Monthly CO2 Emissions

Time Series Analysis

Ashley Son

PSTAT 174

Table of Contents

ABSTRACT:

For this time series project, I will find the best model to forecast the last 10 values of the Atmospheric Carbon Dioxide Emissions dataset using the Box Jenkins approach.. To do so, I have transformed my data using a log transformation and by differencing twice for it to become stationary. I have then analyzed its ACF and PACF. Afterwards, I chose 4 possible SARIMA model candidates. When plotting my residuals, there was a noticeable outlier which may be random abnormal activity, I conducted my diagnostic checking after I extracted the outlier. I concluded that the best model was SARIMA(0,1,1) x (0,1,1)12, and successfully forecasted the next 10 values.

INTRODUCTION:

For this time series project, I have chosen monthly Carbon Dioxide Emissions dataset. This dataset is the monthly mean carbon dioxide measured at Mauna Loa Observatory, Hawaii and was found on Kaggle.

I chose this specific dataset because Mauna Loa Observatory has recently recorded the highest Atmospheric Carbon Dioxide emission record ever in human history in 2021. The increase of CO2 is warming up the planet which can cause instances of droughts, flooding, along with strong hurricanes and potential for unpredictable and dangerous storms and overall, a substantial amount of negative harm to our planet.

The data is reported as a dry air mole fraction defined as the number of molecules of carbon dioxide divided by the number of all molecules in air, including CO2 itself, after water vapor has been removed. The mole fraction is expressed as parts per million (ppm). In the data I will be extracting just the CO2 data from March, 1951 to December, 2021. I have withdrawn the last ten values within the data set to compare those values with my forecasted data.

For my project, the goal will be to accurately forecast the next 10 values of the CO2 emissions dataset. I will be using the software ***R-Studio*** for the analysis with the following packages: "*stats*", "*MASS*", "*astsa*", "*tseries*", "*forecast*", "*qpcR*", and "*data.table*". I will extract the last 10 values to compare with my forecasted values. Afterwards, I will make the transformations necessary for my dataset to be stationary. After proving stationary, I will plot the autocorrelations and partial autocorrelations to find possible SARIMA model candidates to forecast. Once I have done the diagnostic checking for these candidate models, I will find the best fitting model and utilize it to forecast the next 10 values and compare it with the previously extracted 10 values.

DATA ANALYSIS

PLOTTING AND ANALYZING THE TIME SERIES:

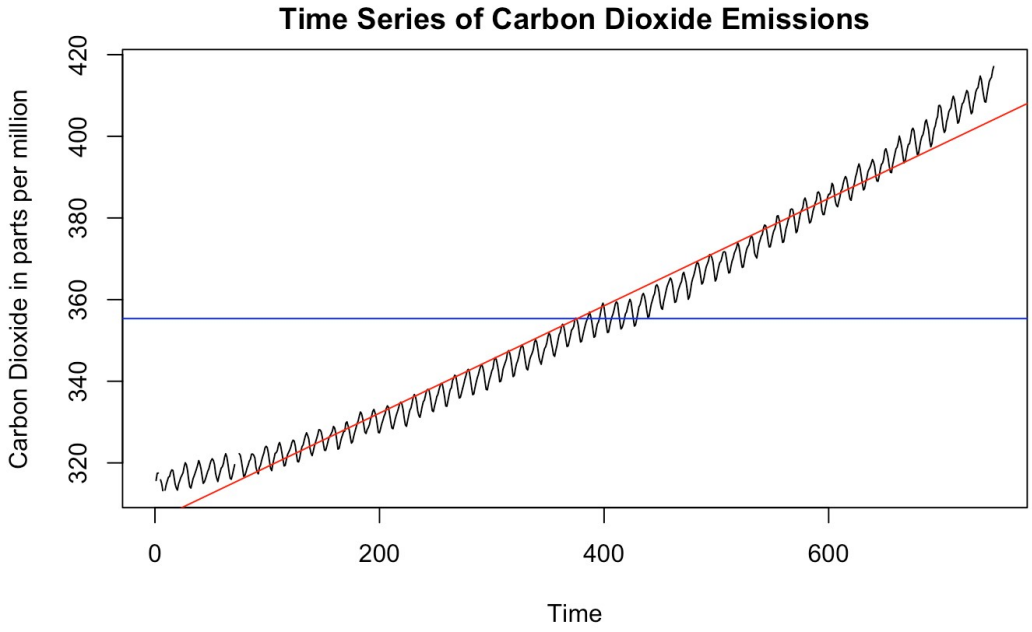

Figure 1: Original Atmospheric Carbon Dioxide Time Series Data

I took the original Atmospheric Carbon Dioxide data and plotted the time series. The linear trend line is plotted in red, and the mean (constant) line is plotted in blue. It is noticeable that this data is not stationary. There is an inconstant variance and an increasing positive trend with seasonality. Next, I will begin to transform the data to make it stationary.

DATA TRANSFORMATION

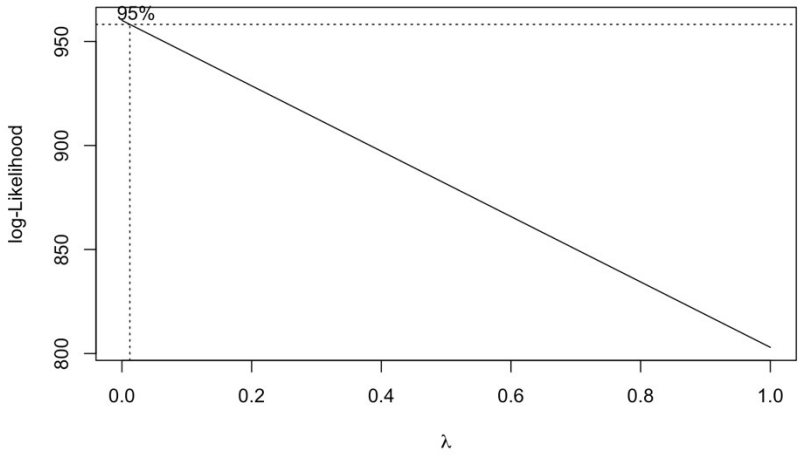

Figure 2: Box Cox Transformation of Original Atmospheric Carbon Dioxide Time Series Data

$$f_\lambda(U_t) = \begin{cases} \ln U_t, & \text{if } U_t > 0,\ \lambda = 0; \\ \lambda^{-1}(U_t^\lambda - 1), & \text{if } U_t \geq 0,\ \lambda \neq 0 \end{cases}$$

To decide on a transformation, I used the box cox transformation. Here, lambda yields to zero, so I will be taking the log transformation of this data.

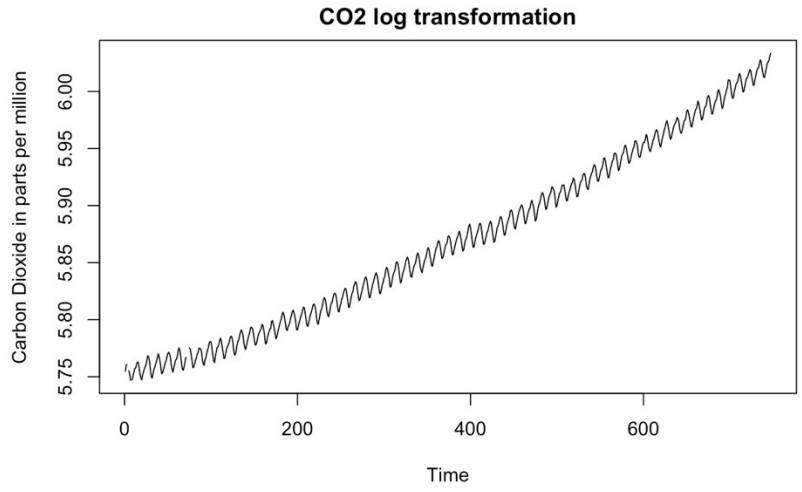

Figure 3: Log Transformation of Original Atmospheric Carbon Dioxide Time Series Data

The log transformation decreased the variance to 0.006345112, however there was still a noticeable increasing positive trend. I will proceed to difference the data at lag = 1.

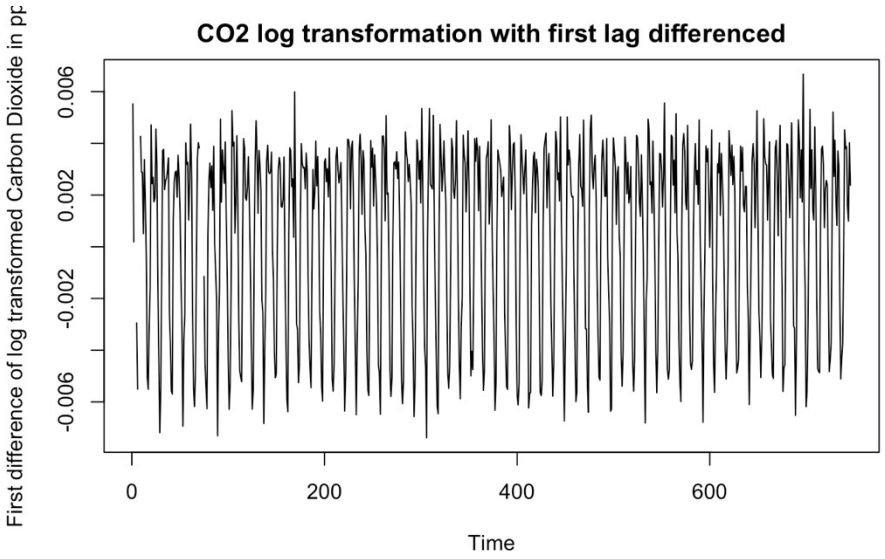

Figure 4: Log Transformation of CO2 Emissions Data with first difference at Lag = 1

To create a more stable variance, I took the first difference at lag = 1. The variance decreased 0.006333697, resulting in the overall variance be 1.241473e-0.5. There is still a noticeable seasonal component. I will now take a second difference at lag = 12

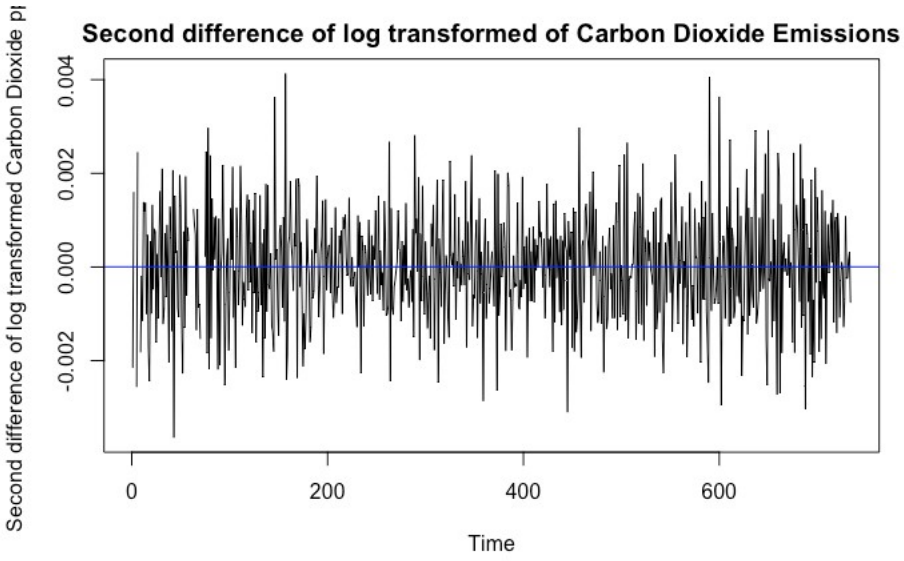

Figure 5: Log Transformation of CO2 Emissions Data with second difference at Lag = 12

I took a second difference at lag = 12 to remove seasonal component. The variance decreased by 1.095733e-05, resulting in the overall variance of 1.457399e-06. The variance has been minimized and the mean has been stabilized, represented by the blue line. The data is now stationary. I will next plot the autocorrelation and partial autocorrelation to find the model candidates.

ACF / PACF ANALYSIS

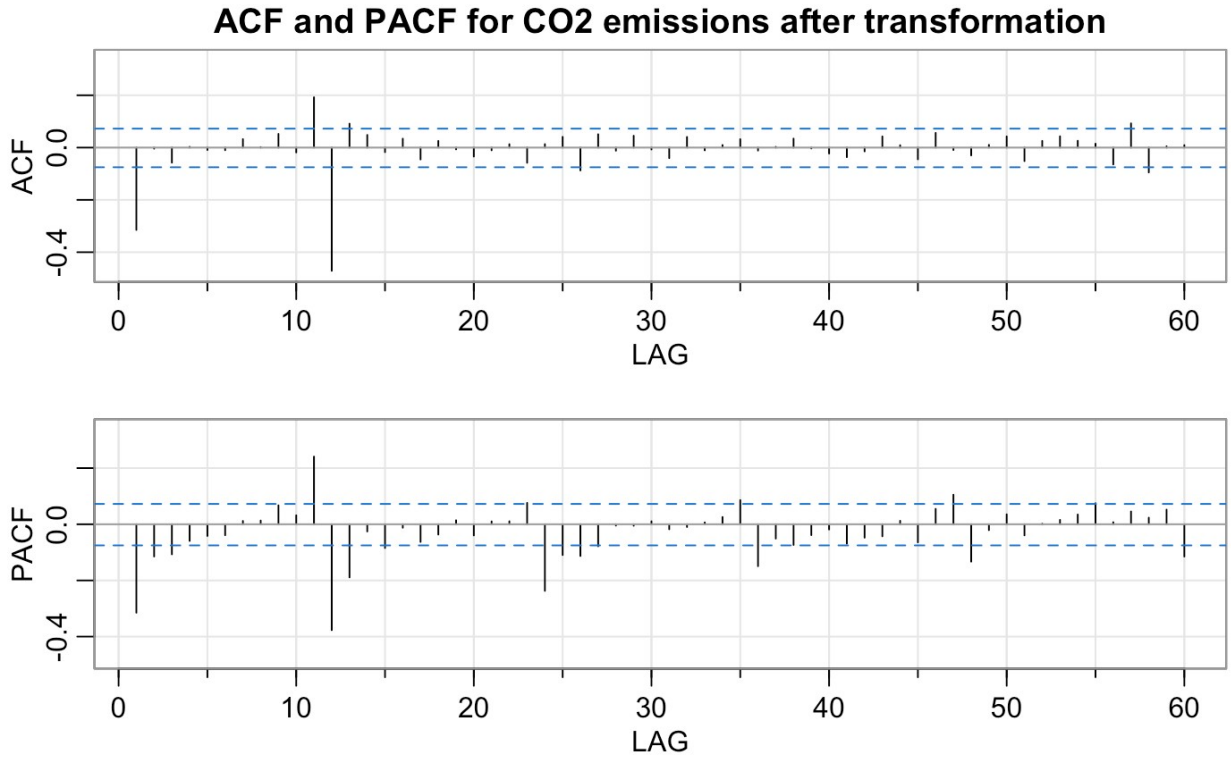

Figure 6: ACF and PACF for Atmospheric Carbon Dioxide emissions after transformations

Looking at the ACF graph there is a peak at lag 11 and 12, then proceeds to cut off. Analyzing from lags 1 through 11, there is also a peak at lag 1. We can assume that Q = 0, 1 and q = 1.

In the PACF graph there is a peak for every multiple of 12 and ends at 60, additionally, there is an exponential decay at these multiples, thus p = 0, 1, 2, 3, 4, 5, 6 however for the sake of modeling, I will limit p to p = 0, 1, 2, 3. Specifically looking from lags 1 through 11, there are peaks at lag 1 through 3 thus P = 0, 1, 2, 3.

Since I took the first difference at lag = 1 and a second difference at lag = 12, D = 1 and d = 1. Moreover, the data is a monthly dataset so S = 12.

Now I will begin model fitting.

MODEL FITTING

After considering all the possible model candidates, I chose the top four by comparing their AICc values and looking at their unit roots. I have limited it down to these four models:

0. SARIMA(0,1,1) X (0,1,1)12
    - AICc: -8198.78
    - $(1 - 0.3912B)(1 - 0.8834B^{12})Z_t = \nabla 12\ \nabla Y_{\#}$
1. SARIMA(3,1,0) X (0,1,1)12
    - AICc: -8198.03
    - $(1 + 0.3557B + 0.1608B^2 + 0.1248B^3)\nabla 12\ \nabla Y_t = (1 - 0.8845B^{12})_{Z_{\#}}$
2. SARIMA(3,1,1) X (0,1,1)12
    - AICc: -8200.16
    - $(1 - 0.2029B - 0.308B^2 + 0.5704B^3)\nabla 12\ \nabla Y_t = (1 - 0.8841B^{12})_{Z_{\#}}$
3. SARIMA(0,1,1) X (1,1,1)12
    - AIC: -8196.96
    - $(1 - 0.0205B^{12})\nabla 12\ \nabla Y_t = (1 - 0.3909B)(1 - 0.8845B^{12})_{Z_{\#}}$

The residuals for these four models all have slight outliers at t = 1 through 20 shown by the zoomed in residual plot in the following model analysis, thus I conducted the diagnostics after extracting the values at 1 through 20. I will now continue to analyze these four models.

MODEL ANALYSIS

Model 0

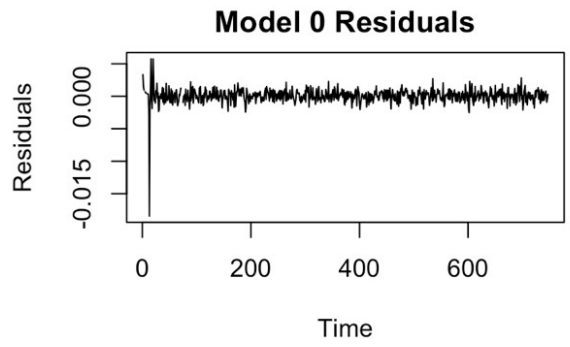

Figure 7: Model 0 Residuals

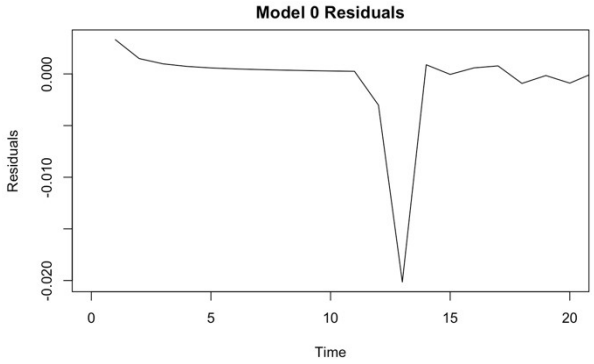

Figure 8: Model 0 Residuals Zoomed In

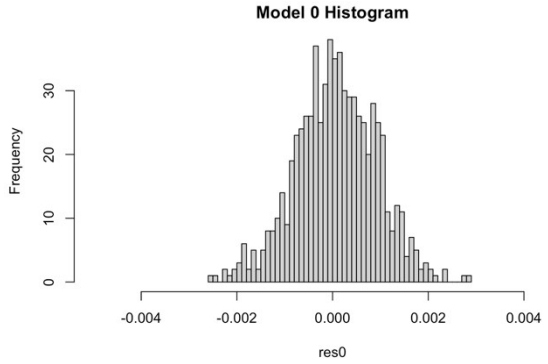

Figure 9: Histogram of Model 0 Residuals

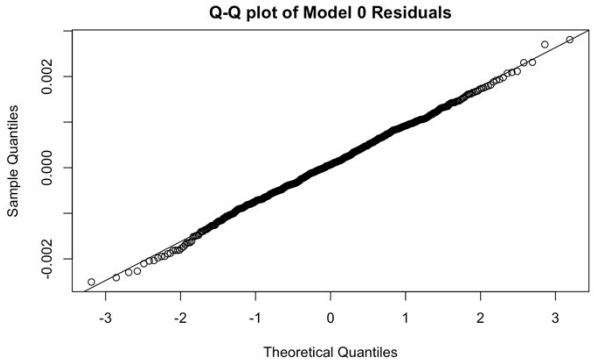

Figure 10: Q-Q plot of Model 0 Residuals

```
        Box-Pierce test

data:  res0
X-squared = 24.078, df = 26, p-value = 0.5715


        Box-Ljung test

data:  res0
X-squared = 24.624, df = 26, p-value = 0.5403


        Box-Ljung test

data:  (res0)^2
X-squared = 22.103, df = 28, p-value = 0.7765


        Shapiro-Wilk normality test

data:  res0
W = 0.9983, p-value = 0.7295
```

Figure 11: Testing for Model 0 Residuals

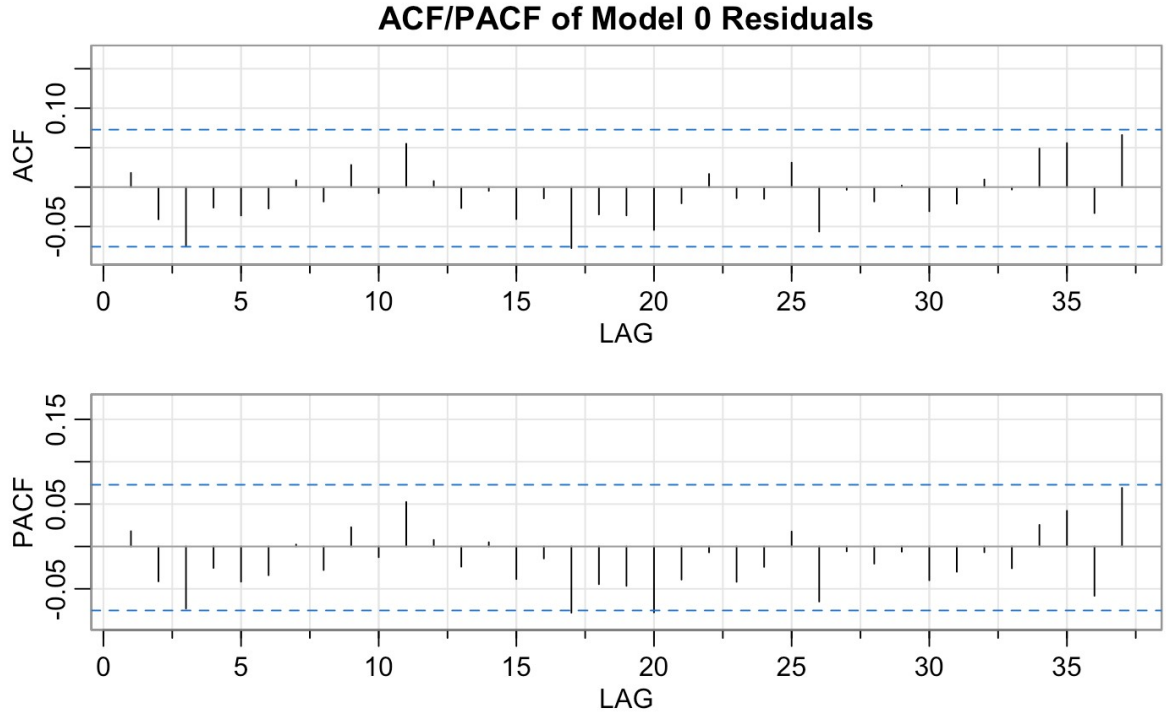

Figure 12: ACF / PACF of Model 0 Residuals

Model 0:
SARIMA(0,1,1) X (0,1,1)12

- AICc: -8198.78
- $(1 - 0.3912B)(1 - 0.8834B^{12})Z_t = \nabla 12\, \nabla Y_{\#}$

In model 0 there is an outlier at t = 1 - 20, there was no trend or seasonal component. The histogram and QQ plot appear to be normally distributed. The model does not reject the null hypothesis for any of the testing shown by the p-value being larger than 0.05. The model passed all the diagnostic checking and the values within the ACF and PACF seem to be inside the confidence interval resembling white noise.

Model 1

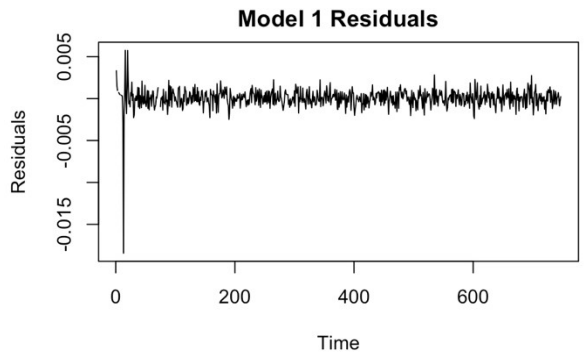

Figure 13: Model 1 Residuals

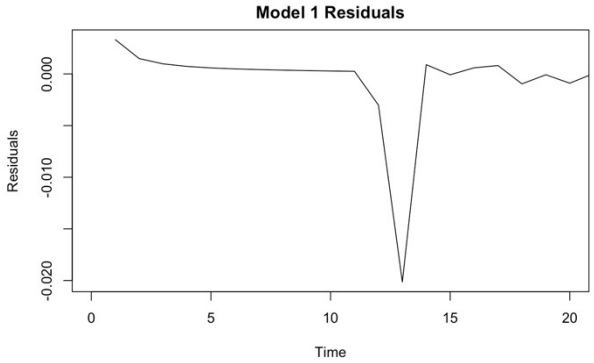

Figure 14: Model 1 Residuals Zoomed In

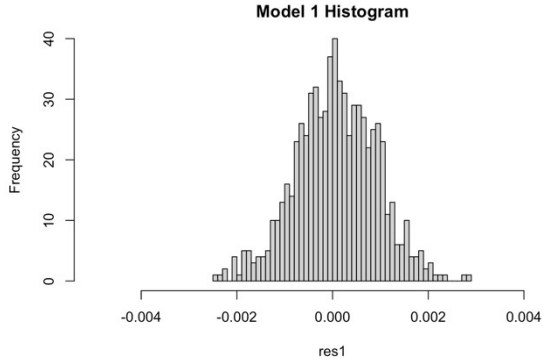

Figure 15: Histogram of Model 1 Residuals

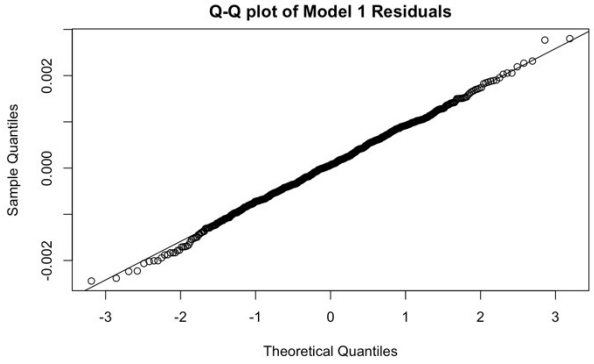

Figure 16: Q-Q plot of Model 1 Residuals

```
        Box-Pierce test

data:  res1
X-squared = 23.34, df = 24, p-value = 0.4998


        Box-Ljung test

data:  res1
X-squared = 23.911, df = 24, p-value = 0.4667


        Box-Ljung test

data:  (res1)^2
X-squared = 22.654, df = 28, p-value = 0.7501


        Shapiro-Wilk normality test

data:  res1
W = 0.99839, p-value = 0.7704
```

Figure 17: Testing for Model 1 Residuals

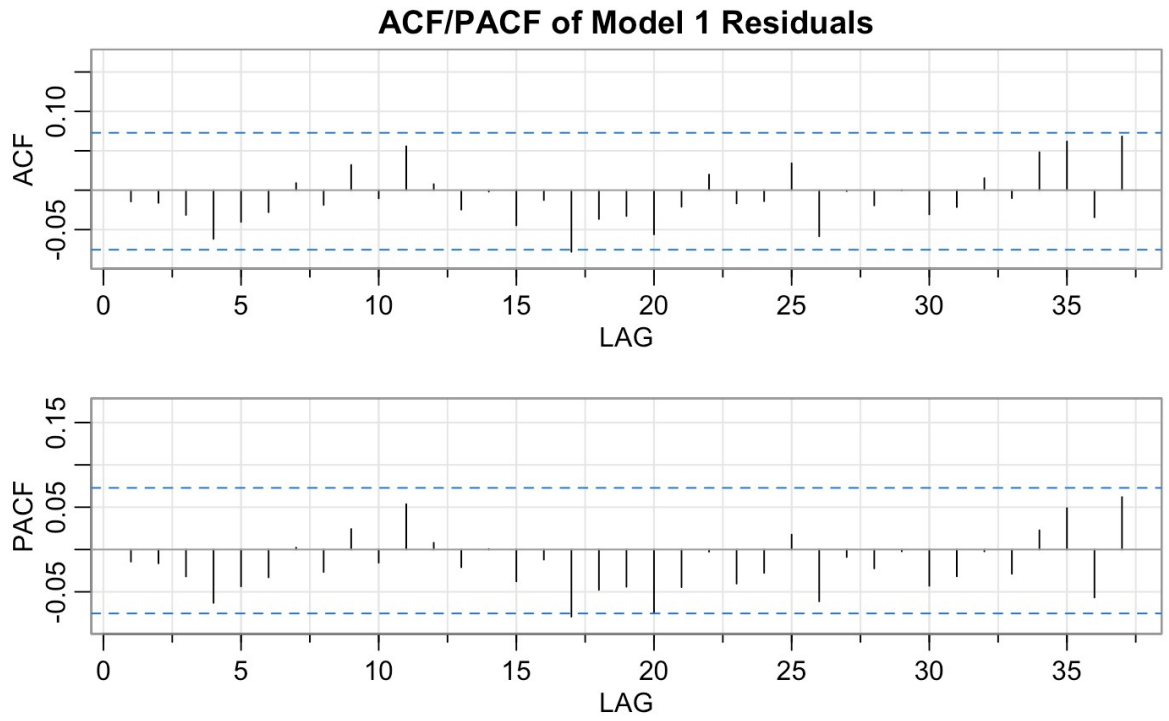

Figure 18: ACF / PACF of Model 1 Residuals

Model 1:
SARIMA(3,1,0) X (0,1,1)12

- AICc: -8198.03
- $(1 + 0.3557B + 0.1608B^2 + 0.1248B^3)\nabla 12\ \nabla Y_t = (1 - 0.8845B^{12})Z_{\#}$

In model 1 there is an outlier at t = 1 - 20, there was no trend or seasonal component. The histogram and QQ plot appear to be normally distributed. The model does not reject the null hypothesis for any of the testing shown by the p-value being larger than 0.05. The model passed all the diagnostic checking and the values within the ACF and PACF seem to be inside the confidence interval resembling white noise.

Model 2

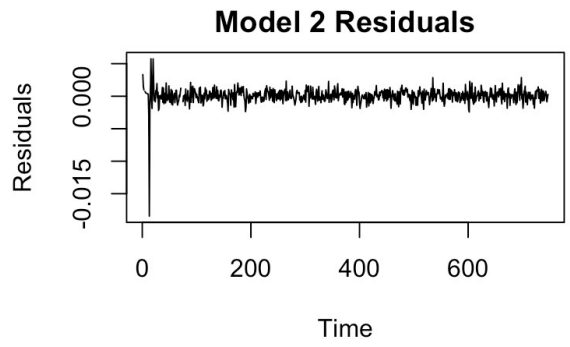

Figure 19: Model 2 Residuals

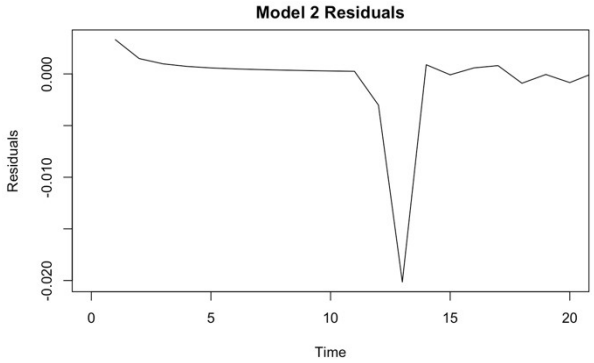

Figure 20: Model 2 Residuals Zoomed In

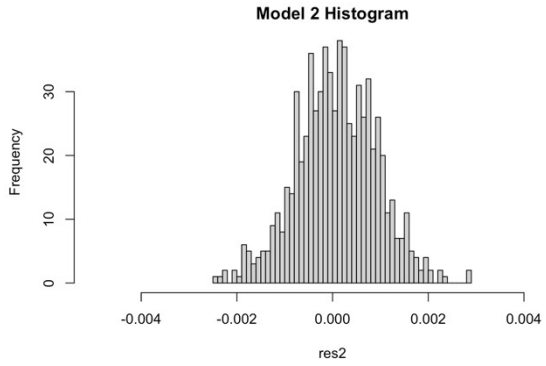

Figure 21: Histogram of Model 2 Residuals

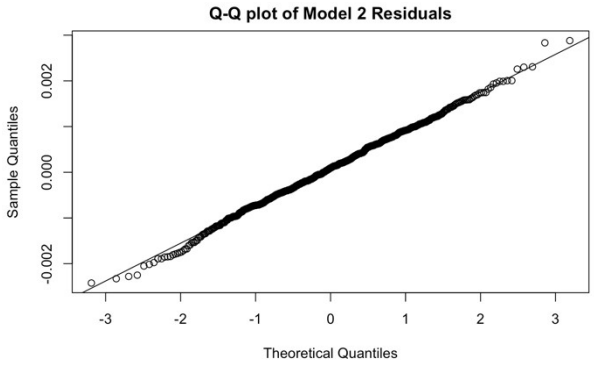

Figure 22: Q-Q plot of Model 2 Residuals

```
        Box-Pierce test

data:  res2
X-squared = 21.893, df = 23, p-value = 0.5267


        Box-Ljung test

data:  res2
X-squared = 22.524, df = 23, p-value = 0.4889


        Box-Ljung test

data:  (res2)^2
X-squared = 23.815, df = 28, p-value = 0.6913


        Shapiro-Wilk normality test

data:  res2
W = 0.99816, p-value = 0.6627
```

Figure 23: Testing for Model 2 Residuals

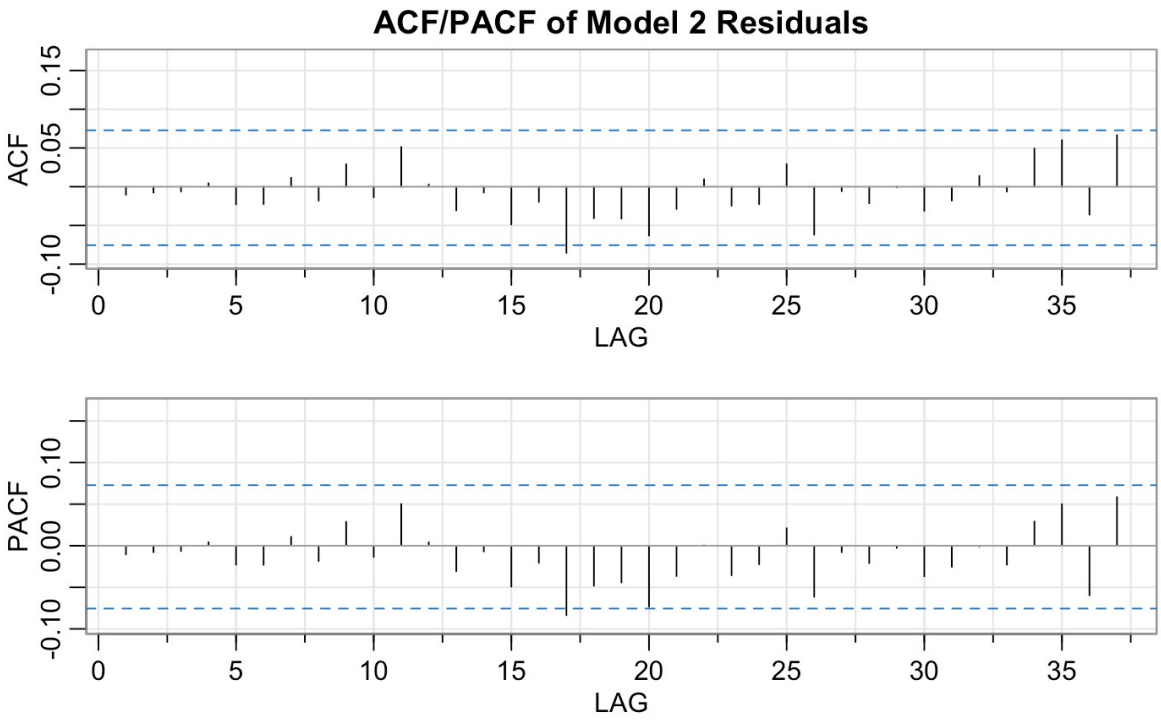

Figure 24: ACF / PACF of Model 2 Residuals

Model 2:
SARIMA(3,1,1) X (0,1,1)12

- AICc: -8200.16
- $(1 - 0.2029B - 0.308B^2 + 0.5704B^3)\nabla 12\ \nabla Y_t = (1 - 0.8841B^{12})Z_{\#}$

In model 2 there is an outlier at t = 1 - 20, there was no trend or seasonal component. The histogram and QQ plot appear to be normally distributed. The model does not reject the null hypothesis for any of the testing shown by the p-value being larger than 0.05. The model passed all the diagnostic checking and the values within the ACF and PACF seem to be inside the confidence interval resembling white noise.

Model 3

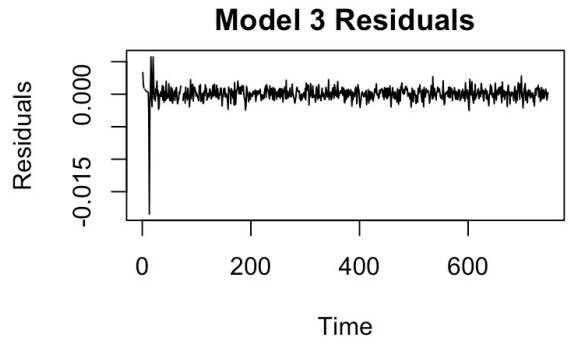

Figure 25: Model 3 Residuals

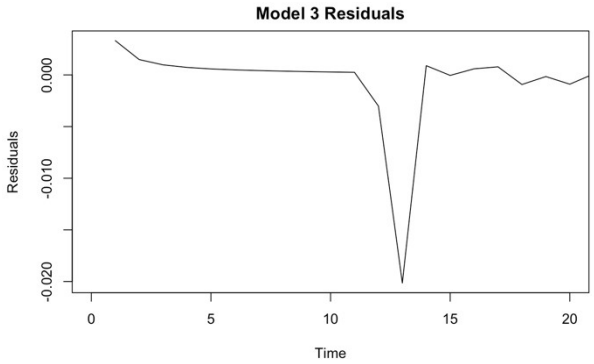

Figure 26: Model 3 Residuals Zoomed In

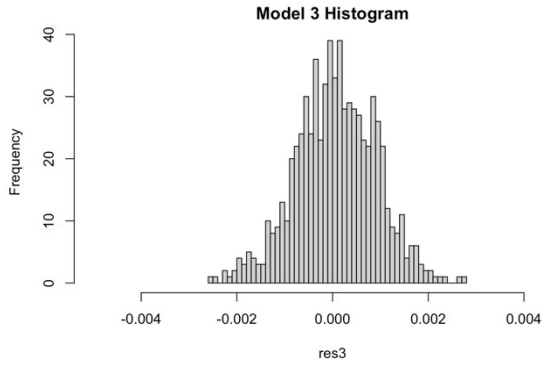

Figure 27: Histogram of Model 3 Residuals

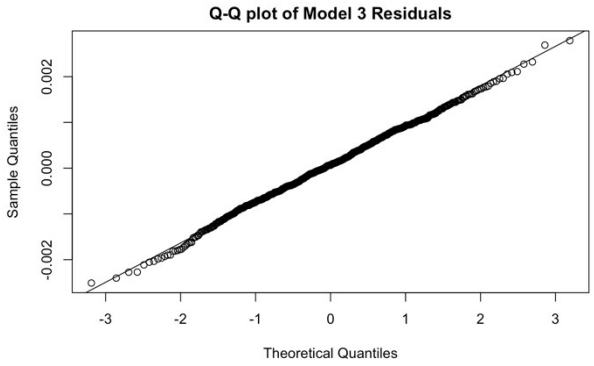

Figure 28: Q-Q plot of Model 3 Residuals

```
        Box-Pierce test

data:  res3
X-squared = 24.12, df = 25, p-value = 0.5125


        Box-Ljung test

data:  res3
X-squared = 24.667, df = 25, p-value = 0.4811


        Box-Ljung test

data:  (res3)^2
X-squared = 21.951, df = 28, p-value = 0.7836


        Shapiro-Wilk normality test

data:  res3
W = 0.99826, p-value = 0.7095
```

Figure 29: Testing for Model 3 Residuals

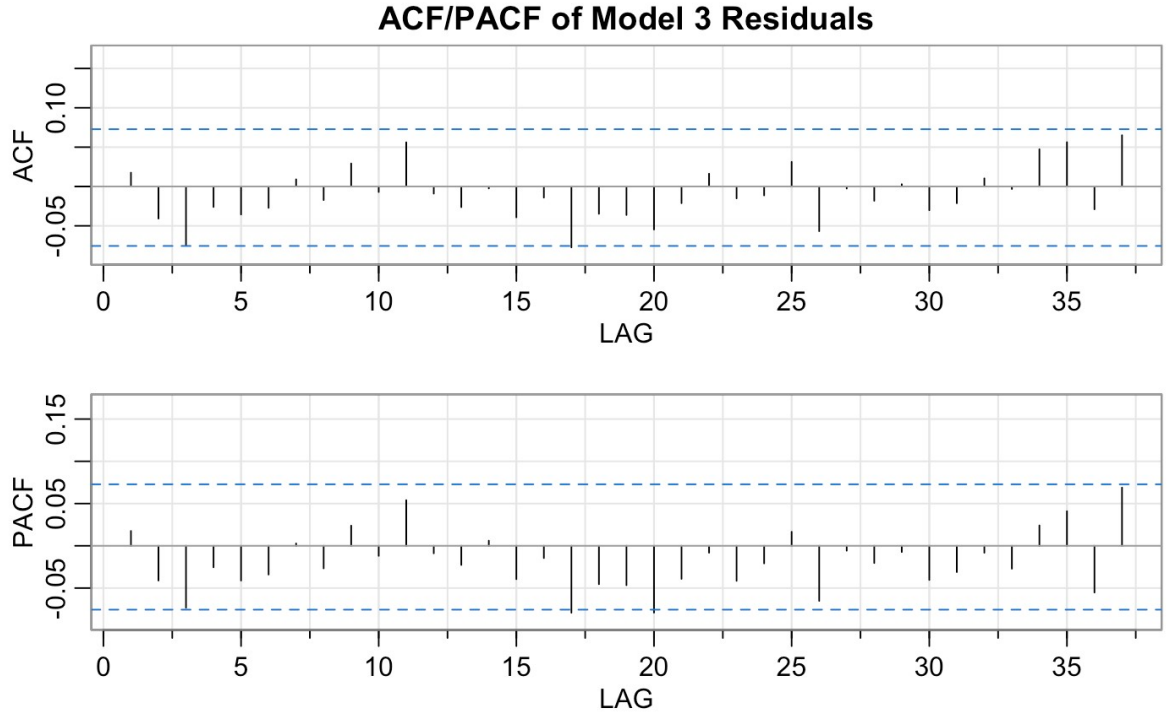

Figure 30: ACF / PACF of Model 3 Residuals Model

3:

SARIMA(0,1,1) X (1,1,1)12

- AIC: -8196.96
- $(1 - 0.0205B^{12})\nabla 12\ \nabla Y_t = (1 - 0.3909B)(1 - 0.8845B^{12})Z_{\#}$

In model 3 there is an outlier at t = 1 - 20, there was no trend or seasonal component. The histogram and QQ plot appear to be normally distributed. The model does not reject the null hypothesis for any of the testing shown by the p-value being larger than 0.05. The model passed all the diagnostic checking and the values within the ACF and PACF seem to be inside the confidence interval resembling white noise.

MODEL CHOICE

All four models have passed diagnostic checking and the residuals for these models are satisfactory. Model 0 had 2 parameters, Model 1 had 4 parameters, Model 2 had 5 parameters, Model 3 had 3 parameters. The AICc values are very similar for all four models differing by +/- 1.5.

From the principle of parsimony, I chose Model 0. I will now proceed to forecast the next 10 values.

Model 0. SARIMA(0,1,1) X (0,1,1)12

- AICc: -8198.78
- $(1 - 0.3912B)(1 - 0.8834B^{12})Z_t = \nabla 12\ \nabla Y_{\#}$

FORECASTING

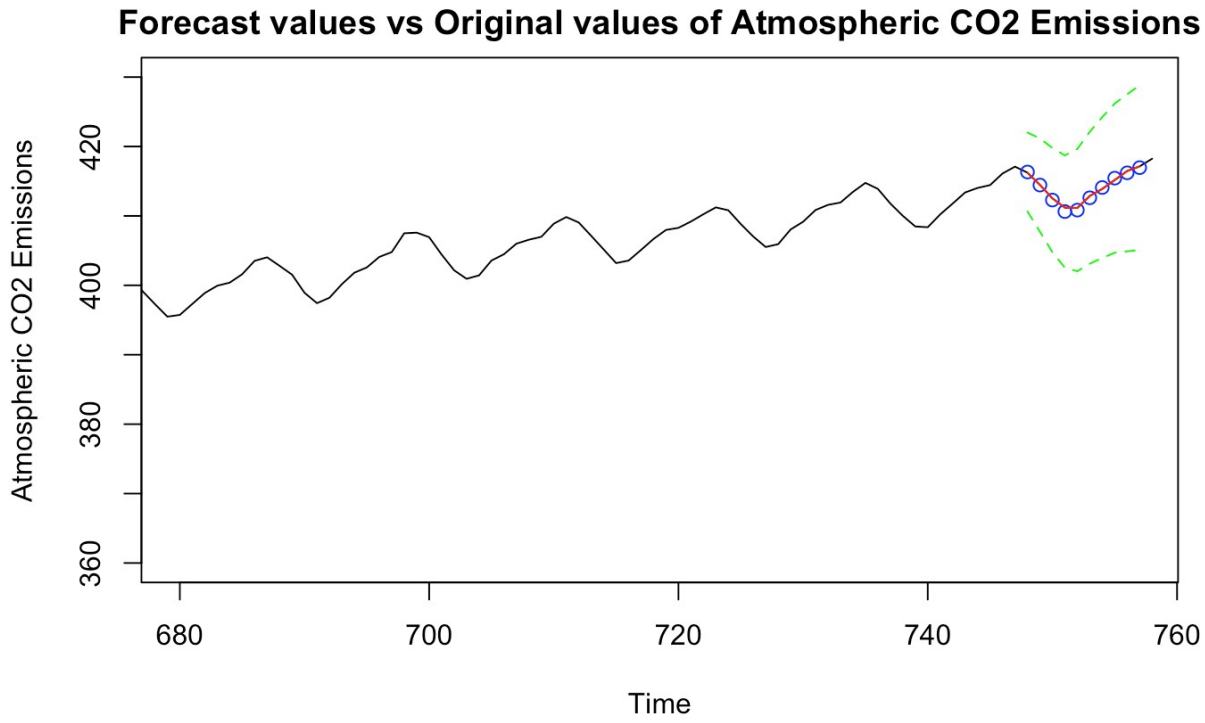

Figure 31: Forecasted Data of Atmospheric Carbon Dioxide Emissions

I forecasted the next 10 points with Model 0. The original Atmospheric Carbon Dioxide Emissions points are represented by the blue circles. The forecasted data is shown by the red line. The confidence interval is shown by the green dashed line. Model 0 accurately forecasts the data.

CONCLUSION

Comparing the four seasonal ARIMA models, all four models passed all diagnostic checking. SARIMA(0,1,1) X (0,1,1)12 was chosen by the Principle of Parsimony.

$$(1 - 0.3912B)(1 - 0.8834B^{12})Z_t = \nabla 12\, \nabla Y_{\#}$$

Using this model, I was able to accurately forecast the next 10 values of the Atmospheric Carbon Dioxide Emissions.

REFERENCES

https://www.kaggle.com/gargilohia1/monthly-co2-from-co2-emissions-mauna-loa

APPENDIX

```
co2emission <- read.csv("monthly_in_situ_co2_mlo.csv",
                        skip =59, header = F)
co2 <- co2emission$V5
co2 <- ifelse(co2 == -99.99, NA, co2)
co2ts <- as.ts(co2, start=c(1958,3), frequency=12)
co2ts <- co2ts[0:747]
plot.ts(co2ts, ylab = "Original Atmospheric Carbon Dioxide (PPM)",
        main = "Atmospheric Carbon Dioxide vs. Time")
plot.ts(co2ts, ylab = "Carbon Dioxide in parts per million",
        main = "Time Series of Carbon Dioxide Emissions")
abline(h=mean(co2ts, na.rm = T), col = "blue")
nt = length(co2ts)
fit<- lm(co2ts~as.numeric(1:nt)); abline(fit, col="red")

# TRANSFORMATION
t <- as.numeric(1:length(co2ts))
fit <- lm(co2ts~t)
bctransform <- boxcox(co2ts~t, lambda = seq(0,1.5))
transformation = bctransform$x[which.max(bctransform$y)]
transformation

co2log <- log(co2ts)
plot.ts(co2log, ylab = "Carbon Dioxide in parts per million",
        main = "CO2 log transformation")

co2log1 <- diff(co2log,1)
plot.ts(co2log1, ylab = "First difference of log transformed Carbon Dioxide in ppm",
        main = "CO2 log transformation with first lag differenced")
varco2log <- var(co2log, na.rm = T)
show(varco2log)
varco2log1 <- var(co2log1, na.rm =T)
show(varco2log1)
vardiff<-varco2log-varco2log1
show(vardiff)

co2log12 <- diff(co2log1,12)
plot.ts(co2log12, ylab = "Second difference of log transformed Carbon Dioxide ppm",
        main = "Second difference of log transformed of Carbon Dioxide Emissions")
abline(h=mean(co2log12, na.rm = T), col = "blue")
varco2log12 <- var(co2log12, na.rm =T)
show(varco2log12)
vardiff2<-varco2log1-varco2log12
```

```r
show(vardiff2)

# ACF / PACF
acf2(co2log12, max.lag = 60, plot = TRUE, main = "ACF and PACF for CO2 emissions after tra
kpss.test(co2log12)

# MODEL FITTING
model0 <- arima(co2log, order = c(0,1,1), seasonal = list(order = c(0,1,1), period = 12))
model1 <- Arima(co2log, order = c(3,1,0), seasonal = list(order = c(0,1,1), period = 12))
model2 <- Arima(co2log, order = c(3,1,1), seasonal = list(order = c(0,1,1), period = 12))
model3 <- Arima(co2log, order = c(0,1,1), seasonal = list(order = c(1,1,1), period = 12))

show(model0)
show(model1)
show(model2)
show(model3)

polyroot(c(1,-0.8834)) #SMA for model 0
polyroot(c(1, -0.8840)) #SMA for model 1
polyroot(c(1, -0.8841)) #SMA for model 2
polyroot(c(1, 0.0205, -0.8877)) #SAR and SMA for model 3
polyroot(c(1, 0.0184, -0.8877)) #SAR and SMA for model 4
polyroot(c(1,-0.8845)) #SMA for model 5

# MODEL ANALYSIS
# MODEL 0
res00<-residuals(model0)
res0<-res00[21:747]
plot(res00, main="Model 0 Residuals",
     ylab = "Residuals")
plot(res00, main="Model 0 Residuals",
     ylab = "Residuals", xlim = c(0,20))

Box.test(res0, lag = 28, type = "Box-Pierce", fitdf=2)
Box.test(res0, lag = 28, type = "Ljung-Box", fitdf=2)
Box.test((res0)^2, lag=28, type="Ljung-Box")
shapiro.test(res0)

hist(res0, xlim = c(-0.005,0.005),
     main = "Model 0 Histogram", breaks = 60)
qqnorm(res0,  main = "Q-Q plot of Model 0 Residuals")
qqline(res0)
acf2(res0, na.action=na.pass,main = "ACF/PACF of Model 0 Residuals")

# MODEL 1
res01<-residuals(model1)
res1<-res01[21:747]
plot(res01, main="Model 1 Residuals",
     ylab = "Residuals")
plot(res01, main="Model 1 Residuals",
     ylab = "Residuals", xlim = c(0,20))

Box.test(res1, lag = 28, type = "Box-Pierce", fitdf=4)
```

```r
Box.test(res1, lag = 28, type = "Ljung-Box", fitdf=4)
Box.test((res1)^2, lag=28, type="Ljung-Box")
shapiro.test(res1)

hist(res1, xlim = c(-0.005,0.005),
     main = "Model 1 Histogram", breaks = 60)
qqnorm(res1,  main = "Q-Q plot of Model 1 Residuals")
qqline(res1)
acf2(res1, na.action=na.pass,main = "ACF/PACF of Model 1 Residuals")

# MODEL 2
res02<-residuals(model2)
res2<-res02[21:747]
plot(res02, main="Model 2 Residuals",
     ylab = "Residuals")
plot(res02, main="Model 2 Residuals",
     ylab = "Residuals", xlim = c(0,20))

Box.test(res2, lag = 28, type = "Box-Pierce", fitdf=5)
Box.test(res2, lag = 28, type = "Ljung-Box", fitdf=5)
Box.test((res2)^2, lag=28, type="Ljung-Box")
shapiro.test(res2)


hist(res2, xlim = c(-0.005,0.005),
     main = "Model 2 Histogram", breaks = 60)
qqnorm(res2,  main = "Q-Q plot of Model 2 Residuals")
qqline(res2)
acf2(res2, na.action=na.pass,main = "ACF/PACF of Model 2 Residuals")

# MODEL 3
res03<-residuals(model3)
res3<-res03[21:747]
plot(res03, main="Model 3 Residuals",
     ylab = "Residuals")
plot(res03, main="Model 3 Residuals",
     ylab = "Residuals", xlim = c(0,20))

Box.test(res3, lag = 28, type = "Box-Pierce", fitdf=3)
Box.test(res3, lag = 28, type = "Ljung-Box", fitdf=3)
Box.test((res3)^2, lag=28, type="Ljung-Box")
shapiro.test(res3)
hist(res3, xlim = c(-0.005,0.005),
     main = "Model 3 Histogram", breaks = 60)
qqnorm(res3,  main = "Q-Q plot of Model 3 Residuals")
qqline(res3)
acf2(res3, na.action=na.pass,main = "ACF/PACF of Model 3 Residuals")

#FORECAST
prediction <- predict(model1, n.ahead = 10)
pred.orig <- exp(prediction$pred)
stderror <- exp(prediction$pred)*prediction$pred*prediction$se
bound = (length(co2ts)+1):(length(co2ts)+10)
```

```r
predval <- co2[748:757]
upperbound = pred.orig+1.96*stderror
lowerbound = pred.orig-1.96*stderror

plot.ts(co2, xlim = c(680,length(co2ts)+10), ylim = c(360,430),
        ylab = "Atmospheric CO2 Emissions",
        main = "Forecast values vs Original values of Atmospheric CO2 Emissions")
points(bound,pred.orig, col="blue")
lines(bound,predval, col = "red")
lines(bound, upperbound,lty=2, col="green")
lines(bound,lowerbound,lty=2, col="green")
```