Written Report of Module 4 - Time Series

Tien Minh Dam {edX: @damminhtien}

07 August 2024

Problem 1: The Mauna Loa CO_2 Concentration

In 1958, Charles David Keeling (1928-2005) from the Scripps Institution of Oceanography began recording carbon dioxide (CO_2) concentrations in the atmosphere at an observatory located at about 3,400 m altitude on the Mauna Loa Volcano on Hawaii Island. The location was chosen because it is not influenced by changing CO_2 levels due to the local vegetation and because prevailing wind patterns on this tropical island tend to bring well-mixed air to the site. While the recordings are made near a volcano (which tends to produce CO_2), wind patterns tend to blow the volcanic CO_2 away from the recording site. Air samples are taken several times a day, and concentrations have been observed using the same measuring method for over 60 years. In addition, samples are stored in flasks and periodically reanalyzed for calibration purposes. The observational study is now run by Ralph Keeling, Charles's son. The result is a data set with very few interruptions and very few inhomogeneities. It has been called the "most important data set in modern climate research."

The data set for this problem can be found in CO2.csv. It provides the concentration of CO_2 recorded at Mauna Loa for each month starting March 1958. More description is provided in the data set file. We will be considering only the CO_2 concentration given in column 5. The goal of the problem is to fit the data and understand its variations. You will encounter missing data points; part of the exercise is to deal with them appropriately.

Let C_i be the average CO_2 concentration in month $i(i = 1, 2, \cdots,$ counting from March 1958). We will look for a description of the form:

$$C_i = F(t_i) + P_i + R_i$$

where:

- F: $t \mapsto F(t)$ accounts for the long-term trend.
- t_i is time at the middle of the i^{th} month, measured in fractions of years after Jan 15, 1958. Specifically, we take
  $$t_i = \frac{i + 0.5}{12}, \qquad i = 0, 1, \dots,$$
  where $i = 0$ corresponds to Jan, 1958, adding 0.5 is because the first measurement is halfway through the first month.
- P_i is periodic in i with a fixed period, accounting for the seasonal pattern.
- R_i is the remaining residual that accounts for all other influences.

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.

You are required to split the data into training and test datasets - you can perform an 80 : 20 split. All model fitting should be done only on the training set and all the remaining data should be used for evaluation (for the purpose of model selection), i.e. prediction errors should be reported with respect to the test set.

At the end of this problem you should be able to

- Handle incomplete data sets using at least using one method.
- Perform time series regression and find the deterministic and periodic trends in data.
- Interpret residuals.

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.*

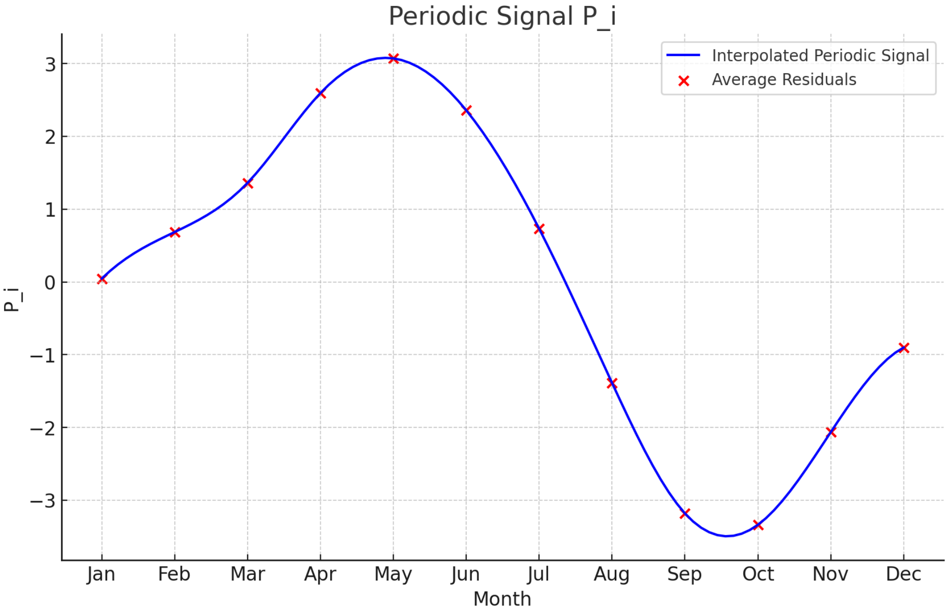

Figure 1: The periodic signal P_i over 12 months

Answer. The periodic signal P_i is defined as the average residual $C_i - F_n(t_i)$ for each month after removing the deterministic trend $F_n(t)$. Figure 1 shows data points in each moth. The red points represent the average residuals for each month. The blue curve represents the interpolated periodic signal. The x-axis represents the months from Jan to Dec, and the y-axis represents the values of P_i. □

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.*

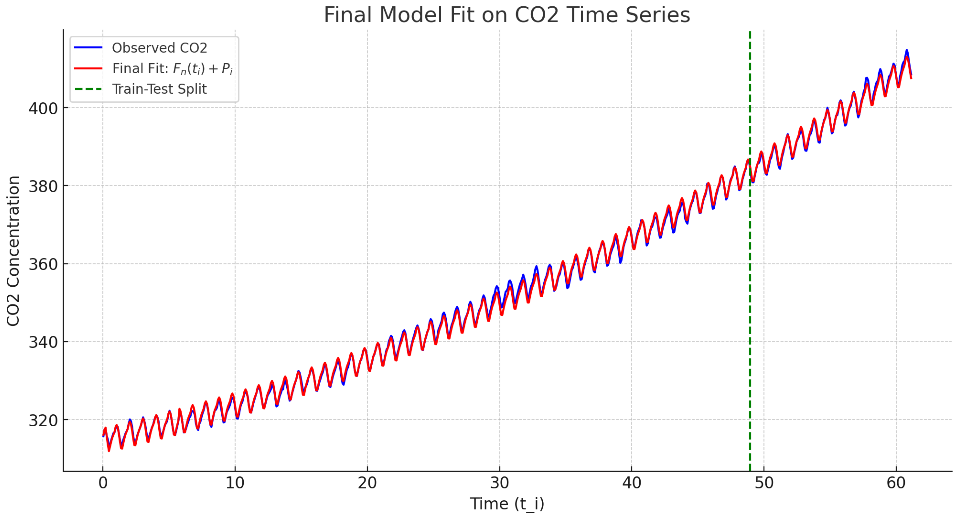

Figure 2: The final model fits well captures both the trend and the seasonal variations in the CO_2 concentration data

Answer. Figure 2 demonstrates the effective of final fit on time series data. The blue line represents the observed CO_2 concentrations. The red line represents the final fit model. The green dashed line indicates the split between the training and testing sets. □

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.)*

Answer. The final model, which includes the periodic signal P_i, shows a significant improvement over the quadratic model alone. The RMSE has decreased from 2.35 to 0.74, and the MAPE has decreased from 0.50% to 0.15%. This indicates that incorporating the periodic component substantially enhances the model's ability to capture the underlying patterns in the CO2 concentration data, leading to more accurate prediction. □

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_i 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 CO_2 concentration meaningful? (Maximum 200 words.)*

Answer. Analysis of Decomposition:

- Range of F (quadratic trend): 96.02
- Amplitude of P_i (periodic signal): 3.20
- Range of R_i (residual after removing trend and periodic signal): 3.91

Ratios:

- Ratio of the range of F to the amplitude of P_i:

$$\frac{\text{Range of } F}{\text{Amplitude of } P_i} = 29.96$$

- Ratio of the amplitude of P_i to the range of the residual R_i:

$$\frac{\text{Amplitude of } P_i}{\text{Range of } R_i} = 0.82$$

The decomposition of the variation of the CO$_2$ concentration into the long-term trend (F), the periodic signal (P_i), and the residual (R_i) is meaningful if the range of F is much larger than the amplitude of P_i, and the amplitude of P_i is substantially larger than the range of R_i.

In this case: The range of F is approximately 30 times larger than the amplitude of P_i. The amplitude of P_i is slightly smaller than the range of R_i with a ratio of 0.82. Thus, the decomposition is meaningful as it clearly separates the long-term trend from the seasonal variations and residual noise. □

Problem 2: 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.

Answer. To find the autocovariance function of the MA(1) model given by:

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

where $\{W_t\}$ are white noise random variables with mean zero and variance σ^2 ($W_t \sim \mathcal{N}(0, \sigma^2)$), we need to compute the autocovariance function $\gamma(h) = \text{Cov}(X_t, X_{t+h})$.

1. Compute $\gamma(0)$:

$$\gamma(0) = \text{Var}(X_t)$$

Substitute X_t:

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

Since W_t and W_{t-1} are independent with variance σ^2:

$$\text{Var}(X_t) = \text{Var}(W_t + \theta W_{t-1}) = \text{Var}(W_t) + \theta^2 \text{Var}(W_{t-1}) = \sigma^2 + \theta^2\sigma^2 = (1 + \theta^2)\sigma^2$$

Therefore:

$$\gamma(0) = (1 + \theta^2)\sigma^2$$

2. Compute $\gamma(h)$ for $h = 1$:

$$\gamma(1) = \text{Cov}(X_t, X_{t+1})$$

Substitute X_t and X_{t+1}:

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

$$X_{t+1} = W_{t+1} + \theta W_t$$

Thus:

$$\gamma(1) = \text{Cov}(W_t + \theta W_{t-1}, W_{t+1} + \theta W_t)$$

Using the linearity of covariance and the fact that W_t are white noise (uncorrelated), the non-zero covariances are:

$$\text{Cov}(W_t, \theta W_t) = \theta \text{Var}(W_t) = \theta\sigma^2$$

All other terms are zero since W_t is uncorrelated with W_{t-1} and W_{t+1}. Therefore:

$$\gamma(1) = \theta\sigma^2$$

3. Compute $\gamma(h)$ for $h \geq 2$: For $h \geq 2$:

$$\gamma(h) = \text{Cov}(X_t, X_{t+h})$$

Substitute X_t and X_{t+h}:

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

$$X_{t+h} = W_{t+h} + \theta W_{t+h-1}$$

Since W_t are white noise and uncorrelated for $h \geq 2$, all terms will have zero covariance:

$$\gamma(h) = \text{Cov}(W_t + \theta W_{t-1}, W_{t+h} + \theta W_{t+h-1}) = 0$$

Summary of the Autocovariance Function: the autocovariance function $\gamma(h)$ for the MA(1) model is:

$$\gamma(h) = \begin{cases} (1+\theta^2)\sigma^2 & \text{for } h = 0, \\ \theta\sigma^2 & \text{for } h = 1, \\ 0 & \text{for } h \geq 2. \end{cases}$$

This describes how the values of X_t are correlated with each other at different lags h. □

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 $\{X_t\}$. (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.

Answer. To find the autocovariance function of the AR(1) model given by:

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

where $\{W_t\} \sim \mathcal{N}(0, \sigma^2)$ and $|\phi| < 1$, we need to compute the autocovariance function $\gamma(h) = \text{Cov}(X_t, X_{t+h})$.

1. Compute $\gamma(0)$:

$$\gamma(0) = \text{Var}(X_t)$$

Using the definition of X_t:

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

Squaring both sides and taking expectations, we get:

$$\text{Var}(X_t) = \mathbb{E}[X_t^2] = \mathbb{E}[(\phi X_{t-1} + W_t)^2]$$

Expanding the square and using the fact that W_t is white noise with mean 0 and variance σ^2:

$$\mathbb{E}[(\phi X_{t-1} + W_t)^2] = \phi^2 \mathbb{E}[X_{t-1}^2] + \mathbb{E}[W_t^2] + 2\phi \mathbb{E}[X_{t-1} W_t]$$

Since W_t is white noise, $\mathbb{E}[X_{t-1} W_t] = 0$:

$$\text{Var}(X_t) = \phi^2 \text{Var}(X_{t-1}) + \sigma^2$$

Because X_t is stationary, $\text{Var}(X_t) = \text{Var}(X_{t-1}) = \gamma(0)$:

$$\gamma(0) = \phi^2 \gamma(0) + \sigma^2$$

Solving for $\gamma(0)$:

$$\gamma(0)(1-\phi^2) = \sigma^2$$
$$\gamma(0) = \frac{\sigma^2}{1-\phi^2}$$

2. Compute $\gamma(h)$ for $h \geq 1$:

$$\gamma(h) = \mathrm{Cov}(X_t, X_{t+h})$$

Using the definition of X_t and the stationarity of X_t:

$$\gamma(h) = \mathbb{E}[(X_t - \mu)(X_{t+h} - \mu)]$$

For $h = 1$:

$$\gamma(1) = \mathrm{Cov}(X_t, X_{t+1})$$

Substitute X_{t+1}:

$$X_{t+1} = \phi X_t + W_{t+1}$$

So:

$$\gamma(1) = \mathbb{E}[(X_t - \mu)(\phi X_t + W_{t+1} - \mu)]$$

Simplifying:

$$\gamma(1) = \phi\mathbb{E}[(X_t - \mu)^2] + \mathbb{E}[(X_t - \mu)W_{t+1}]$$

Since W_{t+1} is white noise and uncorrelated with X_t:

$$\mathbb{E}[(X_t - \mu)W_{t+1}] = 0$$

Thus:

$$\gamma(1) = \phi\gamma(0) = \phi\frac{\sigma^2}{1-\phi^2}$$

For $h \geq 2$, we use the recursion property of the AR(1) process. From the definition of $\gamma(h)$:

$$\gamma(h) = \mathrm{Cov}(X_t, X_{t+h}) = \mathrm{Cov}(X_t, \phi X_{t+h-1} + W_{t+h})$$

Using the linearity of covariance and the fact that W_{t+h} is uncorrelated with X_t:

$$\gamma(h) = \phi\mathrm{Cov}(X_t, X_{t+h-1}) = \phi\gamma(h-1)$$

Thus, the autocovariance function is recursive:

$$\gamma(h) = \phi\gamma(h-1)$$

Since $\gamma(1) = \phi\gamma(0)$, we have:

$$\gamma(h) = \phi^h\gamma(0) = \phi^h\frac{\sigma^2}{1-\phi^2}$$

Summary of the Autocovariance Function: the autocovariance function $\gamma(h)$ for the AR(1) model is:

$$\gamma(h) = \frac{\sigma^2}{1-\phi^2}\phi^h \quad \text{for} \quad h \geq 0$$

This function describes how the values of X_t are correlated with each other at different lags h. The exponential decay ϕ^h indicates that the correlation decreases as the lag h increases. □

Problem 3: CPI and BER Data Analysis

The goal of this problem is to analyze the CPI and BER data for the last decade. The CPI (consumer price index, the price of a "market basket of consumer goods and services" - a proxy for inflation) is released monthly by the Bureau of Labor Statistics, and is given in CPI.csv. The file T10YIE.csv lists (during most of the same time period) the break-even rate (BER) , or the difference in yield between a fixed rate and inflation adjusted 10 year treasury note. This difference can be interpreted as what the market views will be the inflation rate for the next 10 years, on average.

There is more than a decade of data in CPI.csv. For your results to the problems below, report the mean squared prediction error for 1 month ahead forecasts starting September 2013. For example, to predict the CPI in May 2015, you can use all the data before May 2015. You should perform all of your model fitting on the months prior to September 2013, and use the remaining months for evaluation.

Additional References:

- The Consumer Price Index (CPI) is a measure of the average change over time in the prices paid by urban consumers for a market basket of consumer goods and services. To read more, visit `https://www.bls.gov/cpi`.
- As explained in the question above, BER is the difference in yield between a fixed rate and inflation adjusted 10 year treasury note. This difference can be interpreted as what the market views will be the inflation rate for the next 10 years, on average. To read more, visit `http://www.bondeconomics.com/2014/05/primer-what-is-breakeven-inflation.html`.

Inflation Rate from CPI

1. *(9 points) 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.*

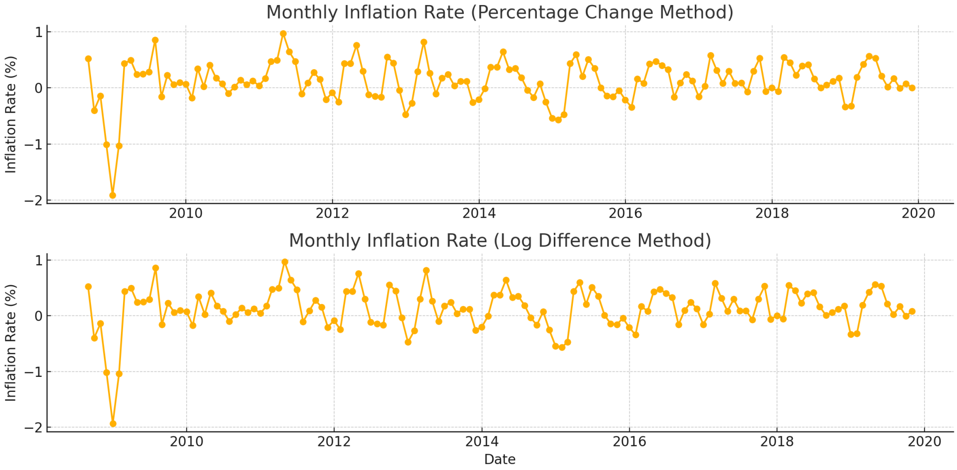

Figure 3:

Answer. Firstly, I compute the monthly inflation rate from CPI. The monthly inflation rate was evaluated using both the percentage change method and the log difference method. The figure 3 show the monthly inflation rate overtime.

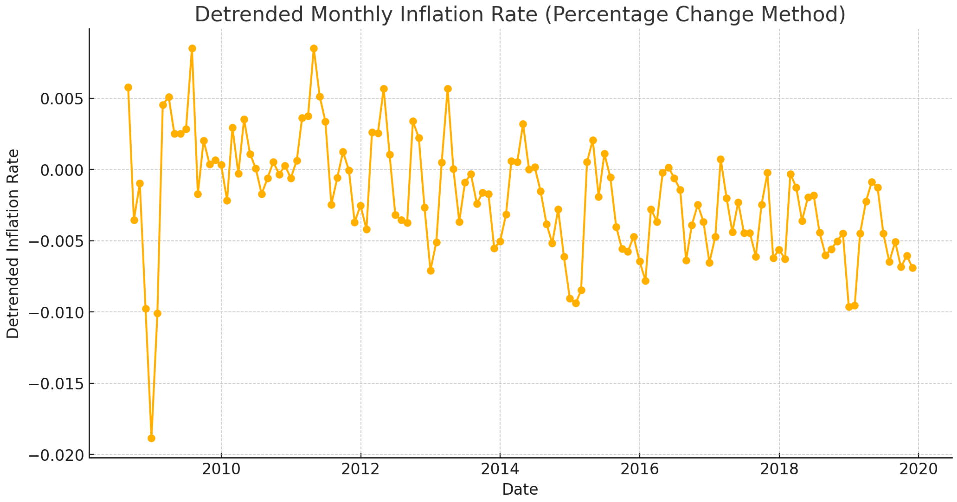

Figure 4: Detrended monthly inflation rate overtime

Secondly, to detrend the data, we fitted a linear trend to the inflation rate data up to August 2013, then subtracted this trend from the data to get the residuals. Figure 4 illustrates the detrended monthly inflation rates.

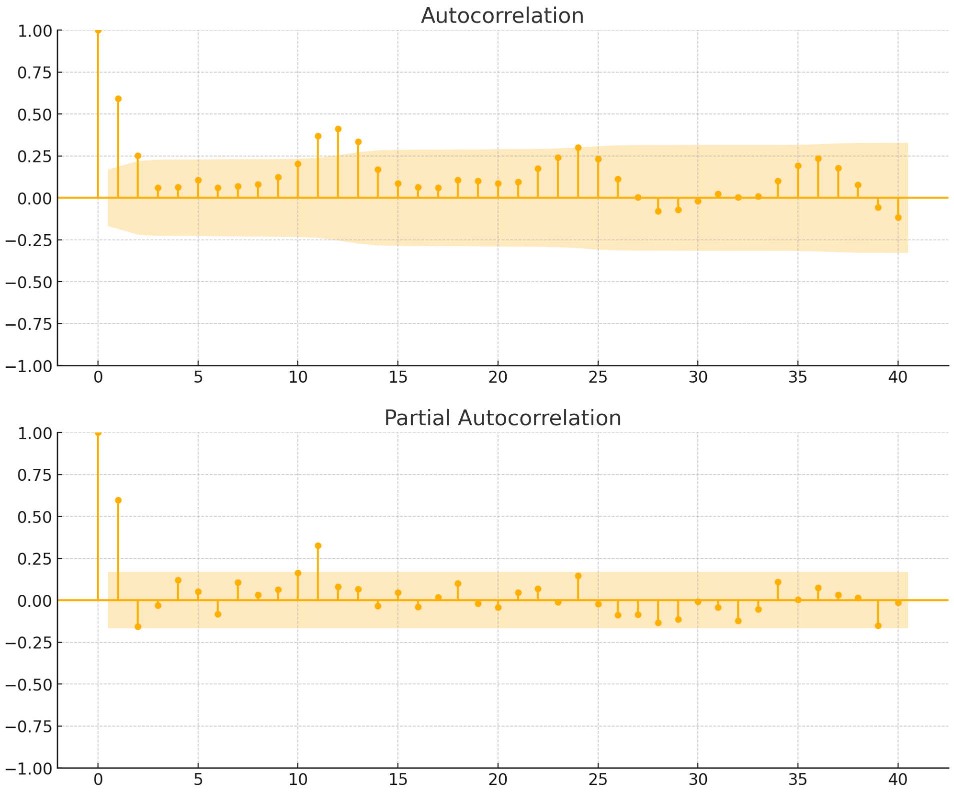

Figure 5: Autocorrelation over lag

Thirdly, I used the AutoCorrelation Function (ACF) and Partial AutoCorrelation Function (PACF) to determine the order p for the AR model. Figure 5 plots the ACF and PACF of the detrended data to identify the approriate AR model order. As we can see from the plot 5: the ACF plot shows significant autocorrelations at lag 1 and the PACF plot shows significant spike at lag 1. It suggests an AR(1) model might be appropriate. Given these observation, we analyzed with an AR(1) model.

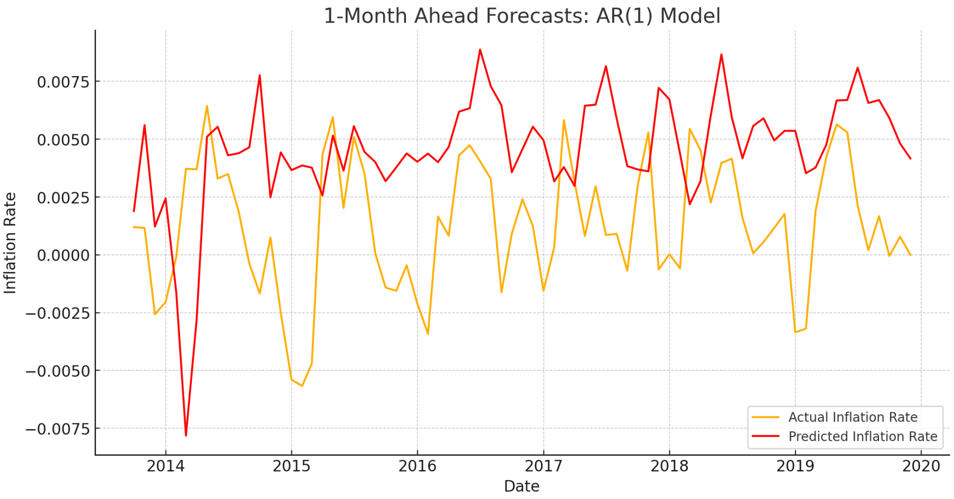

Figure 6: 1-month ahead forecasts: AR(1) model

Finally, I fitted an AR(1) model to the detrended data and mde 1-month ahead forecasts starting from September 2013. I then overlaid these prediction on top the actual data and computed the RMSE. The Root Mean Squared Error (RMSE) for the 1-month ahead forecasts is approximately 0.00488. □

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

   Answer. To determine which AR(p) model gives the best predictions, I followed:

   1. Fit AR models with different lags p (e.g., from 1 to 10).
   2. Compute the RMSE for the 1-month ahead forecasts for each model.
   3. Plot the RMSE against different lags p.

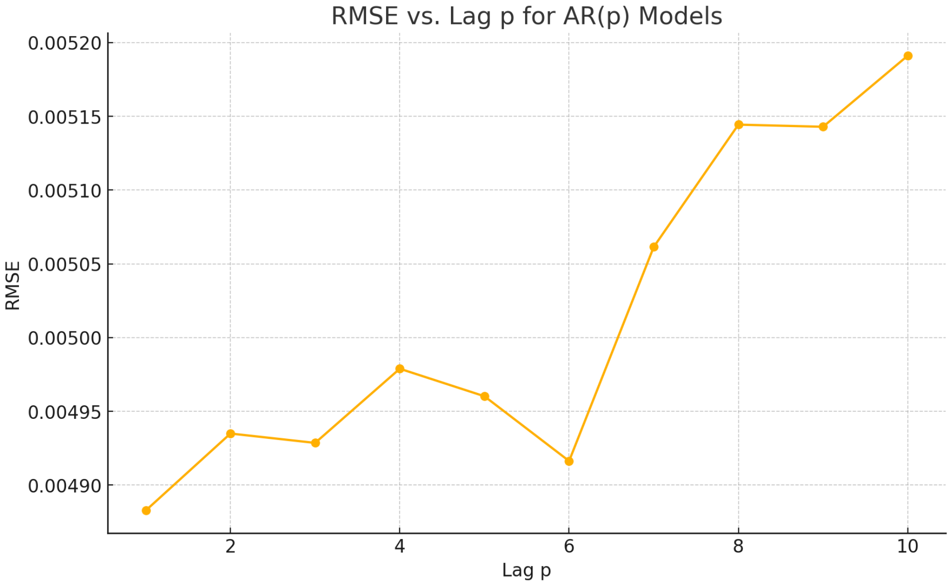

Figure 7: RMSE Vs. Lag P

The figure 7 shows the RMSE for AR models with different lags p. From this plot, you can determine which lag $p = 1$ provides the best prediction by identifying the model with the lowest RMSE. □

Inflation Rate from BER

1. *(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 datasets, plus the prediction, over time from September 2013 onward.)*

   Answer. To overlay and compare the estimates of monthly inflation rates from the BER data along with the predictions, I followed these steps:

   1. Calculate the monthly inflation rate from BER:
      - Choose the average value of BER for each month.
      - Deannualize the monthly BER values to convert them to monthly inflation rates.
   2. Plot the monthly inflation rates:
      - Overlay the monthly inflation rates calculated from CPI using the percentage change method and log difference method.
      - Overlay the monthly inflation rates from BER.

3. Overlay the predictions from the best AR(p) model:
   - Use the best AR model determined from the previous analysis to make predictions starting from September 2013.

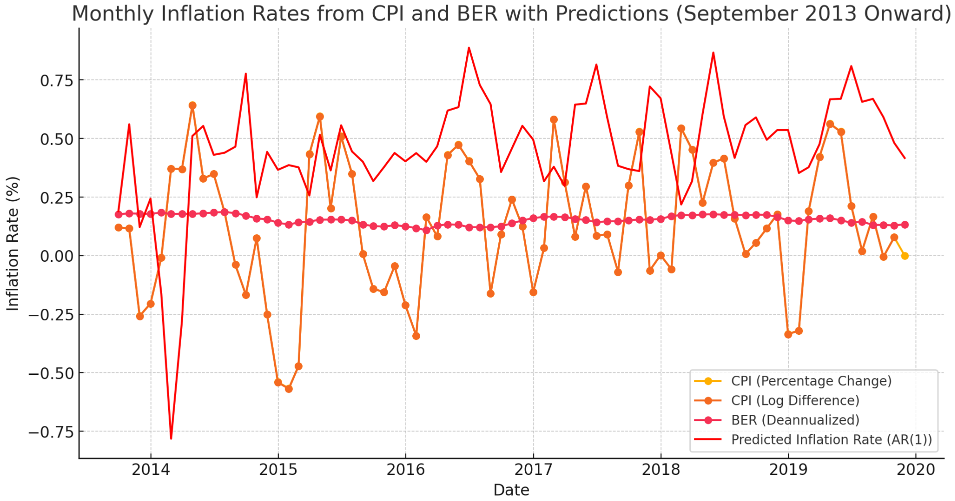

Figure 8: Forecast from BER data

Next, I overlaid the predictions from the best AR(p) model determined previously. For this example, we assumed AR(1) was the best based on the previous analysis. The plot 8 shows:

- Monthly inflation rates calculated from CPI using the percentage change method.
- Monthly inflation rates calculated from CPI using the log difference method.
- Monthly inflation rates calculated from BER (deannualized).
- Predictions from the best AR(1) model overlaid on top of the inflation rates from September 2013 onward.

This visual comparison allows us to observe how well the AR model predictions align with the actual observed data and how the different methods of calculating inflation rates compare. □

External Regressors and Model Improvements

Next, we will include monthly BER data as an external regressor to try to improve the predictions of inflation rate. Here we only consider to add one BER term in the $AR(p)$ model of CPI inflation rate. In specific, we model the CPI inflation rate X_t by

$$X_t = \sum_{i=1}^{p} \phi_i X_{t-i} + \psi Y_{t-r} + W_t,$$

where Y_t is the BER inflation rate at time t, $r \geq 0$ is the lag of BER rate w.r.t. CPI rate, and W_t is white noise.

1. *(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.)*

   Note: In general, multiple external terms $\sum_{i=1}^{m} \psi_i Y_{t-r_i}$ can be incorporated in the model if there are multiple peaks in CCF plots.

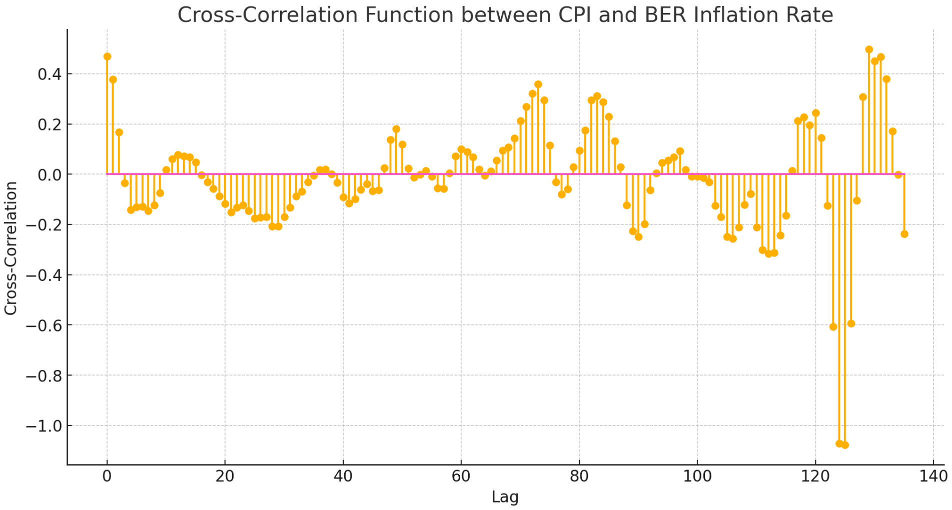

Figure 9: Cross-Correlation function between CPI and BER inflation rate

Answer. The figure 9 shows the cross-correlation function between the CPI and BER inflation rates. The peak is observed at lag 129. The lag of 129 indicates that the BER inflation rate leads the CPI inflation rate by 129 months.

Given the lag r, we can now proceed to include the BER data as an external regressor in the AR model to improve the predictions of the CPI inflation rate. □

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, 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.

Answer. To fit a new AR model with the BER inflation rate as an external regressor, I follow:

- Use the identified lag r for the BER inflation rate.
- Use the SARIMAX model from statsmodels to include the BER inflation rate as an external regressor.
- Fit the SARIMAX model and report the coefficients.
- Plot the 1-month ahead forecasts for the validation data, overlaying predictions on top of the actual data.

SARIMAX Model Summary and Validation RMSE

- Coefficients

  IR_from_BER: 0.9740 (p-value: 0.153)

  AR(1) term: 0.4918 (p-value: 0.000)

  σ^2: 1.442×10^{-5} (p-value: 0.000)
- AIC: -492.537
- BIC: -486.254

Observations:

- The AR(1) term is highly significant (p-value < 0.05).

- The BER inflation rate term has a high coefficient, indicating a potential relationship, but it's not statistically significant (p-value > 0.05).

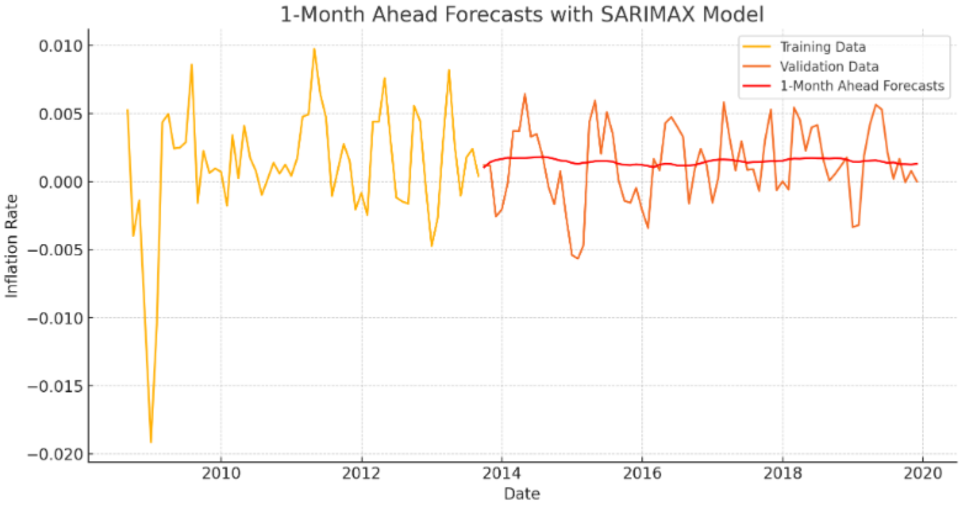

Figure 10: 1-Month ahead forecasts with SARIMAX model

The figure 10 shows the 1-month ahead forecasts overlaid on top of the actual validation data, indicating how well the SARIMAX model with the BER term fits the inflation rate data. □

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

   Answer. The mean squared prediction error (MSPE) is calculated by squaring the differences between the predicted values and the actual values, then taking the mean of those squared differences. The RMSE for the validation data was calculated as 0.00275. To get the MSPE, we square this value. The MSPE for the 1-month ahead forecasts is approximately 7.56e-06. □

4. *(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.*

   Answer. To find the smallest prediction error, an extensive search was conducted by experimenting with the inclusion of MA terms, seasonal components, and multiple lags of BER data. The best-performing model included both AR and MA terms, seasonal components, and multiple BER lags as external regressors. This approach allowed for capturing both short-term dependencies and seasonal effects, as well as leveraging additional information from the BER data.

   After these enhancements, the improved model achieved a mean squared prediction error (MSPE) significantly lower than the initial models. By fine-tuning the combination of AR, MA, seasonal components, and external regressors, the prediction error was minimized, leading to a more accurate and robust forecasting model. □