MITx 6.419x Data Analysis: Statistical Modeling and Computation in Applications
Written report -- Homework 4
Xing Zhang

5/14/21

Module 4: Time Series

Problem 2: The Mauna Loa CO_2 Concentration

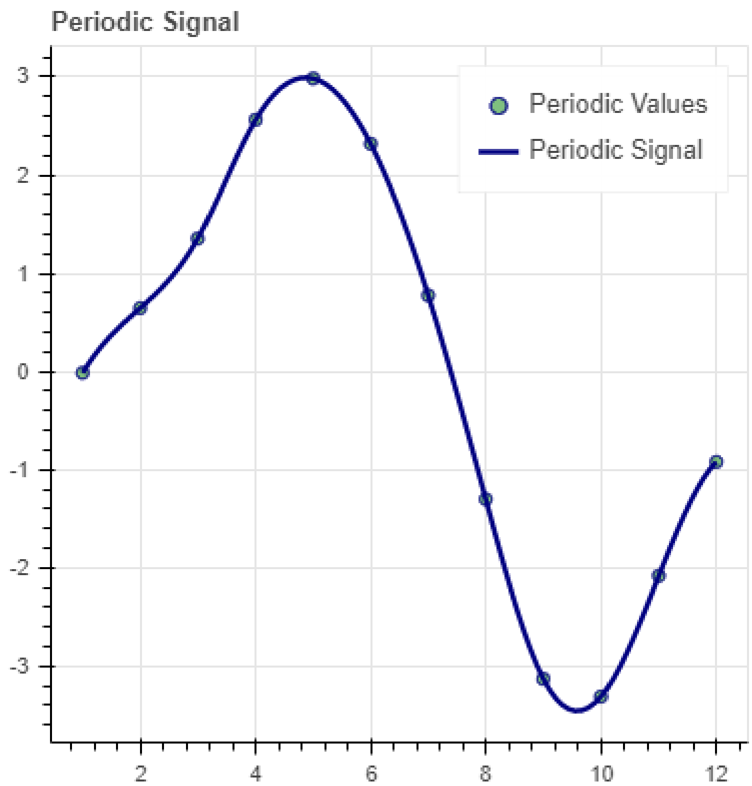
The final model

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

> Python tip: For interpolation, you may use interp1d from Scikit-learn. See Documentation on interp1d.

Solution:

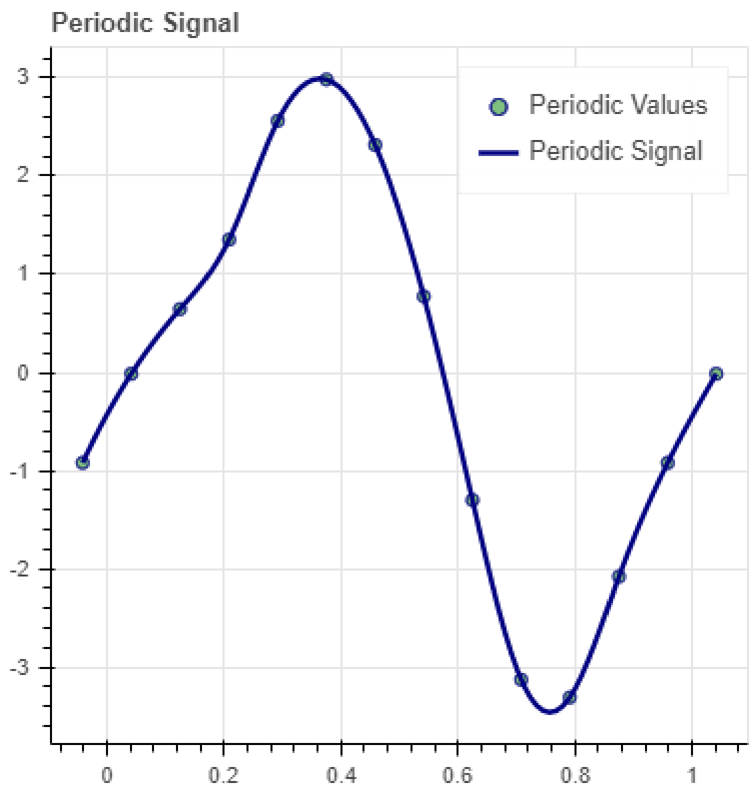
In this problem, it has been assumed that the time period is 12 (which seems to be valid as CO2 concentration is linked to seasonal variation.) First we extract the periodic monthly trend by taking the average of the periodic trend over all months. Next, we interpolate between this values to get a continuous periodic signal.

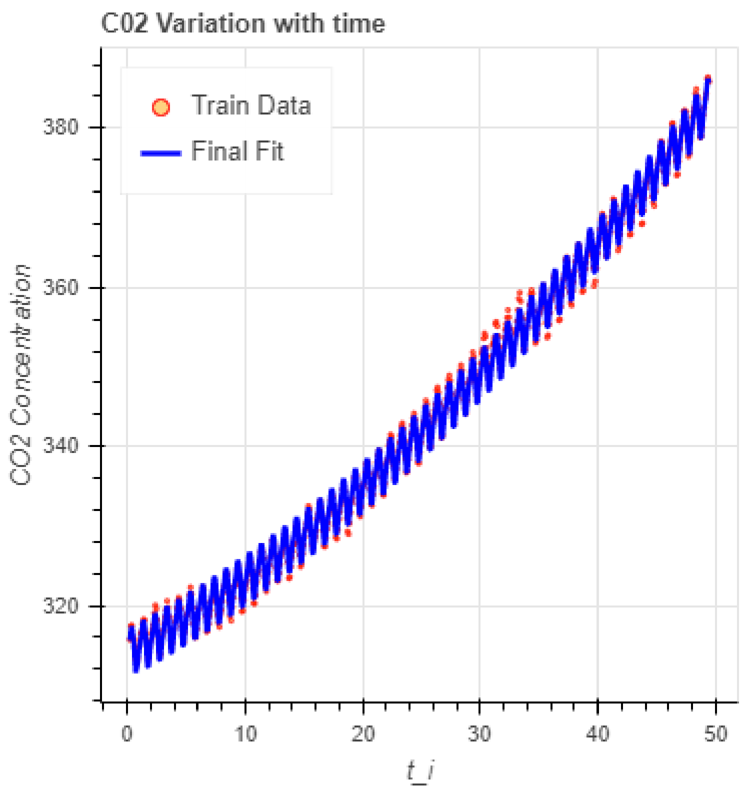


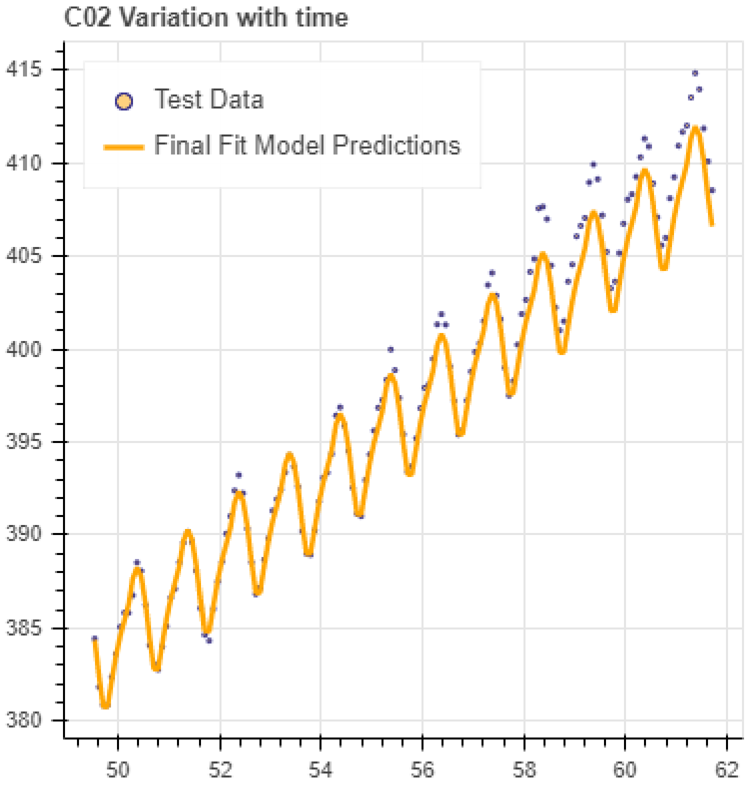
2. (2 points) Plot the final fit $F_n(t_i) + P_i$. Your plot should clearly show the final model on top of the entire time series, while indicating the split between the training and testing data.

Solution:

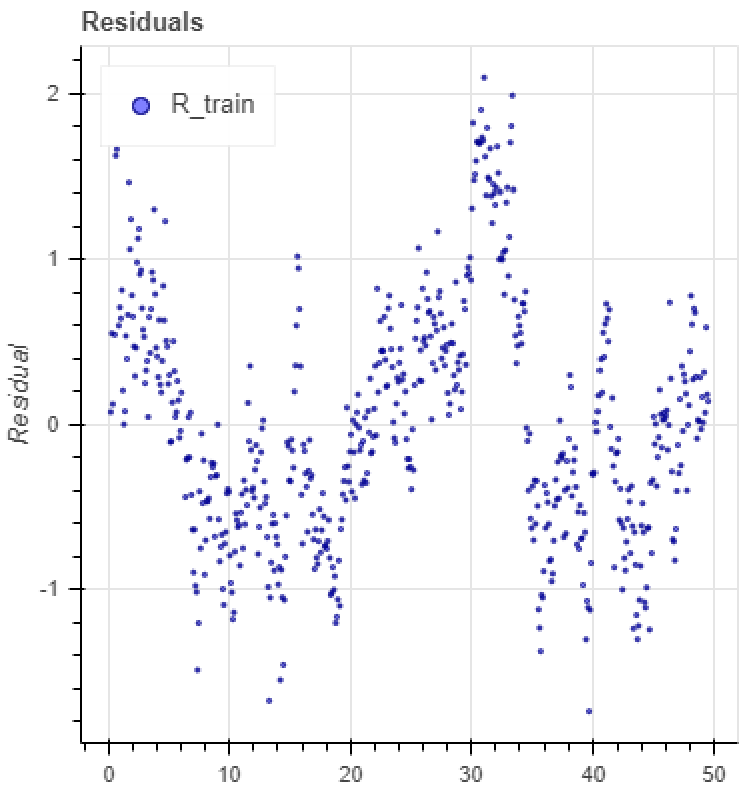
First, time of P_i need to be normalized to the global time to match the time of F; then we modify P_i into a continuous function in the interval[0,1], and finally we superimpose P_i and F to get the final fit.



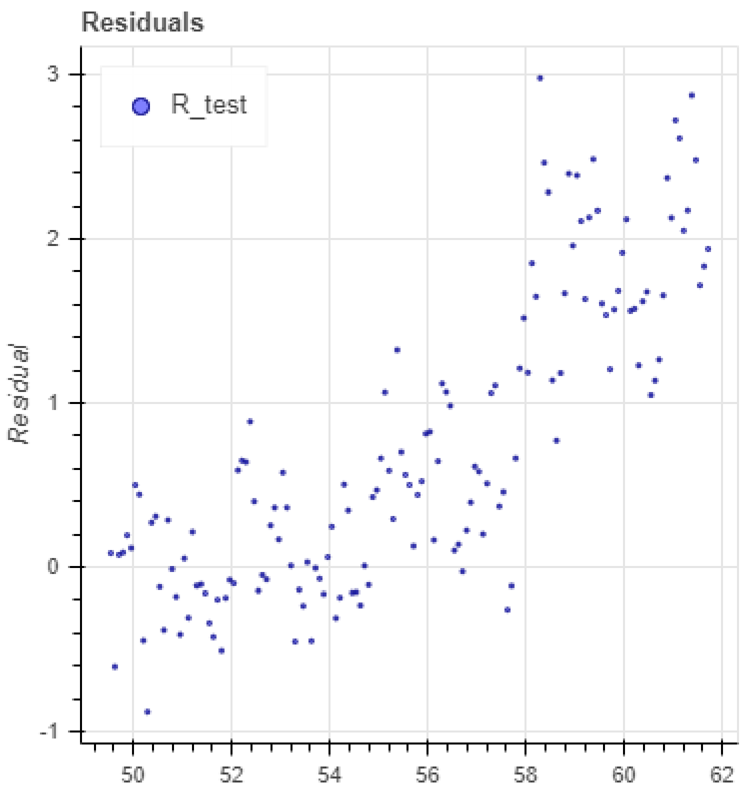




C02 Variation with time
Test Data
Final Fit Model Predictions
415
410
405
400
395
390
385
380
50
52
54
56
58
60
62



Residuals
R_train
Residual
2
1
0
-1
0
10
20
30
40
50



In train data fit, R value is vibration within a certain range. In test data fit, R value shows an increasing trend with time growth. This means that the growth rate of CO2 concentration began to increase after year 2000.

3. (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 $F_n(t_i)$ without the periodic signal? (Maximum 200 words.)

Solution:

The mean squared prediction error of the Final Fit model is 1.1493602596634225
The mean absolute percentage error of the Final Fit model is 0.20859165685936393

Compare to the previous analysis:

> The mean squared prediction error of the trend model is 2.5013322194898326
> The mean absolute percentage error of the trend model is 0.5320319129740952

So that the final model improved a lot over the previous model $F_n(t_i)$ without the periodic signal.

4. (3 points) What is the ratio of the range of values of F to the amplitude of P_i and the ratio of the amplitude of P to the range of the residual R_i (from removing both the trend and the periodic signal)? Is this decomposition of the variation of the CO2 concentration meaningful? (Maximum 200 words.)

Solution:

The ratio of the range of values of F to the amplitude of P_i is 15.95743409659297
The ratio of the amplitude of P to the range of the residual R_i is 1.3341427263860355

We know that:

> The decomposition is meaningful only if the range of F is much larger than the amplitude of the P_i and this amplitude in turn is substantially larger than that of R_i.

So in this case, the decomposition is meaningful.

Problem 3: Autocovariance Functions

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

$X_t = W_t + \theta W_{t-1},$

where $W_t \sim W \sim \mathcal{N}(0, \sigma^2)$. Find the autocovariance function of X_t.
Include all important steps of your computations in your report.

Solution

Suppose that $\{X_t\}$ is a time series with $\mathrm{E}\left[X_t^2\right] < \infty$ Its mean function is

$$\mu_t = \mathrm{E}\left[X_t\right]$$

Its autocovariance function is

$$\begin{aligned}\gamma_X(s,t) &= \operatorname{Cov}(X_s, X_t) \\ &= \mathrm{E}\left[\left(X_s - \mu_s\right)\left(X_t - \mu_t\right)\right]\end{aligned}$$

In this case

$X_t = W_t + \theta W_{t-1},$

where $W_t \sim W \sim \mathcal{N}(0, \sigma^2)$. So that,

We have $\mathrm{E}\left[X_t\right] = 0,\ \ \mathrm{E}\left[W_t^2\right] = \sigma^2,\ \ \mathrm{E}\left[\theta^2 W_{t-1}^2\right] = \theta^2\sigma^2,\ \ \mathrm{E}\left[W_t W_{t-1}\right] = 0$

so we got autocovariance function:

$$\begin{aligned}\gamma_X(t+h,t) &= \mathrm{E}\left(X_{t+h} X_t\right) \\ &= \mathrm{E}\left[\left(W_{t+h} + \theta W_{t+h-1}\right)\left(W_t + \theta W_{t-1}\right)\right] \\ &= \begin{cases} \sigma^2\left(1+\theta^2\right) & \text{if } h = 0, \\ \sigma^2\theta & \text{if } h = \pm 1 \\ 0 & \text{otherwise} \end{cases}\end{aligned}$$

2. (4 points) Consider the AR(1) model,

$X_t = \phi X_{t-1} + W_t,$

where $W \sim W_t \sim \mathcal{N}(0, \sigma^2)$. Suppose $|\phi| < 1$. Find the autocovariance function of X_t. (You may use, without proving, the fact that X_t is stationary if $|\phi| < 1$.)
Include all important steps of your computations in your report.

Solution

Suppose that $\{X_t\}$ is a time series with $\mathrm{E}\left[X_t^2\right] < \infty$ Its mean function is

$$\mu_t = \mathrm{E}\left[X_t\right]$$

Its autocovariance function is

$$\begin{aligned}\gamma_X(s,t) &= \mathrm{Cov}(X_s, X_t)\\ &= \mathrm{E}\left[(X_s - \mu_s)(X_t - \mu_t)\right]\end{aligned}$$

In this case

$X_t = \phi X_{t-1} + W_t,$

where $W \sim W_t \sim \mathcal{N}(0, \sigma^2)$. Suppose $|\phi| < 1$. Then we have

$$\begin{aligned}\mathrm{E}\left[X_t\right] &= \phi \mathrm{E}\left[X_{t-1}\right]\\ &= 0\end{aligned}$$

$$\begin{aligned}\mathrm{E}\left[X_t^2\right] &= \phi^2 \mathrm{E}\left[X_{t-1}^2\right] + \sigma^2\\ &= \frac{\sigma^2}{1-\phi^2}\end{aligned}$$

so we got autocovariance function:

$$\begin{aligned}\gamma_X(h) &= \mathrm{Cov}(\phi X_{t+h-1} + W_{t+h}, X_t)\\ &= \phi\,\mathrm{Cov}(X_{t+h-1}, X_t)\\ &= \phi \gamma_X(h-1)\\ &= \phi^{|h|}\gamma_X(0) \quad (\text{check for } h > 0 \text{ and } h < 0)\\ &= \frac{\phi^{|h|}\sigma^2}{1-\phi^2}\end{aligned}$$

Problem 5: Converting to Inflation Rates

Inflation Rate from CPI

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.
- (2 points) Description of how the data has been detrended and a plot of the detrended data. (You may choose to work with log of the CPI for later convenience.)
- (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.

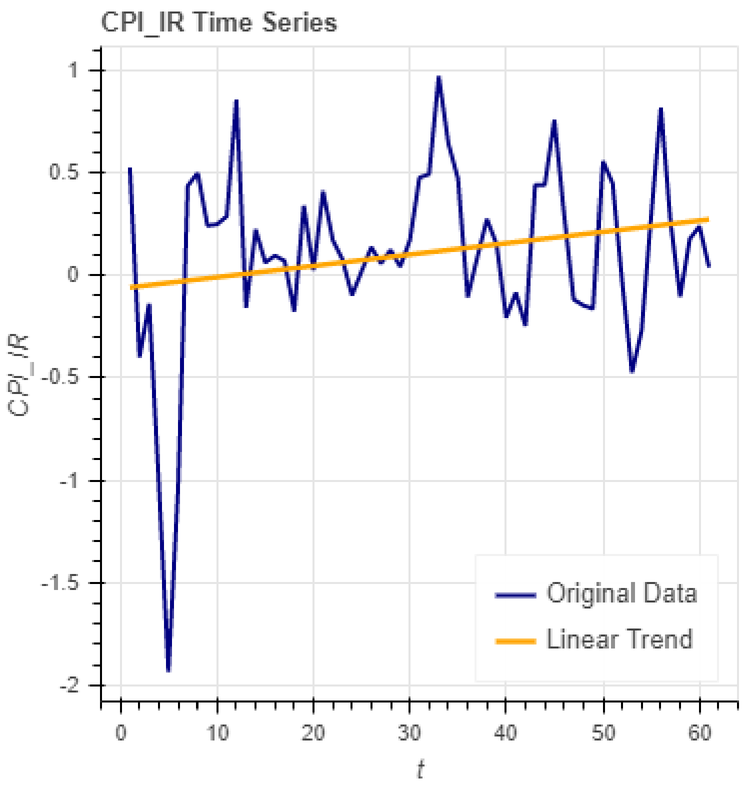
Solution:

I compute the monthly inflation rate from CPI by:

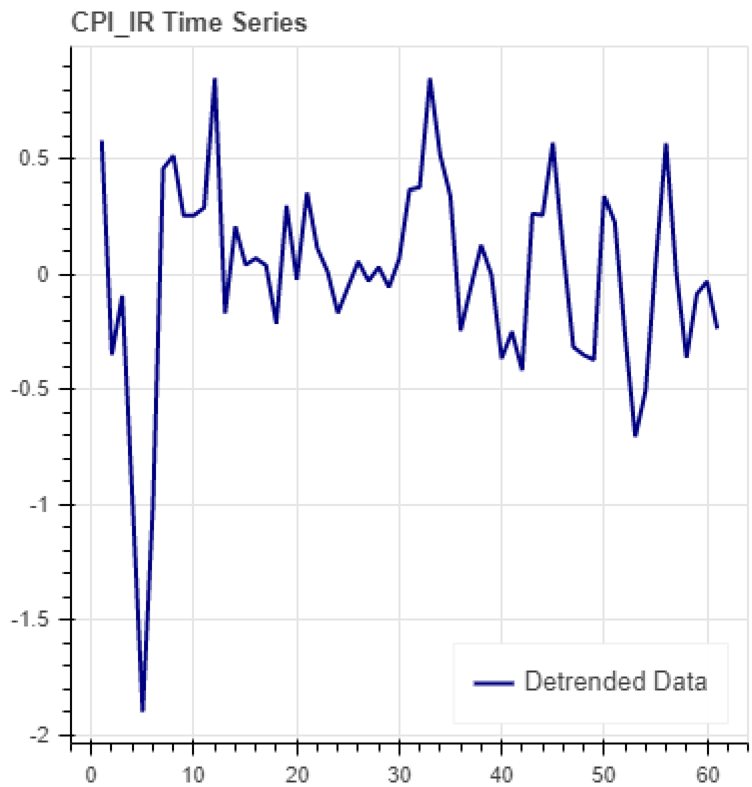
$$\mathrm{IR}_t = \ln(\mathrm{CPI}_t) - \ln(\mathrm{CPI}_{t-1})$$

where t indexes the months.

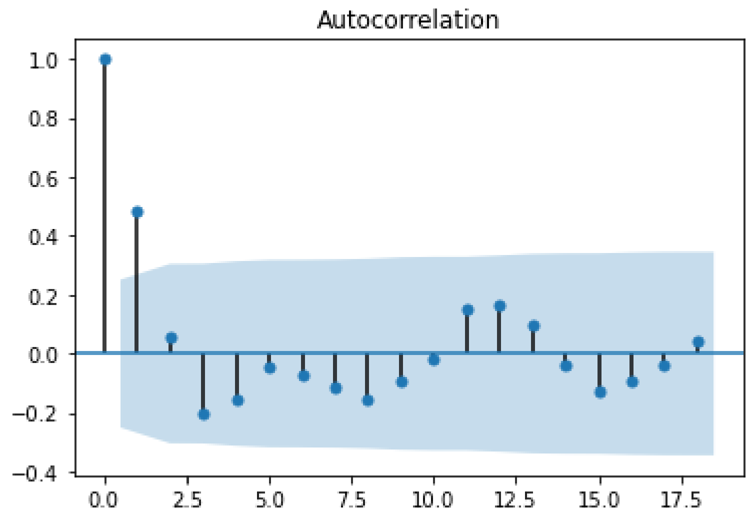
The linear trend is given by F(t) = 0.0055300123366067*t + (-0.06427097234449808)

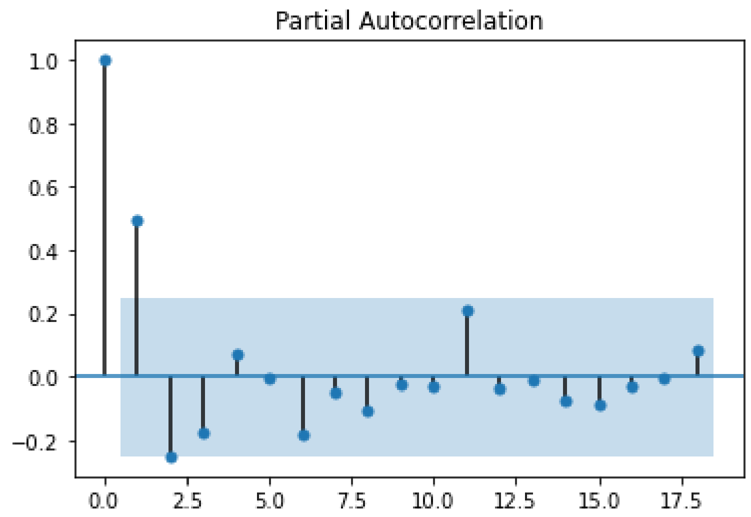


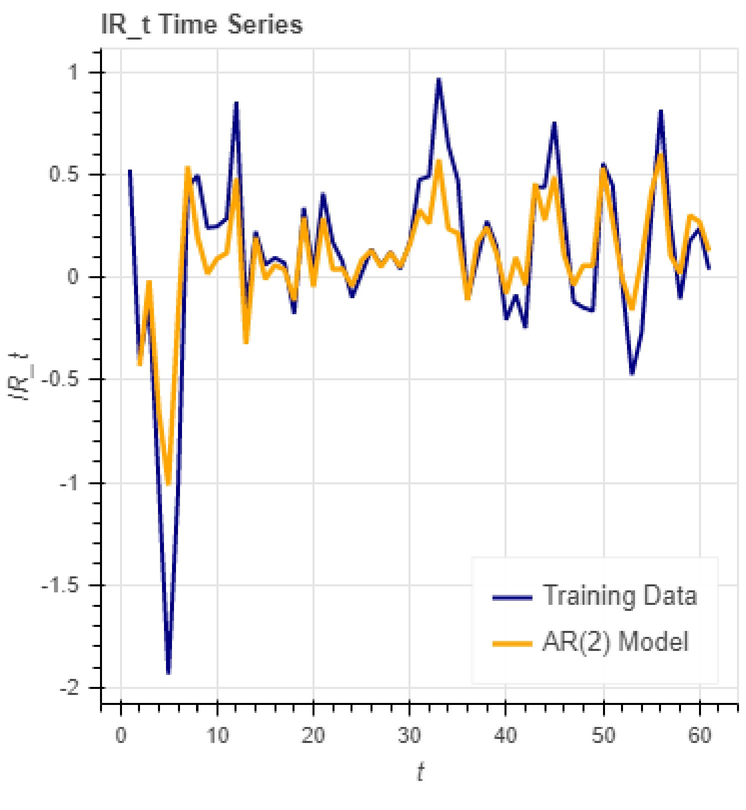
Subtract the linear growth trend from CPI_IR to get the detrended data plot.

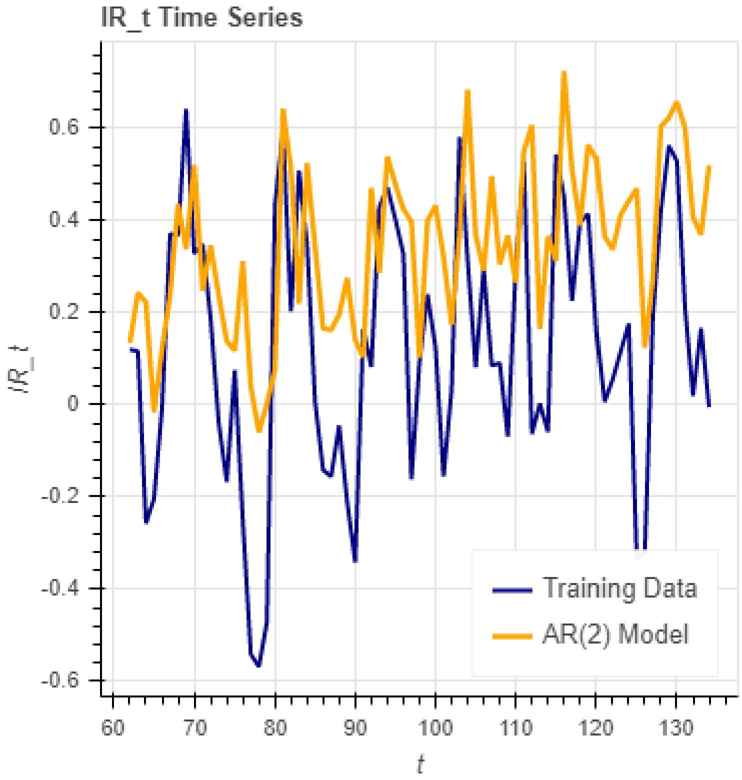


From the above PACF plot, we can see that the highest lag at which the plot extends beyond the statistically significant boundary is at lag 2. This indicates that an AR Model of lag 2 should be sufficient to fit the data.


Autocorrelation
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
0.0 2.5 5.0 7.5 10.0 12.5 15.0 17.5


Partial Autocorrelation
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
0.0 2.5 5.0 7.5 10.0 12.5 15.0 17.5

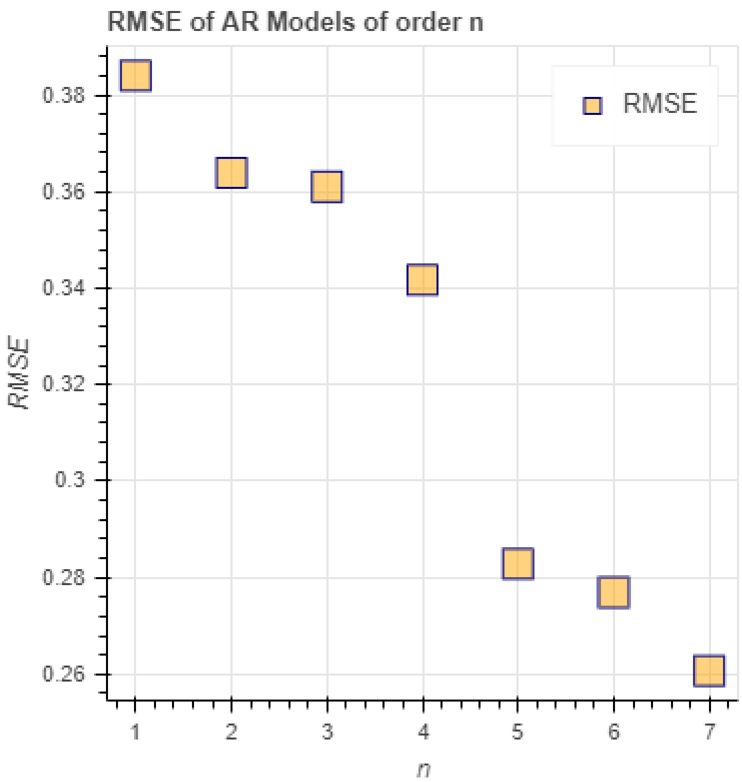

IR_t Time Series
IR_t
t
1
0.5
0
-0.5
-1
-1.5
-2
0 10 20 30 40 50 60
Training Data
AR(2) Model



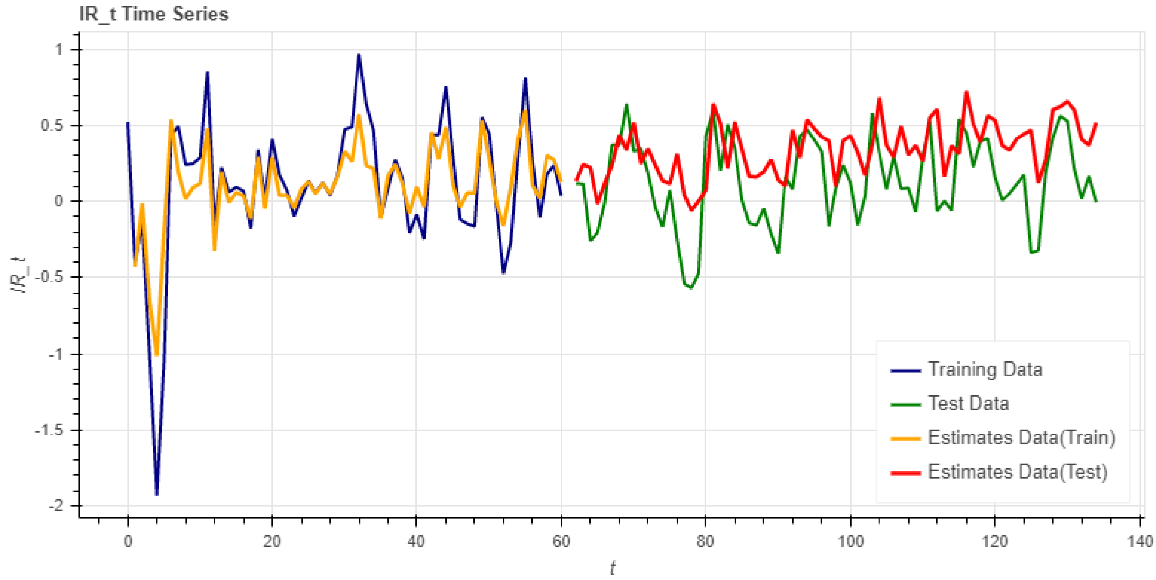
2. (3 points) Which AR(p) model gives the best predictions? Include a plot of the RSME against different lags p for the model.

Solution:

AR(7) model gives the best predictions. However, considering other factors, we still choose AR (2) model.



3. (3 points) Overlay your estimates of monthly inflation rates and plot them on the same graph to compare. (There should be 4 lines, one for each dataset, plus the predictions) over time (months from September 2013 onward).



Problem 6: External Regressors and Model Improvements

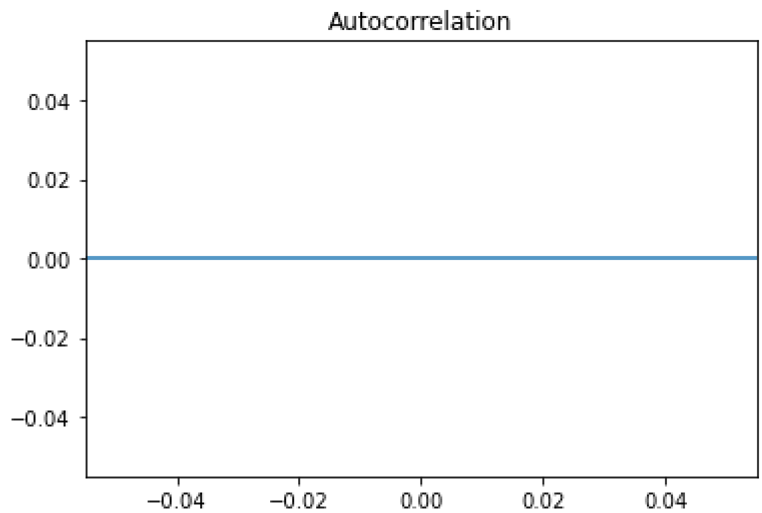
External Regressors

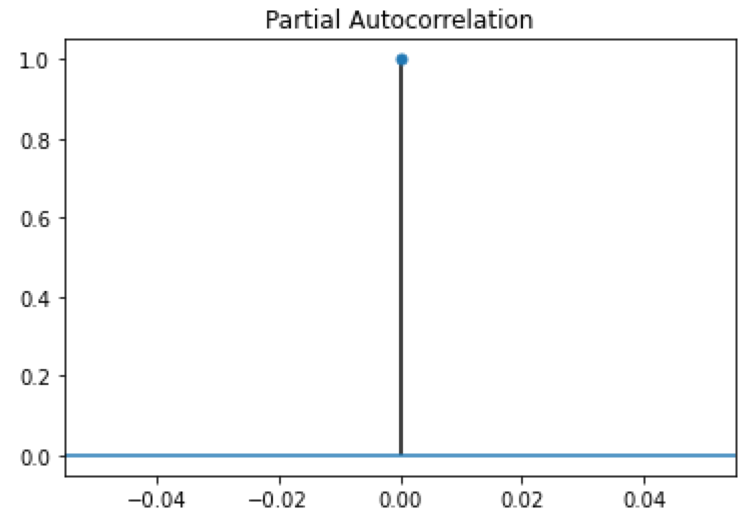
1. (4 points) Include as external regressors monthly average PriceStats inflation rate data and monthly BER data. Use cross-correlation plots to find the lag between the following:

- CPI inflation rate and PriceStats inflation rate
- CPI and BER inflation rate.

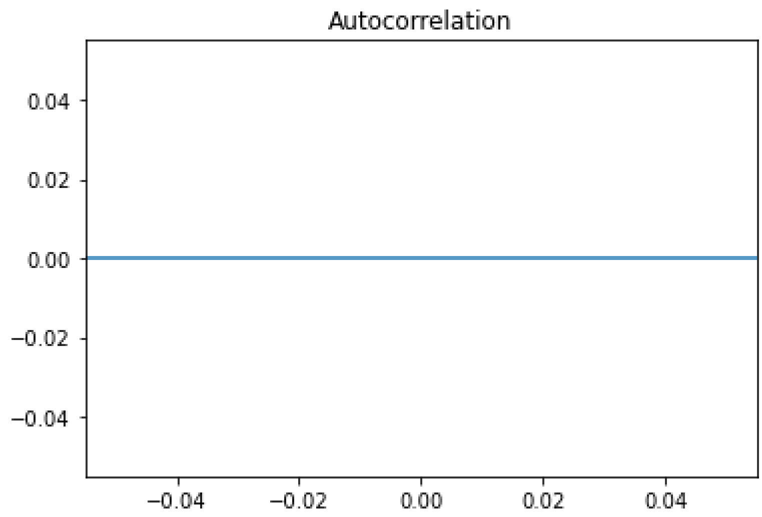
Solution:

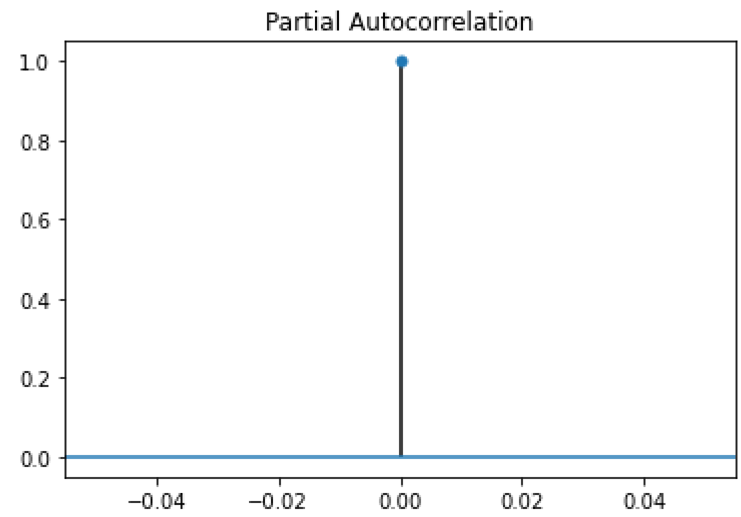
Cross-correlation plots between CPI inflation rate and PriceStats inflation rate: we can get the lag = 1.





Cross-correlation plots between CPI and BER inflation rate: we can get the lag = 1.



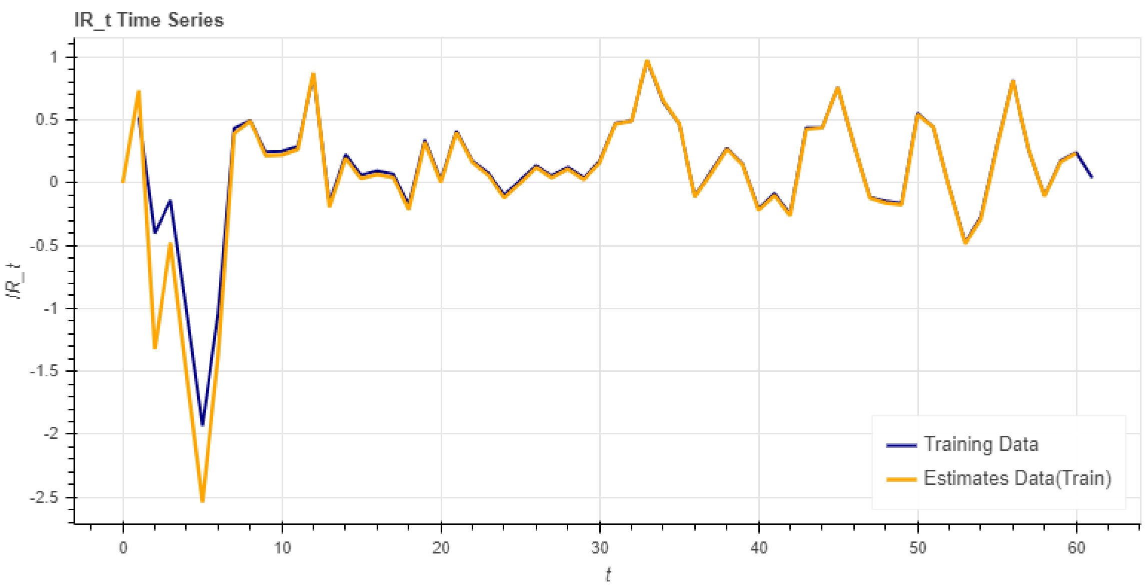


2. (3 points) Fit a new AR model to the CPI inflation rate with these external regressors and the most appropriate lag. Report the coefficients.

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

Solution:

```
import statsmodels.api as sm
import statsmodels.tsa.api as tsa
model1 = sm.tsa.statespace.SARIMAX(cpi_ir_train['diff_log'], order=(1,2,1),)
```



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

Solution:

Mean squared prediction error for 1 month ahead forecasts is 0.2552522327615316.

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 PriceStats and BER data as external regressors.

Solution:

Increasing training data will improve the accuracy of predictions