PSTAT 174 FINAL PROJECT

Residential Construction Costs: A Time Series Analysis

Authors:
Tiffany CHANG
Sriharsha ADDEPALLI
Qi GUO
Permoon HUSSAINI
David ING

Supervised By:
Professor Sudeep BAPAT

Contents

Abstract

The data set of monthly construction costs provides the average cost each month for building or improving structures that have been classified as residential according to the US Census Bureau. These residential buildings include single family homes, apartments and condominiums. The costs represent the total of new material expenses, labor costs, and other administrative costs such as office fees and interest fees. Monthly costs were collected from January 2002 to September 2018, but we removed the last 10 observations to compare our forecast from November 2017 to September 2018 to our original data. We performed exploratory data analysis on our set to remove seasonality, trend and create constant variance. Subsequently, we determined the model that best fit our data to be $SARIMA(2,1,0)\ X\ (1,1,0)_{12}$ and we used this model to forecast costs for the next 10 months. Our predicted values are within the 95% confidence limit and approximately close to the actual observed data for the 10 months until September 2018.

1 Introduction

Forecasting residential construction costs in a particular neighborhood is useful to potential homeowners and investors because it provides pertinent information as to what they can expect to budget for a particular project. Knowing the costs in advance allows investors or homeowners to plan accordingly and set aside the right amount of money that they need to complete the necessary construction or renovations. If the relevant information and estimated values were unknown, the volatility in both material and labor costs could pose as an economic danger between all the parties involved in a contract. Therefore, we hope to hedge against that risk by forecasting costs for future months and by providing sound estimates of future costs using a model that we believe best represents our data.

2 Data Exploration & Analysis

2.1 Initial Data Overview & Time Series Plot

Our data consisted of 201 observations. The two variables analyzed from the data set are the months in which construction spending costs are extracted and the value, or the average construction spending per unit in the United States, for the given month. For our initial data exploration, we plotted the construction spending against the time from January 2002 to September 2018 by month.

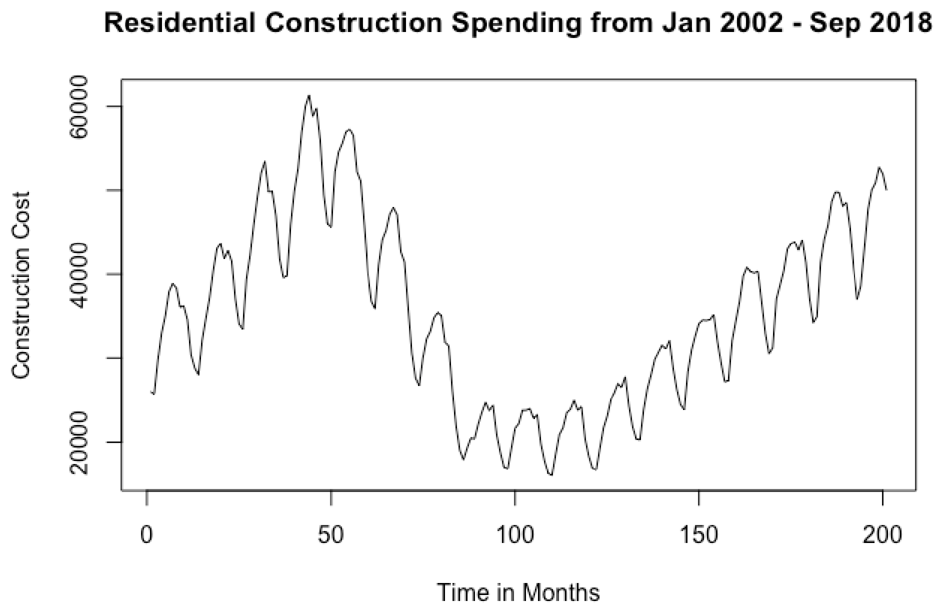


Figure 2.1

From the plot, we can clearly see that there is seasonality in our data set because of the similar spikes at constant intervals throughout the entire length of time observed. We also noticed a trend initially going upward, then downward, and upward again. Both phenomena indicate to us that we will need to de-seasonalize and detrend our data before we can perform analysis.

2.2 Seasonal Plot

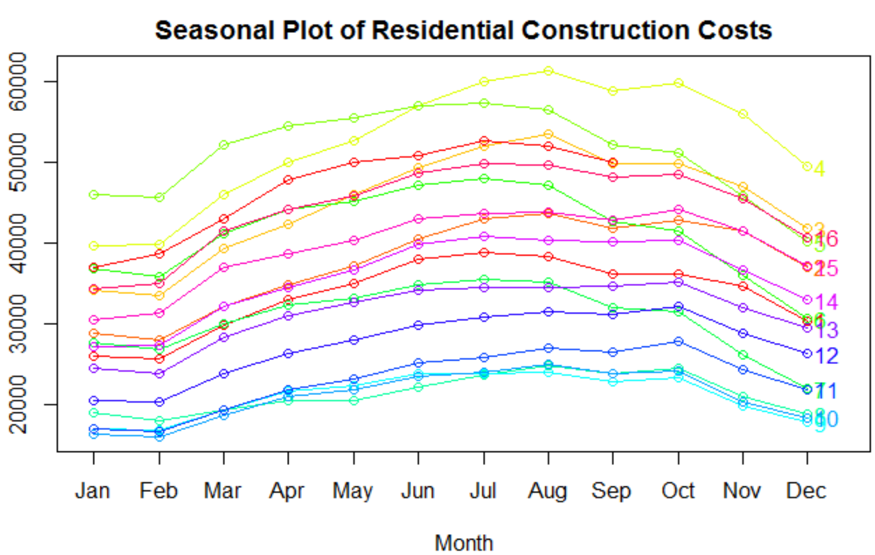


Figure 2.2

From our seasonal plot, we can observe that there is an increase of construction cost from February to August and then a gradual decrease from August to the end of the year for each year. The cost of construction is lowest during the months of December to January, possibly because of the holiday season in which construction work overall is less likely to take place. However, we still need to claim the existence of seasonality and trend for our data. As such, we proceed to transform our data and de-seasonalize and remove the trend as well.

3 Data Transformation

3.1 Box-Cox Transformation

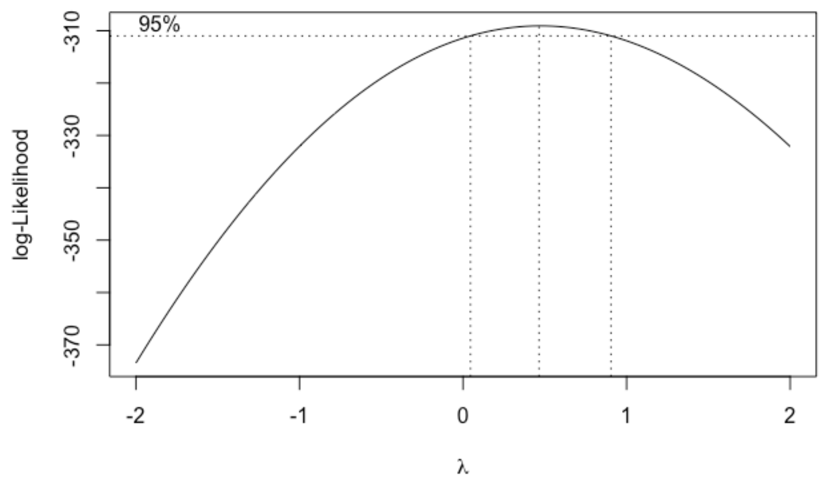


Figure 3.1

From our Box-Cox Transformation plot, we notice that λ is relatively close to $\frac{1}{2}$ (λ ≈0.4646). Following that, we applied the square root transformation since lambda from the Box-Cox transformation tells us what exponent value to raise our data to and raising to $\frac{1}{2}$ is taking the square root of our data. Our transformed data equation would be:

$$V_t = {Y_t}^{0.4646} \approx {Y_t}^{0.5} = \sqrt{Y}$$

Based on the following plot of the square root transformed data, we can see that seasonal variation has been reduced from before.

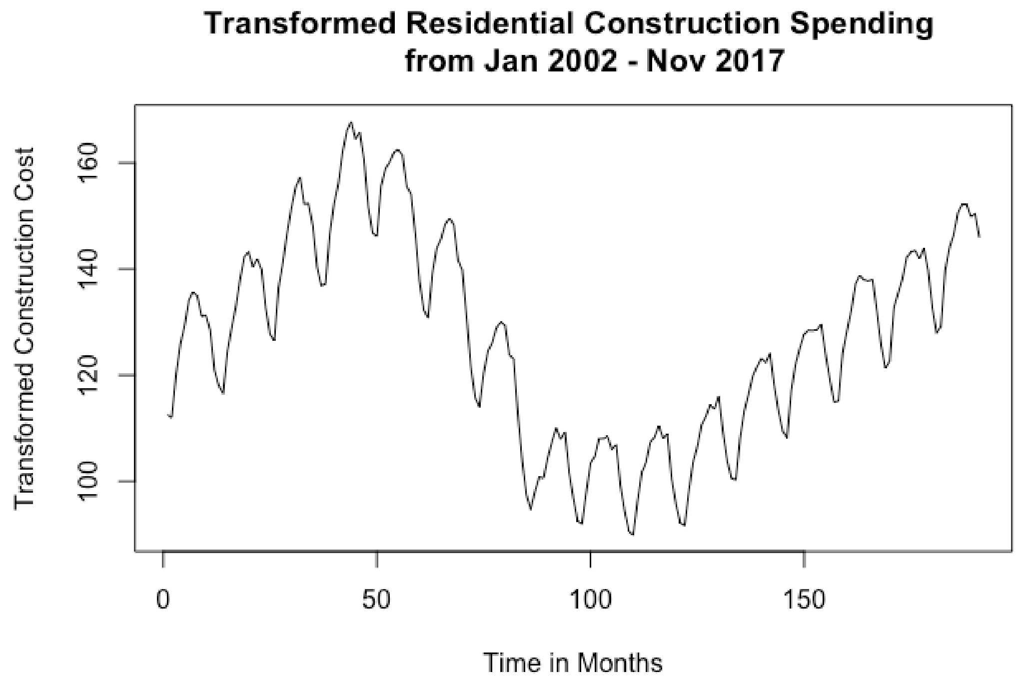


Figure 3.2

In Figure 3.2, we still see the seasonality and trend that we observed in Figure 2.1, so we proceed to de-seasonalize and detrend our transformed data using differencing. In this plot we have removed the last 10 observations from our data set so that we can use those values to compare with our predicted values later in this report.

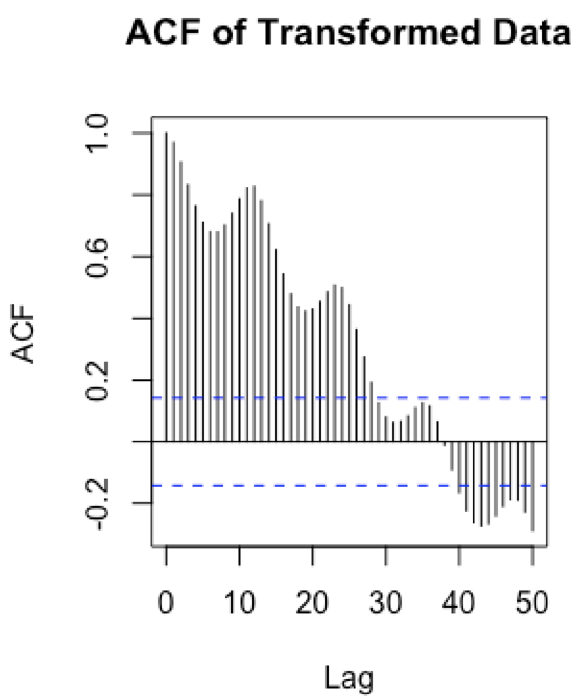


Figure 3.3

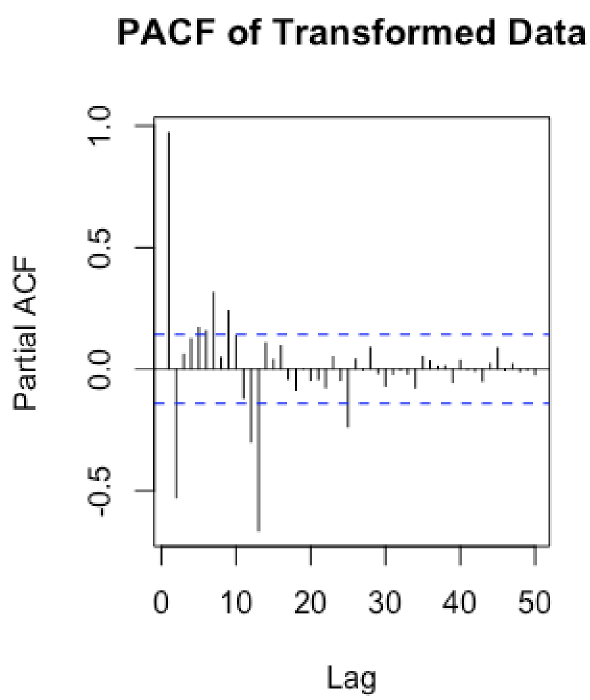


Figure 3.4

Looking at the ACF and PACF plot of the transformed data, we see that the PACF seems to cut off after lag 12 and the ACF is tailing off. This gives us a basic intuition that our seasonal component model may be following an AR(1) process. However, we have many observations that are outside of the confidence intervals, so our model is still not stationary, so we have to de-seasonalize and detrend our data.

3.2 De-seasonalizing and Detrending Data

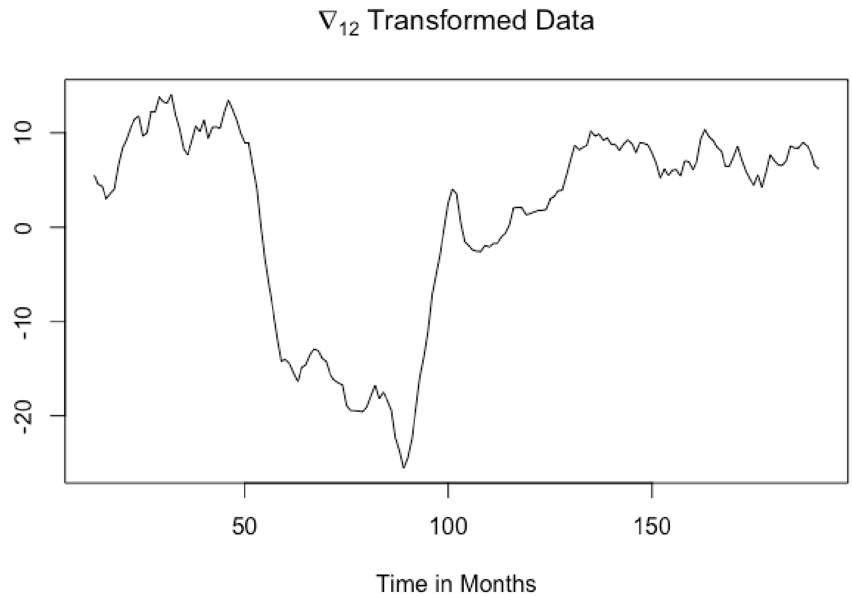


Figure 3.5

In Figure 3.5, we see what our data looks like once we have removed seasonality from it by differencing once at lag 12. The variance of this data is 110.92. There are no more of the cyclical hills and valleys that we witnessed in our original data, but our current data still does not appear to be stationary. As such, we will difference the data again but at lag 1 to see what effect the removal of trend will have on our data.

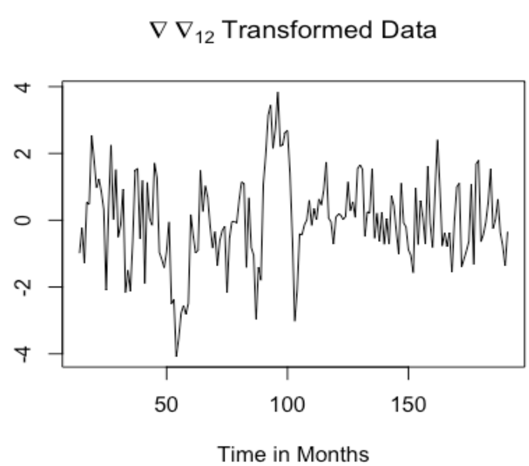


Figure 3.6

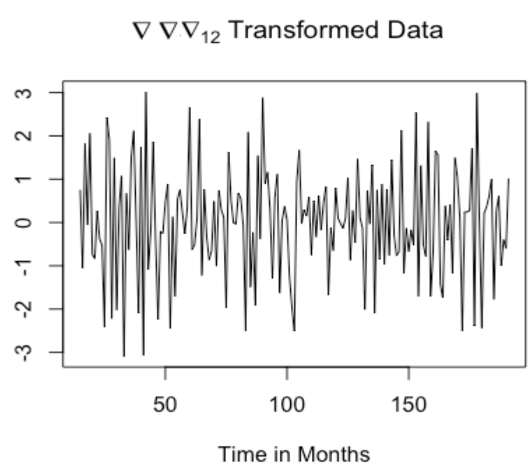


Figure 3.7

As seen in Figure 3.6 above, the data appears to be more stationary that before we had differenced our data at lag 1 to remove the trend. The variance of our data at this point is 1.85, which is tremendously lower than what is was before.

In Figure 3.7 we can see that after differencing our data once again at lag 1, the plot looks even more stationary than when we had only differenced it at lag 1 once. The variance of our data at this point is 1.646 which is also again lower than the variance than when we only differenced at lag 1 once. However, if we were to difference our data yet again at lag 1, for the third time, we would see that the variance has increased to 4.42, an effect of over-differencing which indicates to us that we should have our data be differenced at lag 1 twice and not three times.

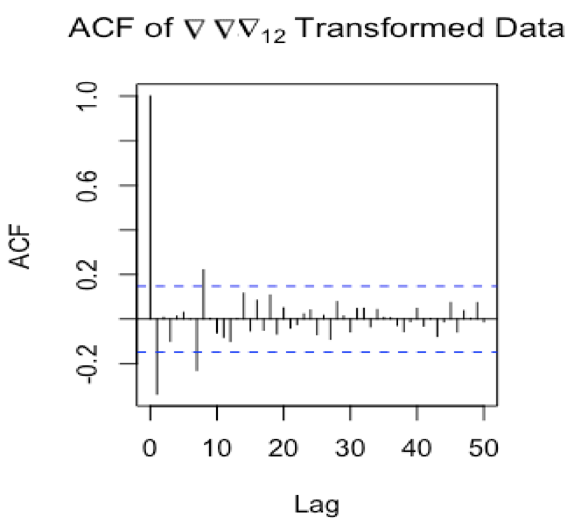


Figure 3.8

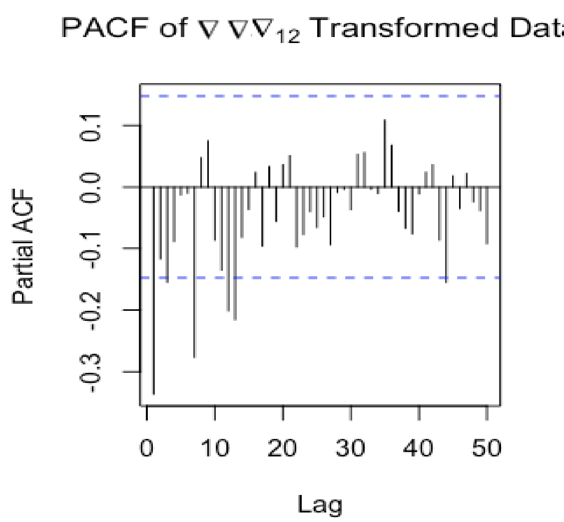


Figure 3.9

In Figure 3.8 & 3.9, we can see the ACF and PACF of our data after de-seasonalizing at lag 12 and differencing twice at lag 1. Both the PACF and ACF graphs appear to tail off after lag 1 so we make a conjecture that our model could possibly be of the form $SARIMA(p, d, q)\ X\ (P, D, Q)_{12}$. These ACF and PACF graphs are a significant improvement from those of our original data set with no differencing at all.

4 Model Identification

Because our data has a seasonal component at lag 12, we are using a $SARIMA(p, d, q)\ X\ (P, D, Q)_{12}$ model to analyze our data. Using Figure 3.3 and 3.4, we notice that the PACF cuts off and the ACF tails off so there is no MA component and there is an AR component. Therefore, we are conjecturing that P = 1 and Q = 0. Because we de-seasonalized once at lag 12, our D = 1. Because we detrended twice at lag 1, our d = 2. We estimate orders p and q below.

4.1 Observed Models and Order

	q.0	q.1	q.2	q.3	q.4
p=0	1.591773	1.381533	1.274167	1.273877	1.264838
p=1	1.251877	1.235650	1.244895	1.254664	1.258670
p=2	1.232643	1.230226	1.231837	1.241982	1.250794
p=3	1.243388	1.233635	1.242440	1.233767	1.249062
p=4	1.249970	1.240915	1.227313	1.195244	1.211414

Figure 4.1

Here in Figure 4.1 we observe the AICc values for all possible combinations of our $SARIMA(p,d,q)\ X\ (P,D,Q)_{12}$ model.

4.2 Model Selection

We observe the following models and their respective AICc values. These 4 models have been chosen based on the criteria that models with lower AICc values are a better fit for our data than models with higher AICc values.

$$\begin{aligned} Model\ 1&: SARIMA(4,2,4)\ X\ (1,1,0)_{12} &= 1.2114 \\ Model\ 2&: SARIMA(2,2,0)\ X\ (1,1,0)_{12} &= 1.2326 \\ Model\ 3&: SARIMA(4,2,3)\ X\ (1,1,0)_{12} &= 1.1952 \\ Model\ 4&: SARIMA(2,2,1)\ X\ (1,1,0)_{12} &= 1.2302 \end{aligned}$$

We rule out Model 1 in compliance with the Law of Parsimony which states that models with fewer terms are better than models with more terms. We chose to rule out Model 1 over Model 3 because Model 3 has a lower AICc value and only 7 $(4+3)$ terms rather than 8 in Fit 1. The final three models that we will be working with and comparing to determine which best represents our data are Models 2, 3, and 4.

4.3 Parameter Estimation

After running the commands to determine the coefficients for these models using maximum likelihood estimation (as seen in the appendix), we find that the models represented in algebraic form are:

$$Model\ 2: SARIMA(2,2,0)\ X\ (1,1,0)_{12} \rightarrow (1-0.4476B-0.1723B^2)(1+0.16B^{12})X_t = Z_t$$

$$\begin{aligned} &Model\ 3: SARIMA(4,2,3)\ X\ (1,1,0)_{12} \\ &\rightarrow (1-0.3296B+0.0689B^2+0.6734B^3-0.5177B^4)(1+0.1722B^{12})X_t \\ &= (1+0.1278B+0.3141B^2+0.8728B^3)Z_t \end{aligned}$$

$$\begin{aligned} &Model\ 4: SARIMA(2,2,1)\ X\ (1,1,0)_{12} \rightarrow (1+0.3553B-0.5599B^2)(1+0.1921B^{12})X_t \\ &= (1+0.8748B)Z_t \end{aligned}$$

5 Diagnostics

5.1 Checking for Stationarity and Invertibility

Based on the root plots generated from these 3 models, found in the appendix, we have determined that Model 3 is eliminated from our possible models. Model 3 has unit roots (red stars) in its MA part and its roots are less than 1 in the AR part, indicating lack of stationarity and invertibility. The remaining two models, Model 2 and Model 4, both pass this check for their roots.

5.2 Checking for Normality

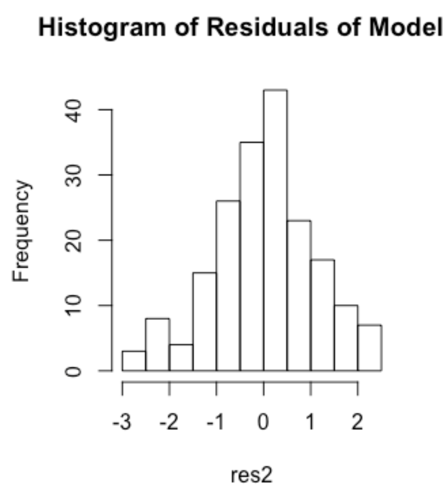


Figure 5.1

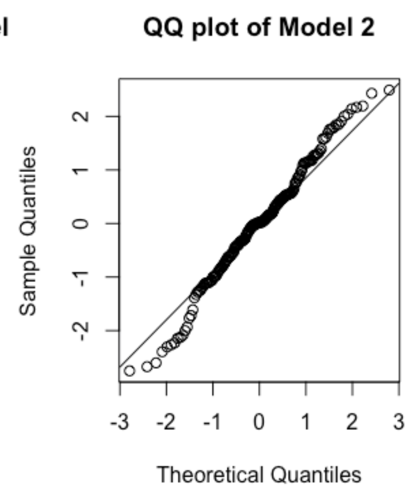


Figure 5.2

As seen in Figure 5.1, the histogram of the residuals from Model 2 are approximately normal and as seen in Figure 5.2, the QQ plot of Model 2 shows that the sample and theoretical quantiles form nearly a straight line with a small amount of deviation at the ends. We then conclude that the residuals of Model 2 pass the Normality Assumption.

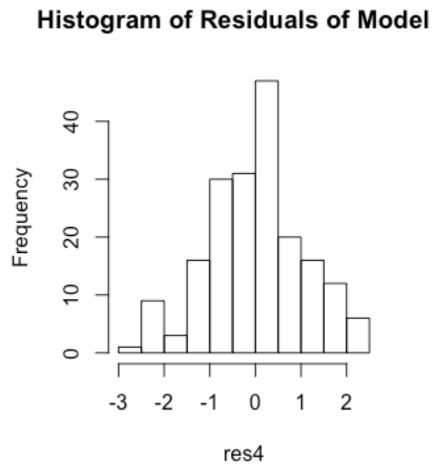


Figure 5.3

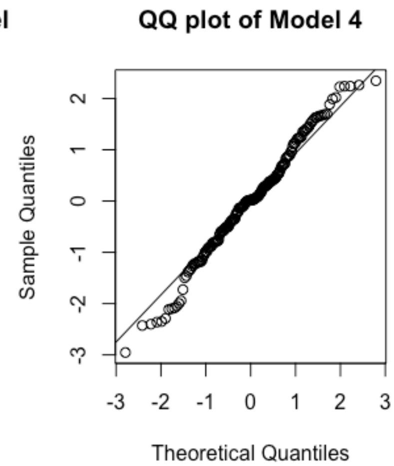


Figure 5.4

As seen in Figure 5.3, the histogram of the residuals from Model 4 are approximately normal and as seen in Figure 5.4, the QQ plot of Model 4 shows that the sample and theoretical quantiles form nearly a straight line with a small amount of deviation at the ends. We then conclude that the residuals of Model 4 pass the Normality Assumption.

$$H_0 = Residuals\ are\ normal$$
$$H_a = Residuals\ are\ not\ normal$$

For the Shapiro-Wilk Test, both Models 2 and 4 pass the test as well. For Model 2, we get a p-value of 0.12 and for Model 4 we get a p-value of 0.14, both of which are greater than our chosen alpha level of 0.05. Thus, we fail to reject the null hypothesis that Models 2 and 4 are normal. There is insufficient support from our data to conclude otherwise. Hence, the residuals of both Models 2 and 4 pass the Normality Check.

5.3 Checking for Serial Correlation

$$H_0 = Residuals\ are\ serially\ uncorrelated$$
$$H_a = Residuals\ are\ not\ serially\ uncorrelated$$

As seen in the appendix where we conducted the Box-Pierce test, both Models 2 and 4 pass the Box-Pierce test. For Model 2, we get a p-value of 0.32 and for Model 4 we get a p-value of 0.24, both of which are greater than our chosen alpha level of 0.05. As such, we fail to reject our null hypothesis that the residuals are serially uncorrelated.

5.4 Checking for Constant Variance

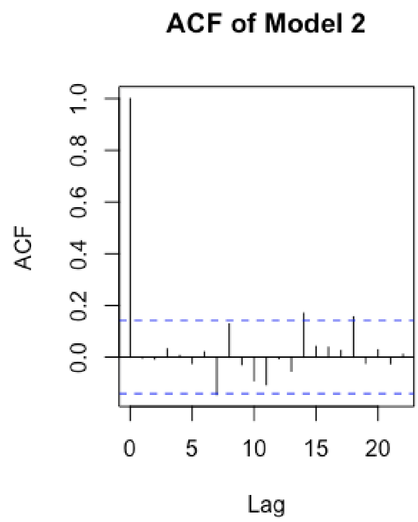


Figure 5.5

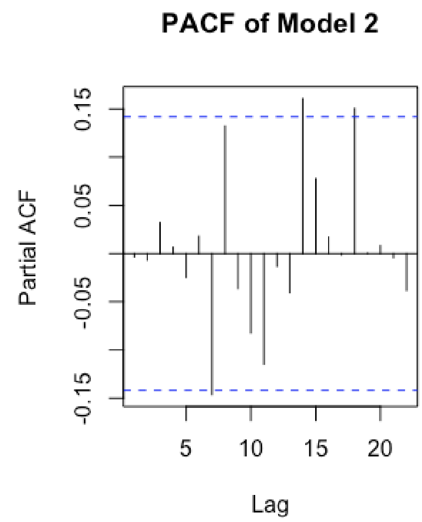


Figure 5.6

Above, we can see how for Model 2, values for both the ACF and the PACF remain relatively within the 95% confidence bounds. However, in the PACF graph, we see that there are actually a few values that exceed the bounds, although not by that much, and these may be considered outliers in our data, leading to the possibility of non-constant variance.

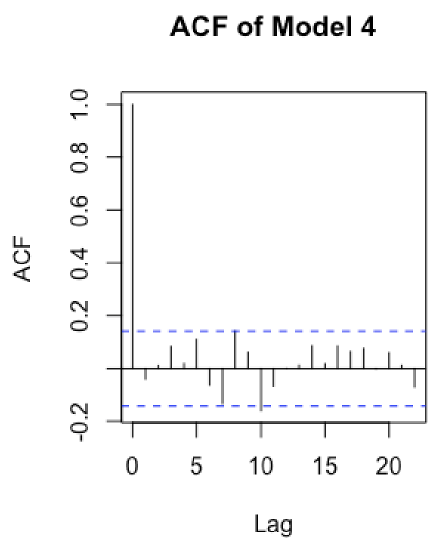


Figure 5.7

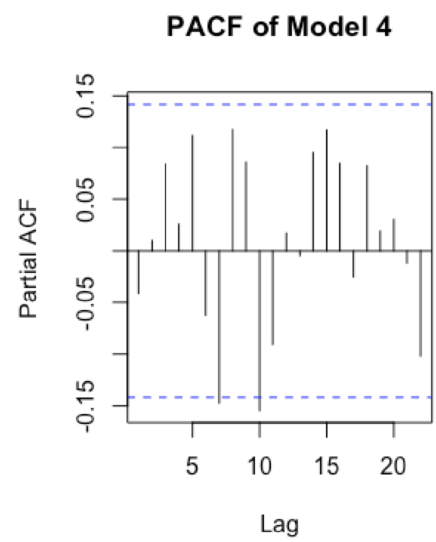


Figure 5.8

Here, in Figures 5.7 and 5.8, we see that the ACF and PACF values for Model 4 also remain relatively within the bounds. Fewer values of the PACF for Model 4 exceed the bounds and by much less than in Model 2, indicating a smaller chance of violating the assumption of constant variance.

Since both models seem to pass the diagnostic checks, for our final model, we choose Model 4 because the ACF's of the residuals seem to be more within the 95% confidence intervals. In addition, we found that Model 4 has a lower AICc than Model 2, further justifying our claim.

Final Model:

$$Model\ 4: SARIMA(2,2,1)\ X\ (1,1,0)_{12}$$
$$\rightarrow (1 + 0.3553B - 0.5599B^2)(1 + 0.1921B^{12})X_t \ = (1 + 0.8748B)Z_t$$

6 Forecasting

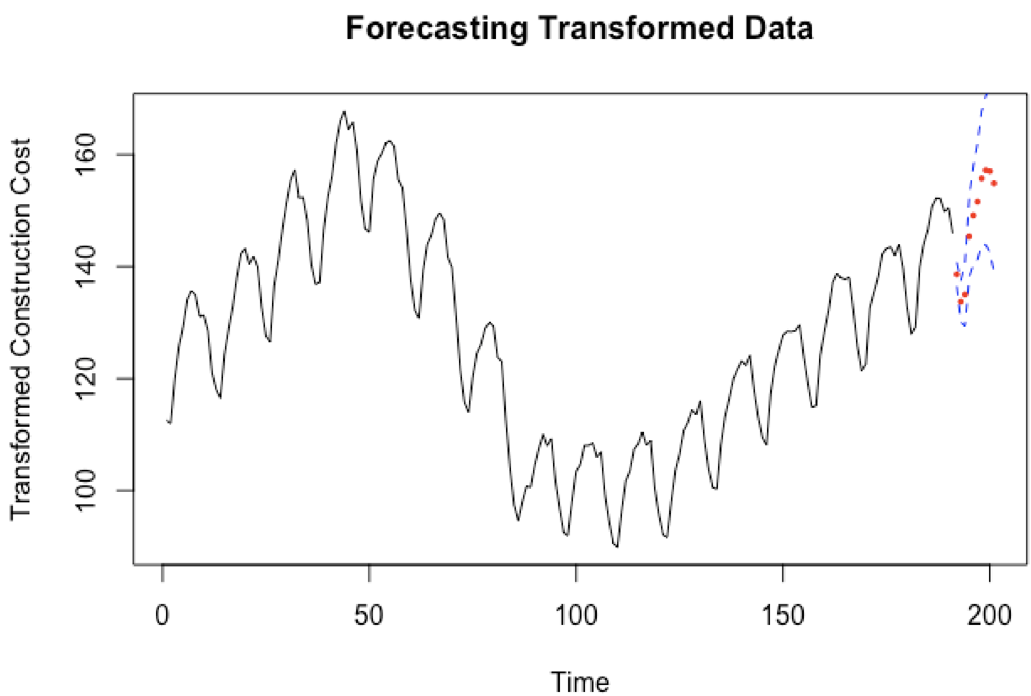


Figure 6.1

In Figure 6.1 we have a graph of our transformed time series data with 10 of the latest observations removed. In their place, we have added our forecasted values of the transformed data as indicated by the red dots and included the 95% confidence intervals around those values as well.

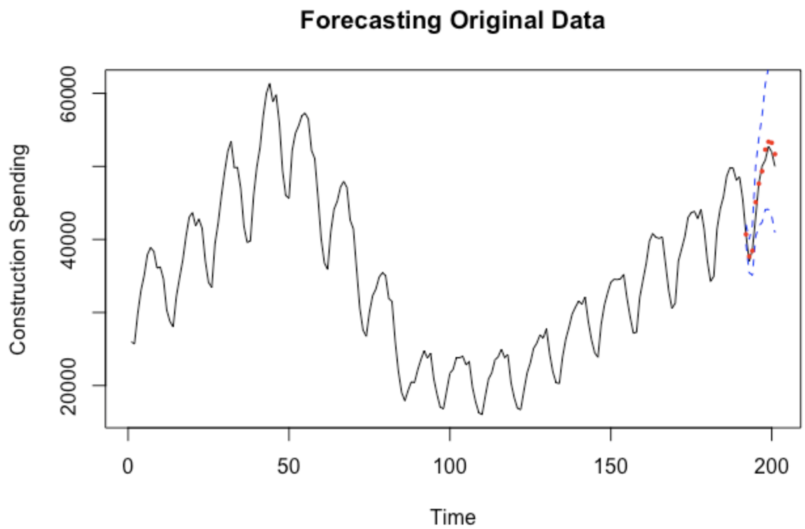


Figure 6.2

In Figure 6.2, we have the forecasted values for our original time series data, which was our values above but squared, since according to our Box-Cox lambda value, we were supposed to transform our data by taking the square root of it and now to transform it back, we take the square of it.

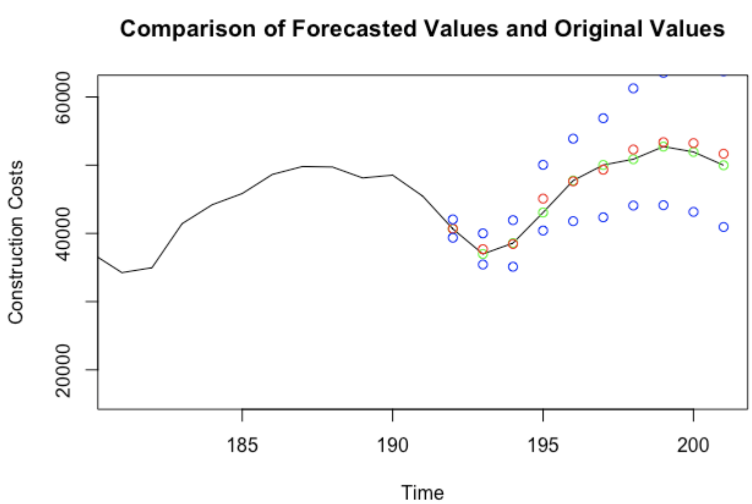


Figure 6.3

In Figure 6.3 we see the forecasted values of our data imposed over the actual data that was observed. This graph has been zoomed in to show the individual points, in which green points represent the actual collected data, red points represent the forecasted data and the blue points indicate the 95% confidence intervals. From this graph we can clearly see that our forecasted values are very close to the actual data and are good estimates of the actual data.

7 Conclusion

Our original goal in this project was to predict values from our data set that would be very close to the actual observed data for the average construction costs for a unit every month from December 2017 to September 2018. We ultimately found that our data, through transformations and differencing, was best represented by a $SARIMA(2,2,1)\ X\ (1,1,0)_{12}$ model:

$$(1 + 0.3553B - 0.5599B^2)(1 + 0.1921B^{12})X_t \ = (1 + 0.8748B)Z_t$$

Using this particular model, we found that it passed all of the necessary diagnostic checks in order to perform forecasting. We then proceeded to predict the values for the next 10 months after November 2017 and found that our forecasted values are accurate with the actual data. Not only were our forecasted data within our bounds but they were very strong estimators for the actual data since there was little to no difference between the two values. With these forecasted values and knowing their high level of accuracy, we hope that we could eventually be able to predict future construction costs that can help to mitigate the risk of volatile costs in constructing residences.

8 Reference

Merryman, Ray, and US Census Bureau. “US Census Bureau Construction Spending Survey.” *Census Bureau QuickFacts*, United States Census Bureau, 10 Aug. 2010, www.census.gov/construction/c30/methodology.html.

Merryman, Ray, and US Census Bureau. “US Census Bureau Construction Spending Survey.” *Census Bureau QuickFacts*, United States Census Bureau, 10 Aug. 2010, www.census.gov/construction/c30/definitions.html.

9 Appendix

Parameter Estimation for Models:

Model 2:

```
Coefficients:
         ar1     ar2     sar1
      0.4476  0.1723  -0.1600
s.e.  0.0745  0.0749   0.0773

sigma^2 estimated as 1.209:  log likelihood = -269.81,  aic = 545.61
```

Model 3:

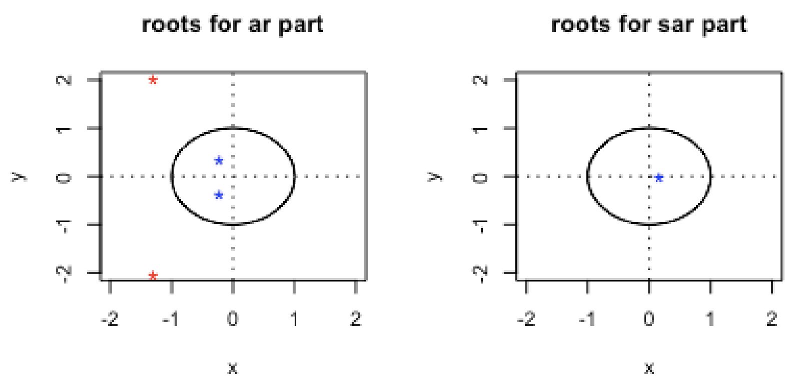
```
Coefficients:
         ar1      ar2      ar3     ar4     ma1     ma2     ma3     sar1
      0.3296  -0.0689  -0.6734  0.5177  0.1278  0.3141  0.8728  -0.1722
s.e.  0.0917   0.0799   0.0846  0.0669  0.0800  0.0681  0.0873   0.0784

sigma^2 estimated as 1.11:  log likelihood = -263.5,  aic = 542.99
```

Model 4:

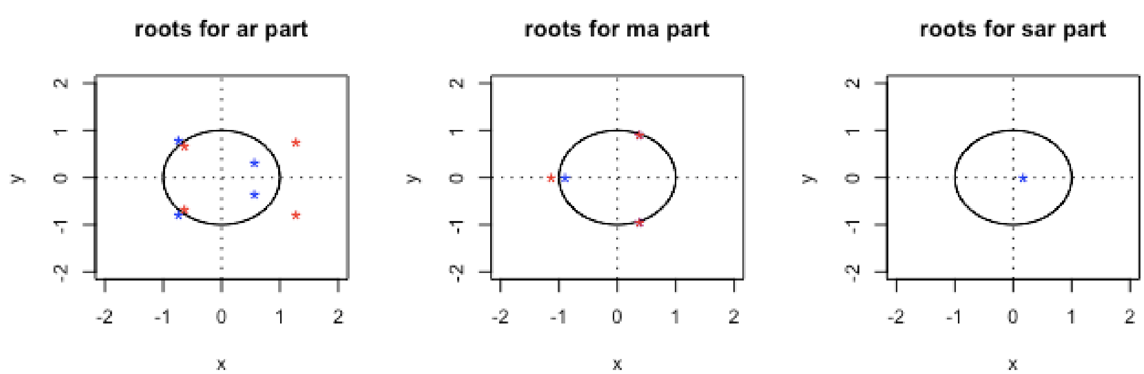
```
Coefficients:
          ar1     ar2     ma1     sar1
      -0.3553  0.5599  0.8748  -0.1921
s.e.   0.0853  0.0629  0.0783   0.0768

sigma^2 estimated as 1.193:  log likelihood = -268.7,  aic = 545.4
```

Root Plots for Models:

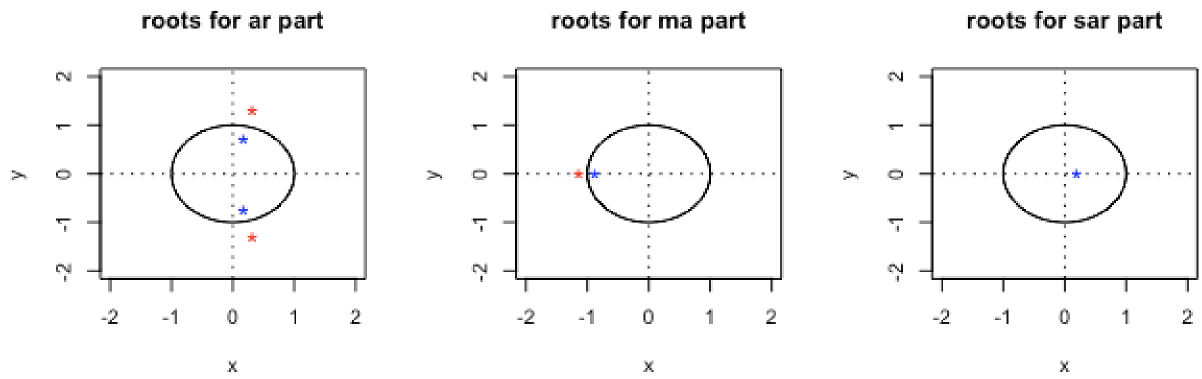
Model 2:



Model 3:



Model 4:



Shapiro-Wilk Test for Normality:

Model 2: Shapiro-Wilk normality test

```
data:  res2
W = 0.98846, p-value = 0.1242
```

Model 4: Shapiro-Wilk normality test

```
data:  res4
W = 0.98882, p-value = 0.1394
```

Box-Pierce Test for Serial Correlation:

Model 2: Box-Pierce test

```
data:  res2
X-squared = 11.514, df = 10, p-value = 0.3189
```

Model 4: Box-Pierce test

```
data:  res2
X-squared = 11.514, df = 9, p-value = 0.2421
```

RCode

```
#Data exploration
const <- read.csv("/Users/sriadd/Downloads/total-construction-spending-residential.csv")
const_res <- ts(const[,2],frequency = 1)
plot.ts(const_res, main = "Residential Construction Spending from Jan 2002 - Sep 2018", xlab = "Time in
Months", ylab = "Construction Cost")
seasonplot(const_res,12, col = rainbow(16), year.labels = T, main = "Seasonal Plot of Residential
Construction Costs")

#Box Cox
require(MASS)
bc <- boxcox(const_res ~ as.numeric(1:length(const_res)))
lambda <- bc$x[which(bc$y == max(bc$y))] #lambda = 0.4646
trans_const <- const_res ^ lambda
trans_const1 <- ts(trans_const [1:(length(const_res) - 10)]) # removing 10 observations
plot(trans_const1, main = "Transformed Residential Construction Spending
     from Jan 2002 - Nov 2017", xlab = "Time in Months", ylab = "Transformed Construction Cost")
op <- par(mfrow = c(1,2))
acf(trans_const1, lag.max = 50, main = "ACF of Transformed Data")
pacf(trans_const1, lag.max = 50, main = "PACF of Transformed Data")
par(op)
const.diff <- diff(trans_const1, lag = 12) #differencing for deseasonalization
plot(const.diff, main=expression(nabla[12] ~ "Transformed Data") , xlab = "Time in Months", ylab = "")
const.diff.nt <- diff(const.diff, differences = 1, lag = 1)# differencing for detrending
var(const.diff.nt)
op <- par(mfrow = c(1,2))
```

```
#Differencing second time because lower variance
plot(const.diff.nt, main=expression(nabla ~ nabla[12] ~ "Transformed Data") , xlab = "Time in Months",
                ylab = "")
plot(diff(const.diff,differences = 2, lag = 1), main=expression(nabla ~ nabla ~ nabla[12] ~ "Transformed
Data") , xlab = "Time in Months", ylab = "")
par(op)
op <- par(mfrow = c(1,2))
acf(diff(const.diff,differences = 2, lag = 1), lag.max = 50, main = expression("ACF of " ~ nabla ~ nabla ~
nabla[12] ~ "Transformed Data"))
pacf(diff(const.diff,differences = 2, lag = 1), lag.max = 50, main = expression("PACF of " ~ nabla ~ nabla
~ nabla[12] ~ "Transformed Data"))
par(op)

#Finding AIC values to select model
AICc <- numeric()
for (p in 0:4) {
  for (q in 0:4) {
    AICc <- c(AICc , sarima(trans_const2, p, 2, q, 1, 1, 0, 12, details=FALSE)$AICc)
  }
}
AICc <- matrix(AICc ,nrow=5,byrow=TRUE)
rownames(AICc) <- c("p=0" , "p=1" , "p=2", "p=3", "p=4")
colnames (AICc) <- c("q=0" , "q=1" , "q=2", "q=3", "q=4" )
AICc <- data.frame(AICc)
AICc #gives dataframe of various p and q values to test

#models with good AIC values and fit law of parsimony
fit2 <- arima(trans_const1, order = c(2,1,0), seasonal = list(order = c(1,1,0), period = 12), method =
"ML")
fit3 <- arima(trans_const1, order = c(4,1,3), seasonal = list(order = c(1,1,0), period = 12), method =
"ML")
fit4 <- arima(trans_const1, order = c(2,1,1), seasonal = list(order = c(1,1,0), period = 12), method =
"ML")
fit2
fit3
fit4

#Fit2 stationary/invertibility plots
op <- par(mfrow = c(2,3))
plot.roots(NULL, polyroot(c(1,0.4476,0.1723)), main = "roots for ar part")
plot.roots(NULL, polyroot(c(1,-0.1600)), main = "roots for sar part")
par(op)

#Fit3 stationarity/invertibility plots
op <- par(mfrow = c(2,3))
plot.roots(NULL, polyroot(c(1,0.3296,-0.0689,-0.6734,0.5177)), main = "roots for ar part")
plot.roots(NULL, polyroot(c(1,0.1278,0.3141,0.8728)), main ="roots for ma part")
plot.roots(NULL, polyroot(c(1,-0.1722)), main = "roots for sar part")
par(op)

#Fit4 stationary/invertibility plots
op <- par(mfrow = c(2,3))
plot.roots(NULL, polyroot(c(1,-0.3553,0.5599)), main = "roots for ar part")
plot.roots(NULL, polyroot(c(1,0.8748)), main ="roots for ma part")
plot.roots(NULL, polyroot(c(1,-0.1921)), main = "roots for sar part")
par(op)

#diagnostic checking using residuals
res2 <- residuals(fit2)
res4 <- residuals(fit4)
```

```
#Normality
op <- par(mfrow = c(1,2))
hist(res2, main = "Histogram of Residuals of Model 2", breaks = 15)
qqnorm(res2, main = "QQ plot of Model 2")
qqline(res2)
par(op)
op <- par(mfrow = c(1,2))
hist(res4, main = "Histogram of Residuals of Model 4", breaks = 15)
qqnorm(res4, main = "QQ plot of Model 4")
qqline(res4)
par(op)
#Shapiro test for normality
shapiro.test(res2)
shapiro.test(res4)

#Box pierce test for serial correlation
box2 <- Box.test(res2, lag = 12, type = "Box-Pierce", fitdf = 2)
box4 <- Box.test(res2, lag = 12, type = "Box-Pierce", fitdf = 3)
box2
box4

#ACF PACF plots of residuals for  model 2 and 4
op <- par(mfrow = c(1,2))
acf(res2, main = "ACF of Model 2")
pacf(res2, main = "PACF of Model 2")
par(op)
op <- par(mfrow = c(1,2))
acf(res4, main ="ACF of Model 4")
pacf(res4, main = "PACF of Model 4")
par(op)

#forecasting- transformed data
pred.tr <- predict(fit2, n.ahead = 10)
upper <- pred.tr$pred + 2*pred.tr$se
lower <- pred.tr$pred - 2*pred.tr$se
ts.plot(trans_const1, xlim = c(1, length(trans_const1)+10), main ="Forecasting Transformed Data", ylab =
"Transformed Construction Cost")
lines(upper, col = "BLUE", lty = "dashed")
lines(lower, col = "BLUE", lty = "dashed")
points((length(trans_const1) +1):(length(trans_const1)+10), pred.tr$pred, col = "red", pch = 16, cex =
0.5)
#forecasting- original data
pred.origin <- pred.tr$pred ^ (1/lambda)
uorig <- upper ^ (1/lambda)
lorig <- lower ^ (1/lambda)
const_res2 <- const_res
ts.plot(const_res2, xlim = c(1, length(const_res2)), main ="Forecasting Original Data", ylab =
"Construction Spending")
lines(uorig, col = "BLUE", lty = "dashed")
lines(lorig, col = "BLUE", lty = "dashed")
points((length(trans_const1) +1):(length(trans_const1)+10), pred.origin, col = "red", pch = 16, cex = 0.5)

#comparison of original and forecasted points
ts.plot((const_res2), xlim = c(length(const_res2) - 20, length(const_res2)), main = "Comparison of
Forecasted Values and Original Values",
        ylab = "Construction Costs")
points(192:201, const_res2[192:201], col = "GREEN")
points(192:201, pred.origin, col = "RED")
points(192:201, uorig, lty = 1, col = "BLUE")
points(192:201, lorig, lty = 1, col = "BLUE")
```