Final Project

Mai Castellano, Shannon Long, Zahlen Zbinden

2024-03-09

Introduction

Analysis

Data Wrangling

```
       X             Date               close            open_24h     
 Min.   :   0.0   Length:2787        Min.   :  109.6   Min.   :  109.6  
 1st Qu.: 696.5   Class :character   1st Qu.:  461.1   1st Qu.:  460.4  
 Median :1393.0   Mode  :character   Median : 2781.2   Median : 2773.4  
 Mean   :1393.0                      Mean   : 6530.9   Mean   : 6514.6  
 3rd Qu.:2089.5                      3rd Qu.: 8507.9   3rd Qu.: 8497.9  
 Max.   :2786.0                      Max.   :63347.8   Max.   :63563.7  
    high_24h          low_24h              date           
 Min.   :  119.7   Min.   :   84.33   Min.   :2014-03-14  
 1st Qu.:  470.7   1st Qu.:  451.82   1st Qu.:2016-02-08  
 Median : 2875.7   Median : 2697.46   Median :2018-01-05  
 Mean   : 6711.1   Mean   : 6303.42   Mean   :2018-01-05  
 3rd Qu.: 8764.3   3rd Qu.: 8222.06   3rd Qu.:2019-12-02  
 Max.   :64802.8   Max.   :62095.63   Max.   :2021-10-29  
```

We can see from the descriptive statistics that all attributes are very similar, this indicates that we will really only need to consider one of the columsn for our analysis as they will yield similar results. Lets drop the remaining columns

Lets get a look at the closing price.

Data Visualization

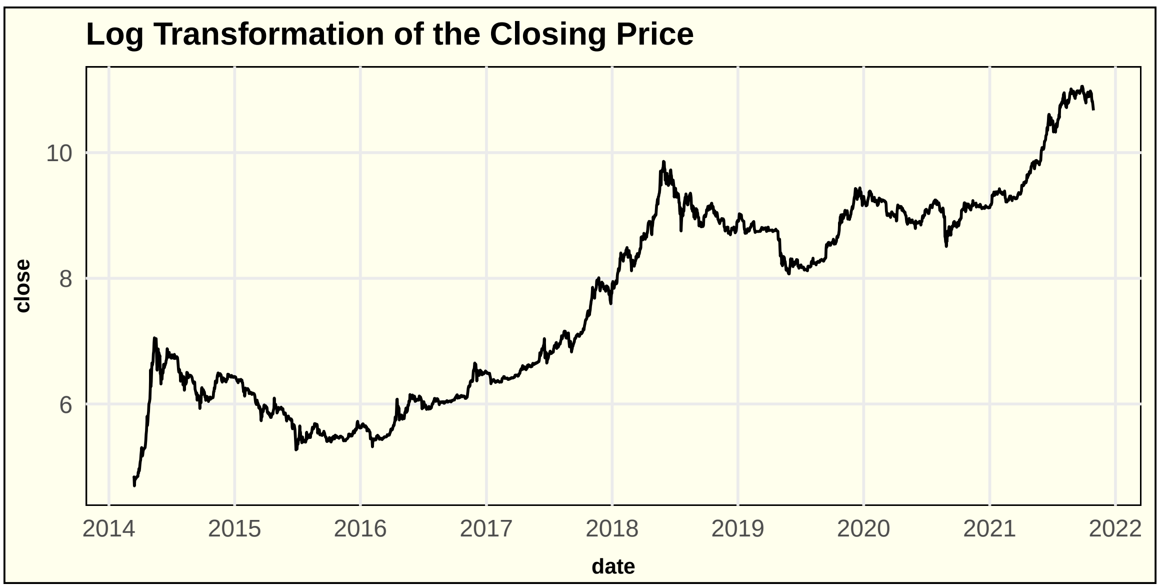
We can see that the date goes from 2014 to 2021, and that the closing price, 24h open, 24h high, and 24h low have roughly the same descriptive statistics, we will only need to consider one of the columns for our analysis as they will yield similar results. This data does not appear to be stationary as is, as it has high volitility (non constant variance). We will need to consider transformations to make it stationary.

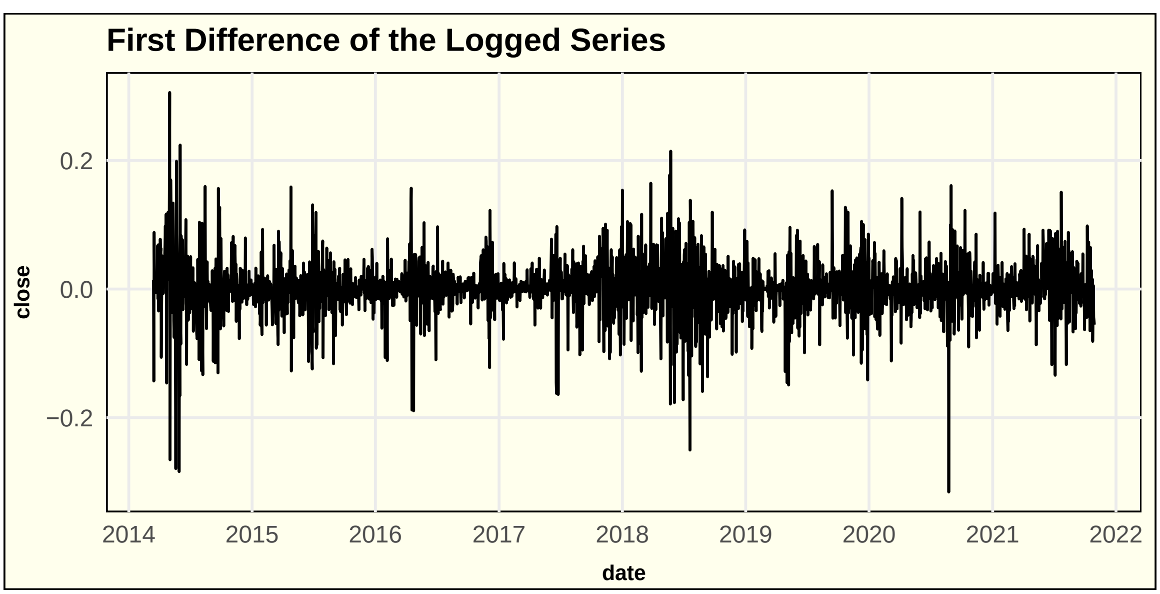
Lets first take a look at the log transformation.

This does appear to help out in making the data stationary, but we shouldn't rule out other transformations. We will also consider the first difference and the square root transformation.

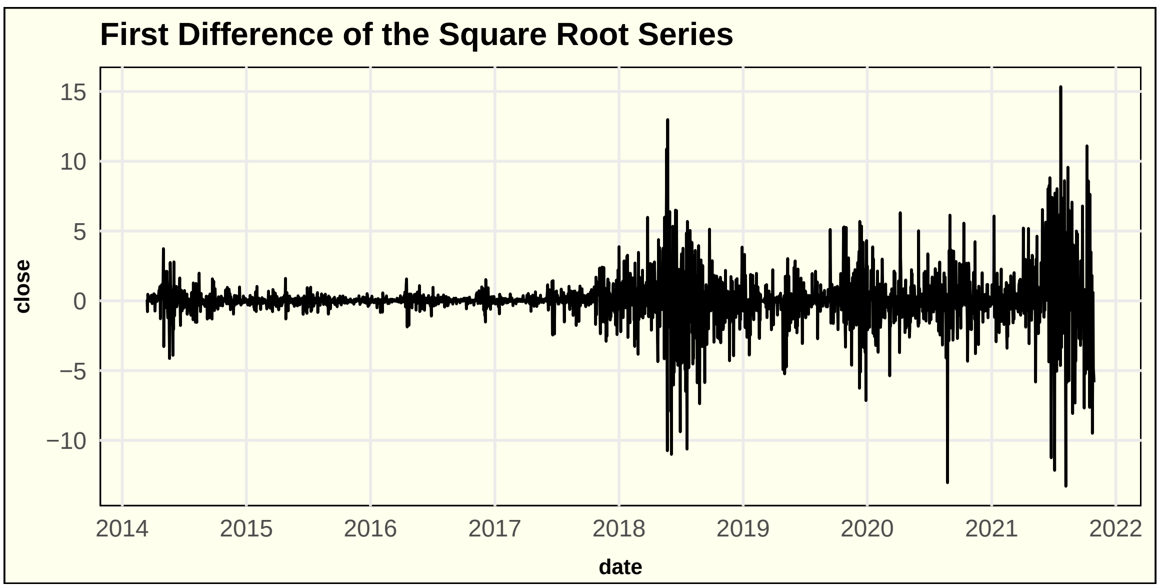
We can see now that the series is much more stationary, and all the trend has been removed. Next we will consider the square root transformation.


Bitcoin Closing Price Over Time
Closing Price (USD)
60000
40000
20000
0
2014
2015
2016
2017
2018
2019
2020
2021
2022
Date


Log Transformation of the Closing Price
close
10
8
6
2014
2015
2016
2017
2018
2019
2020
2021
2022
date


First Difference of the Logged Series
close
0.2
0.0
−0.2
2014
2015
2016
2017
2018
2019
2020
2021
2022
date

At first glance this doesn't appear to be any better then the log transformation, but we will consdier the first difference of the square root transformation.



At this point we can see that the first difference of the logged series is probably the most appropriate transformation for out data, it appears to have less variance over time that the Logged Square Root Transformation. We will use this transformation for the rest of our analysis.

It is important to note that in our investigation up to this point we have not seen any sign of seasonality.

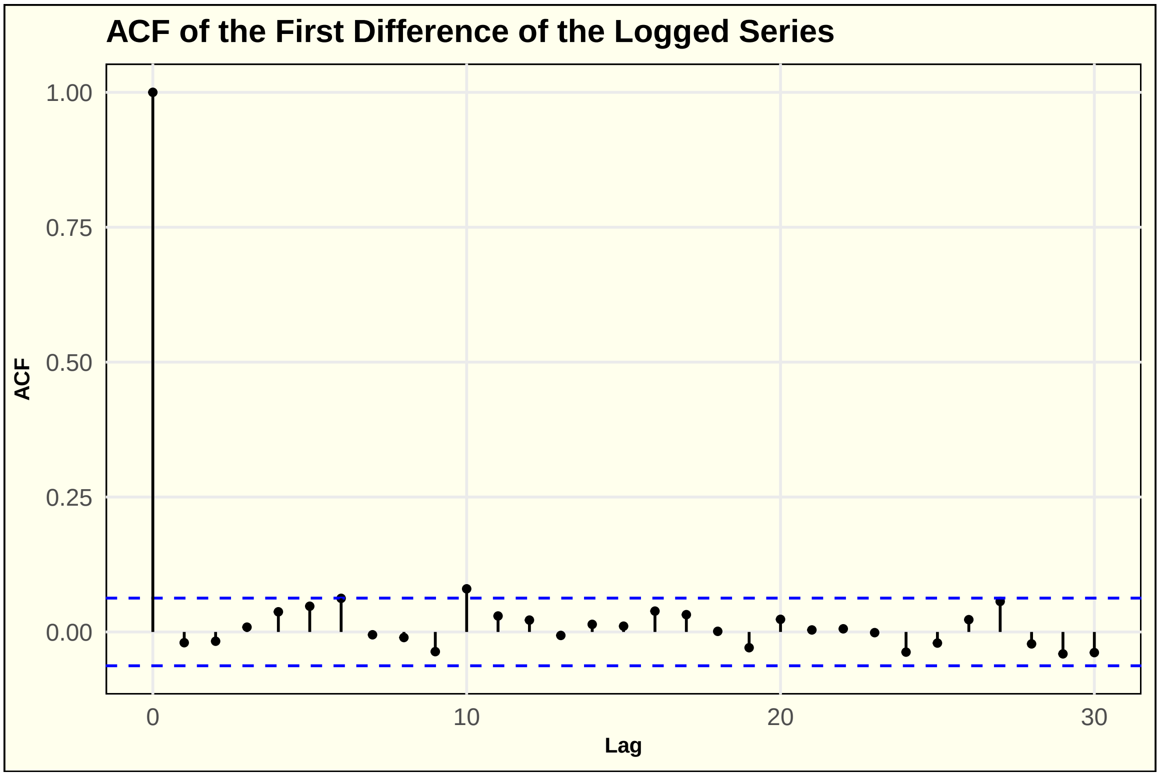
Lets now consider the ACF and PACF for the First Difference of the Logged Series

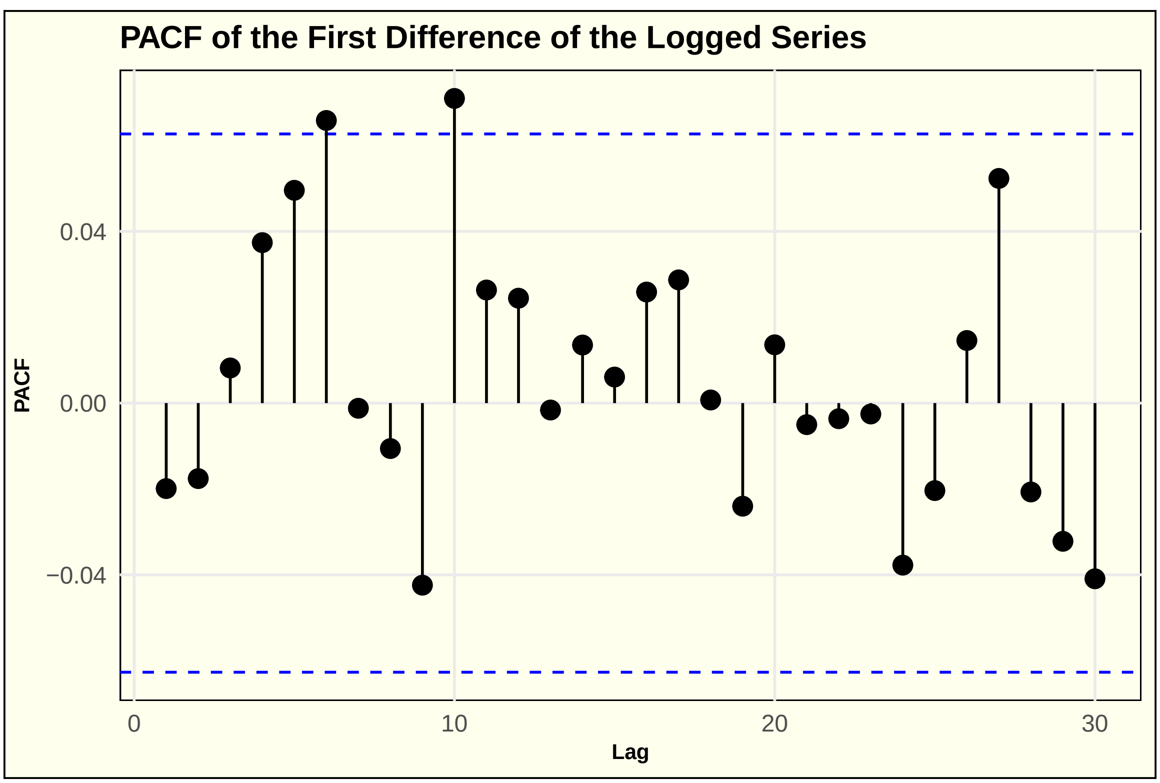
We can see from the ACF that there are no significant values,

Model Fitting

We can see from the ACF that there are significant values at lag 6 and 10. This would suggest that we could try a higher order ARIMA(6,1,0)(0,0,0) with only Auto Regressive features. Let us also check this time series with the forecast package to determine if there is any other models that we should consider.

```
Series: log_dif_df$close
ARIMA(4,0,1) with non-zero mean
```


ACF of the First Difference of the Logged Series
1.00
0.75
0.50
0.25
0.00
ACF
0
10
20
30
Lag


PACF of the First Difference of the Logged Series
0.04
0.00
−0.04
PACF
0
10
20
30
Lag

```
Coefficients:
         ar1     ar2     ar3     ar4      ma1   mean
      0.8198  0.0002  0.0239  0.0364  -0.8457  0.002
s.e.  0.0744  0.0245  0.0246  0.0214   0.0723  0.001

sigma^2 = 0.001809:  log likelihood = 4846.64
AIC=-9679.27   AICc=-9679.23   BIC=-9637.75
```

This suggests that we should consider an ARIMA(4,0,1) model in our analysis as well. As we put in the differenced time series it chose not to do any further differencing which makes its model actually an ARIMA(4,1,1).

It also confirmed that we do not need to consider seasonality components in our model.

We will now consider 3 different model, our higher order AR(6) model, the suggested ARIMA model, and the suggested ARIMA model with an additional MA() order.

ARIMA(6,1,0) ARIMA(4,1,1) ARIMA(4,1,2)

Let us first recall that we will only need to supply the Logged transformation of the series as we are going to be asking it to include the differencing for us.

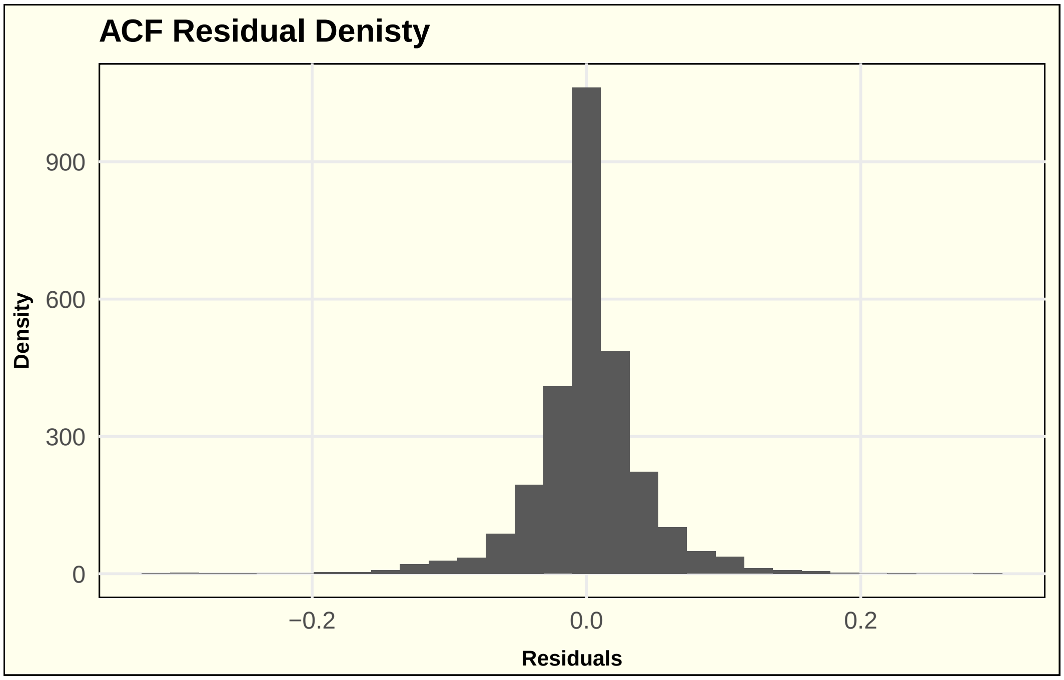
Model Fitting AIC Results

Model	AIC
ARIMA(6,1,0)	-9679.165
ARIMA(4,1,1)	-9677.434
ARIMA(4,1,2)	-9674.832

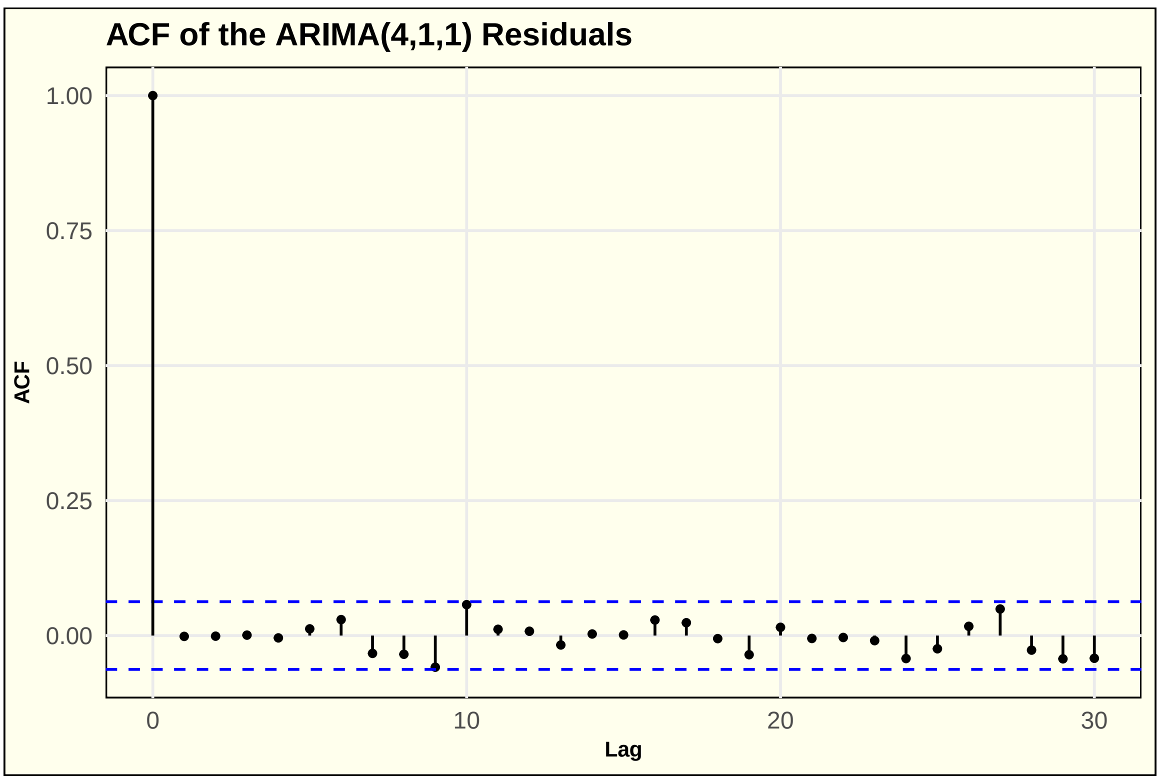
We see from the AIC comparison, that all three models are very similar in their AIC values, but the ARIMA(4,1,1) model has the lowest AIC value. We will use this model for our forecast, next lets look at the residuals and model diagnostics.

Residual Analysis of ARIMA(4,1,1)

```
Don't know how to automatically pick scale for object of type <ts>. Defaulting
to continuous.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

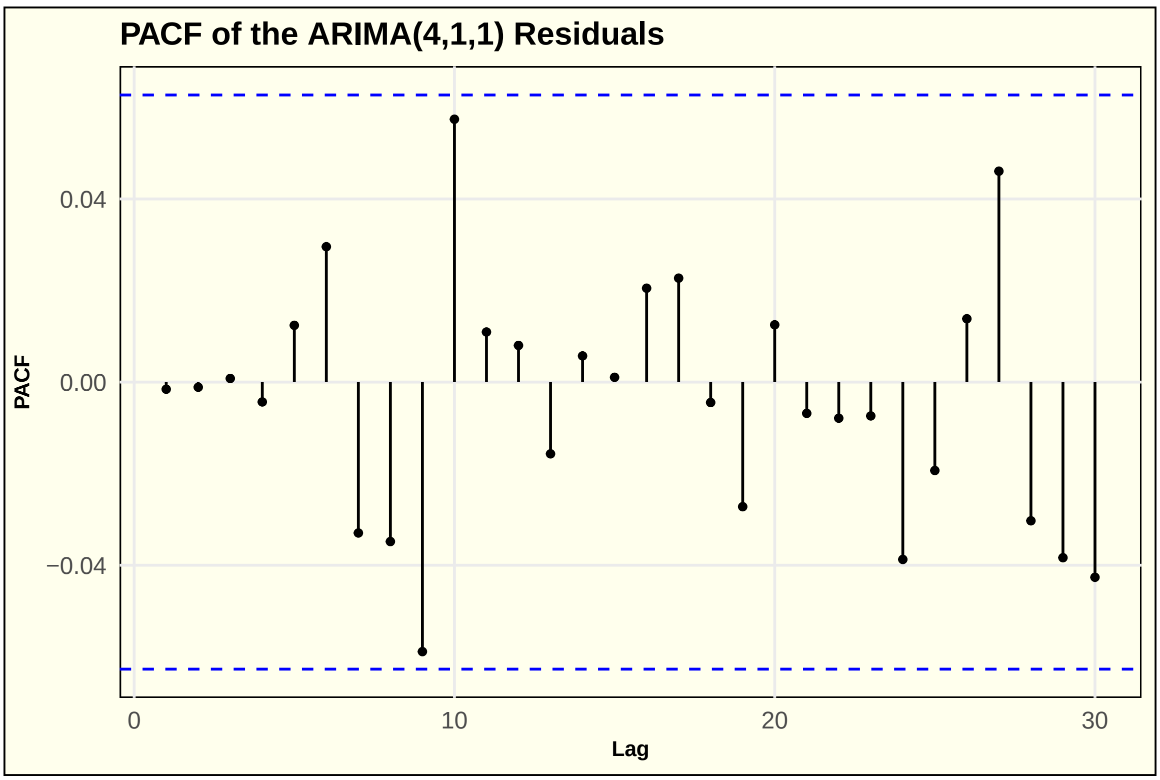


The residuals are normally distributed which is a good sign that our model is behaving well.



Both the ACF of the residuals and the PACF of the residuals are within the confidence intervals, this is a good sign that our model is behaving well.

Prediction


PACF of the ARIMA(4,1,1) Residuals
0.04
0.00
−0.04
PACF
0
10
20
30
Lag