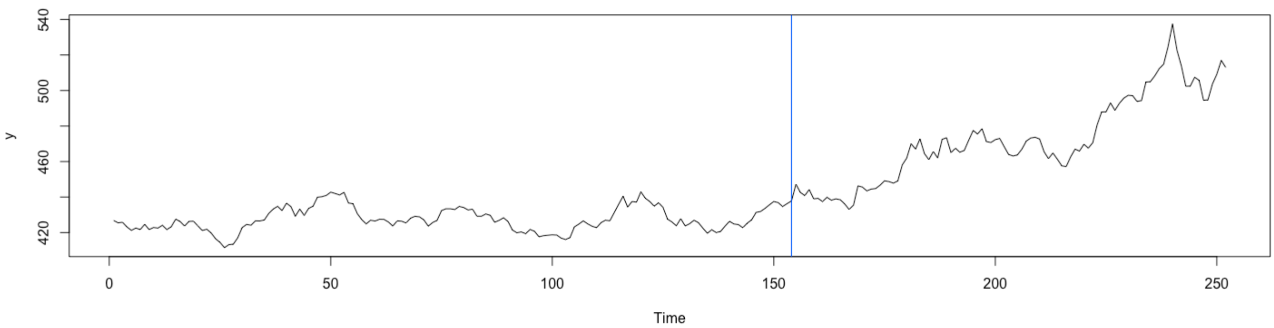
Project work for the teaching Time Series Analysis – Luca Pennella

First Time Series Analysis

Exploratory Data Analysis

The analysis concerns the study of time series relating to the daily price of gold (in $ per troy ounce) for the 252 trading days of 2005, the source of the data is available at the link https://www.lbma.org.uk/.
The dataset presents the information with a daily frequency (from 1 to 252), the data range starts from a minimum value of 411.5 up to a maximum of 537.5 with an average value of 445.

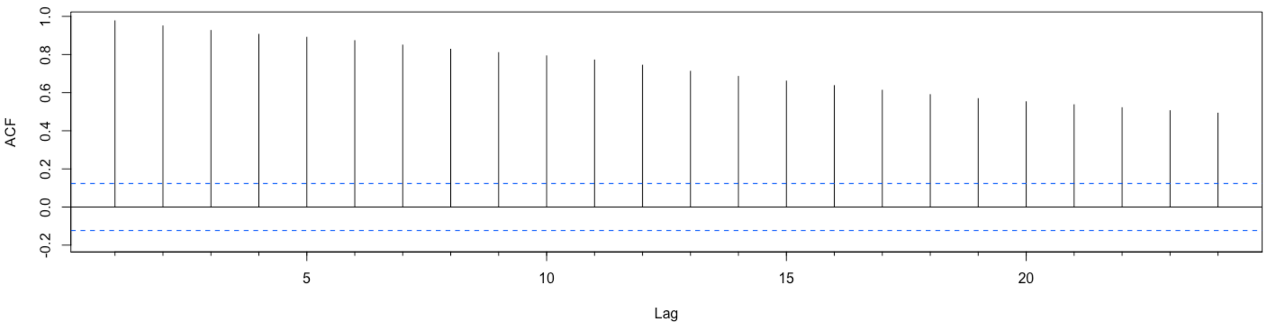


The series seems to show an increasing trend with a probable structural break around t = 154 which leads the values of the series to grow significantly. While the series before t = 154 appears to be quite stable with a stationary variance. Looking at the graph, no cycles or seasonal patterns are highlighted.

Find the Models

To establish the appropriate ARIMA model for a given time series, Box & Jenkins (BJ) proposed the following iterative procedure:
1. Identification
- Transform the time series to make it stationary through differencing, that can be preceded by another transformation useful to stabilize the variance;
- inspect ACF and PACF to deduce plausible values of pp, qq.
2. Estimation
- for each value of pp, qq obtain MLE of parameters based on the transformed time series;
- select the best model based on some fitting statistics.
3. Diagnostic
- check whether the residuals of the best model are compatible with a white noise process.
- If so, the procedure ends, otherwise it is repeated by applying a different transformation

The first phase is the control of the stationarity of the series through the study of the ACF and the application of some tests.



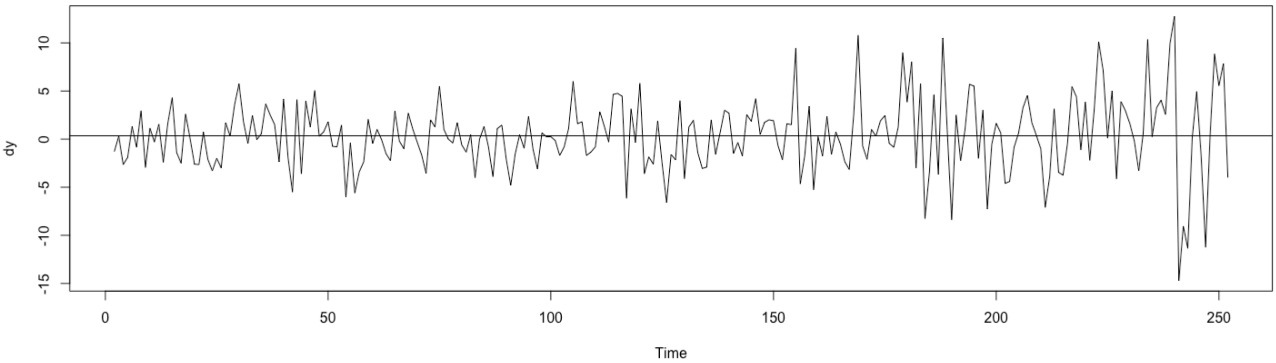
Observing the graph it is possible to see how the decrease of the ACF is linear, this trend highlights a non-stationarity of the model. For further confirmation we will apply two different tests, Augmented Dickey-Fuller Test and KPSS Test for Level Stationarity:

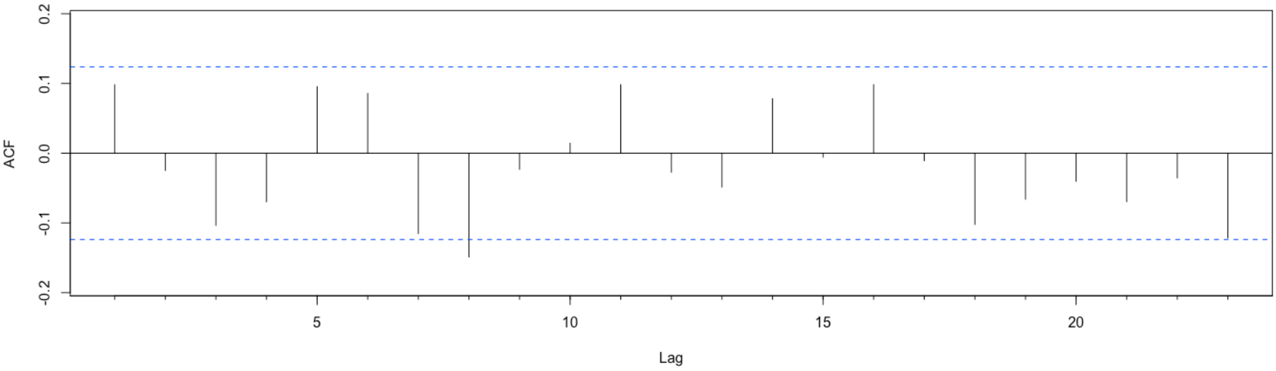
```
        Augmented Dickey-Fuller Test

data:  y
Dickey-Fuller = -1.7846, Lag order = 6, p-value = 0.6665
alternative hypothesis: stationary
```

```
        KPSS Test for Level Stationarity

data:  y
KPSS Level = 3.1043, Truncation lag parameter = 5, p-value = 0.01
```

Both tests confirm what we saw in the previous graph, the series does not appear to be stationary.

In cases where the series is not stationary, a solution can be to apply the first differences with the aim of eliminating the non-stationarity. Once the difference has been applied, we first re-observe the series and the ACF graph.





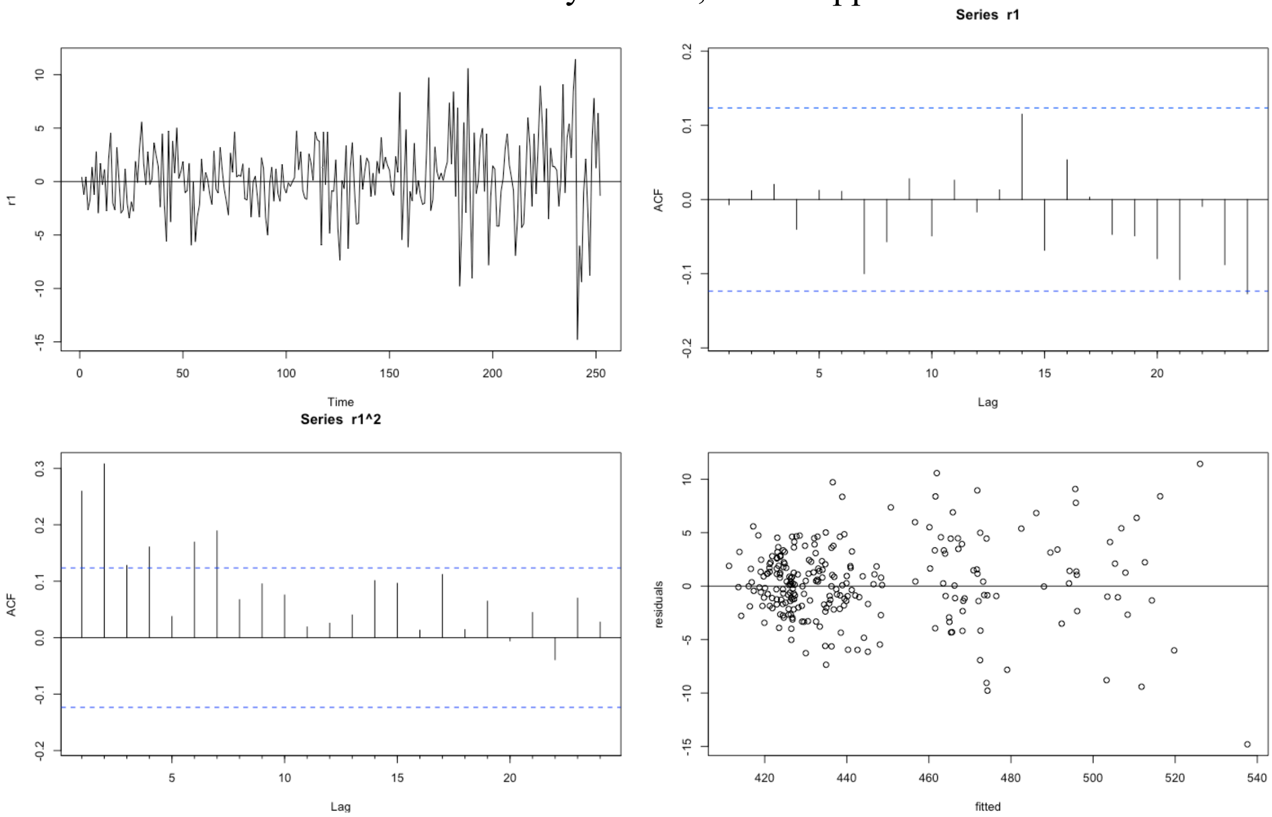
Observing the series it is possible to notice how the variance seems to grow starting from t = 154 while the ACF graph does not show any peak outside the confidence intervals.
The ACF study also seems to recommend the use of a random walk but in my opinion the result cannot be judged given the non-stationary nature.

```
        Augmented Dickey-Fuller Test

data:  dy
Dickey-Fuller = -6.397, Lag order = 6, p-value = 0.01
alternative hypothesis: stationary
```

```
        KPSS Test for Level Stationarity

data:  dy
KPSS Level = 0.23844, Truncation lag parameter = 5, p-value = 0.1
```

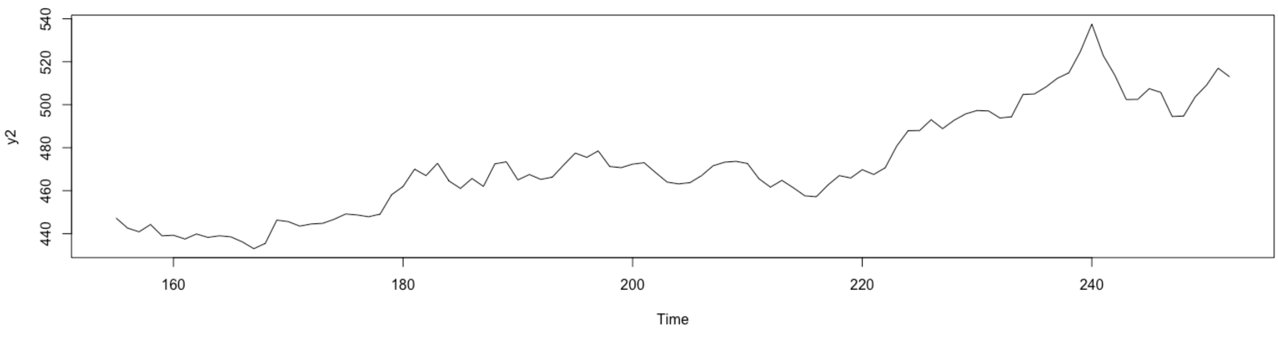
The ADF and KPSS tests reject the hypothesis of non-stationarity. From the observational study of the series and from the results of the tests we can affirm that the presence of the structural break at t = 154 makes the analysis complex. One solution is to divide the series into two parts, 'breaking' it at t = 154 to evaluate the series only after the break.
Further confirmation of this hypothesis is the evaluation of the residues once the

ARIMA model has been automatically chosen, which appear as follows:

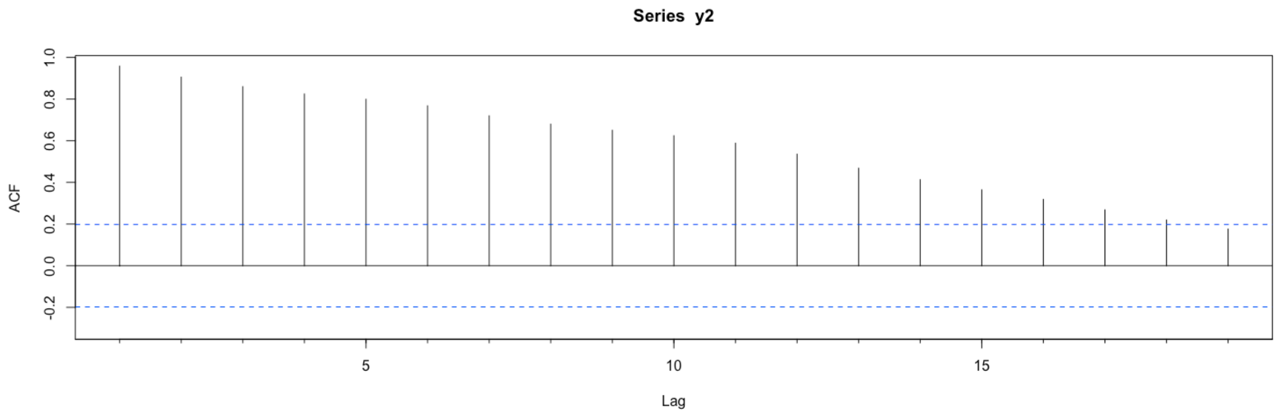


From a general evaluation of the residual plots, we can say that they do not distribute randomly, the situation does not change by applying logarithmic or square root transformations.

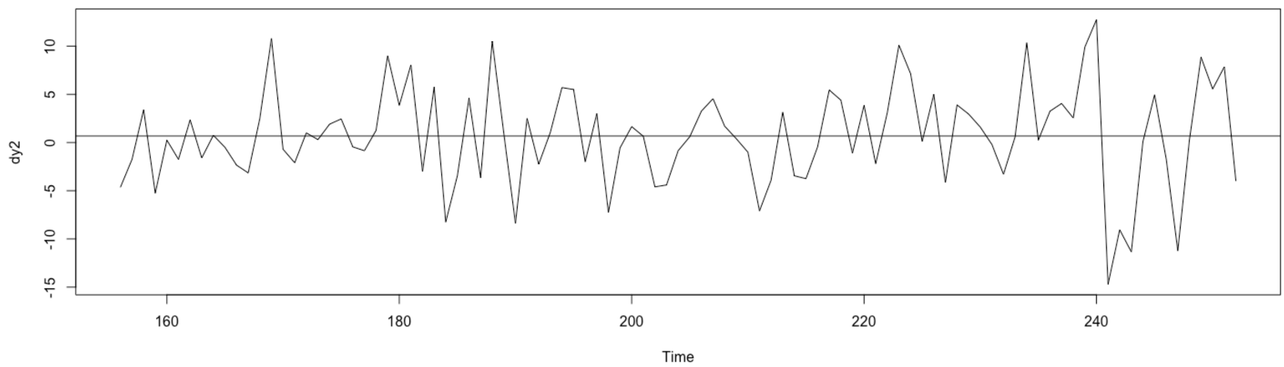
We then proceed by breaking the series at the point t = 154.

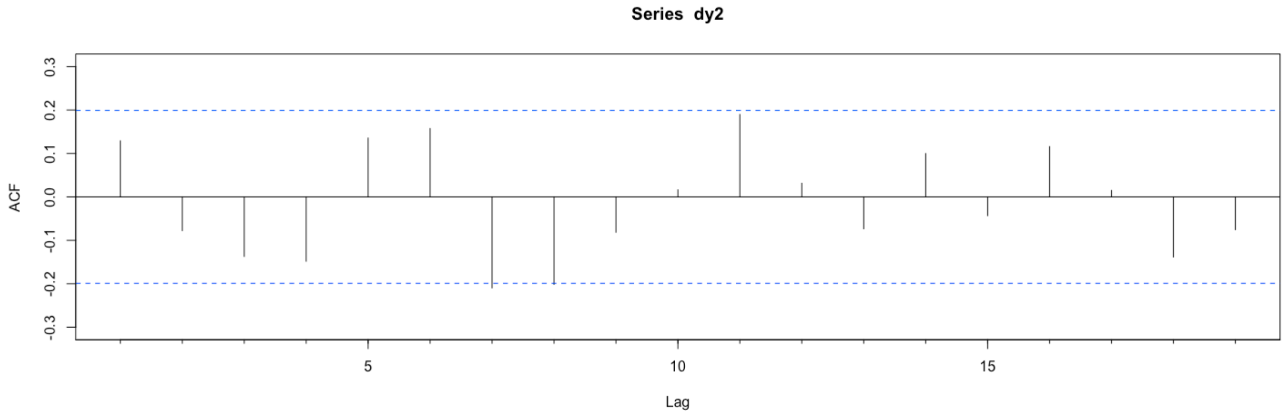


Evaluating the new series (t> 154) we can see how the variance seems to be stationary while the mean trend seems to have an increasing trend.

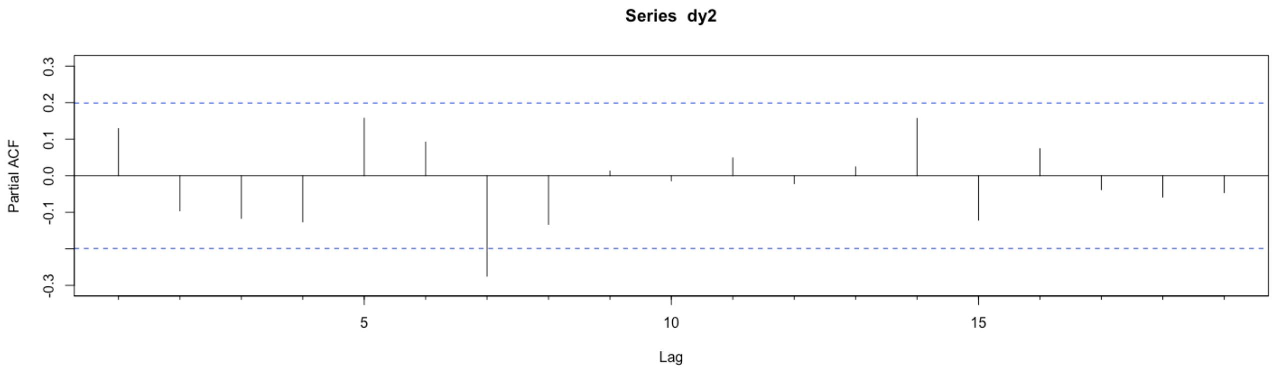


The ACF plus the ADF (0.53) and KPSS (0.01) tests confirm the non-stationarity, so we continue to the first differences with the aim of eliminating the trend observed previously.





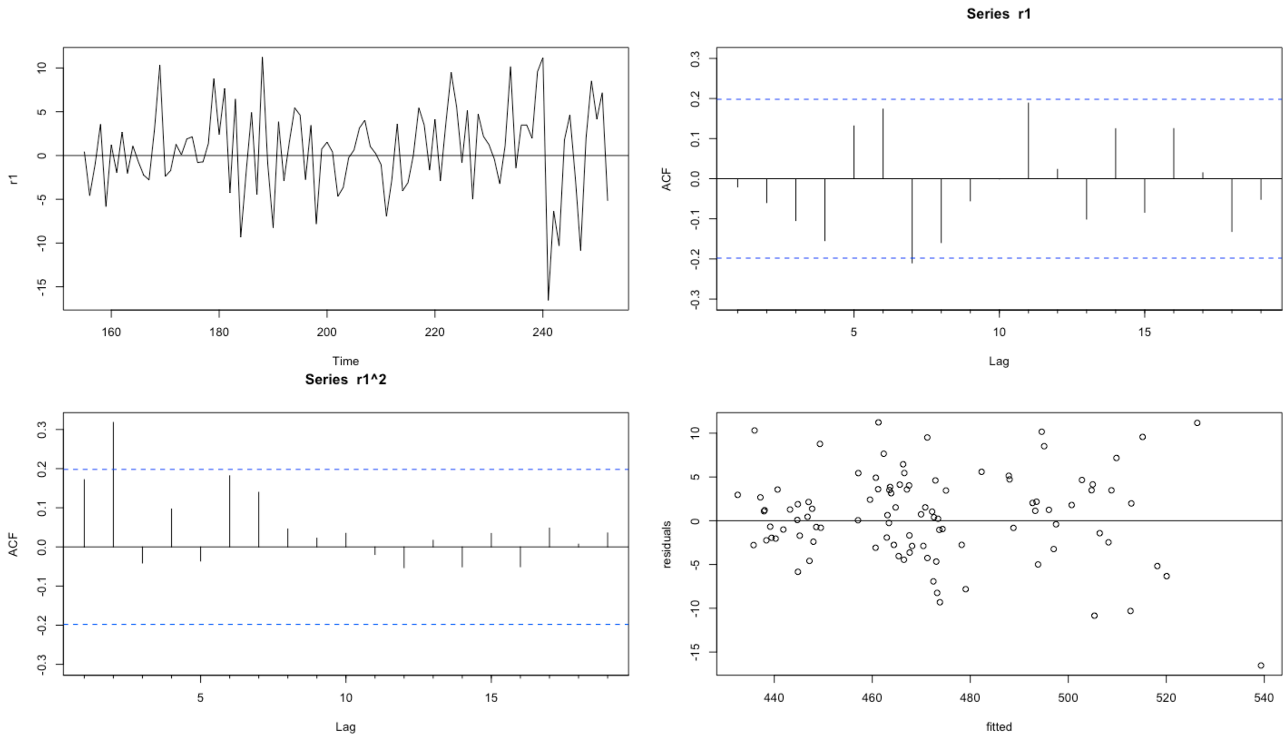
Observing the series, the variance seems to be more stable, however some doubts persist from t = 240.
The ACF graph presents two peaks slightly above the bands which lead to hypothesize an MA(7) or MA(8).
While the ADF (0.01) and KPSS (0.1) tests indicate that the series is stationary, at this point we proceed with the observation of the PACF chart for the evaluation of AR orders.



The PACF graph shows a peak at lag 7, therefore it seems to recommend an AR (7), at this point it is better to proceed with the automatic selection of the best ARIMA through the information criteria.

```
Series: y2
ARIMA(0,1,1)

Coefficients:
         ma1
      0.1644
s.e.  0.1035

sigma^2 = 25.76:  log likelihood = -294.71
AIC=593.42   AICc=593.55   BIC=598.57
```

The automatic selection recommends the use of an ARIMA (0,1,1).

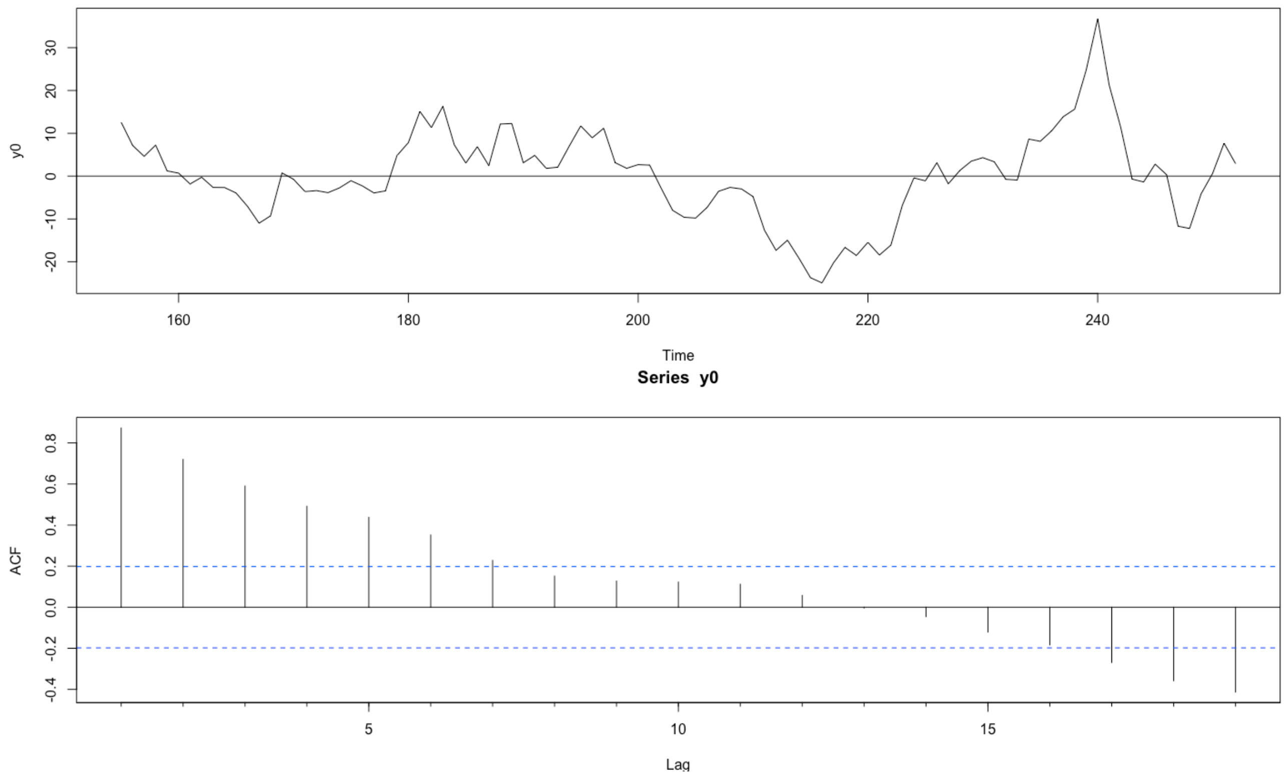


Analyzing the residuals of our model it is possible to see how the situation has improved a lot, however it is necessary to highlight the peak at lag 2 in the squared residuals and a possible increase in the variance in the residuals starting from t = 240. While the graph of fitted vs residuals seems to have a random distribution.

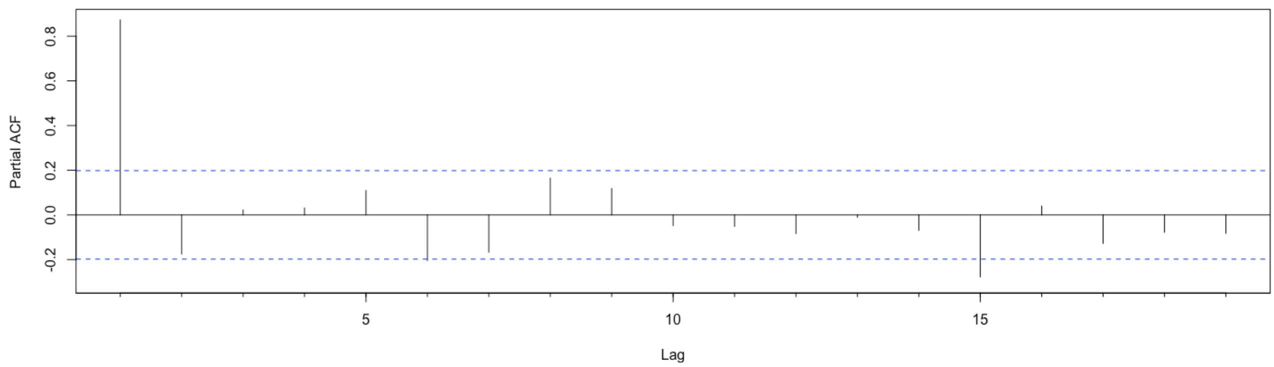
As an alternative to the ARIMA model, we try to model the series with an ARMA with a purely deterministic trend.
The first step to build an ARMA with a deterministic trend is the realization of ARIMA (0,0,0) which has time as a regressor, once this is done it is necessary to check the significance of the model:

```
Series: y2
Regression with ARIMA(0,0,0) errors

Coefficients:
      intercept    xreg
       433.8955  0.7775
s.e.     2.1011  0.0369
```

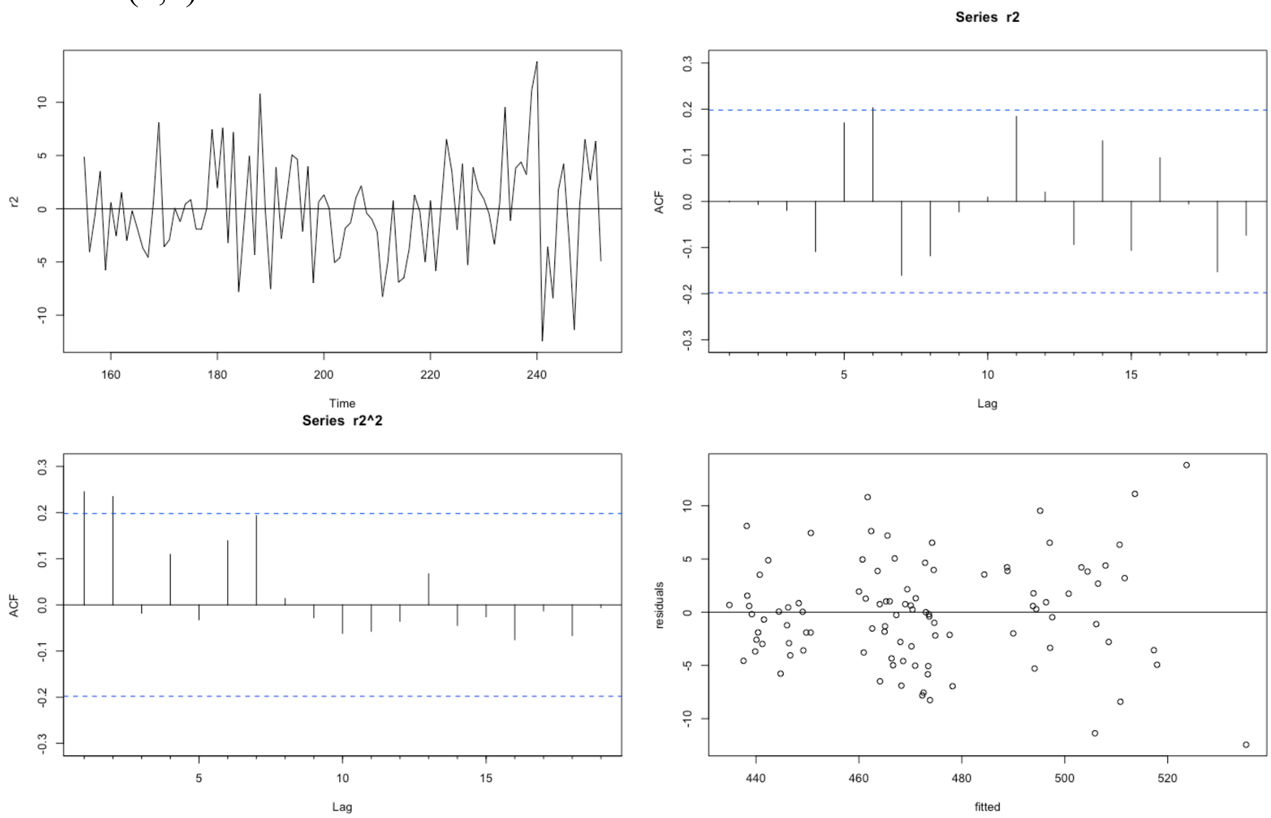
In this case, a linear trend seems plausible. Next it is necessary to check the stationarity of the residuals of the model.



Observing the residual plot, the distribution might seem stationary but what is worrying is the ACF which decays rapidly but at the same time shows a persistent autocorrelation at the highest lags. The result of the ADF test (0.53) also provides evidence on the presence of unit roots.
We proceed with the fit of the ARMA model with pure deterministic trend despite the doubts about stationarity and then make a comparison with the model already estimated.



Observing the ACF (seen previously) and the PACF we can hypothesize an ARMA (6,1), however, to be sure I select the model following the information criteria through automatic selection. The best model provided by automatic selection is an ARMA (2,0) with a deterministic trend.

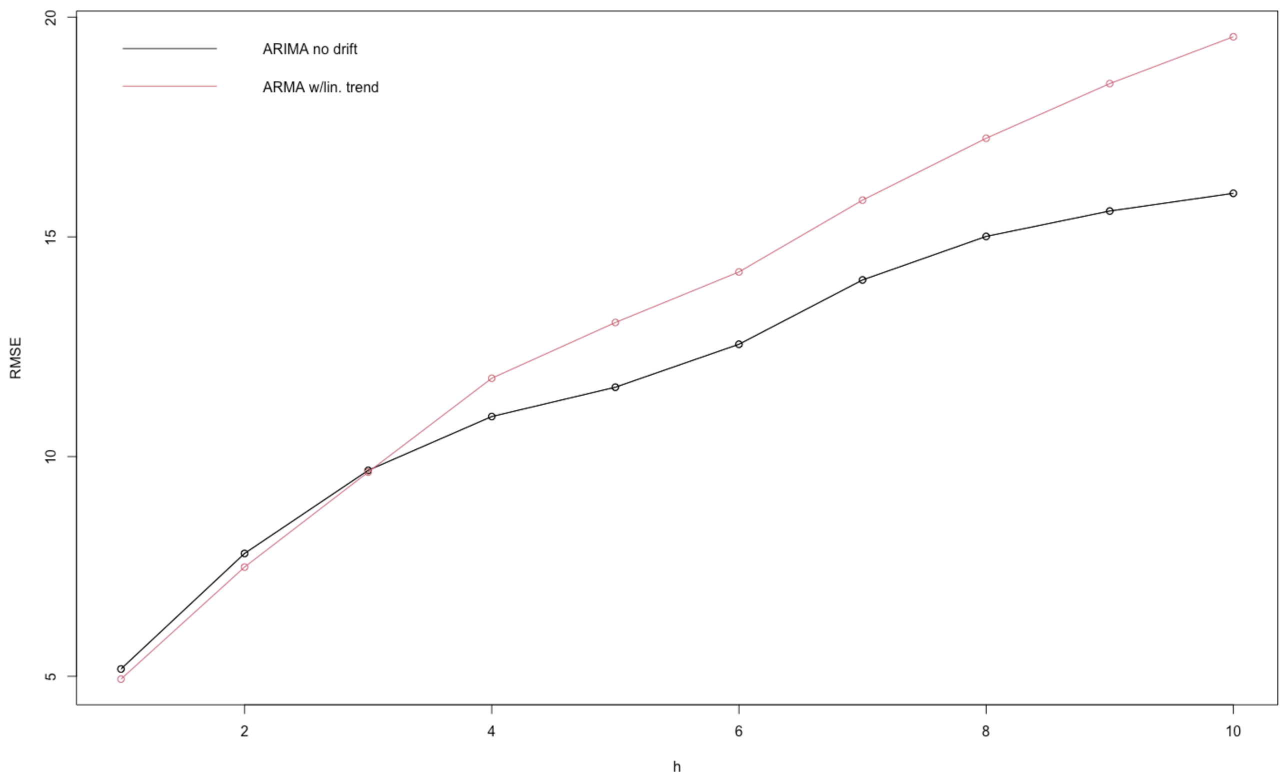


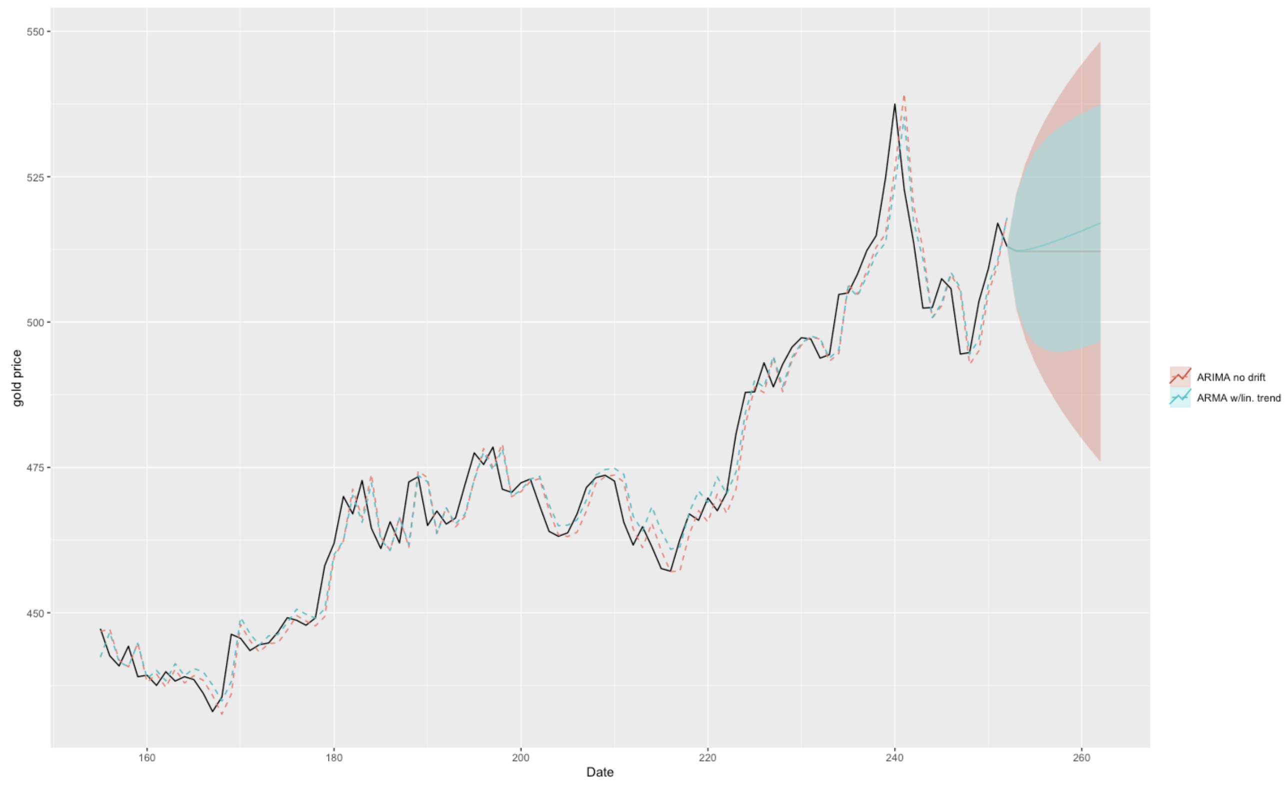
The different plots of the residuals are satisfactory except for the squared residues which show peaks at low lags, this may indicate autocorrelation problems.
In summary, the two models identified are:

- $\Delta Y_t = \psi_1 \varepsilon_{t-1} + \varepsilon_t$
- $Y_t = \omega + \eta_t + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \varepsilon_t$

Model Comparison and Forecast Section

Observing the graph below, the two models for low h have similar performances while as the h value increases, ARIMA returns better values in terms of RMSE. The result is plausible because as the forecast horizon increases, the deterministic trend ARMA has a RMSE that tends to diverge, this is because deterministic trend processes tend to perform better in short-term forecasts.





As for the forecasts we have two very different behaviors, as ARIMA tends to have a convergent forecast, while the ARMA with a deterministic trend as previously

mentioned tends to diverge. As for the variance we can see how the ARMA tends to have a smaller interval than the ARIMA, for the linear trend it is possible to notice a constant interval for the entire forecast horizon, unlike the ARIMA as the he time horizon has an interval that tends to diverge.
As for the choices of the best model, it depends on the time horizon of interest, for close periods (h <4) it is advisable to use the linear model as it is simpler but as h increases the trend with a stochastic process provides strictly better results .