Uppsala University

Gold Price Forecast Using an ARIMA Model

Time Series Analysis Project

Author:
Viktor Ågren

19th September 2022

1 Introduction

In this project, we investigate, analyze and forecast the weekly price of gold from the 30th April 2012 to current time (30th April 2022). In the last couple of years the financial markets have experienced uncertainties in response to macroeconomic concerns around the world such as the recent COVID-19 pandemic, the rising inflation and the crisis in Ukrainian. As a reaction, individual investors are looking towards alternative forms of investments. In the past, a popular solution have been investments into physical gold in a hedge against macro scale troubles.

We fit an *Autoregressive integrated moving average* (ARIMA) model that best described the data in regards to the *autocorrelation function* (ACF) and *partial autocorrelation function* (PACF), and by the use of statistical tests. Lastly, we tried to make predictions for the future price of gold with confidence intervals by dividing the data set to use the first nine years as a training set to predict the last year (52 weeks).

2 Data

We decided to analyze the futures for gold (gc:cmx) as an index for the actual price. The data set was limited to weekly highs on the period between April 2012 and April 2022, resulting in a dataset of 522 data points. It was retrieved from the Nasdaq[1] website which is is an American stock exchange that gathers data on stocks, indices, commodities etc. The gold index follows the futures price of gold in dollars per ounce and it also include figures for analyzing the recessions and inflation adjusted numbers. We could have used other time periods or made comparisons daily/monthly data, but for convenience stick to the aforementioned time period and frequency.

3 Model

In this section, we will try to fit an ARIMA(p, d, q) model to the dataset in the programming language R. In figure 1 we can see the logarithmic transformation of the time series and it has some noticeable periods. Observably, the quick rise in the price of the start of 2020 noting the start of the pandemic. We do not observe any evident seasonality in the time series, therefore there is no reason to consider an *Seasonal Autoregressive Integrated Moving Average* (SARIMA) model.

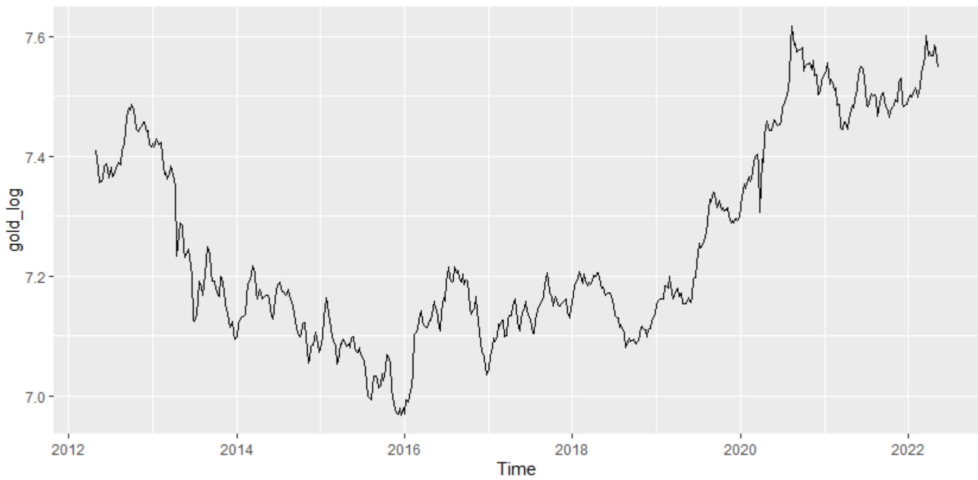
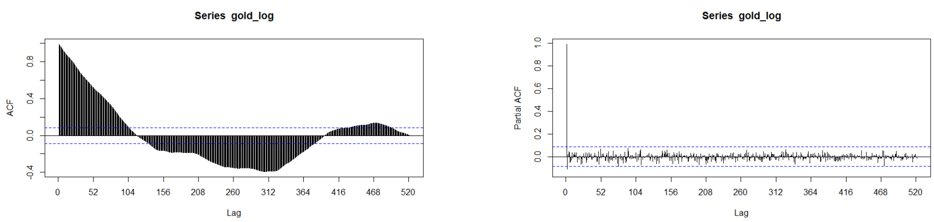


Figure 1: Time series of monthly data.

Initial observations shows a non-stationary time series for the log-transformed data in the given interval. Through out the period in 2013, the price showed a downward trend, while we see an upward trend from early 2015 to late 2020. Another notable observation is the volatility swings in certain periods, such as during the pandemic.

Examining the ACF plot in figure 2 (a), we can see that the data consists of strong autocorrelation up to lag 115 and then it it reverses to weak autocorrelation up to around lag 400. For the PACF in figure 2 (b), we not observe a lot significant autocorrelations except spikes in lag 1, 52 and 68. Assuming the data is not seasonal which we denoted earlier, then we can create a hypothesis that the data is non-stationary.



(a) ACF (b) PACF

Figure 2: ACF and PACF of the time series.

We apply the *Augmented Dickey-Fuller* (ADF) test to this hypothesis. The returned p-value shows 0.623 and we therefore accept the null hypothesis and infer that the time series is indeed non-stationary; Further confirmation is given by the *Kwiatkowski-Phillips-Schmidt-Shin* (KPSS) test. To create an unbiased robust AR-

IMA model, we need to ensure a stationary time series. We achieve this by applying a difference operation with results in figure 3. The logged series differenced at lag 1 oscillates around zero mean indicating stationary. Applying the ADF test, the p-value is printed as less than 0.01 which is significant. Hence, we accept the alternative hypothesis, i.e. that the transformed series is stationary at lag 1 and we proceed to analyze the ACF and PACF for an appropriate ARIMA model.

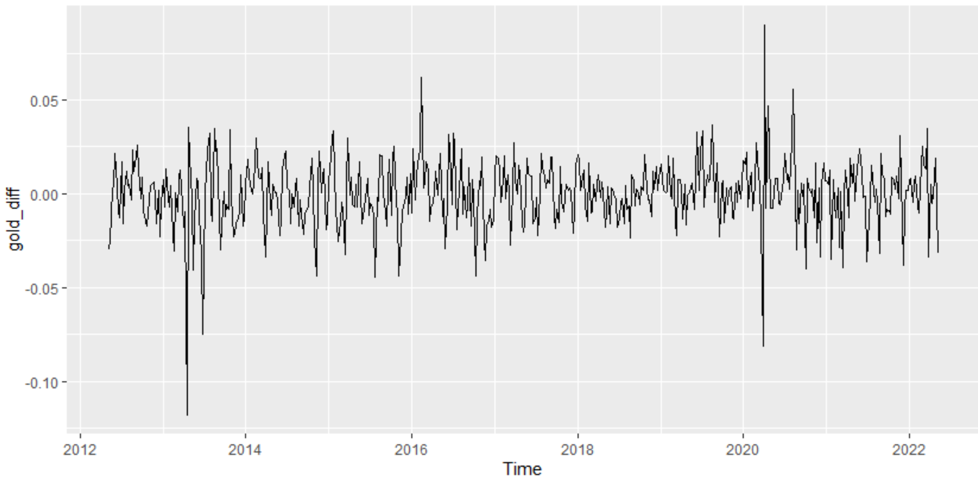
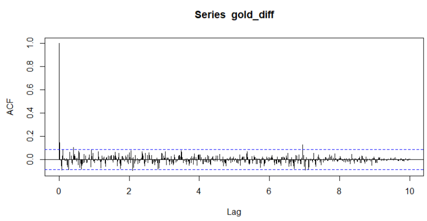
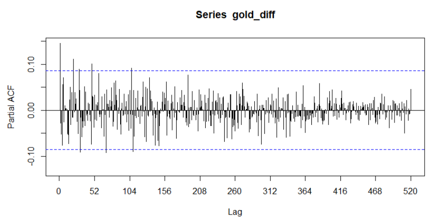


Figure 3: Transformed time series of the monthly data.

Observing the ACF and PACF from figure 4 it is noticeable that the there are some spikes in both figures for smaller lags. The ACF shows cuts off at lower lags indicating the data fits for an AR process. The PACF shows decay from the smaller lags indicating decaying and alternating sequencing, which would fit an MA process. In conclusion, the ARIMA model has characteristics of AR and MA processes in the model. We will take both of this into consideration and test multiple ARIMA models with variations in the AR and MA components.



(a) ACF



(b) PACF

Figure 4: ACF and PACF of the transformed time series.

To get a better understanding on how well the model is at forecasting, we created a data set consisting of 90% of the total series. The final model used for forecasting was the ARIMA(2,1,2) model. We obtained this model by examining several models such as ARIMA(1,1,1) and ARIMA(2,1,1). A determining factor was the estimation of *akaike information criterion* (aic) where the final model had the lowest output of the tested models, as can be inferred below.

```
Coefficients:
         ar1      ar2      ma1     ma2  intercept
      0.5925  -0.7788  -0.4728  0.7465      2e-04
s.e.  0.1075   0.1494   0.1078  0.1744      9e-04

sigma^2 estimated as 0.0003032:  log likelihood = 1236.76,
aic = -2461.53

Training set error measures:
                        ME       RMSE        MAE      MPE
Training set -2.203531e-06 0.01741395 0.01254831 109.3524
    MAPE      MASE       ACF1
163.5583 0.7681409 0.02149901
```

The coefficients of the model are within bounds i.e twice the standard error and none of them are small enough to indicate any over correction or unnecessary components. Further diagnostics were performed on the residuals of the ARIMA(2,1,2) model, this includes a histogram in figure 5 and a normal Q-Q plot in figure 6. Along with the standardized residuals and p-values for Ljung-Box statistic in figure 7. From the histogram we can compare the residuals to the normal distribution line and it suggest that the residuals are indeed normally distributed. We can also see that the figure looks a bit skewed but it is mostly explained by the number of bins shown in the figure. The Normal Q-Q plot can be interpreted to have the same conclusion as the histogram with some deviations in for the extreme values. Figure 7, shows that the standardized residuals look mostly randomized and does not follow a specific pattern with a few outlier spikes. The ACF drops of as expected after lag 0 and does not have any significant lags after. Additionally, the p-values are all well above the 5% cutoff and it follows that the data is independently distributed. All of the points above show that the ARIMA model fits the time series and we will proceed to the forecast.

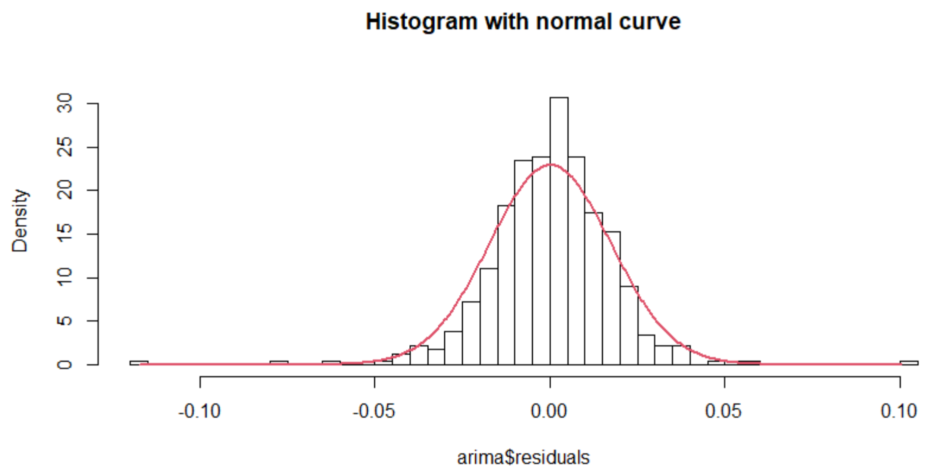


Figure 5: histogram of the residuals

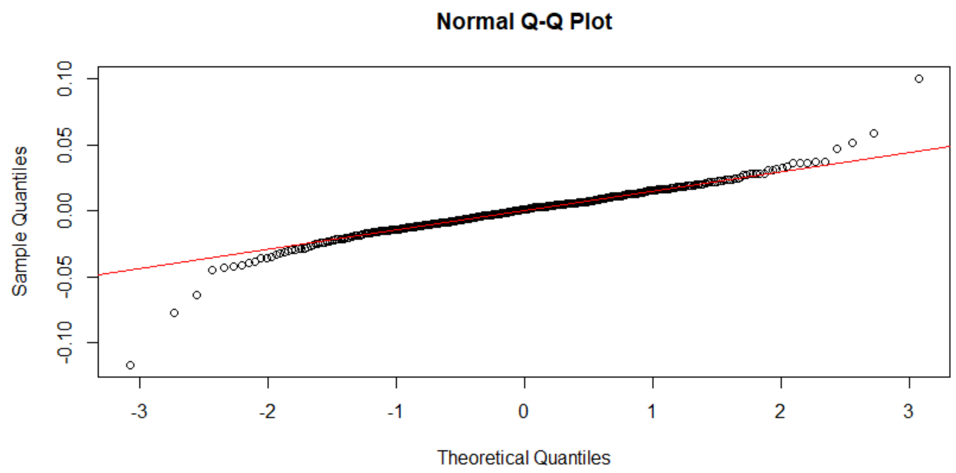


Figure 6: Normal Q-Q Plot of residuals.

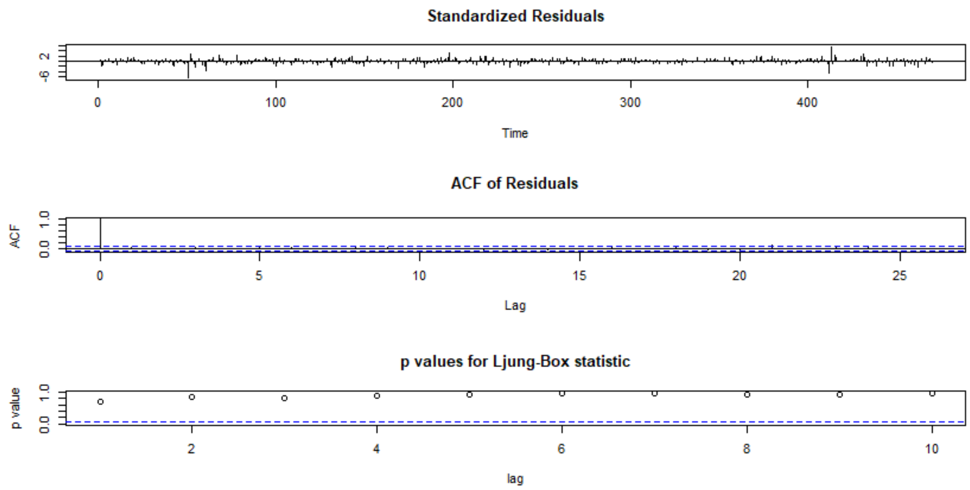


Figure 7: Standarized residuals, ACF of residuals and the p-values for Ljung-Box statistic.

```
        Ljung-Box test

data:  Residuals from ARIMA(2,1,2) with non-zero mean
Q* = 3.7926, df = 5, p-value = 0.5796

Model df: 5.   Total lags used: 10
```

4 Forecast

We decided to forecast the period left out of the training data, i.e. the last 52 weeks of the series. This is performed by fitting the ARIMA(2,1,2) model to the training data and using the forecast function from the forecast package. The resulting figure 8 shows this forecast with the 80% and 95% confidence intervals. Comparing the forecast result in blue to the original time series in red in the figure, we see that the actual series is included in the prediction.

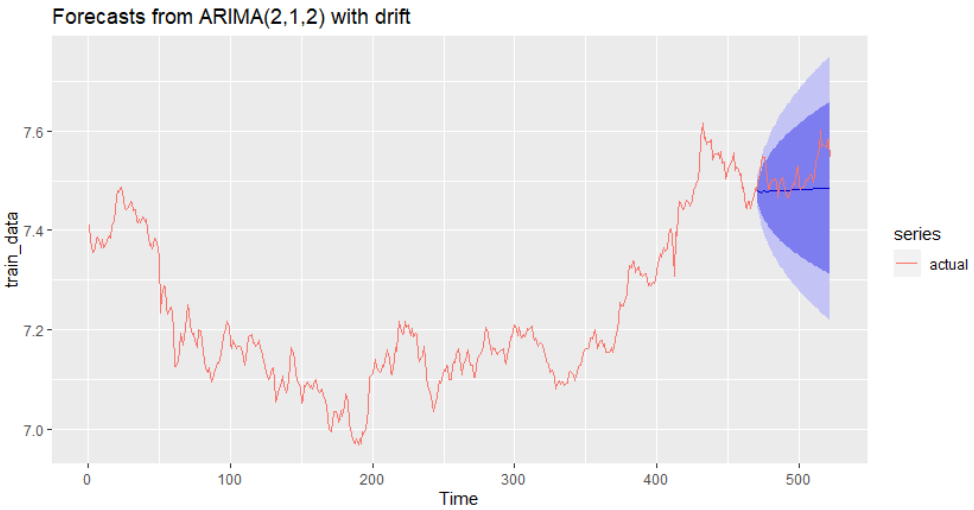


Figure 8: Forecast of the last 52 weeks.

5 Conclusion

The aim of this report was to use an ARIMA model to forecast the future price of gold. We limited the data set to weekly mean values of the series over a ten year time period and utilized the first nine years as a training set to predict the last year (52 weeks). We analyzed the ACF and PACF after the series was transformed to eliminate trend and inconsistent variance. The best model fit for the data set was the ARIMA(2,1,2) model, which was determined by the estimating the AIC and analyzing the residuals.

The predicted price based on the available training data was inside the confidence intervals. The overall forecast yield decent accuracy, with the mean absolute error 0.0125. Unfortunately, with the limitation of ARIMA modeling, the forecast results in mean reversion. Therefore the forecast does not capture random events or deviations that occur. In this case, the result is still within the confidence intervals but forecasting in other time periods yield inconsistent results. We could have compared the results to models such as GARCH or included Fourier transformation. Thus this forecast and model should not be taken as a guaranteed result.

References

[1] "National association of securities dealers automated quotations - gold (gc:cmx)." https://www.nasdaq.com/market-activity/commodities/gc:cmx/historical.

A R Code

```
library(forecast)
library(tidyverse)
library(lmtest)
library(tseries)
library(lubridate)
library(dplyr)

# Read data and arrange
df<-read.csv("GOLD.csv")
df<-df %>% arrange(desc(row_number()))

# Change to weekly data (max of each week)
df$week<-floor_date(as.Date(df$Date),"week")
df<-df %>%
  group_by(week) %>%
  summarize(Close = mean(Close.Last))


# Create and plot time series
gold_data<-ts(df$Close,frequency = 52.2,
              start=decimal_date(ymd("2012-04-30")))

# Log of time series
gold_log <- log(gold_data)
autoplot(gold_log)

# ACF and PACF log data
Acf(gold_log, lag.max = 522)
Pacf(gold_log, lag.max = 522)

# ADF and KPSS test log data
adf.test(gold_log, alternative = c("stationary", "explosive"),
         k = 0)
kpss.test(gold_log)

# Difference of lag 1
gold_diff <- diff(gold_log, lag = 1)
autoplot(gold_diff)

# ADF and KPSS test diff data
adf.test(gold_diff, alternative = c("stationary", "explosive"),
         k = 0)
kpss.test(gold_diff)
# ACF and PACF diff data
Acf(gold_diff,lag.max = 522)
Pacf(gold_diff,lag.max = 522)
```

```r
# Train data set (90%)
train_data <- gold_log[1:470]

# Creating ARIMA model
arima <- Arima(train_data, order = c(2, 1, 2),include.drift = TRUE)
summary(arima)
checkresiduals(arima)
tsdiag(arima)

# Histogram with normal curve
normal <- seq(min(arima$residuals), max(arima$residuals), length = 600)
fun <- dnorm(normal, mean = mean(arima$residuals),
             sd = sd(arima$residuals))
hist(arima$residuals, prob = TRUE, col = "white",
     ylim = c(0, max(fun)+10),
     main = "Histogram with normal curve",
     breaks = 50)
lines(normal, fun, col = 2, lwd = 2)

# Normal Q-Q plot
qqnorm(arima$residuals)
qqline(arima$residuals, col = "red",
       distribution = qnorm)

# Final forecast
forecast_ori <- forecast(arima, h = 52)
actual <- ts(gold_log)
forecast_ori %>% autoplot() + autolayer(actual)
```