Predicting the Phillips Curve Using Machine Learning

Albert Choi, Henry Price, Sean Sanatana

Boston University

CDS DS 110: Introduction to Data Science With Python

Dr. Kevin Gold

May 2023

Introduction

The purpose of this research is to find the relationship between inflation and unemployment rate. This is data only from the US and dates all the way back to 1961. We have gathered the data from 1960 to the present to understand unemployment rates as inflation increases because of economic growth. As urban consumers the CPI (Consumer Price Index) increase means consumer goods and services increase in value. The relationship between CPI and unemployment rate could be seen as the Phillips Curve. The Phillips Curve is an economic concept that was developed by A.W. Phillips, which stated that inflation and unemployment rate have a stable and inverse relationship. What this means is that as the economy grows, inflation also grows with it. This creates more jobs and unemployment decreases. It is a concave curve which applies to the inverse relationship between unemployment and inflation. If flipped the opposite would happen, if inflation decreases unemployment increases. The Phillips Curve works in theory but what would happen if the economy becomes stagnant? In the 1970s the economy faced stagnant growth. This meant that there was high unemployment rates and inflation rates.

Methodology

In order to begin predicting parameters with machine learning regression models, the underlying data must be cleaned and formatted in such a way that it can be used for such prediction. In this specific case, the unemployment data and inflation data are in different comma separated value (CSV) files. So the relevant data from both files must be collected and merged together into a single table for use with the machine learning models. Additionally, the inflation data must be converted from monthly CPI into yearly inflation rate using the formula:

$\frac{CPI_x - CPI_{x-12}}{CPI_{x-12}} * 100$.

To begin, the CPI data will be cleaned and reformatted to indicate over year inflation rate. The CPI data set has two columns: CPI and Date in YYYY-MM-DD format, with each column corresponding to a different month. The format of this data set will remain the same, the only differences being that the CPI values will be replaced with inflation, and the first 12 months' rows will be dropped as there is no previous year from which to calculate inflation. Practically, this is done by first using a loop to iterate through each row starting from the most recent month and working backward to the 12th month, and applying the aforementioned formula to each subsequent month, then dropping the first year's rows. This results in a table identical to the original, the only differences being that the CPI values have been replaced with inflation values and the first row is what was formerly the 12th.

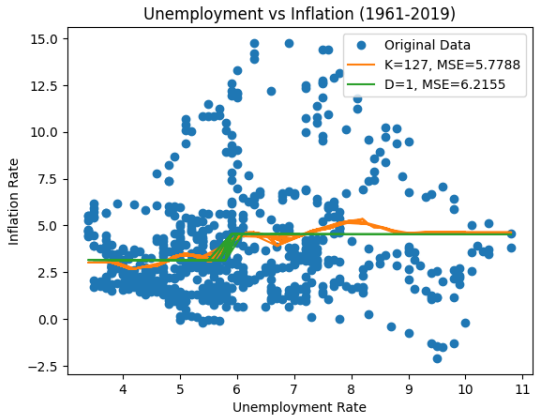
The format of the unemployment data is quite different from that of the CPI data. The data is structured with 12 columns, each corresponding to a different month, and each row corresponds to a different year. This data will be reformatted such that it matches that of the CPI data set. Empirically, this is done with two nested loops, the outer loop iterating over each row, and the inner loop appending each column to a new dataframe as a row indexed with the date in YYYY-MM-DD format. Once complete, the two cleaned data frames can be merged together, using the date as the index column. This merged data frame is what will be used for machine learning modeling.

With this reformatted data, the machine learning models can now be fitted to the data. To begin, the k nearest neighbors model is fitted to the data with unemployment rate as the X variable and inflation rate as the y. In order to maximize the fit of the model, the k parameter that yields the best fitting model must be found. This is done with a loop that, for each value of k from 1 to the number of samples in the data set (N), fits a model using this k value, takes the

MSE of said model, and stores both values in a dictionary with the MSE as the key and the value as k. Once this loop has terminated, the function returns the predicted values for the model with the lowest MSE. A very similar function is used for the decision tree regressor, the only difference is that the maximum depth is varied instead of k, and the greatest maximum depth that is used is 10.

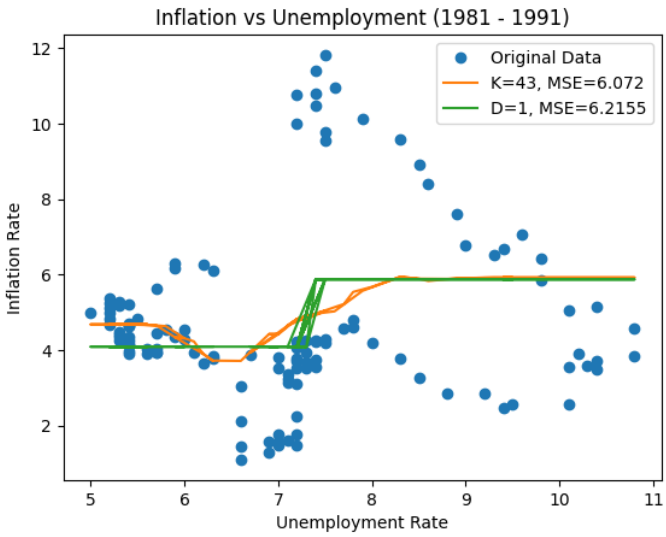
Results

To begin analyzing the results, we will start with the models fitted to the entire dataset. When visualized, this results in the following graph:



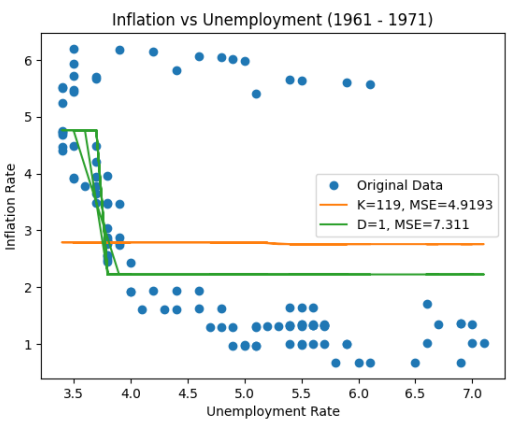
As can be seen in the graph, the KNN model with a k value of 127 resulted in a regression line with a MSE of 5.7788, while the decision tree model with a depth of 1 resulted in a regression line with a MSE of 6.2155. From this we can conclude that the best fitting KNN model fits the data better than the best fitting decision tree model, as indicated from the lower MSE, subsequently resulting in more accurate predictions.

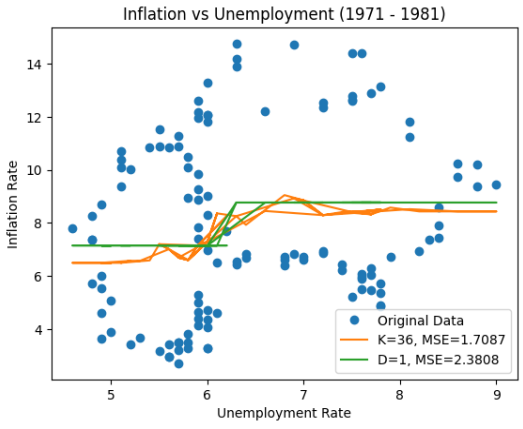
Now we will fit the data to a smaller time frame of 10 years, 1981-1991. This results in the following graph:

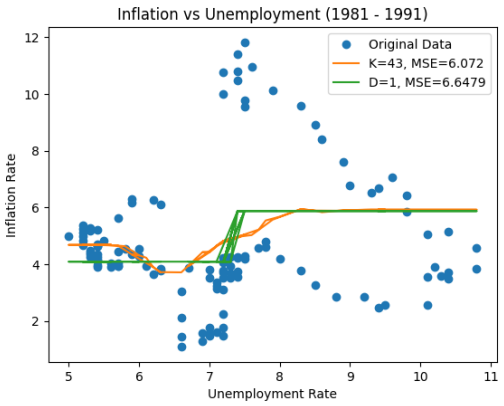


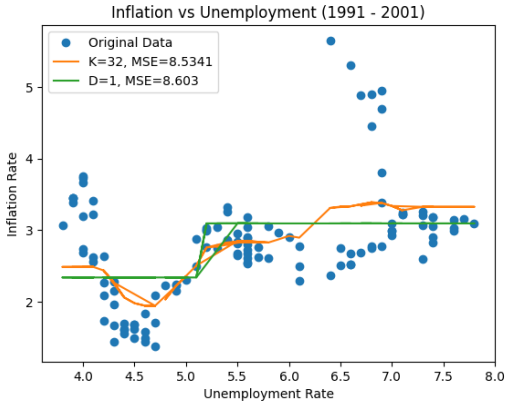
Once again, the KNN model, this time with k equal to 43 and a MSE of 6.072, fits the data better than the decision tree model with a maximum depth of 1 and a MSE of 6.2155.

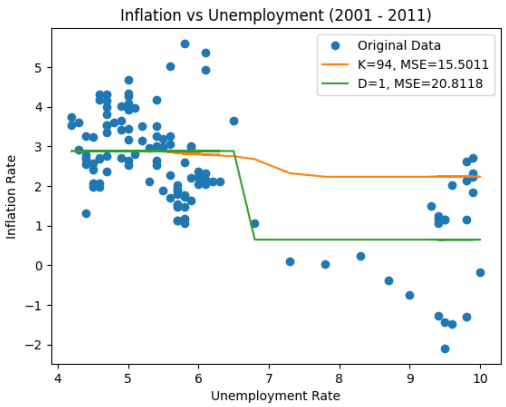
In both of these predictions, a positive relationship is observed between inflation and unemployment. This directly contradicts the Phillips Curve that is treated as a textbook macroeconomic theory, rarely ever disputed. To investigate this more thoroughly, we will run the same regression procedure as above for all 5 decades from 1961 to 2011. This results in the following graphs:











As can be seen above, a negative relationship is only observed in two out of five decades, 2001-2011 and 1961-1971. Once again this result contradicts the Phillips Curve theory.

The final thing that will be examined is predicting contemporary inflation rates using current unemployment rates. For this we will use both the 1961-2019 model and the 2001-2011 model. The most recent unemployment rate as of March 2023 is 3.5%(bls.gov). Plugging this into the full time frame model results in the predicted inflation rate of 3.01902835%. Comparing this to the actual inflation rate 5%(bls.gov), we can see that the model was about 2% inaccurate. Using the same procedure for the short time frame model, we get a predicted inflation rate of 2.88411596%, slightly less accurate than the previous model.

Conclusions

While conducting this research, we encountered unexpected discrepancies between our data and the Phillips Curve Theory. As the theory is seemingly concrete, and rarely ever disputed, we had originally assumed that there was an issue with our data or the cleaning process that it underwent. However, after editing our code to no avail, we decided to take a different perspective, looking into the Phillips Curve Theory itself. In doing so, we understood that our main issue actually lay within our broad time frame. The Phillips Curve theory was established as a short-run model, only holding true for time periods of around 10 years or less. Since our dataset covered a broad span of around 50 years, the cause of our previous discrepancies became evident. Thus, we decided to split our datasets into decades, investigating the relationship between our dataset and the Phillips curve further. Coincidentally, we encountered more discrepancies. Only two out of the five decades we analyzed seemed to correlate with the Phillips Curve Theory, directly contradicting this macroeconomic theory that had been long established as sound. Throughout this research process, we came to numerous realizations. Firstly, we were able to understand that machine-learning regression is not always as accurate as standard OLS regression. This was made clearly evident by the lower MSE values within the linear regression model in comparison to the KNN model. Next, we learned that in order to find the root cause of an issue, it helps to approach it from varying perspectives, in spite of its seeming improbability. Finally, we realized that even theories that have been proven to be concrete can have discrepancies/contradictions and that they cannot always be relied upon to be a “one size fits all” model for every dataset.

References

U.S. Bureau of Labor Statistics, U.S. Bureau of Labor Statistics, 2 May 2023, https://www.bls.gov/.