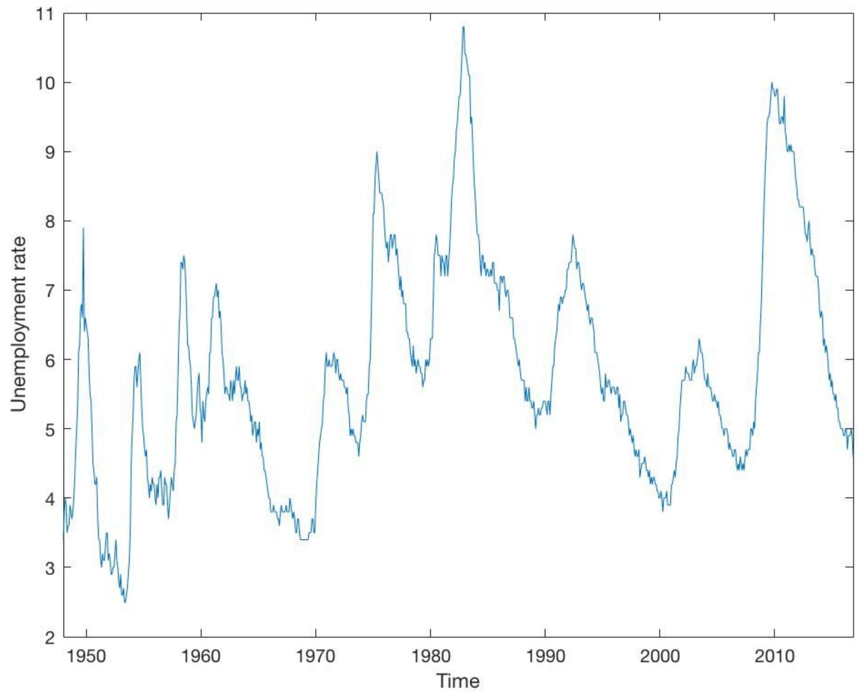
MgtF 405: Business Forecasting
Thomas Murphey A53044035, Anton Prokopyev A53105867
Assignment 1

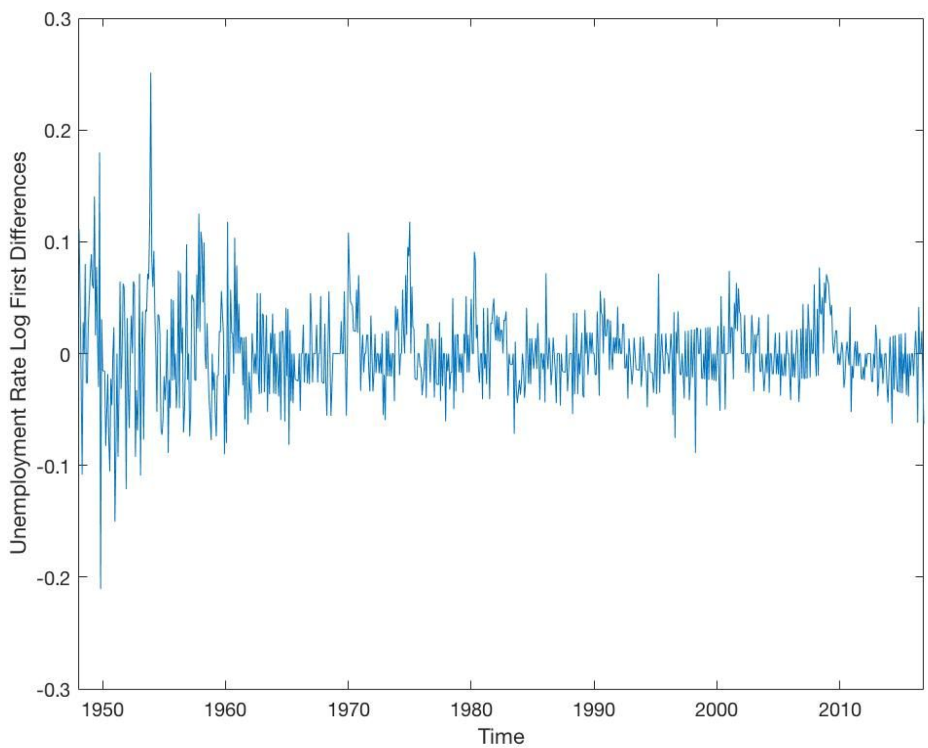
Part I. Constructing prediction models for different variables

Unemployment Rate (UNRATE)

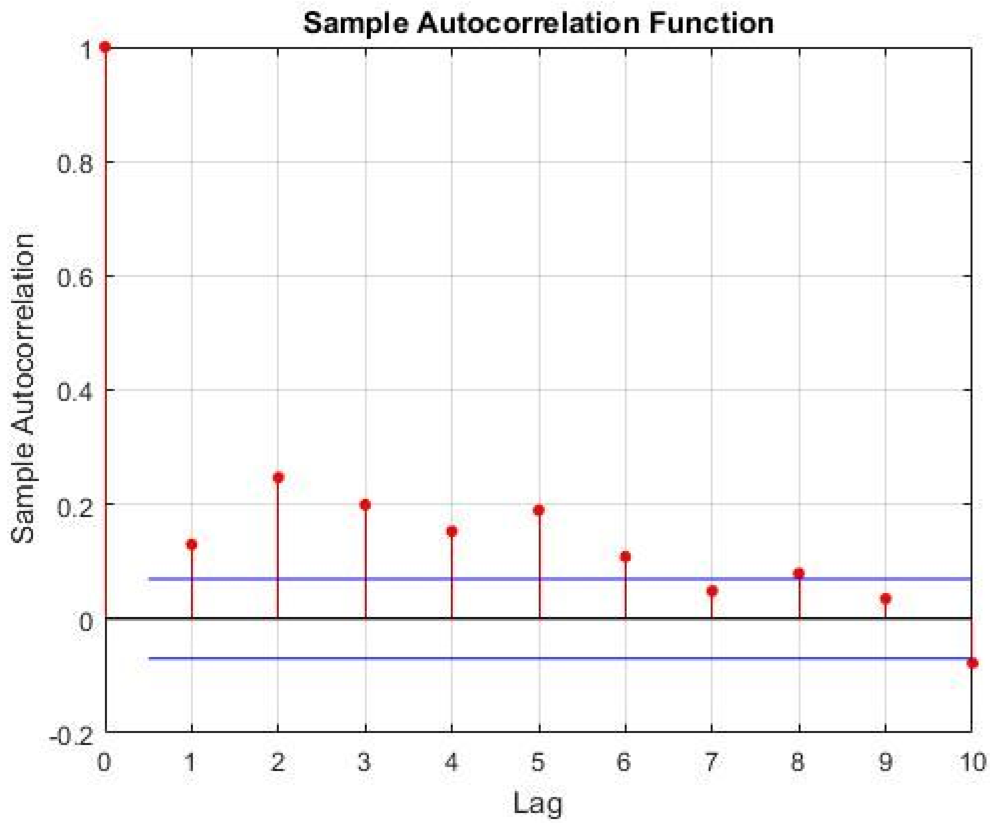
1.



Raw unemployment rate appears to have a slight upward trend. An Augmented Dickey-Fuller test results in an h=0 result with a p-value of 0.5162, failing to reject the null hypothesis of a unit root, and suggesting this series is not stationary. We therefore take the unemployment rate log first differences, $\Delta \log(y_t) = \log(y_t) - \log(y_{t-1})$, and we plot the transformed series below.



2. See the first 10 autocorrelations below.



Is the variable persistent? The graph above implies presence of a long-term influence of a shock, i.e. that has a discernible influence on the log difference in unemployment level for 6 periods. The first ten autocorrelations are:

Lag 1: 0.1297
Lag 2: 0.2469
Lag 3: 0.1992
Lag 4: 0.1527
Lag 5: 0.1896
Lag 6: 0.1085
Lag 7: 0.0484
Lag 8: 0.0791
Lag 9: 0.0348
Lag 10: -0.0778

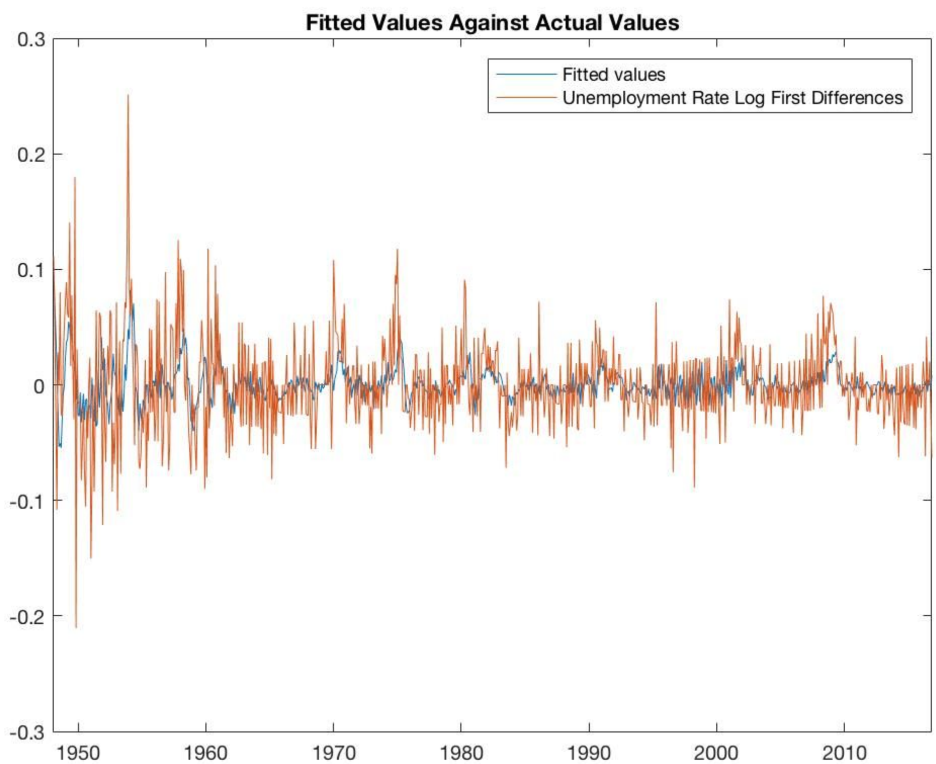
Is the serial correlation statistically significant? A Ljung-Box Q Test reveals that serial correlation is statistically significant at a p-value of 1.8720e-04 or 0.0001872 at 99% confidence level.

3.

```
ARIMA(4,0,4) Model:
--------------------
Conditional Probability Distribution: Gaussian

                                  Standard          t
 Parameter        Value            Error        Statistic
-----------    -----------    ------------    -----------
 Constant     0.000100393    0.000819589        0.122492
    AR{1}         0.54597      0.0480525         11.3619
    AR{2}        0.100583      0.0484218         2.07723
    AR{3}        0.667351      0.0454705         14.6766
    AR{4}       -0.680824       0.033373        -20.4004
    MA{1}       -0.551802      0.0443503        -12.4419
    MA{2}       0.0741709      0.0467022         1.58817
    MA{3}        -0.64827      0.0408566         -15.867
    MA{4}        0.740536      0.0329171          22.497
 Variance      0.00125317    4.18557e-05         29.9403
```

The best fit model is ARMA(4,4) based on t-Values for AR{4} and MA{4} having the largest magnitudes and thus furthest from zero in either direction. The AICBIC function confirms this, as the information criteria are minimized for AR(4) and MA(4). A Ljung-Box Q Test results in h=0 and a p-value of 0.3883, meaning that we cannot reject the claim that there is no auto-correlation, which suggests that the residuals are not serially correlated.

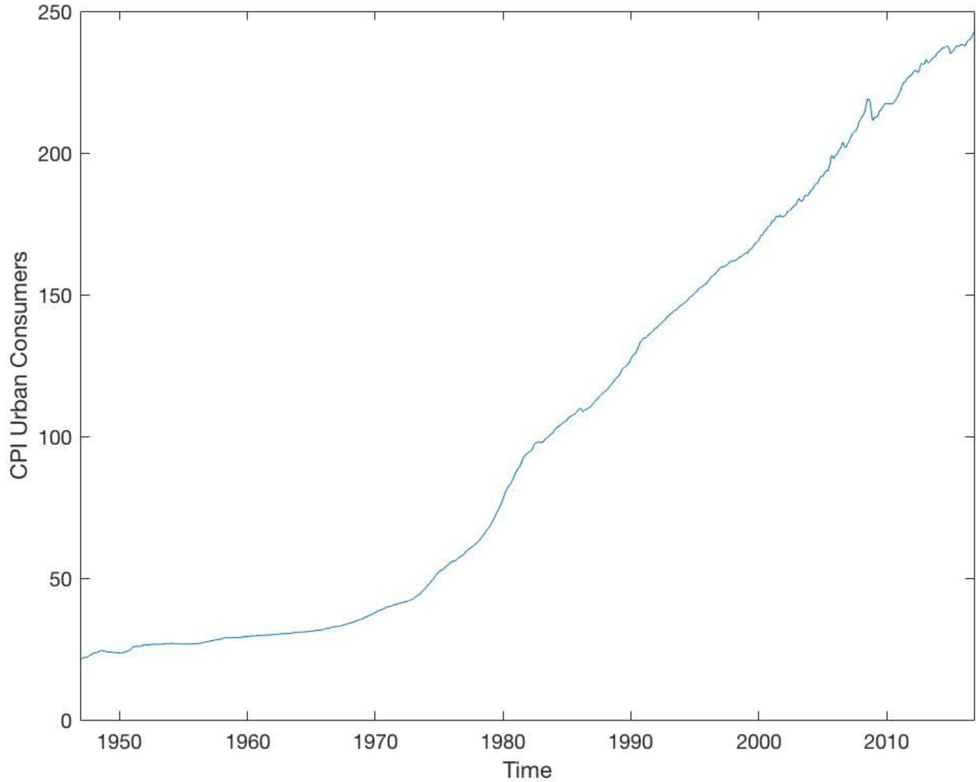
4.



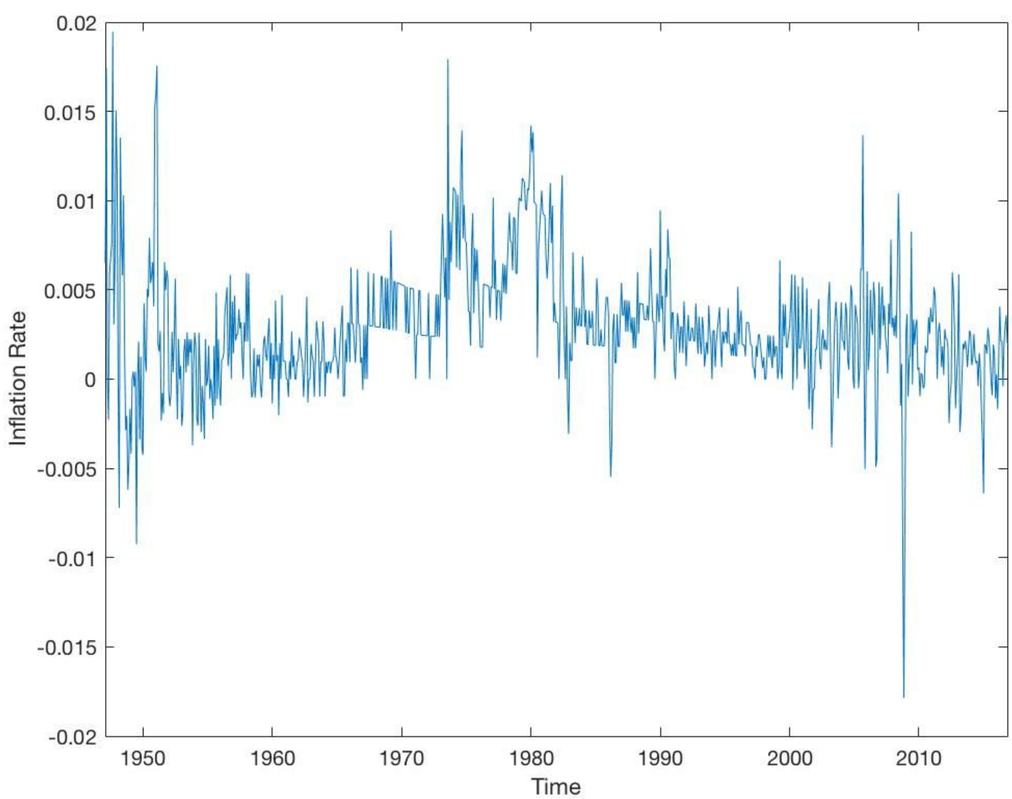
From the first glance, our forecasting model demonstrates a good fit as it follows the realized values closely. There are sporadic overestimations, however the model still holds well on to the general pattern. Moreover, we fail to reject null hypothesis for the Ljung-Box Q-test for residual autocorrelation, therefore we do not have serial correlation (persistence) in our error term, a sign of a good forecasting model.

US consumer prices (CPIAUCSL)

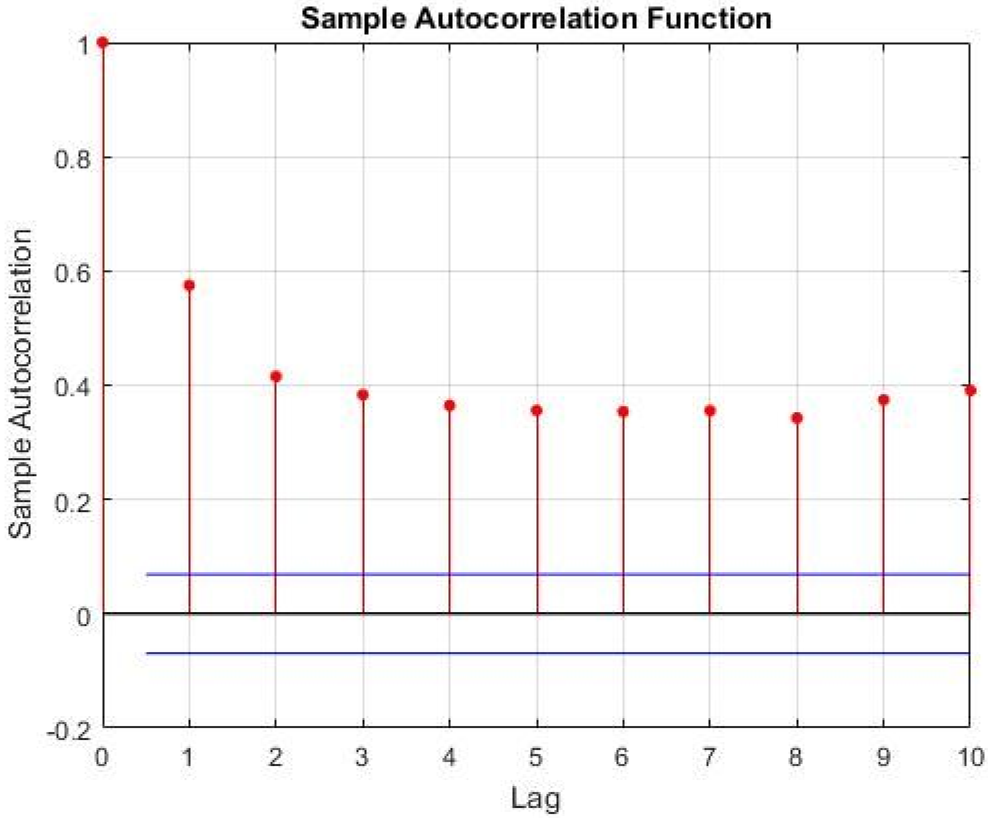
1.



Consumer Price Index appears to have a steep upward trend over time. An Augmented Dickey-Fuller test results in an h=0 result with a p-value of 0.999, failing to reject the null hypothesis of a unit root, and suggesting this series is not stationary. We therefore take the log-first difference, $\Delta \log(y_t) = \log(y_t) - \log(y_{t-1})$, which is the inflation rate, and we plot the transformed series below.



2. See the first 10 autocorrelations below.



The values of the first ten autocorrelations are:

Lag 1: 0.575
Lag 2: 0.4154
Lag 3: 0.3832
Lag 4: 0.3647
Lag 5: 0.3560
Lag 6: 0.3539
Lag 7: 0.3555
Lag 8: 0.3428
Lag 9: 0.3745
Lag 10: 0.3908

Is the variable persistent? The graph above implies a strongly persistent, long-term influence of a shock, as it remains in the series and does not go away..

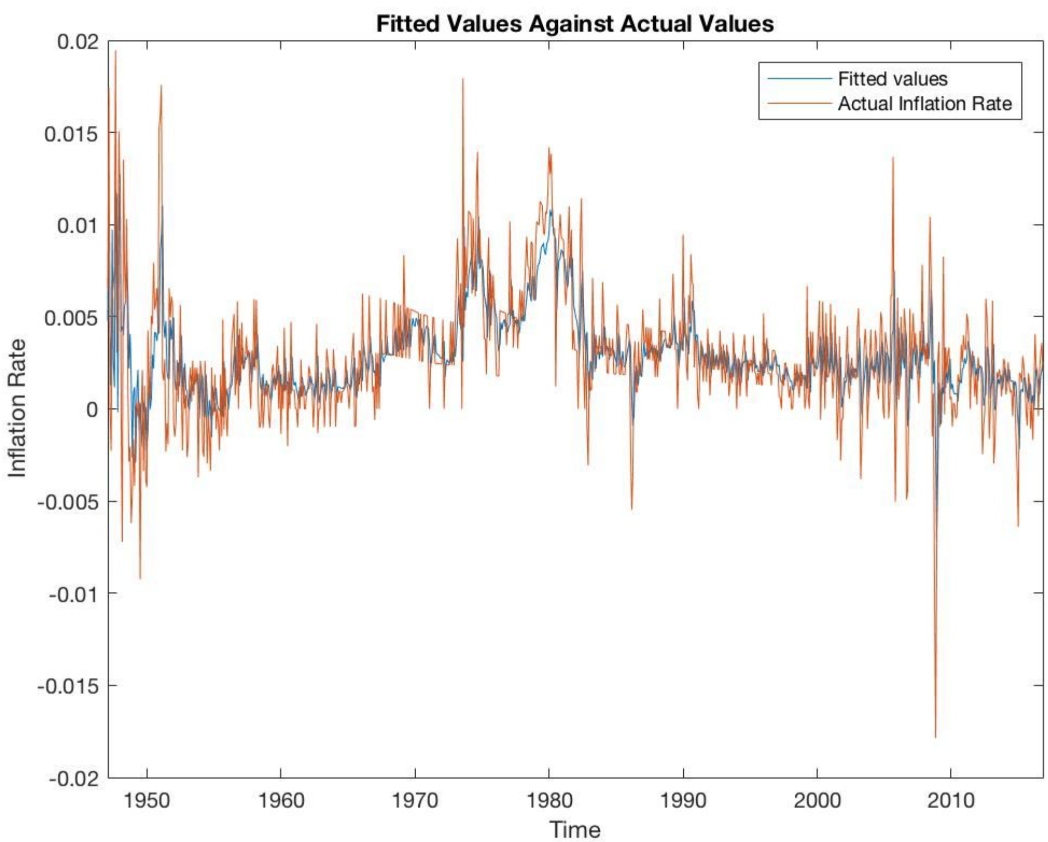
Is the serial correlation statistically significant? A Ljung-Box Q Test reveals that serial correlation is statistically significant at p-values of 0 at 99% confidence level both at 1 lag and 10 lags.

3.

```
ARIMA(4,0,3) Model:
--------------------
Conditional Probability Distribution: Gaussian

                                  Standard          t
 Parameter       Value             Error         Statistic
-----------    -----------    ------------    -----------
 Constant      0.00015612     8.07796e-05         1.93266
    AR{1}         0.59099       0.0387665         15.2449
    AR{2}       -0.310695       0.0160842        -19.3167
    AR{3}        0.919499       0.0163456         56.2536
    AR{4}       -0.256769       0.0295865        -8.67861
    MA{1}       -0.148676       0.0337224        -4.40881
    MA{2}        0.310224       0.0216848          14.306
    MA{3}        -0.77429       0.0289827        -26.7156
 Variance     6.77184e-06     3.50815e-08         193.032
```

The best fit model is ARMA(4,3) based on lowest AICBIC Information Criteria for lags. The information criteria are minimized for AR(4) and MA(3). A Ljung-Box Q Test results in h=0 and a p-value of 0.1466, meaning that we cannot reject the claim that there is no auto-correlation, which suggests that the residuals are not serially correlated.

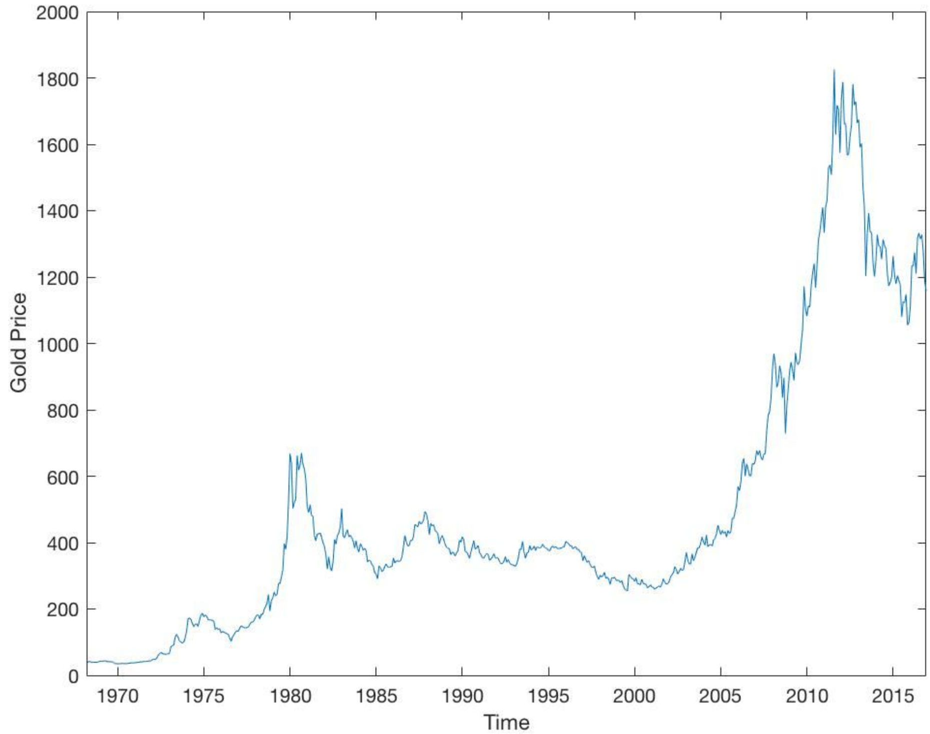
4.



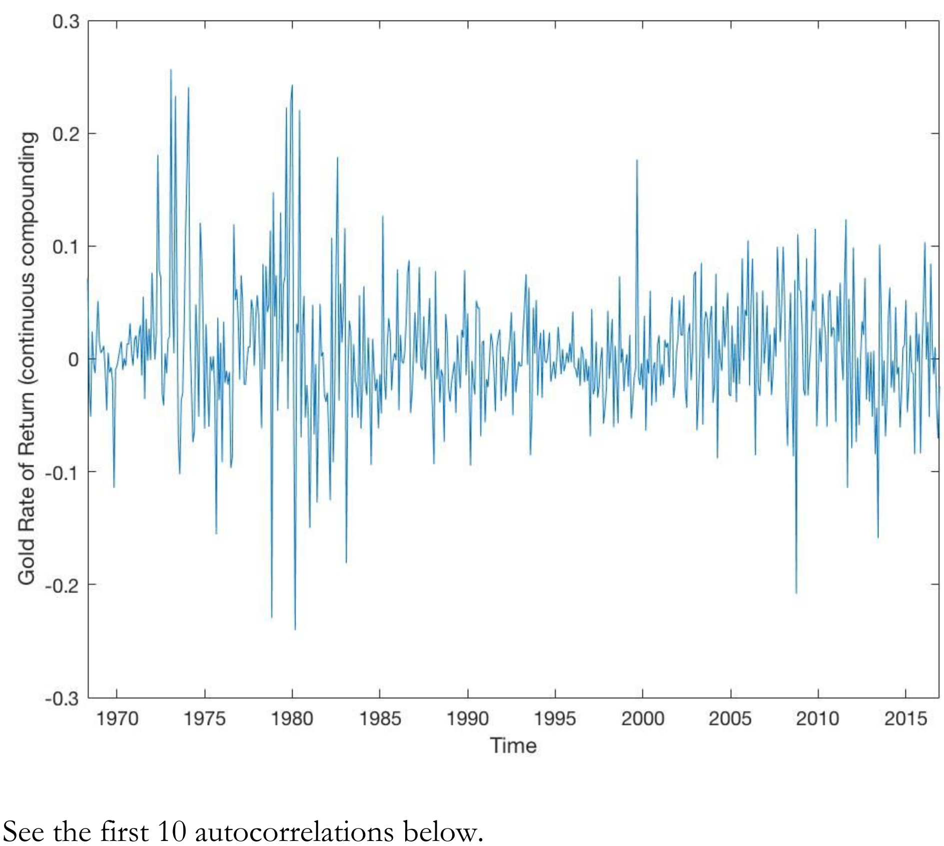
This forecasting model demonstrates a good fit as it follows the realized values closely. There are sporadic overestimations, however the model still holds well on to the general pattern. We fail to reject null hypothesis for the Ljung-Box Q-test for residual autocorrelation. Therefore, we do not have serial correlation in our error term – a sign of a good forecasting model.

Gold prices (GOLD variable)

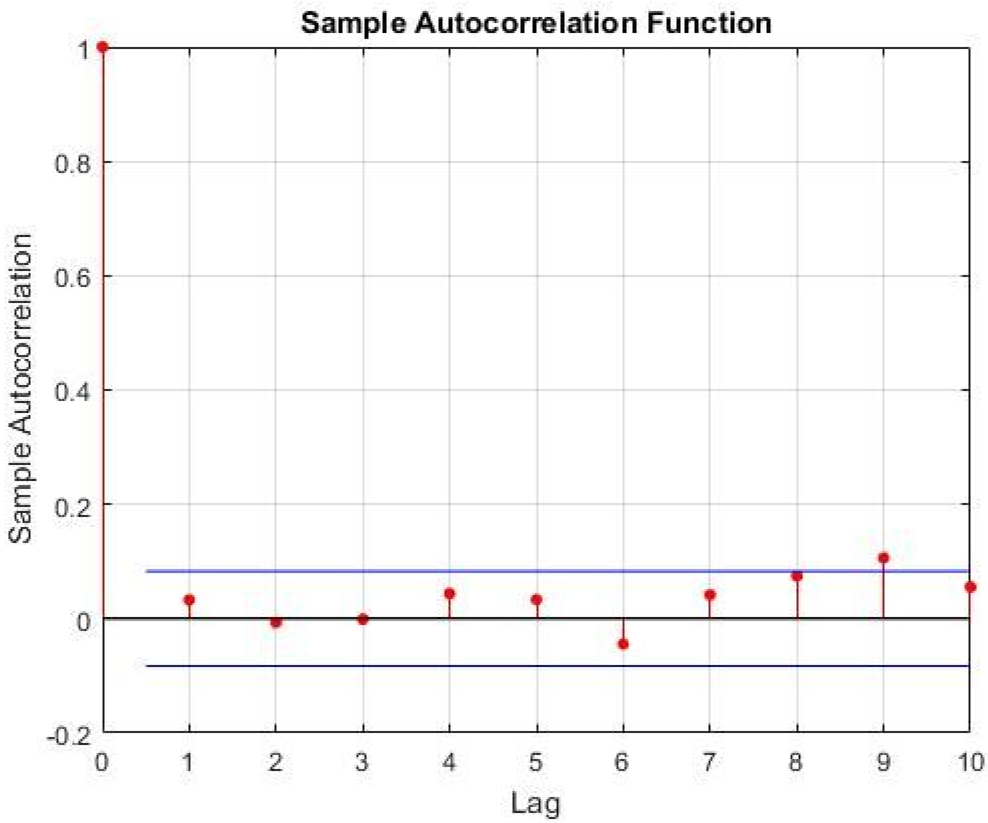
1.



Gold prices appear to have a steep upward trend. An Augmented Dickey-Fuller test results in an h=0 result with a p-value of 0.5162, failing to reject the null hypothesis of a unit root, and suggesting this series is not stationary. We therefore take the gold price log first differences, $\Delta\log(y_t) = \log(y_t) - \log(y_{t-1})$, which is the continuously compounded rate of return, and we plot the transformed series below.



2. See the first 10 autocorrelations below.



Is the variable persistent? The graph above implies no presence of a long-term influence of a shock, i.e. it does not significantly influence the rate of return from gold across 10 lag periods. The first 10 autocorrelations are:

Lag 1: 0.0328
Lag 2: -0.0067
Lag 3: -0.0009
Lag 4: 0.0439
Lag 5: 0.0334
Lag 6: -0.0444
Lag 7: 0.0417
Lag 8: 0.0746
Lag 9: 0.1063
Lag 10: 0.0554

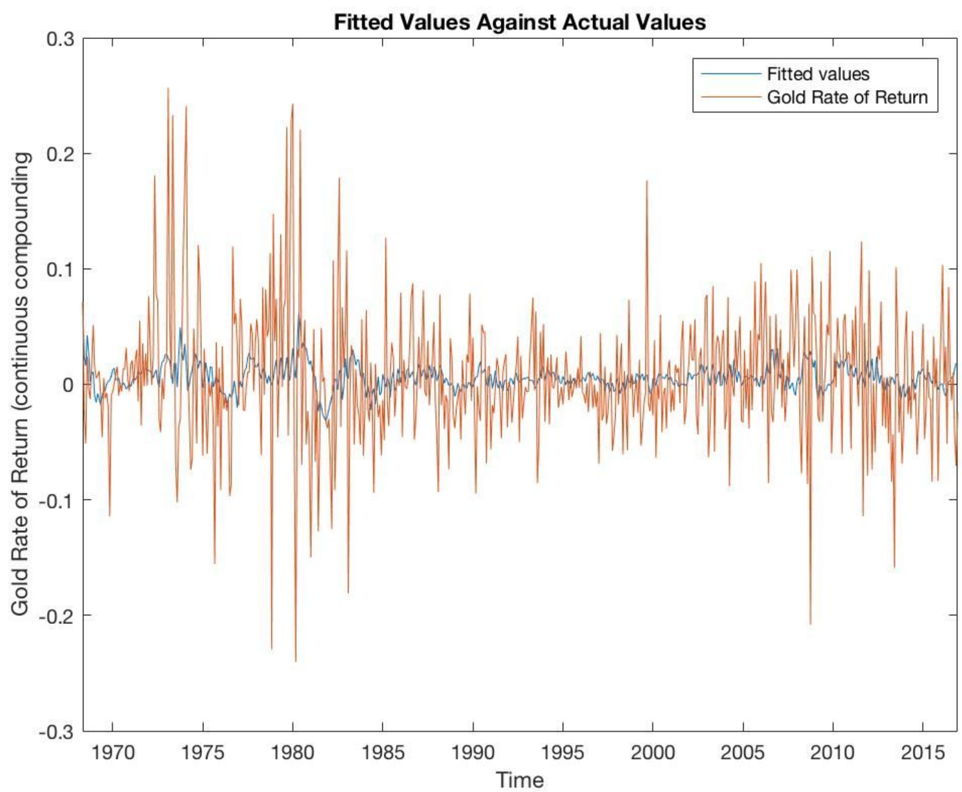
Is the serial correlation statistically significant? A Ljung-Box Q Test reveals that serial correlation is not statistically significant at a p-value of 0.068 at 95% confidence level at 10 lags.

3.

```
ARIMA(4,0,4) Model:
--------------------
Conditional Probability Distribution: Gaussian

                                  Standard          t
 Parameter       Value             Error        Statistic
-----------   -----------      ------------    -----------
 Constant     0.000598563     0.000404572         1.4795
    AR{1}         1.30392        0.174305         7.4807
    AR{2}       -0.757068        0.348313       -2.17353
    AR{3}        0.985842        0.300078        3.28528
    AR{4}       -0.634553        0.125578       -5.05308
    MA{1}        -1.30252        0.173652       -7.50076
    MA{2}        0.723588        0.337797        2.14208
    MA{3}              -1        0.279663       -3.57574
    MA{4}        0.721235        0.121038        5.95875
 Variance       0.0031691      0.00012738        24.8791
```

The best fit model is ARMA(4,4) based on the lowest AICBIC Information Criteria for lags. The information criteria are minimized for AR(4) and MA(4). A Ljung-Box Q Test results in h=0 and a p-value of 0.7699, meaning that we cannot reject the claim that there is no auto-correlation, which suggests that the residuals are not serially correlated.

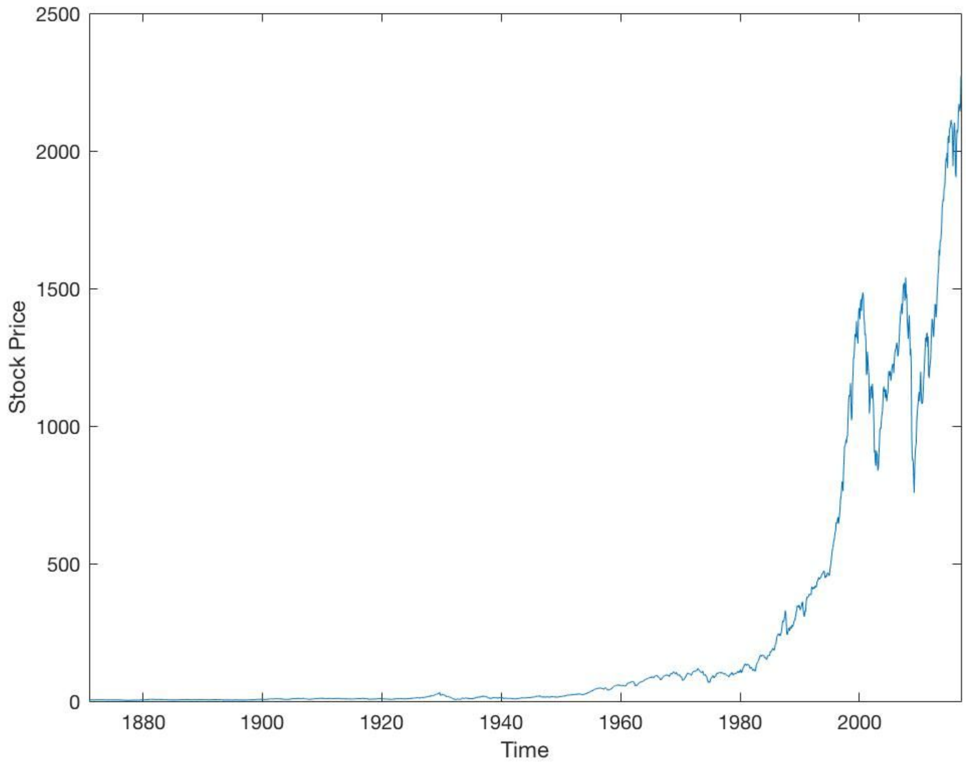
4.



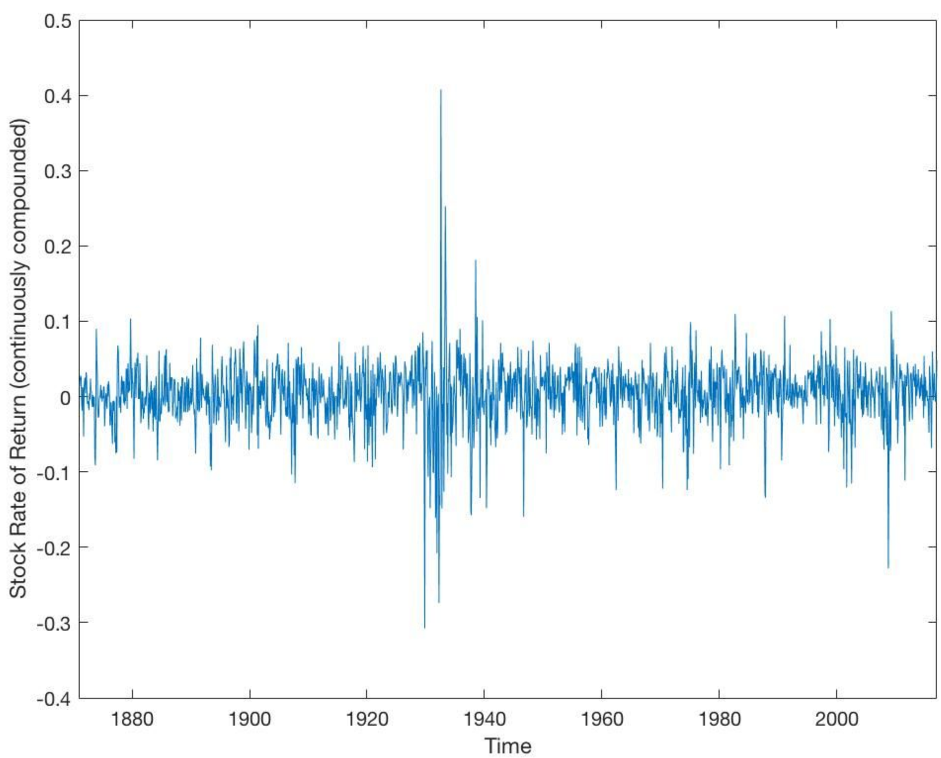
This forecasting model does not demonstrate the best fit as it continuously overestimates the values. We fail to reject null hypothesis for the Ljung-Box Q-test for residual autocorrelation. Therefore, we do not have serial correlation in our error term.

S&P500 Index

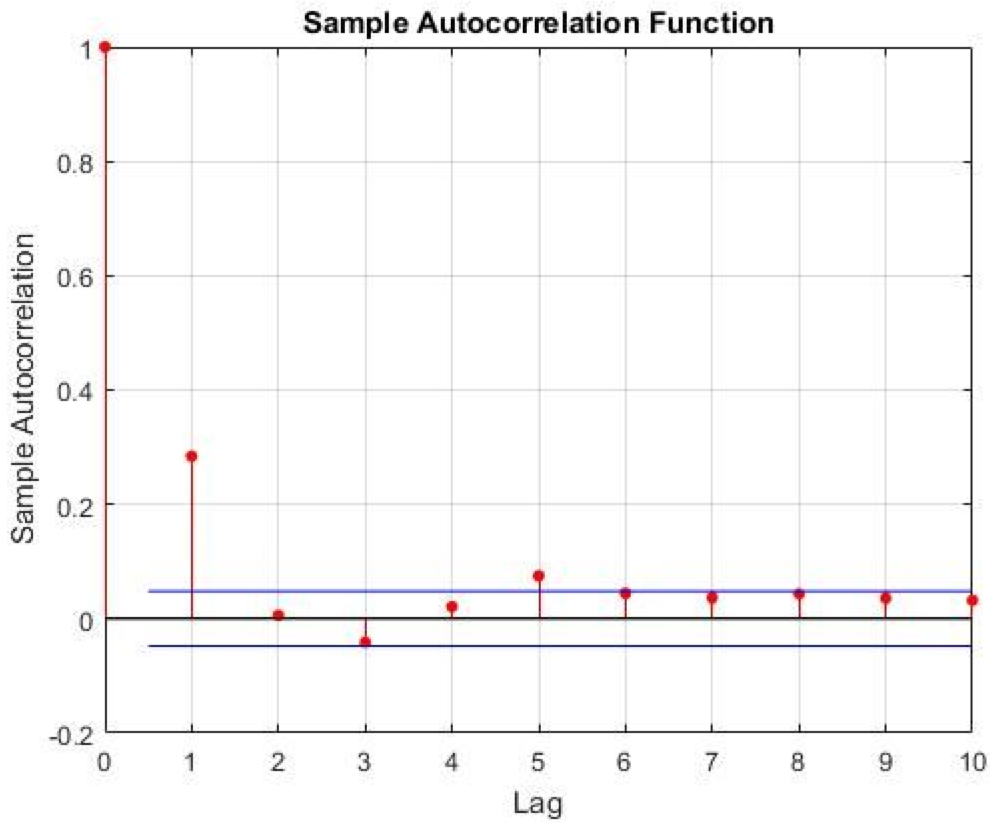
1.



An Augmented Dickey-Fuller test results in an h=0 result with a p-value of 0.9990, failing to reject the null hypothesis of a unit root, and suggesting this series is not stationary. We therefore take the stock price log first differences, $\Delta \log(y_t) = \log(y_t) - \log(y_{t-1})$, which is the continuously compounded rate of return, and we plot the transformed series below.



2. See the first 10 autocorrelations below.



Is the variable persistent? The graph above implies that the series is not particularly persistent, as it exhibits the presence of shock which quickly dies out after one period. The first 10 autocorrelations are:

Lag 1: 0.2841
Lag 2: 0.0057
Lag 3: -0.0414
Lag 4: 0.0209
Lag 5: 0.0750
Lag 6: 0.0445
Lag 7: 0.0363
Lag 8: 0.0434
Lag 9: 0.0358
Lag 10: 0.0320

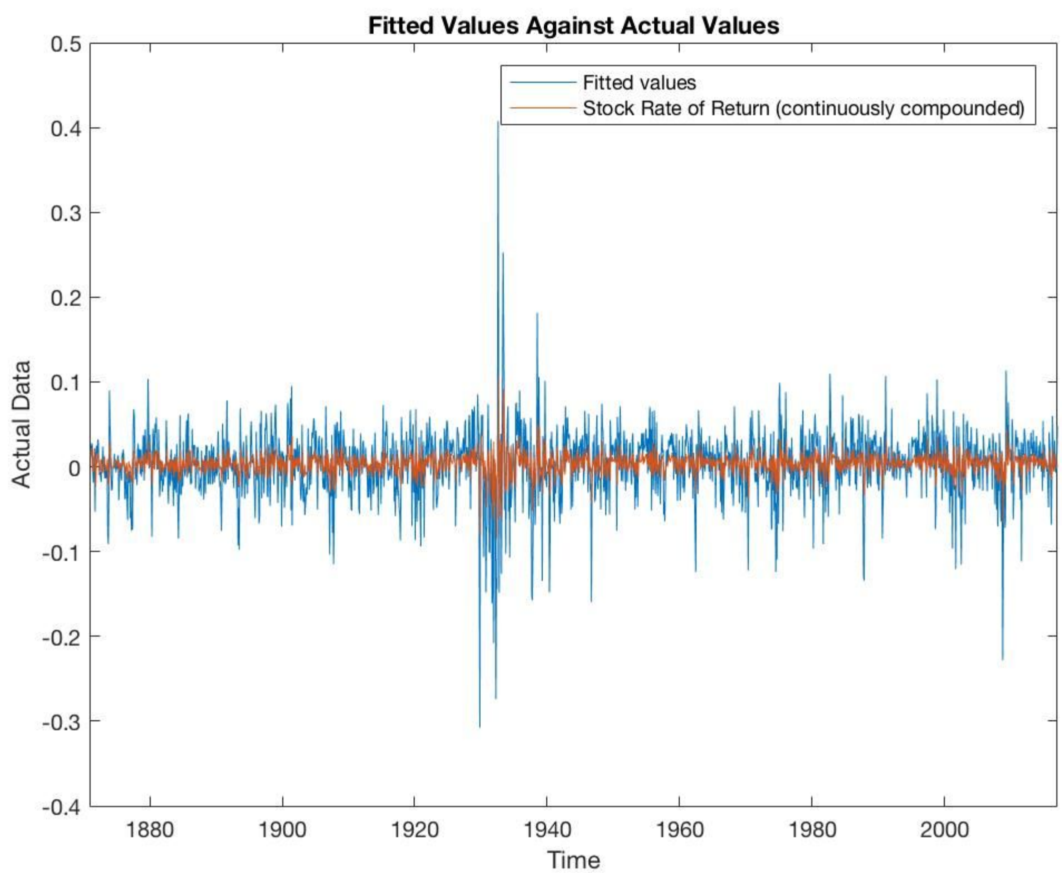
Is the serial correlation statistically significant? Autocorrelation is statistically significant at p-value of 0 at 99% confidence level at 10 lags.

3.

```
ARIMA(2,0,4) Model:
--------------------
Conditional Probability Distribution: Gaussian

                                Standard          t
 Parameter        Value           Error        Statistic
-----------   -----------   ------------   -----------
 Constant      0.00363356    0.00126971        2.86172
    AR{1}        0.969492     0.0203617        47.6135
    AR{2}       -0.959834     0.0192448        -49.875
    MA{1}       -0.666452     0.0253422       -26.2981
    MA{2}        0.694696     0.0265453        26.1702
    MA{3}        0.235791     0.0232948        10.1221
    MA{4}       0.0591092     0.0170663        3.46351
 Variance      0.00149621   2.32674e-05        64.3051
```

The best fit model is ARMA(2,4) based on the lowest AICBIC Information Criteria for lags. The information criteria are minimized for AR(2) and MA(4). A Ljung-Box Q Test results in h=0 and a p-value of 0.1432, meaning that we cannot reject the claim that there is no auto-correlation, which suggests that the residuals are not serially correlated.

4.

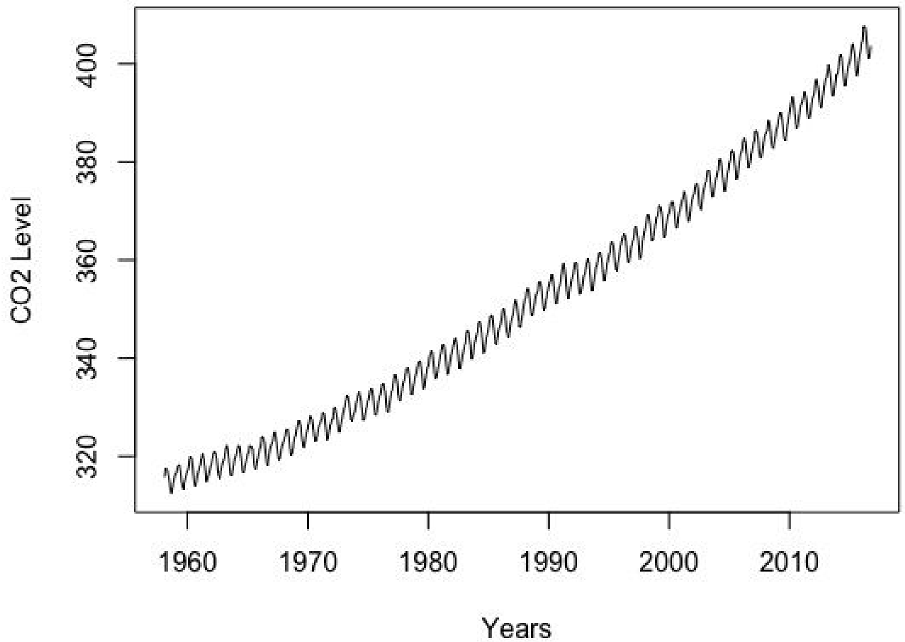


This forecasting model does not demonstrate the best fit as it continuously underestimates the values. We fail to reject null hypothesis for the Ljung-Box Q-test for residual autocorrelation. Therefore, we do not have serial correlation in the model, which supports the model's veracity.

Part II. Modeling Trend and Seasonality

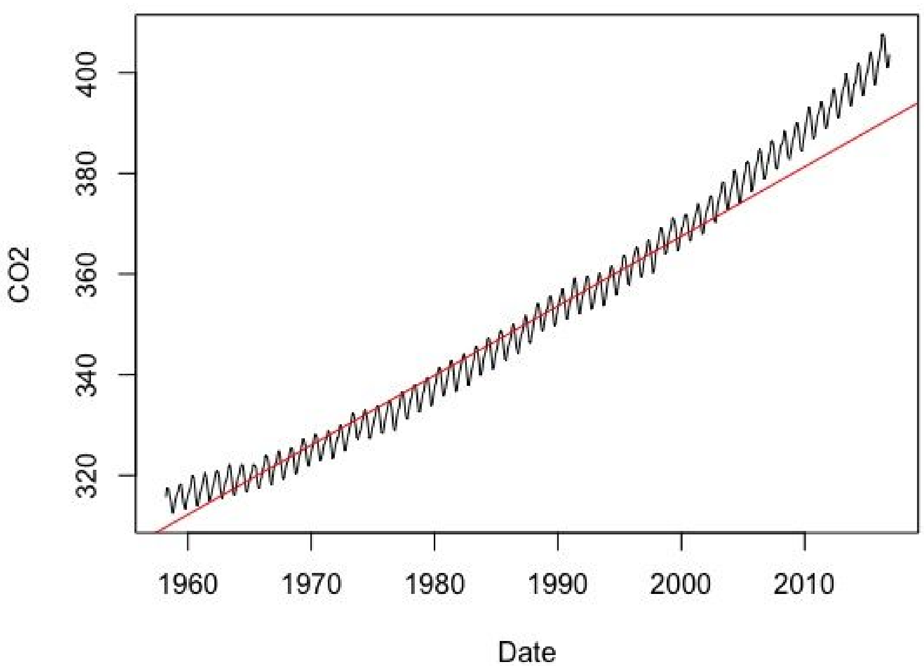
This assignment uses data on COS emissions from the file Keeling_CO2data_2017.xlsx available in the Assignment 1 folder on TED. The data was collected by Dave Keeling who worked at the Scripps Institute of Oceanography for many years. This is a famous data set that shows monthly measurements of CO2 in Hawaii from 1958 through 2016:10.

1. Plot the CO2 time series in column E. Briefly summarize the evidence of seasonal effects and trends in the time-series plot.



The data exhibits strong upward trend over time. Seasonal aspects are evident as well as the CO2 values seem to move up and down in a cyclical pattern. This likely corresponds to seasonal emissions during the winter in the developed world, in which emissions come from heating devices.

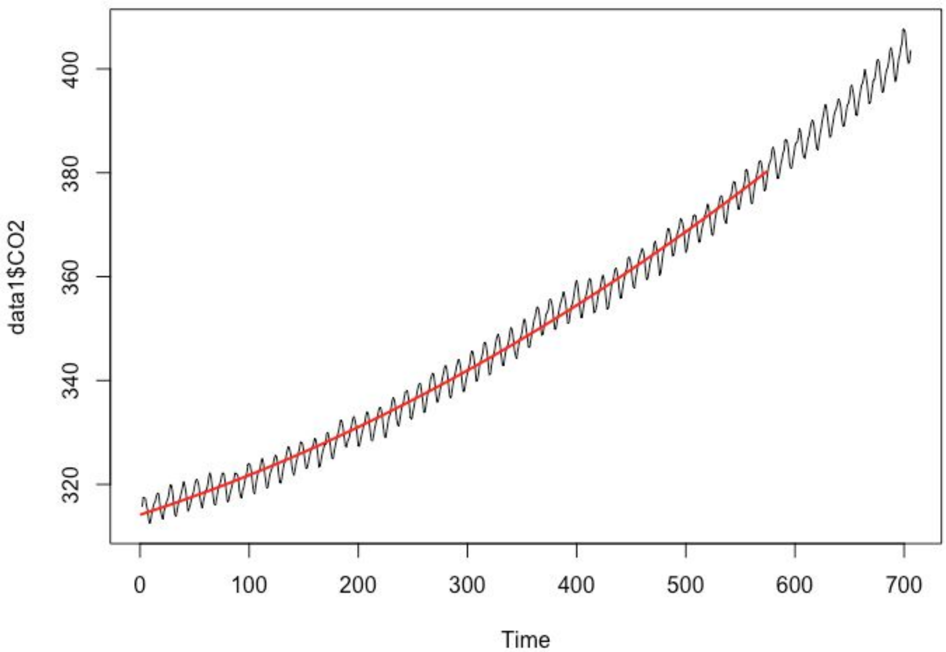
2. Using data up to 2005, estimate a linear trend model for the CO2 measurements. Is the trend significant? Is the linear trend a good specification that yields reliable forecasts?



```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -2.396e+03  1.783e+01  -134.4   <2e-16 ***
Time Trend       1.382e+00  8.993e-03   153.7   <2e-16 ***
```

The trend is highly significant at 95% level. However, the linear trend model does not yield a reliable forecast as the red line does not capture any of the seasonal variation. Furthermore, the linear trend does not capture the non-linear upward trend out of sample, and is highly biased towards fitting the in-sample data.

3. Using data up to 2005, develop a better trend specification. Report your trend estimates and explain what your specification is and what makes it better.

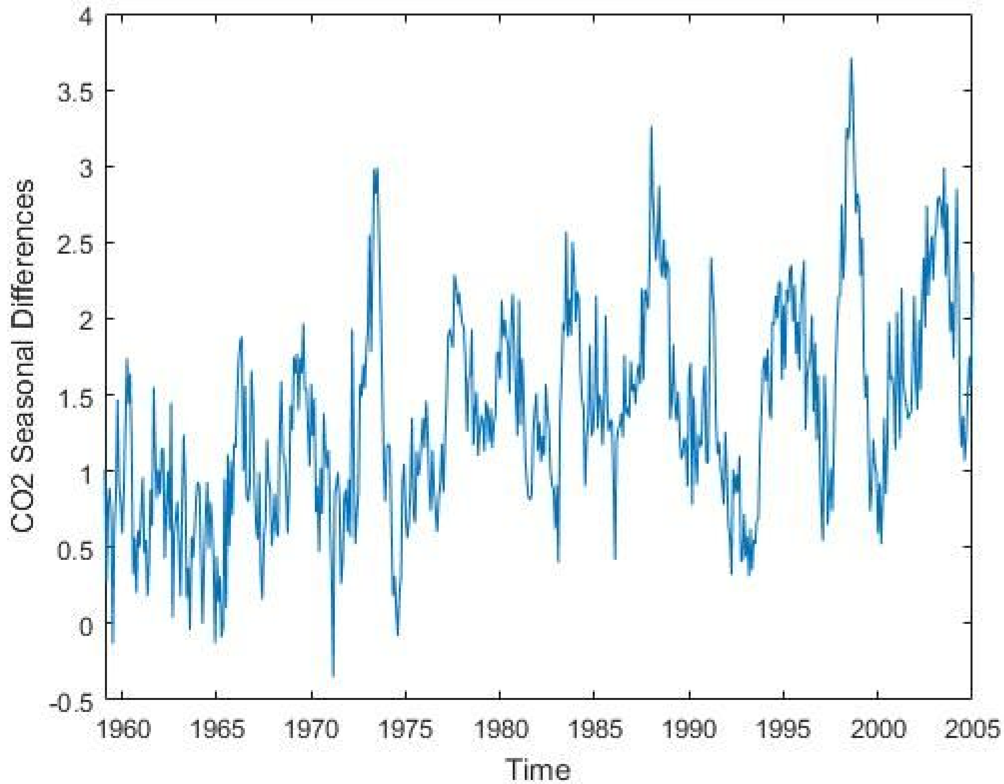



```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.444e+03  1.813e+01  -134.8   <2e-16 ***
t            1.406e+00  9.150e-03   153.7   <2e-16 ***
```

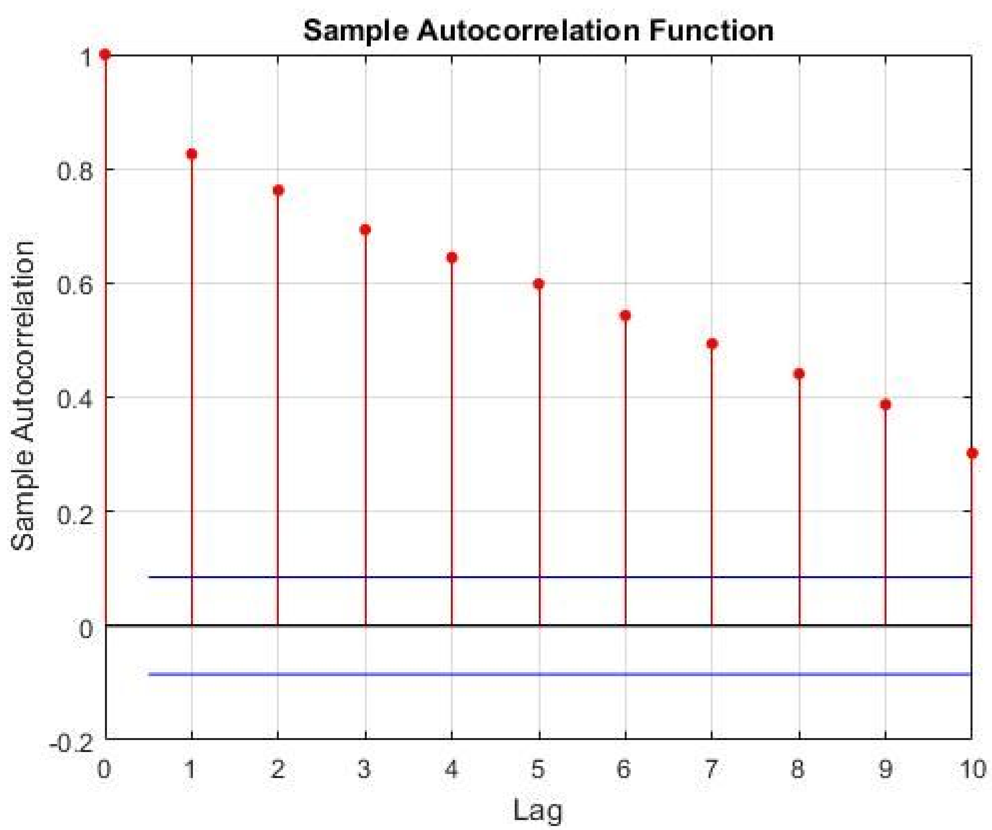
A quadratic trend specification is statistically significant at 95% level. It fits the curve better, but may be susceptible to overfitting in the long term. It also fails to capture the seasonal variation in the model.

4. Include seasonal effects and report results for your preferred model that includes both trend and seasonal effects. Again, use data only up to 2005.

An Augmented Dickey Fuller Test on CO_2 levels up till 2005 results in h=0, with a p-value of 0.9936. This means we fail to reject the null hypothesis of a unit root, suggesting we have a non-stationary process. Because this appears to be seasonally trended, we first account for seasonality by taking seasonal first differences, that measuring the difference of each observation versus 12 months ago. The seasonally differenced series is presented below:



An ACF of the differenced series exhibits a highly persistent series:

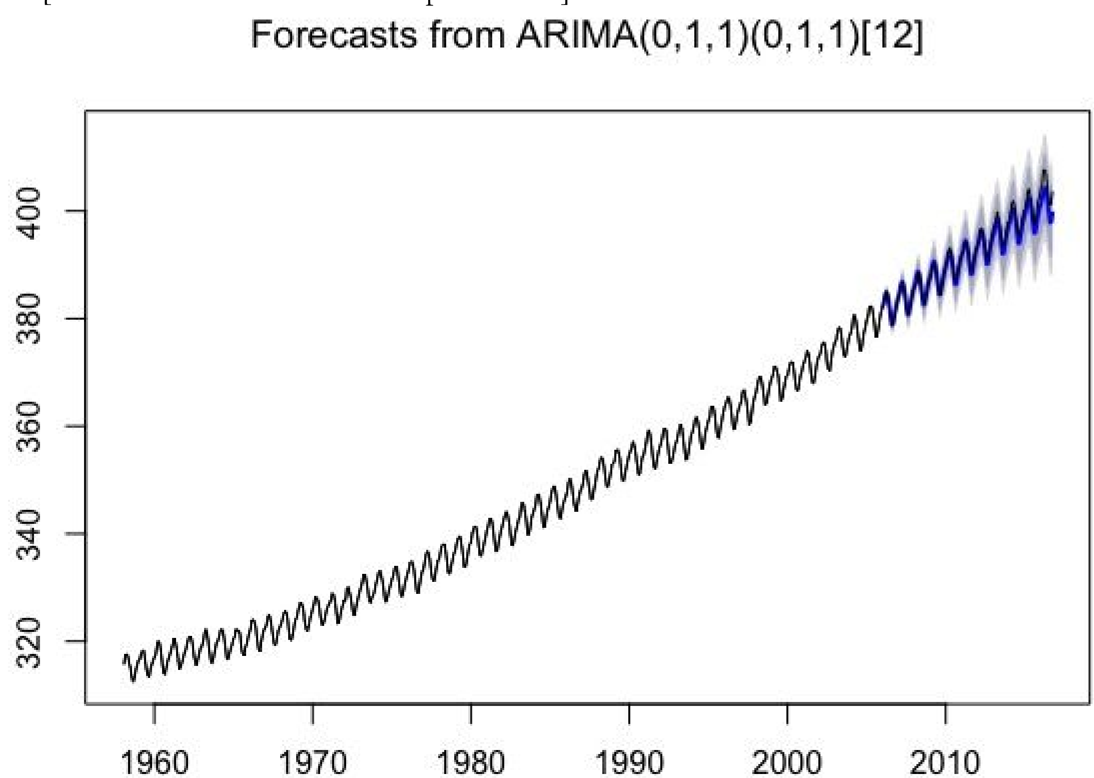


To determine the appropriate model for the series, we would set AR and MA maximum limits of 4 lags, and tested all combinations against AICBIC criteria. The lowest scores occurred at AR(4) and MA(4), so we could proceed to model this process with an ARMA(4,4).

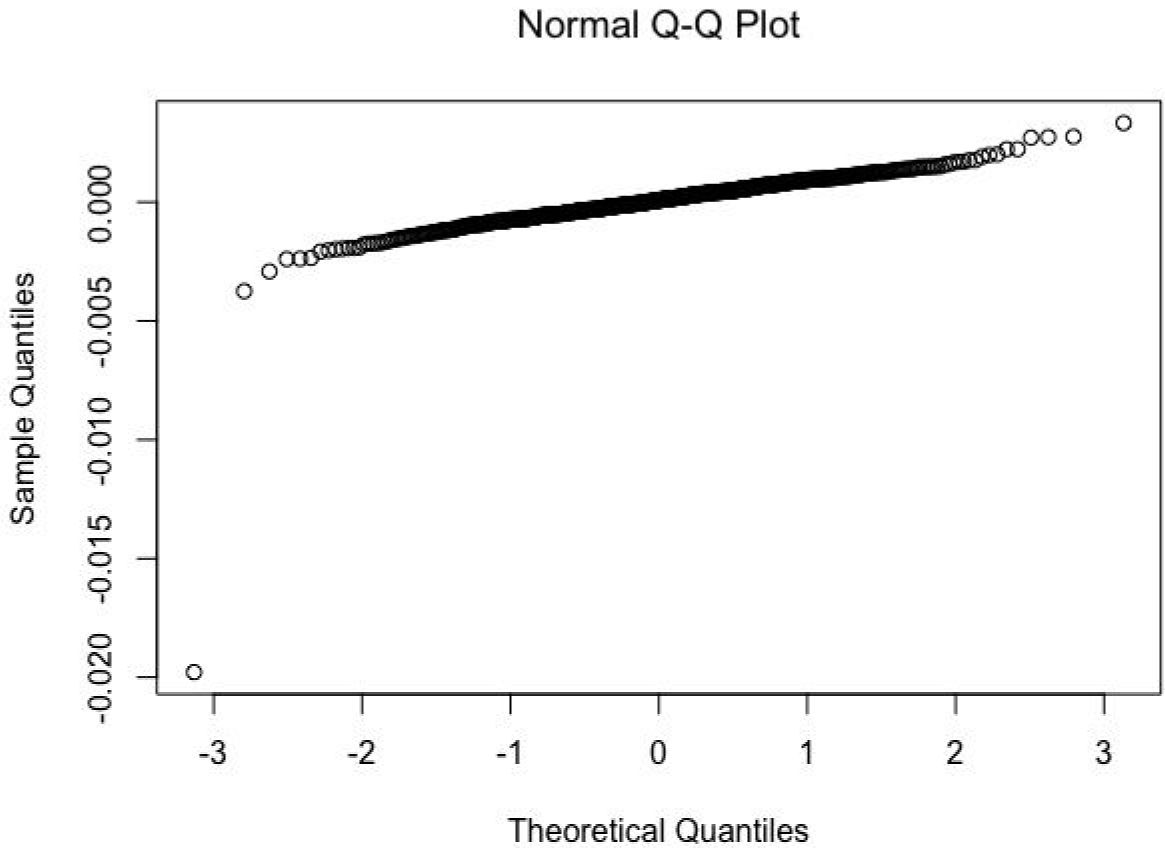
However, having done further research we find an ‘auto.arima’ function in R package ‘forecast’ that allows us to choose ARIMA models with a seasonal component by going through each one of them automatically. Based on the information criteria, the reported best Seasonal ARIMA (SARIMA) model is ARIMA(0,1,1)(0,1,1)[12]. Results below.

```
ARIMA(0,1,1)(0,1,1)[12]

Coefficients:
          ma1     sma1
      -0.3623  -0.8602
s.e.   0.0428   0.0220

sigma^2 estimated as 0.08814:  log likelihood=-115.22
AIC=236.44   AICc=236.48   BIC=249.44

Training set error measures:
                     ME     RMSE       MAE         MPE       MAPE      MASE       ACF1
Training set 0.01830858 0.293255 0.2313094 0.005121923 0.06764993 0.1678164 0.02520205
```

5. Using data up to 2005m12, forecast future CO2 for the period 2006m01 to 2016m10. Evaluate how good the forecasts from this model are. Are there any obvious problems with your forecasts [hint: are the forecast errors unpredictable]?



The actual data is shown in black. Visual inspection of the prediction made until 2016m10 [blue] shows that the model does a fairly good job following the seasonal pattern of the data, while starting to increasingly underestimate the level of CO2 with time. Forecast error margins are increasing over time, which is not good, but also is unavoidable in most cases. The residuals, however, seems to be normally distributed, as shown below.

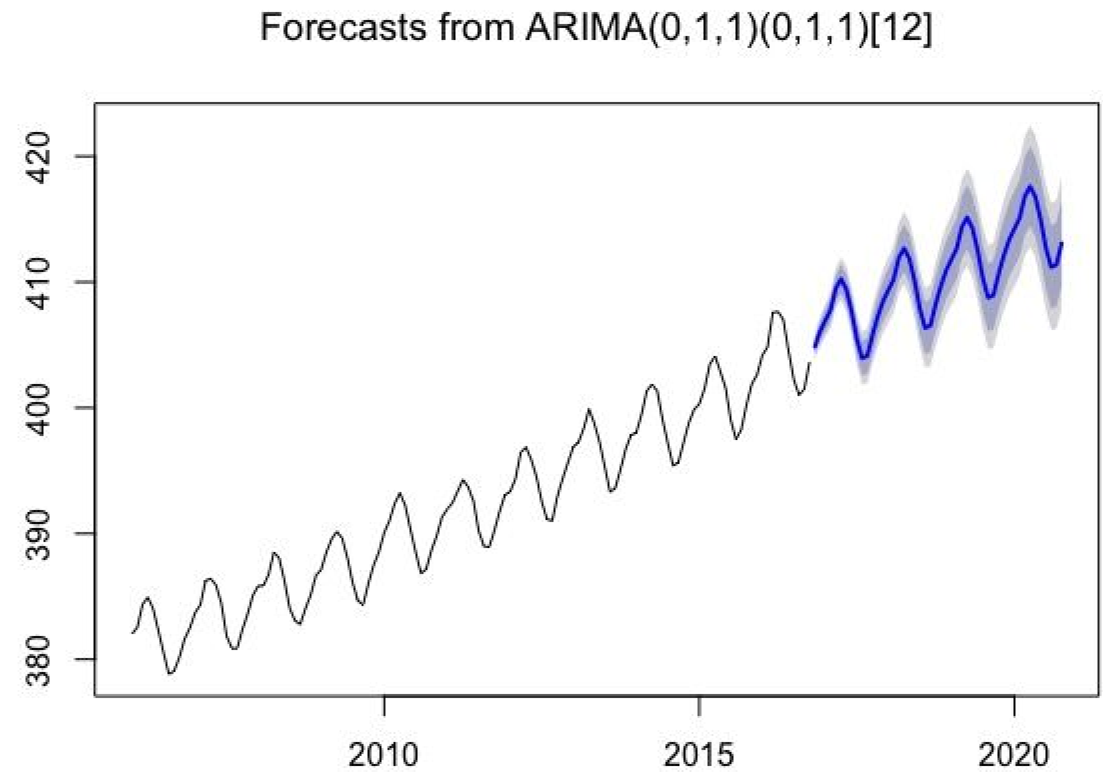


6. Include one additional forecasting variable in your model. Argue why you think it may help forecast CO2 and test if your intuition is right. [NB this is a variable of your own choosing and so you need to add it to the data set. To do so, add data covering the same period as the CO2 data (monthly from 1958-2016:10)]

We have accessed Carbon Dioxide Analysis Center to include 'Monthly Surface Air Temperature Time Series Area-Averaged Over the 30-Degree Latitudinal Belts of the Globe' variable. We thought that given that global warming is believed to be correlated with CO2 levels, this indicator would be statistically significant and close to show collinearity. As we run a naive OLS regression with it, we do see that this variable is not significant at 95% confidence and also shows a negative sign, which was unexpected. We proceed without using this variable.

```
lm(formula = CO2 ~ Time + Time^2 + Temperature)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.4880 -2.1033  0.0184  1.9767  7.5888 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  4.479e+02  1.184e+02   3.782 0.000172 ***
Time         4.098e-03  1.706e-04  24.023  < 2e-16 ***
Temperature -1.150e-01  6.231e-02  -1.846 0.065368 .  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.975 on 572 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.9765,	Adjusted R-squared:  0.9764 
F-statistic: 1.189e+04 on 2 and 572 DF,  p-value: < 2.2e-16
```

7. Produce a forecast of CO2 for December 2020. How reliable do you think your forecast is? Be as specific as you can in discussing this point.



This figure using all of available data and the same SARIMA model specification from above to produce a forecast until December 2020. RMSE looks relatively small, so we tend to think this is a reasonably plausible model. However, it is susceptible to increasing variance in errors over time, thus long term predictions, such as this one, may pose little practical value to the scientific and climate communities. See descriptive statistics below.

```
ARIMA(0,1,1)(0,1,1)[12]
Box Cox transformation: lambda= 0

Coefficients:
         ma1     sma1
      -0.368  -0.9048
s.e.   0.000   0.0000

sigma^2 estimated as 1.477e-06:  log likelihood=638.77
AIC=-1275.55   AICc=-1275.51   BIC=-1272.78

Training set error measures:
                      ME      RMSE       MAE          MPE       MAPE      MASE      ACF1
Training set -0.008377616 0.8100432 0.3650793 -0.002531108 0.09332173 0.1620804 0.1149829
```