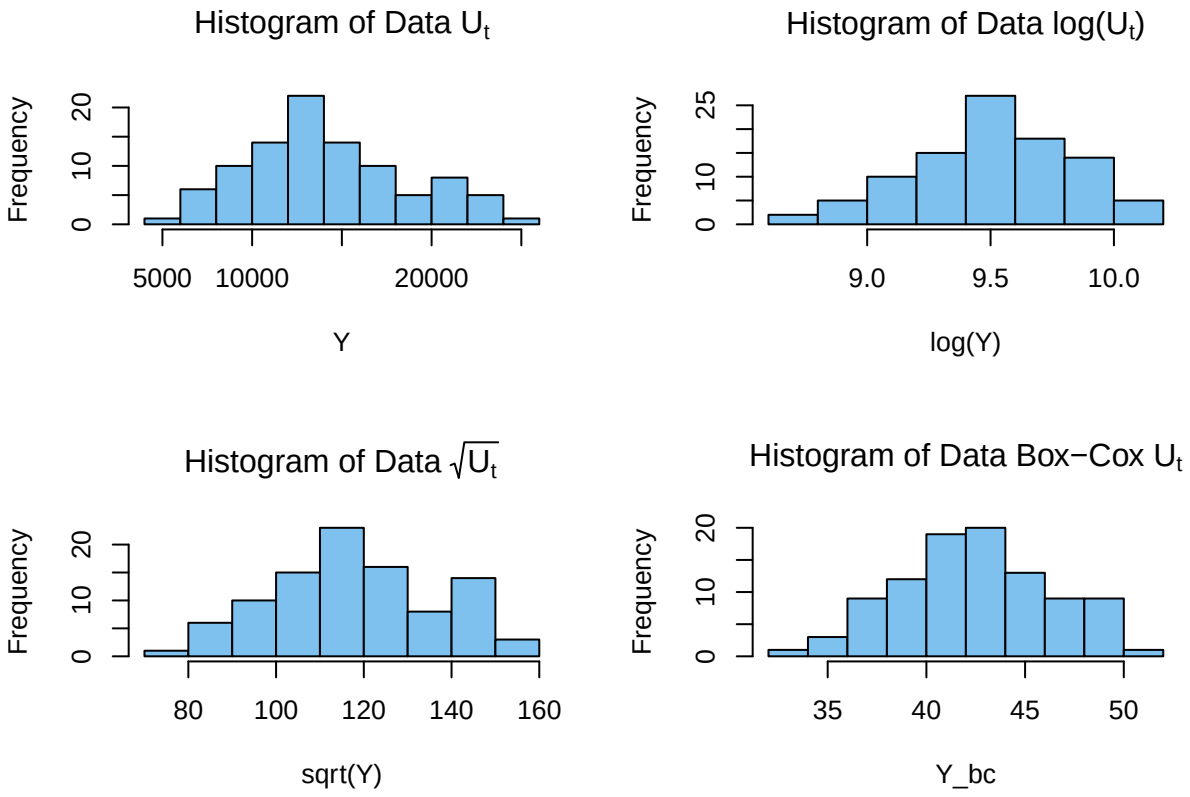


By looking at this plot you could see obvious positive trend and seasonality. Our blue line represents the trend and our red dashed line represents the sample mean of our data which is sitting around 15,000. I have concluded that this dataset represents a time series with trend and seasonality.

Before I move on to checking for transformations I want to split my data up into a training set and test set for model validation for later on during forecasting. I will leave out the year of 1968(12 data points) and continue model building with the year 1960-1967 (96 data points).

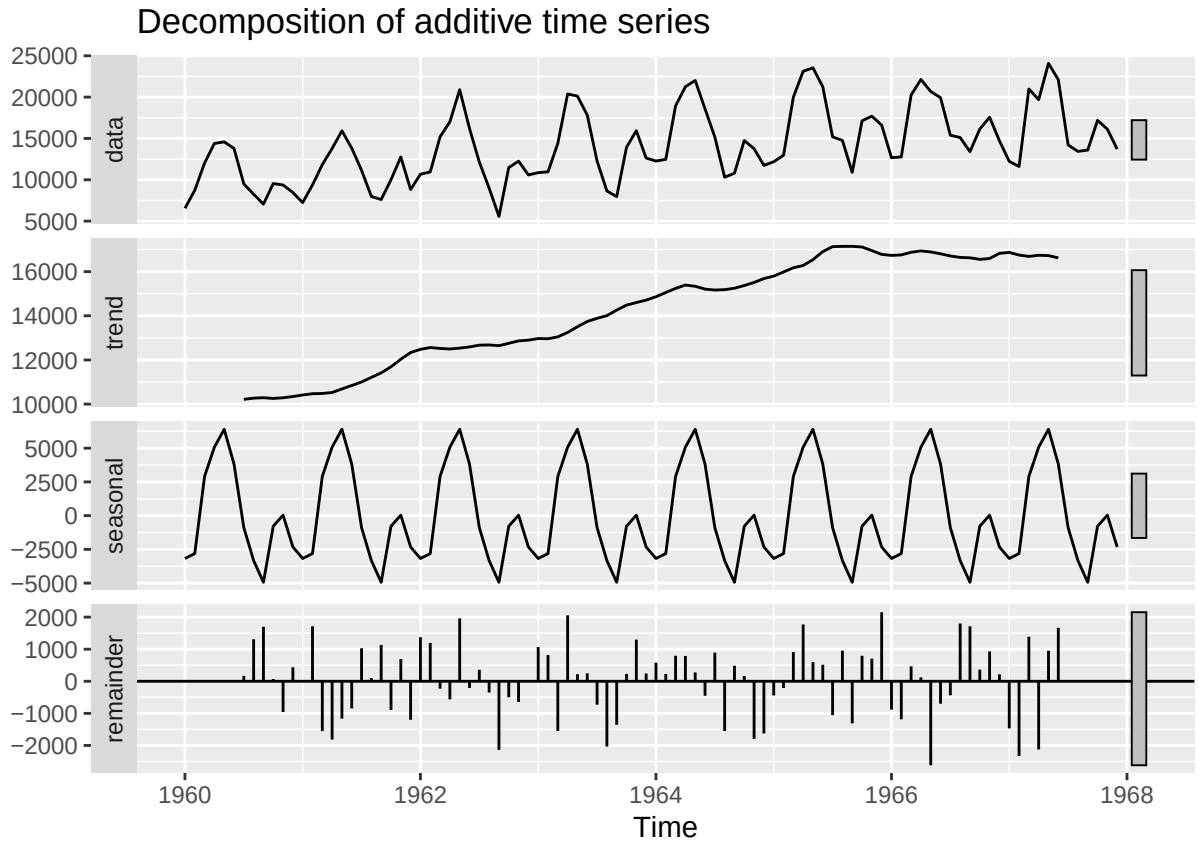


Before we handle trend and seasonality we want to check if we have skewed data and if it is necessary to transform it to stabilize our variance. I will check if our data is skewed by plotting histograms of the data and possible transformations that we could apply.



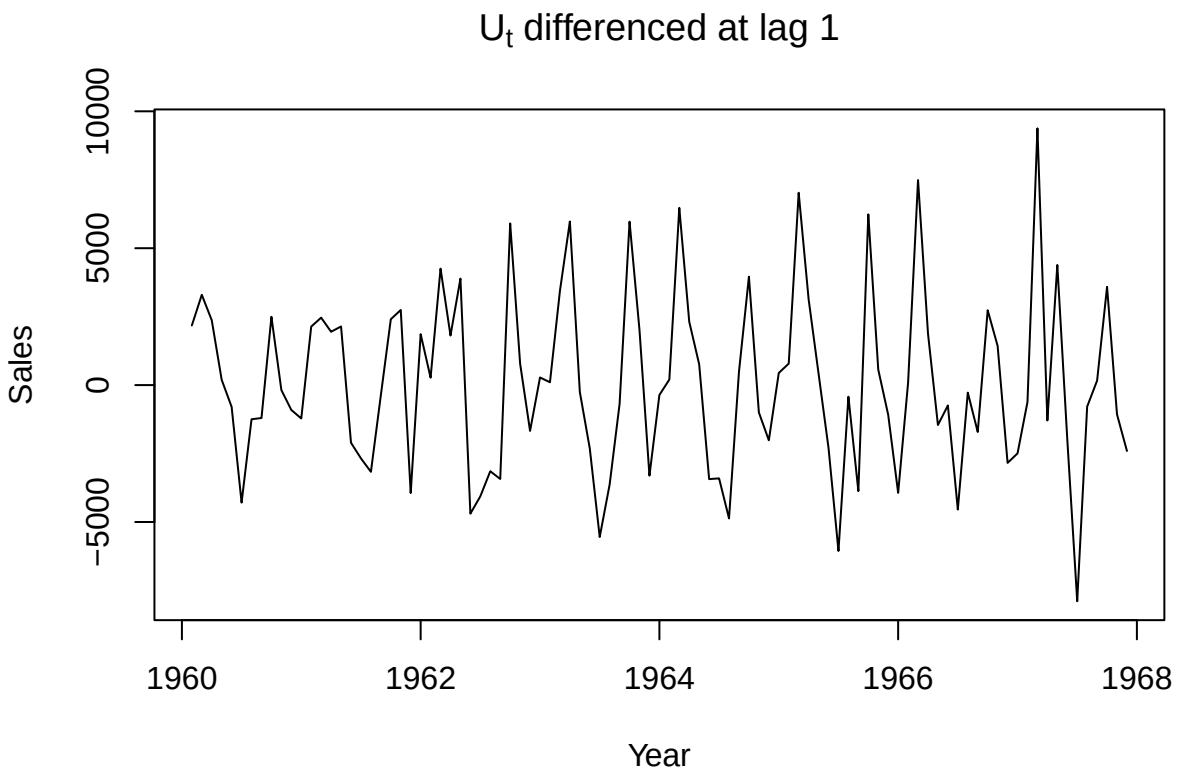
Our original data gives us a symmetric histogram and our variance is stabilized so a transformation may be unnecessary. I will continue without a transformation but if I run into any issues I can apply one of these and try model building with my chosen transformation.

To get another view of trend and seasonality we can create a decomposition plot that will show us individually trend and seasonality. This is a good way to visualize trend and seasonality if you're not sure if either existed. I will also show the variance as well because once we start differencing we will want to keep an eye on our variance to make sure it decreases because if not then we may be overdifferencing.




```
##          Sales
## Sales 19418113
```

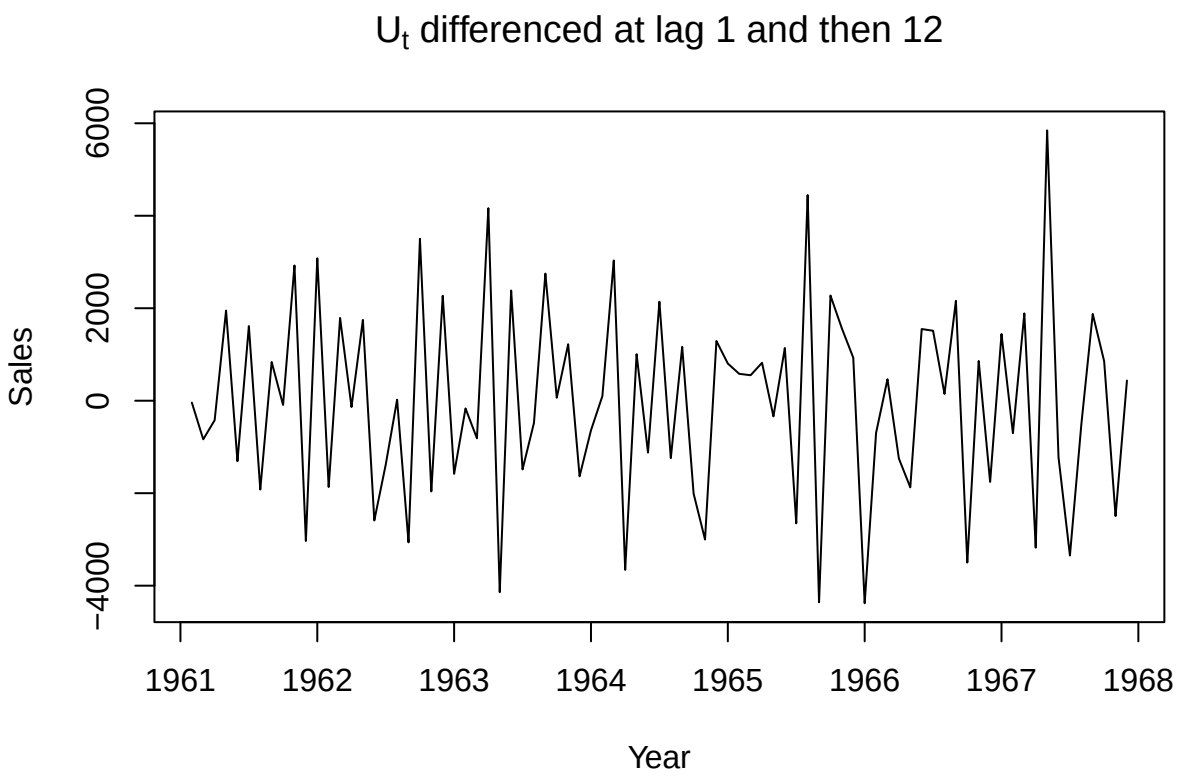
First we will handle the issue of trend by differencing at lag 1 once.




```
##          Sales
## Sales 10664848
```

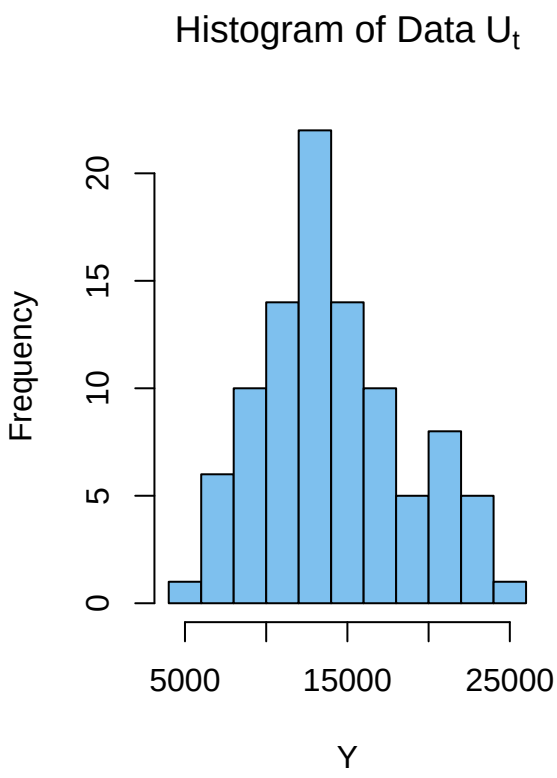
As you can see our trend is no longer present and our variance decreased as well. We will now have to deal with the seasonality issue that is still present.

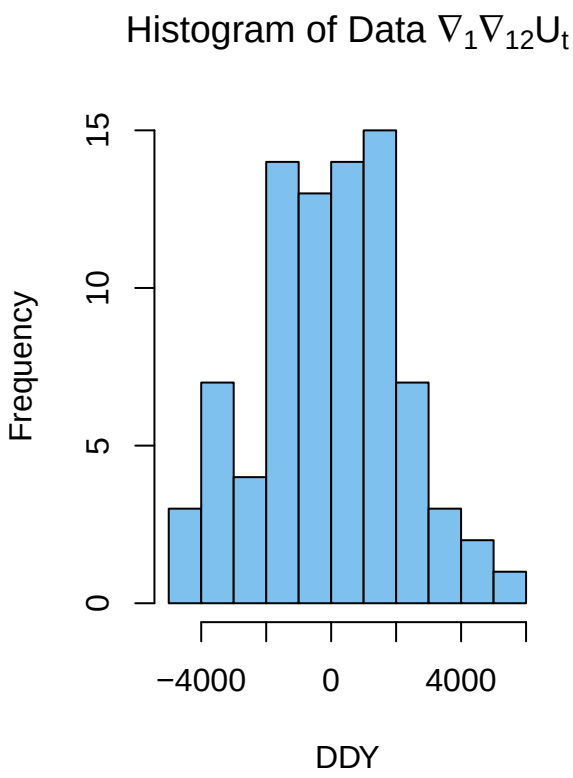
We will handle the issue of seasonality by differencing at lag 12 once. We difference at lag 12 because our data is split into 12 periods (Months of the year). We include the variance as well to check if we overdifferenced.




```
##         Sales
## Sales 4595505
```

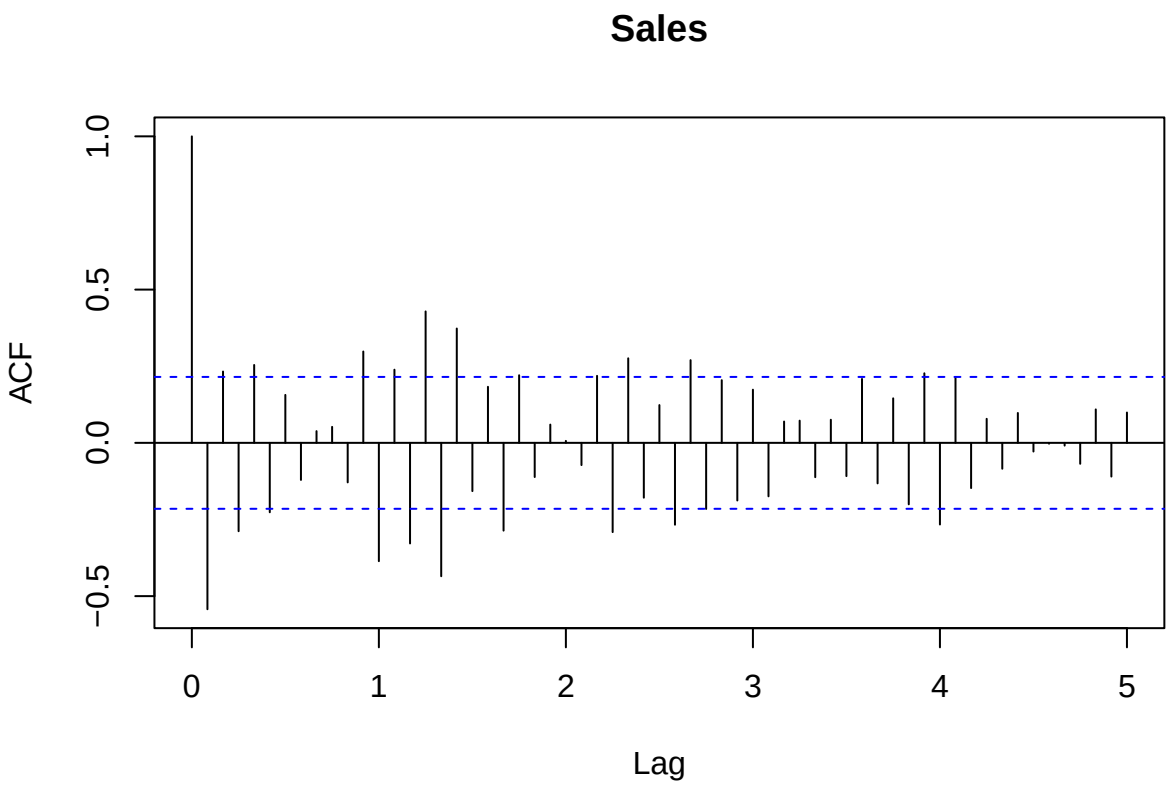
Our data now looks stationary and the variance has decreased as well so now we should check if our data is symmetric and gaussian.



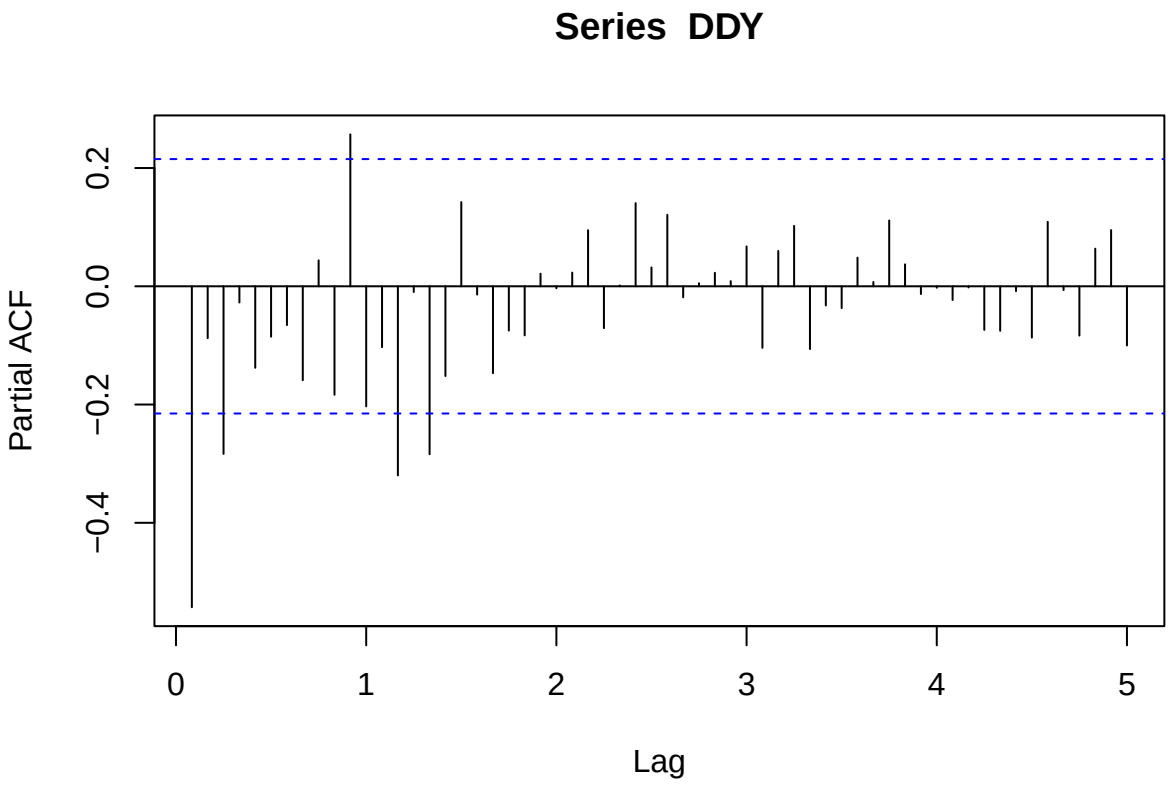


By looking at our stationary data you could see that the histogram is symmetric and almost gaussian with the mean centered around 0.

We will continue to look at our ACF and PACF to preliminary identify possible models.



ACF outside of confidence interval may be 1,3, or 4.



PACF outside of confidence intervals may be 1,3,8,11, or 14.

The possible models based off of our ACF and PACF plots may be $SARIMA(0,1,1)\text{x}(0,1,1)s_{12}$ or also $SARIMA(0,1,3)\text{x}(0,1,1)s_{12}$. I will look at each models AICc and compare to the two.

```
## [1] 1464.197
```

```
## [1] 1465.473
```

The AICc for for Model A and and Model B are shown above respectively. The AICc are relatively close to one another and I will continue with diagnostic checking for both models even though Model A has the lower AICc I would like to check if they both pass diagnostic checking.

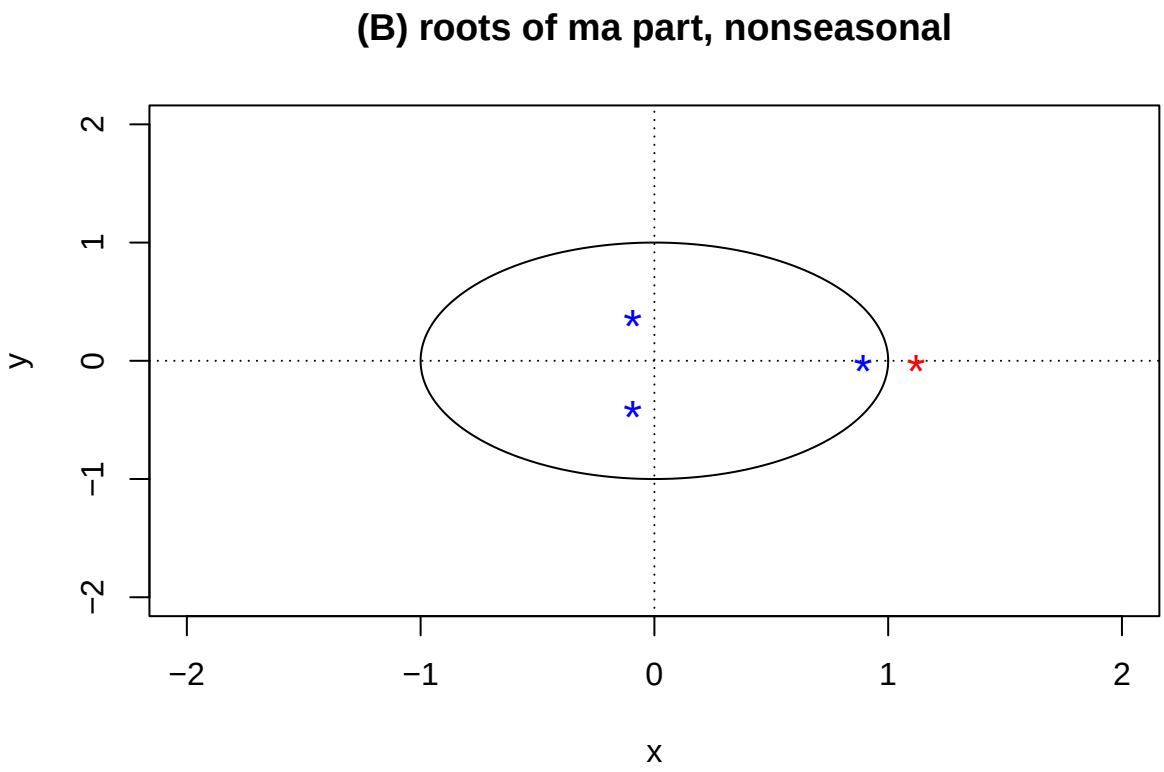
Model A: $(1 - 0.7833_{(0.1123)}B)(1 - 0.5335_{(0.1390)}B^{12})Z_t$

Model B: $(1 - 0.7077_{(0.1102)}B - 0.0095_{(0.1337)}B^2 - 0.1398_{(0.1049)}B^3)(1 - 0.4981_{(0.1401)}B^{12})Z_t$

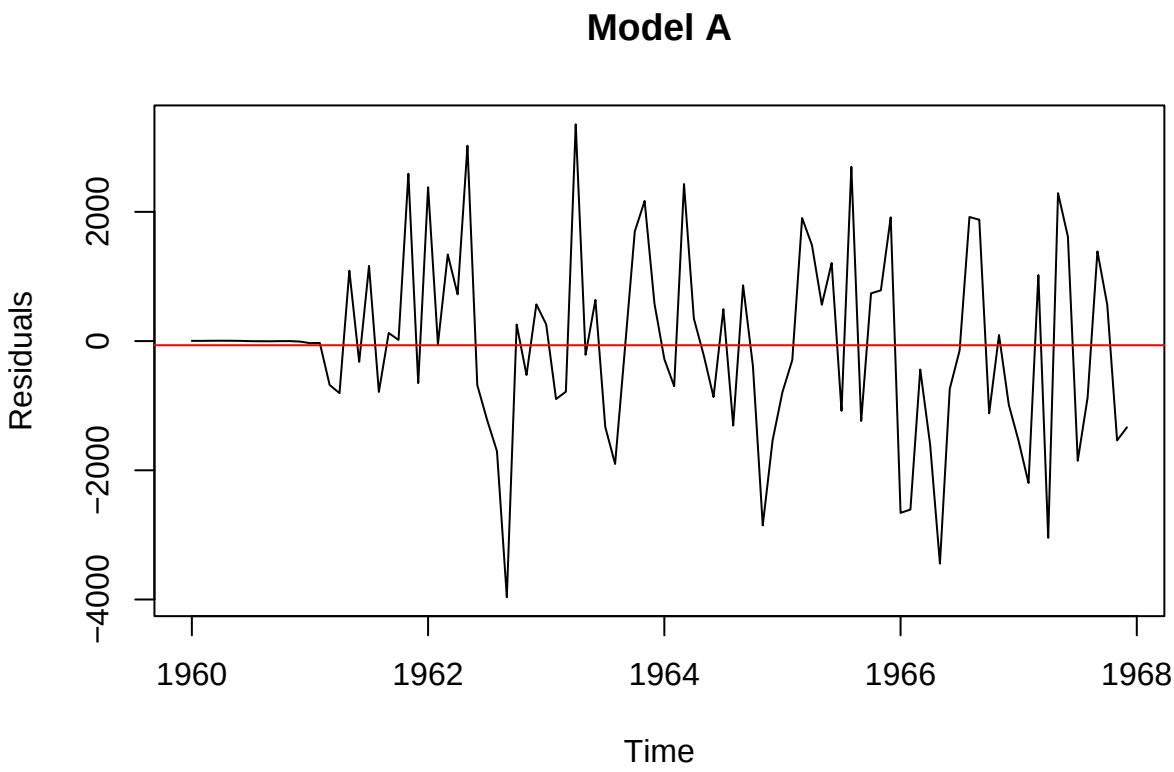
Both models are stationary because they are pure MA.

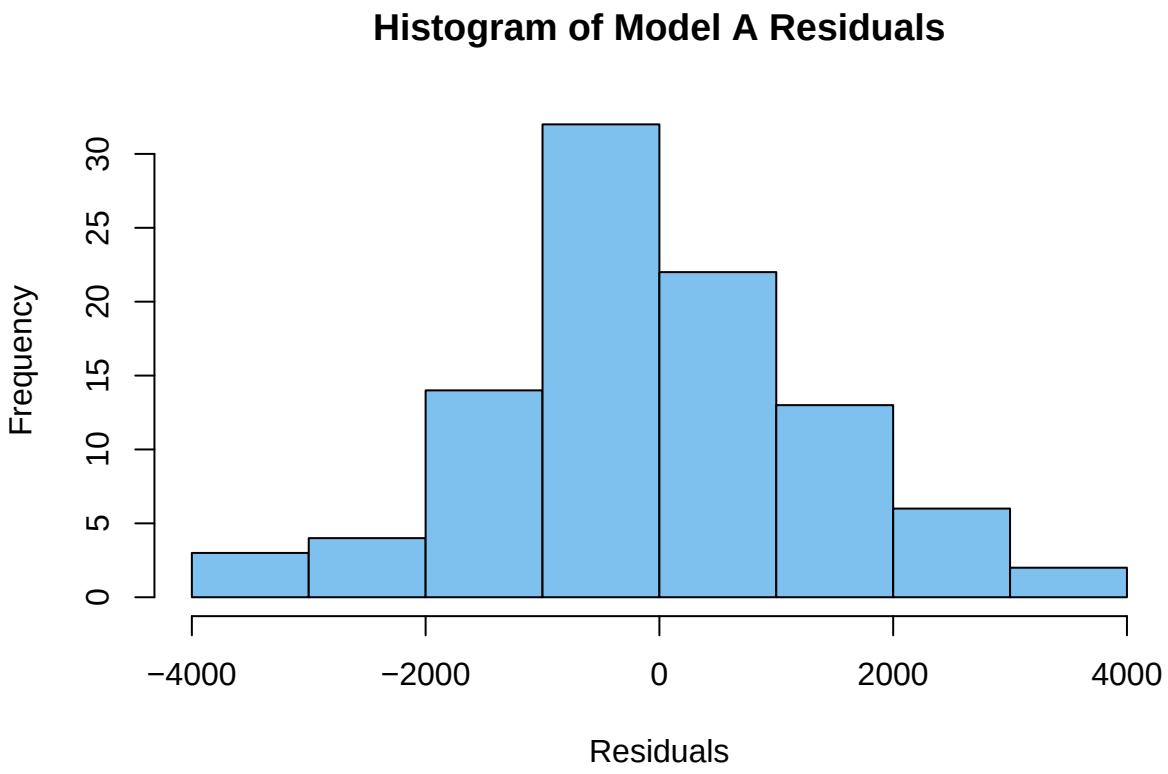
Model (A) is invertible because $|\theta_1| < 1$; $|\Theta_1 < 1|$

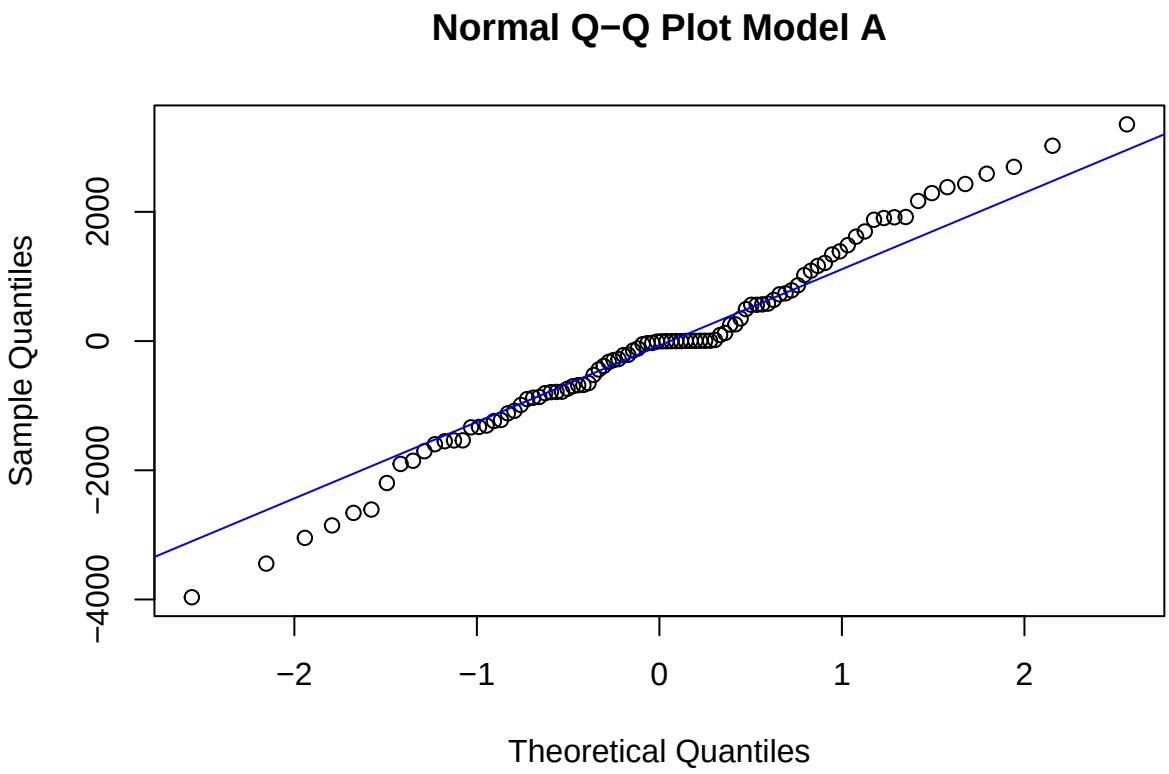
Model (B) is not invertible, all roots are not outside the unit circle:

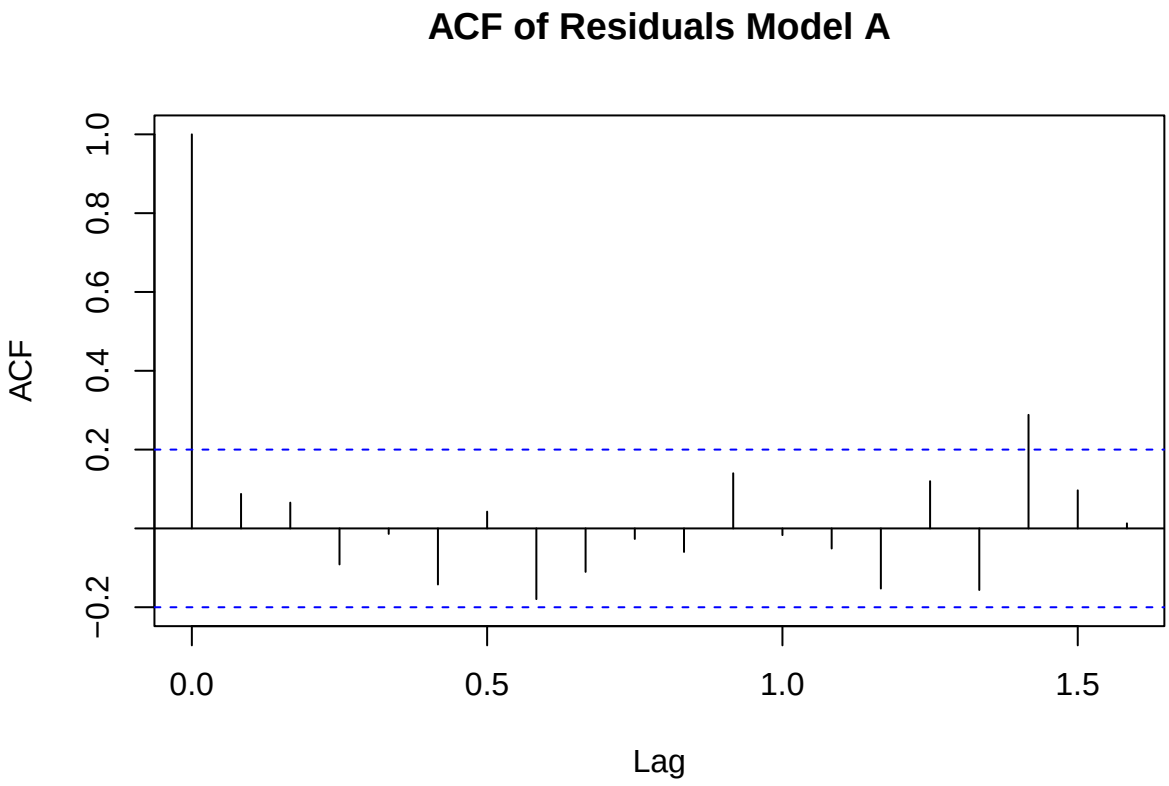


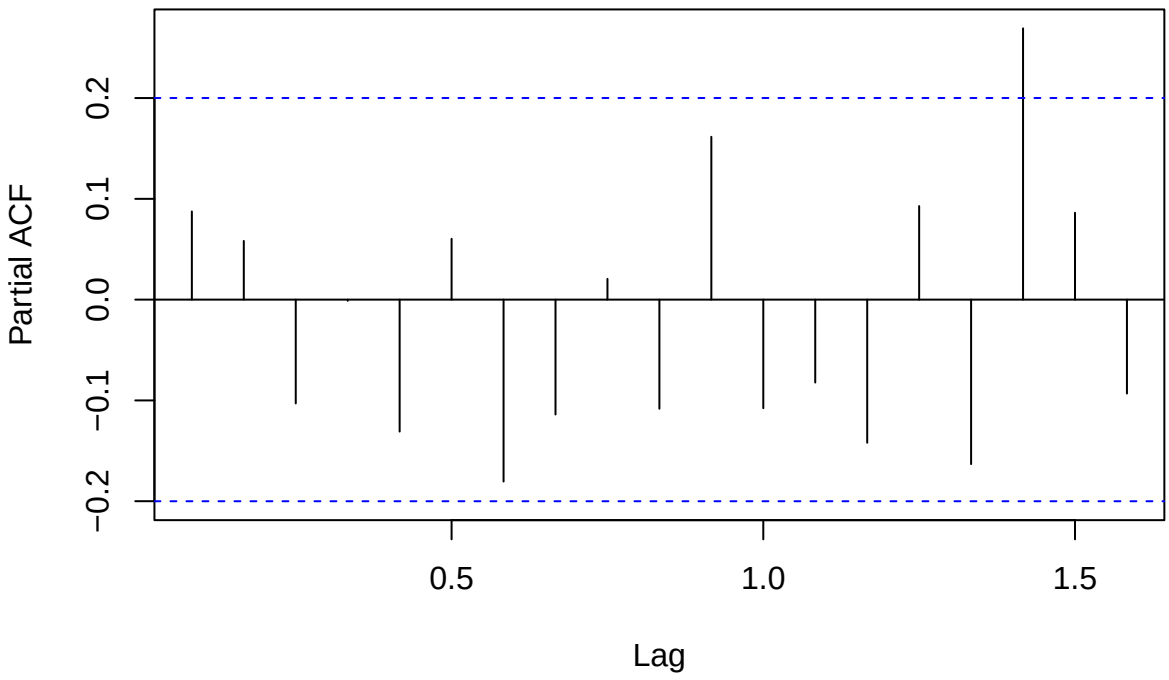
Now we will continue to diagnostic checking on each model to help us conclude which model will be more adequate for forecasting.


Model A
Residuals
2000
0
−2000
−4000
1960
1962
1964
1966
1968
Time


Histogram of Model A Residuals
Frequency
0
5
10
15
20
25
30
−4000
−2000
0
2000
4000
Residuals


Normal Q−Q Plot Model A
Sample Quantiles
Theoretical Quantiles
2000
0
−2000
−4000
−2
−1
0
1
2

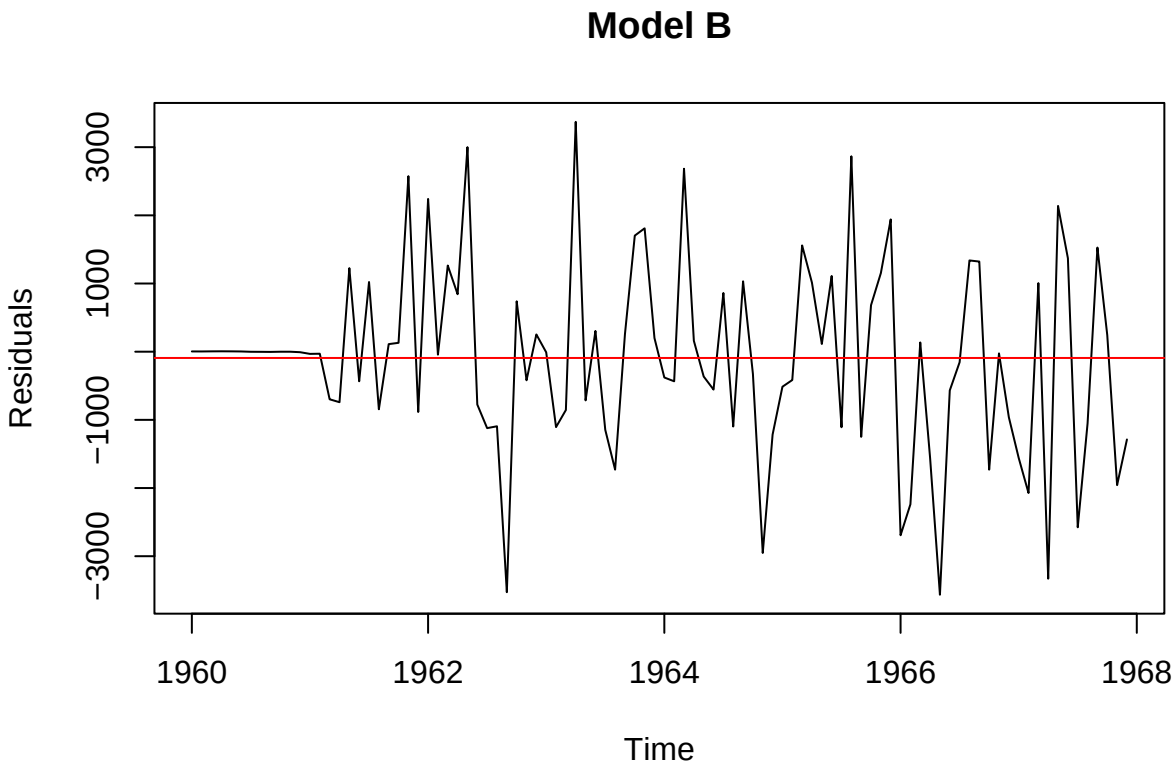

ACF of Residuals Model A
ACF
1.0
0.8
0.6
0.4
0.2
−0.2
0.0
0.5
1.0
1.5
Lag

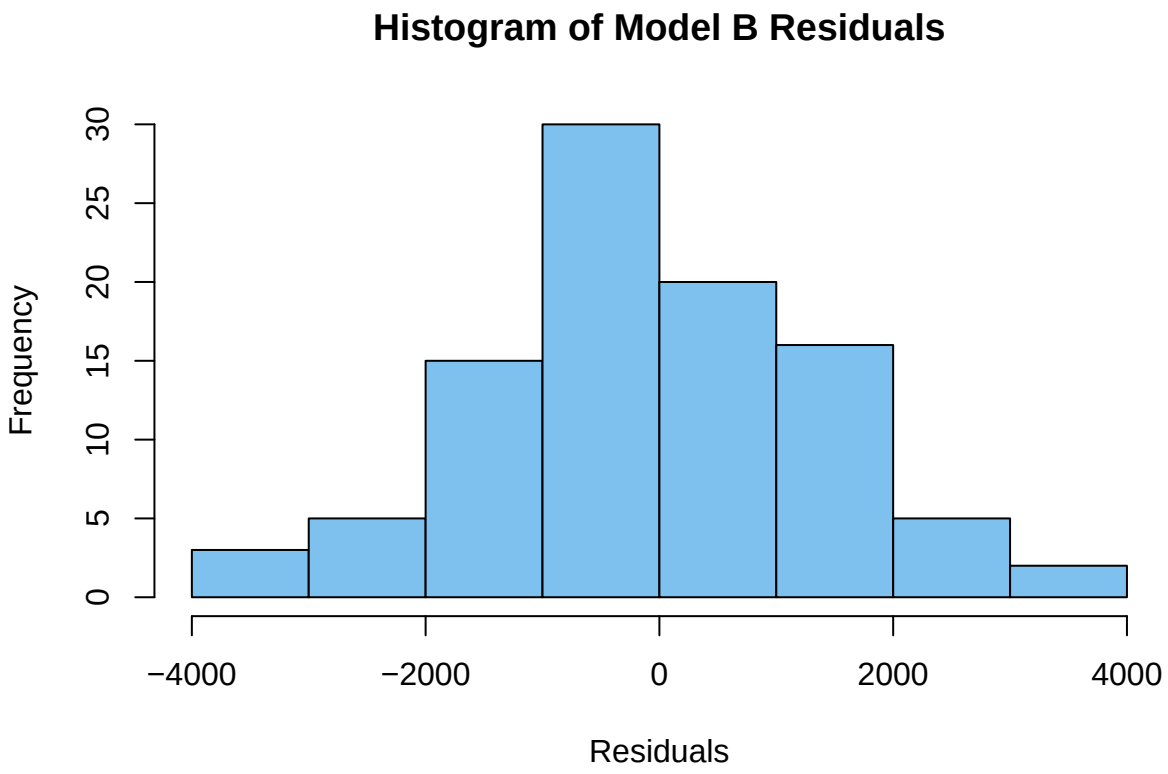


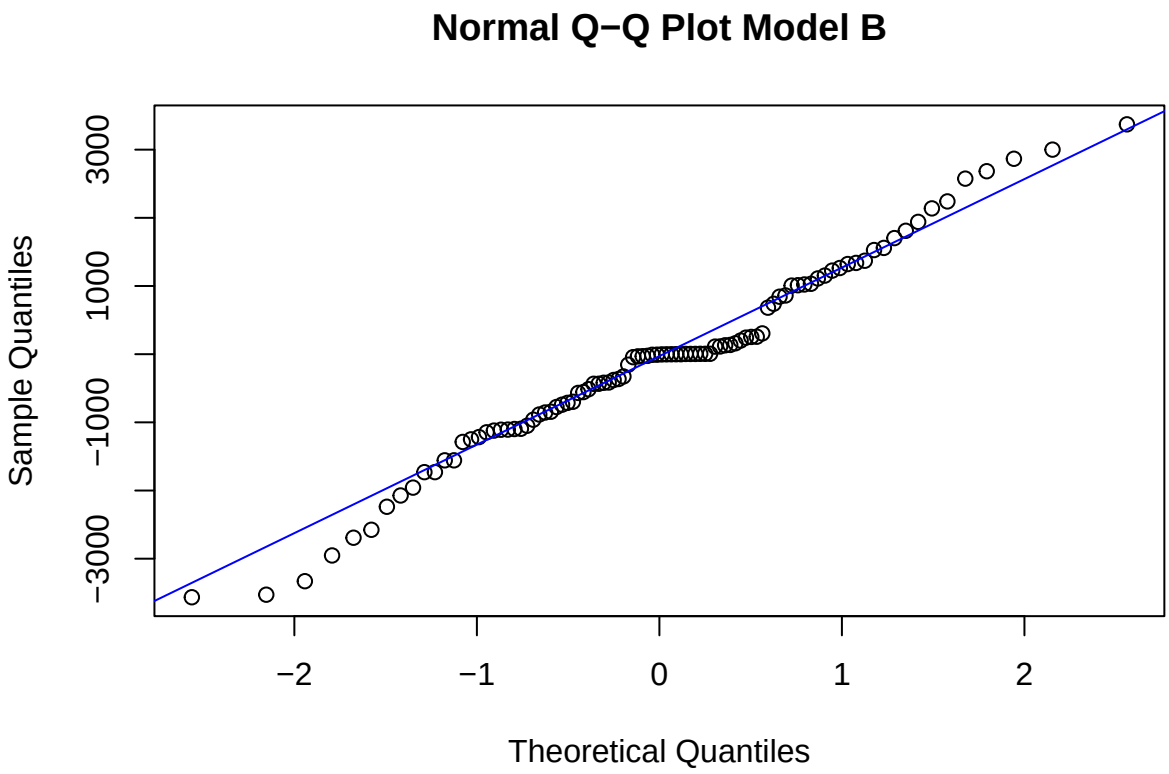
No trend, no visible change of variance, no seasonality, and our sample mean is close to zero. Histogram and Q-Q plot look acceptable. All ACF and PACF residuals are within confidence intervals and may be counted as zero.

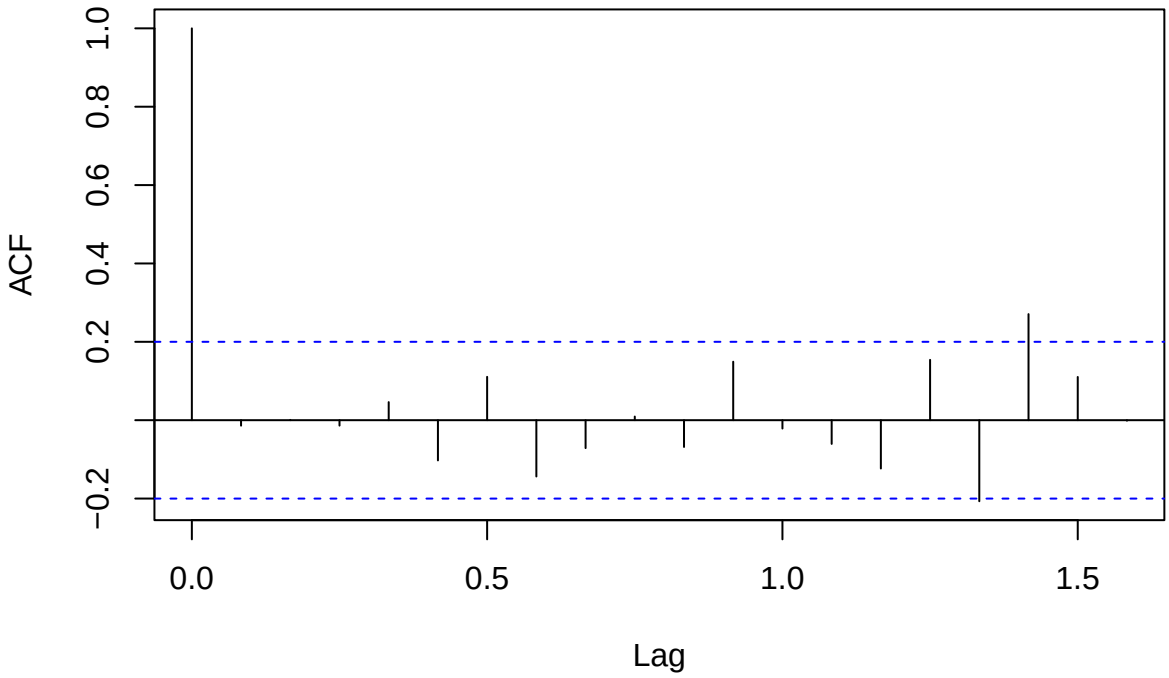
```
##
##  Shapiro-Wilk normality test
##
## data:  fit1$residuals
## W = 0.98613, p-value = 0.4113

##
##  Box-Pierce test
##
## data:  fit1$residuals
## X-squared = 10.66, df = 11, p-value = 0.4722

##
##  Box-Ljung test
##
## data:  fit1$residuals
## X-squared = 11.704, df = 11, p-value = 0.3863

##
##  Box-Ljung test
##
```

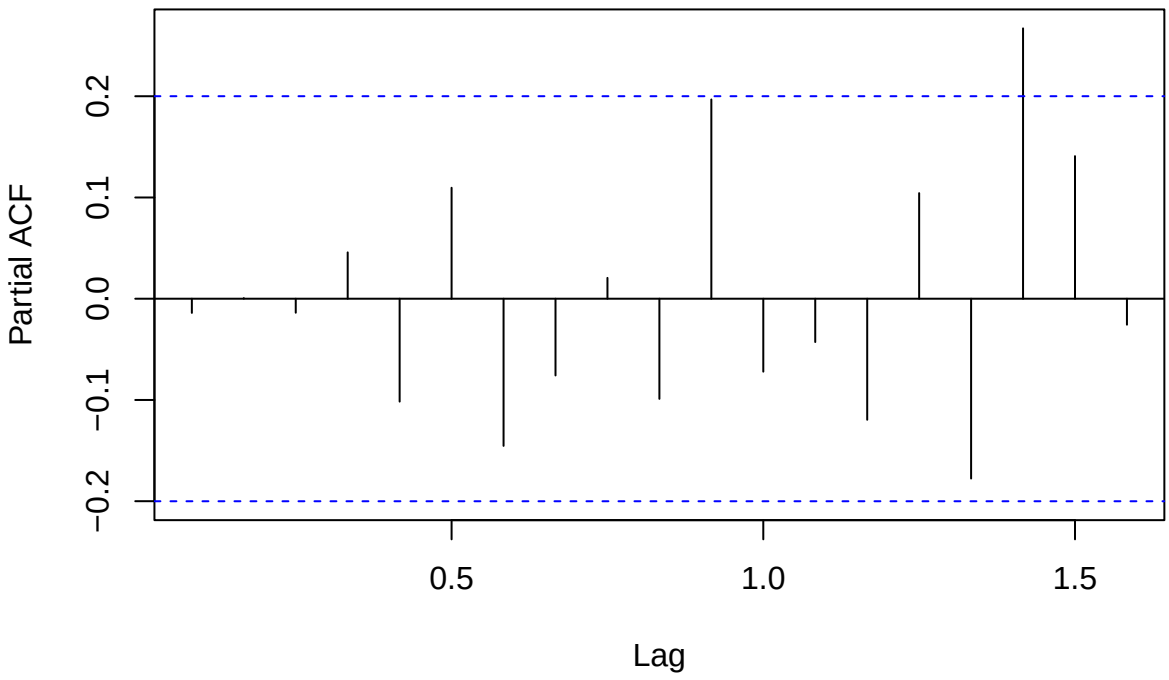
```
## data:  (fit1$residuals)^2
## X-squared = 14.968, df = 12, p-value = 0.2432
```




Histogram of Model B Residuals
Frequency
0
5
10
15
20
25
30
−4000
−2000
0
2000
4000
Residuals


Normal Q−Q Plot Model B
Sample Quantiles
3000
1000
−1000
−3000
−2
−1
0
1
2
Theoretical Quantiles


ACF of Residuals Model B
ACF
1.0
0.8
0.6
0.4
0.2
−0.2
0.0
0.5
1.0
1.5
Lag



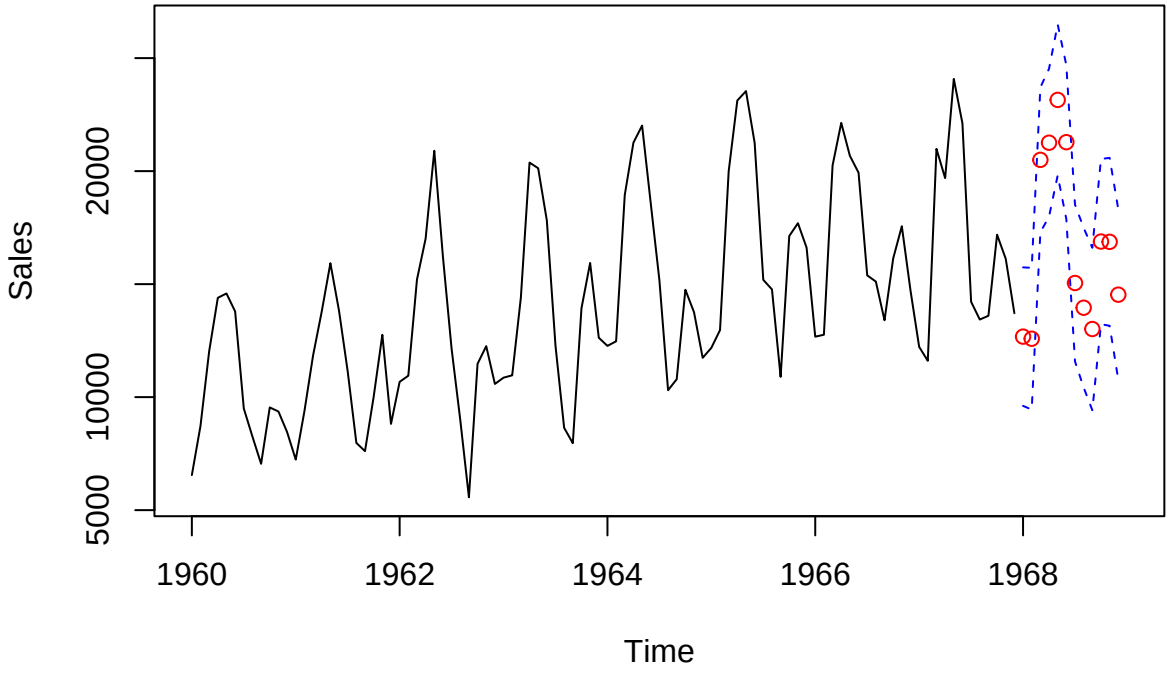
No trend, no visible change of variance, no seasonality, and our sample mean is close to zero. Histogram and Q-Q plot look acceptable. All ACF and PACF residuals are within confidence intervals and may be counted as zero.

```
##
##  Shapiro-Wilk normality test
##
## data:  fit2$residuals
## W = 0.98088, p-value = 0.1748

##
##  Box-Pierce test
##
## data:  fit2$residuals
## X-squared = 7.5394, df = 9, p-value = 0.5811

##
##  Box-Ljung test
##
## data:  fit2$residuals
## X-squared = 8.3906, df = 9, p-value = 0.4953

##
##  Box-Ljung test
##
```

```
## data:  (fit2$residuals)^2
## X-squared = 12.997, df = 12, p-value = 0.3693
```

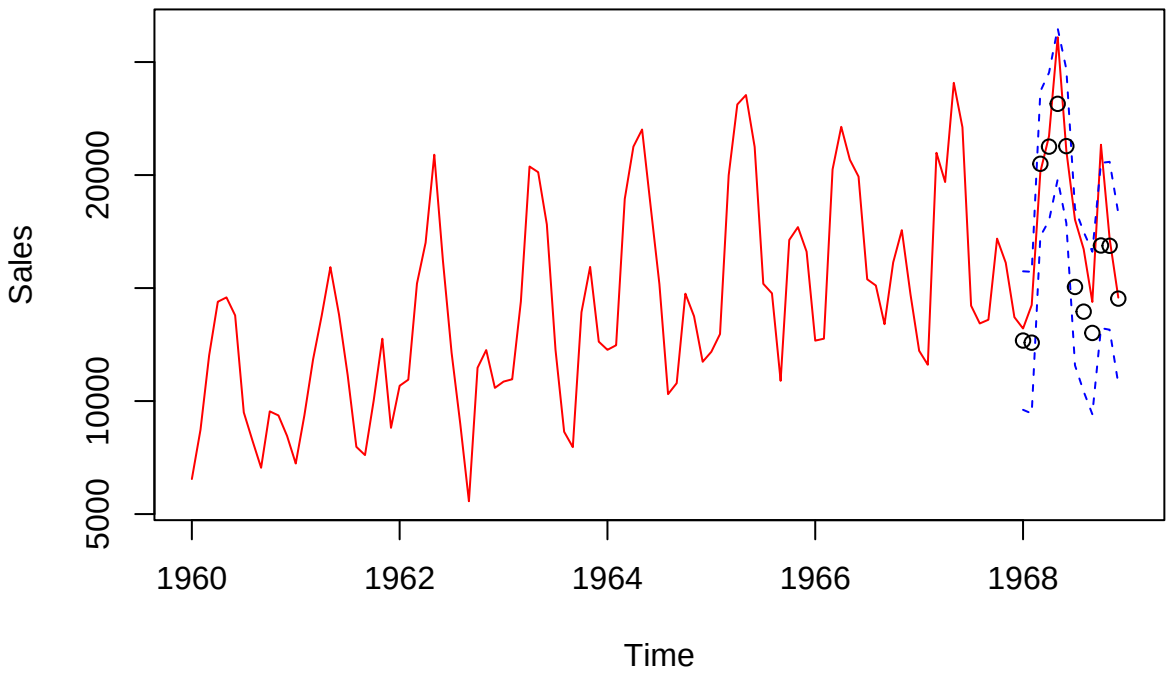
Model A and B both passed all diagnostic tests but Model A did have a lower AICc and uses less parameters so I feel confident that Model A will be the most accurate final model. My final model for my original data U_t follows a $SARIMA(0,1,1)\text{x}(0,1,1)s_{12}$ model.

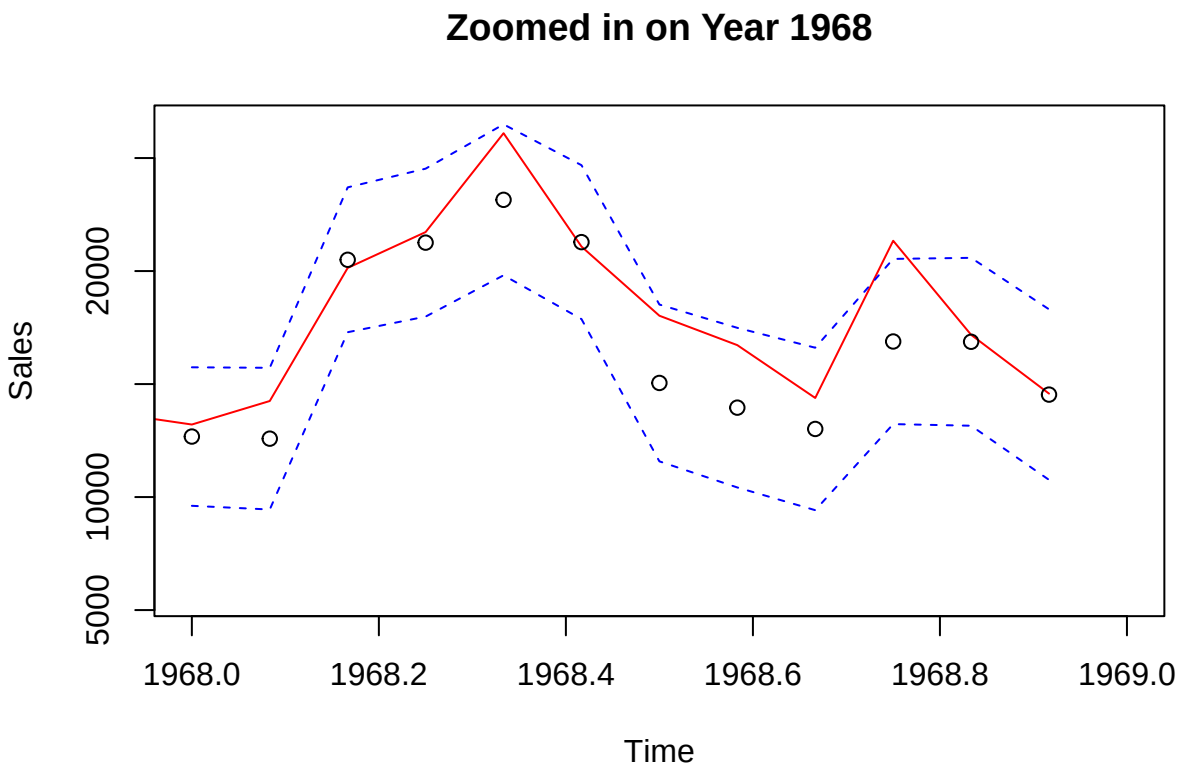
Now we can continue to forecasting the next 12 data points and create the respective intervals based off of our newly created model.



The red points in our plot represent our forecasted data points and the blue dashed lines represent our predictive intervals using Model A.

We will now include the original data and focus in on our test set which is the year 1968.


Sales
5000
10000
20000
1960
1962
1964
1966
1968
Time



We can see in our plots that our test set lies within our predictive interval so we can confidently say that our model is successful in forecasting for this dataset.

Conclusion

The goal of this project was to be able to create a model that could successfuly forecast future observations of our Car Sales data and we have successfully achieved this goal through the various techniques applied and shown above. The math formula used to achieve this goal was $(1-0.7833_{(0.1123)}B)(1-0.5335_{(0.1390)}B^{12})Z_t$.

This was achieved through a variety of techniques and testing learned out of University of California, Santa Barbara from Dr.Feldman and Ph.D. candidate in Statistics Chao Zhang.

References

Tatman, R. (2017, November). US-Car-Sales-1960-1968: Version 1.

Retrieved May 20, 2021 from https://www.kaggle.com/rtatman/us-car-sales.

Appendix

```
library(fpp2)
library(readxl)
library(MASS)
library(forecast)
library(latex2exp)
library(qpcR)
library(ggplot2)
library(tidyverse)
source("Plot.Roots.R")
#U.S. Car sales data 1960 - 1968
mydata = read_xlsx("~/Documents/R/PSTAT 174/Cars.xlsx")
#Plot with trend line and mean of data line
ggplot(mydata,
       aes(x = Month, y = Sales)) +
  geom_line() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(x = "Year", y = "Sales") +
  theme_classic() +
  geom_hline(yintercept=mean(mydata$Sales),
             col = "red", linetype = "dashed", size = 1.5) +
  ggtitle("Monthly Sale Totals of U.S. Cars 1960 - 1968 (in millions)")
#Split data up into training set and test set.
train = mydata[1: 96,]# 8 years
test = mydata[97:108,]# 1 year
#Plot of training set
ggplot(train,
       aes(x = Month, y = Sales)) +
  geom_line() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(x = "Year", y = "Sales") +
  theme_classic() +
  geom_hline(yintercept=mean(train$Sales), col = "red",
             linetype = "dashed", size = 1.5) +
  ggtitle("Monthly Sale Totals of U.S. Cars (Training Set: 96 Observations)")
# Create Time series data with ts() function.
Y = ts(train[,2],start = c(1960,1), frequency = 12)
Y2 = ts(mydata[,2], start = c(1960,1), frequency = 12)
#Histogram to check for skewed data and possible transformations.
#Box-cox Transformation
t <- 1:length(Y)
fit <- lm(Y ~ t)
bcTransform <- boxcox(Y ~ t,plotit = TRUE)
lambda <- bcTransform$x[which(bcTransform$y == max(bcTransform$y))]
Y_bc <- (1/lambda)*(Y^lambda - 1)
#Histogram of 4 transformations
par(mfrow = c(2,2))
hist(Y, col = "skyblue2",
     main = TeX(r'(Histogram of Car Sales Data $U_{t}$)'),
     breaks = 10)
hist(log(Y), col = "skyblue2",
     main = TeX(r'(Histogram of Car Sales Data $log(U_{t})$)'),
     breaks = 10)
hist(sqrt(Y), col = "skyblue2",
```

```
     main = TeX(r'(Histogram of Car Sales Data $\sqrt{U_{t}}$)'),
     breaks = 10)
hist(Y_bc, col = "skyblue2",
     main = TeX(r'(Histogram of Car Sales Data Box-Cox $U_{t}$)'),
     breaks = 10)
#Decomposition plot to see seasonality and Trend.
Y %>%
  decompose() %>%
  autoplot()
var(Y)
#diff at lag 1 for trend
DY = diff(Y, lag = 1)
plot(DY, xlab = "Year", main = TeX(r'($U_{t}$ differenced at lag 1)'))
var(DY)
#diff at lag 12 for seasonality
DDY = diff(DY, lag = 12)
plot(DDY, xlab = "Year",
     main = TeX(r'($U_{t}$ differenced at lag 1 and then 12)'))
var(DDY)
"Seasonality is no longer present and variance has decreased and
our data now looks stationary"
par(mfrow = c(1,2))
hist(Y, col = "skyblue2",
     main = TeX(r'(Histogram of Car Sales Data $U_{t}$)'), breaks = 10)
hist(DDY, col = "skyblue2",
     main = TeX(r'(Histogram of Car Sales Data $\nabla_{1}\nabla_{12}U_{t}$)'),
     breaks = 10)
#ACF and PACF differenced at lag 1 then 12
acf(DDY, lag.max = 60)
"ACF outside of confidence interval may be 1,3, or 4"
pacf(DDY, lag.max = 60)
"PACF outside of confidence intervals may be 1,3,8,11, or 14"
#Possible models based off our ACF/PACF plots
#SARIMA (0,1,1)x(0,1,1)s=12
fit1 = arima(Y, order=c(0,1,1),
             seasonal = list(order = c(0,1,1), period = 12), method="ML")
AICc(fit1)
#SARIMA (0,1,3)x(0,1,1)s=12
fit2 = arima(Y, order=c(0,1,3),
             seasonal = list(order = c(0,1,1), period = 12), method="ML")
AICc(fit2)
plot.roots(NULL,polyroot(c(1, -0.7077, -0.0095, -0.1398)),
           main="(B) roots of ma part, nonseasonal")
#Model A
par(mfrow = c(3,1))
plot(fit1$residuals, ylab = "Residuals", main = "Model A")
abline(h = mean(fit1$residuals), col = "red")
hist(fit1$residuals, col = "skyblue2", xlab = "Residuals",
     main = "Histogram of Model A Residuals")
qqnorm(fit1$residuals,main = "Normal Q-Q Plot Model A")
qqline(fit1$residuals,col="blue")
par(mfrow = c(2,1))
acf(fit1$residuals, main = "ACF of Residuals Model A")
```

```
pacf(fit1$residuals, main = "PACF of Residuals Model A")
#Diagnostic Checking Model A
shapiro.test(fit1$residuals)
Box.test(fit1$residuals, lag =12,
         type = c("Box-Pierce"), fitdf = 1)
Box.test(fit1$residuals, lag = 12,
         type = c("Ljung-Box"), fitdf = 1)
Box.test((fit1$residuals)^2, lag = 12,
         type = c("Ljung-Box"), fitdf = 0)
#Model B
par(mfrow = c(3,1))
plot(fit2$residuals, ylab = "Residuals", main = "Model B")
abline(h = mean(fit2$residuals), col = "red")
hist(fit2$residuals, col = "skyblue2", xlab = "Residuals",
     main = "Histogram of Model B Residuals")
qqnorm(fit2$residuals,main = "Normal Q-Q Plot Model B")
qqline(fit2$residuals,col="blue")
par(mfrow = c(2,1))
acf(fit2$residuals, main = "ACF of Residuals Model B")
pacf(fit2$residuals, main = "PACF of Residuals Model B")
#Diagnostic Checking Model B
shapiro.test(fit2$residuals)
Box.test(fit2$residuals, lag =12,
         type = c("Box-Pierce"), fitdf = 3)
Box.test(fit2$residuals, lag = 12,
         type = c("Ljung-Box"), fitdf = 3)
Box.test((fit2$residuals)^2, lag = 12,
         type = c("Ljung-Box"), fitdf = 0)
#Forecast
pred.tr <- predict(fit1, n.ahead = 12)
U.tr = pred.tr$pred + 2*pred.tr$se #Upper bound of prediction interval
L.tr= pred.tr$pred - 2*pred.tr$se #Lower bound of prediction interval
ts.plot(Y, xlim = c(1960,1969), ylim = c(min(Y),
                                         max(U.tr)), ylab = "Sales")
lines(U.tr, col="blue", lty="dashed")
lines(L.tr, col="blue", lty="dashed")
points(pred.tr$pred, col="red")
#With training set
ts.plot(Y2, xlim = c(1968,1969), ylim = c(min(Y2),
                                          max(U.tr)), col = "red", ylab = "Sales")
lines(U.tr, col="blue", lty="dashed")
lines(L.tr, col="blue", lty="dashed")
points(pred.tr$pred, col="black")
```