Analysis of CPI For Inflation Prediction

Channing Che (3034120912), George Henderson (3032960543), Jeremy Levitt (3032683747)

Introduction

Our group was motivated by an interest in applying methods in time series analysis to questions of immediate relevance in the news and in our social and political experience. To us, one form of time-series data that has had a sudden and personal impact has been inflation. At the beginning of the COVID-19 pandemic, inflation was in the news as a speculative concern, not only because of the rupture of the global supply-chain, but also as a potential consequence of government aid programs. Now, reporting on inflation has changed from speculative to descriptive, as we have seen a sharp increase in inflation rates.

To model inflation, we chose to consider the Consumer Price Index. Our data comes from usinflationcalculator.com,[1] which cites the Bureau of Labor Statistics, Reserve Monetary Policy, and Reuters Inflation News as sources of its data. The full record of CPI data ranges from 1913 to 2021, and it is calculated using a 1982 reference base of 100 index units. The 1929 stock market crash and the economic boom following the Second World War introduced local variance that seems unlikely to occur in the future. Therefore, we decided to pursue the project using data from the postwar era, as this era seems more representative of current conditions. With this dataset, our goal was to model how the CPI would change in the future, from which we could directly deduce how inflation rates would change.

Exploratory Data Analysis

Figure 1 below shows the time series data.

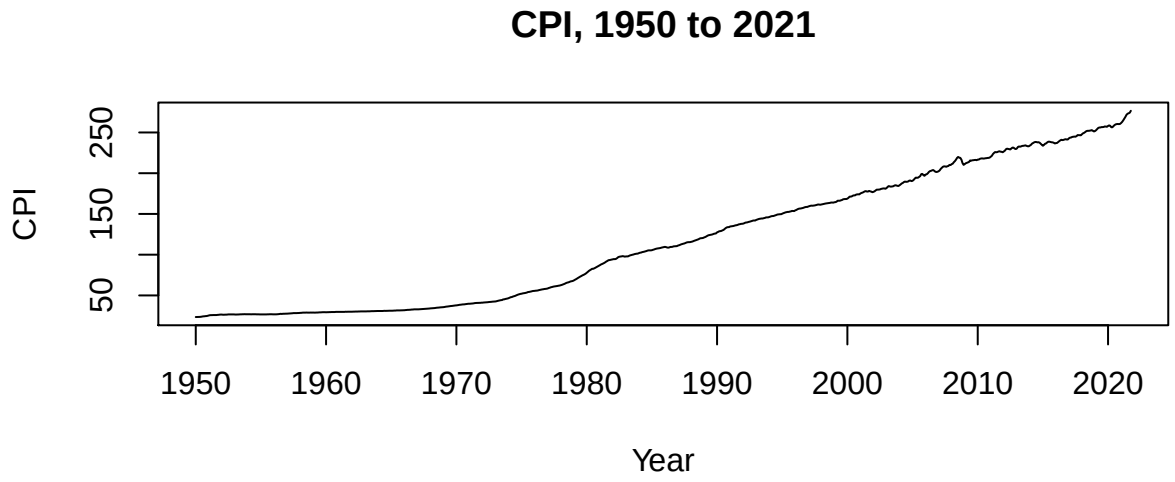

Figure 1: Time series plot of the data

There is no obvious seasonality we can observe from the plot above. We note three time segments with distinct behavior. From 1950 to 1982, we see a gradual quadratic curve; the trend then levels and appears to be linear until around 2000. At this point, we begin to see noise at the macro-level. This is unique because we have noise throughout the data; our scope, covering over 70 years, makes the data appear smoother than it actually is. In the 2000s, the data retains its apparent linearity, but its variance increases dramatically, so heteroscedasticity is present. We see this with the spike in 2008, along with other smaller spikes; we also see the spike beginning around 2020, which has yet to level or to crash. Since the financial era after 2000

[1] https://www.usinflationcalculator.com/inflation/consumer-price-index-and-annual-percent-changes-from-1913-to-2008/

began with the dot-com crisis, we speculate that the data's strong heteroscedaticity is due to the basis of the financial economy around increasingly prospective investments in technology companies. We also expect that the CPI may exhibit such variance because the US economy at this point will have become highly globalized and thus increasingly dependent on price-determinations through the global supply chain.

Models Considered

The time series plot of the original data shows that the variance of the data is higher in the later years than in the earlier years, so we will take a variance-stabilizing log transform of the data. The resulting plot is as follows.

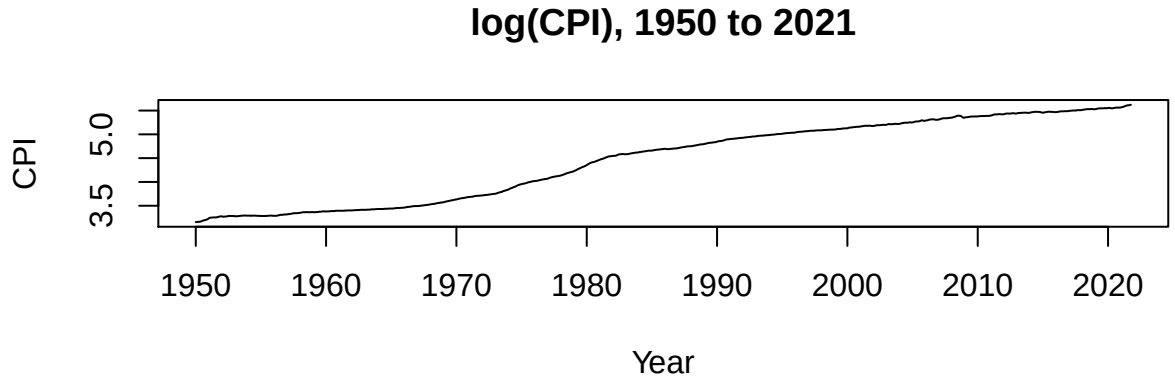

Figure 2: Time series plot of the data after log transform

From Figure 2, we can see an increasing trend which looks more quadratic than linear, but no clear seasonality. The variance is also stabilized compared to before. The two signal models we will consider in our pursuit of stationarity will be based on the removal of this quadratic trend.

Second-Order and Lag-12 Differencing

Since the trend looks quadratic, we take a second-order differencing in order to remove this trend. From the ACF of the differenced data, we saw that it exhibited annual seasonality, so we further took a lag-12 difference to remove this seasonality. The resulting plot of residuals is shown below as Figure 3, and looks fairly stationary overall.

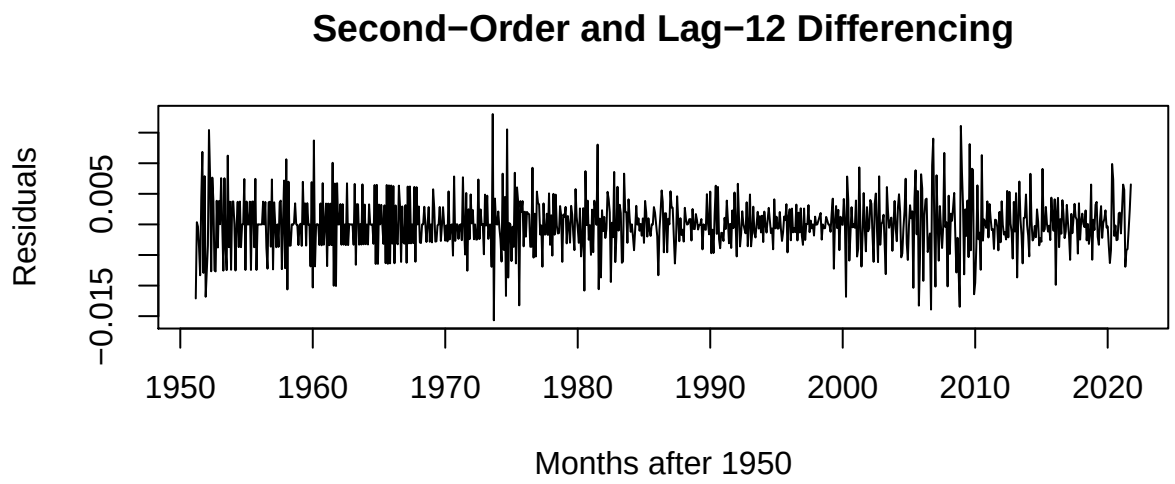

Figure 3: Plot of residuals after differencing

Second-Order and Lag-12 Differencing with MSARMA(p = 1, q = 3, P = 1, Q = 1)

The ACF and PACF of the second-order and lag-12 differenced time series are shown below in Figure 4.

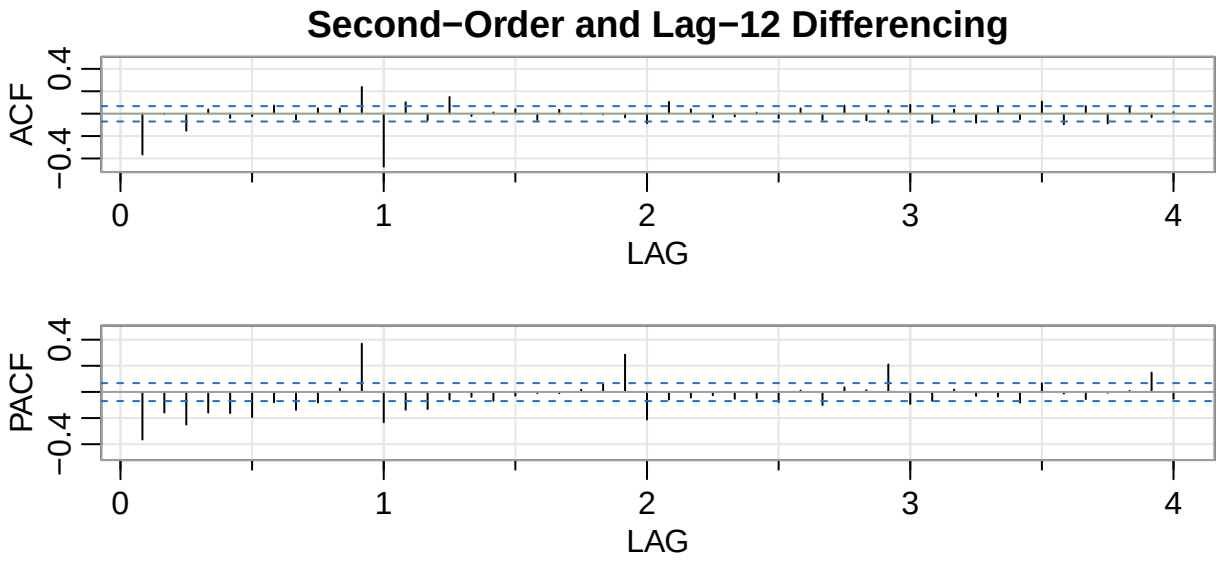

Figure 4: ACF (top) and PACF (bottom) for the differenced series

From the ACF, we see significant lags at $h = 1, 3, 11, 12, 13, 15$ (these correspond to months; note that the labels on the ACF and PACF correspond to years). An multiplicative MA model of the form $(1 + \theta_1 B + \theta_2 B^2 + \theta_3 B^3)(1 + \Theta B^{12})$ has nonzero theoretical ACF at lags $h = 1, 2, 3, 12, 13, 14, 15$, so it closely matches what we see in the sample ACF. To try to fit the match the sample PACF more closely, we also include an AR parameter ϕ and a seasonal AR parameter of order 12, Φ. A pure multiplicative AR model with just these parameters would have nonzero PACF values at lags $h = 1, 12, 13$, so combining this multiplicative AR model with our multiplicative MA model in the form of an MSARMA model will more closely match those lags in the PACF, as well as the lags in the PACF at months 11, 23, 35, and so on.

The fitted values are shown below in Figure 5, along with the original data. Upon inspection, the fitted values do look good.

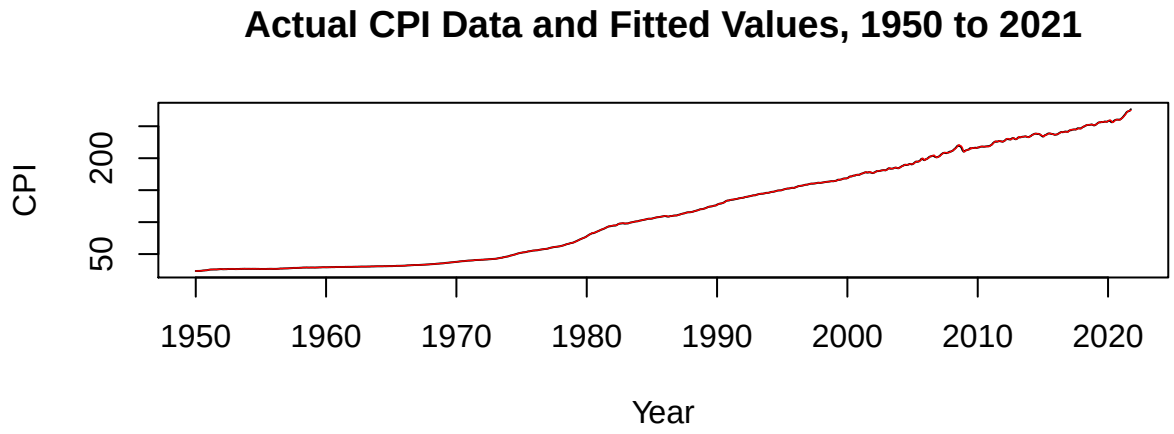

Figure 5: Fitted values (red) and actual data (black)

The SARIMA diagnostics are shown below in Figure 6.

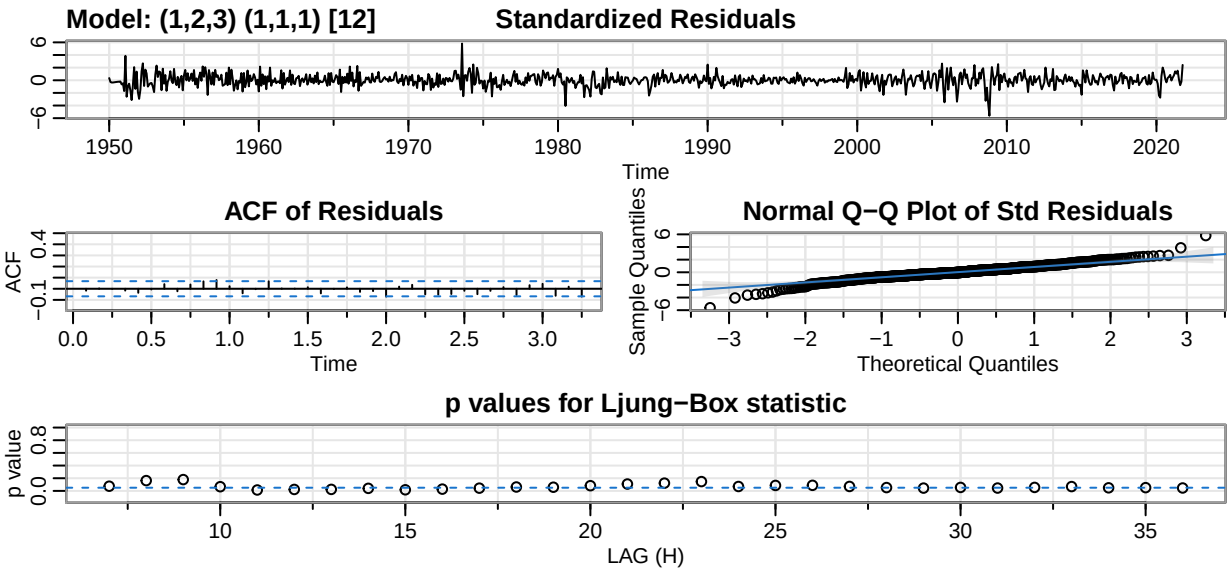

Figure 6: Model diagnostics for SARIMA(p = 1, d = 2, q = 3, P = 1, D = 1, Q = 1, S = 12)

The standardized residuals and ACF of residuals do look like white noise in shape. The results of the Ljung-Box test support the idea that the residuals are white noise, however, as at around lag 20, the p-values are all not significant.

Second-Order and Lag-12 Differencing with MSARMA(p = 0, q = 3, P = 3, Q = 1)

With this model, we place a greater emphasis on trying to capture the significant lags in the PACF at years 1, 2, and 3. In that vein, we keep the multiplicative MA model from before, but include three seasonal components for the seasonal AR model, so that the seasonal AR part is now $(1 + \Phi_1 B^{12} + \Phi_2 B^{24} + \Phi_3 B^{36})$. We don't include the regular AR parameter ϕ in this case for fear of overfitting.

The fitted values for this model are shown below in Figure 7, and again, it appears by observation that the fit is quite good.

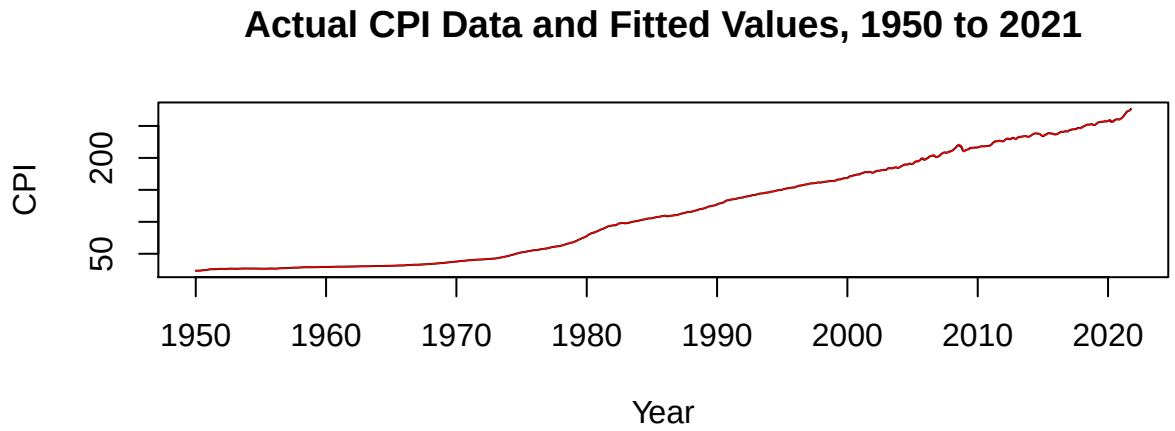

Figure 7: Fitted values (red) and actual data (black)

The SARIMA model diagnostics are below in Figure 8.

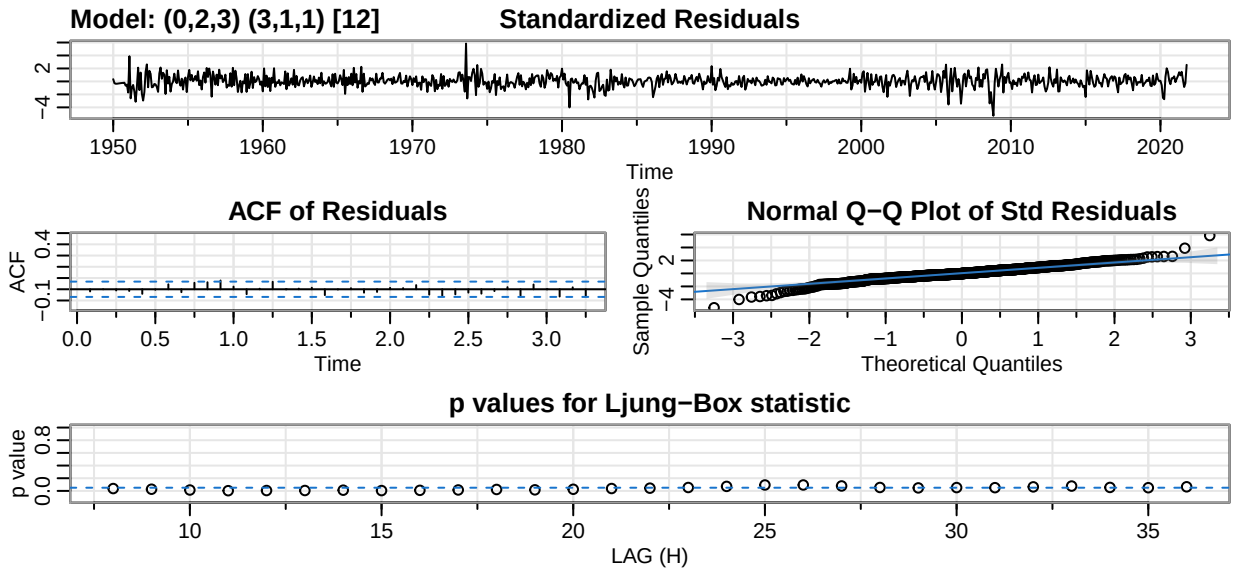

Figure 8: Model diagnostics for SARIMA(p = 0, d = 2, q = 3, P = 3, D = 1, Q = 1, S = 12)

Based on the plot of standardized residuals and the ACF of residuals, the residuals do look like white noise in shape. The p-values for the Ljung-box test right around lag 20 are significant, which suggests that this model may not fully fit, but the residuals do look good, so we still believe that this model is appropriate.

Parametric Trend Fitting

Another way to remove a quadritic trend is to estimate it and then subtract the estimates from the data. We do this by fitting a quadratic model using least squares. The plot of the residuals looks cubic, so we further do a third-order difference. Based on the ACF and PACF of this differenced data, we observe annual seasonality, so we further perform a lag-12 difference, at which point the residuals look stationary overall, as shown below in Figure 9.

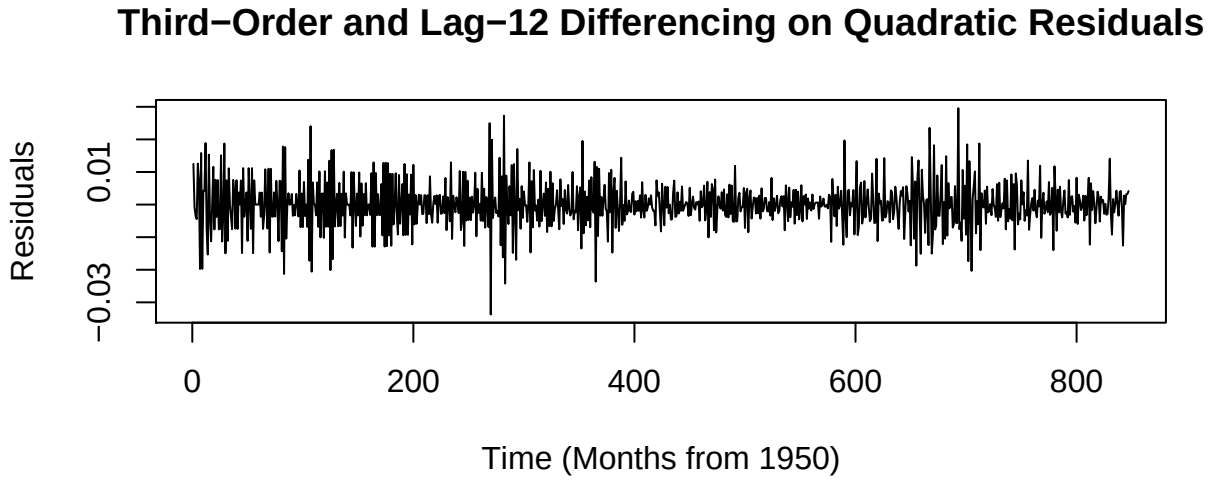

Figure 9: Time series after third-order and lag-12 differencing on residuals obtained by estimating quadratic trend

Quadratic Model with SARIMA(p = 0, d = 3, q = 4, P = 0, D = 1, Q = 1, S = 12)

The ACF and PACF plots for the stationary-looking residuals are shown below in Figure 10.

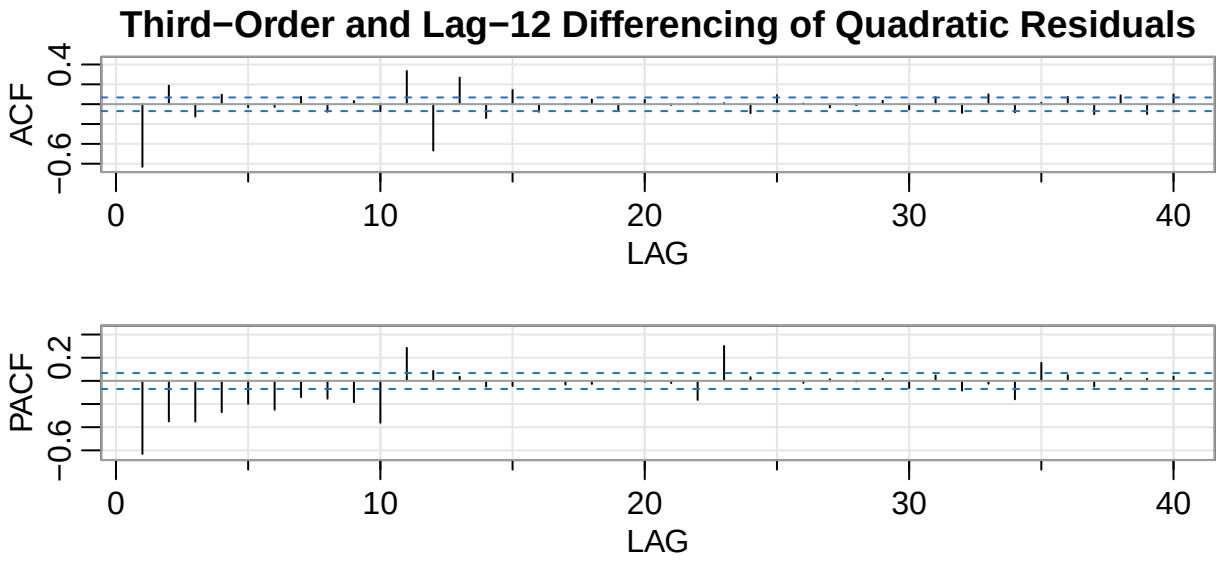

Figure 10: Time series after third-order and lag-12 differencing on residuals from estimating quadratic trend

Based on the plot of the sample ACFs, we try fitting the multiplicative MA model using the product $(1 + \theta_1 B + \theta_2 B^2 + \theta_3 B^3 + \theta_4 B^4)(1 + \Theta B^{12})$. The theoretical ACFs of this model are nonzero at lags $h = 1, 2, 3, 4, 12, 13, 14, 15, 16$, which closely matches the sample ACFs for the first 20 lags (we ignore the later lags for now).

The fitted values are shown below in Figure 11, and as before, the model does appear to fit the data well.

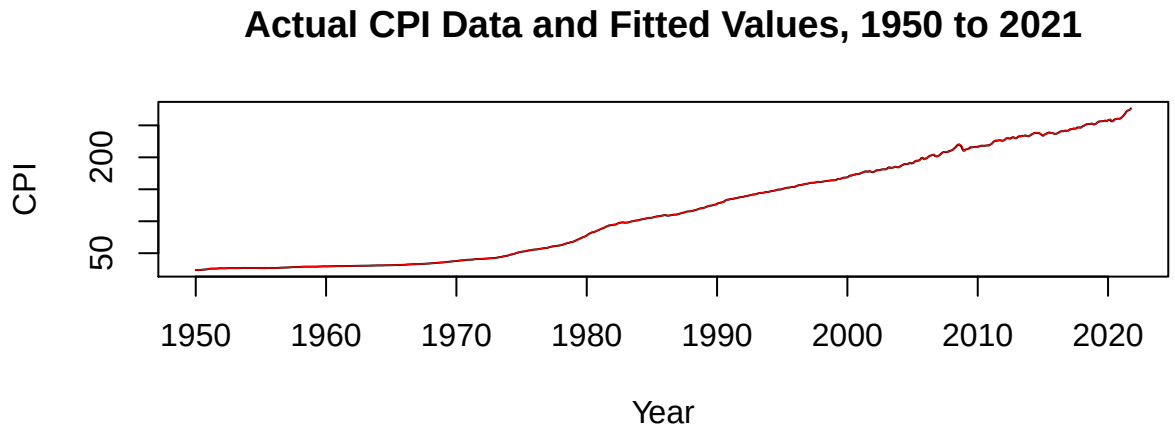

Figure 11: Fitted values (red) and actual data (black)

Figure 12 below shows the SARIMA diagnostics.

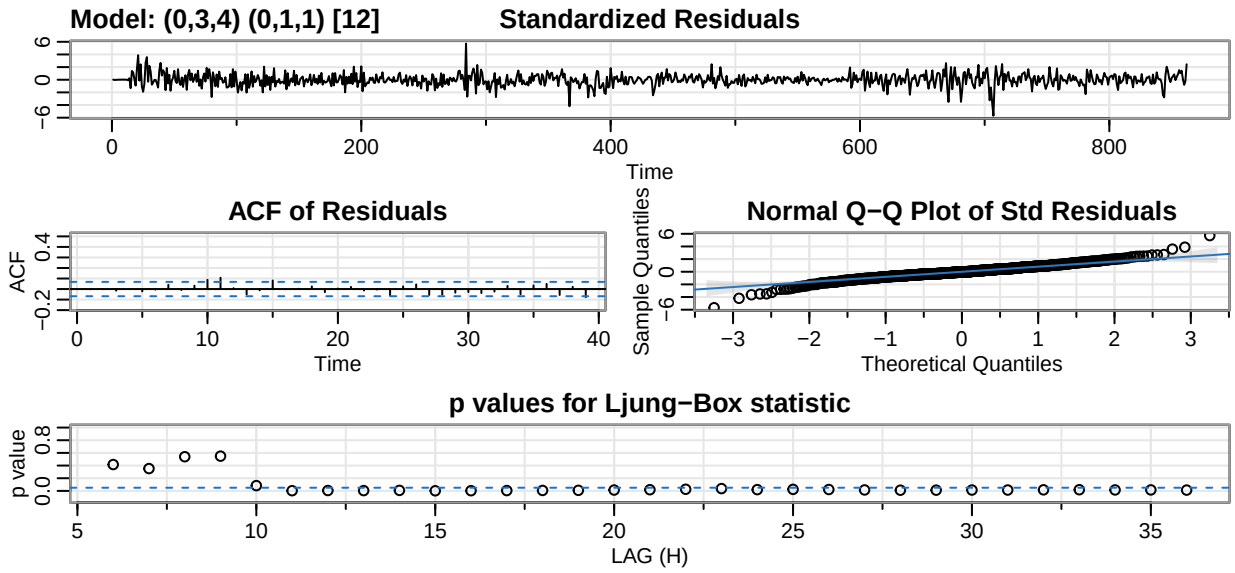

Figure 12: Model diagnostics for SARIMA(p = 0, d = 3, q = 4, P = 0, D = 1, Q = 1, S = 12) model for residuals of quadratic trend estimation

Based on the plot of standardized residuals and the ACF of residuals, the residuals do not look too different from white noise in shape, but the fit is not perfect. There are a few values of the sample ACF that lie outside the blue 95% confidence bands for white noise. The p-values for the Ljung-Box test at around lag 20 also are significant, which suggests that the model may not be a perfect fit. Still, based on the first two plots mentioned, we believe the model is still reasonable for prediction.

Quadratic Model with SARIMA(p = 1, d = 3, q = 4, P = 0, D = 1, Q = 5, S = 12)

For this model, we examine the ACF plot for larger lags than shown before, and we observe that there are many lags where the sample ACF is substantial. This is shown in Figure 13.

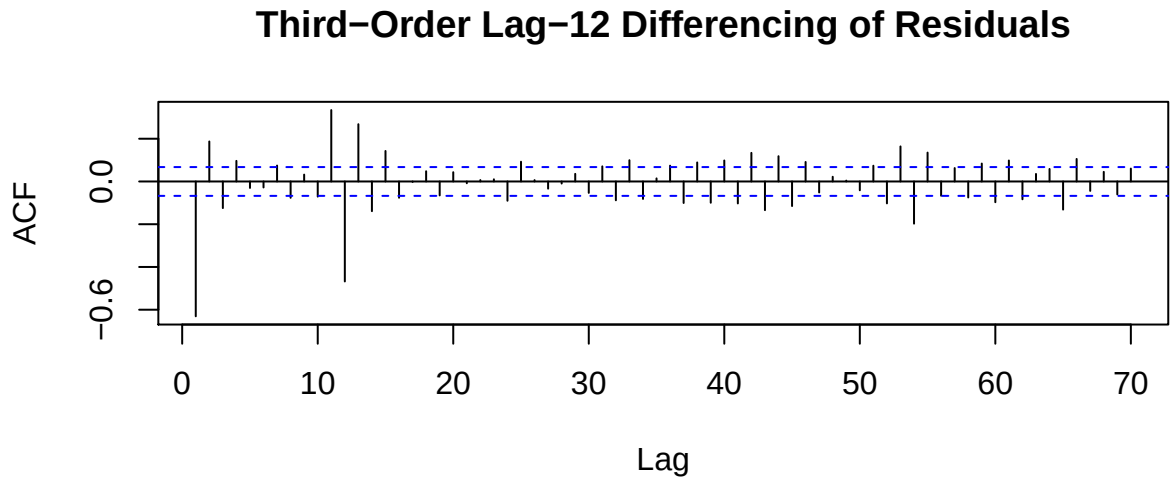

Figure 13: Sample ACF up to a lag of 70 months

So, we try to account for this by incorporating additional seasonal MA terms in our model. Specifically, we model the MA component with the product $(1+\theta_1 B+\theta_2 B^2+\theta_3 B^3+\theta_4 B^4)(1+\Theta_1 B^{12}+\Theta_2 B^{24}+\Theta_3 B^{36}+\Theta_4 B^{48}+\Theta_5 B^{60})$ to try to capture more of the nonzero lags. We will also include an AR parameter ϕ to try

to capture the more substantial PACF lag at $h = 1$. The fitted values for this model are shown below in Figure 14. As it always was previously, it appears by this type of inspection that the model fits the data well.

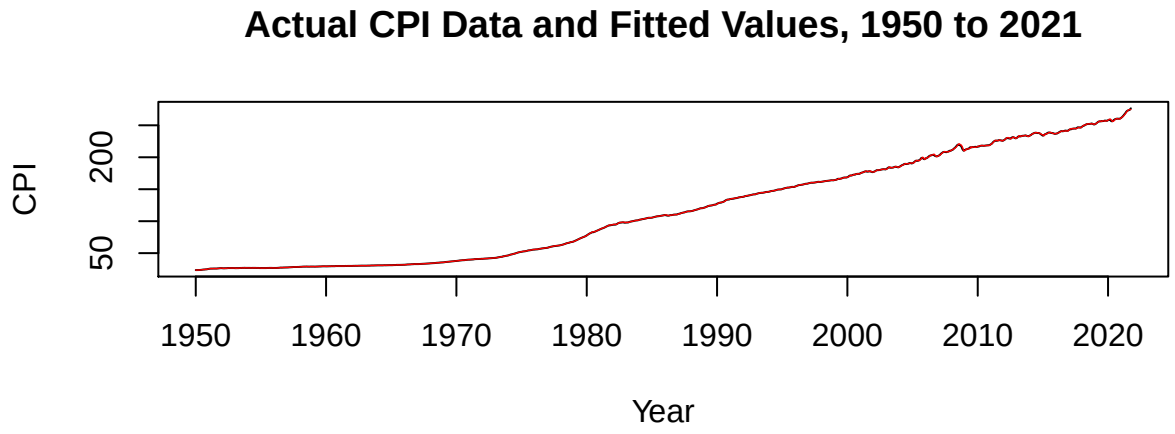

Figure 14: Fitted values (red) and actual data (black)

The SARIMA model diagnostics appear below as Figure 15.

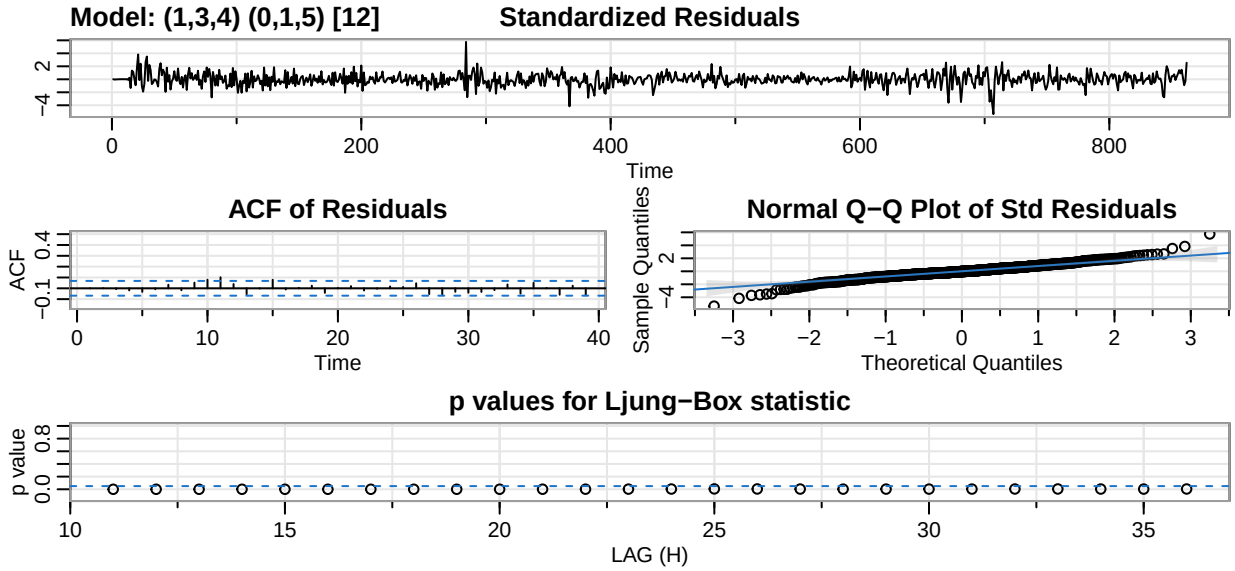

Figure 15: Model diagnostics for SARIMA(p = 1, d = 3, q = 4, P = 0, D = 1, Q = 5, S = 12) model for residuals of quadratic trend estimation

As before, the plot of the standardized residuals and the ACF of residuals look fairly close to white noise. However, the p-values for the Ljung-Box test at around lag 20 are significant, which raises doubt that this is actually white noise, suggesting that the model fit may not be perfect. We conclude that the models based on removing the trend by way of second-order differencing may be better than the models based on estimation of the trend by least squares.

Model Comparison and Selection

Since the goal of this project is to find the best model for forecasting future values, our method of comparing the four models will be to choose the model with the lowest cross-validation MSE score (CV-MSE). The

CV-MSE is calculated as follows:

1. For each year in $\{2001, 2002, \ldots, 2021\}$, train each of the models on all of the data before the selected year.
2. For each of the trained models, predict the next 12 values (or 10, in the case of year 2021, since we only have data up until October 2021) and compute the sum of the square errors for the predictions.
3. Sum together the sum of the square errors for each year (21 years in total), and take the average to yield the final CV-MSE score.

The CV-MSE scores are reported below.

Table 1: CV-MSE scores for the four models considered

Model	CV-MSE
SARIMA$(p = 1, d = 2, q = 3, P = 1, D = 1, Q = 1, S = 12)$	0.001482588
SARIMA$(p = 0, d = 2, q = 3, P = 3, D = 1, Q = 1, S = 12)$	0.001579242
Quadratic + SARIMA$(p = 0, d = 3, q = 4, P = 0, D = 1, Q = 1, S = 12)$	0.001529539
Quadratic + SARIMA$(p = 1, d = 3, q = 4, P = 0, D = 1, Q = 5, S = 12)$	0.001600821

It follows that the SARIMA$(p = 1, d = 2, q = 3, P = 1, D = 1, Q = 1, S = 12)$ (which is the first model discussed is the best model for prediction, and it will be the model we use for forecasting.

Results

Let the data used in this project be denoted as X_t, let W_t denote white noise, and let μ denote the mean of the second-order lag-12 differenced series. Then, the mathematical formula for our chosen model is

$$(1 - \phi B)(1 - \Phi B^{12})(\nabla_{12}\nabla^2 \log(X_t) - \mu) = (1 + \theta_1 B + \theta_2 B^2 + \theta_3 B^3)(1 + \Theta B^{12})W_t.$$

Estimation of Model Parameters

The parameter estimates and their standard errors are recorded in the table below. Note that the μ parameter is an estimate of the mean of the data after second-order and lag-12 differencing. It is essentially 0, which is what is desired, but we include it in the table nonetheless.

Table 2: Parameter estimates for the selected model, SARIMA(p = 1, d = 2, q = 3, P = 1, D = 1, Q = 1, S = 12)

Parameter	Estimate	SE
ϕ	-0.1228	0.2791
θ_1	-0.4932	0.2762
θ_2	-0.1858	0.1833
θ_3	-0.1648	0.0598
Φ	-0.0392	0.0406
Θ	-0.8935	0.0208
μ	0.0086	0.0981

Forecasting

In Figure 16 below, we plot the next 10 forecasted data points (red). We also add the ± 1 SE and ± 2 SE confidence intervals (shown in blue and green, respectively). For clarity, we only show the data starting from

2020 (otherwise, the graph would be difficult to clearly read); note that the horizontal axis is the year.

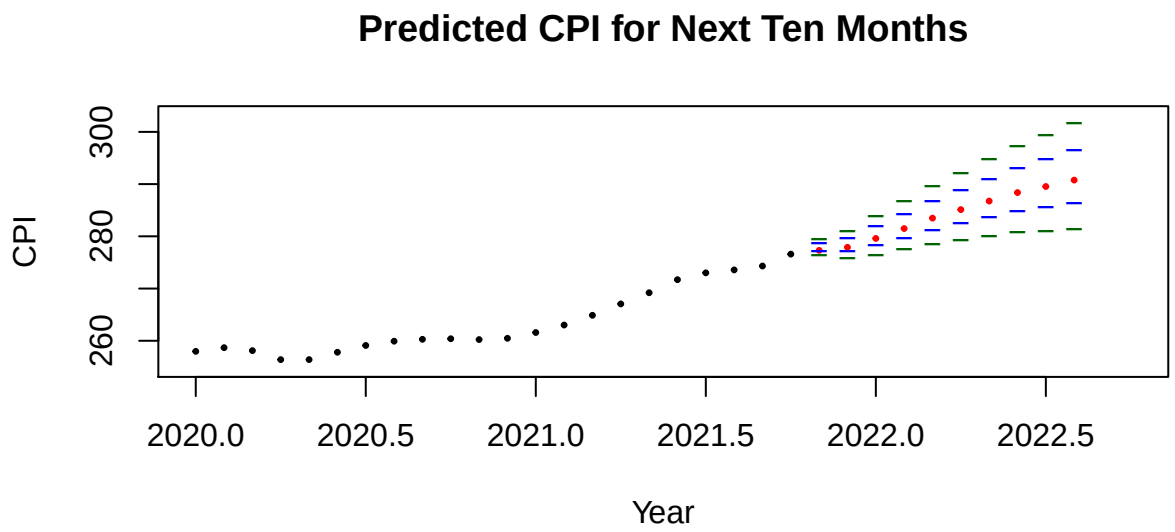

Figure 16: Forecasted values and confidence intervals for the next 10 months (November 2021 to August 2022)

From the graph, we can see that the uncertainty associated with these estimates increases as we predict further in time, which is logical. Although the first few CPI predictions have low uncertainty associated with them, at the end the predictions are so uncertain to the point that trying to model inflation using these predictions would be very difficult. For example, consider the last two forecasted values, which correspond to July and August of 2022. Suppose the prediction for July 2022 is perfect, and that the CPI is actually 289.5389 at that time. Consider the ± 1 SE confidence interval for August, which is (285.7421, 295.8787) and is "centered" at 290.7662 (the CI is not symmetric since we had to back transform using the exponential). This corresponds to monthly inflation from $\frac{285.7421-289.5389}{289.5389} \approx -1.3\%$ (i.e., 1.3% deflation) to $\frac{295.8787-289.5389}{289.5389} \approx 2.2\%$ from July to August. Since we are working on a monthly scale, this is a very wide range for possible values of inflation. And since the prediction for July 2022 has uncertainty associated with it as well, the range of inflation is actually even greater than in the example. As a result, we should not trust the predictions of this model (for the purposes of calculating the inflation rate) for the later months.

Conclusion

When we began our analysis, we wished to use complete CPI data from the mid-20th century to today in order to better try to understand trends of inflation. Given the complexity of the many underlying causes of change in the CPI over time, we were pleased to find models in which the residuals were generally stationary. Out of the four models generated, we determined which model best fit the data by evaluating each in terms of their CV-MSE. We used our best model according to this measure to forecast monthly inflation rates up to August 2022, hoping that our model would be accurate enough to predict the future inflation rate. We generated predictions of a rough average monthly inflation rate of 0.53%, which would imply an average 6.3% annual inflation rate, which at first glance seems consistent with today's expectations, as the inflation rate from October 2020 to October 2021 was 6.1%. However, as discussed previously, the standard errors were large, especially for the later predictions of the CPI, so there was a lot of uncertainty associated with these estimates. Thus, through our own experience, we concluded that having a model that fits the data well, and even performs well under cross-validation, does not necessarily imply that the predicted future values be accurate. Generating predictions on the inflation rate is as difficult as it is important to overall economic stability. While our model may not generate reliable predictions far in the future, the groundwork we laid may be useful in the development of more sophisticated models.