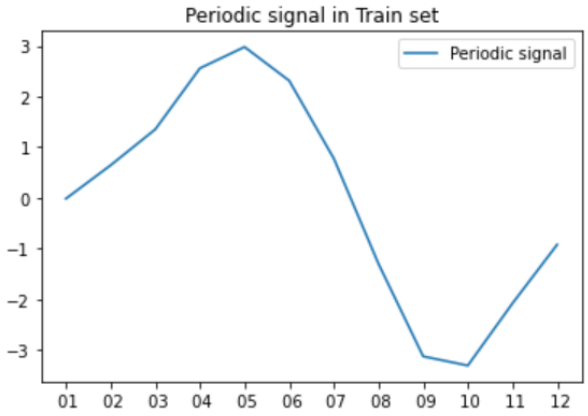
Written Report – 6.419x Module 4

Name: (luis_go95)

2. The Mauna Loa CO_2 Concentration

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

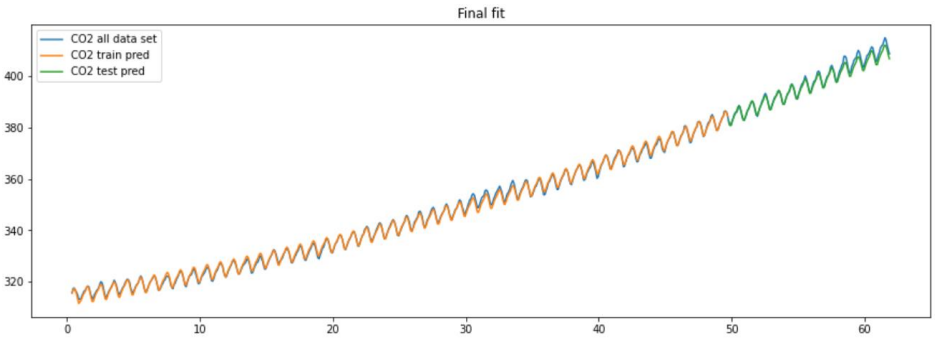
Solution:



A periodic signal is a signal that represents the behavior of a measure of each specific period. In the case of C02 concentration, the periodic signal is the behavior on average of C02 for each month throughout 1958 to 2007.

2. (2 points) Plot the final fit. 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:

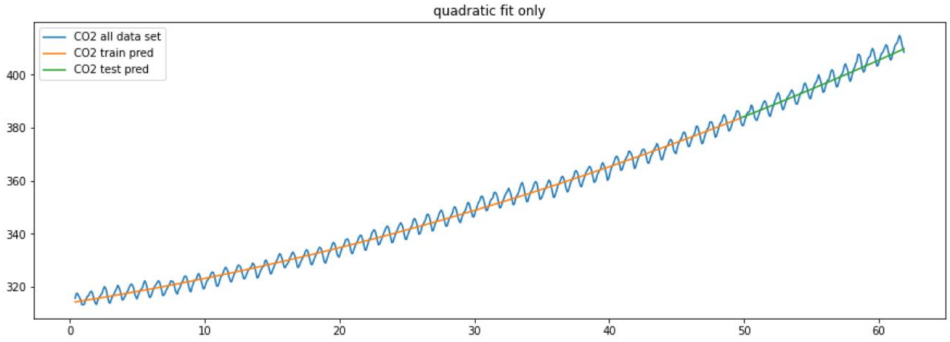


We can see that when combining quadratic and periodic signal, the final models seem to be close to the real values of CO2.

3. (4 points) Report the root mean squared prediction error and the mean absolute percentage error with respect to the test set for this final model. Is this an improvement over the previous model without the periodic signal? (Maximum 200 words.)

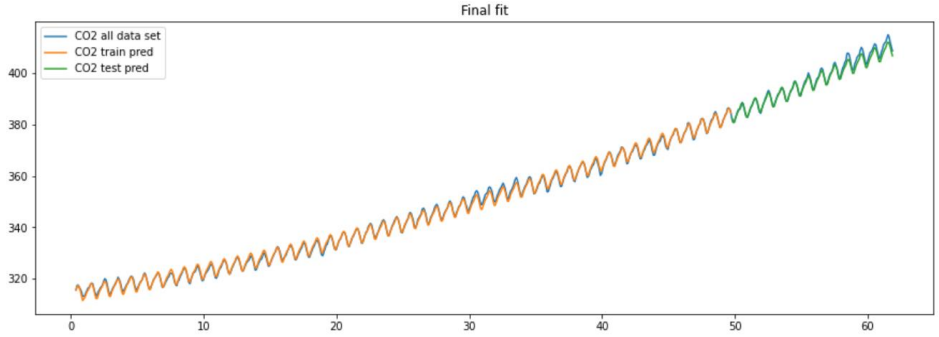
Solution:

Only quadratic model




```
RMSE: 2.5018943915296408
MAPE: 0.5327130356940849
```

Quadratic model + periodic signal




```
RMSE: 1.1385017014361456
MAPE: 0.20719330410998646
```

In the previous results we can clearly see that both visually and in metrics comparison, that the final fitted model is totally better. More concretely, the RMSE and MAPE reduce considerably after including the periodic signal. Honestly, the prediction is visually extremely like the original data.

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

Solution:

```
F (CO2_pred) = max: 409.7676592766418, min 314.27265404575314, range: 95.49500523088864
Pi (CO2_dtrend_avg) = max: 2.981335251839811, min -3.3108615787136877, amplitud: 3.1460984152767493
Ri (quadratic_residuals_without_Pi) = max: 2.955615723190952, min -1.7417010546004352, range: 4.697316777791388

Ratio range_F/amplitud_Pi = 30.353470434105393
Ratio amplitud_Pi/range_F = 0.03294516197648333
Ratio amplitud_Pi/range_Ri = 0.6697650092817459
```

In my opinion, the descomposition of the variaton of the CO concentration is meaninful because the ration “amplitud_Pi/range_Ri = 0.669” which suggest that Residuals are well represented in the average Pi.

3. Autocovariance Functions (Written Report)

(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 Xt. Include all important steps of your computations in your report.*

Solution:

$$\gamma(1) = Cov(W_t + \theta \cdot W_{t-1}, W_{t-1} + \theta \cdot W_{t-2})$$
$$\gamma(1) = \theta \cdot 1 \cdot Cov(W_{t-1}, W_{t-1})$$
$$\gamma(1) = \theta \cdot \sigma^2$$

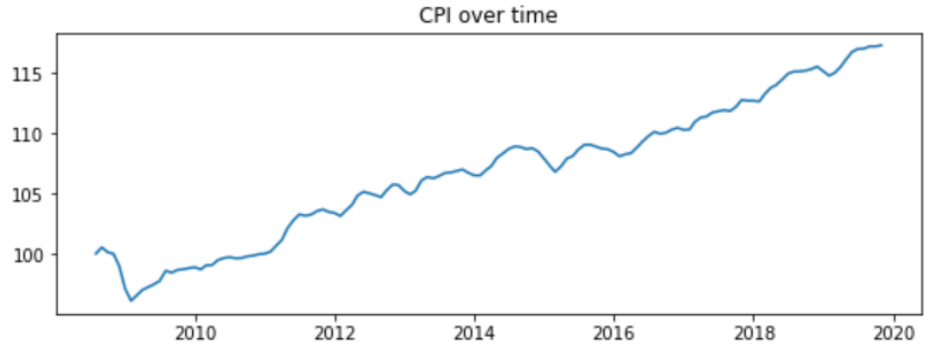
(4 points) Consider the AR(1) model, $X_t = \phi X_{t-1} + W_t,$ *, where* $\{W_t\} \sim W \sim \mathcal{N}(0, \sigma^2)$*, Suppose* $|\phi| < 1$*. Find the autocovariance function of Xt. (You may use, without proving, the fact that Xt is stationary if* $|\phi| < 1$*)*

Solution:

$$\gamma(1) = Cov(X_t, X_{t-1})$$
$$\gamma(1) = Cov(\phi \cdot X_{t-1} + W_t, X_{t-1})$$
$$\gamma(1) = \phi \cdot Cov(X_{t-1}, X_{t-1}) + Cov(Wt, X_{t-1})$$
$$\gamma(1) = \phi^1 \cdot \gamma(0)$$
$$\gamma(1) = \phi^1 \cdot \sigma^2$$

4. CPI and BER Data Analysis

Detrend CPI
Solution:

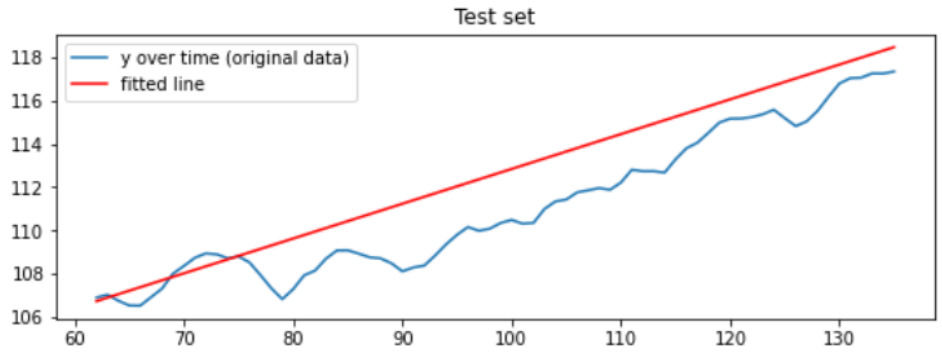


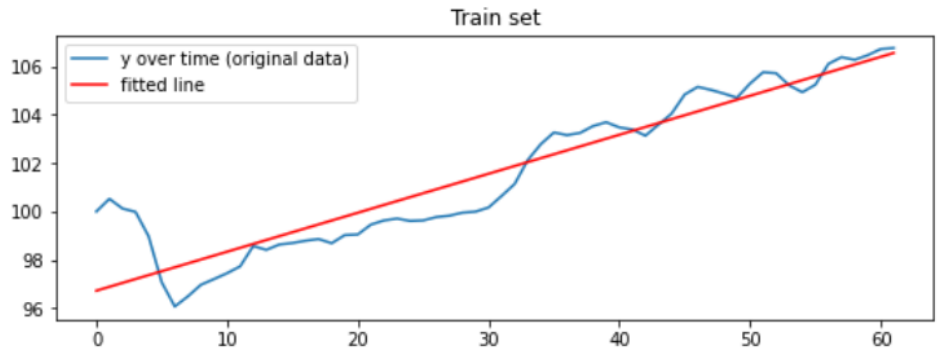
First, we plot the CPI over time for all the data and we can see that probably it may have a linear trend.

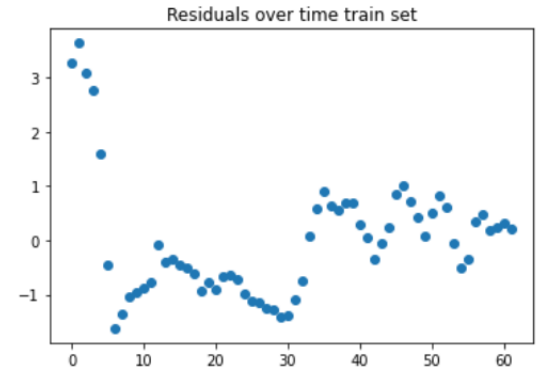
Detrend CPI with linear trend
Solution:

First, we split the data in train and test taking as reference September 2013 as the starting point for the testing set. Then, we fit different trends like linear, quadratic cubic, etc. And we determined that the liner trend is the one where we get a better RMSE and MAPE score in the testing set. As a result, we proceed in getting the linear residuals (y_train – prediction_train) that is equivalent to the "Detrended CPI"

```
Intercept (B0): 96.72932632872502
Weight (B1): 0.16104348366951224
RMSE: 1.8007900862687192
MAPE: 1.435202517195968
```





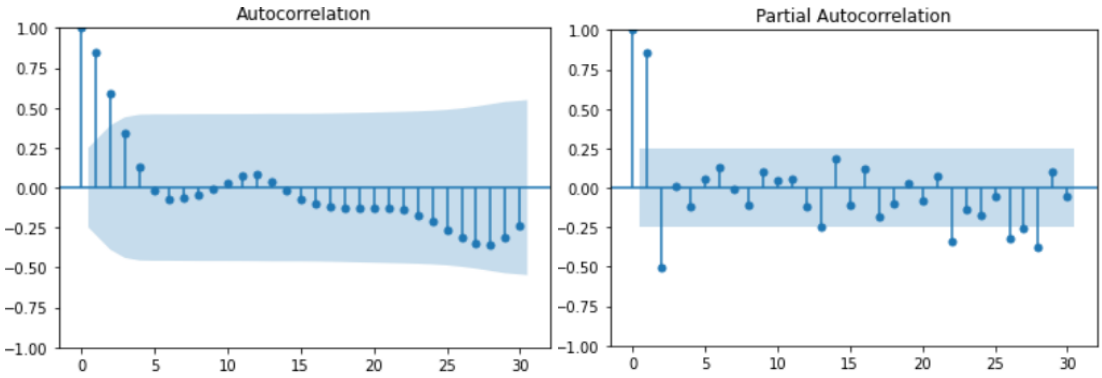



```
max(abs(linear_residuals))
```

```
3.634730187605456
```

AR Model: Determine the Lag

Solution:



From the partial autocovariance function we can see that an order p=2 for the AR model is good enough for fitting because it is the last significant value over all the possible lags plotted.

Find the Parameters for AR model, AR(2)

Solution:

We fit an AR model of order 2 using the linear residuals from previous exercise and get next results.

AutoReg Model Results

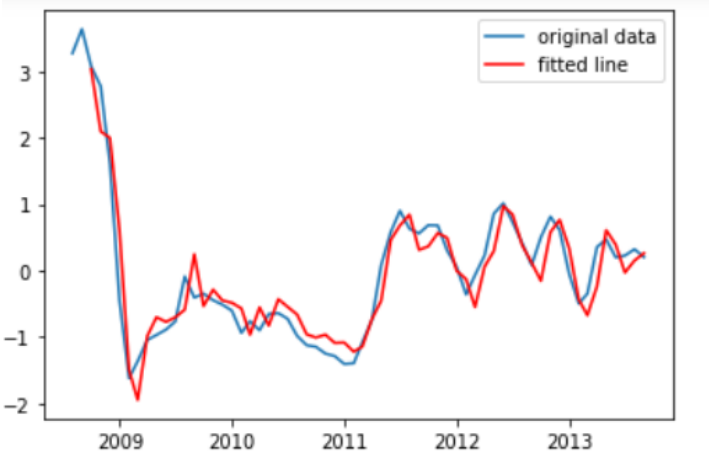
Dep. Variable:	CPI	No. Observations:	62
Model:	AutoReg(2)	Log Likelihood	-17.470
Method:	Conditional MLE	S.D. of innovations	0.324
Date:	Tue, 11 Apr 2023	AIC	42.939
Time:	23:40:04	BIC	51.317
Sample:	09-30-2008	HQIC	46.216
	- 08-31-2013		

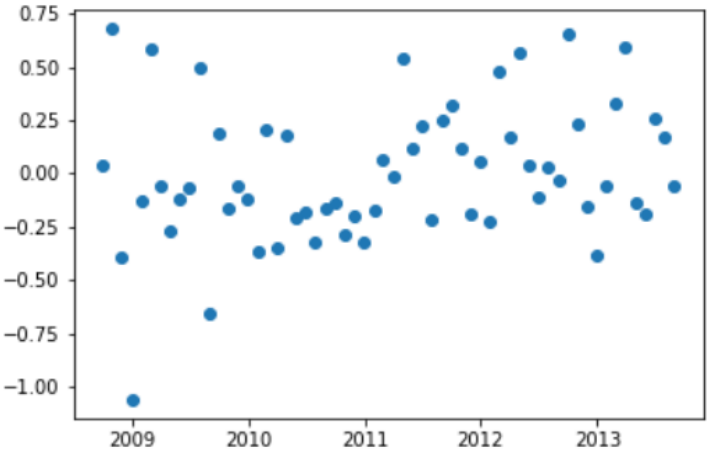
	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.0431	0.042	-1.024	0.306	-0.126	0.039
CPI.L1	1.3237	0.098	13.461	0.000	1.131	1.516
CPI.L2	-0.5308	0.091	-5.824	0.000	-0.709	-0.352

Roots

	Real	Imaginary	Modulus	Frequency
AR.1	1.2469	-0.5738j	1.3726	-0.0686
AR.2	1.2469	+0.5738j	1.3726	0.0686

Then we predict values for the train data and plot it with the original data





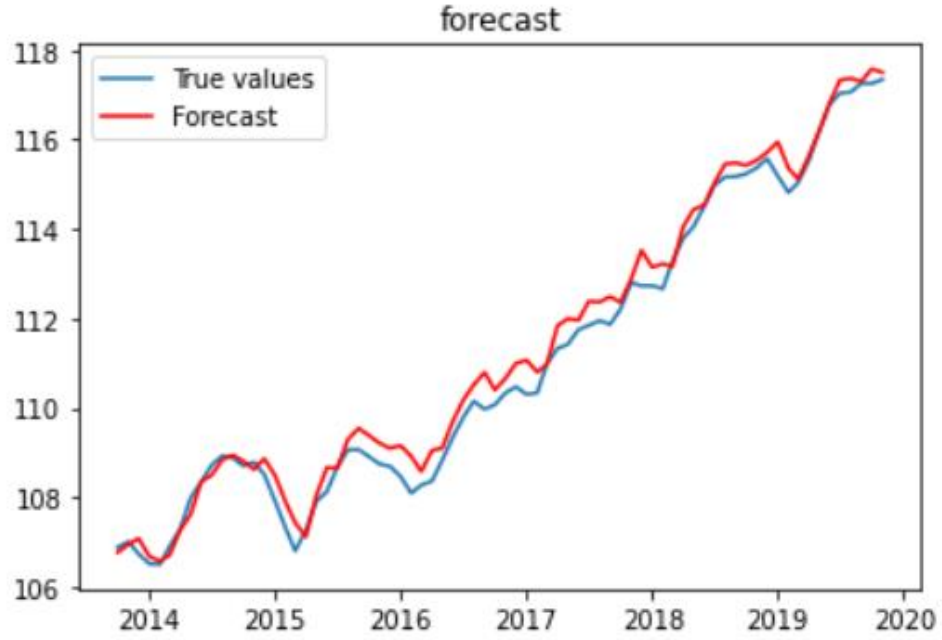
Forecast

Solution:

For this part, we want to get a forecast by using all the dataset, the linear model and the AR model that we fitted in our training set. In order to perform this, first with our fitted linear model, we predict all the values for the entire dataset. Second, we obtain linear residuals (y – y_pred) for the entire dataset. Third, we use the coefficients we get when fitting our AR(2) model to get the predictions manually:

AR_pred = -0.0431 + 1.3237*linear_residuals.shift(1) - 0.5308*linear_residuals.shift(2)

And for the final prediction, we sum our "AR_pred + Linear_pred" and plot only the values from September 2013 and so on.




```
RMSE: 0.39100025178623227
```

5. Converting to Inflation Rates

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

(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.)

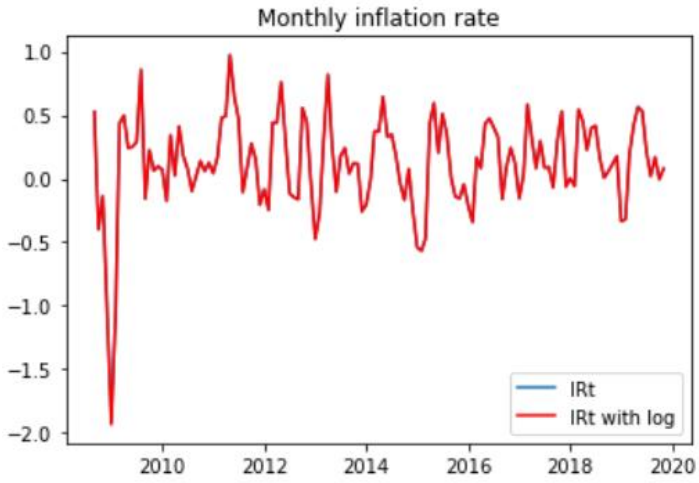
Solution:

By Following the suggested formula of the statement, we get the IRt with log, and IRt without log:

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

$$\mathrm{IR}_t = \frac{\mathrm{CPI}_t - \mathrm{CPI}_{t-1}}{\mathrm{CPI}_{t-1}}$$

```
df_cpi["diff"] = (y - y.shift(1))/y.shift(1)*100
df_cpi["diff_ln"] = (np.log(y) - np.log(y.shift(1)))*100
```

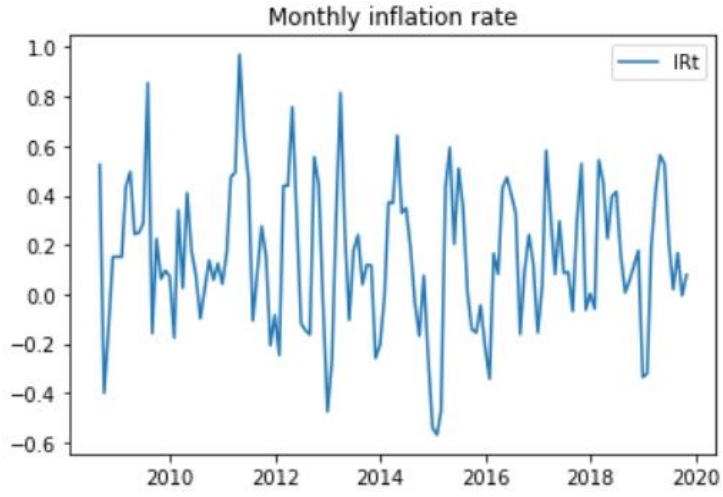


As a result, we can validate that both approaches give approximately the same result. But as the statement suggests, we will use IRt with log.

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

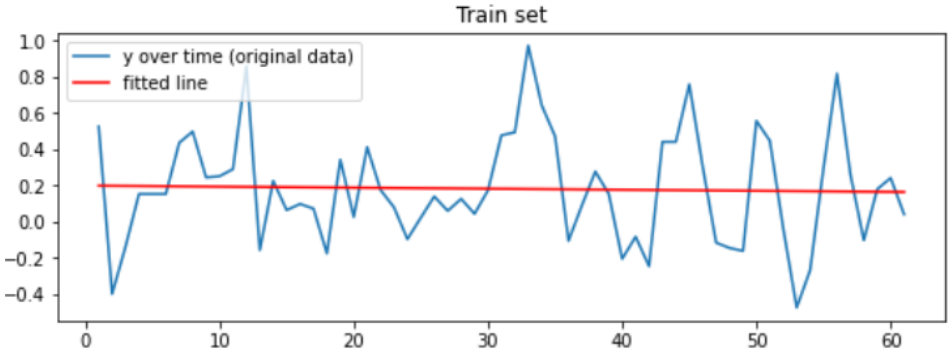
Solution:

We clearly see that there are important outliers that can have a big impact in the next steps. Therefore, we proceed by identifying the outliers that are outside 3 standard deviations and impute them with a trimmed mean when not including them. After these steps, we get next plot:



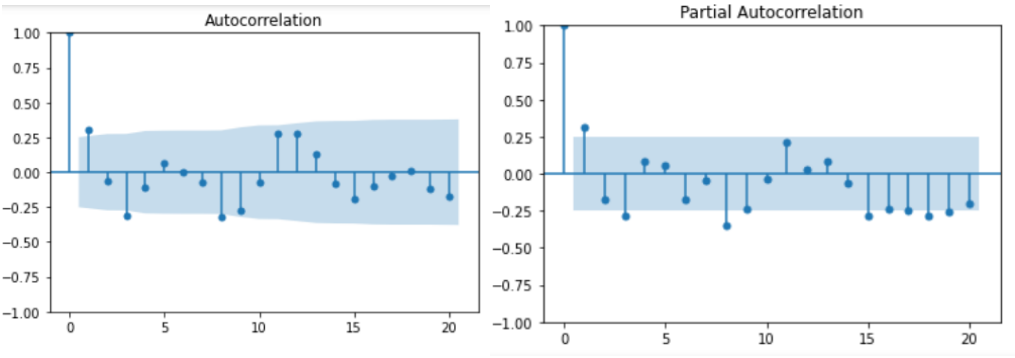
From this graph we can see that probably we do not have a trend, and this could be because of the step of transforming the CPIt to IRt. We are applying a kind of differencing and therefore is like a way of removing trending already. Also, if we fit a linear regression on the data, we get so low coefficients that suggest us that it is not necessary to fit one

```
Intercept (B0): 0.19789036138550314
Weight (B1): -0.0005773312405061398
```



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

Solution:



From the ACF and PACF we can see that probably a good order for the AR model could be 2 or 1. Asa result, I decided to test with both values and decide the correct one by comparing the AIC and BIC values.

P = 1

AutoReg Model Results

Dep. Variable:	diff_ln	No. Observations:	61
Model:	AutoReg(1)	Log Likelihood	-10.037
Method:	Conditional MLE	S.D. of innovations	0.286
Date:	Sun, 16 Apr 2023	AIC	26.074
Time:	12:58:05	BIC	32.357
Sample:	09-30-2008	HQIC	28.532
	- 08-31-2013		

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.1182	0.043	2.745	0.006	0.034	0.203
diff_ln.L1	0.3073	0.122	2.527	0.012	0.069	0.546

P = 2

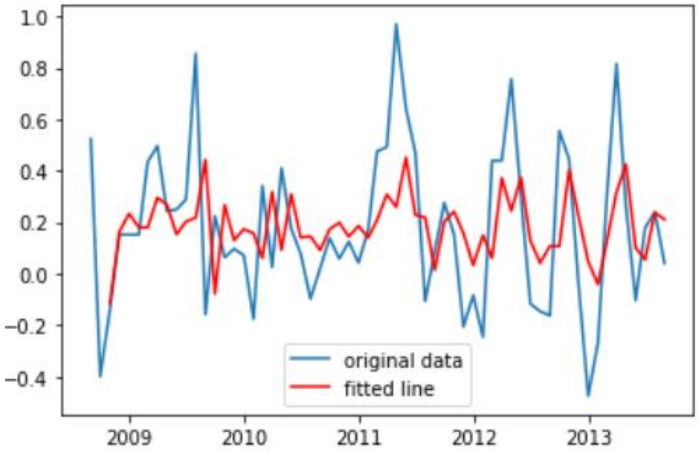
AutoReg Model Results

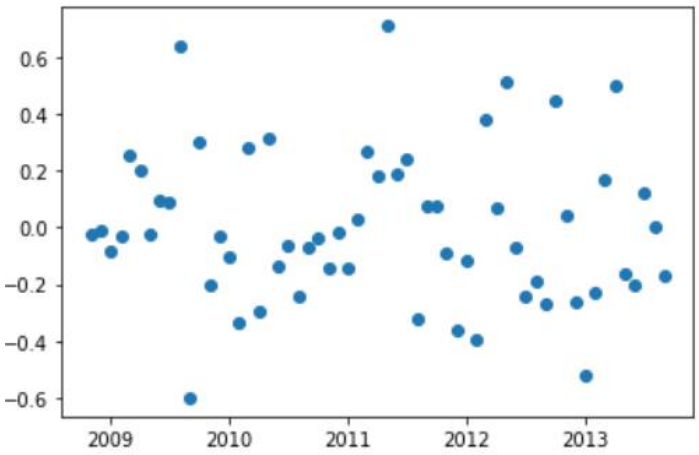
Dep. Variable:	diff_ln	No. Observations:	61
Model:	AutoReg(2)	Log Likelihood	-6.163
Method:	Conditional MLE	S.D. of innovations	0.269
Date:	Sun, 16 Apr 2023	AIC	20.325
Time:	12:58:02	BIC	28.635
Sample:	10-31-2008	HQIC	23.569
	- 08-31-2013		

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.1454	0.043	3.366	0.001	0.061	0.230
diff_ln.L1	0.4095	0.122	3.370	0.001	0.171	0.648
diff_ln.L2	-0.1859	0.120	-1.546	0.122	-0.422	0.050

Roots

As we can see, it seems that with P = 2 we get better values for AIC and BIC. As a result, we chose order 2 for fitting the AR model.





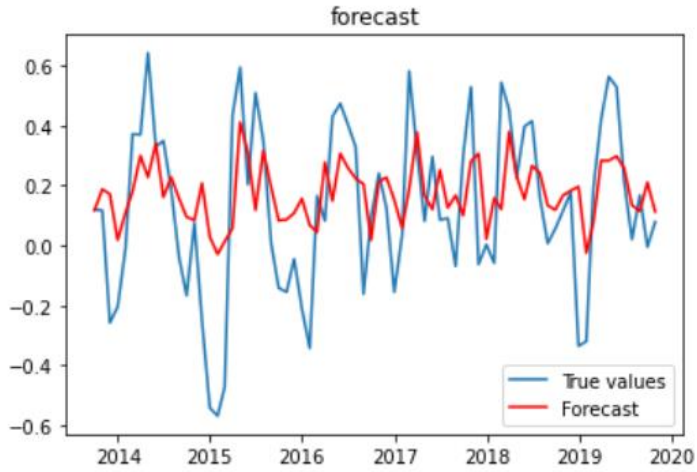
In previous graphs we can see the fitted line of the AR(2) model in contrast with the real data in the training set. Besides, we can see the residuals plot

(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:

In this case, the predictions were only made by using the AR model coefficients for the entire dataset and then filter for only the validation set:

AR_pred = 0.1454 + 0.4095*diff_ln.shift(1) - 0.1859*diff_ln".shift(2)



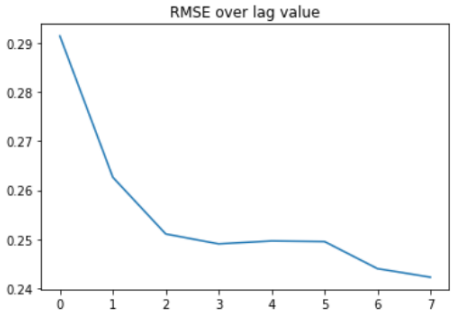
(3 points) Which model gives the best predictions? Include a plot of the RMSE against different lags p for the model.

Solution:

```
_, _, _, _, _, rmse_matrix = grid_search_AR(df_IRt.diff_ln, range(8), verbose=True)
```

```
Minimizing RMSE order:  7
Minimizing AIC order:  3
Minimizing BIC order:  2
matrix of RMSE: [0.29135387 0.26262506 0.25108902 0.24906809 0.24967054 0.24954504
 0.24403769 0.24229423]
matrix of AIC [54.14489401 27.95205346 17.83954556 17.63212192 20.20726818 22.013898
 18.19453453 18.34203399]
Matrix of BIC [59.95544357 36.64557285 29.40094207 32.04613154 37.45845212 42.08663915
 41.07303376 44.01030637]
```

If we perform a grid search and save the value of RMSE, AIC, BIC, we can have the best order value for each metric. If we consider the AIC value, it seems that p = 3 is the best one, however, if we analyze BIC, it suggests that order 2 is the best one.

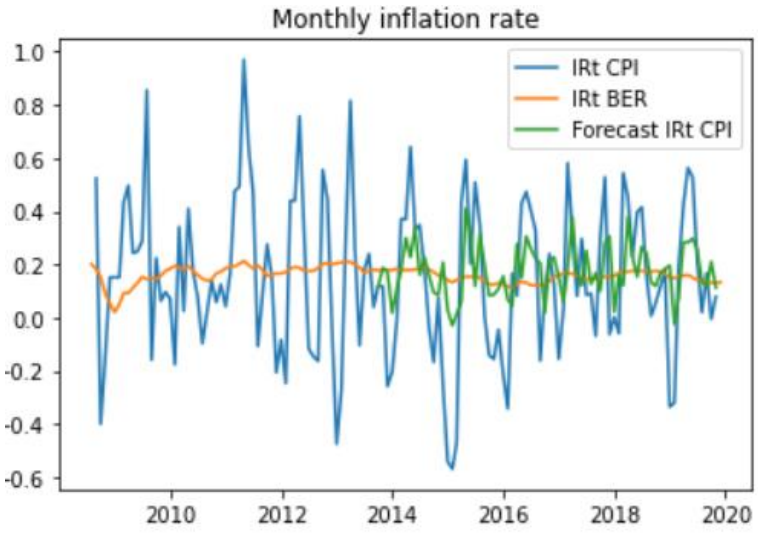


In contrast, if we only consider RMSE over different values of p, we can see that as a higher value we have, the RMSE is better. That's probably why consider RMSE alone is not the best option because as

a more complex model (higher p) we are more prone to overfit and not necessary get the best order for AR model. As a result, in my opinion, I would get p=2 taking into consideration the BIC value and visualizing the ACF and PACF plots.

(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.)

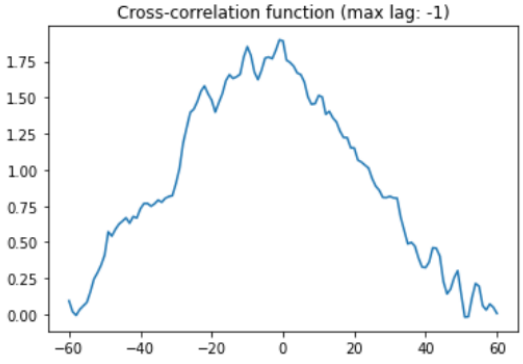
Solution:



6. External Regressors and Model Improvements (Written Report)

(4 points) Plot the cross-correlation function between the CPI and BER inflation rate, by which find r, 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.)

Solution:



As we can observe in the cross-correlation function, the suggested lag between two inflation rates is –1. This means that the BER value of the previous month is most correlated with the current month value of CPI.

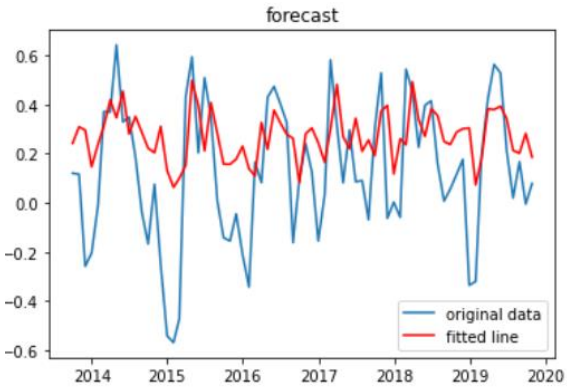
(3 points) Fit a new model to the inflation rate with these external regressors and the most appropriate lag. Report the coefficients and plot the 1 month-ahead forecasts for the validation data. In your plot, overlay predictions on top of the data.

Solution:

SARIMAX Results

Dep. Variable:	diff_ln	No. Observations:	61
Model:	SARIMAX(2, 0, 0)	Log Likelihood	-9.217
Date:	Mon, 17 Apr 2023	AIC	26.434
Time:	22:52:34	BIC	34.877
Sample:	08-31-2008 - 08-31-2013	HQIC	29.743
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
BER_M	1.0716	0.270	3.970	0.000	0.543	1.601
ar.L1	0.3963	0.130	3.050	0.002	0.142	0.651
ar.L2	-0.1913	0.142	-1.349	0.177	-0.469	0.087
sigma2	0.0790	0.014	5.486	0.000	0.051	0.107



After fitting the model, we get the coefficients for predicting the test set by next formula:

Pred = 0.3963*diff_ln.shift(1) - 0.1913*diff_ln.shift(2) + 1.0716*ber.lag1 + 0.0790

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

Solution:

`RMSE: 0.2771791370089299`

(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 data as external regressors.

Solution:

The approach I decided to take was to not work with exogenous variables and iterate over a range of ARIMA models. After performing grid search, and with the comparison of AIC and BIC, the best options to test were next ones:

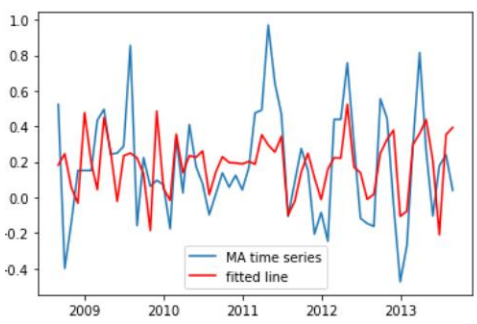
```
Minimizing AIC order:  (1, 0, 3)
Minimizing BIC order:  (0, 0, 3)
```

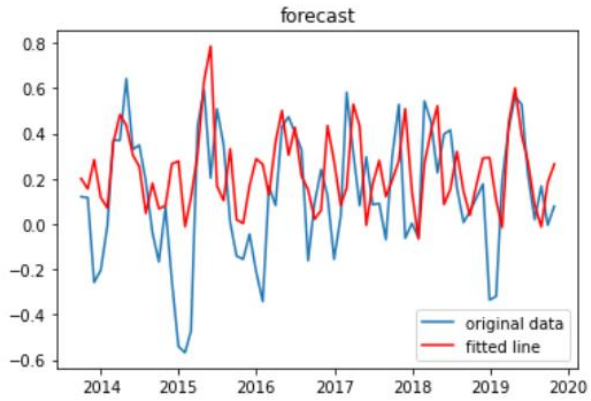
(0, 0, 3)

SARIMAX Results

Dep. Variable:	diff_ln	No. Observations:	61
Model:	ARIMA(0, 0, 3)	Log Likelihood	-4.738
Date:	Tue, 18 Apr 2023	AIC	19.476
Time:	20:12:41	BIC	30.030
Sample:	08-31-2008	HQIC	23.612
	- 08-31-2013		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.1821	0.034	5.434	0.000	0.116	0.248
ma.L1	0.3186	0.115	2.767	0.006	0.093	0.544
ma.L2	0.1760	0.126	1.395	0.163	-0.071	0.423
ma.L3	-0.5768	0.172	-3.348	0.001	-0.914	-0.239
sigma2	0.0661	0.014	4.858	0.000	0.039	0.093






```
RMSE: 0.2888181229681028
```

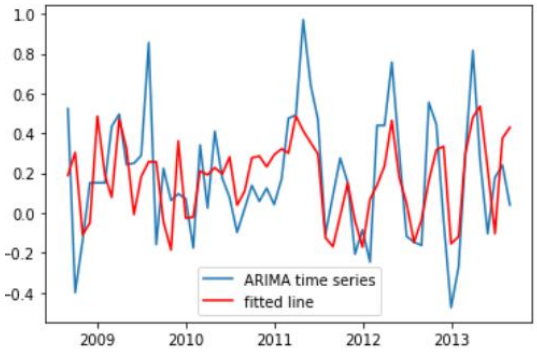
After fitting and measuring RMSE for test set, now we got a higher RMSE value and therefore, this model is good but is not better than the previous one with exogenous values.

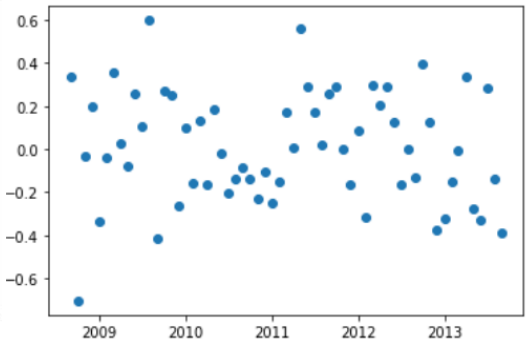
(1, 0, 3)

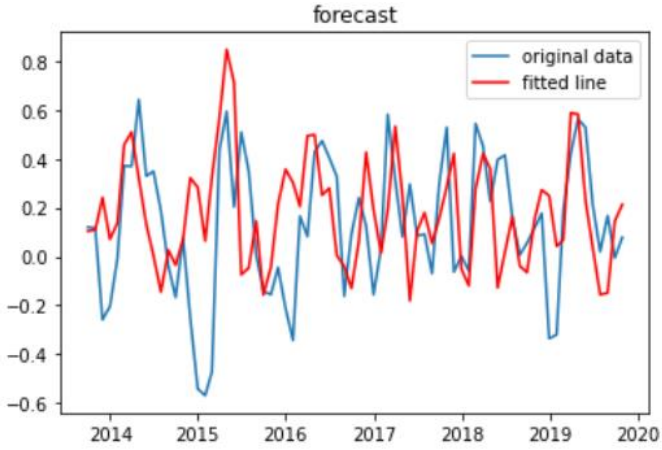
SARIMAX Results

Dep. Variable:	diff_ln	No. Observations:	61
Model:	ARIMA(1, 0, 3)	Log Likelihood	-2.922
Date:	Tue, 18 Apr 2023	AIC	17.844
Time:	20:55:37	BIC	30.509
Sample:	08-31-2008 - 08-31-2013	HQIC	22.808
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.1889	0.011	17.479	0.000	0.168	0.210
ar.L1	0.4580	0.168	2.733	0.006	0.130	0.787
ma.L1	-0.1544	2.163	-0.071	0.943	-4.394	4.086
ma.L2	-0.1275	1.841	-0.069	0.945	-3.735	3.480
ma.L3	-0.7147	1.561	-0.458	0.647	-3.775	2.346
sigma2	0.0605	0.132	0.457	0.648	-0.199	0.320







RMSE: 0.3139276061568122

After fitting and measuring the RMSE for the testing set, we again got a higher RMSE value and therefore, the model with the exogenous values is still the best one. As a result, the best model that I got what the initial one with BER as exogenous values with

RMSE: 0.2771791370089299