US Home Sales 1965 - 1975; A Monthly Timeseries and Forecast

Riley Zamora

2022-11-21

Abstract

In this report, we will be analyzing a data set from the `Time Series Data Library (tsdl)`, provided by `Abraham & Ledolter (1983)`. The data set is `Monthly sales of U.S. houses (thousands) 1965 - 1975`, which is under the `Sales` subject in tsdl and contains `132 observations` of monthly home sales in thousands. The goal of analyzing this data set is to be able to confidently forecast future sales of homes. This is important because homes are always going to be necessary to the human population and being able to forecast the sales will allow a person to know when more homes are available. From an economic view the future supply and demand of homes will be known after forecasting allowing us to predict when homes will be more or less expensive. I will use time-series techniques to build a model and forecast up to 12 months ahead using that model.

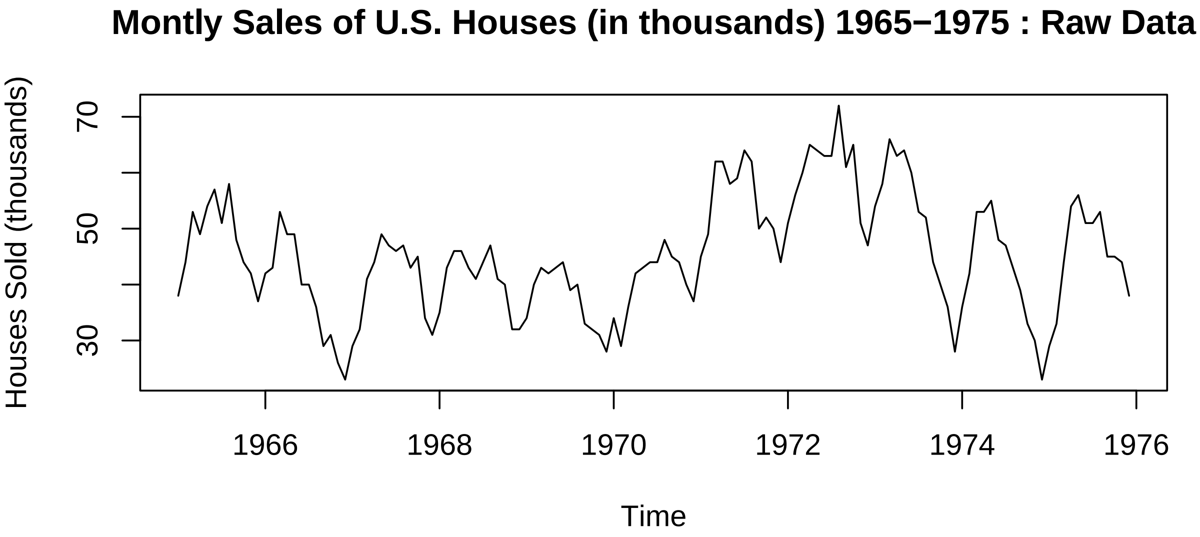
Introduction:

In this report, I addressed how many homes will be purchased in the next 12 months. To do this I used a variety of time series techniques including analyzing ACF and PACFs, checking box-cox transform, differencing, checking AICc, diagnostic checking and forecasting. The main question at hand is how many homes will be sold in upcoming months. I will first plot the raw data, in order to see any sudden changes or a linear trend in the data. Then I will split the data set into a training set and a test set to verify if the forecast is valid. I will check if any transformations or differencing is needed to make a stationary model. Next I will find preliminary p's and q's to fit a few models. The models with the lowest AICc I will find the algebraic equation and run diagnostic checking on them. Finally, I will forecast and make sure the raw data fit inside the prediction interval.

Testing this forecast to the original data, I found that a chosen model I built and tested gave me an accurate prediction. One issue I ran into throughout the project was that my data was not normal. This is likely from my chosen data set. Not all data will be completely normal. I conclude that I was successful in predicting future US home Sales.

Sections:

- Plot and Analyze the time series:



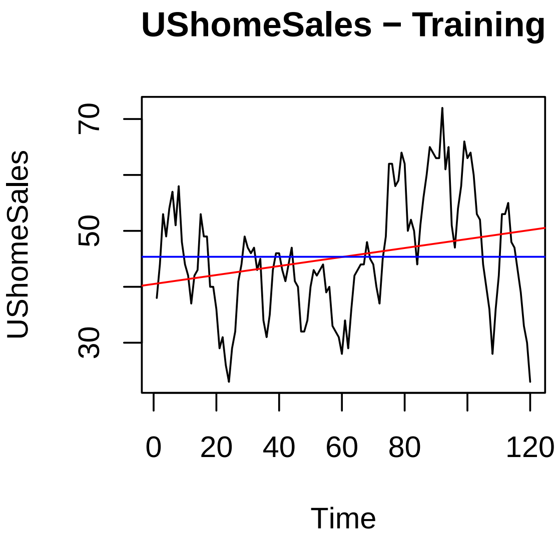
i. Looking at the plot of the raw data I can see that there are years when people bought more houses and years when people bought less houses. I plan to difference to deal with any trend that was observed in the time series.

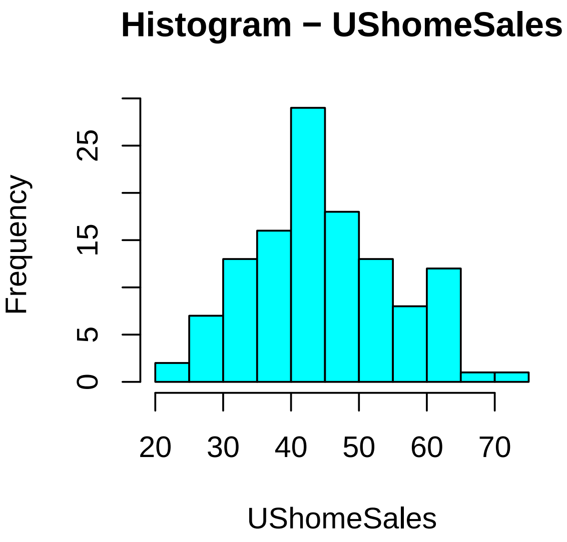
ii. There is some seasonality to be noted in the plot. The seasonality is probably due to changes in house prices, this could be due to cost of materials if the home was built recently or not. Economic changes would cause seasonality in house sales. To adjust for seasonality, I plan to run a preliminary check on the ACF to find which lags to difference at.

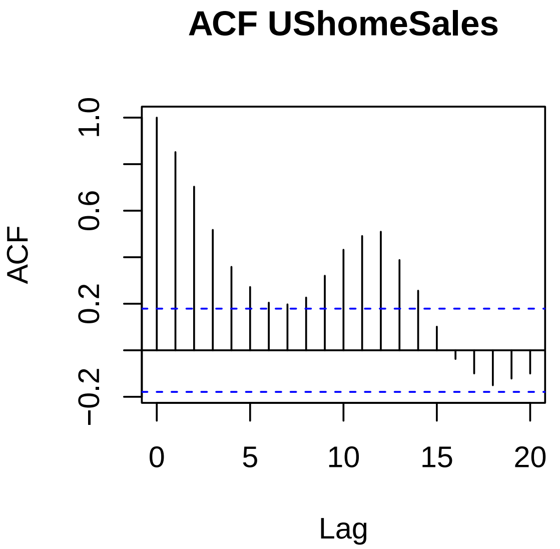
iii. The plot is filled with jagged spikes, but I would say it is notable that there is some upward trend in 1971. Even with seasonality the least amount of houses sold increases from 1971 - 1973 then drops much lower in 1974. This could be due to a housing bubble or some other economic effect.

- Necessary Differencing an Transform:

Now I split my data to allow for testing. The original data set is 132 observations and I will spit at 120 to allow 12 observations to be used for testing.

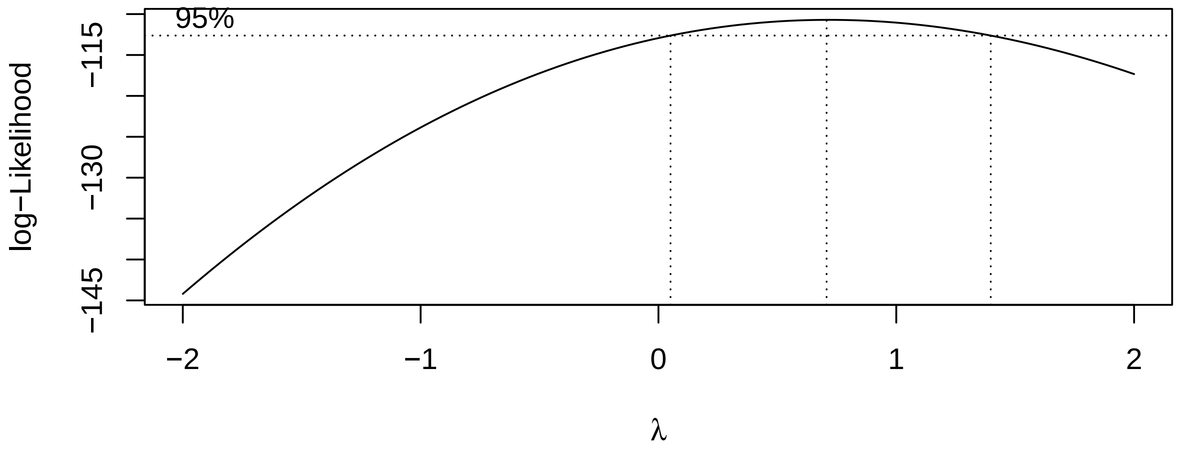






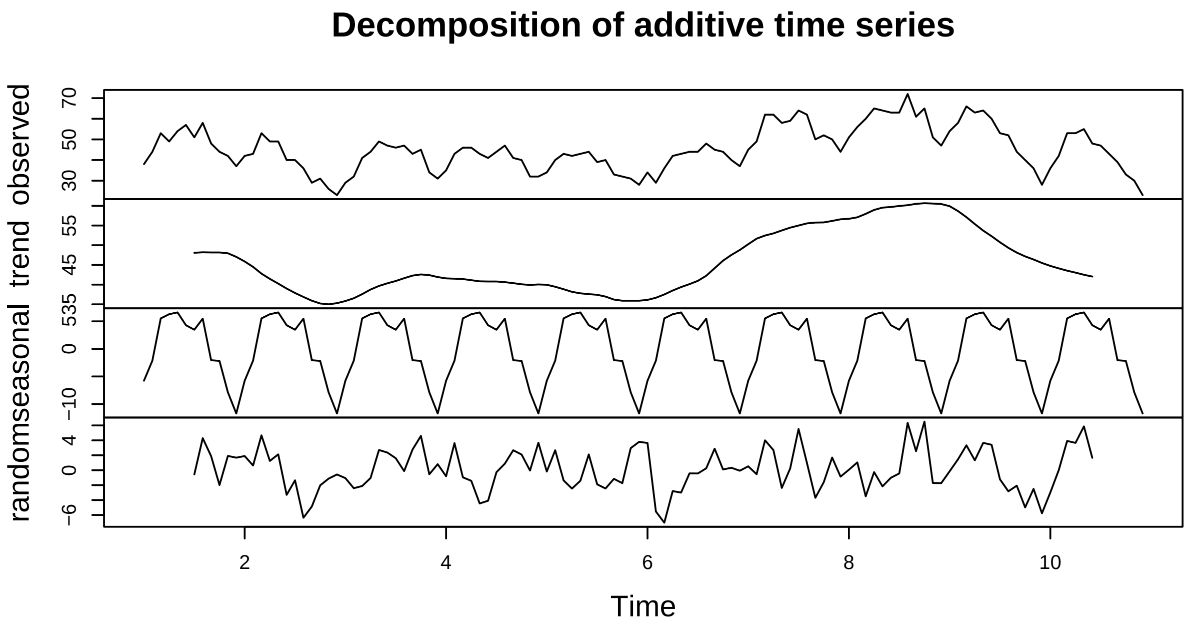
I can see that the training data set has trend (red line) so we must difference until it matches up with the mean (blue line). The histogram looks pretty standard but has a jump after 60. Looking at the acf we can see that the lags are large and periodic. I conclude non-stationary and proceed with a checking if transformation is needed before differencing. Our variance at this stage is 111.898, with a mean of 45.36667.

Transformation



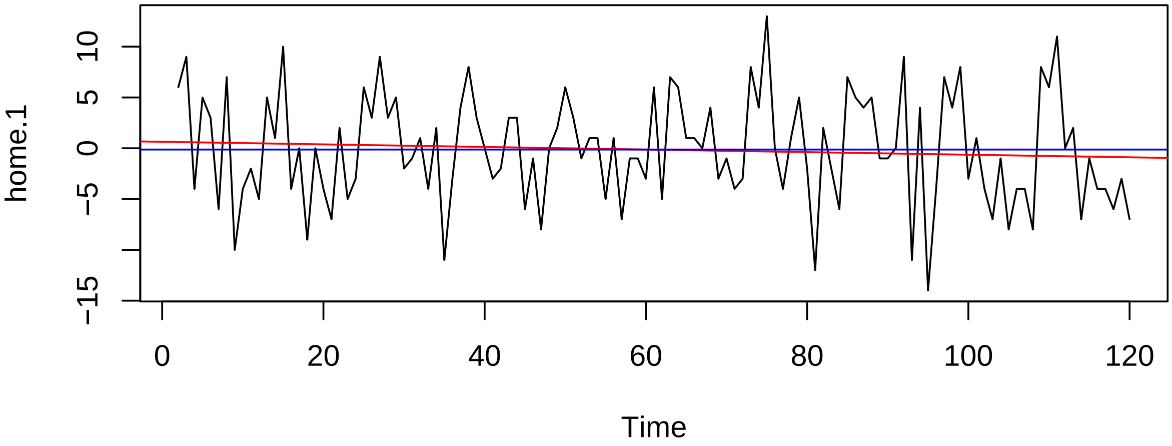
Since **0** does not fall in the confidence interval, I will not proceed with a transformation. The value of $\lambda = 0.7070707$, which is pretty high to assume a transformation is needed. Now, I will look at the decomposition of the training set.

Decomposition of Training Set

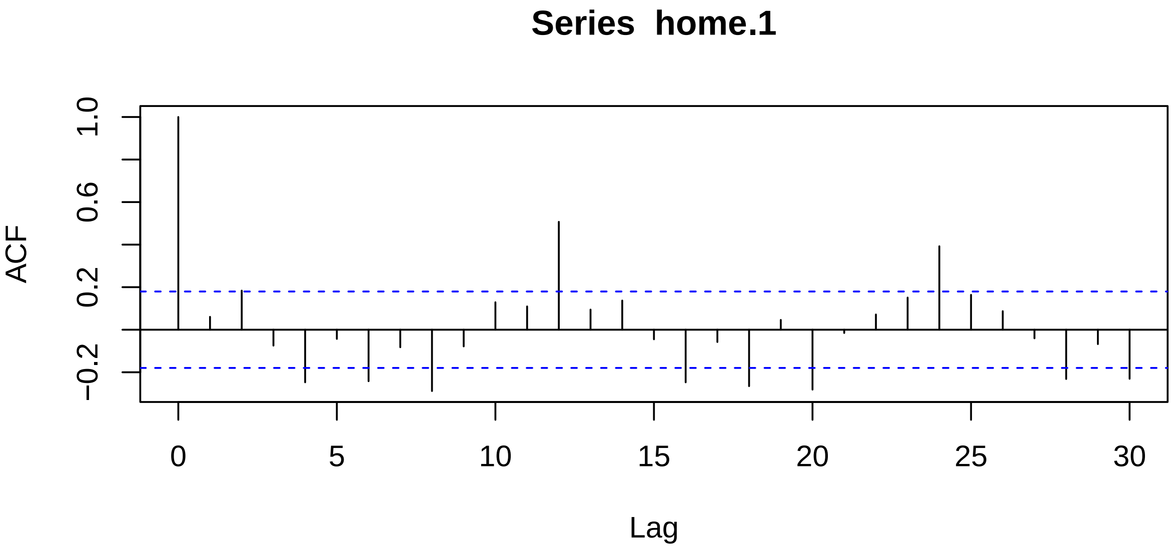


From the plot, I can see that there is a trend factor that is not linear. We will still need to difference our training set, `UShomeSales`, at lag 1 to deal with trend, but our data might not be normal. There is also a seasonal component so differencing at lag 12 might prove to be useful.

Differencing UShomeSales at Lag 1

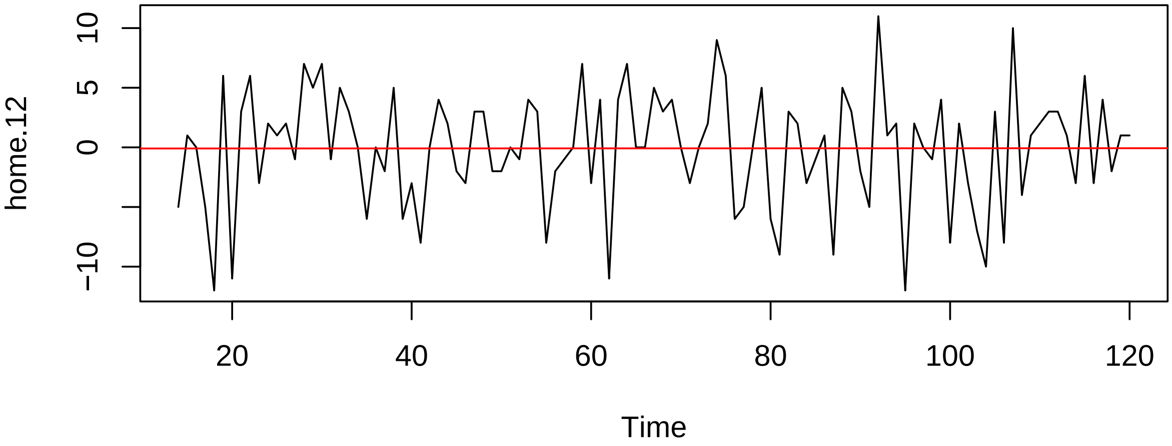


I have not reached a stationary state yet so I will run a preliminary check on the acf to see if I can gather anymore information. The variance is now 28.73821 with a mean of -0.1260504 which is lower than before differencing so we are moving in the right direction.



Using a lag.max of 30, I can see that there is still some seasonal component. At lags 12 and 24 we have significant data that decays so I can assume it would be helpful to difference at lag 12 as well as lag 1.

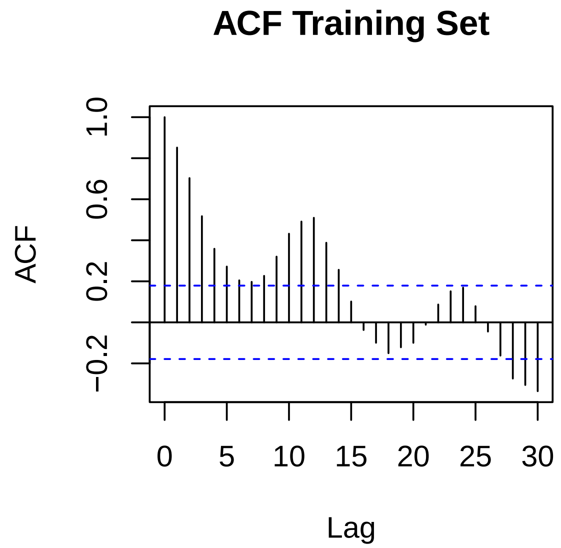
Differencing home.log at Lag 1 then 12

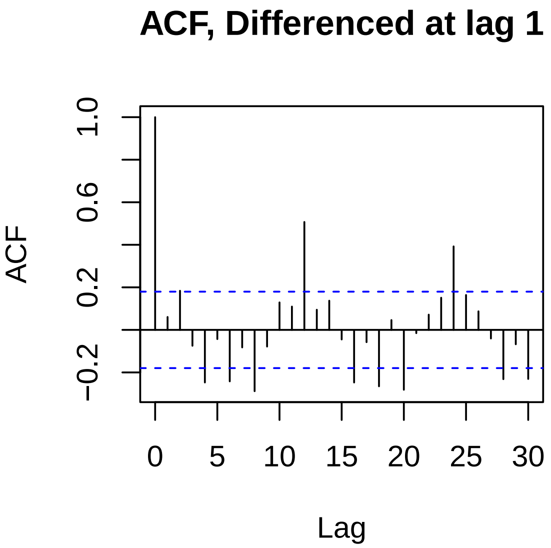


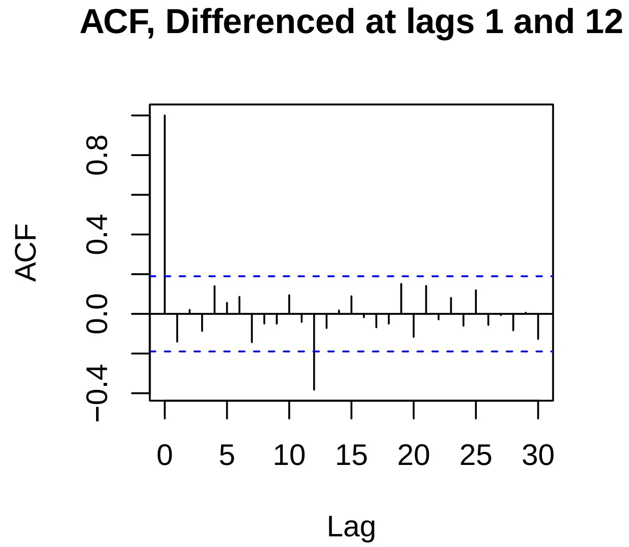
Since the trend (red line) is now level with the mean (blue line obscured) I conclude that the mean and trend are stationary and I can move on to looking at the ACF and PACF. Just to double check I can look at the acf in each phase of differencing. The mean is -0.08411215 and the variance is lower again with a value of 23.75701.

All mean and var

Model	VAR(model)	Mean(model)
UShomeSales	111.898	45.36667
home.1	28.73821	-0.1260504
home.12	23.75701	-0.08411215

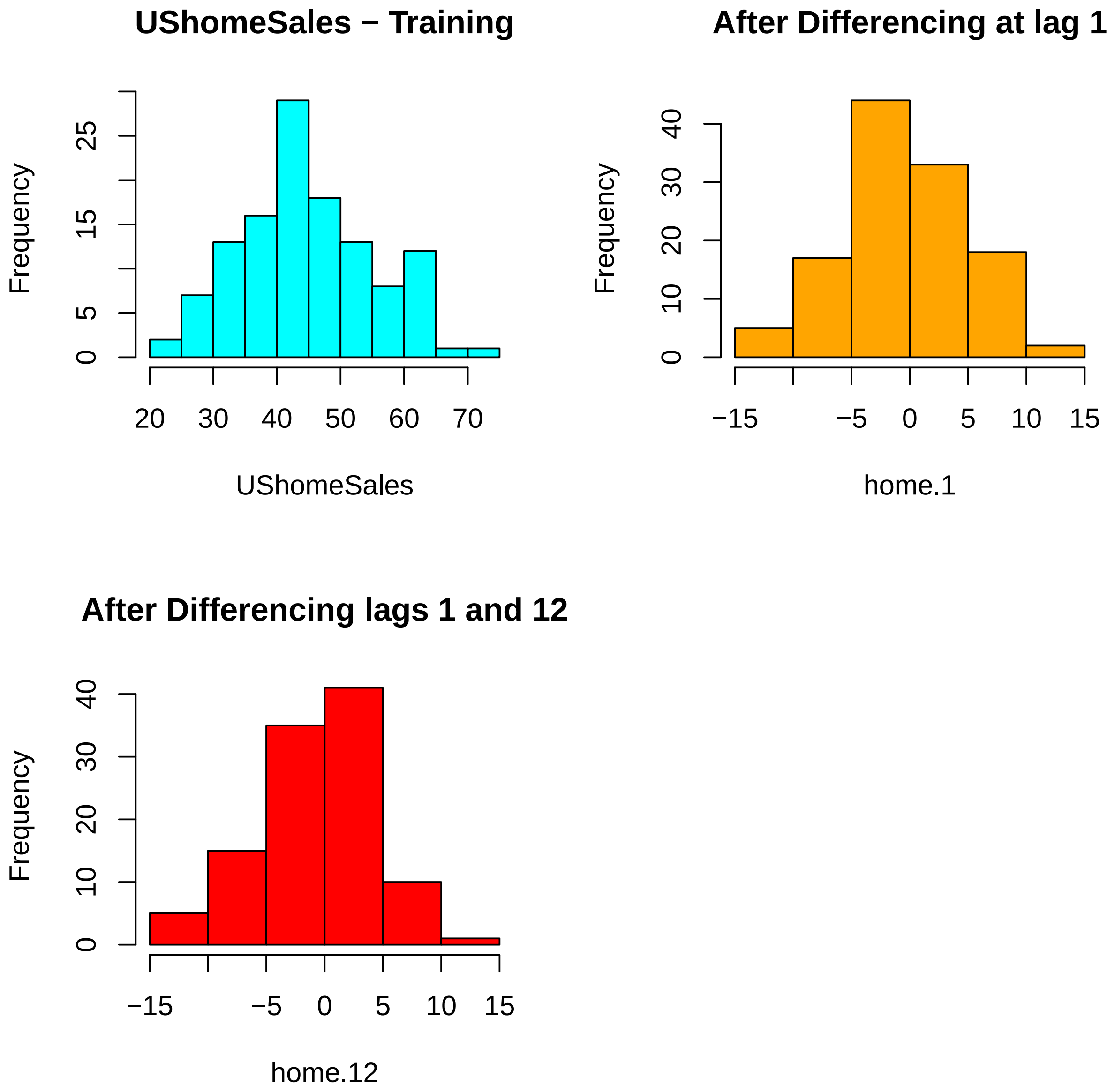






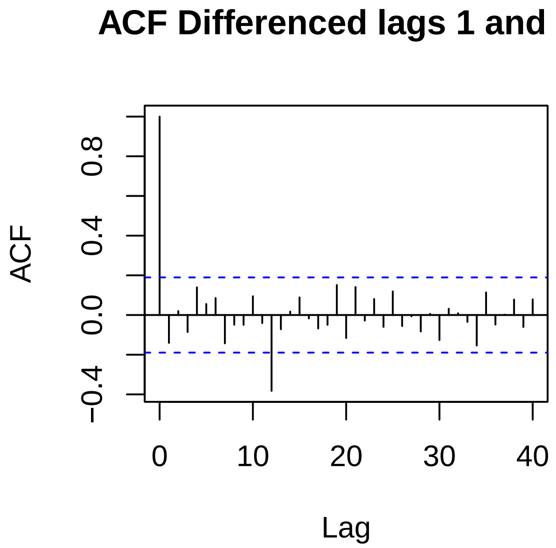
In our first phase, there is significant seasonality and trend so we difference at lag 1. In the second phase we still see some seasonality so another difference at lag 12 is useful to achieve a stationary model. Finally, after differencing at lags 1 and 12, we have a stationary model and we can move forward to choosing p's and q's.

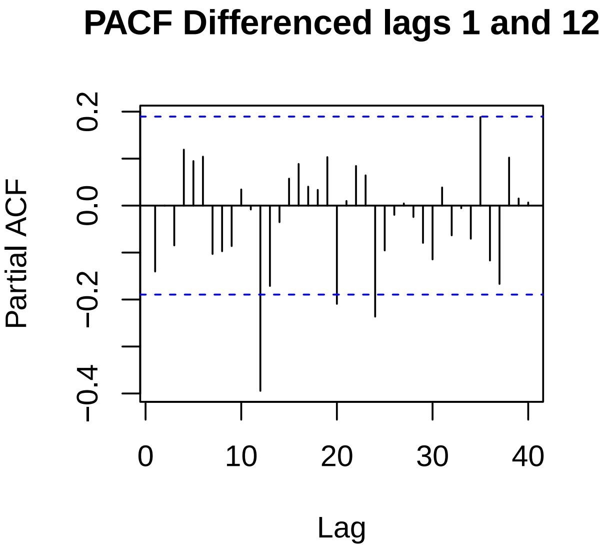
Comparing Histograms



Comparing the UShomesales to home.1, I can see that the range of values decreased and that the histogram follows a more normal curve. After differencing again at lag 12, home.12, I can see that the histogram from differencing at lag 1, home.1, seems to have flipped. I am happy with the final histogram and move on to looking at the ACF and PACF.

- Plot ACF and PACF:





Analyze ACF and PACF:

In the ACF there is significance at lags: 0 and 12. In the PACF there is significance at lags: 12, 20, 24, and maybe 35. From these significant lags, I would say that preliminary p,q,P,Q would be:
- p: 0 or 1
- d: 1 because I differenced once at lag 1
- q: 0 or 1
- P: 2
- D: 1 because I differenced once at lag 12
- Q: 0

Model Fitting

I fit the following models using the arima() function.

$SARIMA(0,1,0)(2,1,0)_{12}$	sar1	sar2
$\rho(h)$	-0.5897	-0.3646
s.e.	0.0987	0.1022

AICc(model1): 617.8837

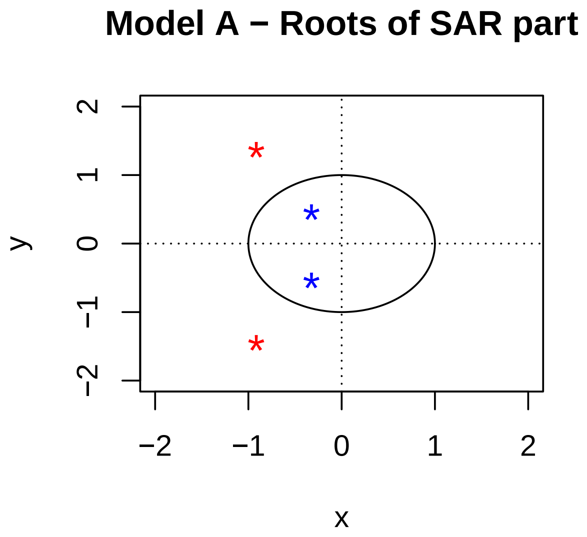
$SARIMA(1,1,0)(2,1,0)_{12}$	ar1	sar1	sar2
$\rho(h)$	-0.2544	-0.6484	-0.3540
s.e.	0.0963	0.0987	0.1008

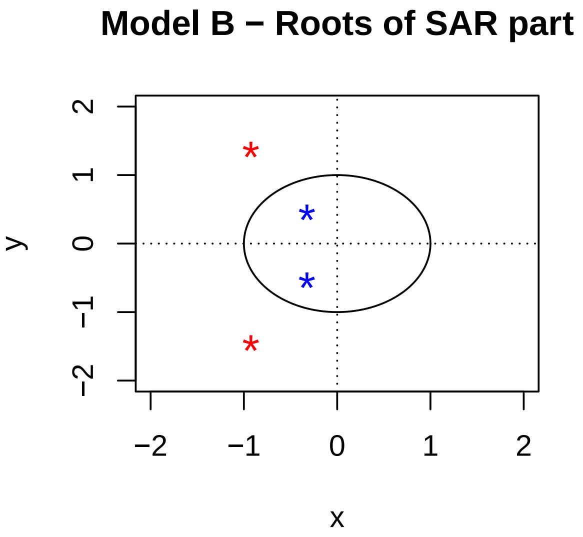
AICc(model2): 613.3684

$SARIMA(1,1,1)(2,1,0)_{12}$	ar1	ma1	sar1	sar2
$\rho(h)$	-0.5128	0.2821	-0.6492	-0.3507
s.e.	0.4946	0.5643	0.0987	0.1014

AICc(model3): 615.3009

The lowest AICc is model2 and the second lowest is model3. I will now run diagnostic testing on models 2 and 3, and call these models Model A and Model B, respectively from now on.

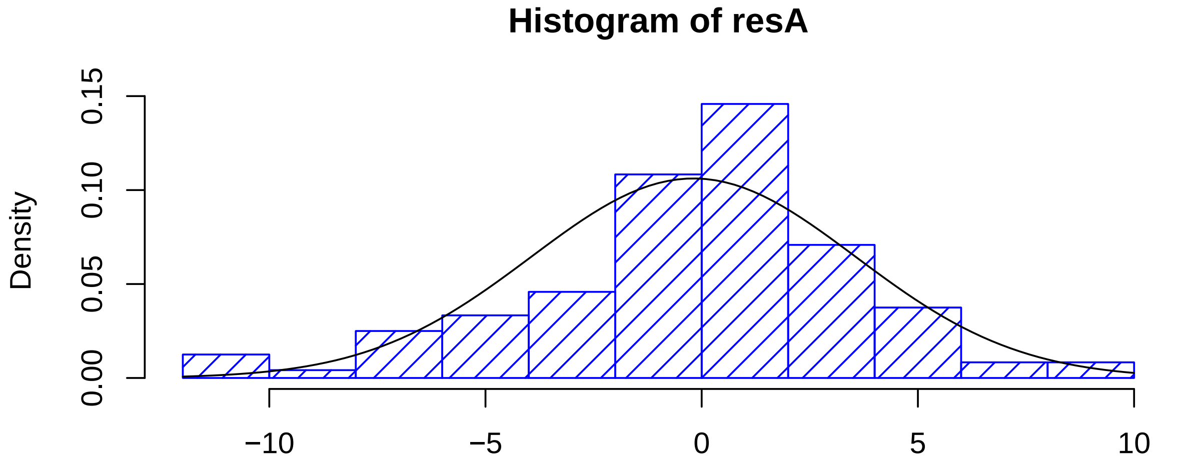

Model A − Roots of SAR part


Model B − Roots of SAR part

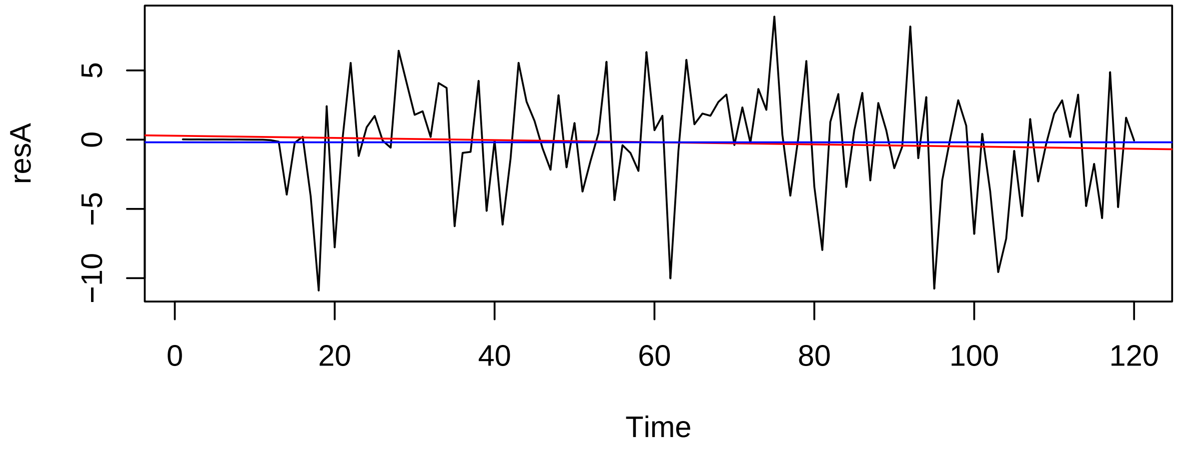
Model A and B both have roots outside the unit circle meaning they are casual and invertible. The inverse roots are inside the unit circle and no roots touch the unit circle. We proceed to diagnostic checking for these models.

Diagnostic Checking

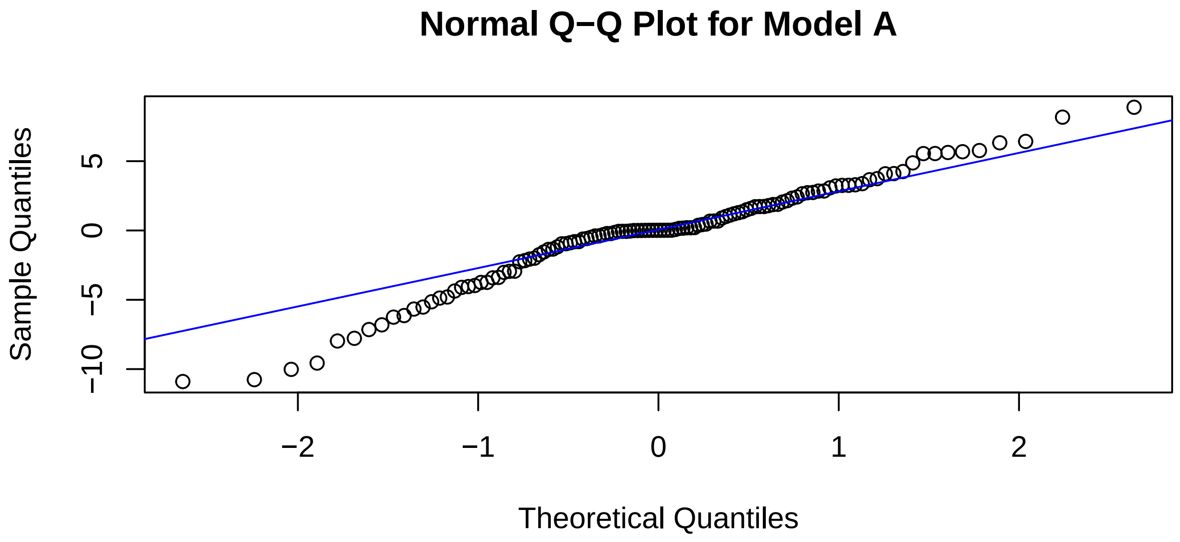
Model A:



The histogram of residuals for model A fits a normal curve.

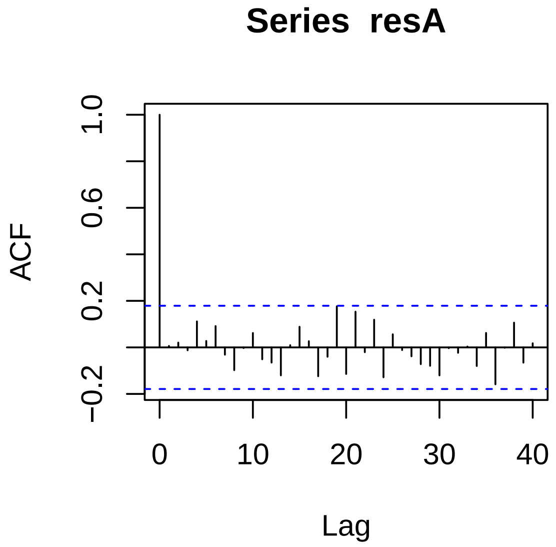


The plot of residuals with a regression line is near horizontal.

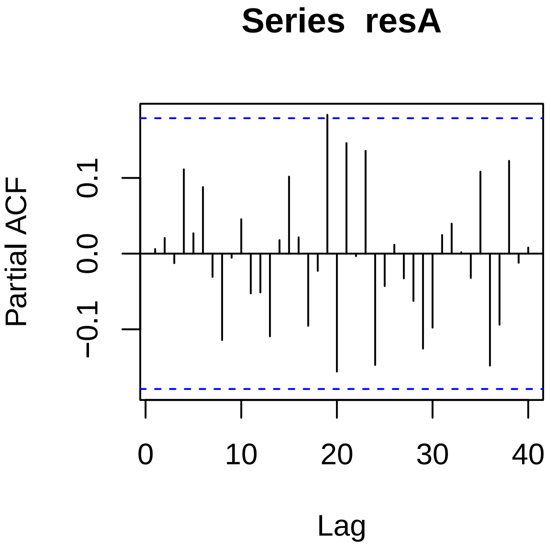


Normal Q−Q Plot for Model A

Our data does not seem to be normal here but we can proceed as we differenced at lag 1 when the trend was not linear. Reference decomposition section.



Series resA



Series resA

The ACF and PACF are within the confidence interval. Lag 19 in the PACF touches the CI, but this is nothing to worry about in our model.

Shapiro-Wilk normality test	resA
w	0.96479
p-value	0.003129

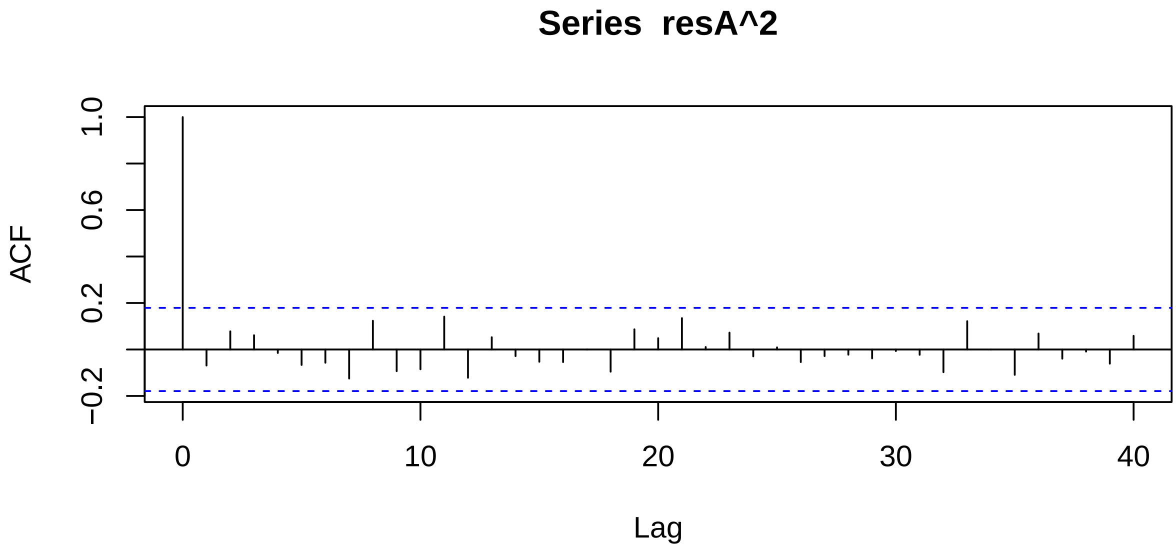
Does not pass Shapiro-Wilk test because p-value is below 0.05. Thus, the data is not normal.

Box-Pierce test	resA
X-squared	4.7098
p-value	0.7881
df	8

Box-Ljung test	resA
X-squared	5.0656
p-value	0.7505
df	8

Box-Ljung test	resAˆ2
X-squared	11.661
p-value	0.3896
df	11

All p-values above 0.05, these tests are a pass.



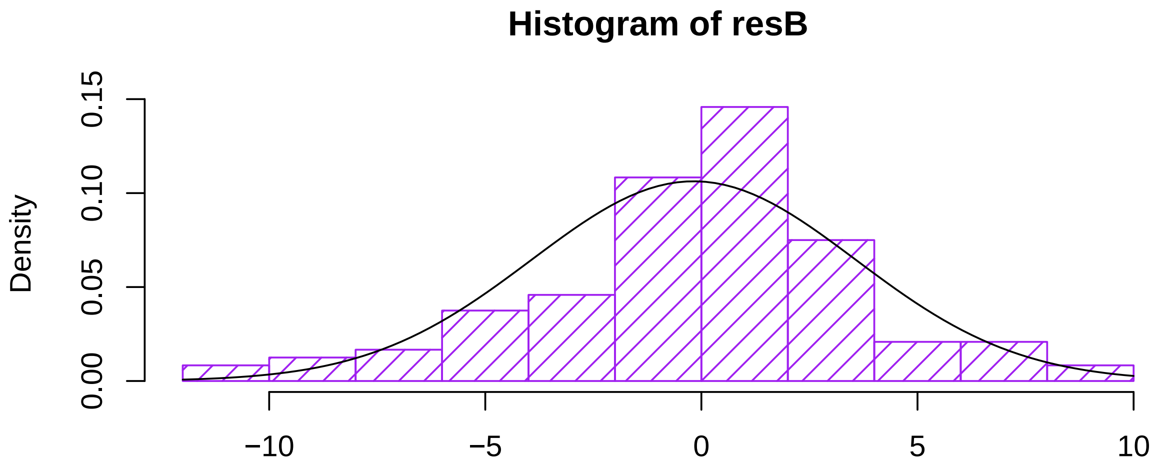
The ACF of residuals squared is all with in the confidence interval.

```
ar(resA, aic = TRUE, order.max = NULL, method = c("yule-walker"))
```

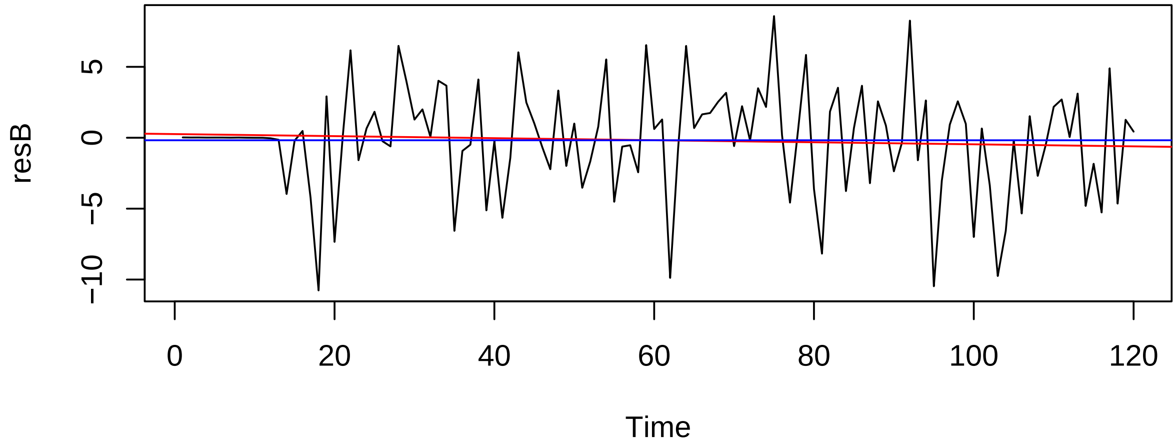
```
##
## Call:
## ar(x = resA, aic = TRUE, order.max = NULL, method = c("yule-walker"))
##
##
## Order selected 0  sigma^2 estimated as  14.12
```

Order selected is 0, the model is white noise.

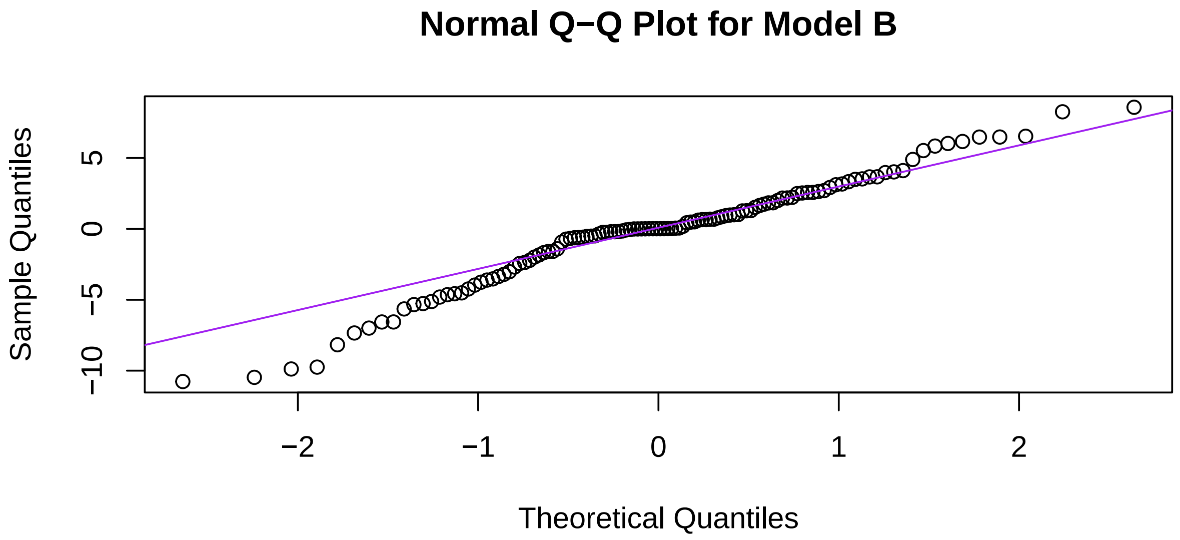
Model B:



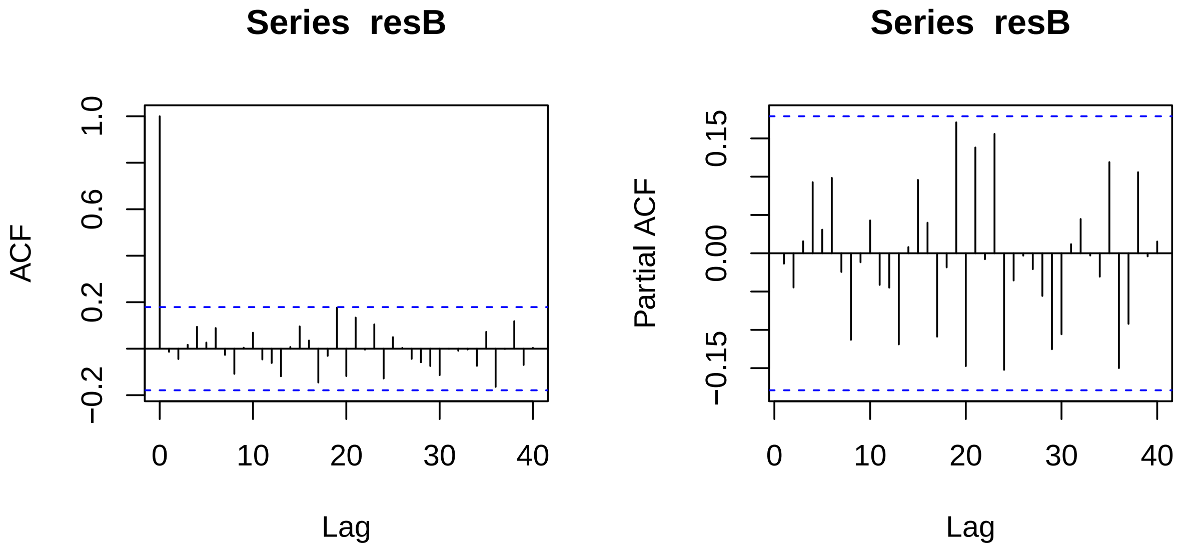
The histogram of model B residuals has a normal curve.



The plot of the residulas has a regression line that is nearly horizontal and this looks okay.



Our data in this model is not normal, thus the Normal Q-Q Plot will not be fitted correctly.



The ACF and PACF seem to check out. Lag 19 in the ACF is close to the line but this is irrelevant.

Shapiro-Wilk normality test	resB
w	0.96479
p-value	0.003129

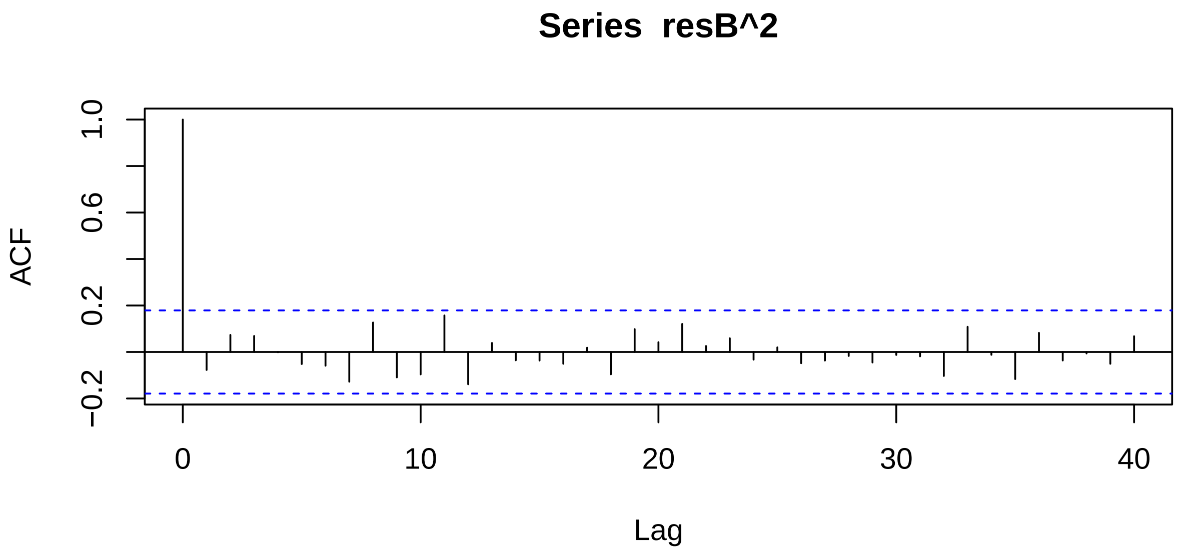
P-value below 0.05. Fails the Shapiro-Wilk test because our model is not normal.

Box-Pierce test	resB
X-squared	4.7039
p-value	0.696
df	7

Box-Ljung test	resA
X-squared	5.0656
p-value	0.652
df	7

Box-Ljung test	resA^2
X-squared	11.661
p-value	0.2833
df	11

All p-values above 0.05, these tests are a pass.



The ACF of residuals squared looks good.

```
ar(resB, aic = TRUE, order.max = NULL, method = c("yule-walker"))
```

```
##
## Call:
## ar(x = resB, aic = TRUE, order.max = NULL, method = c("yule-walker"))
##
##
## Order selected 0  sigma^2 estimated as  14.09
```

Order selected is 0, the model is white noise.

Recap of Diagnostic Checking

Tests	Model A	Model B
Histogram	PASS	PASS
Plot	PASS	PASS
Q-Q Norm	FAIL	FAIL
ACF/PACF	PASS	PASS
Shapiro-Wilk	FAIL	FAIL
Box-Pierce	PASS	PASS
Box-Ljung	PASS	PASS
Box-Ljung(res^2)	PASS	PASS
ACF (res^2)	PASS	PASS
AR	PASS	PASS

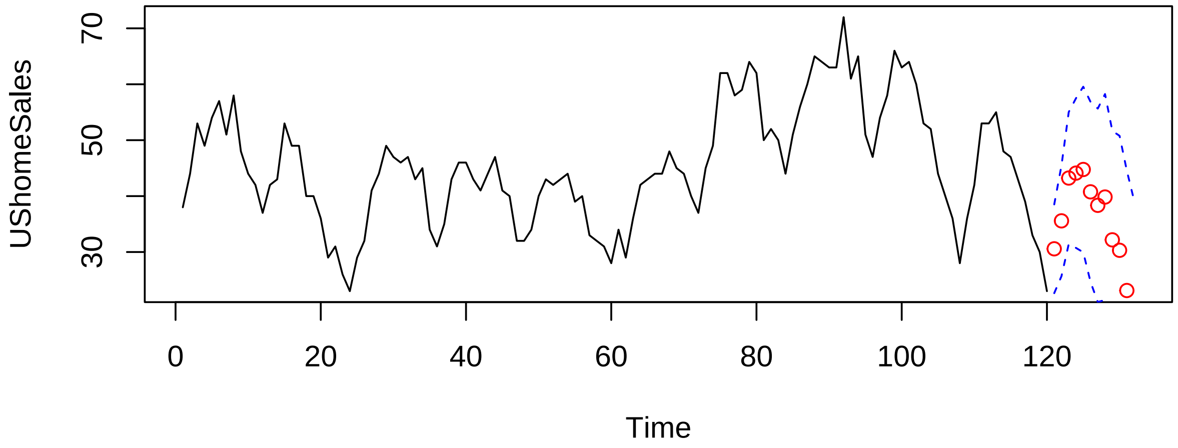
Model A passed each test with a greater value than Model B. The normality tests were closer to passing for Model A than Model B. Since Model A proved better than model B and the AICc was lower, we proceed to forecast with `Model A`.
Model A in algebraic form:

$$SARIMA(1,1,0)(2,1,0)_{12}$$
$$(1-\phi_1 B)(1-B)(1-\Phi_1 B^{12}-\Phi_2 B^{24})(1-B^{12})X_t = Z_t$$
$$(1-(-0.25)B)(1-B)(1-(-0.65)B^{12}-(-0.35)B^{24})(1-B^{12})X_t = Z_t$$
$$(1+0.25B)(1-B)(1+0.65B^{12}+0.35B^{24})(1-B^{12})X_t = Z_t$$

Forecasting Model A

First we will print the forecasts with prediction bounds to gather some basic information about the forecast.

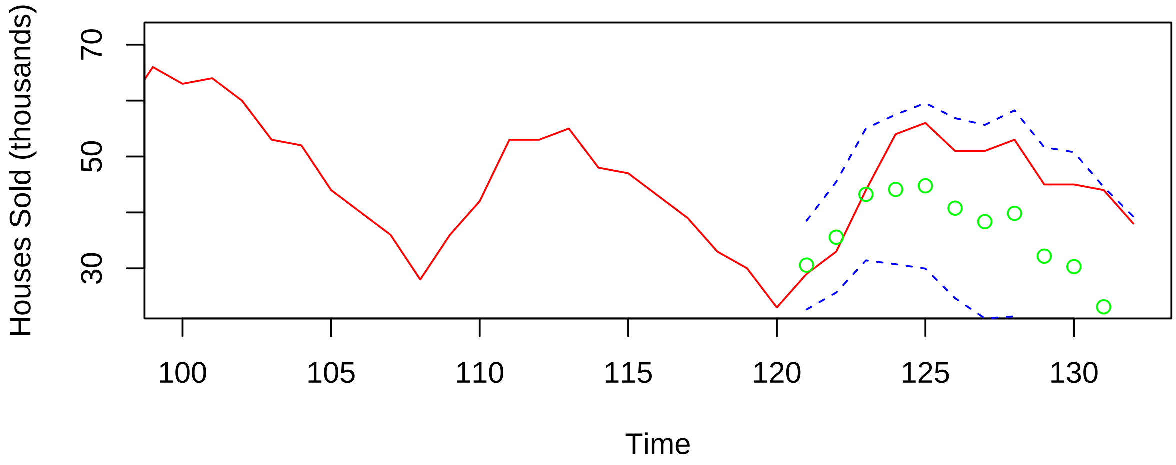
Producing a graph with 12 forecasts



The 12 forecasts fit within the confidence interval for the training set. Now, I can check to see how accurate my model building is by testing with the true data set names sales. I can predict how many homes will be bought in the next 12 months within a 95% confidence interval.

Plot forecasts and Sales(raw) data

Essentially we are testing with `new` data even though the data was gathered a while back.



Here the true values and the predicted values fit within the same confidence interval so I have successfully found a model to forecast the amount of homes bought in thousands for future months. The discrepancy here is that as the prediction time gets longer the model becomes less successful because we are predicting based on other predictions. Essentially, our confidence interval will get wider and wider eventually deeming our prediction useless as we try to predict further ahead.

Conclusion

In forecasting `Monthly sales of U.S. houses (thousands) 1965 - 1975`, provided by `Abraham, B. and Ledolter, J. (1983)`, I found that a $SARIMA(1,1,0)(2,1,0)_{12}$ model, Model A, was the best fit to predict home sales for the next 12 months. I ran into some normality errors that are most likely due to not having linear trend and differencing at lag 1. However, differencing at lag 1 proved to be more helpful than having normality in my data. I found a second model, Model B, that I ran diagnostic checking on and passed all the same tests but the AICc was lower, and the diagnostic tests did not pass as well as Model A. Being able to forecast US home sales will always be a helpful technique because as long as humans live in the US we will need a place to live. Being able to predict the market will allow for home buyers to know when to buy or sell. I would like to acknowledge Professor Raisa Feldmen and the TAs for giving me advice in the development of my project.

References

Statistical Methods for Forecasting, by Abraham, B. and Ledolter, J. (1983).

Appendix

Load Libraries

```
library(tsdl)
library(MuMIn)
library(ggplot2)
library(ggfortify)
library(MASS)
library(forecast)
```

Read Data

```
length(tsdl[[6]])
attr(tsdl[[6]], "subject")
attr(tsdl[[6]], "source")
attr(tsdl[[6]], "description")
```

Plot Raw Data

```
sales <- (tsdl[[6]])[1:132]
plot.ts(sales, main = 'Montly Sales of U.S. Houses (in thousands) 1965-1975 : Raw Data',
        ylab='Houses Sold (thousands)')
```

Split into Training Set

```
UShomeSales <- (sales)[1:120]        # training set
UShomeSales <- ts(UShomeSales)           # Make the vector a time series
UShomeSales.test <- (sales)[121:132] # testing set
UShomeSales.test <- ts(UShomeSales.test)
```

Plot Training Set

```
par(mfrow=c(1,2)) # formats charts and plots
plot.ts(UShomeSales, main = 'UShomeSales - Training')
nt=length(UShomeSales)
fit <- lm(UShomeSales ~ as.numeric(1:nt)); abline(fit, col="red") # regression
abline(h=mean(UShomeSales), col="blue") # adds regression line
hist(UShomeSales, main='Histogram - UShomeSales', col='cyan') # plots histogram
acf(UShomeSales, main = 'ACF UShomeSales') # Fits an ACF
```

Variance, Mean, Box Cox check

```
var(UShomeSales) # variance of training set
mean(UShomeSales) # Mean of training set
```

```
bcTransform <- boxcox(UShomeSales~ as.numeric(1:length(UShomeSales)))  # plots the graph
lam <- bcTransform$x[which(bcTransform$y == max(bcTransform$y))]  # finds lambda
```

Decomposition

```
y <- ts(as.ts(UShomeSales), frequency = 12)
plot(decompose(y)) # plots decomposition
```

Differencing

```
home.1 <- diff(UShomeSales, lag=1) # differences at lag 1
plot.ts(home.1)
nt=length(home.1)
fit <- lm(home.1 ~ as.numeric(1:nt)); abline(fit, col="red") # regression and trend line
abline(h=mean(home.1), col="blue") # adds mean as a constant


var(home.1) # checking lower variance and mean
mean(home.1)


acf(home.1, lag.max = 30)

home.12 <- diff(home.1, lag = 12) # differences at lag 12 using lag 1 difference
plot.ts(home.12)
nt=length(home.12)
fit <- lm(home.12 ~ as.numeric(1:nt)); abline(fit, col="red") # adds trend line


var(home.12) # lower again
mean(home.12)
```

Plots ACF and Histogram of differenced data

```
par(mfrow=c(1,2))
acf(UShomeSales, lag.max = 30, main='ACF Trainig Set')
acf(home.1, lag.max = 30, main='ACF, Differenced at lag 1')
acf(home.12, lag.max = 30, main='ACF, Differenced at lags 1 and 12')


par(mfrow=c(1,2))
hist(UShomeSales, main='Original Truncated', col='cyan')
hist(home.1, main='After Differencing at lag 1', col = 'orange')
hist(home.12, main='After Differencing lags 1 and 12', col='red')


par(mfrow=c(1,2))
acf(home.12, lag.max = 40, main='ACF Differenced lags 1 and 12')
pacf(home.12, lag.max = 40, main='PACF Differenced lags 1 and 12')
```

Model buliding

```
model1 <- arima(UShomeSales, order=c(0,1,0), seasonal=list(order=c(2,1,0),
                                                     period=12), method='ML')
```

```
model1
AICc(model1) # Akaike information criterion corrected

model2 <- arima(UShomeSales, order=c(1,1,0), seasonal=list(order=c(2,1,0),
                                                   period=12),method='ML') #This one works
model2
AICc(model2)

model3 <- arima(UShomeSales, order=c(1,1,1), seasonal=list(order=c(2,1,0),
                                                   period=12),method='ML')
model3
AICc(model3)
```

Check Invertibility

```
source("plot.roots.R.txt") # Loads functions from file
par(mfrow=c(1,2))

# plots roots w/ unit cirlce
plot.roots(NULL, polyroot(c(1,0.6484,0.3540)), main='Model A - Roots of SAR part')
plot.roots(NULL, polyroot(c(1,0.6492,0.3507)), main='Model B - Roots of SAR part')
```

Diagnostic Checking

```
modelA <- model2
resA <- residuals(modelA)
hist(resA,density=10,breaks=10, col="blue", xlab="", prob=TRUE)
m <- mean(resA)
std <- sqrt(var(resA))
curve( dnorm(x,m,std), add=TRUE )
plot.ts(resA)
fitt <- lm(resA ~ as.numeric(1:length(resA))); abline(fitt, col="red")
abline(h=mean(resA), col="blue")
qqnorm(resA,main= "Normal Q-Q Plot for Model A")
qqline(resA,col="blue")
par(mfrow=c(1,2))
acf(resA, lag.max=40)
pacf(resA, lag.max=40)
shapiro.test(resA)
Box.test(resA, lag = 11, type = c("Box-Pierce"), fitdf = 3)
Box.test(resA, lag = 11, type = c("Ljung-Box"), fitdf = 3)
Box.test(resA^2, lag = 11, type = c("Ljung-Box"), fitdf = 0)
acf(resA^2, lag.max=40)
ar(resA, aic = TRUE, order.max = NULL, method = c("yule-walker"))
```

Forecasting

```
forecast(modelA) # plots forecast with prediction interval
pred.tr <- predict(modelA, n.ahead=12) # Prediction on transformed data
U.tr <- pred.tr$pred+2*pred.tr$se      # Upper bound
L.tr <- pred.tr$pred-2*pred.tr$se      # Lower bound
ts.plot(UShomeSales, xlim=c(1,length(UShomeSales)+12)) # plots forecasted values from training set
```

```
lines(U.tr, col="blue", lty="dashed")
lines(L.tr, col="blue", lty="dashed")
points((length(UShomeSales)+1):(length(UShomeSales)+12), pred.tr$pred, col="red")
ts.plot(sales, xlim = c(100,length(UShomeSales)+12),ylab=’Houses Sold (thousands)’
        , col="red")                   # plots forecasts with raw data
lines(U.tr, col="blue", lty="dashed")
lines(L.tr, col="blue", lty="dashed")
points((length(UShomeSales)+1):(length(UShomeSales)+12), pred.tr$pred, col="green")
```