**Written Report – 6.419x Module 4**

**Name:** 4sowmya

**2. The Mauna Loa CO\_2 Concentration**

1. **(3 points) Plot the periodic signal, Pi. (Your plot should have 1 data point for each month, so 12 in total.) Clearly state the definition the Pi, and make sure your plot is clearly labeled.**

Let Ci be the average CO2 concentration in month i (i=1,2, 3, …12). Ci can be described as

**Ci = F(ti) + Pi +Ri**

Where,

F: t 🡪 F(t) accounts for the long-term trend.

ti = time at the middle of *i*th month

Pi = periodic in *i* with a fixed period, accounting for the seasonal pattern.

Ri = the remaining residual that accounts for all other influences.

Based on the residual plots drawn in the auto graded exercise and the prediction errors reported, the lowest degree polynomial that seems to be sufficient to represent the data is quadratic model (degree 2). Hence the periodic signal fitting is done using quadratic model.

Figure-1 shows the predicted periodic signal Pi for training data. This value is the monthly average of CO2 concentration residual for the quadratic fit of the training data.

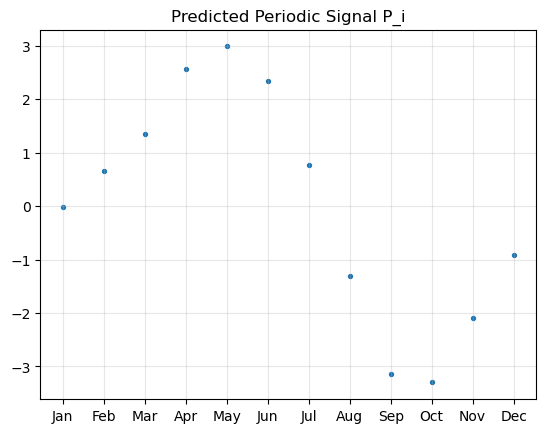
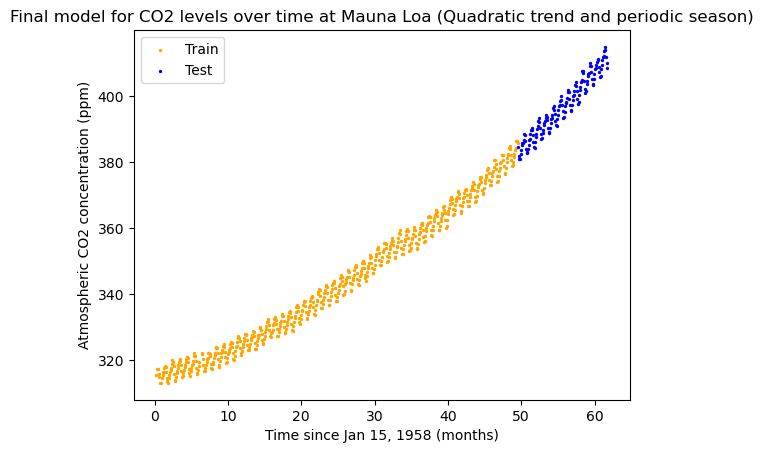


Figure 1 Periodic signal of CO2 concentration (1958-2007)

1. **(2 points) Plot the final fit Fn(ti) +Pi . Your plot should clearly show the final model on top of the entire time series, while indicating the split between the training and testing data.**



1. **(4 points) Report the root mean squared prediction error RMSE and the mean absolute percentage error MAPE with respect to the test set for this final model. Is this an improvement over the previous model Fn(ti)without the periodic signal? (*Maximum 200 words*.)**

Here are the results of the test data with both final models :

RMSE for model with Fn(ti)= 2.502807

RMSE for model with Fn(ti)+Pi = 0.8574

MAPE for model with Fn(ti) = 0.53%

MAPE for model with Fn(ti)+Pi = 0.1820%

The model used for the prediction is quadratic model of degree 2. From these two metrics, we infer that the prediction result from the final model Fn(ti)+Pi is an improvement over the model Fn(ti) without the periodic signal.

1. **(3 points) What is the ratio of the range of values of F to the amplitude of Pi and the ratio of the amplitude of P to the range of the residual Ri (from removing both the trend and the periodic signal)? Is this decomposition of the variation of the CO2 concentration meaningful? (*Maximum 200 words*.)**

The decomposition of the variation of the CO2 concentration is meaningful only if the range of F is much larger than the amplitude of the Pi and this amplitude is much larger than the Ri. The results for the quadratic model with degree 2 over the training data are as follows:

range of F = 69.1437

amplitude of P = 3.147

range of residual, R = 9.84455

From the above values, we can infer that the decomposition of the CO2 is not meaningful as the amplitude of P is much smaller than the range of residual R

**3. Autocovariance Functions**

1. **(4 points) Consider the model, MA (1) model,**

**Xt = Wt +** *Ɵ* **Wt-1**

**Where, {Wt} ~ W ~ Normal\_dist (0,** σ2**).**

**Find the autocovariance function of {Xt}. Include all important steps of your computations in your report.**

Since the sum of white noise with expectation is 0,

**mean (E[Xt]) = 0**

A first order moving-average process, written as MA (1), has the general equation

*xt = wt + Ɵwt−1*

where wt is a white-noise series distributed with constant variance σw2.

We must compute γ(k), which is defined as the autocovariance of the process at lag k. Since xt has zero mean,

γ(k) = E (xt,xt−k)

γ(k) = E [(wt + *Ɵ*wt−1) (wt−k + *Ɵ*wt−k−1)]

= E (wtwt−k + *Ɵ*wtwt−k−1 + *Ɵ*wt−1wt−k + *Ɵ*2wt−1wt−k−1)

= E(wtwt−k) + E(*Ɵ*wtwt−k−1) + E (*Ɵ*wt−1wt−k) + E (*Ɵ*2wt−1wt−k−1)

For **k=0**, γ (0) = σMA2 = variance of the series

Γ (0) = σMA2

= E (w t 2) + *Ɵ* E(wtwt−1) + *Ɵ* E(wt−1wt) + *Ɵ*2 E (w t−12)

= σw2 + 0 + 0 + *Ɵ*2σ w2

**Γ (0) = (1 + *Ɵ*2) σ w2**

For **k = 1**.

γ (1) = E(wtwt−1) + *Ɵ* E(wtwt−1) + *Ɵ* E(wt−12wt-1) + *Ɵ*2 E(wt−1wt-2)

**γ (1) = *Ɵ* σ w2**

For **k > 1**, we will obtain **γ(k) = 0**, since E [(wt + *Ɵ* wt−1) (wt−k + *Ɵ* wt−k−1)] will contain only terms, whose expected value is zero.

For an MA (1), the autocovariance function truncates (i.e., it is zero) after lag 1.

1. **(4 points) Consider the AR (1) model,**

**Xt = Ø\* Xt - 1 + Wt,**

**Where, {Wt} ~ W ~ Normal\_dist (0, σ2). Suppose | Ø |<1. Find the autocovariance function of {Xt}. (You may use, without proving, the fact that {Xt} is stationary if | Ø |<1.) Include all important steps of your computations in your report.**

Since the sum of white noise with expectation is 0, and Xt is stationary

**E[Xt] =** Ø **E[Xt-1] = 0**

E [X t 2] = σAR2 = γ0 = φ2E [X t-1 2] + σ2

**E[WtWs]** =

If φ < 0 then the autocorrelation function oscillates and has negative correlation at lag 1.

The autocovariance for this autoregressive process is

for h>=1

**Hence the white noise, E[WtWs] =**  when s=t

**E[WtWs] = 0 when s ≠ t**

**5. Converting to Inflation Rates**

1. **Repeat the model fitting and evaluation procedure from the previous page for the monthly inflation rate computed from CPI.**

**Your response should include:**

**(1 point) Description of how you compute the monthly inflation rate from CPI and a plot of the monthly inflation rate. (You may choose to work with log of the CPI.)**

**(2 points) Description of how the data has been detrended and a plot of the detrended data.**

**(3 points) Statement of and justification for the chosen AR(p) model. Include plots and reasoning.**

**(3 points) Description of the final model; computation and plots of the 1 month-ahead forecasts for the validation data. In your plot, overlay predictions on top of the data.**

Read the two files given in the dataset (CPI and BRE). The monthly CPI can be obtained by getting the first value for each month in a year. Plotting the graph for monthly CPI in the time series way shows that the data follows a linear trend (fig ure-2). The data is split into train and test data as mentioned in the problem statement. Running a linear regression model on the training data and visualizing in the graph (figure 3) to see the linear graph.

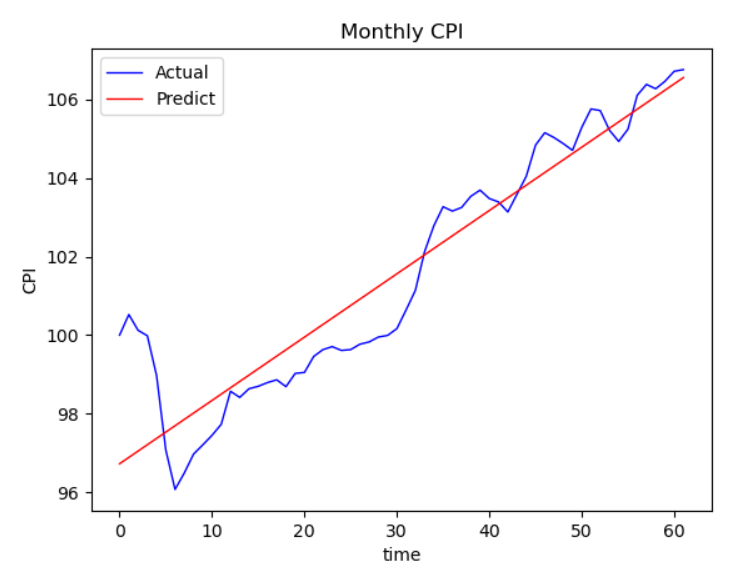
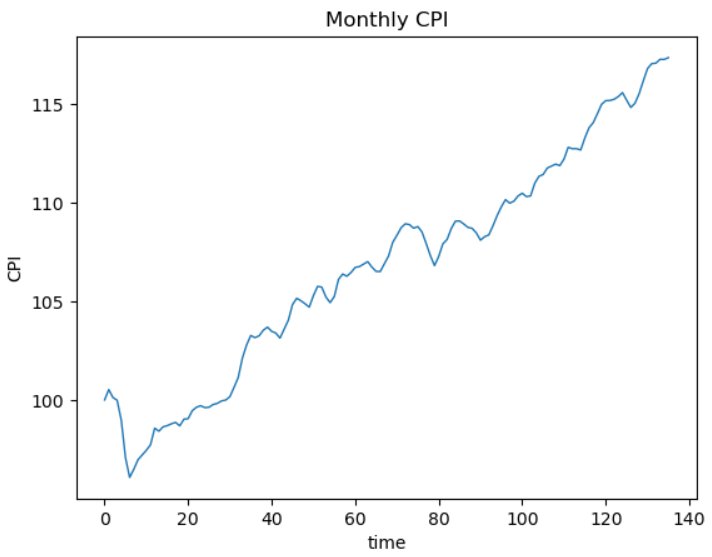
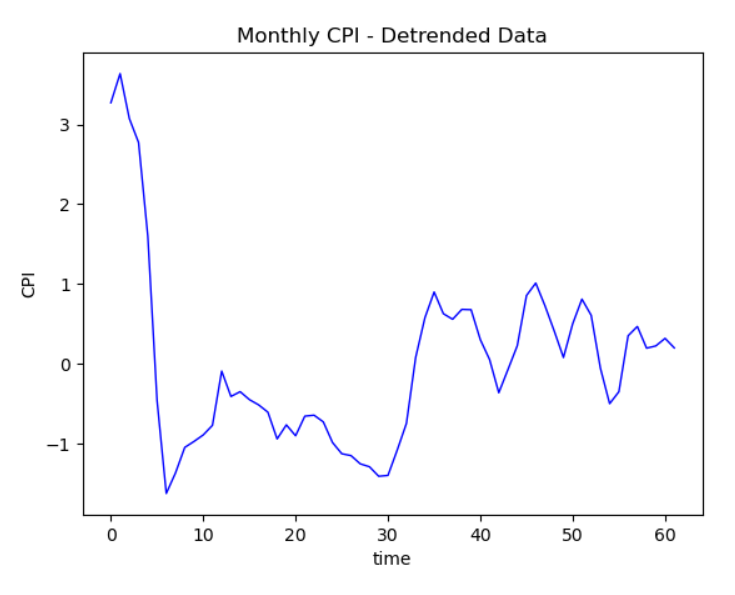
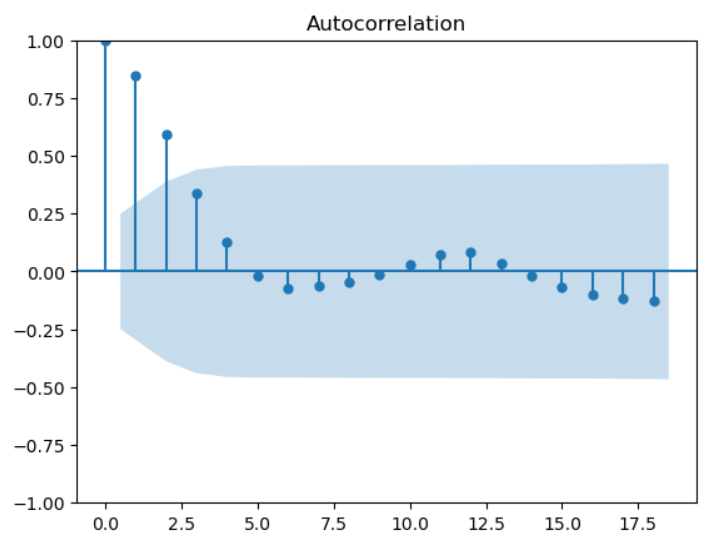
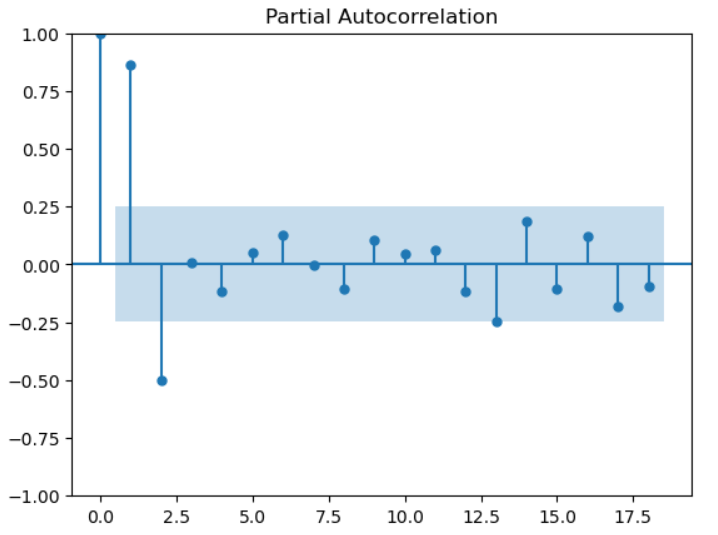


Figure 2 Monthly CPI Figure 3 Monthly CPI (Actual vs Predicted)

We now subtract this linear trend from the data to get the residue (Detrended data)



The above figure does not show any sufficient info for trend. We can continue with linear trend to de-seasonalize the data. Since the visual does not provide information on the seasonality, we perform the AR model on the residue. We can determine the order p of the AR model using the autocorrelation and partial correlation functions of the residuals.

Now we calculate AR(n) for n in the range of 1 to 5 and get the RMSE for test data for each AR(n). Here are the RSME for each n

RMSE for AR(1):0.36992170332386093

RMSE for AR(2):0.39104043069964806

RMSE for AR(3):0.3762508376939184

RMSE for AR(4):0.4399093616668139

RMSE for AR(5):0.4321160798236862

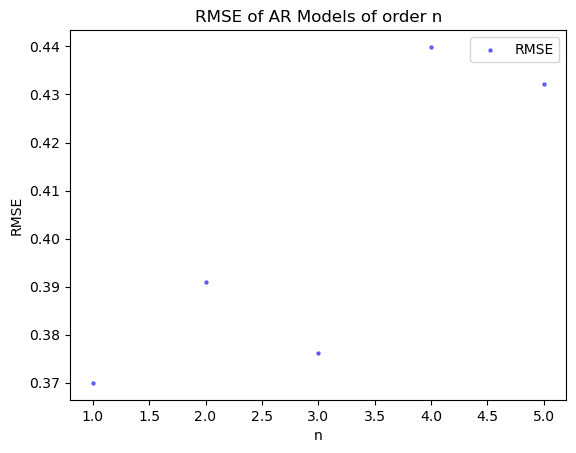
From the above values, AR(1) has the least RSME. This is a discrepancy with our inference from the results of the partial autocorrelation where the lag is 2. We will continue to work with lag value of 2 to avoid missing any lag terms.

1. **(3 points) Which AR(p) model gives the best predictions? Include a plot of the RSME against different lags for the model.**

AR(2) model will give the best prediction. Creating the partial autocorrelation for that model is shown below.

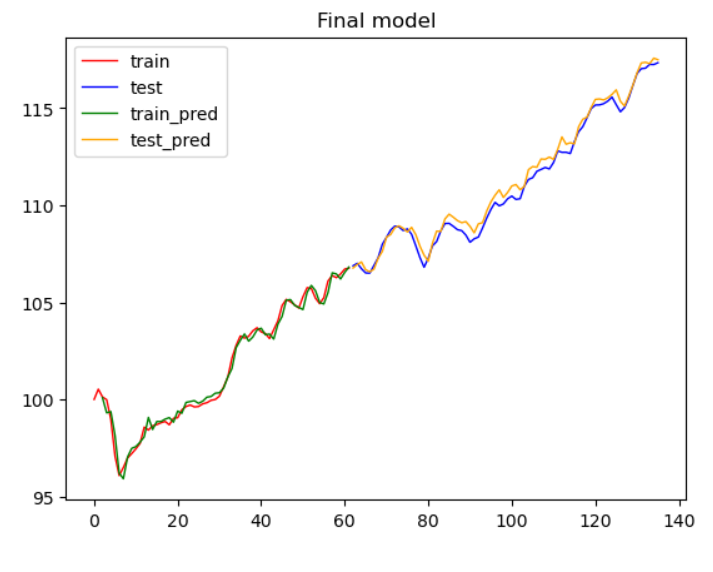
The RSME for the test data is 0.3932

We will consider that the order of the data is 10^2.



### **Inflation Rate from BER**

**(3 points) Overlay your estimates of monthly inflation rates and plot them on the same graph to compare. (There should be 3 lines, one for each dataset, plus the prediction, over time from September 2013 onward.)**

: 

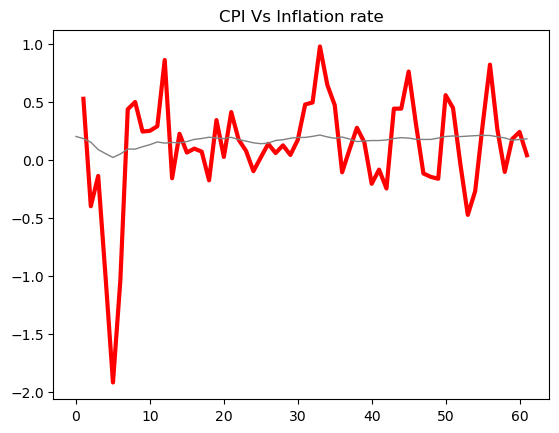
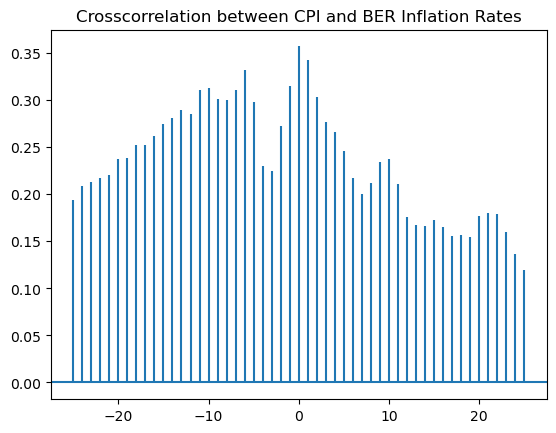
We see that the AR(2) Model does predict pretty well and the mean squared prediction error is small. We can also reaffirm this conclusion by plotting the residuals after the AR(2) predictions are subtracted from the detrended data

**External Regressors**

1. **(4 points) Plot the cross correlation function between the CPI and inflation rate, by which find, i.e., the lag between two inflation rates. (As only one external regressor term is involved in the model, we only consider the peak in the CCF plot.)**

**Note: In general, multiple external terms can be incorporated in the model if there are multiple peaks in CCF plots.**

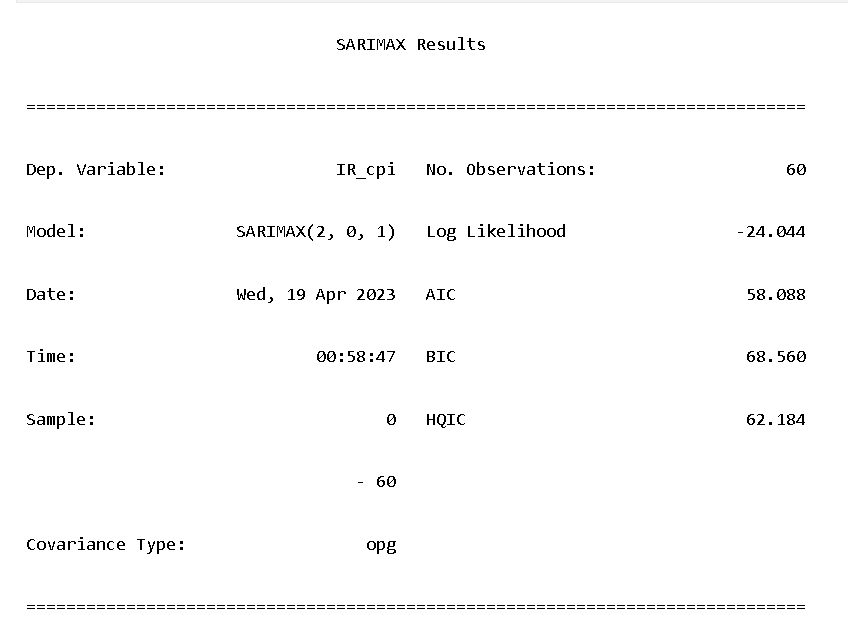
Comparing the lag between the CPI and inflation rate. In order to identify the lag between the external regressor and the CPI timeseries, we plot the cross-correlation plots between BER and CPI.

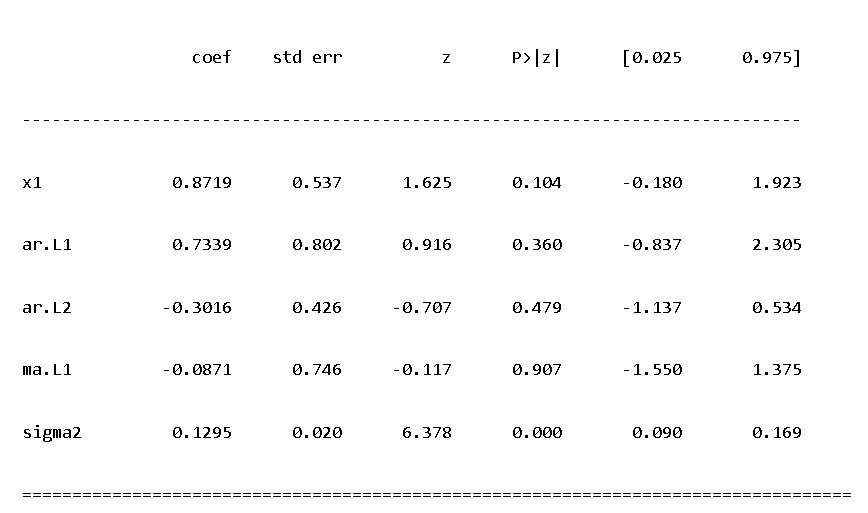
 

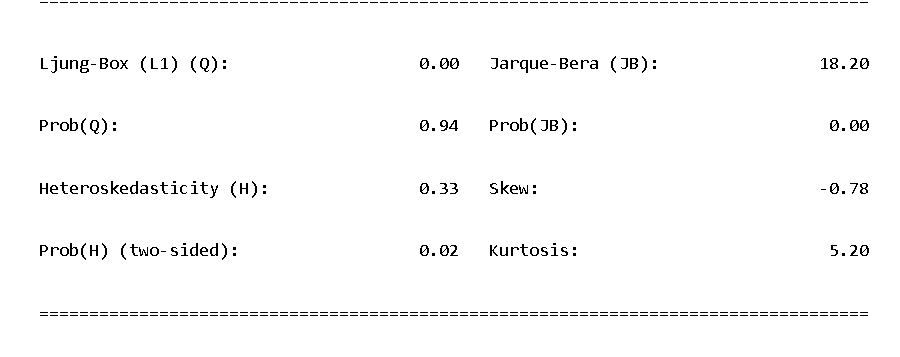
1. **(3 points) Fit a new  model to the  inflation rate with these external regressors and the most appropriate lag. Report the coefficients and plot the 1 month-ahead forecasts for the validation data. In your plot, overlay predictions on top of the data.**

**Python Tip: You may use sm.tsa.statespace.SARIMAX.**

Here is the result of the new model

******

******

******

1. **(3 points) Report the mean squared prediction error for 1 month ahead forecasts.**

RMSE of the final model = 0.2360

**Improving your Model**

**(5 points) What other steps can you take to improve your model from part III? What is the smallest prediction error you can obtain? Describe the model that performs best. You might consider including MA terms, adding a seasonal AR term, or adding multiple daily values (or values from different months) of BER data as external regressors.**

Based on the results so far, we see that even though the RSME has increased, the percentage error MAPE has halved. Here is the prediction of the new improved model compared to the actual data.

: 