M.Sc. in Data Science

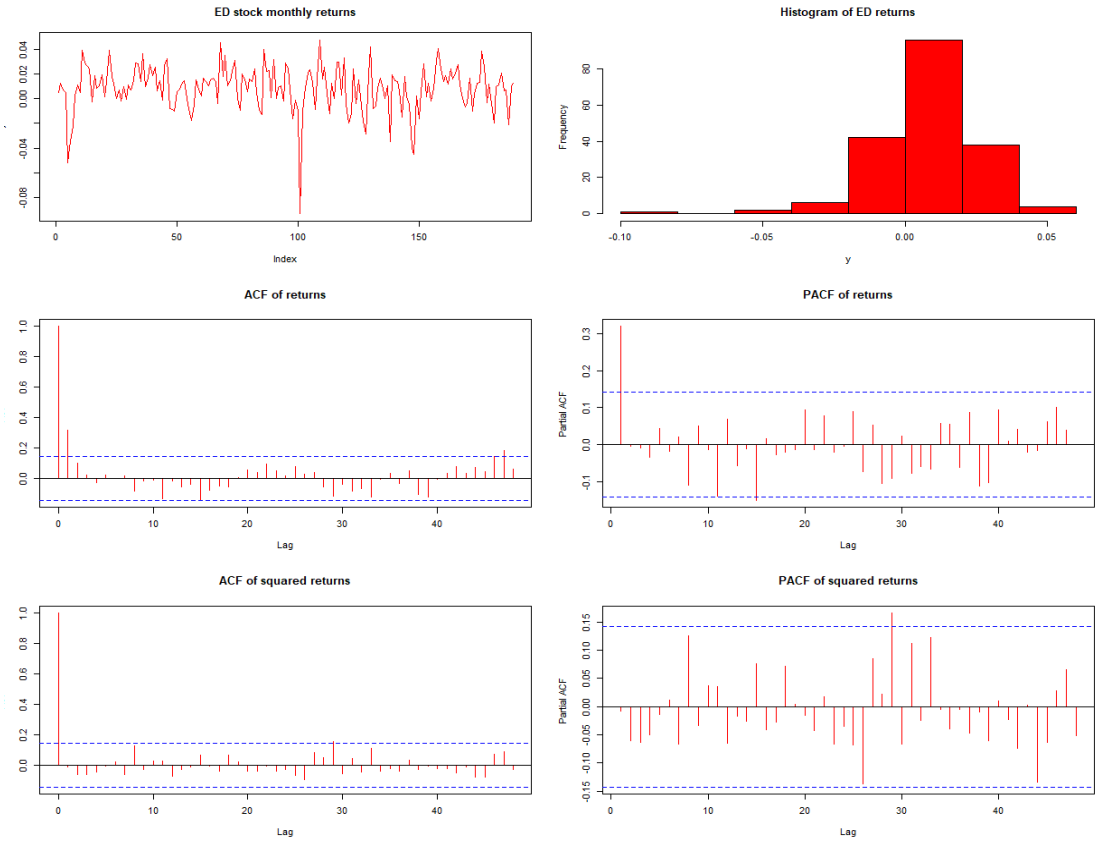
Athens University of Economics & Business

Time Series and Forecasting Methods

Assignment: November 2022

Task 1: Develop Time Series Models (AR,MA).

After the insertion of the taget data column y5 or ED, containing the monthly returns of the stock the data are analyzed as a timeseries. We observe that the timeseries is stationary with a frequency of 12. Then we create a visual representation through a histogram, the autocorrelation and partial autocorrelation functions of the returns and their squared returns.



It is observed that we lack normality and a present autocorrelation for lag 1.

For the ACF graph we use MA(1).

```
Call:
arima(x = y, order = c(0, 0, 1))

Coefficients:
         mal  intercept
      0.2976     0.0083
s.e.  0.0654     0.0016

sigma^2 estimated as 0.0002973:  log likelihood = 499.17,  aic = -992.35
```

MA(1) shows statistical significance, sigma squared is small ang log likelihood big.

I also tried different solutions such as:

MA(2,3)

```
Call:
arima(x = y, order = c(0, 0, 3))

Coefficients:
          ma1     ma2     ma3  intercept
      0.3230  0.1039  0.0476     0.0083
s.e.  0.0733  0.0794  0.0760     0.0018

sigma^2 estimated as 0.0002944:  log likelihood = 500.09,  aic = -990.18
```

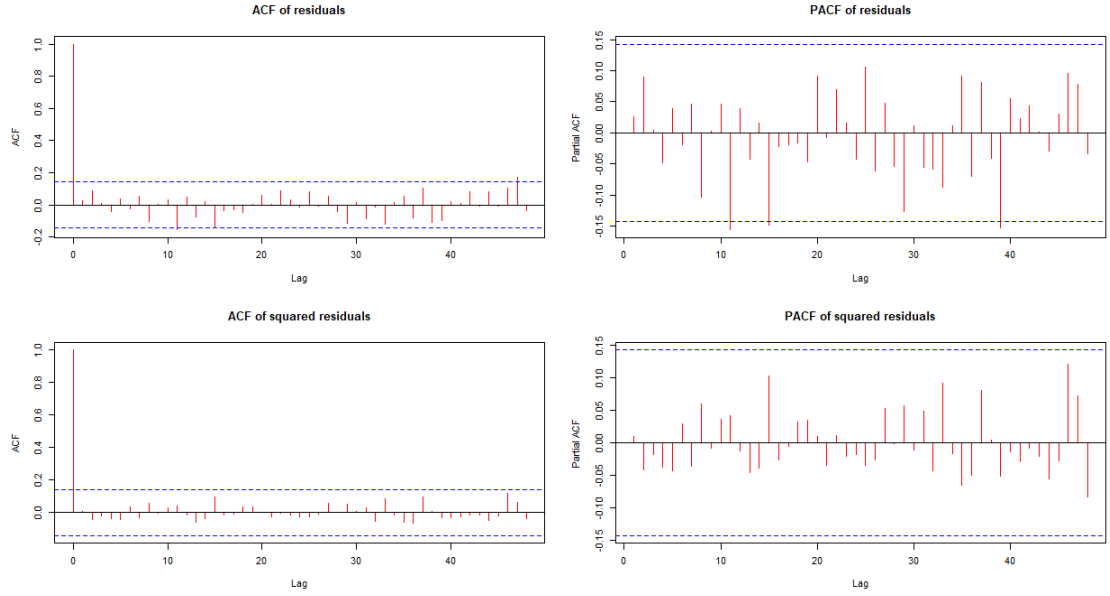
Moving average for lags 2 and 3 don't show importance.

AR(1)

```
Call:
arima(x = y, order = c(1, 0, 0))

Coefficients:
          ar1  intercept
      0.3179     0.0083
s.e.  0.0687     0.0018

sigma^2 estimated as 0.0002948:  log likelihood = 499.98,  aic = -993.97
```

The autoregressive model for lag 1 does not show statistical significance.

We conclude that the right model for the data is the MA(1).



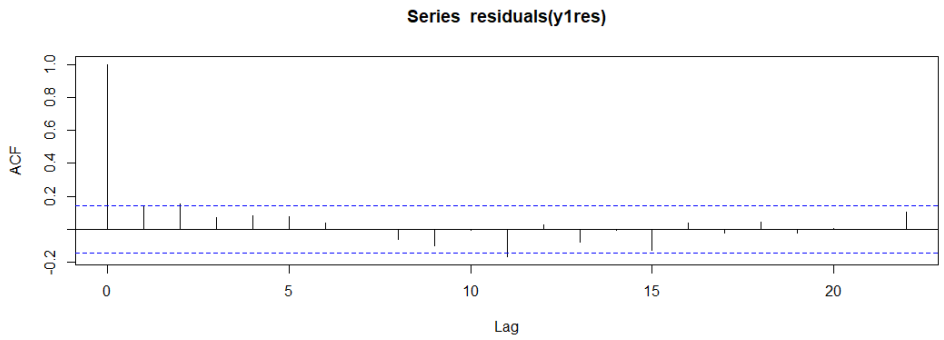
Indeed, the use of the model solves the autocorrelation problem. Beyond that we observe 3 values close and above the limit in the PACF graph of the residuals, but they are far away from zero and are not important enough to need to solve. For real data is tolerable.

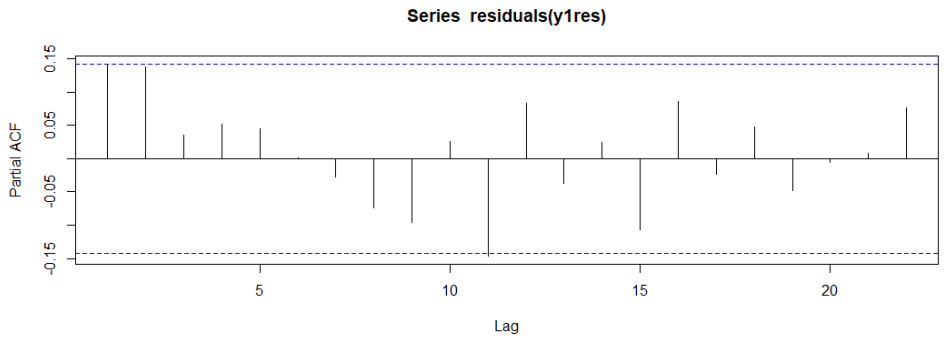
Task 2: Develop some multiple regression models.

For the first model I decided to incorporate every piece of information we get and I added all variables offered.

```
Call:
lm(formula = y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 +
    x10 + x11 + x12 + x13 + x14 + x15)

Residuals:
      Min        1Q    Median        3Q       Max
-0.029058 -0.005361 -0.000754  0.005376  0.032321

Coefficients:
             Estimate Std. Error t value         Pr(>|t|)
(Intercept)  0.003488   0.001801   1.937         0.054435 .
x1           0.198072   0.032100   6.170 0.00000000468564 ***
x2           0.052378   0.018992   2.758         0.006442 **
x3           0.027876   0.023962   1.163         0.246275
x4           0.071195   0.016001   4.449 0.00001537790861 ***
x5           0.203438   0.026806   7.589 0.00000000000191 ***
x6           0.084837   0.024992   3.395         0.000852 ***
x7           0.035548   0.014923   2.382         0.018298 *
x8           0.173664   0.074038   2.346         0.020130 *
x9          -0.042894   0.058719  -0.730         0.466076
x10          0.035276   0.025579   1.379         0.169642
x11         -1.317212   0.668417  -1.971         0.050360 .
x12          0.062307   0.070603   0.882         0.378733
x13          0.010268   0.012297   0.835         0.404842
x14          0.012393   0.026819   0.462         0.644592
x15          0.718525   0.481921   1.491         0.137793
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.008947 on 173 degrees of freedom
Multiple R-squared:  0.7768,    Adjusted R-squared:  0.7575
F-statistic: 40.15 on 15 and 173 DF,  p-value: < 0.00000000000000022
```

As expected not all variables carry useful information. Through our analysis we find those that are statistically significant. We also observe that residuals have few and small errors, R squared is large so we are in the right path.

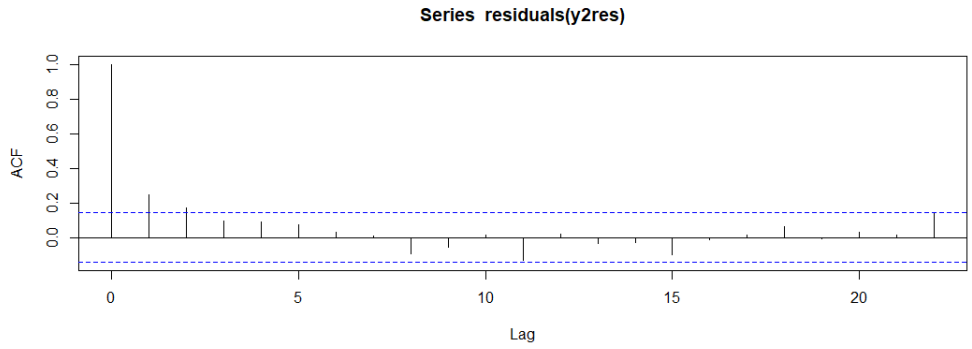


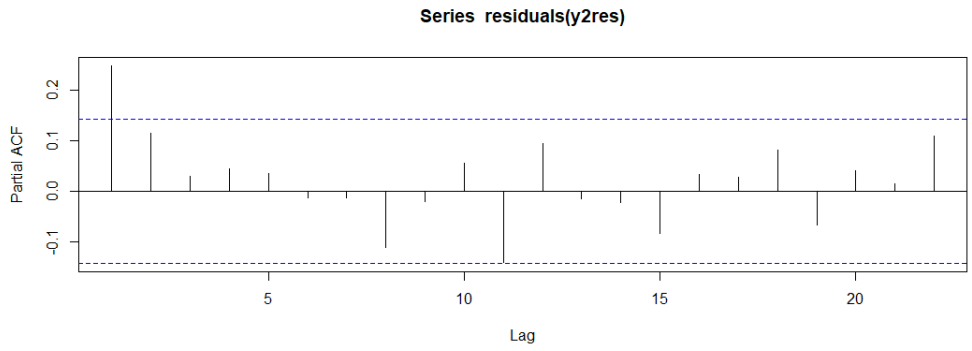


To improve upon the first model, we keep the variables showing statistical significance.

```
Call:
lm(formula = y ~ x1 + x2 + x4 + x5 + x6 + x8)

Residuals:
      Min        1Q    Median        3Q       Max
-0.030782 -0.005388 -0.000113  0.005617  0.035115

Coefficients:
             Estimate Std. Error t value            Pr(>|t|)
(Intercept) 0.0060139  0.0006956   8.645 0.00000000000000273 ***
x1          0.1997461  0.0233845   8.542 0.00000000000000518 ***
x2          0.0682601  0.0174444   3.913            0.000129 ***
x4          0.0754380  0.0148755   5.071 0.00000096803942010 ***
x5          0.1955356  0.0238669   8.193 0.00000000000004396 ***
x6          0.0518590  0.0193711   2.677            0.008104 **
x8          0.1437123  0.0545533   2.634            0.009156 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.00921 on 182 degrees of freedom
Multiple R-squared:  0.7512,    Adjusted R-squared:  0.743
F-statistic: 91.58 on 6 and 182 DF,  p-value: < 0.00000000000000022
```

I slight drop of R squared but an important gain from the great reduction from the number of variables and an increase of the F statistic.



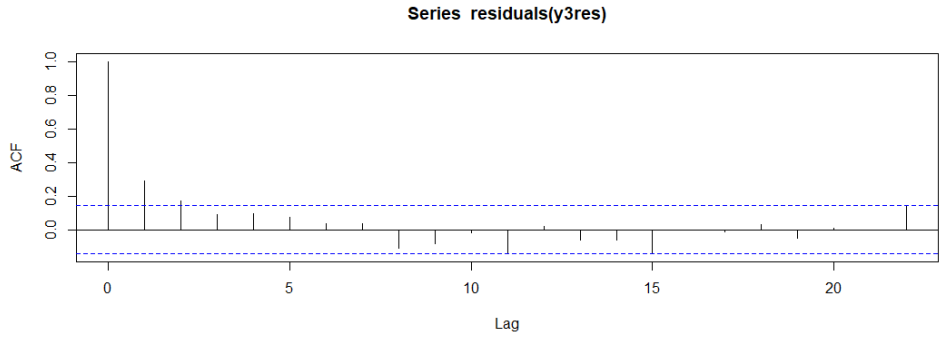


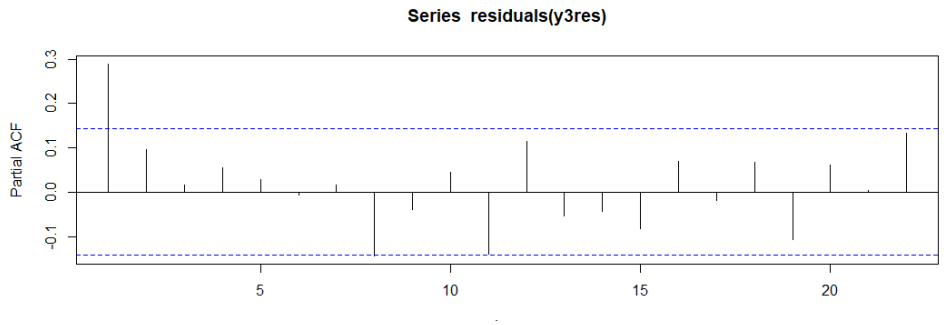
Unfortunately, we observe the appearance of autocorrelation for lag 1 ang 2.

To have an alternative solution I kept the 4 variables for a<0.005.

```
Call:
lm(formula = y ~ x1 + x2 + x4 + x5)

Residuals:
      Min        1Q    Median        3Q       Max
-0.030197 -0.005412 -0.000003  0.005617  0.036092

Coefficients:
            Estimate Std. Error t value             Pr(>|t|)
(Intercept) 0.006579   0.000699   9.412 < 0.0000000000000002 ***
x1          0.196116   0.023349   8.400   0.0000000000000118 ***
x2          0.077646   0.017364   4.472   0.0000135574232701 ***
x4          0.074827   0.015211   4.919   0.0000019189465749 ***
x5          0.165794   0.022724   7.296   0.0000000000085798 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.009516 on 184 degrees of freedom
Multiple R-squared:  0.7315,    Adjusted R-squared:  0.7256
F-statistic: 125.3 on 4 and 184 DF,  p-value: < 0.00000000000000022
```

The F statistic significantly improved while the R squared barely moved. Meaning that the these 4 variables carry the main load of information.





The problem of autocorrelation for lag1 and 2 remains.

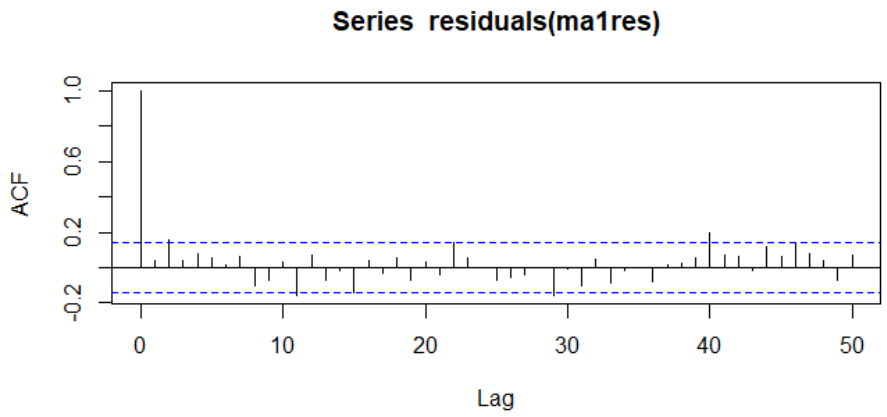
Both the 6 variable and 4 variable models are useful, note that before any use it would be necessary to solve the issue of autocorrelation.

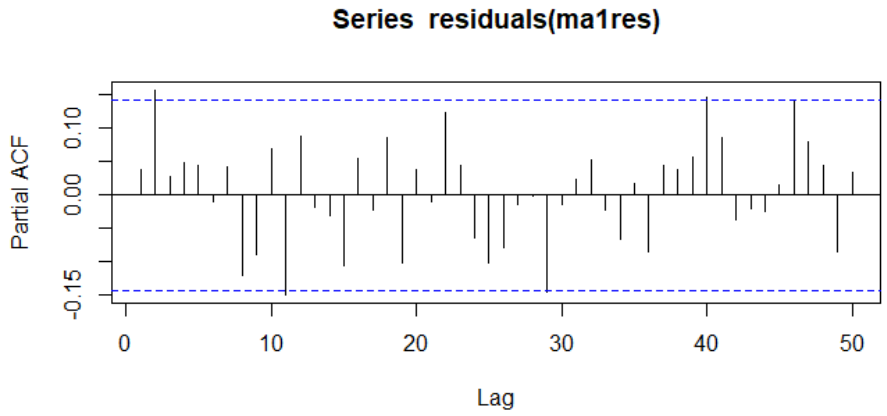
Task 3: Develop a fitting multiple regression model.

(i) In case of autocorrelation with the residues use AM, AR, ARMA to solve the issue.

(ii) In case of a heteroscedasticity problem with the residues from the regression solve the issue using constrained heteroscedasticity models

Starting with the regression from task 2. Our first task is to solve the issue of the residue autocorrelation for lag 1. The simplest way is to use MA(1).

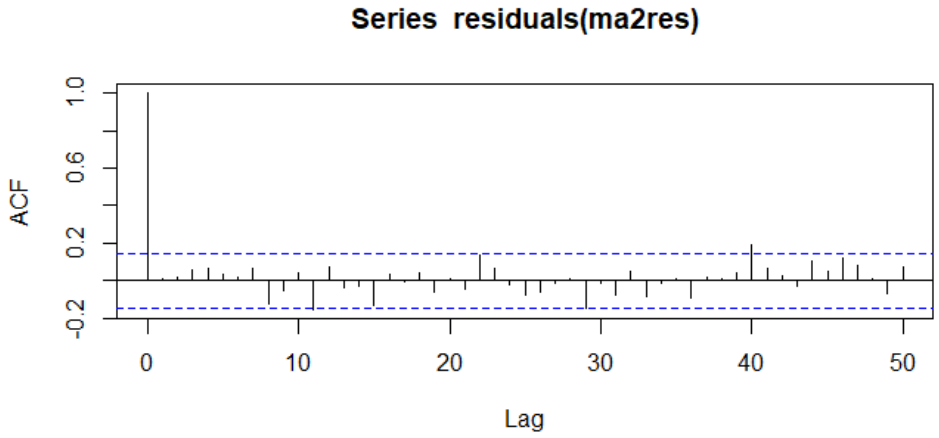


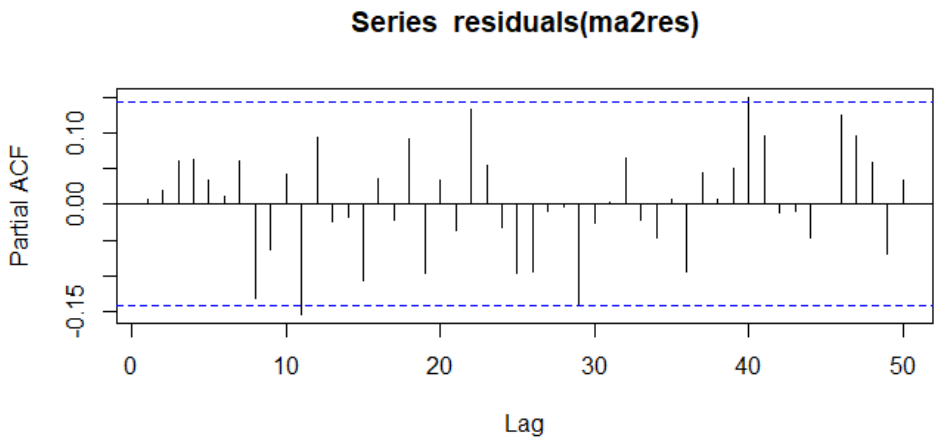


The use of MA(1) worked for lag 1 but for lag 2 the problem remains and there is a slight problem of partial autocorrelation for lag 2. Since MA(1) was useful we retain its use and improve upon our approach by using MA() for lag 2 as well.

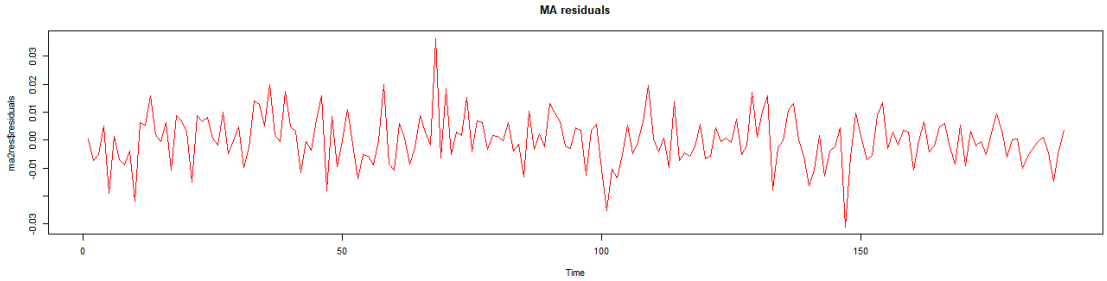
```
Call:
arima(x = residuals(y3res), order = c(0, 0, 2), include.mean = FALSE)

Coefficients:
         ma1     ma2
      0.2587  0.1388
s.e.  0.0721  0.0675
```

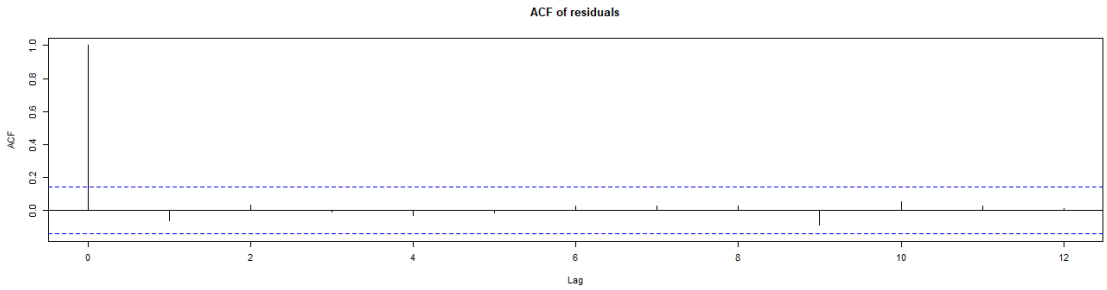
Both MA(1), MA(2) are statistically significant.

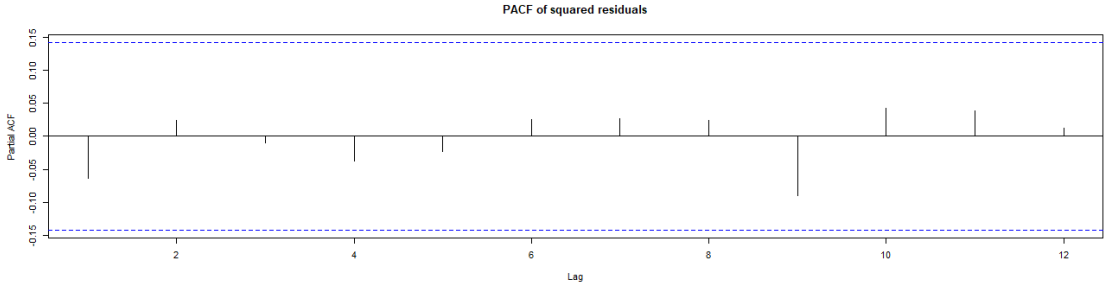




MA(2) done wonders as it solved both the autocorrelation and the partial autocorrelation problem for MA(2). PACF shows high partial autocorrelation for lags 11, 22, 29,… close to or above the limit but those are away from 0 and are not important enough so we need to do something about them. For real data it is good enough.



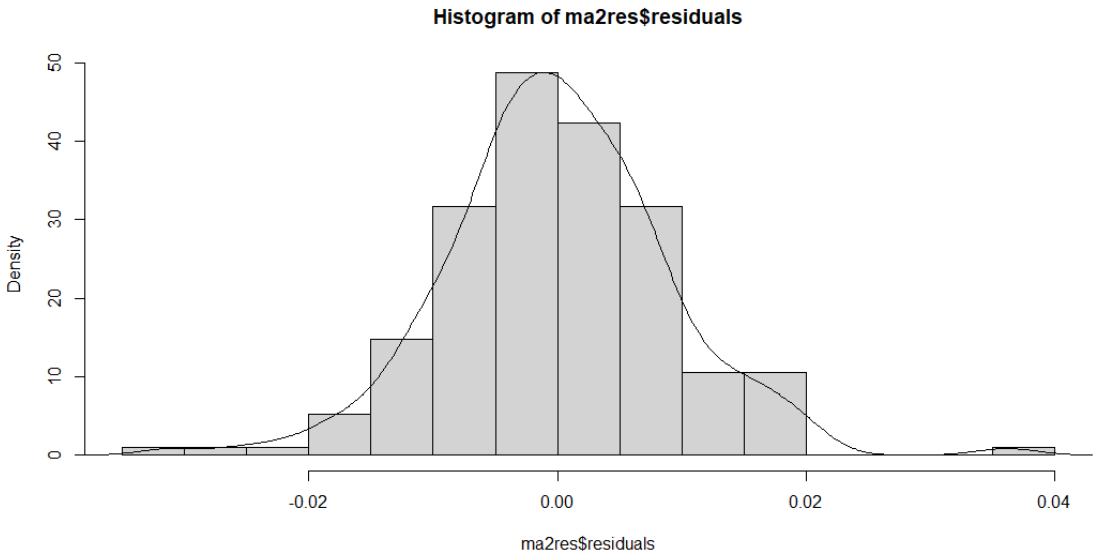


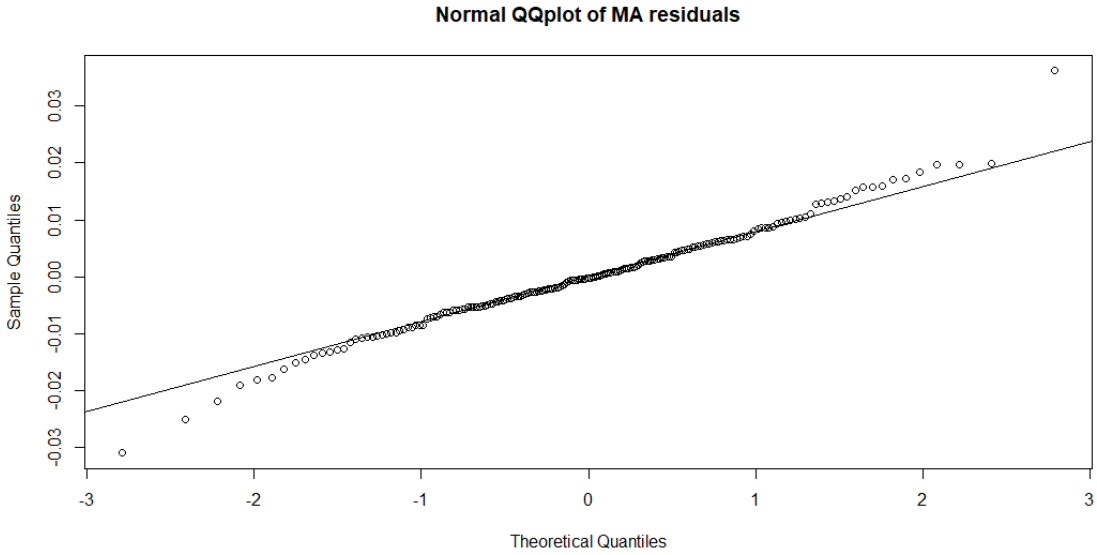


The squared residuals are within bounds meaning that there are no heteroscedasticity problems. To be safe I used the Ljung-Box test.

```
> Box.test(ma2res$residuals,lag=12,type="Ljung")

        Box-Ljung test

data:  ma2res$residuals
X-squared = 12.664, df = 12, p-value = 0.3939

> Box.test(ma2res$residuals^2,lag=12,type="Ljung")

        Box-Ljung test

data:  ma2res$residuals^2
X-squared = 3.9209, df = 12, p-value = 0.9848
```

The results show a lack of heteroscedasticity problems. Since we were lucky, I wanted to observe if the data are normal. I started with a histogram, for the density and Quantile-Quantile plot. Even as the histogram looks promising the Shapiro-Wilk test reveals that there is a slight violation of normality.






```
> shapiro.test(ma2res$residuals)

        Shapiro-Wilk normality test

data:  ma2res$residuals
W = 0.98505, p-value = 0.04201
```