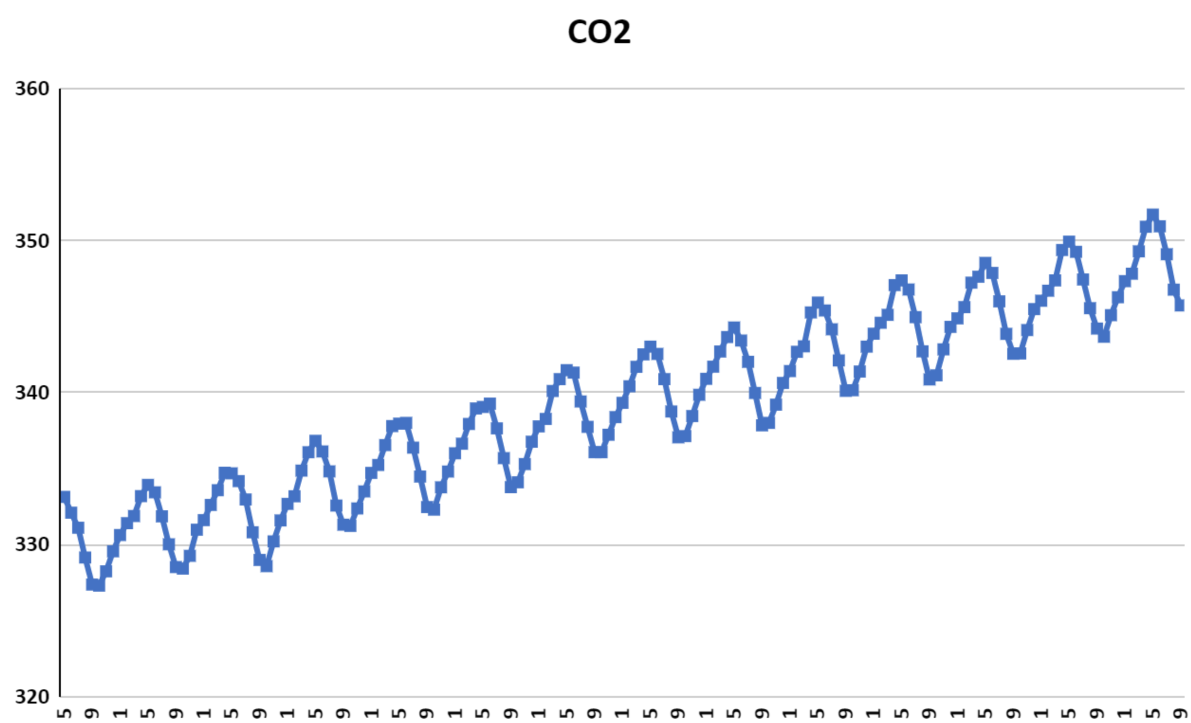
Tuan Dang
05/13/2021

Project 4: Time Series Forecasting on Monthly CO_2 Levels

Part 1:



This time series has a clear positive trend and seasonality. It looks like the CO_2 concentration increases from September/October to May then decreases from June to September.

Part 2:

MAE Naïve	MAE Average	MAE 2-MA
1.112125	5.200621019	1.608050314

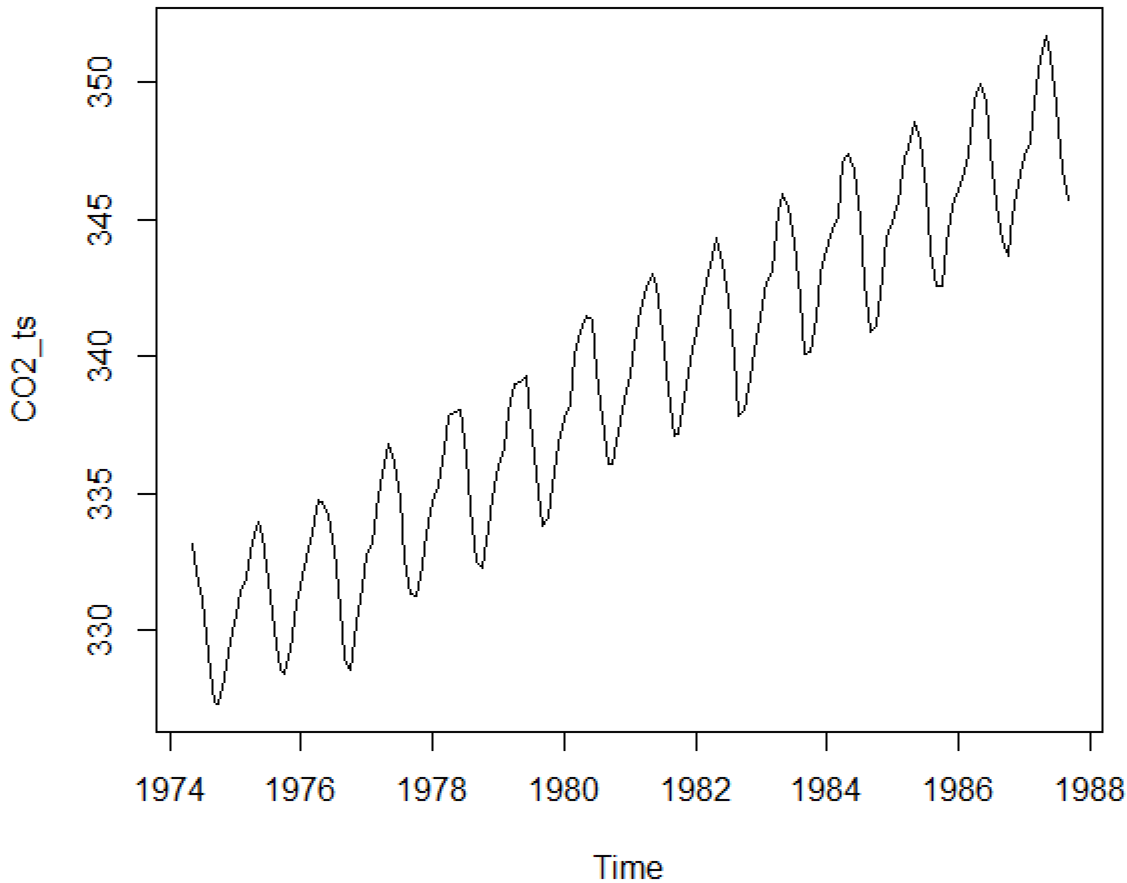
(Calculated on Excel)

We want the MAE to be as small as possible. Based on this criterion, the naive forecasting method gives the most accurate predictions, since its MAE is the lowest, followed by the 2-MA then the average method. This makes sense since seasonality is a big factor, so the predecessor values are closer to the observed value, while the average value presents too much smoothing.

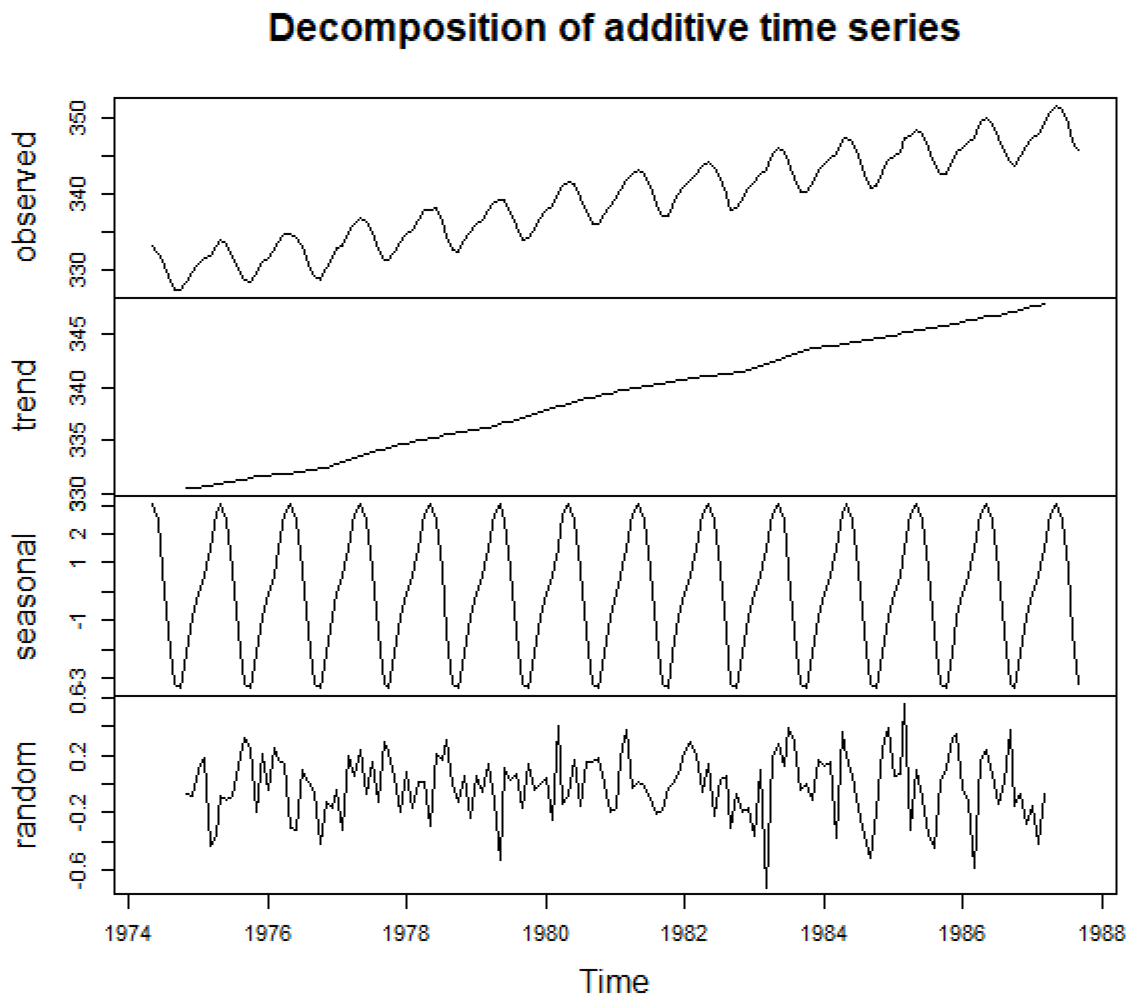
Part 3:

R Code:

```
co2 <- read.csv("MonthlyCO2Concentrations.csv")
CO2_ts <- ts(co2$CO2, frequency = 12, start = c(1974,5))
plot.ts(CO2_ts)
```




```
decompCO2 <- decompose(CO2_ts)
decompCO2$trend
decompCO2$seasonal
decompCO2$random
plot(decompCO2)
```



The trend component is consistently increasing. The seasonal component is very constant during each period, so much so that its values in each month are actually identical. As expected, the general pattern is that it increases from October to May then decreases. The random component looks very stationary, where the mean and variance seem constant over time.

```
times <- c(1:161)
model1 <- lm(CO2 ~ times, co2)
summary(model1)
```

Call:

```
lm(formula = CO2 ~ times, data = co2)

Residuals:
   Min          1Q           Median        3Q            Max
-4.392       -1.866        0.199         2.190         3.677

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.293e+02  3.462e-01  951.40   <2e-16 ***
times       1.210e-01  3.707e-03   32.64   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’  0.1 ‘ ’ 1

Residual standard error: 2.186 on 159 degrees of freedom
Multiple R-squared:  0.8701,     Adjusted R-squared:  0.8693
F-statistic:  1065 on 1 and 159 DF,  p-value: < 2.2e-16
```

The model seems to be good despite not accounting for seasonality. The p-values and R-squared values look good. The coefficient of times is positive, so there is an increase over time, but the change does not seem significant, since it only increases by around 0.1 each month. However, this change becomes more significant when measuring the CO2 concentration over a few years.

```
month <- factor(c(c(5,6,7,8,9,10,11,12), rep(c(1,2,3,4,5,6,7,8,9,10,11,12),12),
c(1,2,3,4,5,6,7,8,9)))
model2 <- lm(CO2 ~ times + month, co2)
summary(model2)


Call:
lm(formula = CO2 ~ times + month, data = co2)

Residuals:
```

```
     Min       1Q   Median       3Q      Max
-1.06869 -0.25190  0.01348  0.27987  0.82000

Coefficients:
             Estimate Std. Error  t value Pr(>|t|)
(Intercept)  3.292e+02  1.215e-01 2709.911  < 2e-16 ***
times        1.207e-01  6.661e-04  181.146  < 2e-16 ***
month2       6.247e-01  1.539e-01    4.058 7.98e-05 ***
month3       1.619e+00  1.540e-01   10.519  < 2e-16 ***
month4       2.783e+00  1.540e-01   18.074  < 2e-16 ***
month5       3.217e+00  1.512e-01   21.280  < 2e-16 ***
month6       2.583e+00  1.512e-01   17.085  < 2e-16 ***
month7       8.979e-01  1.512e-01    5.940 1.96e-08 ***
month8      -1.264e+00  1.512e-01   -8.362 4.18e-14 ***
month9      -3.051e+00  1.512e-01  -20.178  < 2e-16 ***
month10     -3.217e+00  1.540e-01  -20.897  < 2e-16 ***
month11     -2.046e+00  1.540e-01  -13.287  < 2e-16 ***
month12     -8.032e-01  1.539e-01   -5.217 6.03e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’  0.1 ‘ ’ 1

Residual standard error: 0.3925 on 148 degrees of freedom
Multiple R-squared:  0.9961,     Adjusted R-squared:  0.9958
F-statistic:  3153 on 12 and 148 DF,  p-value: < 2.2e-16
```

As expected, this model is much better than the previous one, with the R-squared values almost perfect.

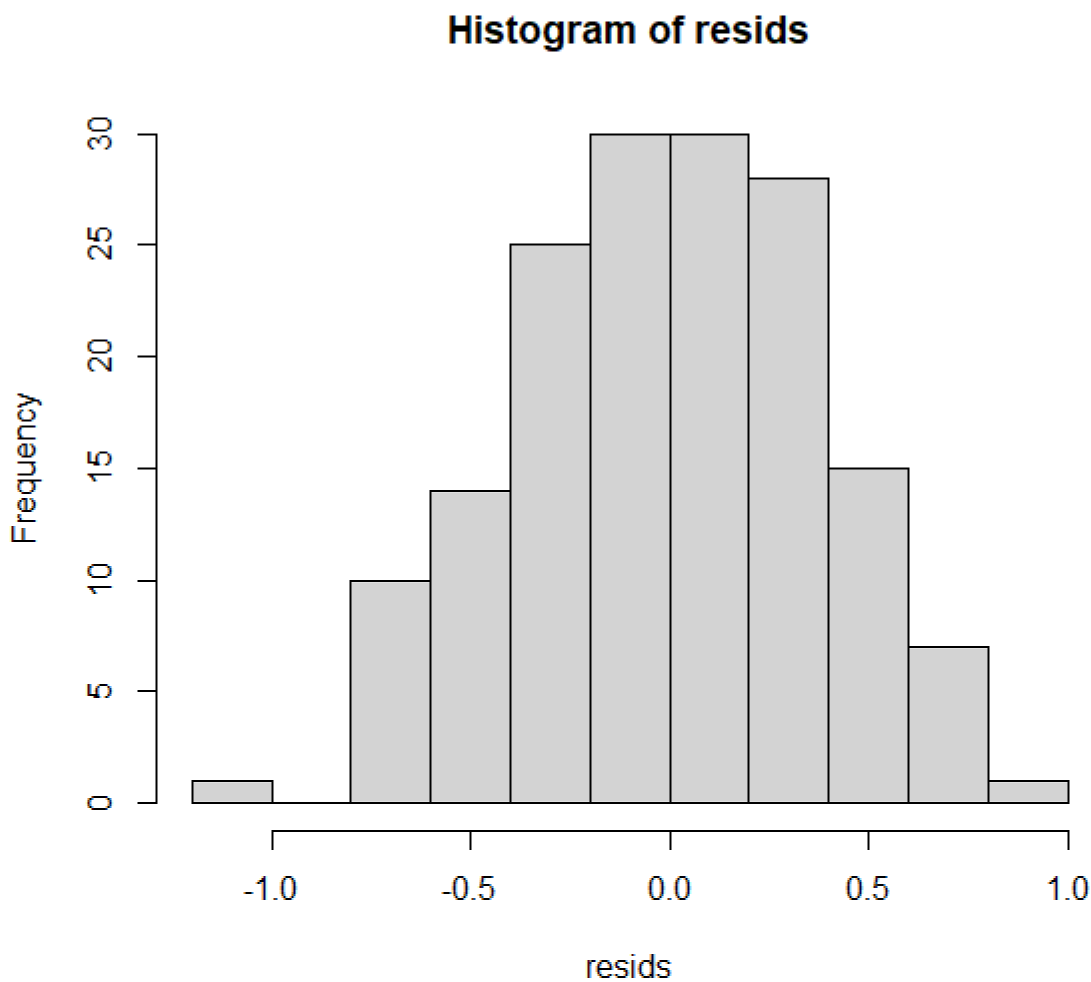
Using this model to predict CO2 concentration for December 1987, which will be the 164th month:

329.2 + 0.1207*164 - 0.8032 = 348.1916

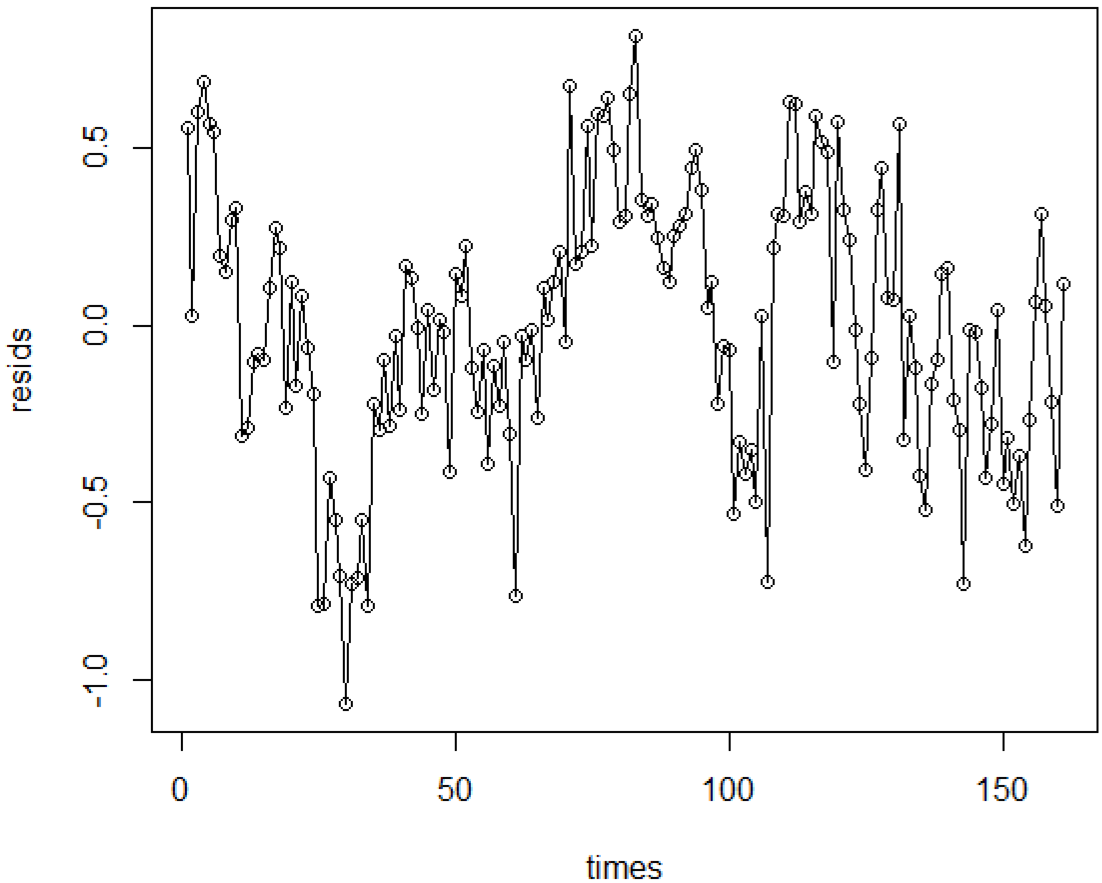
Part 4:

resids <- model2$residuals

hist(resids)



plot(times, resids, type = "o")

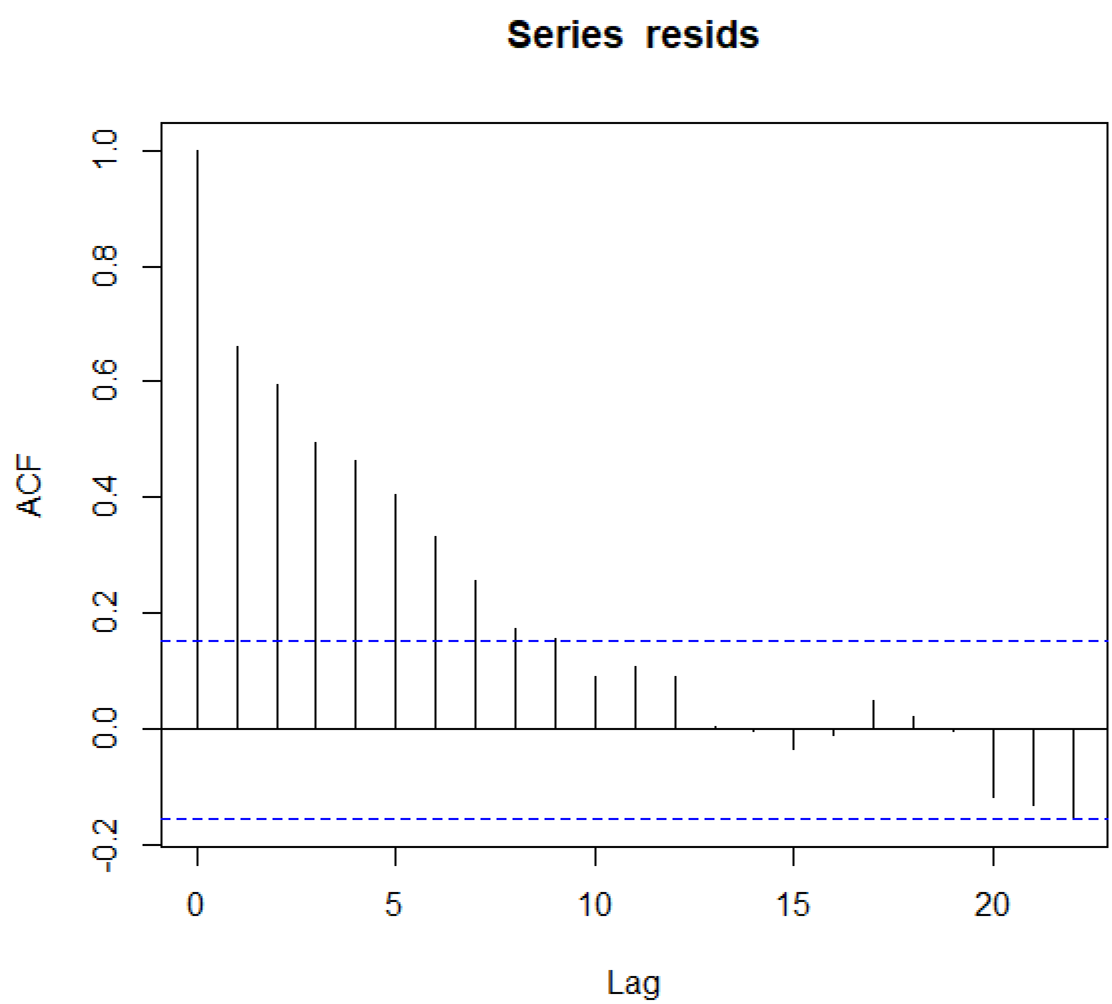


adf.test(resids)

Augmented Dickey-Fuller Test

data: resids
Dickey-Fuller = -3.1569, Lag order = 5, p-value = 0.09752
alternative hypothesis: stationary

acf(resids)



Based on the resids-times graph, the residuals do appear to be stationary, with a mean of 0. However, this is not supported by the Dicky-Fuller test with a 0.5 significance level. The p-value is 0.09752 so we fail to reject the null hypothesis, meaning that this time series could be a random walk. A further look at the autocorrelation function of this series supports this idea. The residuals do not seem to be just white noise.