Seasonal ARIMA Model and Forecasting Monthly CPI in China

1. Overview

This is the term project I took in Time Series Analysis in my postgraduate studies at the University of Southern California in 2021 Fall.

1. Background & Project Goal

The Consumer Price Index (CPI) forecast has received wide attention around the world due to its excellent reflection on the state of the economy. The reason we chose China as our study area is that China is the second-largest economy in the world, an important manufacturing country, and an important link in the global supply chain. Our project goal is to forecast the monthly CPI in China using data from 1995 to 2018.

1. Data Sources

The monthly data from January 1995 to September 2021 are collected for our SARIMA model to predict China's CPI, and then to study the changes in China's macro price level. What we want most is CPI data with January 1995 as the base period, whose advantage is that it can more directly reflect the trend of China's macroeconomic growth.

However, there is no existing fixed-base CPI that can be selected without any preprocessing. Thus, we download the month-on-month data (last month=100), from the National Bureau of Statistics of China, and convert it into CPI with a fixed base (Jan 1995=100) with the following formula:

*CPI with a fixed base in this period = month-on-month CPI in this period \* CPI with a fixed base in the last period*

Then we have divided our data into training datasets from January 1995 to December 2018, where the last three months are the validation dataset and the testing datasets from January 2019 to September 2021. To be more specific, we regard the period of 2019 as the normal period and the period of 2020-2021 as the recession period, when COVID-19 happens.

1. Methodology
2. Seasonal ARIMA Model
3. Testing Stationery
4. Model Identification and Estimation
5. Diagnostic Checking
6. Model Forecasting and Accuracy Measures
7. Extension --- Time Series Split Cross-Validation for Robustness Check
8. Sample Results
9. Modeling Result

We could detect the possible values of order p and order q are {3,4,5,6}, and the values of order P and order Q are {1,2}. Thus, we implement trial and error approach to set different ranges as input in the AIC. Let us assume the symbol (p, q, P, Q) indicates the maximum p, q, P, Q that we put in the information criteria.

Thus, the AIC presents the results with training dataset from Jan 1995 to Sep 2018.

|  |  |  |
| --- | --- | --- |
| AIC Input | Best Model in This Input | AIC value |
| (3, 3, 2, 2) |  | 436.60 |
| (4, 4, 2, 2) |  | 435.85 |
| (5, 5, 2, 2) |  | 435.85 |
| (6, 6, 2, 2) |  | 435.85 |

Table 1

We construct SARIMA model with these parameters and test their performance in the validation sample from Oct 2018 to Dec 2018. Their Mean Square Error and Ljung-Box Test results are shown in Table 2.

|  |  |  |
| --- | --- | --- |
| Model | Ljung-Box p-value | Validation Sample MSE |
|  | 0.6741 | 0.2848 |
|  | 0.8375 | 0.1368 |

Table 2

1. Forecasting Result with Different Steps

We use 3-step forecastand 12-step forecast in the test dataset for both normal period and recession period to measure the performance of our model and compare the forecasting result with different steps. We will re-estimate the model with updated dataset every quarter for 3-step forecast to make them comparable in the 12-month performance. Notice that the blue scatter plot represents 3-step forecast, the red triangle plot represents 12-step forecast and the black line indicates the actual value of CPI in both Figures.

Chart, line chart, scatter chart

Description automatically generatedChart, line chart

Description automatically generated

1. Robustness Check for Model Selection

We implement the cross-validation strategy mentioned in the extension part of methodology section to check the robustness of model selection.

|  |  |  |
| --- | --- | --- |
| Model / Validation MSE |  |  |
| Baseline | 0.9236 | 1.1981 |
| Extension | 0.8186 | 0.7484 |

|  |  |
| --- | --- |
| Model | Out of Sample MSE |
|  | 0.9236736 |
|  | 0.748400766 |

Table 3

If we use Baseline strategy with only one period training-testing scheme, it will choose due to the lower MSE in the validation sample, while the Extension strategy with Cross-validation scheme chooses after averaging all the MSE in the multiple validation sample. The result of out of sample forecast indicates that would perform better.

This project uses Seasonal Autoregressive Integrated Moving Average (SARIMA) to forecast the monthly CPI in China using data from 1995 to 2018. The study presents the SARIMA model is best for predicting CPI in China and it measures the performance of SARIMA model in both the normal period from Jan 2019 to Dec 2019 and the recession period from Jan 2020 to Sep 2021. The result of our study shows that the SARIMA model can track the actual trend of CPI in both periods.