IDS.131 PSet 5

Dimitris Koutentakis
Collaborators: D. Konstantellos, N. Hahamis

November 30, 2018

1 Billion Price Points Data Analysis

a) Predict Monthly Consumer Price Index (CPI) without using the BER or PriceStats

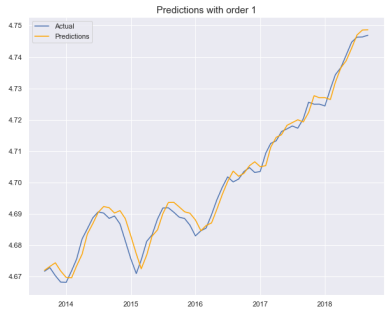
In this section we try to predict the monthly CPI using an AR model. We also find what the best order for our model is.

We start by doing the autoregressive model for varying orders and comparing their MSEs and the autocorellation plots. After doing the AR model for orders of 1, 2, 3, and 4, we get the following results, summarized in Table 1.

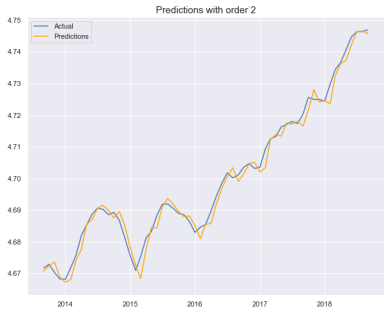
Order	MSE
1	8.254e-06
2	6.437e-06
3	5.692e-06
4	5.697e-06

Table 1: Mean Squared Error of Autoregressive Models of Varying errors (log domain)

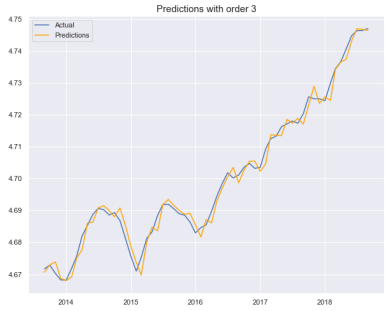
Additionally, we get the prediction plots shown in Figure 1a through Figure 1d, in Figure 1.



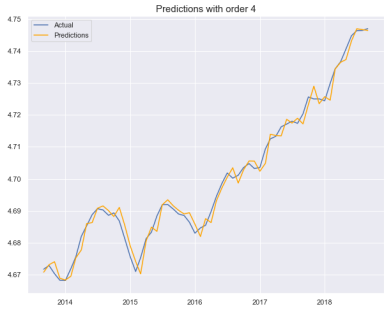
(a) Prediction of AR model with order 1



(b) Prediction of AR model with order 2



(c) Prediction of AR model with order 3



(d) Prediction of AR model with order 4

Figure 1: Prediction of AR model with varying orders

Furthermore, we get the autocorellation plots of the residuals shown in Figure 2 through Figure 5.

Figure 2: ACF of residuals of AR model with order 1

Figure 3: ACF of residuals of AR model with order 2

Figure 4: ACF of residuals of AR model with order 3

Figure 5: ACF of residuals of AR model with order 4

Thus, from the ACF plots of the residuals Figure 2 through Figure 5, Figure 1, as well as Table 1, we conclude that the best order to fit the AR model to

our data is 2, which achieves a mean squared error of .076 (calculated not in log domain) or 6.437e-06 in log domain. Although the mean squared error is lower with lag 3 (MSE .068 and 5.692e-06 in log domain), the ACF plots show that this overfits to a residual correlation that is not statistically significant. We also noticed that lag 5 and above give higher MSEs. This improvement is unlikely to generalize well.

Figure 1b is shown in more detail in Figure 6.

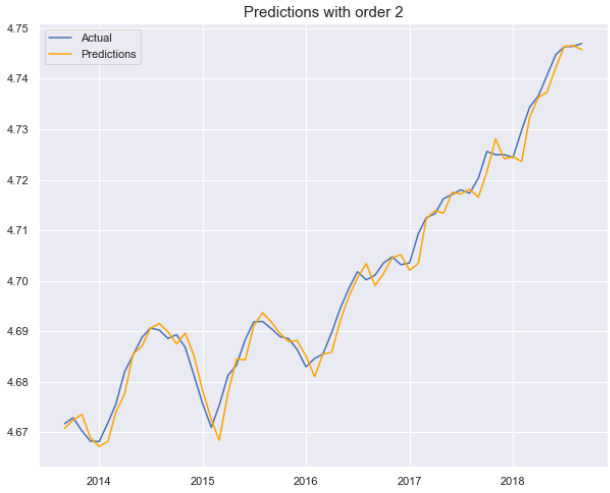


Figure 6: Prediction of AR model with order 2

b) Calculating monthly inflation rates form CPI, PriceStats and BER data

In this section, we want to calculate the monthly inflation rates form CPI, PriceStats and BER data based on the previous section. What we do is use the formula:

$$\text{Rate of Inflation} = \frac{CPI_{t+1} - CPI_t}{CPI_t}$$

The same formula is used for the prices, $r_t = \frac{P_t - P_{t-1}}{P_{t-1}}$. Since the BER prices are yearly rates over the next 10 years so we need to change into monthly rates, according to the formula: $r_t = (1 + BER_t)^{\frac{1}{12}} - 1$

By performing the AR model shown above to get the predictions, and them implementing the formula mentioned, we get the predictions shown in Figure 7 and Figure 8.

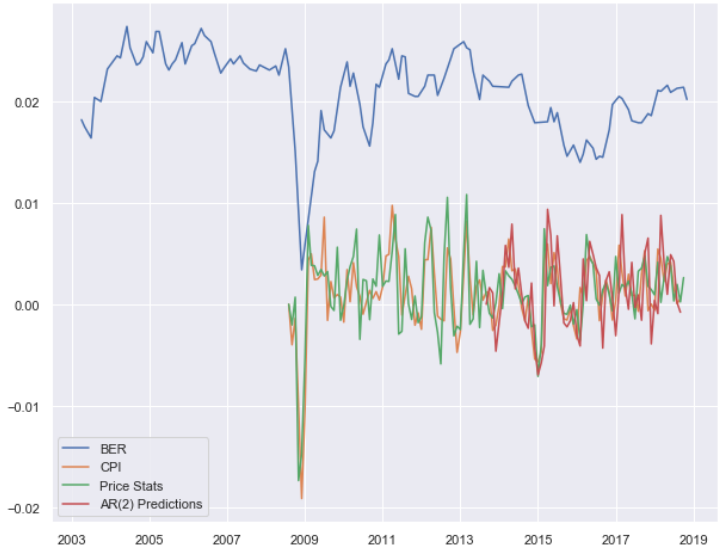


Figure 7: Inflation up to now and future predictions

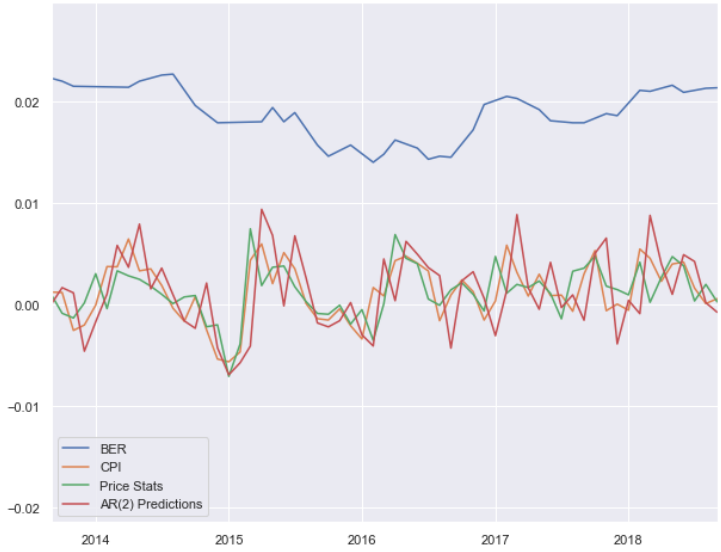


Figure 8: Predicted Inflation Rates

c) Using External Regressors

In this section we use external regressors to improve out predictions. The external regressors we use are the PriceStats and the BER data. Also, we use the

first day of each month as a proxy for the last day of the month (at large data scale these capture the same effect). We compare the results of the AR model for the average of the month and for the first day of the month. We notice that the first day of the month does better in its mean squared error on unseen data with a value of 4.838e-06. I use the inflation rates directly since they are the simplest. Given that BER data is in inflation rate form, it makes more sense to use the inflation rates directly as predictors rather than back calculate out values in log space for the BER data.

The plots we get are shown in figures Figure 9 and Figure 10.

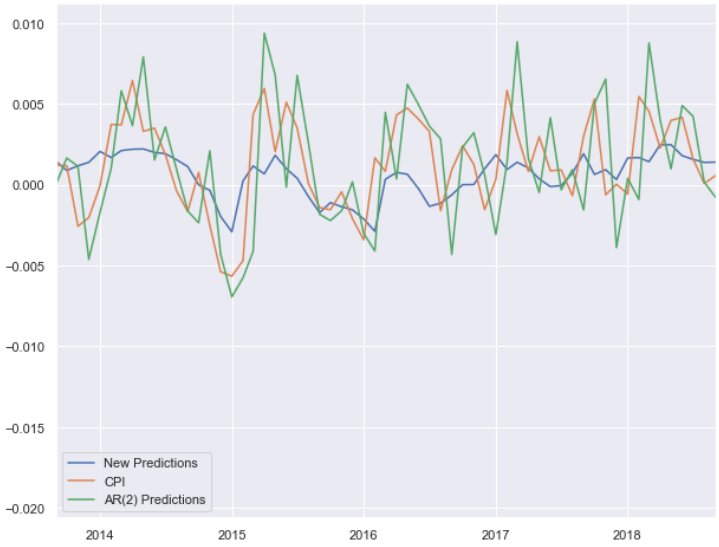


Figure 9: Monthly Averages, MSE: 6.5402e-06

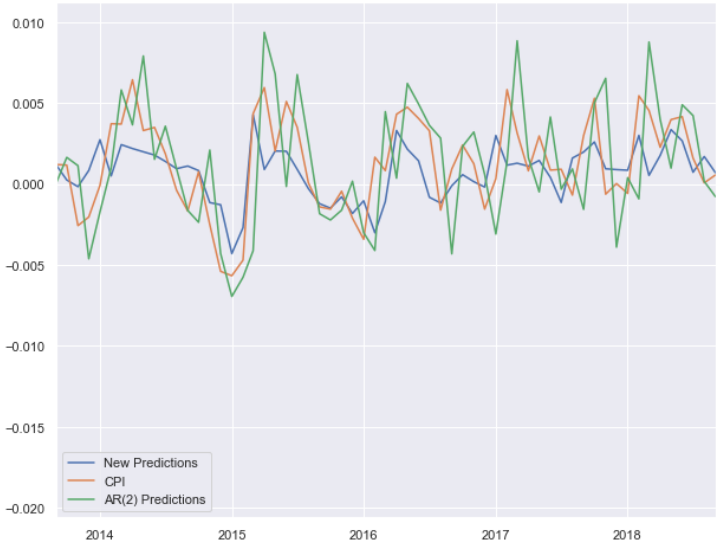


Figure 10: First day of each month, MSE: 4.766e-06

We can see that the monthly average version generate a smoother prediction curve. The actual CPI data has higher variance and the predictions based on the first day of each month capture this higher variance better.

d) Model Improvements

Below we can see the figures for different MAs and ARs. We see that an MA of order 4 gives rise to a slightly better forecasting accuracy, although this seems like it might just be noise. We also see that the model that uses both the average values and first value of each month results in performance about average between the two independently, so it is better to leave the averages out.

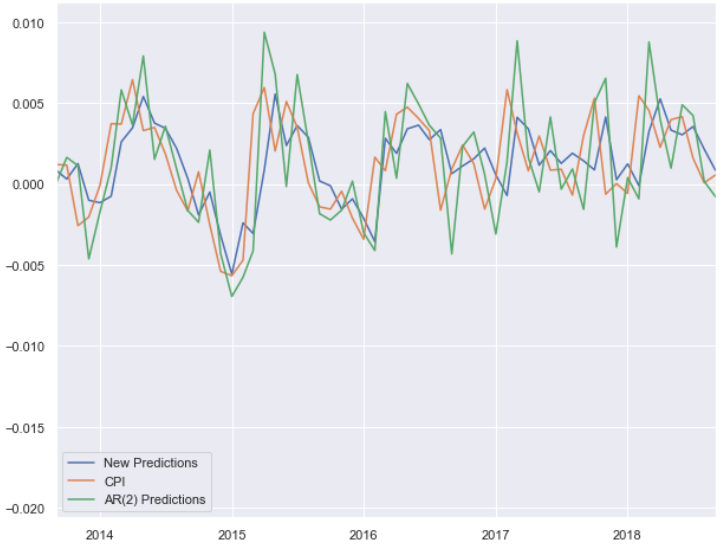


Figure 11: Monthly values, MSE: 7.311e-06

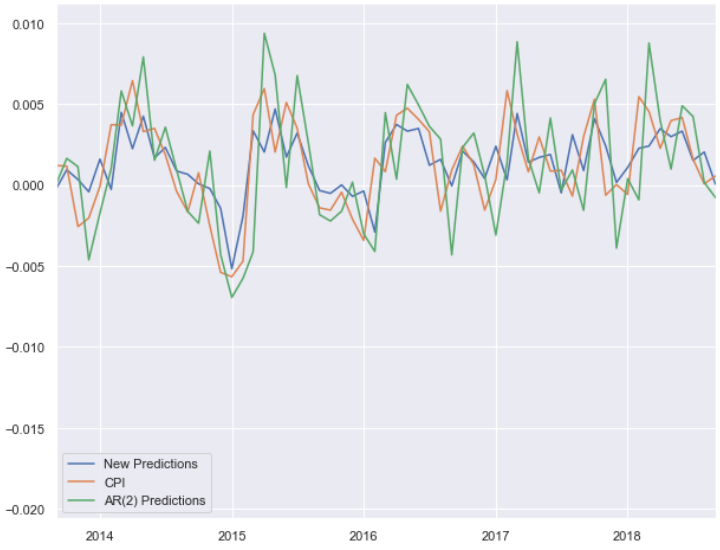


Figure 12: Average values, MSE: 4.632e-06

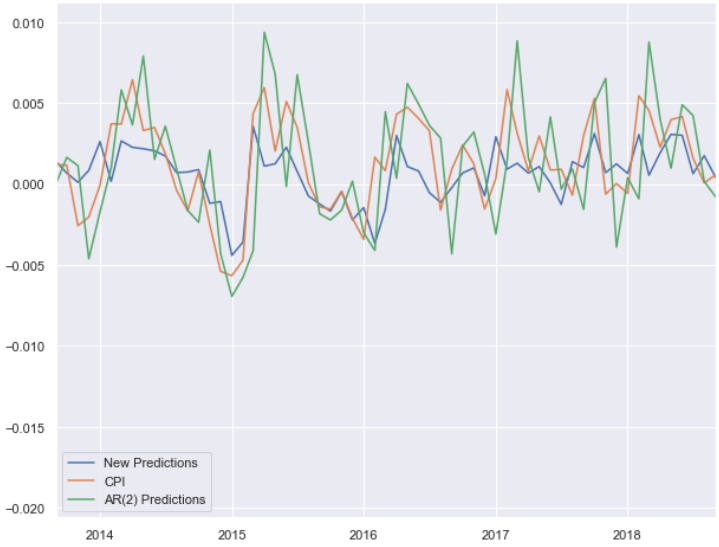


Figure 13: Both, MSE: 4.935e-06

e) Autocovariance function of $X_t = W_t + \theta W_{t+1}$

Let $X_t = W_t + \theta W_{t-1}$, where $W_t \sim N(0, \sigma^2)$. Then, we have the following:

- $E(X_t) = E(W_t + \theta W_{t-1}) = E(W_t) + \theta E(W_{t-1}) = 0$ since $W_t \sim N(0, \sigma^2)$
- $Var(X_t) = Var(W_t + \theta W_{t-1}) = Var(W_t) + Var(\theta W_{t-1}) = \sigma^2 + \theta^2\sigma^2 = (1 + \theta^2)\sigma^2$ since $W_t \sim N(0, \sigma^2)$

Consider the covariance between X_t and X_{t-h}.We have,

$$\begin{aligned} E(X_t X_{t-h}) &= E[(W_t + \theta W_{t-1})(W_{t-h} + \theta W_{t-1-h})] \\ &= E[W_t W_{t-h} + \theta W_t W_{t-h} + \theta W_{t-1} W_{t-h} + \theta^2 W_{t-h} W_{t-1-h}] \end{aligned}$$

When $h = 1$, the above expression reduces to $E(X_t X_{t-1}) = E(\theta W_{t-1}^2) = \theta(Var(W_{t-1}) + E(W_{t-1}^2)) = \theta\sigma^2$ since $W_t \sim N(0, \sigma^2)$.

When $h \geq 2$, $E(X_t X_{t-h}) = 0$ since $E(W_i W_j) = 0$ for $i \neq j$ by definition of independence of $W_i's$.

Therefore, $\rho_{ACF} = \frac{E(X_t X_{t-h})}{Var(X_t)} = \frac{\theta}{1+\theta^2}$ when $h = 1$ and 0 otherwise.

When $h = 0$, $E(X_t X_{t-h}) = E(X_t^2) = Var(X_t) = (1 + \theta^2)\sigma^2$.

Also, $\rho_{ACF} = \frac{E(X_t X_{t-h})}{Var(X_t)} = \frac{\theta}{1+\theta^2}$ when $h = 1$ and 0 when $h \geq 2$

f)

Let $X_t = \phi X_{t-1} + W_t$, where $W_t \sim N(0, \sigma^2)$. To find the covariance $E[X_t X_{t-h}]$, we multiply each side of the model for X_{t-h}, then take expectations. We have:

$$X_t X_{t-h} = \phi X_{t-1} X_{t-h} + W_t X_{t-h}$$
$$E[X_t X_{t-h}] = E[\phi X_{t-1} X_{t-h}] + E[W_t X_{t-h}]$$
$$E[X_t X_{t-h}] = E[\phi X_{t-1} X_{t-h}]$$

since $E[W_t] = 0$.

Let $\gamma_h = E[X_t X_{t-h}]$. Then, the above formula becomes $\gamma_h = \phi \gamma_{h-1}$. By moving recursively, we obtain $\gamma_h = \phi^h \gamma_0$. By definition, $\gamma_0 = Var(X_t)$. We also have that $Var(X_t) = \phi^2 Var(X_{t-1}) + \sigma^2$. Since, the series is stationary (that is $Var(X_t) = Var(X_{t-1})$), we get that $Var(X_t) = \frac{\sigma^2}{1-\phi^2}$.

Therefore,

$$\gamma_h = \phi^h \frac{\sigma^2}{1 - \phi^2}.$$

2 The Mauna Lua CO_2

a) Linear Model Fit

We used linear regression to fit the data. The estimated parameters were $\hat{\alpha_1} = 306.96$ and $\hat{\alpha_2} = 1.52$, with an R2 of 97.8%. Below we have printed the predicted vs observed values, and also the residuals. We can observe that we manage to capture the general trend, however we can still improve, since there is a pattern in the residuals and they are not random.

We can see the data and the fit in Figure **??**, and the residuals in Figure 15.

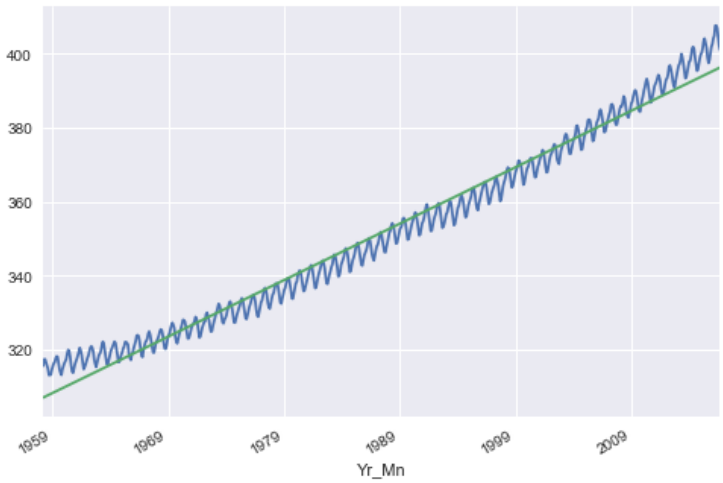


Figure 14: Fit of the data

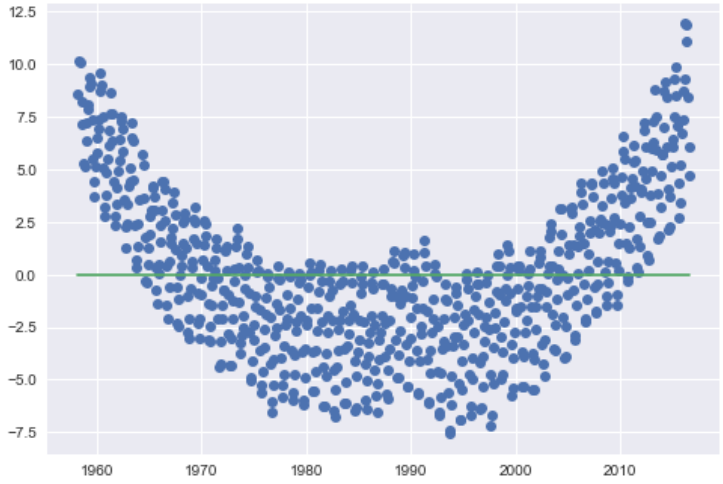


Figure 15: Residuals

b) Quadratic Model Fit

Here we fit the data to a quadratic model. We again, used linear regression. The estimated parameters were $\hat{\beta}_1 = 314.30$,$\hat{\beta}_2 = 0.78$ and $\hat{\beta}_3 = 0.0125$, with an R^2 of 99.3%.

Also, we have plotted the predicted vs observed values, and also the residuals. We can see that the residuals are much more random than before which means that fitting is much better.

We can again see the data and the fit in Figure **??**, and the residuals in Figure 15.

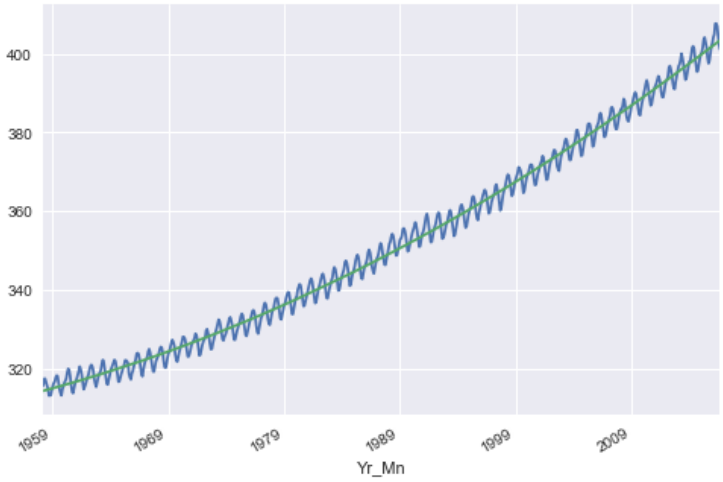


Figure 16: Fit of the data

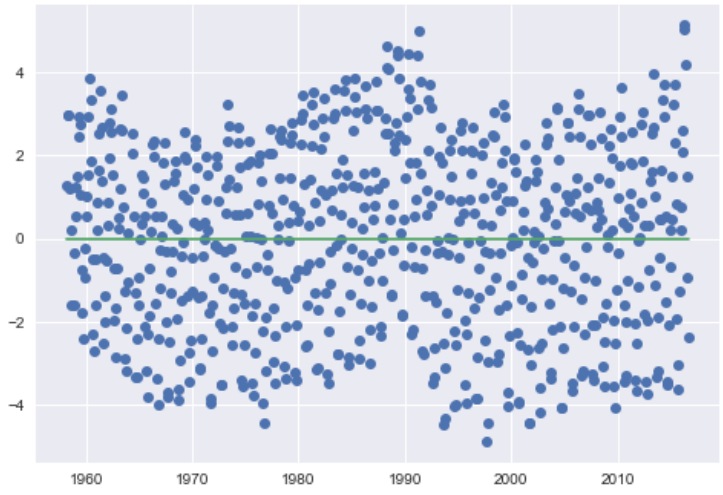


Figure 17: Residuals

c) Model Difference

We can see that the second fit is much better, since it follows the observed data more closely. Additionally, the residuals are much smaller. The R^2 values also

show that the second model explains more of the sample variance compared to the first model.

d) Removing Seasonality

The periodic signal we need to extract is shown in Figure 18.

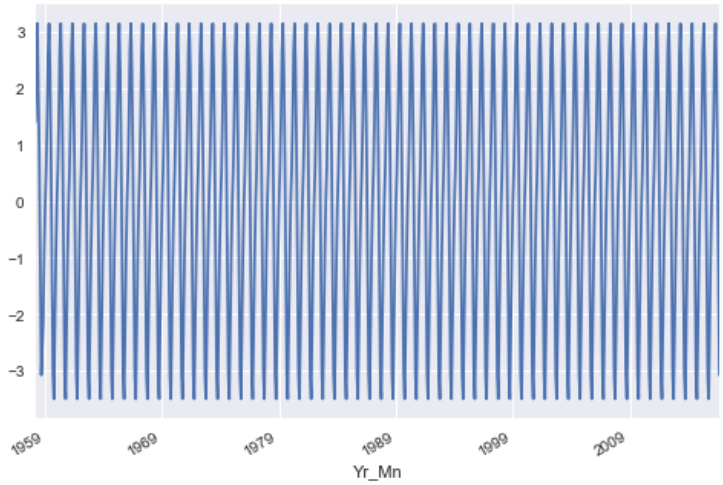


Figure 18: Periodic Signal to be exctracted

e) Variation of CO_2

The plot vs the predicted curves are shown in Figure 19.

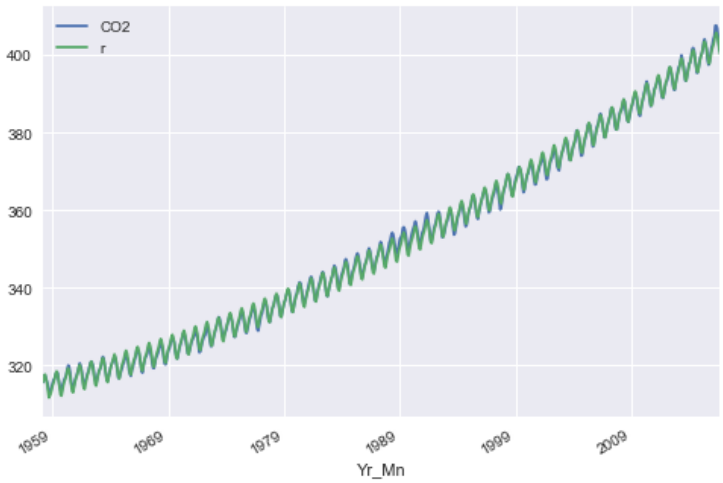


Figure 19: Fit of model

We can see that the predicted values follow the actual data very closely. Thus, we conclude that the seasonal variation is similar across years, while there is a clear upward trend in the same season values.