CPI of Used Vehicles
Time Series Analysis

KENZIE BAKER & RYAN COLE

21 APRIL, 2023

Background, Data Selection, Research Questions

What is Consumer Price Index (CPI)?

This statistic can be defined as the measure of overall change in consumer prices paid by urban consumers for a ‘market basket’ of consumer goods and services.

The data for used cars/trucks is a part of the transportation group of the CPI. The subset of that group had a weight of 2.668% of the total CPI as of December 2022.

TIMMERMAN
INDUSTRIES

CPI Usage

CPI is one of the best and most commonly used tools to measure inflation/deflation, an essential indicator of economic health.

CPI INCREASES

INFLATION INCREASES

CPI DECREASES

INFLATION DECREASES

TIMMERMAN
INDUSTRIES

Data Description

- Measure of overall change in urban consumer prices (CPI) for used cars & trucks for cities in the United States.
- Monthly data from January 1970 - Feb 2023, gathered from the U.S. Bureau of Labor Statistics.
- Vehicles considered for this data are used vehicles that are between 2 and 7 years of age. Subcompact, compact or sporty, intermediate, full, and luxury cars are all included. Light trucks in the index include pickup trucks, vans, and specialty vehicles like sport/cross utility vehicles.

Research Questions

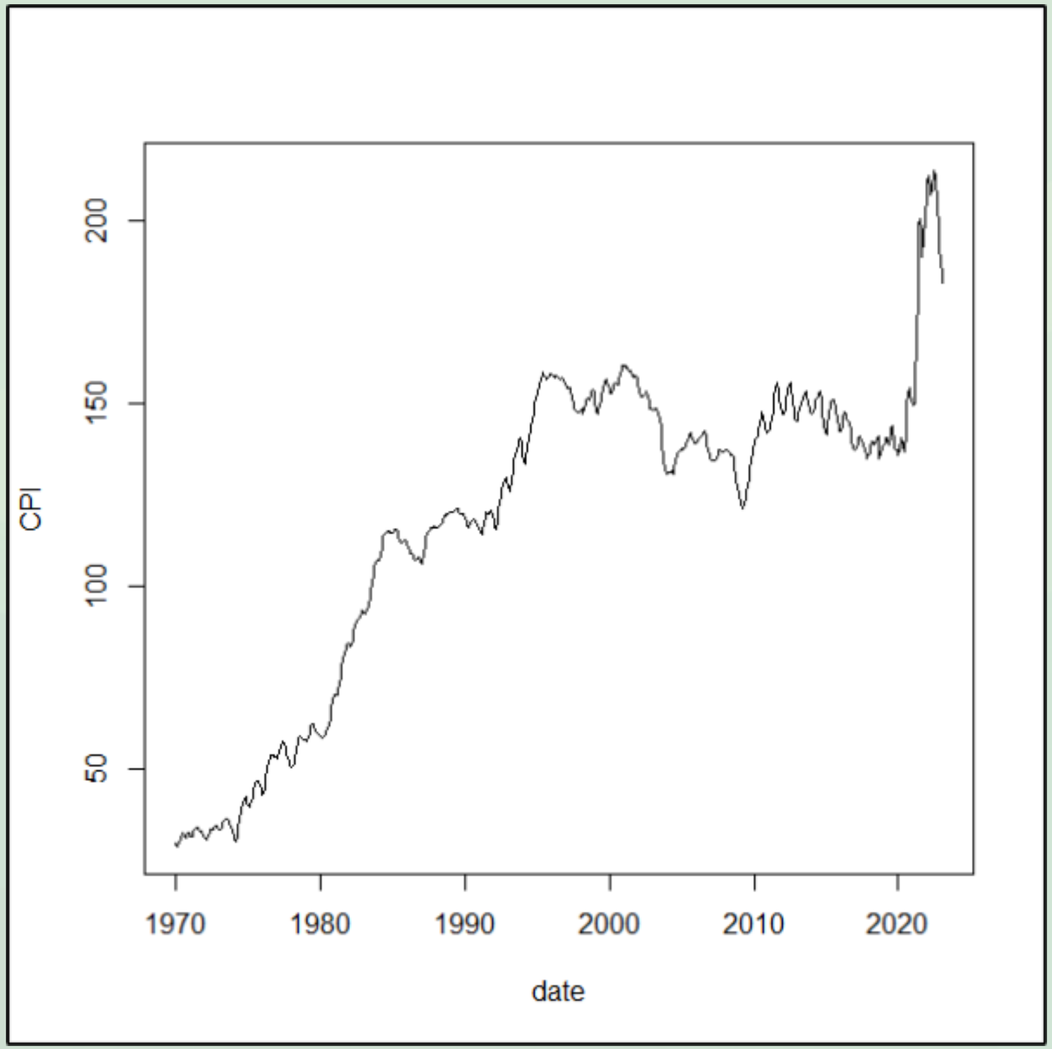
What is the forecasted CPI for used cars/trucks in U.S. cities over the next 3 years?

(Mar. 2023 - Feb. 2026)

Can the model's forecasted data be trusted? Why or why not?

What effects did the COVID-19 pandemic have on the CPI of used vehicles?

Raw Time Plot

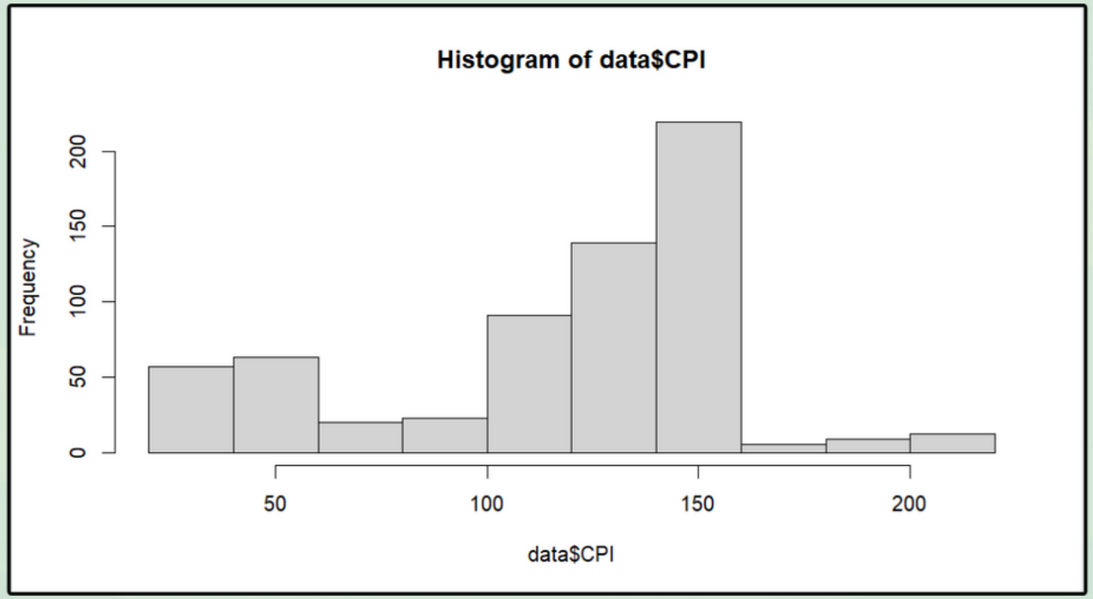


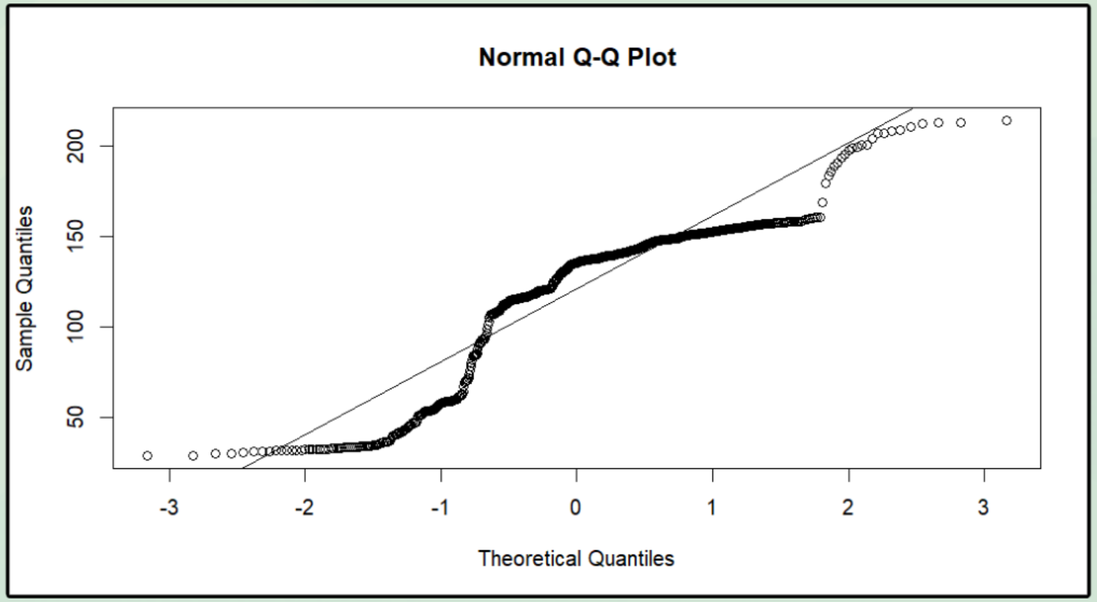
Notes:

- If deterministic, try a linear trend. Could be stochastic as well.
- Try log transformation to reduce variance.
- Can't tell if there is seasonality immediately.

Data Transformation, Trend Removal

Check for Normality





Shapiro-Wilk normality test

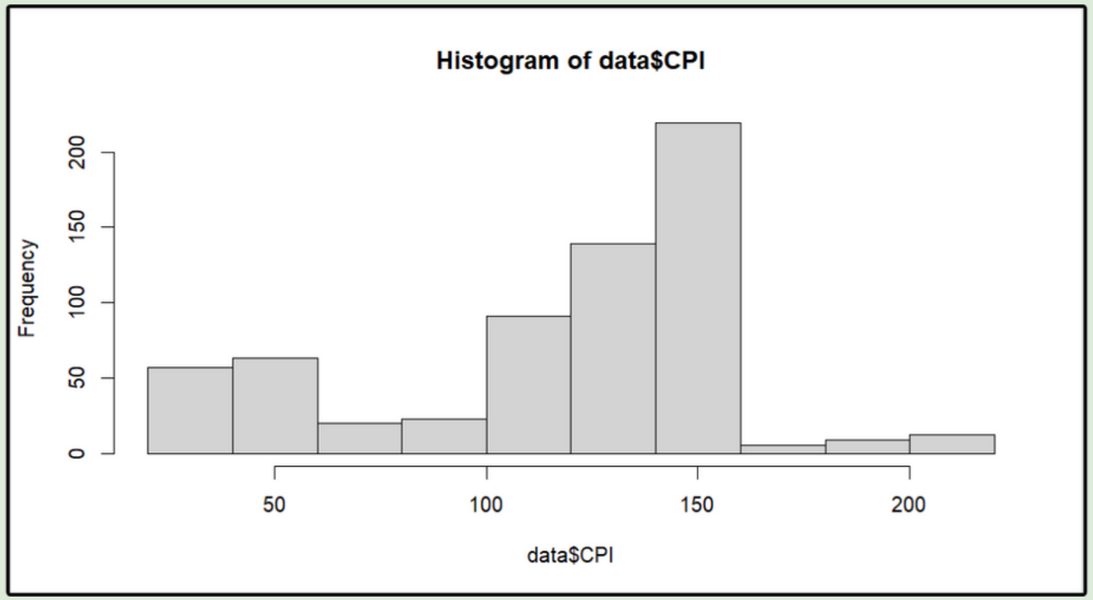
data: data$CPI
W = 0.88655, p-value < 2.2e-16

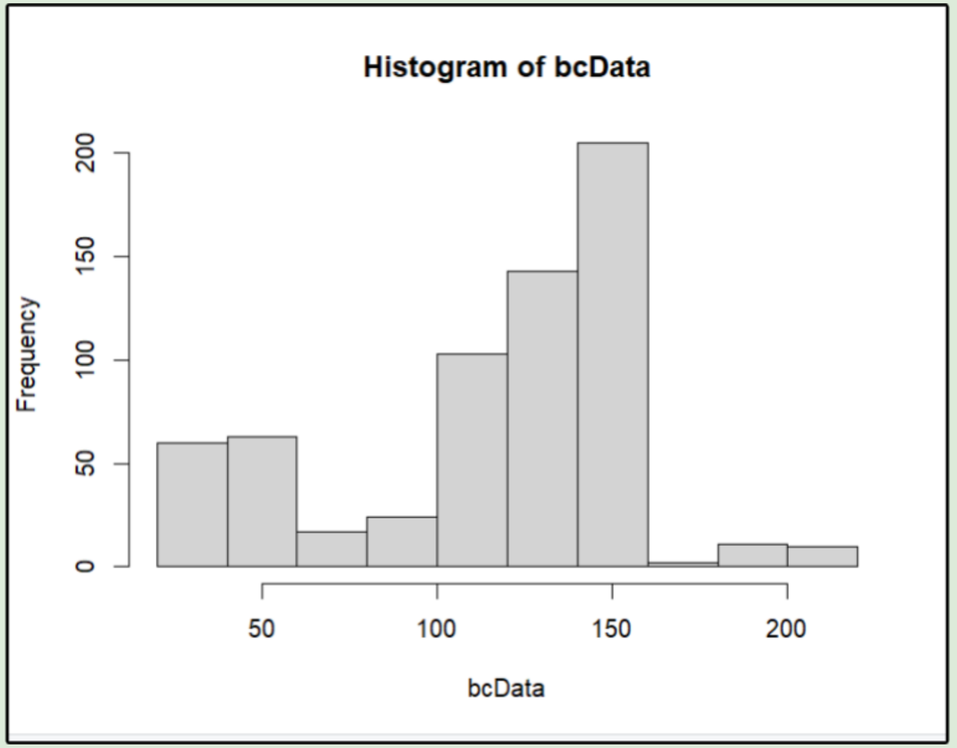
Normality

TIMMERMAN INDUSTRIES

Box-Cox for Normality

lambda = 1.3939





Shapiro-Wilk normality test

data: bcData
W = 0.88655, p-value < 2.2e-16

Normality, so not used

Check for Stationarity (d = 0)

```
> adf.test(data$CPI)

        Augmented Dickey-Fuller Test

data:  data$CPI
Dickey-Fuller = -2.1741, Lag order = 8, p-value = 0.5046
alternative hypothesis: stationary

> pp.test(data$CPI)

        Phillips-Perron Unit Root Test

data:  data$CPI
Dickey-Fuller Z(alpha) = -5.986, Truncation lag parameter = 6, p-value = 0.7758
alternative hypothesis: stationary

> kpss.test(data$CPI)

        KPSS Test for Level Stationarity

data:  data$CPI
KPSS Level = 7.0543, Truncation lag parameter = 6, p-value = 0.01

Warning message:
In kpss.test(data$CPI) : p-value smaller than printed p-value
```

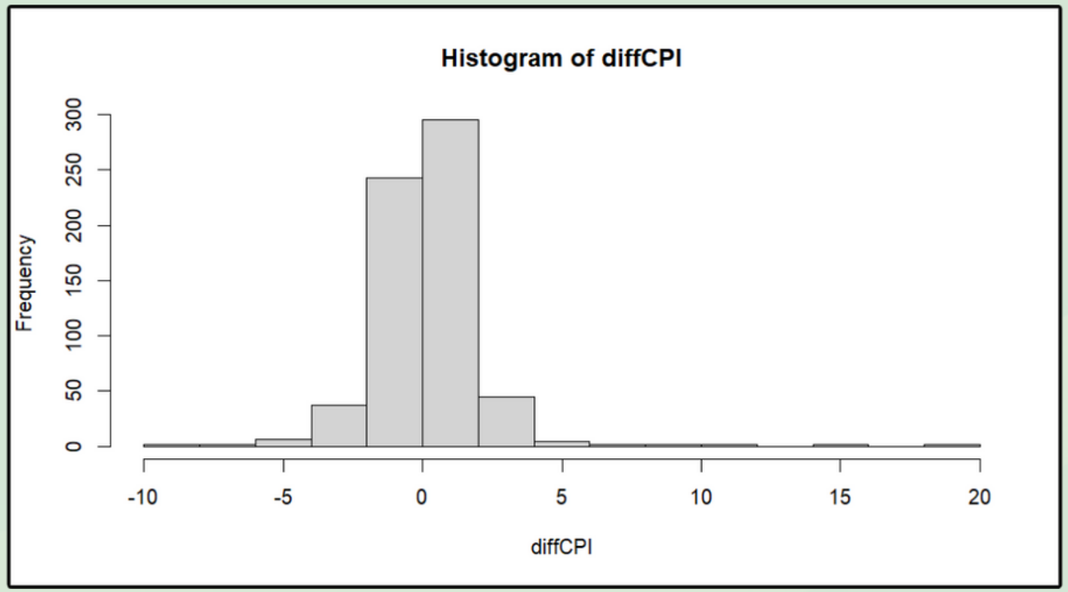
- All three tests claim not stationary.
- Take the first difference of the data and try again.

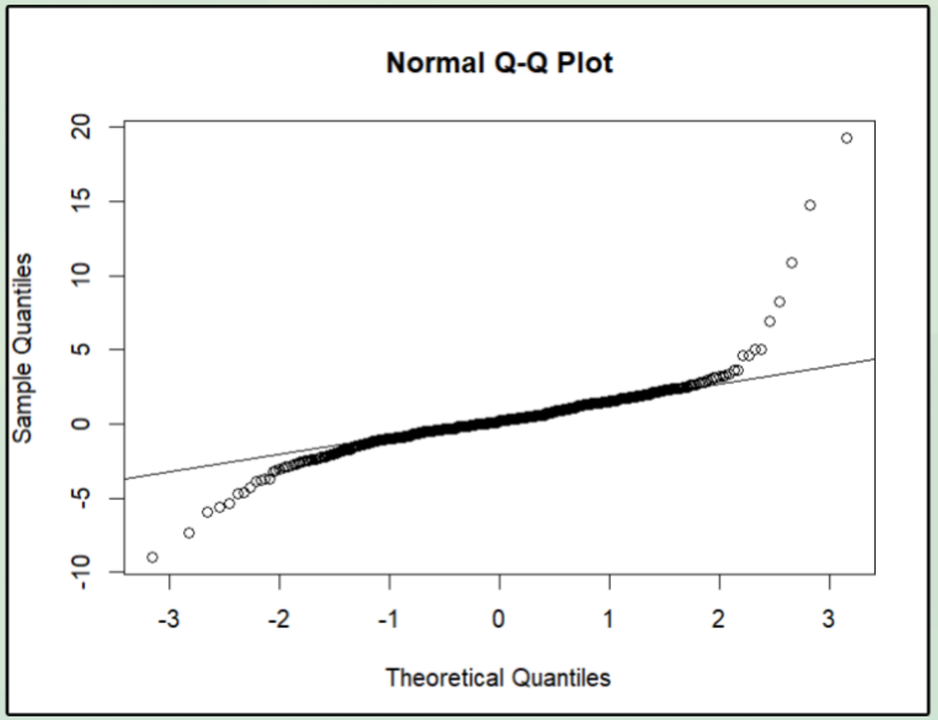
Check for Stationarity (d = 1)

```
> adf.test(diffCPI)

        Augmented Dickey-Fuller Test

data:  diffCPI
Dickey-Fuller = -7.4288, Lag order = 8, p-value = 0.01
alternative hypothesis: stationary

Warning message:
In adf.test(diffCPI) : p-value smaller than printed p-value
> pp.test(diffCPI)

        Phillips-Perron Unit Root Test

data:  diffCPI
Dickey-Fuller Z(alpha) = -200.41, Truncation lag parameter = 6, p-value = 0.01
alternative hypothesis: stationary

Warning message:
In pp.test(diffCPI) : p-value smaller than printed p-value
> kpss.test(diffCPI)

        KPSS Test for Level Stationarity

data:  diffCPI
KPSS Level = 0.12993, Truncation lag parameter = 6, p-value = 0.1

Warning message:
In kpss.test(diffCPI) : p-value greater than printed p-value
```

- All three tests now claim the data is stationary.
- We select d = 1 and proceed with the first differencing data.

Diagnostics for First Difference





Homoscedasticity

Zero Mean Assumption

Stationary (ADF, PP, KPSS)

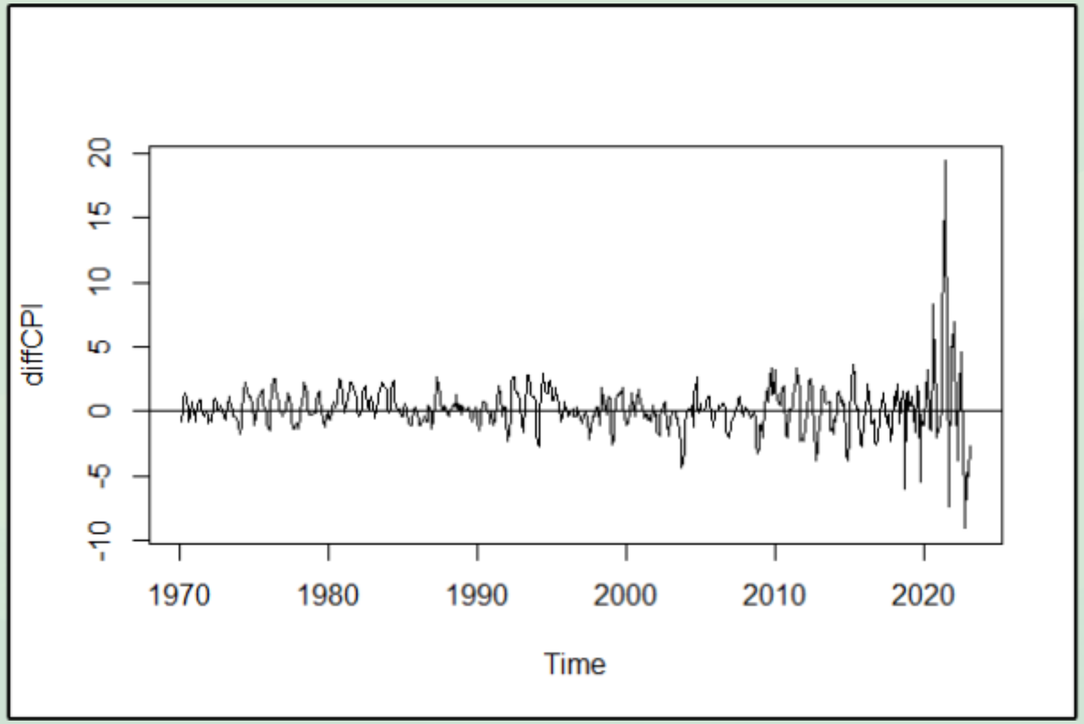
Independence (Runs Test)

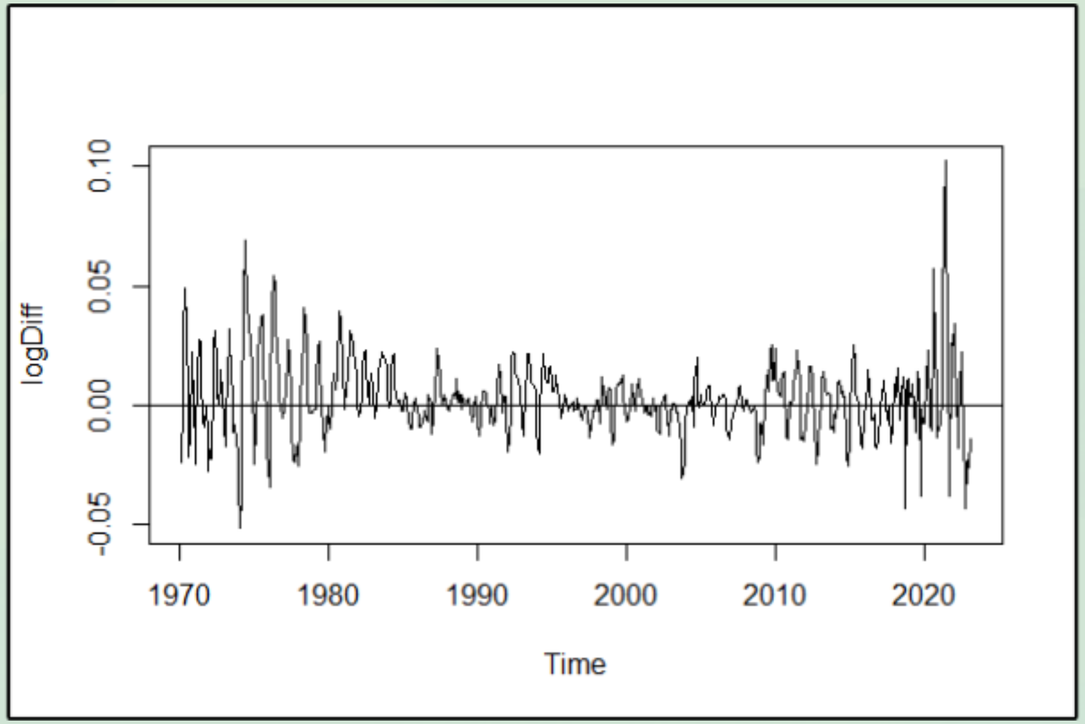
$pvalue [1] 2.68e-50

Normality, but better

W = 0.80486, p-value < 2.2e-16

Is logged data any better?



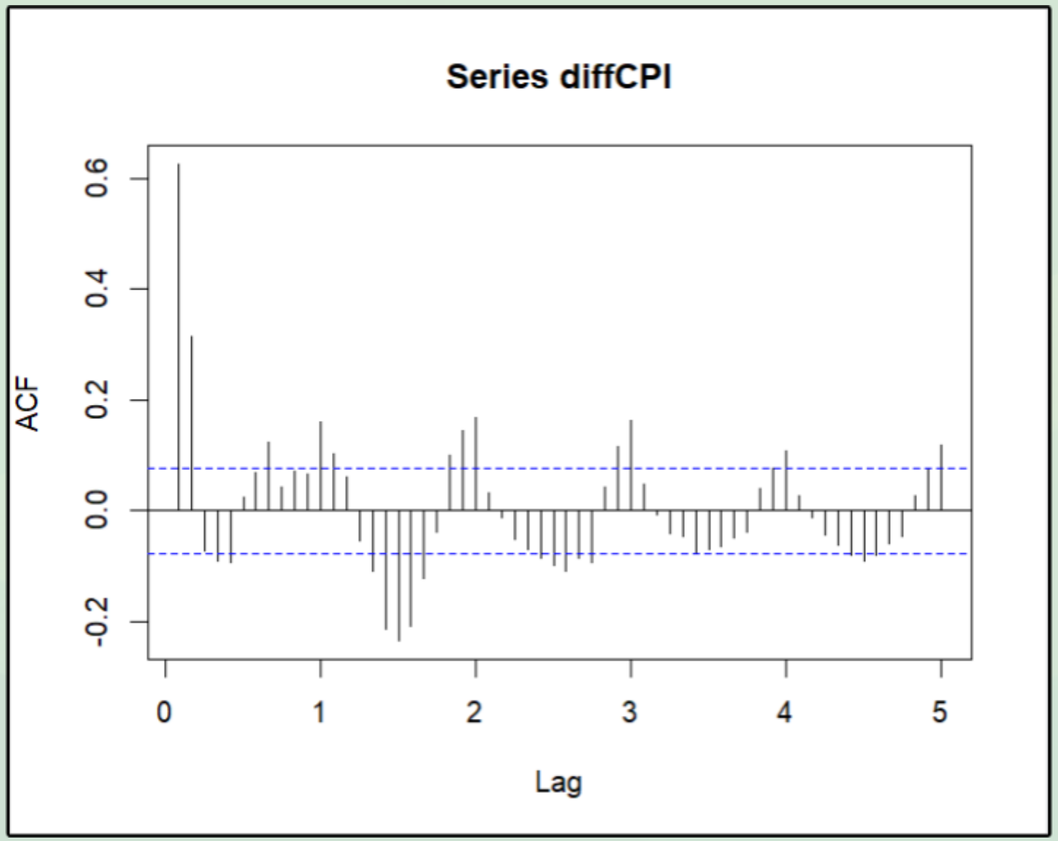


Homoscedasticity, but only 2020-onwards

Homoscedasticity, but a bit better than without log

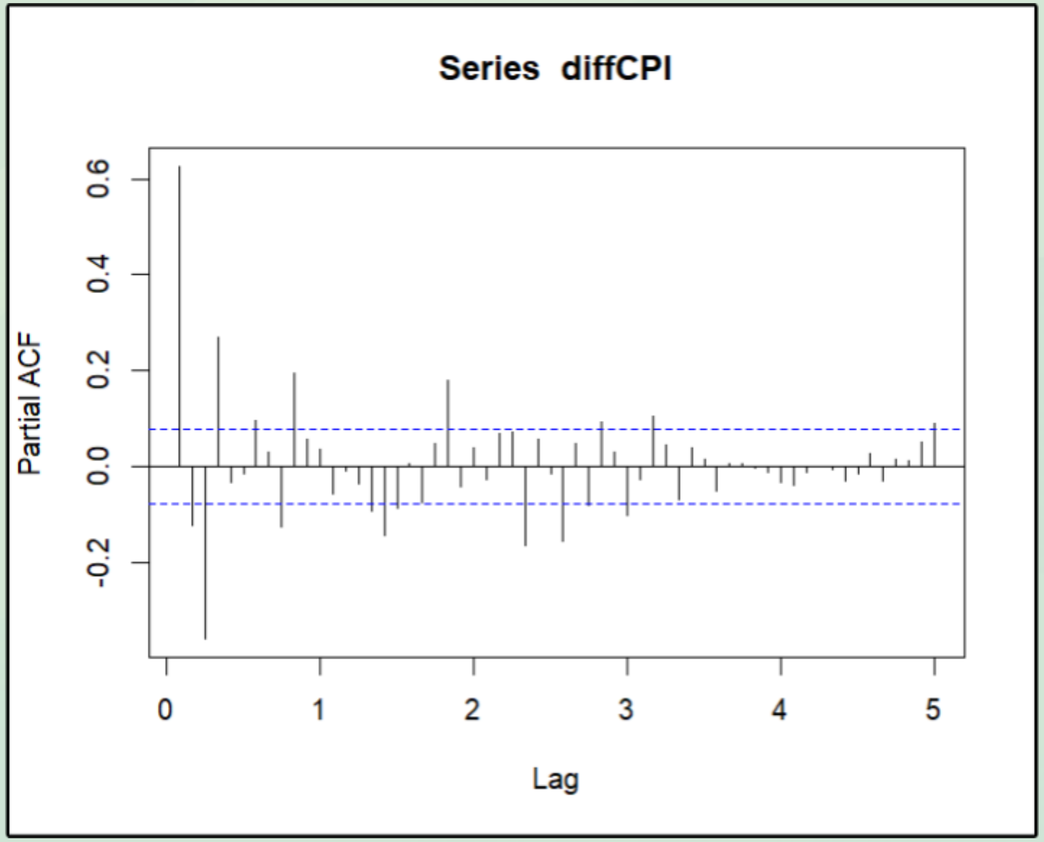
Model Selection

ACF & PACF



Suggests MA(2)

Seasonality at yearly lags suggests SARIMA



Suggests AR(3)

Seasonality cuts off after year 2, suggesting SAR(2)

EACF & auto.Arima

```
AR/MA
  0 1 2 3 4 5 6 7 8 9 10 11 12 13
0 x x o x x o o x o o o  x  x  o
1 x x x o x o o x x o o  x  o  o
2 x x x o x x o x x o o  x  o  x
3 x x x x x o o x o o o  x  o  o
4 x o x o x x o x o o o  x  o  x
5 x o x o x x x o o o o  x  o  o
6 x x o x o x x x o o o  x  o  o
7 x x x x o x x x o o o  x  o  o
```

Suggests:

- MA(2)
- ARMA(4, 1)
- ARMA(1, 3)

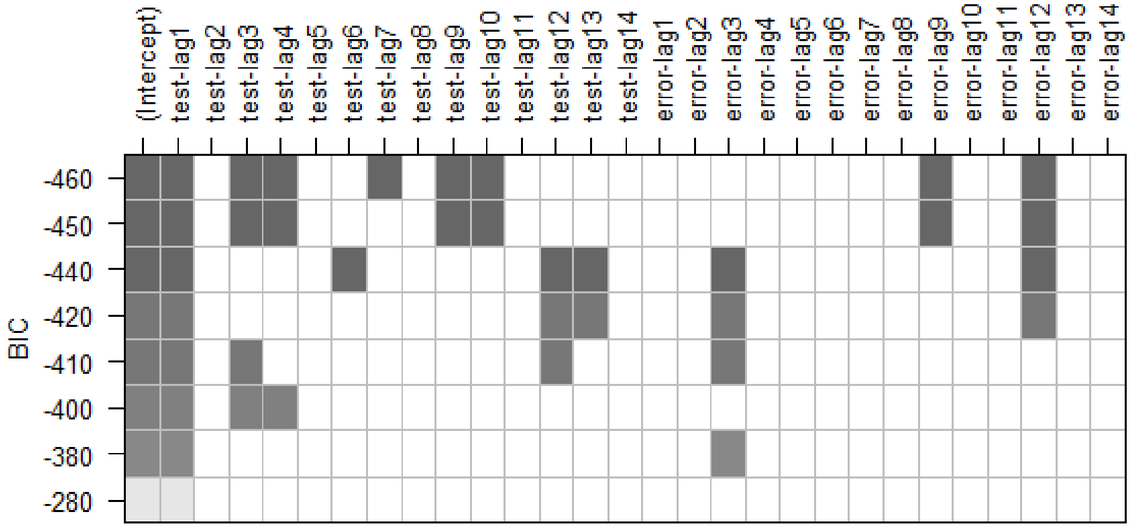
```
Series: diffCPI
ARIMA(4,0,0)(2,0,0)[12] with zero mean

Coefficients:
         ar1     ar2      ar3     ar4    sar1    sar2
      0.8119  0.0363  -0.5289  0.3171  0.1452  0.3777
s.e.  0.0380  0.0453   0.0449  0.0382  0.0393  0.0504

sigma^2 = 1.492:  log likelihood = -1031.11
AIC=2076.21   AICc=2076.39   BIC=2107.41
```

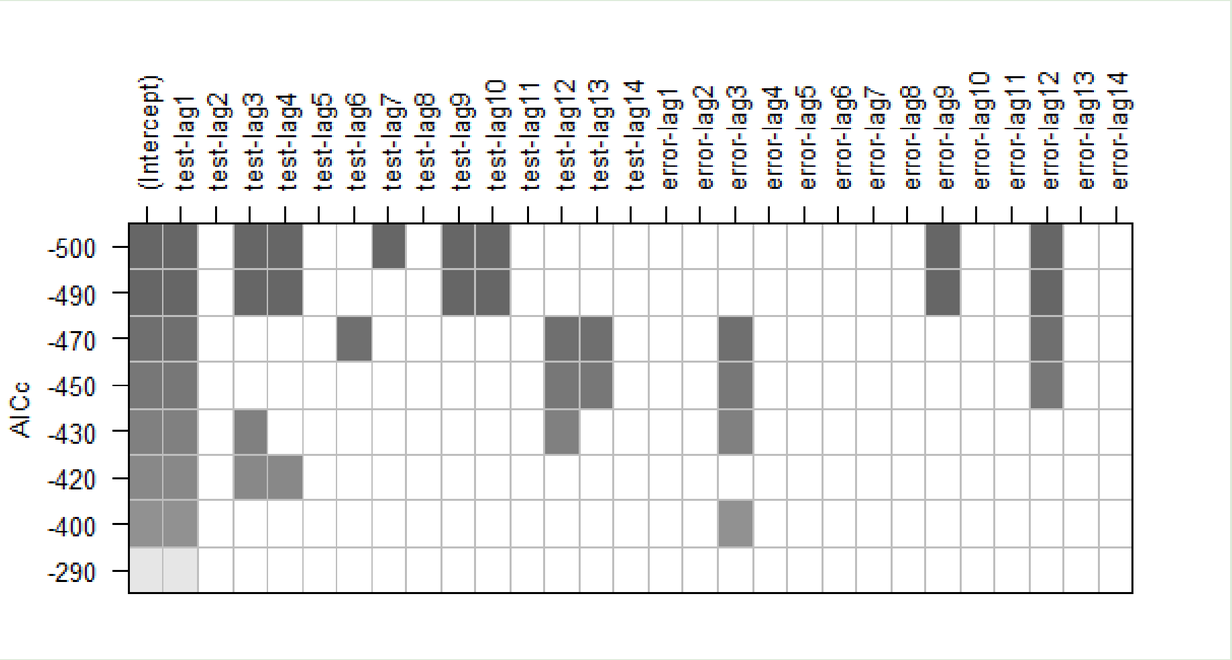
Suggests:

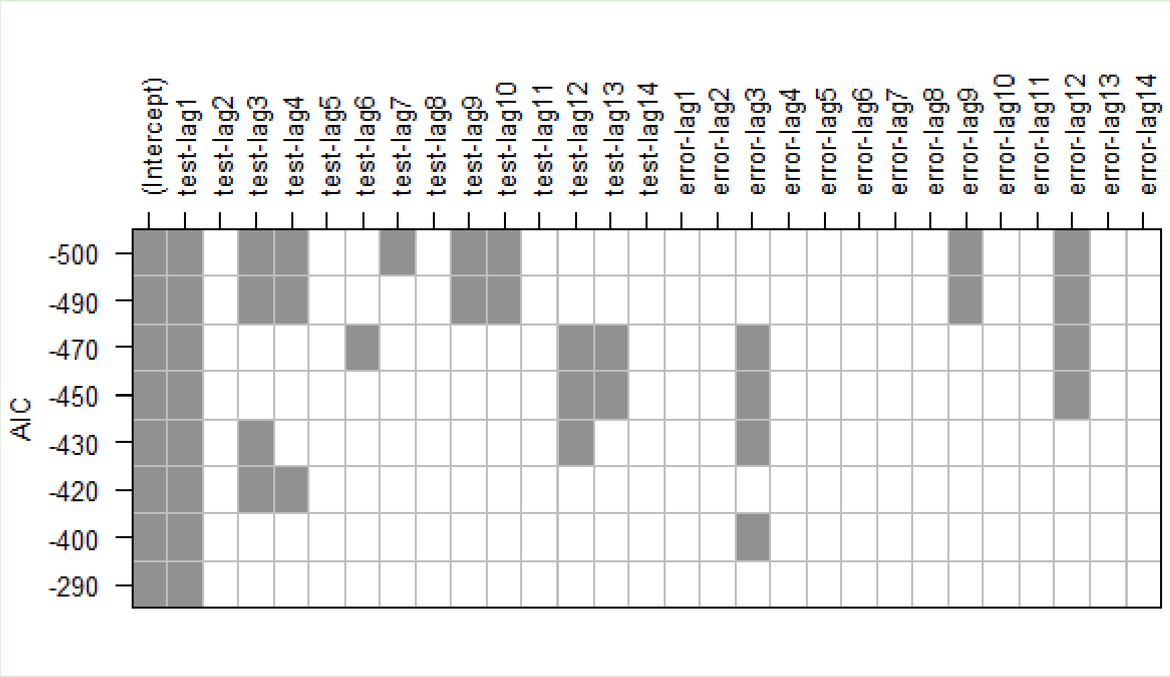
- ARMA(4, 0)
- SAR(2)

Suggests:

- ARMA(1, 3) as the simplest model.
- ARMA(4, 1) as the next-best next-simplest model.







Candidate Models

- **SARIMA(0,0,2)(2,0,0)**
  (from ACF, EACF)
- **SARIMA(1,0,3)(2,0,0)**
  (from EACF, BIC plot)
- **SARIMA(4,0,0)(2,0,0)**
  (from PACF, auto.Arima)
- **SARIMA(4, 0, 1)(2, 0, 0)**
  (from EACF)
- **SARIMA(4, 0, 3)(2, 0, 0)**
  (from BIC plot)

Model Comparison

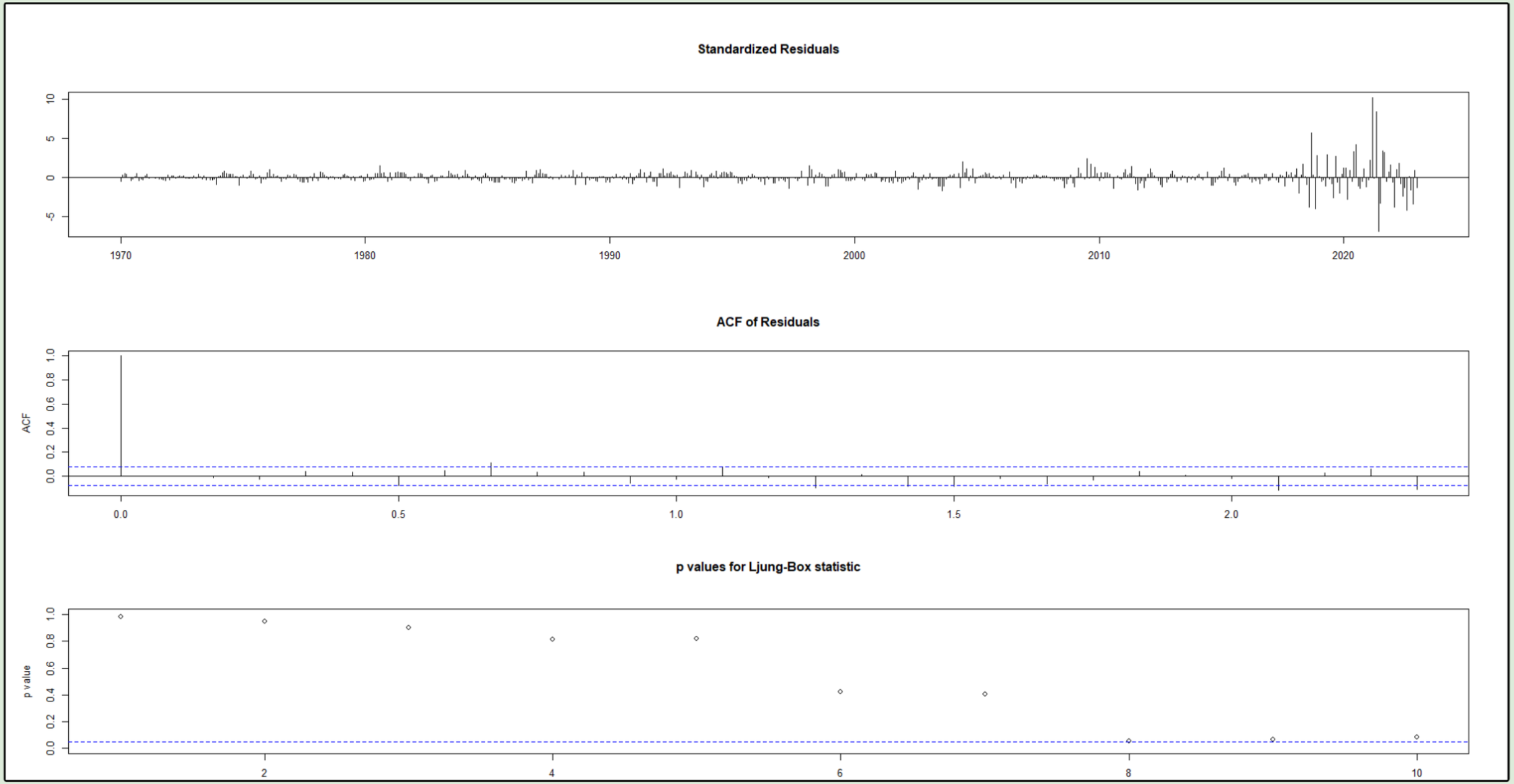
Model	AIC	AICc	BIC
SARIMA(4,0,0)(2,0,0)	**2077.43**	**2077.66**	2113.08
SARIMA(0,0,2)(2,0,0)	2080.33	2080.46	**2107.07**
SARIMA(4,0,1)(2,0,0)	2079.43	2079.71	2119.54
SARIMA(1,0,3)(2,0,0)	2084.12	2084.35	2119.77
SARIMA(4,0,3)(2,0,0)	2078.19	2078.62	2127.22

Model Selection

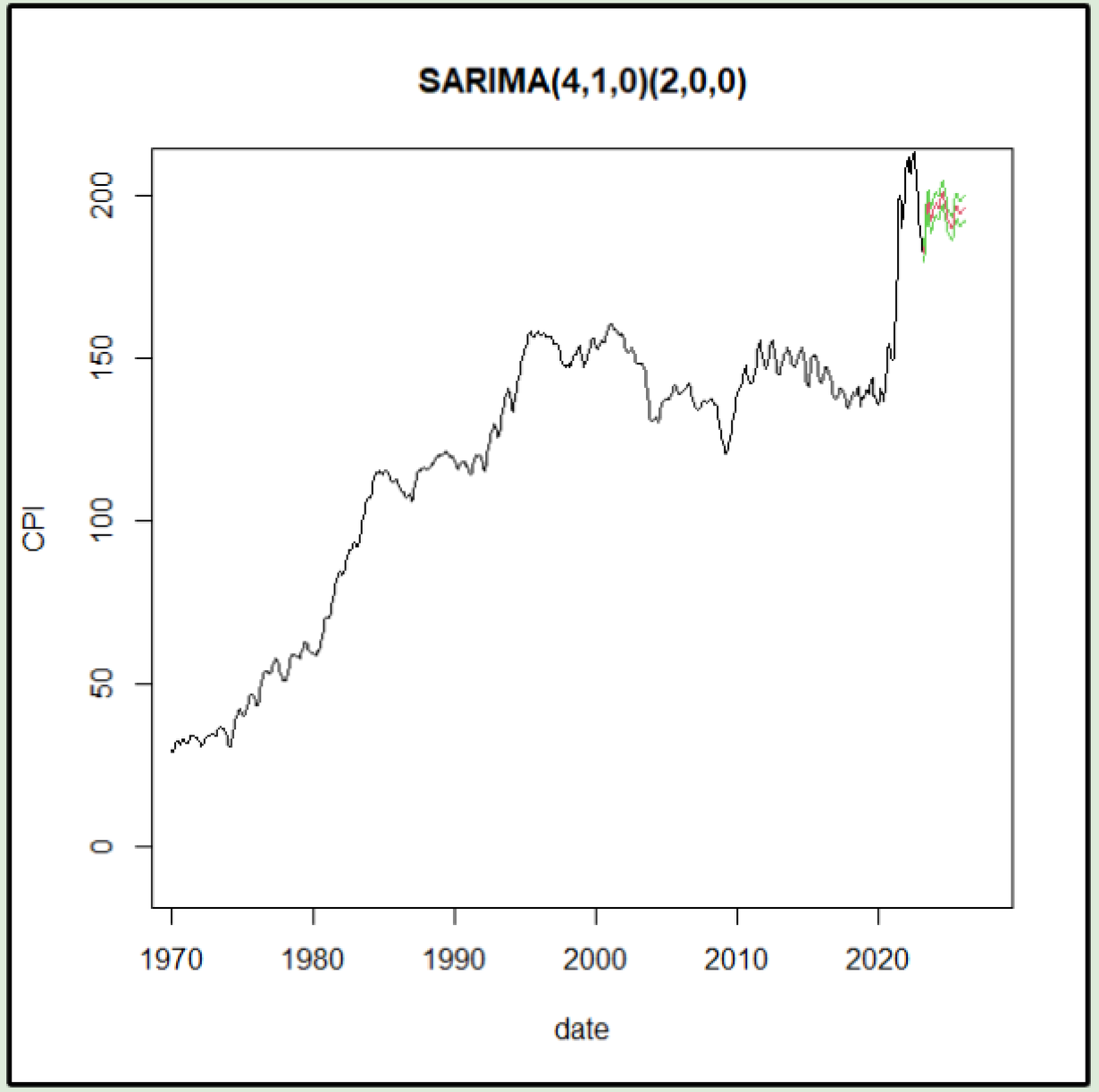
Since we are interested in prediction, we will judge by AIC/AICc.

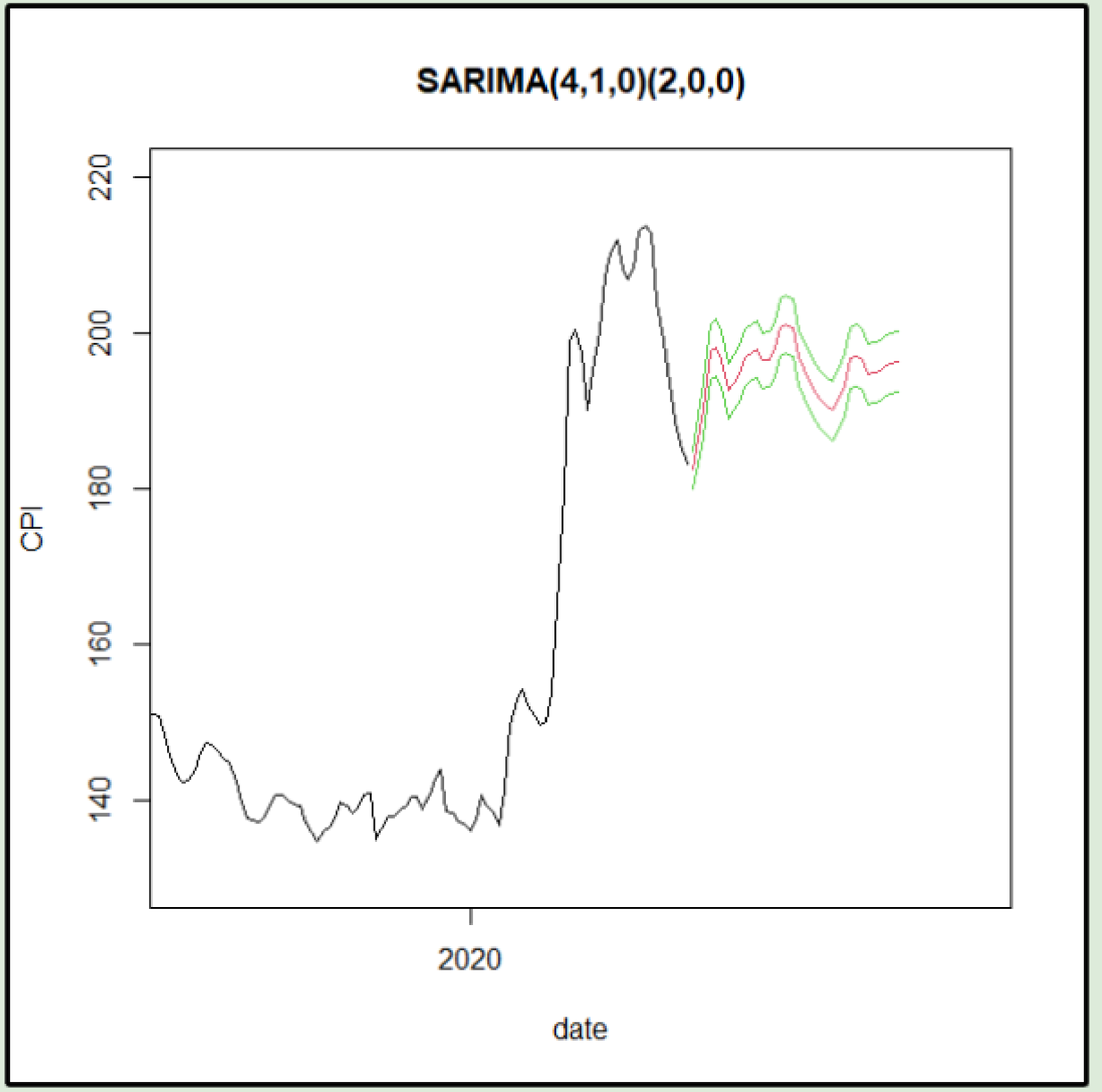
SARIMA(4,0,0)(2,0,0) had the most significant AIC & AICc values.

Also had 2nd best BIC value.


Standardized Residuals
10
5
0
-5
1970
1980
1990
2000
2010
2020
ACF of Residuals
ACF
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.5
1.0
1.5
2.0
p values for Ljung-Box statistic
p value
1.0
0.8
0.6
0.4
0.2
0.0
2
4
6
8
10

Forecasting & Conclusions


SARIMA(4,1,0)(2,0,0)
CPI
0
50
100
150
200
1970
1980
1990
2000
2010
2020
date


SARIMA(4,1,0)(2,0,0)
CPI
140
160
180
200
220
2020
date

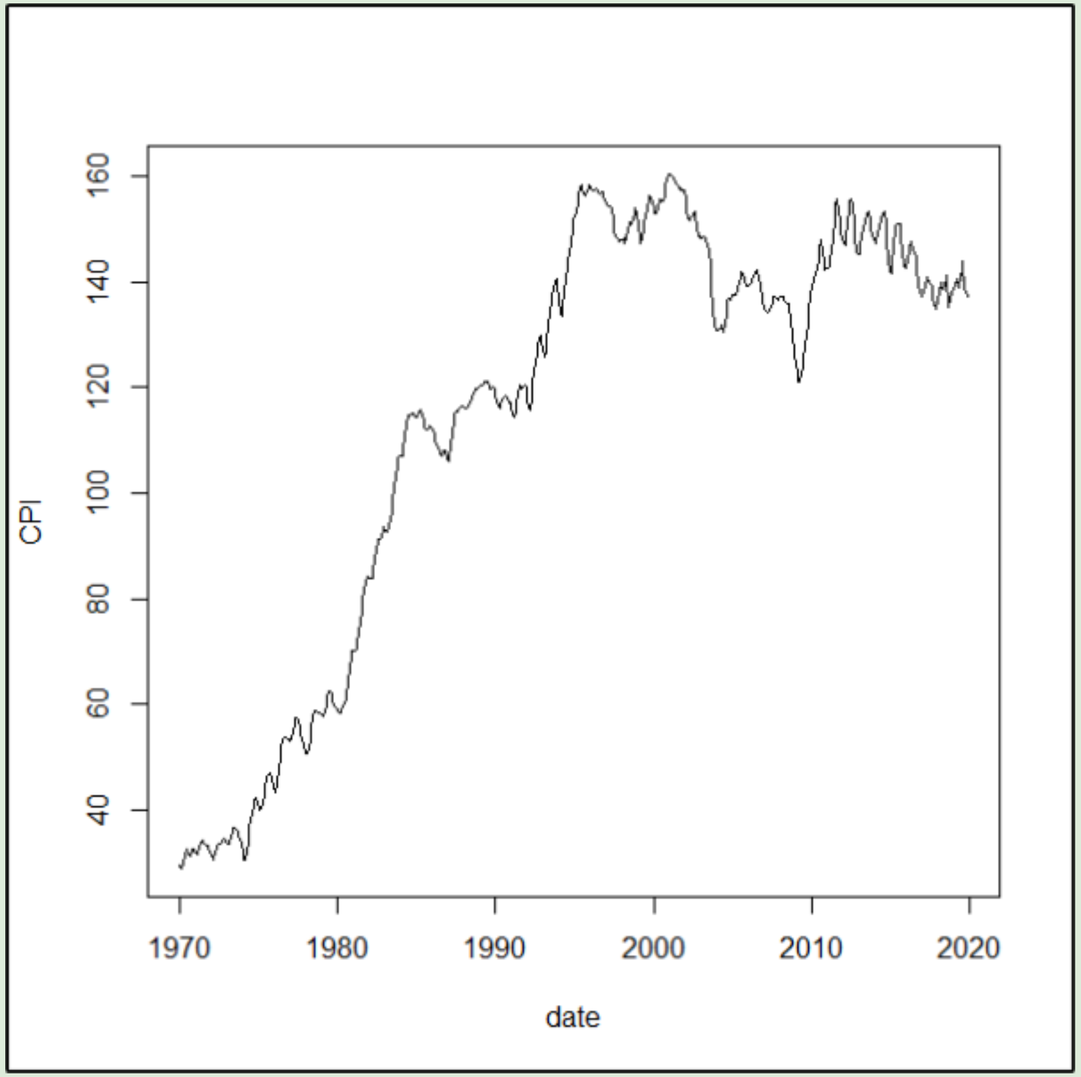
Can we trust this forecast?

No, we should not.

- The heteroscedasticity of the data implies an extremely volatile CPI from 2020-onwards.
- This volatility could also be the reason why other assumptions like normality failed.
- What if we took out the 2020-present data?

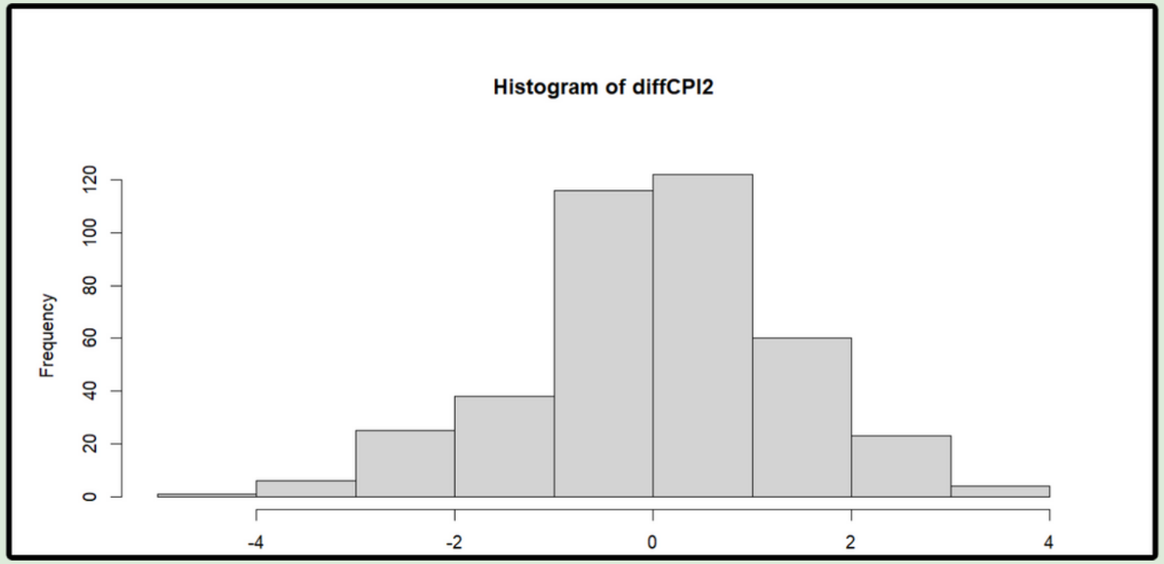
Model 2

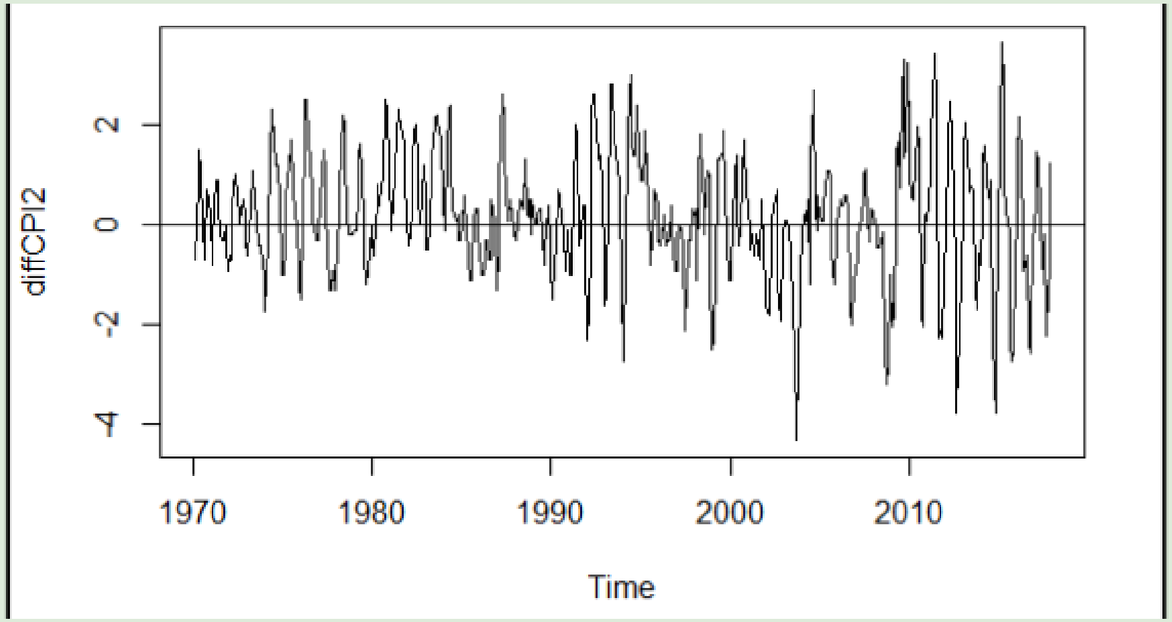
Raw Data Plot, Stationarity




```
> adf.test(data2$CPI)

        Augmented Dickey-Fuller Test

data:  data2$CPI
Dickey-Fuller = -0.57554, Lag order = 8, p-value = 0.9784
alternative hypothesis: stationary

> pp.test(data2$CPI)

        Phillips-Perron Unit Root Test

data:  data2$CPI
Dickey-Fuller Z(alpha) = -1.7148, Truncation lag parameter = 6, p-value = 0.976
alternative hypothesis: stationary

> kpss.test(data2$CPI) # Very much not stationary

        KPSS Test for Level Stationarity

data:  data2$CPI
KPSS Level = 6.7937, Truncation lag parameter = 6, p-value = 0.01

Warning message:
In kpss.test(data2$CPI) : p-value smaller than printed p-value
```

First Difference Diagnostics





Homoscedasticity

Zero Mean Assumption

Stationary (ADF, PP)

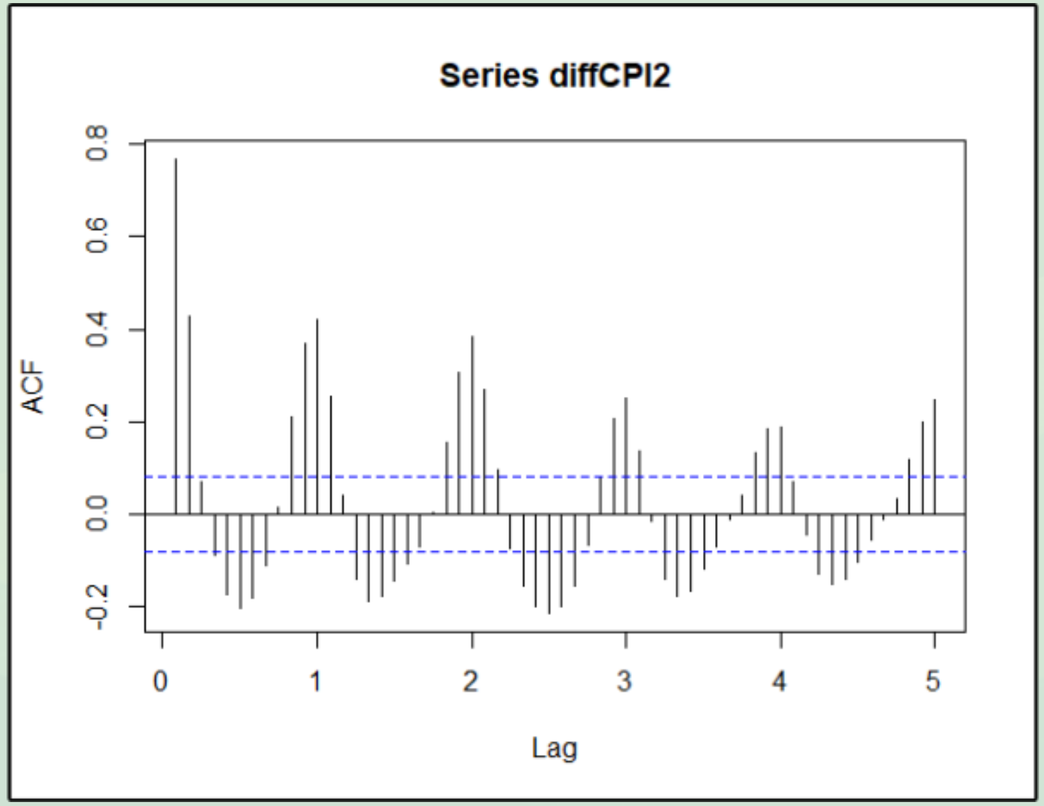
Normality (very close to it)

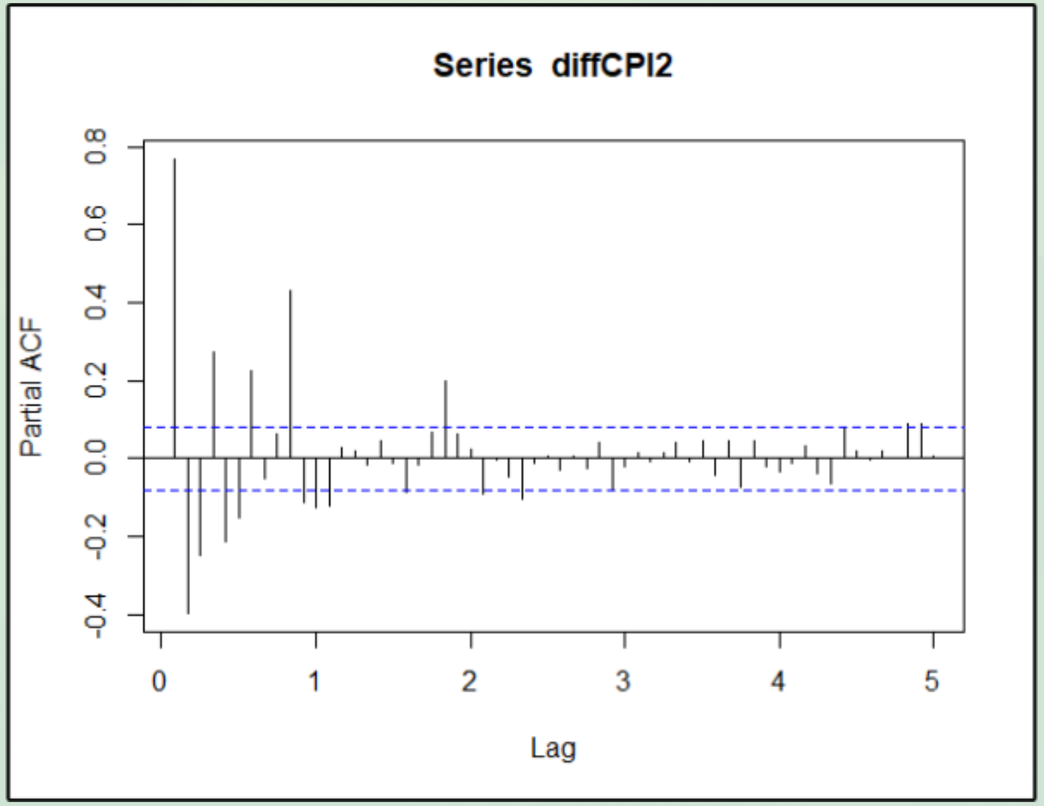
W = 0.99046, p-value = 0.0008906

Independence

$pvalue [1] 6.38e-48

ACF & PACF





Suggests MA(2)

Seasonality at yearly lags suggests SARIMA

Suggests AR(7)

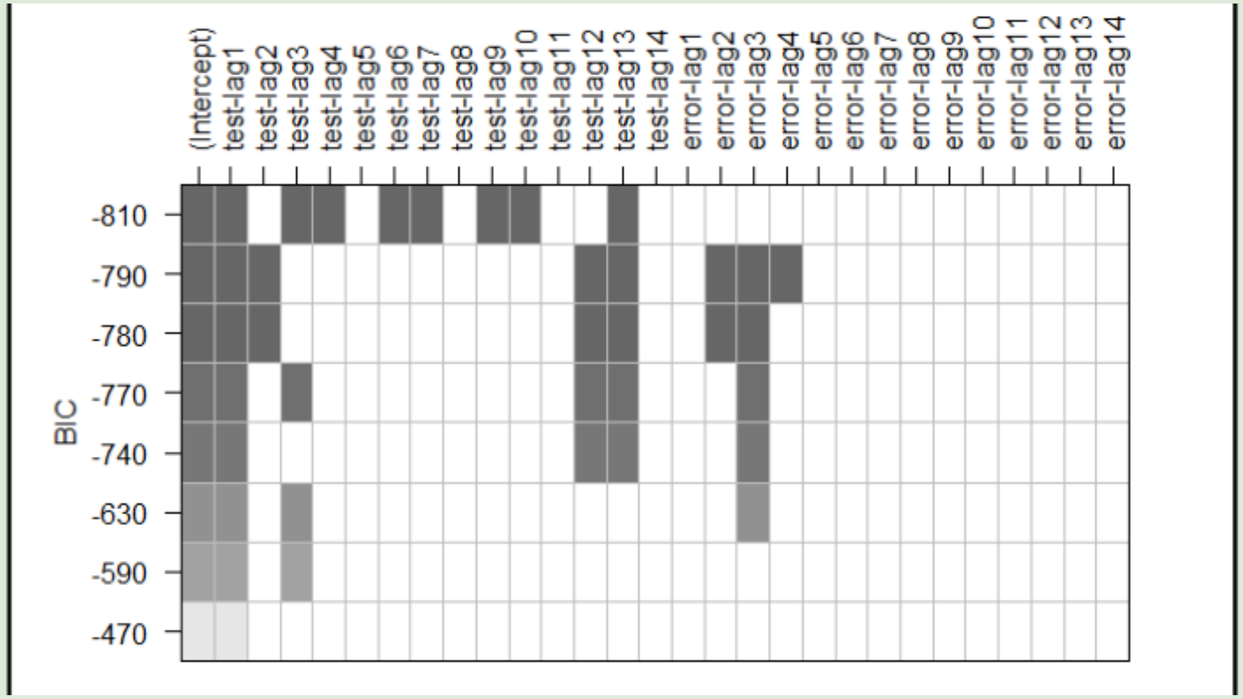
Seasonality cuts off after year 2, suggesting SAR(2)

```
AR/MA
  0 1 2 3 4 5 6 7 8 9 10 11 12 13
0 x x o x x x x x o x x  x  x  o
1 x x x x x x x x o x x  x  x  o
2 x x x x o o o o o o o  x  x  x
3 x x x x o o o o x o o  x  x  o
4 x x x x o o o o x o o  x  x  x
5 x x x x x x o o o o o  x  x  x
6 x x x x o x o x o o o  x  o  x
7 x x x x x o x x o o o  x  o  x
```

Suggests MA(2), ARMA(2, 4)

```
Series: diffCPI2
ARIMA(5,0,0)(2,0,0)[12] with zero mean

Coefficients:
          ar1      ar2      ar3     ar4      ar5     sar1     sar2
      1.0733  -0.1075  -0.5548  0.4538  -0.1477  0.3447  0.2388
s.e.  0.0414   0.0583   0.0533  0.0583   0.0418  0.0408  0.0416

sigma^2 = 0.3445:  log likelihood = -509.01
AIC=1034.02   AICc=1034.28   BIC=1068.86
```

Suggests SAR(2), ARMA(5, 0)

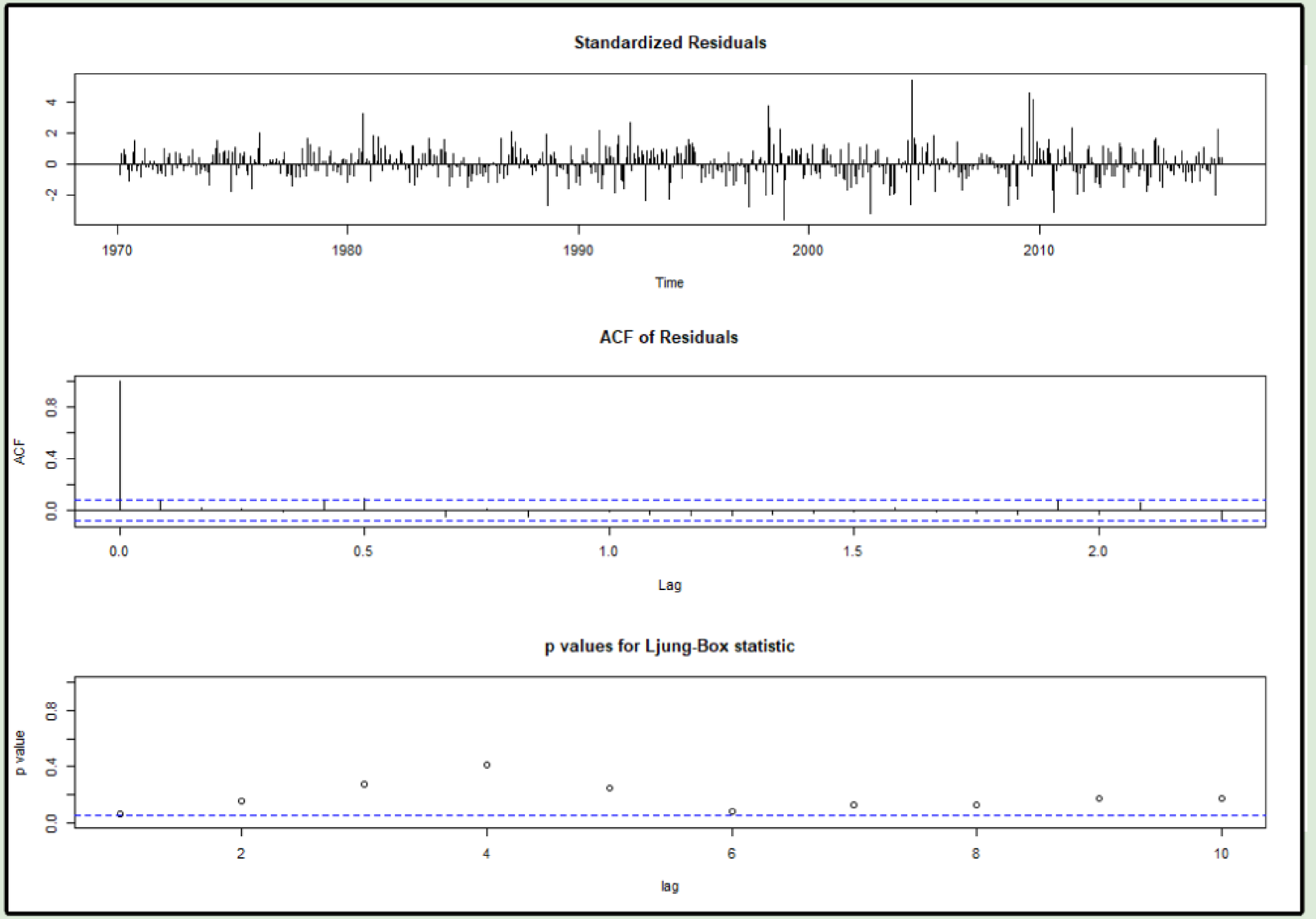


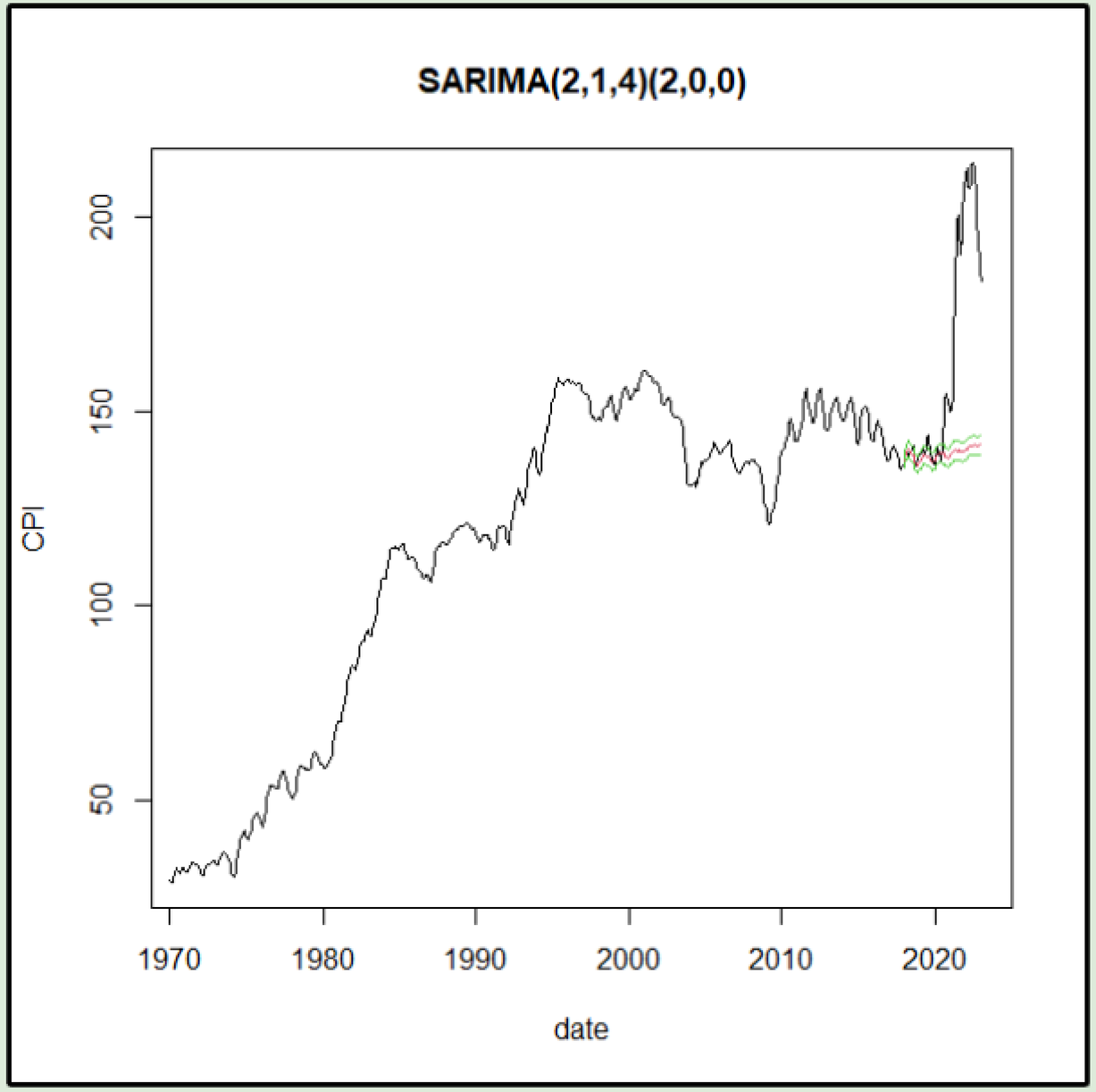
Suggests:

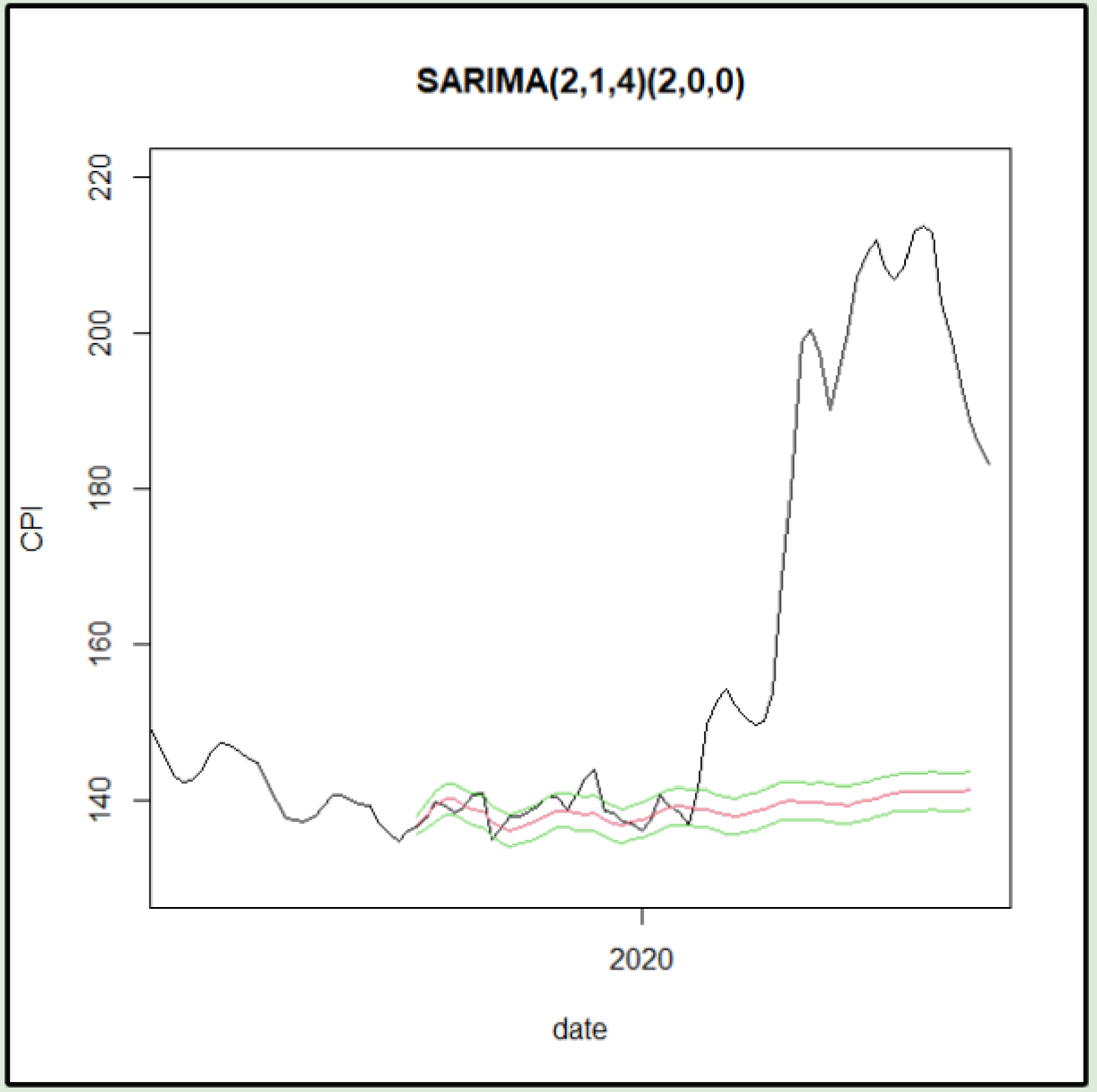
- ARMA(3, 3)
- AR(3)

Candidate Models

- SARIMA(0,0,2)(2,0,0)
- SARIMA(3,0,0)(2,0,0)
- SARIMA(7,0,0)(2,0,0)
- SARIMA(3,0,3)(2,0,0)
- SARIMA(5,0,0)(2,0,0)
- **SARIMA(2,0,4)(2,0,0)**


Standardized Residuals
-2
0
2
4
1970
1980
1990
2000
2010
Time
ACF of Residuals
ACF
0.0
0.4
0.8
0.0
0.5
1.0
1.5
2.0
Lag
p values for Ljung-Box statistic
p value
0.0
0.4
0.8
2
4
6
8
10
lag


SARIMA(2,1,4)(2,0,0)
CPI
50
100
150
200
1970
1980
1990
2000
2010
2020
date


SARIMA(2,1,4)(2,0,0)
CPI
140
160
180
200
220
2020
date

What does this forecast tell us?

Either...

- The 1970-2018 model is wildly inaccurate, and shouldn't be used to predict CPI.

Or...

- The data from 2020-onwards is so volatile that it necessitates the use of a completely different model altogether just to account for this.