Time Series & Forecasting Methods

Assignment 2020-2021

Galanos Vasileios - p3351902

November 29, 2020

In this assignment we perform statistical analysis on the data that we imported from the file data-assignment.txt into a dataframe in **R** . The file contains monthly returns of nine alternative investment funds. Below we can see the correspondence between the column headers and the names of the funds for the time period $4/1990 - 12/2005$.

Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8	Y9
HFRI	EH	M	RVA	ED	CA	DS	EMN	MA

The file also contains the independent variables that we will use:

x1	RUS-Rf
x2	RUS(-1)-Rf(-1) lagged Russel index
x3	MXUS-Rf
x4	MEM-Rf
x5	SMB
x6	HML
x7	MOM
x8	SBGC-Rf
x9	SBWG-Rf
x10	LHY-Rf
x11	DEFSPR
x12	FRBI-Rf
x13	GSCI-Rf
x14	VIX
x15	Rf

Exercise 1

In this exercise we will try to build time series models on some of the dependent variables ($Y1 \ldots Y9$). We will first test the hypothesis that the series are stationary using **unit root testing** along with other tests. We will then use the Box-Jenkins [1] methodology in order to find the AR(I)MA [2] model that best describes our stationary time series.

Y8 / EMN

First of all, we plot the Y8 variable in order to get a first glimpse of our data.

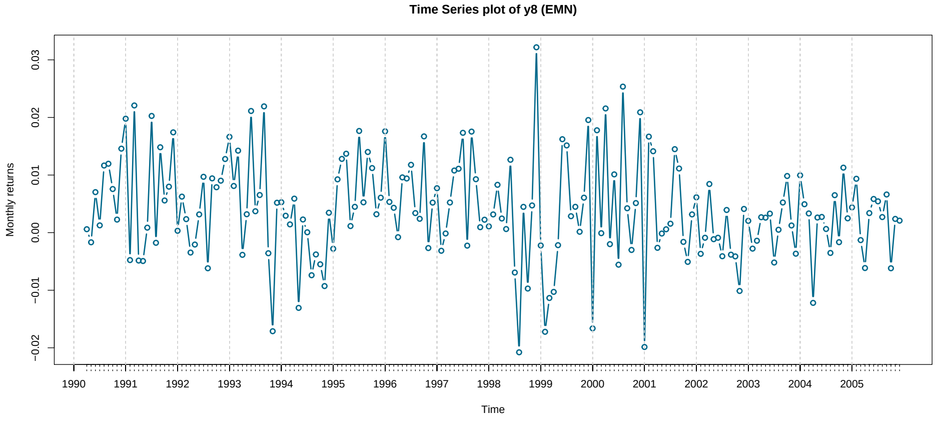


Figure 1: Y8 plot

We can see that the mean of the data is around zero and constant over time, the series appears to have no trend. Moreover there isn't any major problem with heteroscedasticity (variance roughly remains constant over most of the time). From a visual point of view, our series appears stationary.

[1] https://en.wikipedia.org/wiki/Box%E2%80%93Jenkins_method

[2] https://en.wikipedia.org/wiki/Autoregressive%E2%80%93moving-average_model#ARMA_model

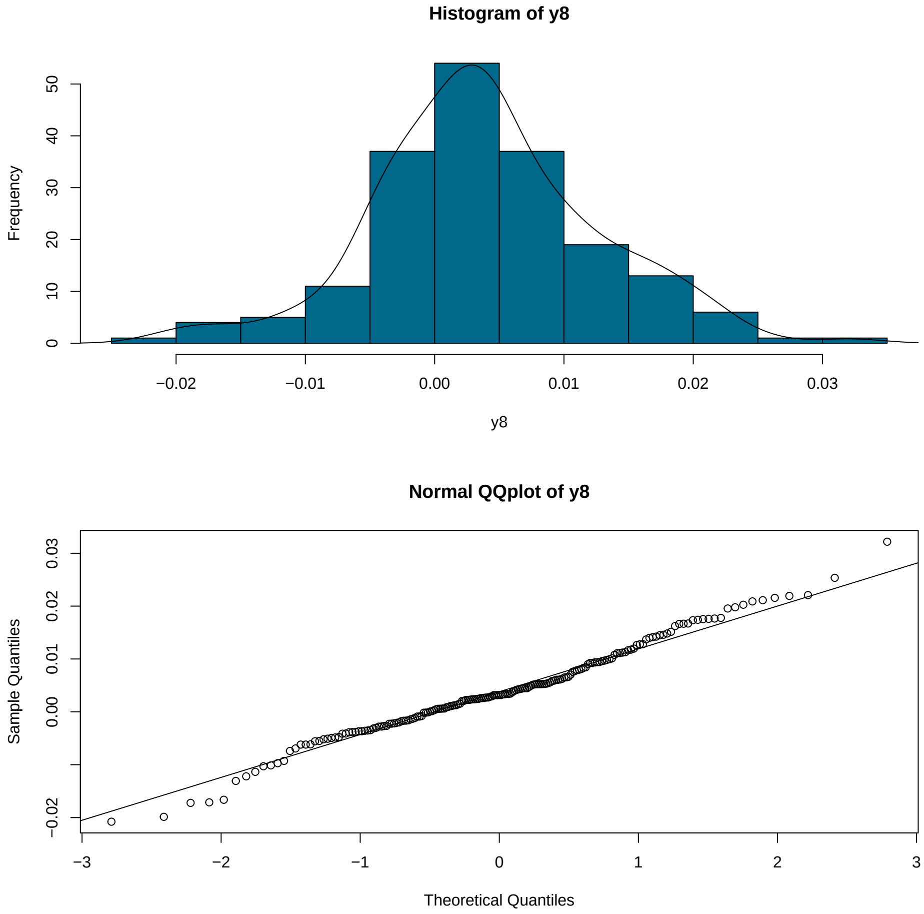


Figure 2: Y8 Histogram & QQ plot

With the previous graphs (histogram and QQ-plot), we confirm that our data appear to roughly follow a zero-mean normal distribution. We can also test for normality using a Shapiro-Wilk normality test which we fail to reject. Thus, our time series is normally distributed.

```
> shapiro.test(y8)
        Shapiro-Wilk normality test
data:  y8
W = 0.98633, p-value = 0.06395
```

We also perform a One Sample t-test [3] in order to test the hypothesis that the mean of our time series is zero.

```
 > t.test(y8)
        One Sample t-test
data:  y8
t = 6.1809, df = 188, p-value = 3.87e-09
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 0.002655555 0.005145186
sample estimates:
 mean of x
0.00390037
```

As we see, we reject the Null hypothesis, thus our data mean is not zero. But we can see from the confidence intervals that it is very close to zero.

Next, using the **R** function Box.test() we can compute the Box-Pierce and the Ljung-Box [4] test statistic for examining the null hypothesis that the autocorrelations of a given time series are zero. We use 36 lags (3 years) and we reject the null hypothesis (for $\alpha = 0.05$ significance level) on both tests, that means the "overall" autocorrelations of the time series are different from zero for at least the last 3 years (36 months).

```
> Box.test(y8,36,type="Box-Pierce")
        Box-Pierce test
data:  y8
X-squared = 56.905, df = 36, p-value = 0.01469

> Box.test(y8,36,type="Ljung-Box")
        Box-Ljung test
data:  y8
X-squared = 62.754, df = 36, p-value = 0.003765
```

Now it is time to test the stationarity of our time series. First we try to fit the series, as best as we can, using an Autoregressive (AR) model. With the default option the model finds the best order (complexity) of the fitted model by minimizing the "Akaike" Information Criterion (AIC).

[3] https://en.wikipedia.org/wiki/Student%27s_t-test

[4] https://en.wikipedia.org/wiki/Ljung%E2%80%93Box_test#Box-Pierce_test

```
Call:
ar(x = y8)
Coefficients:
      1        2        3        4        5        6
 0.0128   0.0177   0.0671   0.0369  -0.0926   0.2807
Order selected 6  sigma^2 estimated as  7.017e-05
> m$order
[1] 6
```

The model that we found that best fits the series can be described by the following equation:

$$\begin{aligned}y_t &= \phi_1 y_{t-1} + \phi_2 y_{t-2} + \phi_3 y_{t-3} + \phi_4 y_{t-4} + \phi_5 y_{t-5} + \phi_6 y_{t-6} + \epsilon_t \\ &= 0.0128 y_{t-1} + 0.0177 y_{t-2} + 0.0671 y_{t-3} + 0.0369 y_{t-4} - 0.0926 y_{t-5} + 0.2807 y_{t-6} + \epsilon_t\end{aligned}$$

We use this *order* in order to perform an augmented Dickey-Fuller test of unit root based on Random Walk with Drift. We chose this model because we proved that the mean of our series is statistically different from zero although very close to it. The testing procedure is applied to the model

$$\Delta y_t = \mu + \beta y_{t-1} + \sum_{j=1}^{5} \lambda_j \Delta y_{t-j}$$

because we found that the best fitted model had an order of 6, thus we need 5 lagged differences for the augmented version of the test.

```
> m1=ur.df(y8,type="drift",lags=m$order-1)
> m1

###############################################################
# Augmented Dickey-Fuller Test Unit Root / Cointegration Test
    #
###############################################################

The value of the test statistic is: -4.0125 8.0654
```

We can see that the test statistic has a value (-4) that is significantly lower than the critical value (-2), thus we reject the null hypothesis that the model has a unit root ($\beta = 0$). In conclusion, we have statistical evidence that our series is stationary, thus we can proceede with the Box-Jenkins methodology.

1. Identification step

We will examine the autocorrelation and partial autocorrelation plots in order to get a first impression of which lags are more heavily correlated.

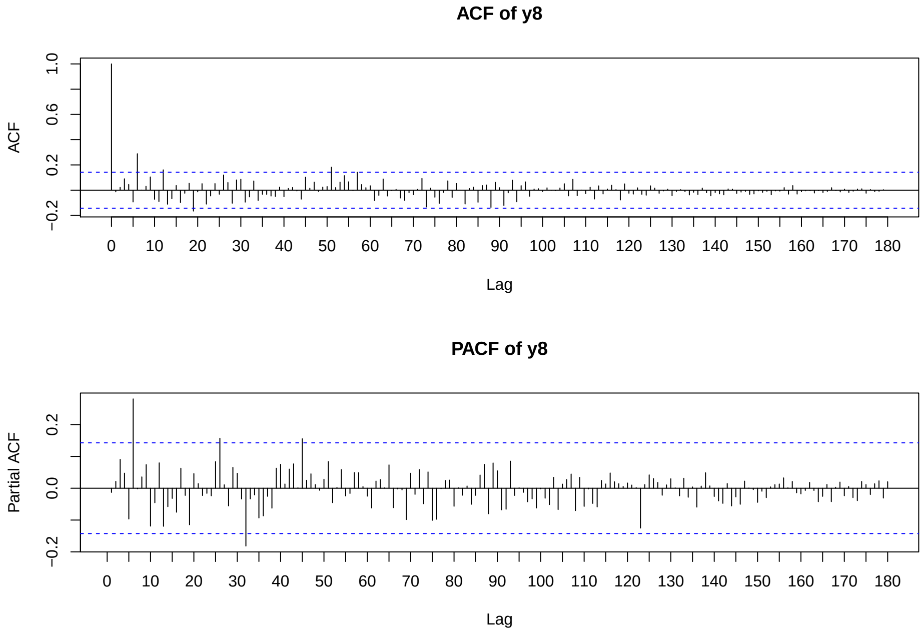


Figure 3: Y8 ACF & PACF plot

- From the **acf** plot we see that we have statistically significant peaks at lags 6,12,19,51 (they are outside of the bounds).
- From the **pacf** plot we see that we have statistically significant peaks at lags 6,26,32,45 .

In both graphs, only lag 6 is significantly larger than the bound. All the others are very close to it.

2. Estimation step

From the acf and pacf graphs, we have hints to use both ϵ_{t-6} and y_{t-6}. Ofcourse we will try to add more terms like ϵ_{t-12},y_{t-26}, e.t.c if needed.

- **MA(1)**
  We will first estimate a Moving Average MA(1) model as a point of reference.

```
Call:
arima(x = y8, order = c(0, 0, 1))
Coefficients:
          ma1  intercept
      -0.0123     0.0039
s.e.   0.0708     0.0006
sigma^2 estimated as 7.485e-05:  log likelihood =
    629.57,  aic = -1253.15
```

  We calculate a rough estimate of how statistically significant the estimated parameter is:

$$\frac{\hat{\theta_1}}{s.e(\hat{\theta_1})} = \frac{-0.0123}{0.0708} \approx 0.17 \in (-2, 2)$$

  As we expected, it is not statistically significant. The MA(1) model does not seem to fit our time series well.

- **MA(6)**
  We have a strong hint to use lag 6 both as a restricted MA(6) with the first five parameters fixed to zero.

```
Call:
arima(x=y8, order=c(0,0,6), fixed = c(0,0,0,0,0,NA,NA))
Coefficients:
      ma1  ma2  ma3  ma4  ma5     ma6  intercept
        0    0    0    0    0  0.2348     0.0039
s.e.    0    0    0    0    0  0.0627     0.0007
sigma^2 estimated as 6.972e-05:  log likelihood =
    636.11,  aic = -1266.22
```

  The log-likelihood has been increased from the previous model. σ^2 is small and the aic has been reduced. We very strong evidence that this model fitted our data better.

Also we see that the estimated parameter is statistically significant

$$\frac{\hat{\theta_6}}{s.e(\hat{\theta_6})} = \frac{0.2348}{0.0627} \approx 3.74 > 2$$

The estimated model can be written in the form:

$$y_t = 0.0039 + 0.2348\epsilon_{t-6} + \epsilon_t$$

- **AR(6)**
  We will estimate a restricted Autoregressive AR(6) model.

```
Call:
arima(x=y8, order=c(6,0,0), fixed = c(0,0,0,0,0,NA,NA))
Coefficients:
      ar1  ar2  ar3  ar4  ar5     ar6  intercept
        0    0    0    0    0  0.2854     0.0039
s.e.    0    0    0    0    0  0.0690     0.0008
sigma^2 estimated as 6.851e-05:  log likelihood =
    637.68,  aic = -1269.36
```

  The log-likelihood is slightly larger, the aic is slightly lower and the estimated parameter is statistically significant [5]. Thus, this is our best model so far. The estimated model can be written in the form:

$$y_t = 0.0039(1 - 0) + 0.2854y_{t-6} + \epsilon_t$$

- **ARMA(6,6)**
  We will estimate a restricted Autoregressive Moving Average ARMA(6,6) model.

[5] $\frac{\hat{\phi_6}}{s.e(\hat{\phi_6})} = \frac{0.2854}{0.0690} \approx 3.74 > 2$

```
Call:
arima(x = y8, order = c(6, 0, 6), fixed = c(0, 0, 0, 0, 0,
    NA, 0, 0, 0, 0, 0,
    NA, NA))

Coefficients:
  ar1 ar2 ar3 ar4 ar5   ar6 ma1 ma2 ma3 ma4 ma5   ma6 intercept
  0 0 0 0 0 0.5439 0 0 0 0 0 -0.2835  0.0039
s.e.  0 0 0 0 0 0.2064 0 0 0 0 0 0.2367  0.0009

sigma^2 estimated as 6.799e-05:  log likelihood = 638.36,
    aic = -1268.71
```

Although we used a more complex model, we got a lower aic score than the previous (AR(6)) model and the same σ^2, log-likelihood.

The ar6 term is almost statistically significant [6], but the ma6 term is not [7]. The estimated model can be written in the form:

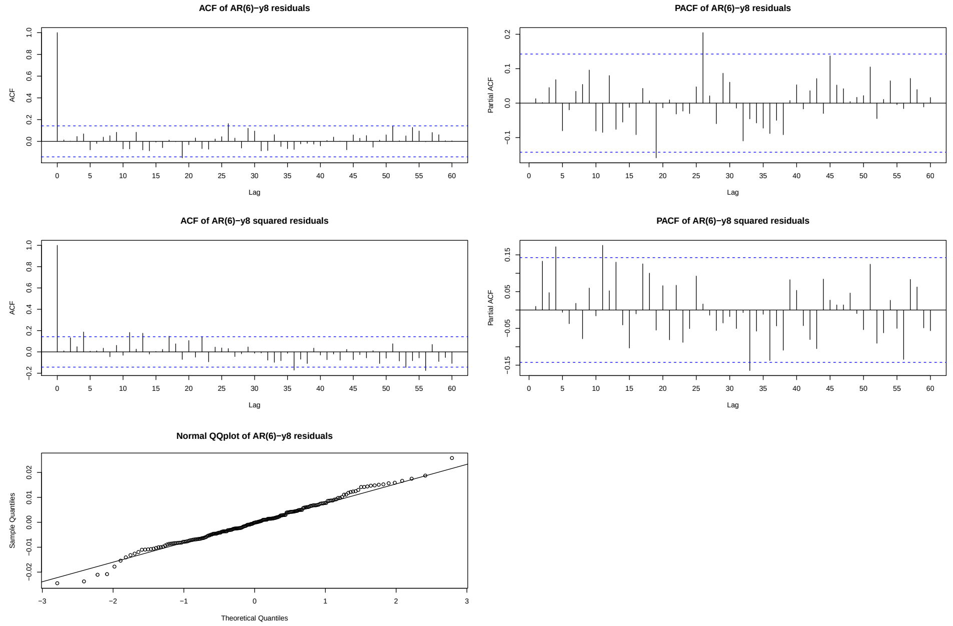
$$y_t = 0.0039(1-0) + 0.5439y_{t-6} - 0.2835\epsilon_{t-6} + \epsilon_t$$

3. Diagnostic plots

First we use the AR(6) model that we found was the best candidate. After fitting the model we test the residuals using again the acf and pacf plots. We are looking for important autocorrelations that we did not account for.

[6] $\frac{\hat{\phi_6}}{s.e(\hat{\phi_6})} = \frac{0.5439}{0.2064} \approx 2.63 > 2$

[7] $\frac{\hat{\theta_6}}{s.e(\hat{\theta_6})} = \frac{-0.2835}{0.2367} \approx -1.19 \in (-2, 2)$


ACF of AR(6)−y8 residuals

PACF of AR(6)−y8 residuals

ACF of AR(6)−y8 squared residuals

PACF of AR(6)−y8 squared residuals

Normal QQplot of AR(6)−y8 residuals

Figure 4: Y8 Residuals ACF & PACF plot for AR(6)

From the previous graph we have strong indication to add an ar(26) term. We return to the estimation step and add it.

- **AR(26)** (y_6, y_{26})

```
Call:
arima(x = y8, order = c(26, 0, 0), fixed = fx)
...
sigma^2 estimated as 6.716e-05:  log likelihood =
    639.32,  aic = -1270.64
```

The estimated model can be written in the form:

$$y_t = 0.0039(1-0) + 0.2856y_{t-6} + 0.1285y_{t-26} + \epsilon_t$$

Again, we return to the diagnostic plots to check if the autocorrelations have been taken care of.

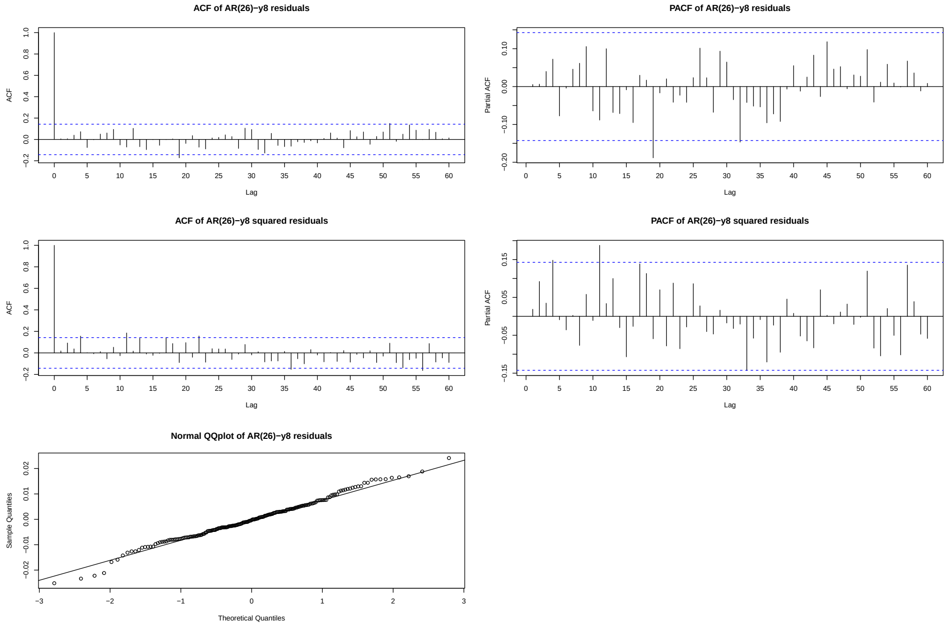


Figure 5: Y8 Residuals ACF & PACF plot for AR(26)

Lastly, we will try to add on last term, ar19 in order to reduce the correlation on pacf plot on lag 19.

- **AR(26)** (y_6, y_{19}, y_{26})

```
Call:
arima(x = y8, order = c(26, 0, 0), fixed = fx)
...
sigma^2 estimated as 6.566e-05:  log likelihood =
    641.28,  aic = -1272.55
```

$$y_t = 0.0039(1-0) + 0.2727y_{t-6} - 0.139y_{t-19} + 0.1271y_{t-26} + \epsilon_t$$

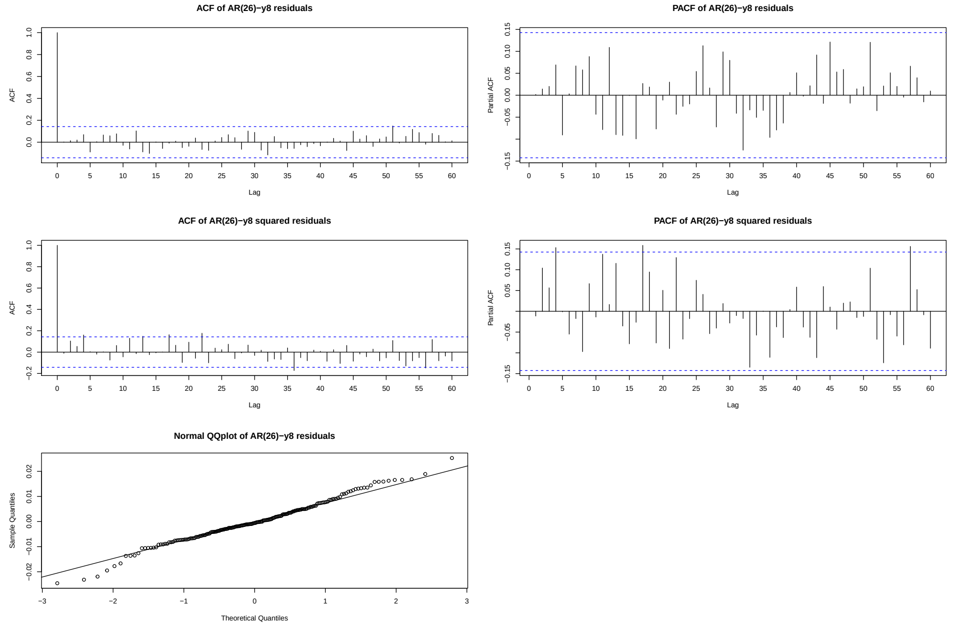


Figure 6: Y8 Residuals ACF & PACF plot for AR(26)

In conclusion, we are satisfied with the above graph, we have taken account for all the important autocorrelations. Thus, we proceede to the next step of the Box-Jenkins methodology with this AR(26) model.

Nevertheless, we tried to insert other terms even moving average terms, but the overhead in complexity was not worth the slight improvement in the results. Below we can see a table with all the different models and their corresponding measurements that we experimented with.

Model	**terms**	**AIC**	**Log-Likelihood**	σ^2
MA(1)	ϵ_{t-1}	-1253	629	7.48e-05
MA(6)	ϵ_{t-6}	-1266	636	6.97e-05
AR(6)	y_{t-6}	-1269	637	6.85e-05
ARMA(6,6)	y_{t-6}, ϵ_{t-6}	-1268	638	6.79e-05
AR(26)	y_{t-6}, y_{t-26}	-1270	639	6.71e-05
AR(26)	$y_{t-6}, y_{t-19}, y_{t-26}$	-1272	641	6.56e-05
AR(32)	$y_{t-6}, y_{t-19}, y_{t-22}, y_{t-26}, y_{t-32}$	-1275	644	6.56e-05

Overall, we still choose the restricted AR(6) with y_6, y_{19}, y_{26} as our best model.

4. Predictions

We compute predictions based on an estimated AR model. We predict for 8 months ahead.

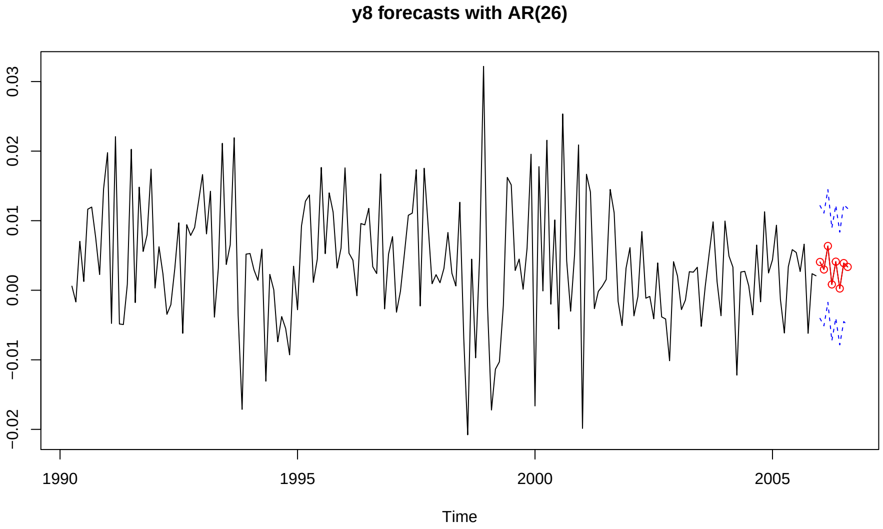


Figure 7: Y8 Predictions for AR(26)

As we can see, the confidence intervals are pretty wide, meaning we do not have very good predictions. But that is expected, since we are dealing with real-world data in the stock market, which are basically unpredictable.

Y5 / ED

We will follow exactly the same methodology as before. First of all, we plot the Y5 variable.

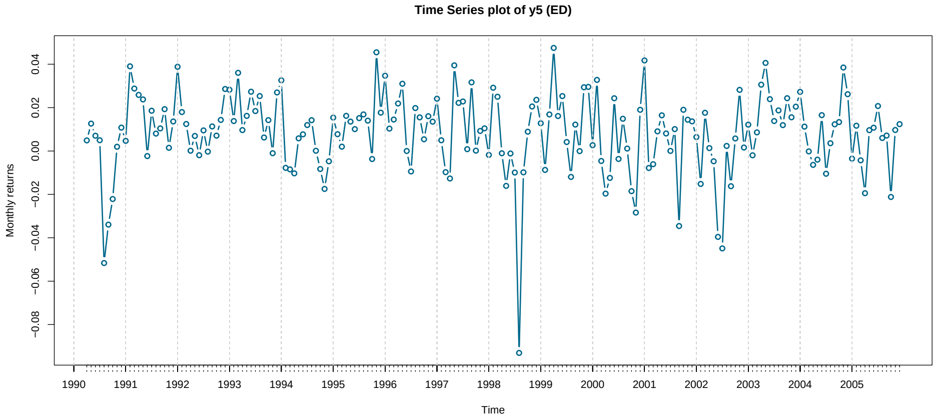


Figure 8: Y5 plot

We can see that the mean of the data is not around zero and but it is constant over time, the series appears to have no trend. Moreover there isn't any major problem with heteroscedasticity (variance roughly remains constant over most of the time). From a visual point of view, our series appears stationary.

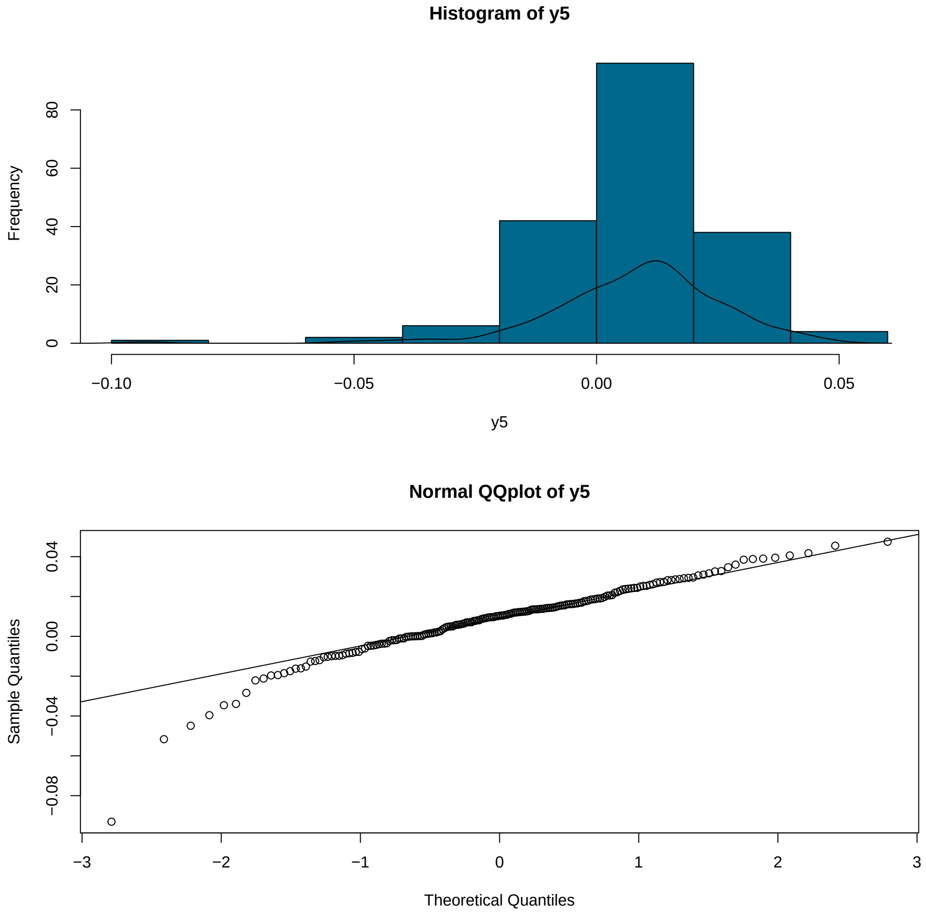


Figure 9: y5 Histogram & QQ plot

With the previous graphs (histogram and QQ-plot), we confirm that our data do not appear to follow a strictly zero-mean normal distribution but a slightkly skewed one. We can also test for normality using a Shapiro-Wilk normality test which we reject. Nevertheless, we are satisfied with our data and we believe that no further transformation is needed (log,diff-logs).

```
        Shapiro-Wilk normality test
data:  y5
W = 0.93113, p-value = 8.385e-08
```

Next, using the **R** function Box.test() we can compute the Box-Pierce and the Ljung-Box test statistic . We use 50 lags and we reject the null hypothesis (for $\alpha = 0.05$ significance level) on both tests, that means the "overall" autocorrelations of the time series are different for at least the last 4 years .

Now it is time to test the stationarity of our time series. First we try to fit the series, as best as we can, using an Autoregressive (AR) model. We use the estimated *order* in order to perform an augmented Dickey-Fuller test of unit root based on Random Walk with Drift.

```
###############################################################
# Augmented Dickey-Fuller Test Unit Root / Cointegration Test
    #
###############################################################

The value of the test statistic is: -9.7928 47.9495
```

We can see that the test statistic has a value (-9.7) that is significantly lower than the critical value (-2), thus we reject the null hypothesis that the model has a unit root . In conclusion, we have statistical evidence that our series is stationary, thus we can proceede with the Box-Jenkins methodology.

1. Identification step

We will examine the autocorrelation and partial autocorrelation plots in order to get a first impression of which lags are more heavily correlated.

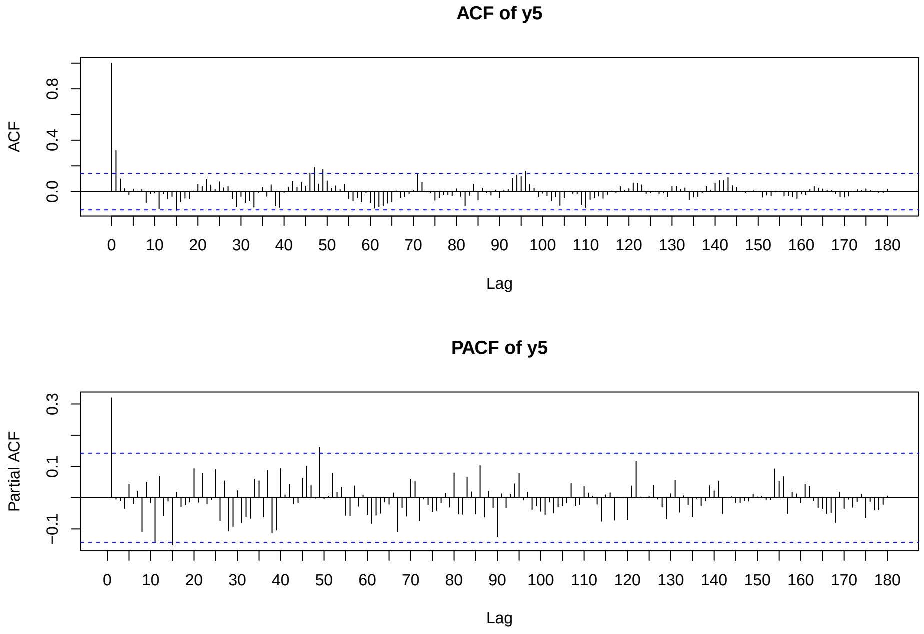


Figure 10: y5 ACF & PACF plot

- From the **acf** plot we see that we have statistically significant peaks at lags 1,47,49.
- From the **pacf** plot we see that we have statistically significant peaks at lags 15,49 .

In the acf graphs, only lag 1 is significantly larger than the bound. All the others are very close to it.

2. Estimation step

- **MA(1)**
  We will first estimate a Moving Average MA(1) model as a point of reference.

```
Call:
arima(x = y5, order = c(0, 0, 1))
Coefficients:
         ma1  intercept
      0.2976     0.0083
s.e.  0.0654     0.0016
sigma^2 estimated as 0.0002974:  log likelihood =
    499.17,  aic = -992.33
```

We calculate a rough estimate of how statistically significant the estimated parameter is:

$$\frac{\hat{\theta}_1}{s.e(\hat{\theta}_1)} = \frac{0.2976}{0.0654} \approx 4.54 > 2$$

which indicates that the estimated parameter is statistically significant.

- **AR(1)** Now, an AR(1) model gives us the following results

```
Call:
arima(x = y5, 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 = -993.95
```

$$y_t = 0.0083(1 - 0.3179) + 0.3179 y_{t-1} + \epsilon_t$$

We the previous scores in mind, we prefer the AR(1) from MA(1) and proceed to the estimation step

3. Diagnostic plots

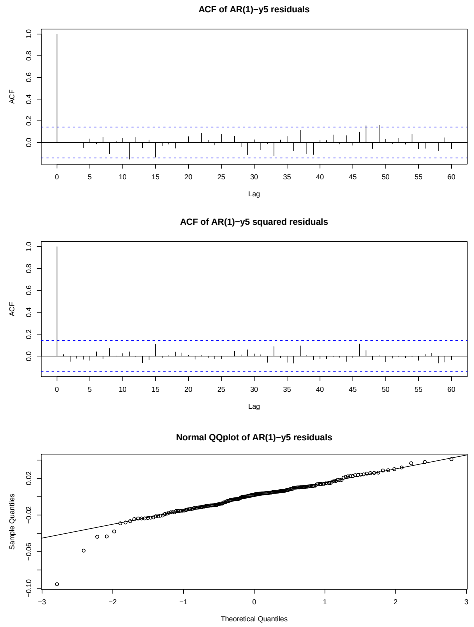


Figure 11: y5 Residuals ACF & PACF plot for AR(1)

Given the previous graphs, we could just stop here and proceed to the predictions because all the autocorrelations of the residuals are rougly inside the bounds. Nevertheless, we add the ar49 term

- **AR(49)**

```
Call:
arima(x = y5, order = c(49, 0, 0), fixed = fx)
...
sigma^2 estimated as 0.0002819:  log likelihood =
    503.17,  aic = -998.34
```

$$y_t = 0.0083(1 - 0.3036) + 0.3036y_{t-1} + 0.1936y_{t-49} + \epsilon_t$$

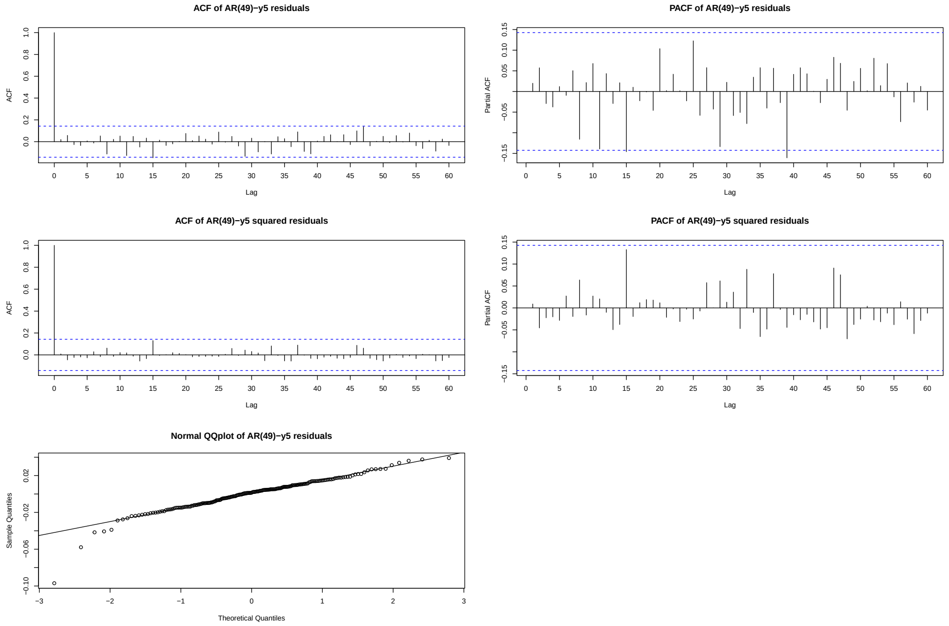


Figure 12: y5 Residuals ACF & PACF plot for AR(49)

- **ARMA(49,1)**

```
Call:
arima(x = y5, order = c(49, 0, 1), fixed = fx)
...
sigma^2 estimated as 0.0002831:  log likelihood = 502.7,
     aic = -997.4
```

$$y_t = 0.0083 + 0.3024\epsilon_{t-1} + 0.2098y_{t-49} + \epsilon_t$$

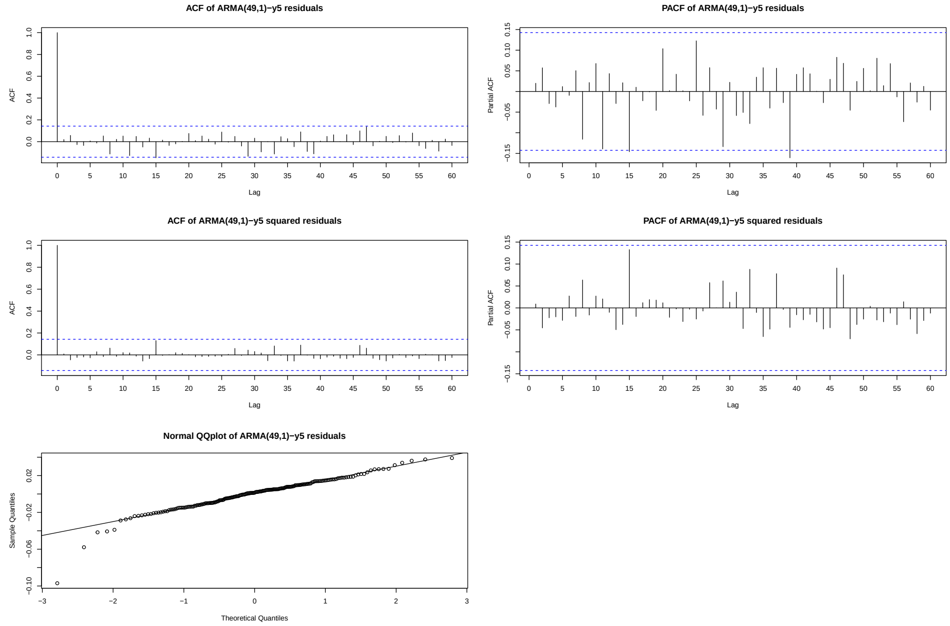


Figure 13: y5 Residuals ACF & PACF plot for ARMA(49,1)

We will stop here, although we could have continued and added ar39. Our best model is AR(49).

Model	**terms**	**AIC**	**Log-Likelihood**	σ^2
MA(1)	ϵ_{t-1}	-992	499	0.0002974
AR(1)	y_{t-1}	-993	499	0.0002948
ARMA(1,1)	y_{t-1}, ϵ_{t-1}	-991	499	0.0002948
AR(49)	y_{t-1}, y_{t-49}	-998	503	0.0002819
ARMA(49,1)	y_{t-49}, ϵ_{t-1}	-997	502	0.0002831

4. Predictions

We compute predictions based on an estimated AR model. We predict for 8 months ahead.

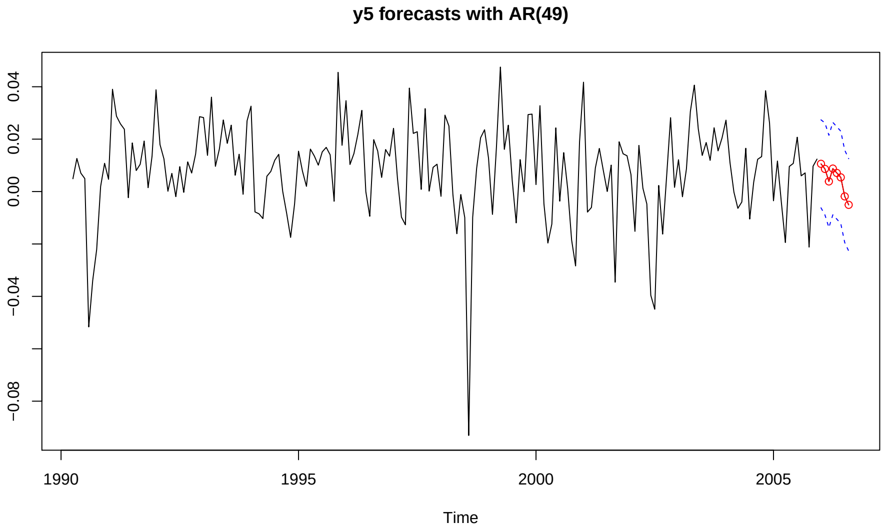


Figure 14: y5 Predictions for AR(49)

As we can see, the confidence intervals are pretty wide, meaning we do not have very good predictions. But that is expected, since we are dealing with real-world data in the stock market, which are basically unpredictable.

Exercise 2

In this exercise, we will try to develop multivariate regression models for two of the dependent variables using multiple predictors from the independent variables ($x_1 \rightarrow x_{15}$)

Y8 / EMN

We will first plot some scatterplot graphs and explore if there is any correlations between certain independent variables and our depedent variable y_8.

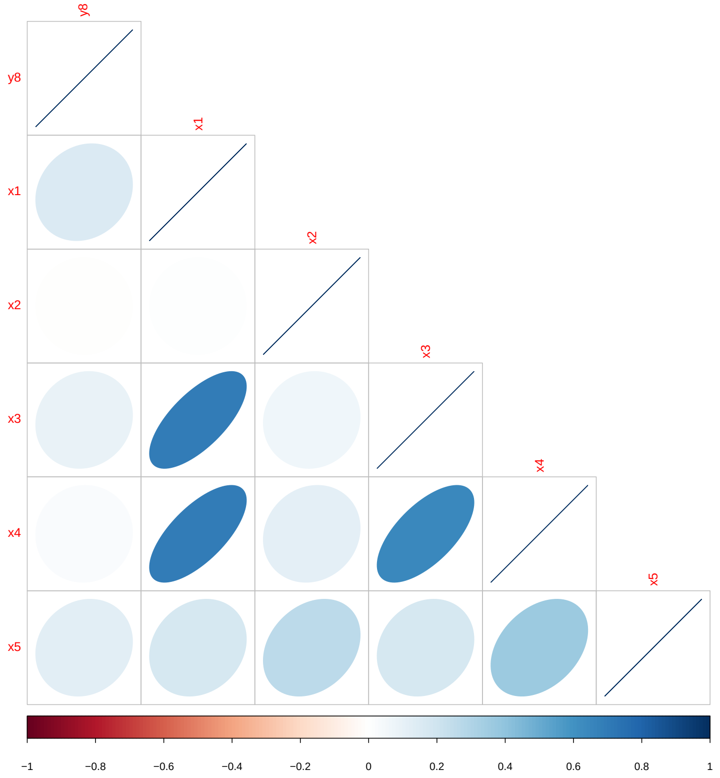


Figure 15: y8 correlation plot with $x_1 \rightarrow x_5$

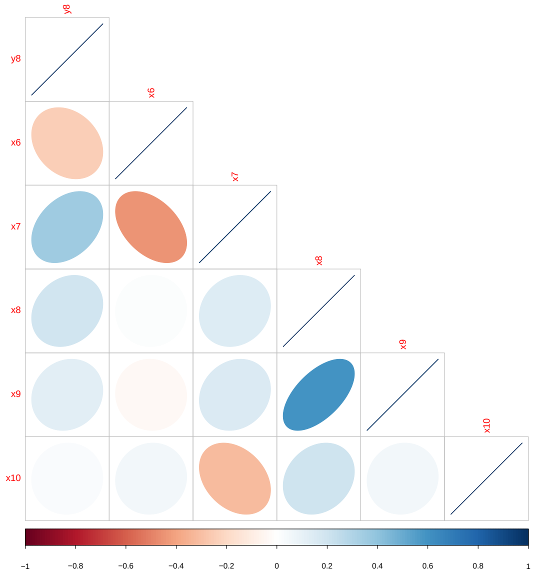


Figure 16: y8 correlation plot with $x_6 \rightarrow x_{10}$

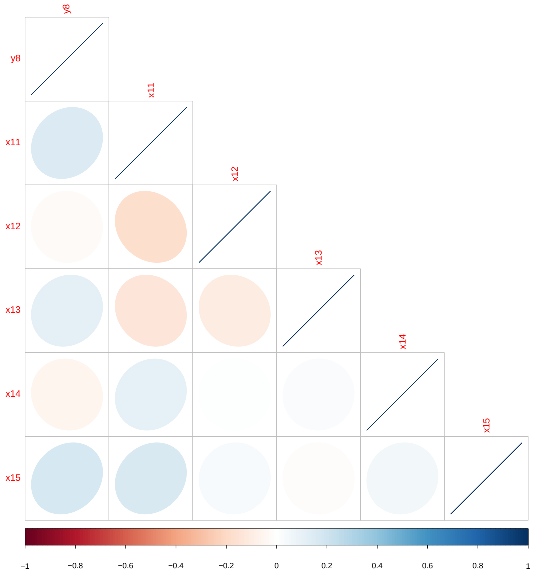


Figure 17: y8 correlation plot with $x_{10} \rightarrow x_{15}$

From the correlation plots we notice that no particular regressor is really correlated to y8. But $x_1, x_3, x_5, x_6, x_7, x_8, x_9, x_{11}, x_{13}, x_{15}$ are more related than the others.

We will start with a linear regression using all independent variables, for a point of reference.

```
Call:
lm(formula = y8 ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 +
    x10 + x11 + x12 + x13 + x14 + x15, data = ts_data)

Residuals:
       Min         1Q     Median         3Q        Max
-0.0182020 -0.0045284 -0.0004771  0.0044112  0.0226891
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.0008281  0.0015074  -0.549 0.583470
x1           0.0863401  0.0268521   3.215 0.001555 **
x2           0.0159894  0.0158870   1.006 0.315608
x3          -0.0038931  0.0200443  -0.194 0.846226
x4          -0.0147831  0.0133854  -1.104 0.270943
x5           0.0836747  0.0224237   3.732 0.000258 ***
x6           0.0398218  0.0209061   1.905 0.058466 .
x7           0.0724953  0.0124832   5.807 2.97e-08 ***
x8           0.0783604  0.0619365   1.265 0.207511
x9          -0.0122433  0.0491200  -0.249 0.803461
x10         -0.0050010  0.0213979  -0.234 0.815484
x11          0.8400574  0.5591602   1.502 0.134828
x12         -0.0057518  0.0590607  -0.097 0.922531
x13          0.0166690  0.0102864   1.620 0.106948
x14          0.0106063  0.0224348   0.473 0.636979
x15          0.9641552  0.4032211   2.391 0.017870 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.007484 on 173 degrees of freedom
Multiple R-squared:  0.3151,    Adjusted R-squared:  0.2557
F-statistic: 5.306 on 15 and 173 DF,  p-value: 1.078e-08
```

We reached an adjusted R-squared 0.25, which is not good. We used all variables and it appears that they may not contain any important information about our dependent variable. We will use the backwards, forward and stepwise elimination methods in order to insert or remove independent variables from our model according to **bic**.

- **Stepwise elimination method**

```
> stepSR$anova
    Step Df     Deviance Resid. Df  Resid. Dev       AIC
1          NA           NA       173 0.009690403 -1783.143
2  - x12  1 5.312606e-07       174 0.009690934 -1788.375
3   - x3  1 2.165593e-06       175 0.009693100 -1793.574
4  - x10  1 3.530398e-06       176 0.009696630 -1798.747
5   - x9  1 7.198168e-06       177 0.009703828 -1803.849
6  - x14  1 9.724977e-06       178 0.009713553 -1808.901
7   - x2  1 7.401645e-05       179 0.009787570 -1812.708
8   - x4  1 8.893502e-05       180 0.009876505 -1816.240
9  - x13  1 1.098318e-04       181 0.009986337 -1819.392
10 - x11  1 8.623876e-05       182 0.010072575 -1823.008
11  - x6  1 2.181089e-04       183 0.010290684 -1824.201
```

All methods agreed that the best regression model is the following:

$$\begin{aligned} y8_t = &-0.0008034 + 0.0439172 x1_t + 0.0578252 x5_t \\ &+ 0.0587388 x7_t + 0.1063517 x8_t + 1.07406 x15_t + \epsilon_t \end{aligned}$$

```
Call:
lm(formula = y8 ~ x1 + x5 + x7 + x8 + x15, data =
    ts_data)
...
Residual standard error: 0.007499 on 183 degrees of
    freedom
Multiple R-squared:  0.2727,    Adjusted R-squared:
    0.2528
F-statistic: 13.72 on 5 and 183 DF,  p-value: 2.223e-11
```

Below, we can visualize the residuals:

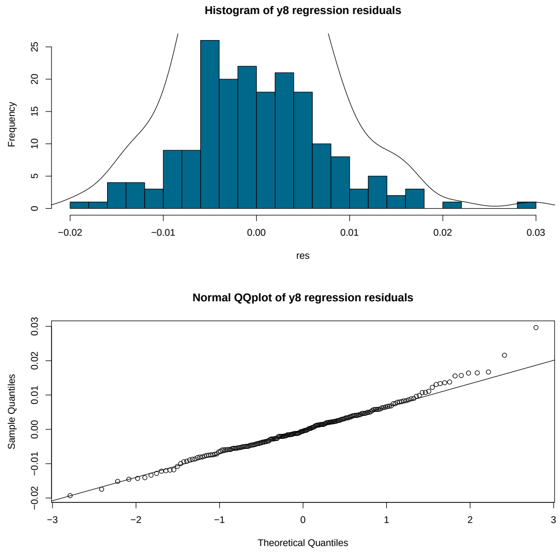


Figure 18: y8 residuals histogram and qq plot

The residuals seem to sligthly follow a slightly skewed normal distribution but with a right fat tail, so the normality problem is not yet treated. We can also visualize the autocorrelation plots for the residuals:

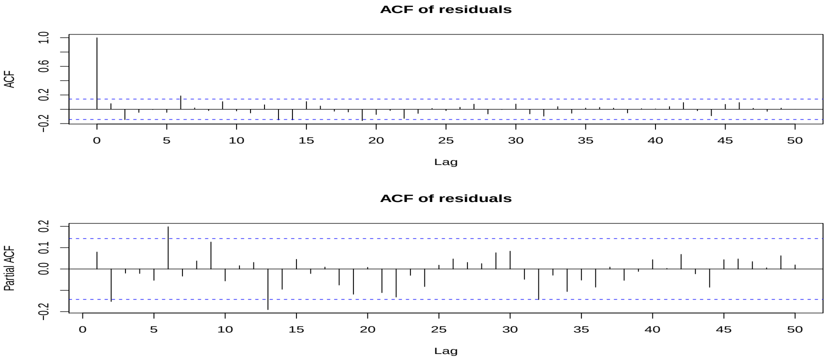


Figure 19: y8 residuals histogram and qq plot

The most important autocorrelations seem to be at lags 6,13

Y5 / ED

We will first plot some scatterplot graphs and explore if there is any correlations between certain independent variables and our depedent variable y_5.

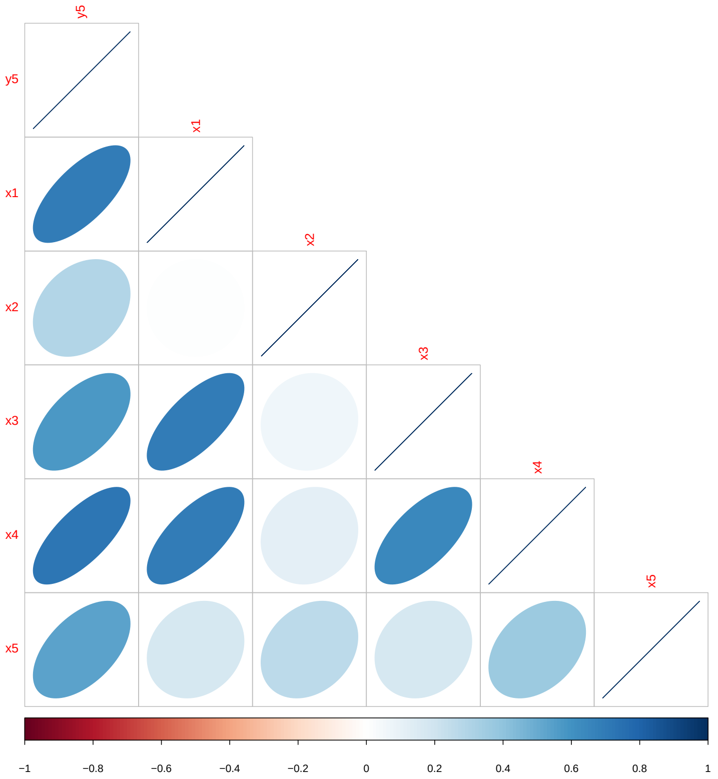


Figure 20: y5 correlation plot with $x_1 \rightarrow x_5$

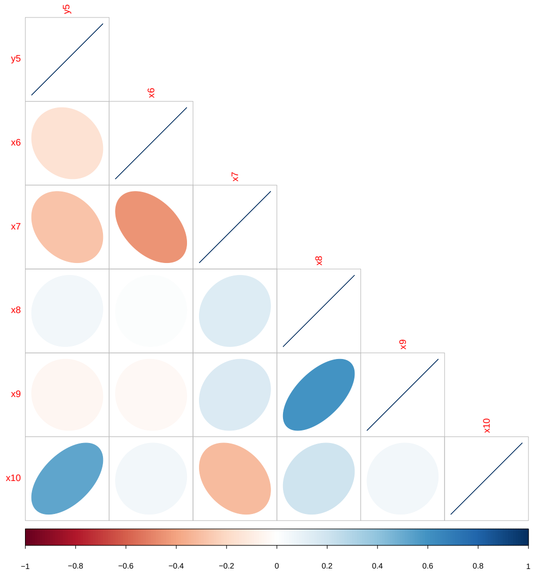


Figure 21: y5 correlation plot with $x_6 \rightarrow x_{10}$

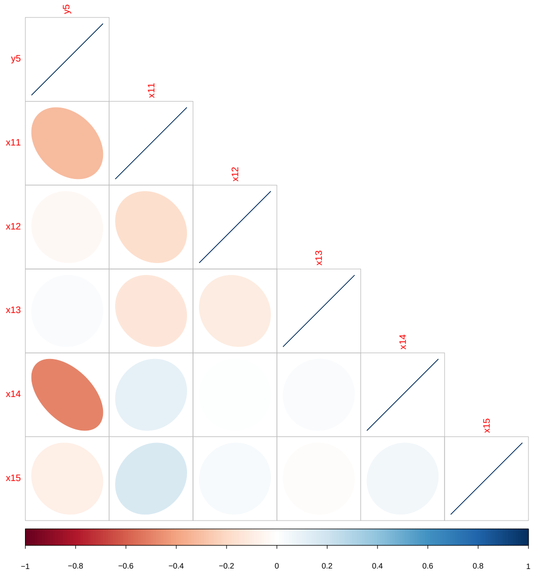


Figure 22: y5 correlation plot with $x_{10} \rightarrow x_{15}$

From the correlation plots we notice that $x_1, x_3, x_4, x_5, x_7, x_{10}, x_{11}, x_{14}$ are more heavily related to y_5. The other indepedent variables are more randomly scattered.

We will start with a linear regression using all independent variables, for a point of reference.

```
Call:
lm(formula = y5 ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 +
    x10 + x11 + x12 + x13 + x14 + x15, data = ts_data)
Residuals:
      Min        1Q    Median        3Q       Max
-0.029063 -0.005366 -0.000760  0.005380  0.032318
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.003484   0.001802   1.933 0.054812 .
x1           0.198069   0.032100   6.170 4.69e-09 ***
x2           0.052381   0.018992   2.758 0.006440 **
x3           0.027868   0.023962   1.163 0.246434
x4           0.071202   0.016001   4.450 1.54e-05 ***
x5           0.203445   0.026806   7.589 1.91e-12 ***
x6           0.084835   0.024992   3.394 0.000853 ***
x7           0.035551   0.014923   2.382 0.018290 *
x8           0.173611   0.074042   2.345 0.020174 *
x9          -0.042846   0.058720  -0.730 0.466586
x10          0.035296   0.025580   1.380 0.169428
x11         -1.317569   0.668447  -1.971 0.050309 .
x12          0.062326   0.070604   0.883 0.378597
x13          0.010274   0.012297   0.836 0.404567
x14          0.012381   0.026820   0.462 0.644920
x15          0.719747   0.482030   1.493 0.137217
---
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: < 2.2e-16
```

We reached an adjusted R-squared 0.75, which is satisfying enough. But we used all variables and some of them, as we saw in the graphs, may not contain any important information about our dependent variable. We will run stepwise methods in order to insert or remove independent variables from our model according to **bic**

- **Backwards elimination method**

```
stepBE<-step(y5res_fitall, scope=list(lower = ~ 1,
                                upper= ~
   x1+x2+x3+x4+x5+x6+x7+x8+x9+x10+x11+x12+x13+x14+x15),
           direction="backward",k = log(n), criterion
   = "BIC", data=ts_data)
> stepBE$anova
   Step Df     Deviance Resid. Df Resid. Dev       AIC
1       NA           NA       173 0.01384850 -1715.662
2 - x14  1 1.705937e-05       174 0.01386556 -1720.672
3  - x9  1 3.619523e-05       175 0.01390176 -1725.421
4 - x13  1 5.555499e-05       176 0.01395731 -1729.909
5  - x3  1 7.760502e-05       177 0.01403492 -1734.102
6 - x12  1 9.235978e-05       178 0.01412728 -1738.104
7 - x10  1 1.752784e-04       179 0.01430256 -1741.016
8 - x15  1 1.702909e-04       180 0.01447285 -1744.020
```

As we see, the process starts with the full model. In the first step it removes x_{14} because it offers the greater reduction of BIC in comparison to the other variables. Then it removes x_9 e.t.c The estimation model is of the following form:

$$
\begin{aligned}
y_5 = {} & 0.00558 + 0.22065x_1 + 0.05977x_2 + 0.07104x_4 \\
& + 0.20678x_5 + 0.08547x_6 + 0.03601x_7 + 0.16004x_8 - 1.58502x_{11} + \epsilon
\end{aligned}
$$

- **Forward elimination method**

```
> stepFS$anova
  Step Df     Deviance Resid. Df Resid. Dev       AIC
1      NA           NA       188 0.06205517 -1510.818
2 + x4 -1 0.0323427278       187 0.02971244 -1644.766
3 + x5 -1 0.0053885178       186 0.02432392 -1677.345
4 + x1 -1 0.0058485463       185 0.01847538 -1724.082
5 + x2 -1 0.0018110678       184 0.01666431 -1738.339
6 + x6 -1 0.0006361916       183 0.01602812 -1740.454
7 + x7 -1 0.0006307108       182 0.01539741 -1742.800
```

The process starts with an empty model. In the first step it adds x_4 because it offers the greater reduction of BIC in comparison to the other variables. Then it adds x_5 e.t.c The estimation model is of the following form:

$$\begin{aligned} y_5 = 0.00558 + 0.22889x_1 + 0.06895x_2 + 0.07255x_4 \\ + 0.21483x_5 + 0.09309x_6 + 0.04026x_7 + \epsilon \end{aligned}$$

- **Stepwise elimination method**

```
> stepSR$anova
   Step Df     Deviance Resid. Df Resid. Dev       AIC
1       NA           NA       173 0.01384850 -1715.662
2 - x14  1 1.705937e-05       174 0.01386556 -1720.672
3  - x9  1 3.619523e-05       175 0.01390176 -1725.421
4 - x13  1 5.555499e-05       176 0.01395731 -1729.909
5  - x3  1 7.760502e-05       177 0.01403492 -1734.102
6 - x12  1 9.235978e-05       178 0.01412728 -1738.104
7 - x10  1 1.752784e-04       179 0.01430256 -1741.016
8 - x15  1 1.702909e-04       180 0.01447285 -1744.020
```

The process starts with a full model. In the first step it removes x_{14} e.t.c

In conclusion, the best regression model is the following:

$$\begin{aligned} y_5 = 0.00558 + 0.22065x_1 + 0.05977x_2 + 0.07104x_4 \\ + 0.20678x_5 + 0.08547x_6 + 0.03601x_7 + 0.16004x_8 - 1.58502x_{11} + \epsilon \end{aligned}$$

```
Call:
lm(formula = y5 ~ x1 + x2 + x4 + x5 + x6 + x7 + x8 + x11,
    data = ts_data)
...
Residual standard error: 0.008967 on 180 degrees of freedom
Multiple R-squared:  0.7668,    Adjusted R-squared:  0.7564
F-statistic: 73.97 on 8 and 180 DF,  p-value: < 2.2e-16
```

Below, we can visualize the residuals:

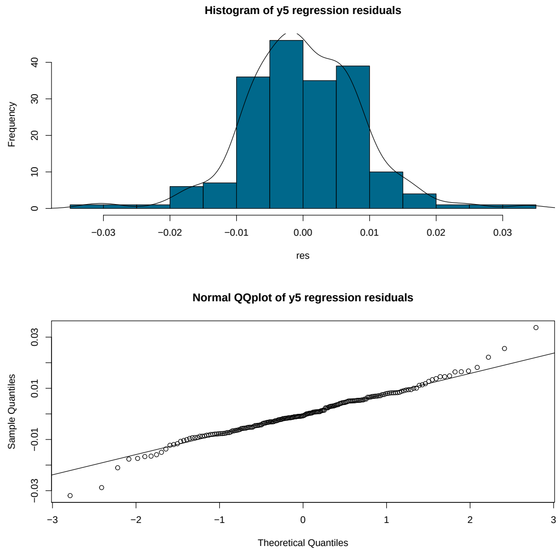


Figure 23: y5 residuals histogram and qq plot

The residuals seem to follow the normal distribution except the extreme values. We can also visualize the autocorrelation plots for the residuals:

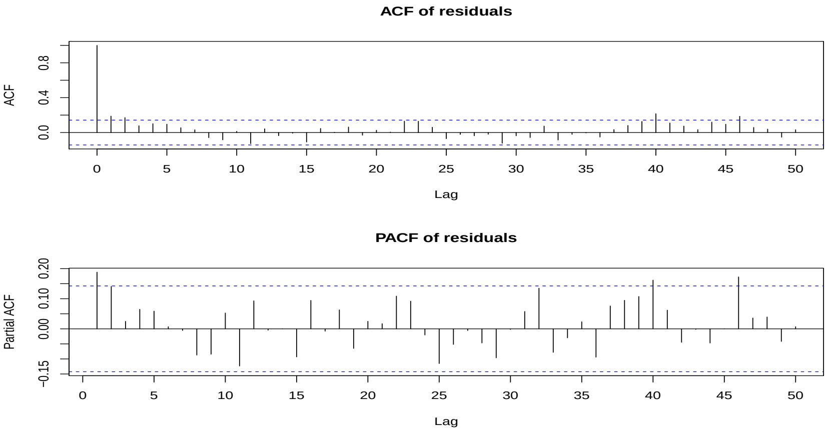


Figure 24: y5 residuals acf & pacf plot

The most important autocorrelations seem to be at lags 1,2,40,46.

Exercise 3

3.a. Regression with ARMA

y5/ ED

We will continue from exercise 2 and will try to treat the autocorrelation problems at the simple regression using time series models at the residuals. As we saw at **Figure 24**, we have peaks at lags 1,2,40,46 at the acf plot and at lags 1,40,46 at the pacf plot.

Experimenting with different models, we first tried to model the residuals with a simple MA(2), but in order to treat the statistically significant peaks at lags 40 and 46 we eventually used a restricted MA(46).

```
Call:
arima(x = y5, order = c(0, 0, 46), xreg = cbind(x1, x2, x4,
    x5, x6, x7, x8,
    x11), fixed = fx)
...
sigma^2 estimated as 6.81e-05:  log likelihood = 637.38,  aic
    = -1246.77
```

Below we have the model equation:

$$\begin{aligned} y5_t = {} & 0.00558 + 0.22065x1_t + 0.05977x2_t + 0.07104x4_t \\ & + 0.20678x5_t + 0.08547x6_t + 0.03601x7_t + 0.16004x8_t - 1.58502x11_t \\ & + \epsilon_t + 0.1393\epsilon_{t-1} + 0.1229\epsilon_{t-2} + 0.1491\epsilon_{t-40} + 0.1576\epsilon_{t-46} \end{aligned}$$

In the next figure we can see that we no longer have autocorrelations at the residuals

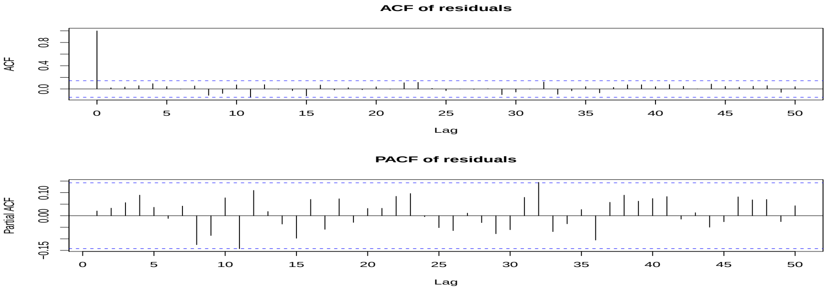


Figure 25: y5 regression with ma(46) residuals

We compute predictions based on an estimated regression model with restricted ma(46) residuals. We predict for 8 months ahead.

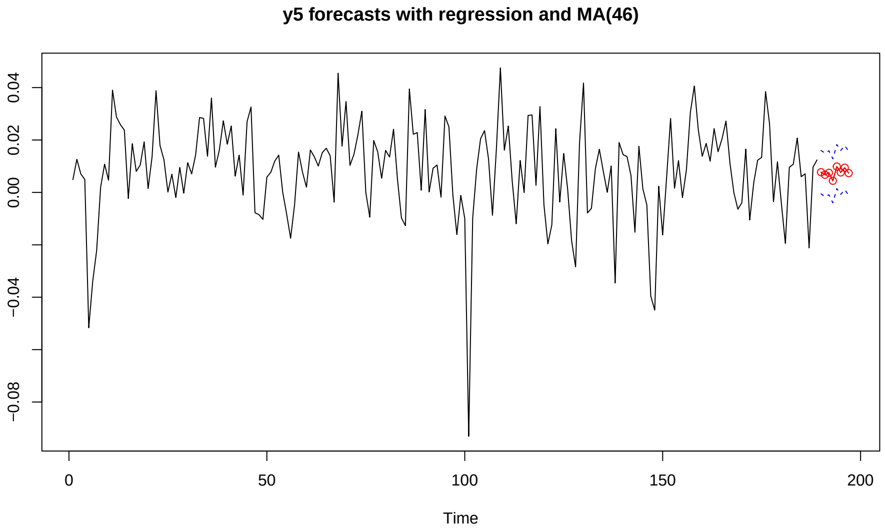


Figure 26: y5 Predictions for Regression with MA(46)

As we can see, we achieved narrower confidence intervals than before, meaning we have greater certainty predictions.

Y8/ EMN

We will continue from exercise 2 and will try to treat the autocorrelation problems at the simple regression using time series models at the residuals. As we saw at **Figure 19**, we have a small peak at lags 6 at the acf plot and a bigger peak at lags 13 at the pacf plot.

Experimenting with different models, we first tried to model the residuals with a simple AR(2), but in order to treat the statistically significant peak at lag 6 we eventually used a restricted AR(6). We did not treat lag 13 because it was not significant.

```
Call:
arima(x = y8, order = c(6, 0, 0), xreg = cbind(x1, x5, x7,
    x8, x15), fixed = fx)
...
sigma^2 estimated as 5.229e-05:  log likelihood = 663.34,
    aic = -1310.69
```

Below we have the model equation:

$$\begin{aligned} y8_t &= -0.0007 + 0.0420x1_t + 0.0545x5_t \\ &\quad + 0.0527x7_t + 0.1089x8_t + 1.0597x15_t + 0.2039y8_{t-6} + \epsilon_t \end{aligned}$$

In the next figure we can see that we no longer have important autocorrelations at the residuals.

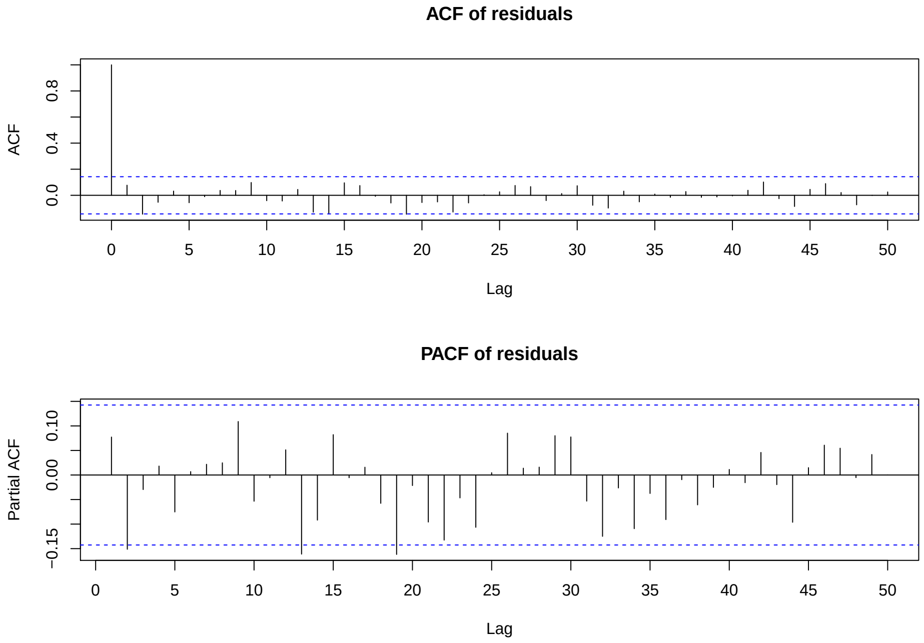


Figure 27: y8 regression with AR(6) residuals

We compute predictions based on an estimated regression model with restricted AR(6) residuals. We predict for 8 months ahead.

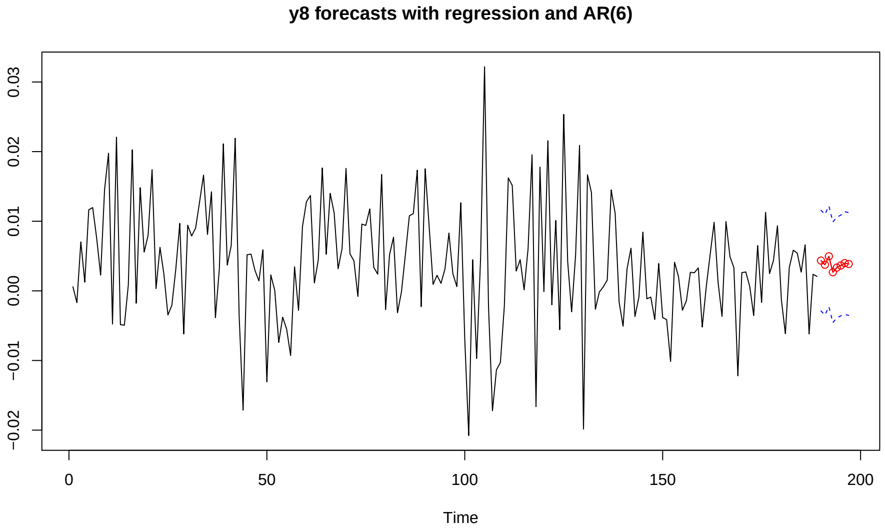


Figure 28: y8 Predictions for Regression with AR(6)

In this case, we did not achieve anything more. Perhaps because the autocorrelations were not that significant, and the regression was very weak. It appears that y8 is a difficult time series to model.

3.b. Regression with ARCH

y5/ ED

Now that we treated autocorrelations problems, in this section we will try to treat heteroscedasticity problems. With various experiments on our $y5$ time series, an ARMA(49,1) model treated both autocorrelation and heteroscedasticity. But this is not the best model in terms of in-sample metrics like the AIC that we used so far. Thus, we will continue improving the best model so far, the restricted AR(46) model with external regressors. After modelling our series, we plot the following graph:

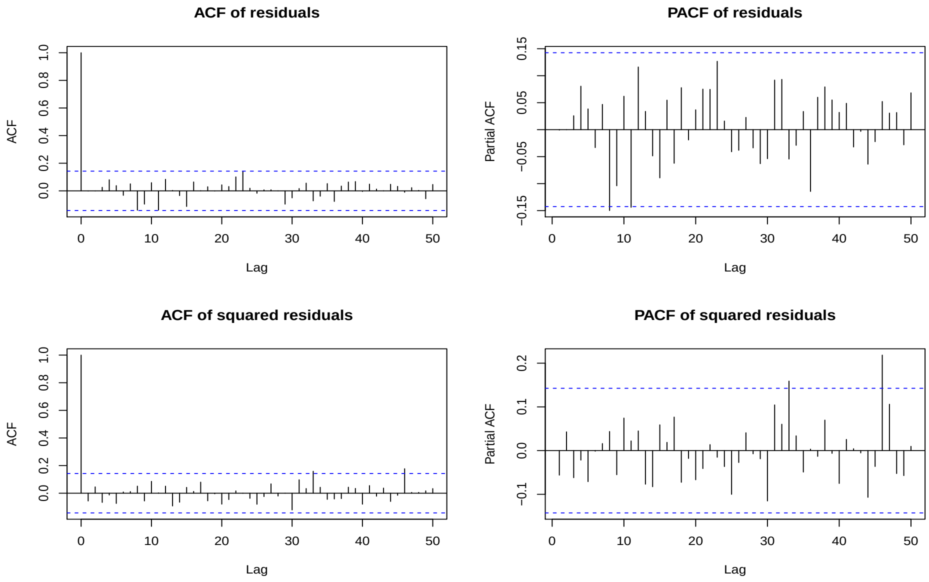


Figure 29: y5 residuals with AR(46)+Regression

```
        Box-Ljung test
data:  residuals(y5res_ar46)
X-squared = 38.031, df = 50, p-value = 0.8926

        Box-Ljung test
data:  residuals(y5res_ar46)^2
X-squared = 42.079, df = 50, p-value = 0.7795
```

As we can see from the acf and pacf on the residuals, we treated the autocorrelation problem. Also, there isn't any significant heteroscedasticity problem, as seen from the Box-tests. We could have stopped our analysis here. However we will continue, we will try to care for the peaks at lags 33 and 46 on the squared residuals pacf graph.

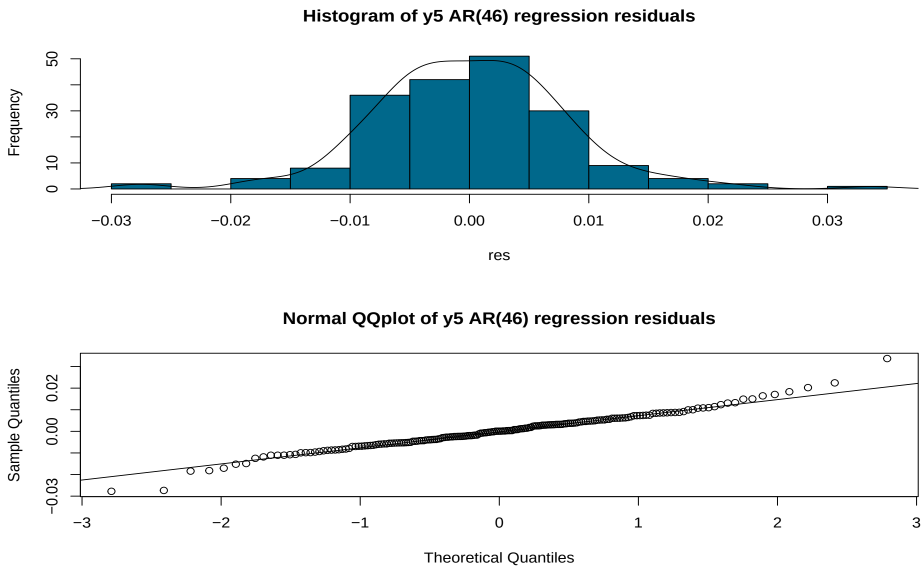


Figure 30: y5 residuals with AR(46)+Regression

As we model the conditional variance of the time series, we hope that we fix the normality problem that we saw in the previous QQ-plot.

There are two ways that we can implement this in **R**. The first way is a two step process. First we model our series with ARMA and regression. In the second step we model the residuals with ARCH or GARCH. The second way is using an all-in-one function, where we will model everything at once. We achieved this by the rugarch package .

After several experiments, we figured that a **restricted ARCH(46) - restricted AR(46) -Regression** treated every problem. Below we can see the equation that best describes our time series:

$$y5_t = \mu + \theta_1 x1_t + \theta_2 x2_t + \theta_4 x4_t + \theta_5 x5_t + \theta_6 x6_t + \theta_7 x7_t + \theta_8 x8_t - \theta_{11} x11_t$$
$$+\phi_1 y5_{t-1} + \phi_2 y5_{t-2} + \phi_{32} y5_{t-32} + \phi_{40} y5_{t-40} + \phi_{46} y5_{t-46} + \epsilon_t \qquad , \epsilon_t \sim \mathrm{GED}(0, \sigma_t^2)$$
$$\sigma_t^2 = \omega + \alpha_{33}\epsilon_{t-33}^2 + \alpha_{46}\epsilon_{t-46}^2$$

Next, the acf, pacf plots on the squared residuals and the histogram, QQ-plot:

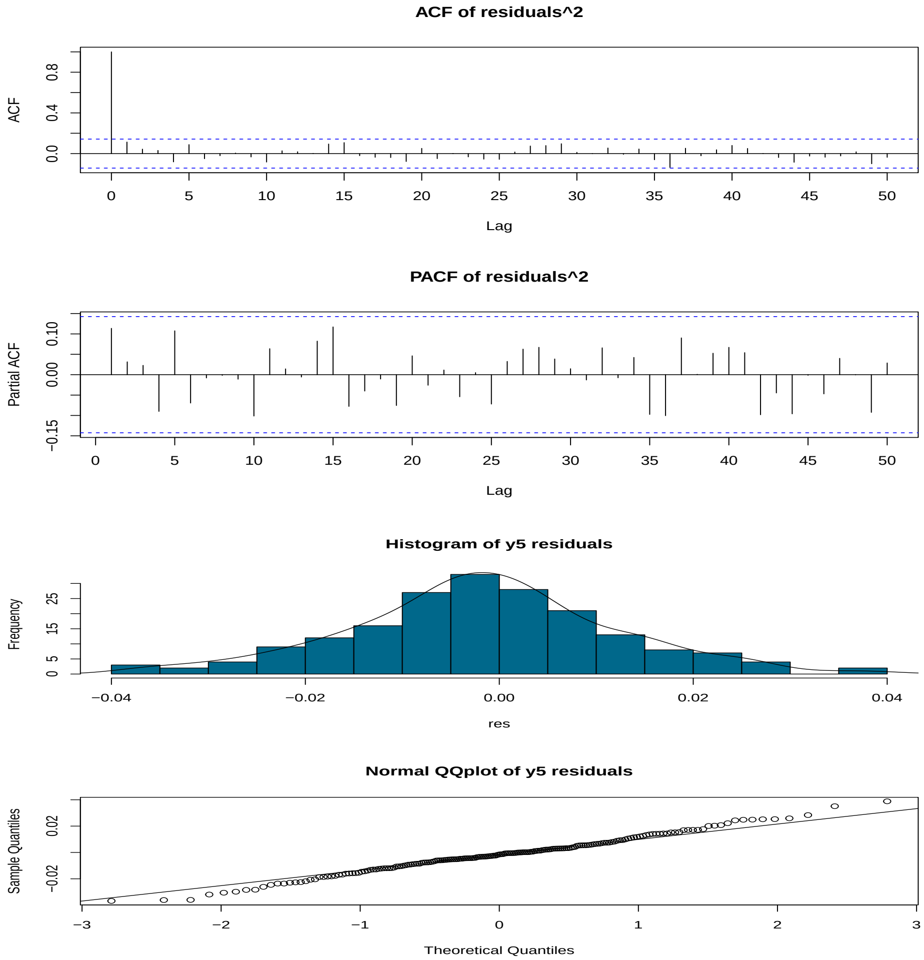


Figure 31: y5 squared residuals with AR(46)+Regression+GARCH

As we see, using a Generalized Error Distribution (Generalized Normal Distribution) also fixed the normality problem that we had with the fat-tails. We can also make forecasts for 8 months ahead using the new model :

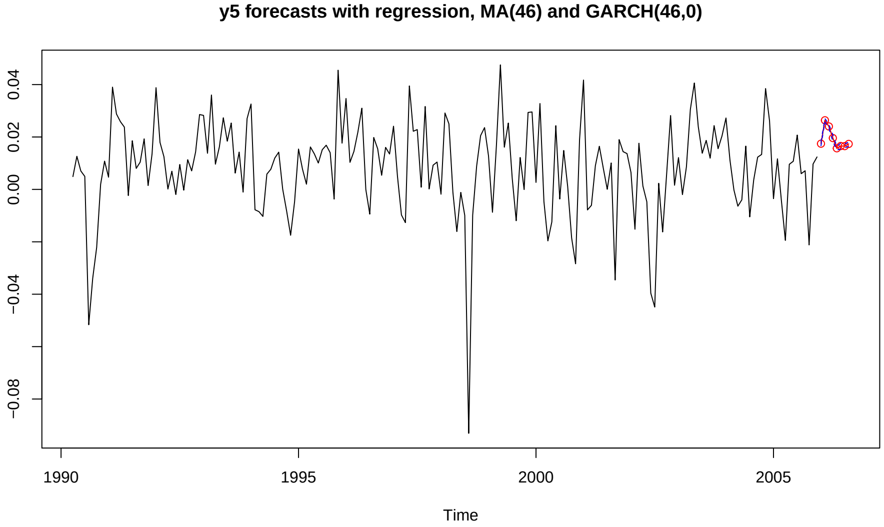


Figure 32: y5 predictions with AR(46)+Regression+GARCH

y8/ ED

Thus, we will continue improving the best model so far, the restricted AR(6) model with external regressors. After modelling our series, we plot the following graph:

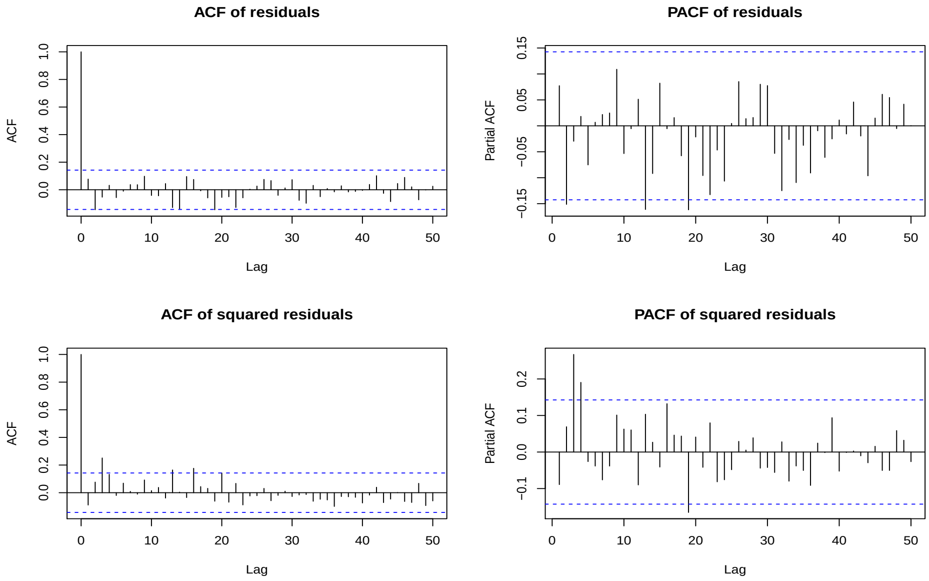


Figure 33: y8 residuals with AR(6)+Regression

```
        Box-Ljung test
data:  residuals(y8res_ar6)
X-squared = 26.387, df = 20, p-value = 0.1534

        Box-Ljung test
data:  residuals(y8res_ar6)^2
X-squared = 39.739, df = 20, p-value = 0.005389
```

As we can see from the acf and pacf on the residuals, we treated the autocorrelation problems. However, it appears that we have heteroscedasticity problems as seen in the squared residuals graphs at lags 3 & 4. We confirm this using the Box-tests as well.

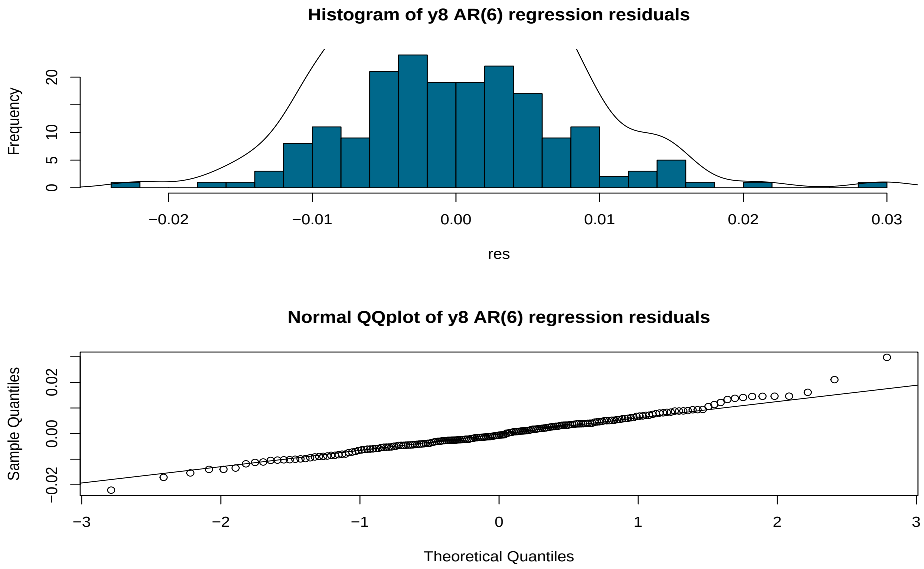


Figure 34: y8 residuals with AR(6)+Regression

We don't seem to have a big normality problem, but we notice a right fat-tail.

After several experiments, we figured that a **restricted GARCH(0,6) - restricted AR(6)** treated every problem. [8] Below we can see the equation that best describes our time series:

$$
\begin{aligned}
y8_t &= \mu + \phi_6 y8_{t-6} + \epsilon_t \qquad , \epsilon_t \sim \text{student-t}(0, \sigma_t^2) \\
\sigma_t^2 &= \omega + \beta_1 \sigma_{t-1}^2 + \beta_2 \sigma_{t-2}^2 + \beta_3 \sigma_{t-3}^2
\end{aligned}
$$

Next, the acf, pacf plots on the squared residuals and the histogram, QQ-plot:

[8] We removed the regressors because it affected our convergence and the quality of the model

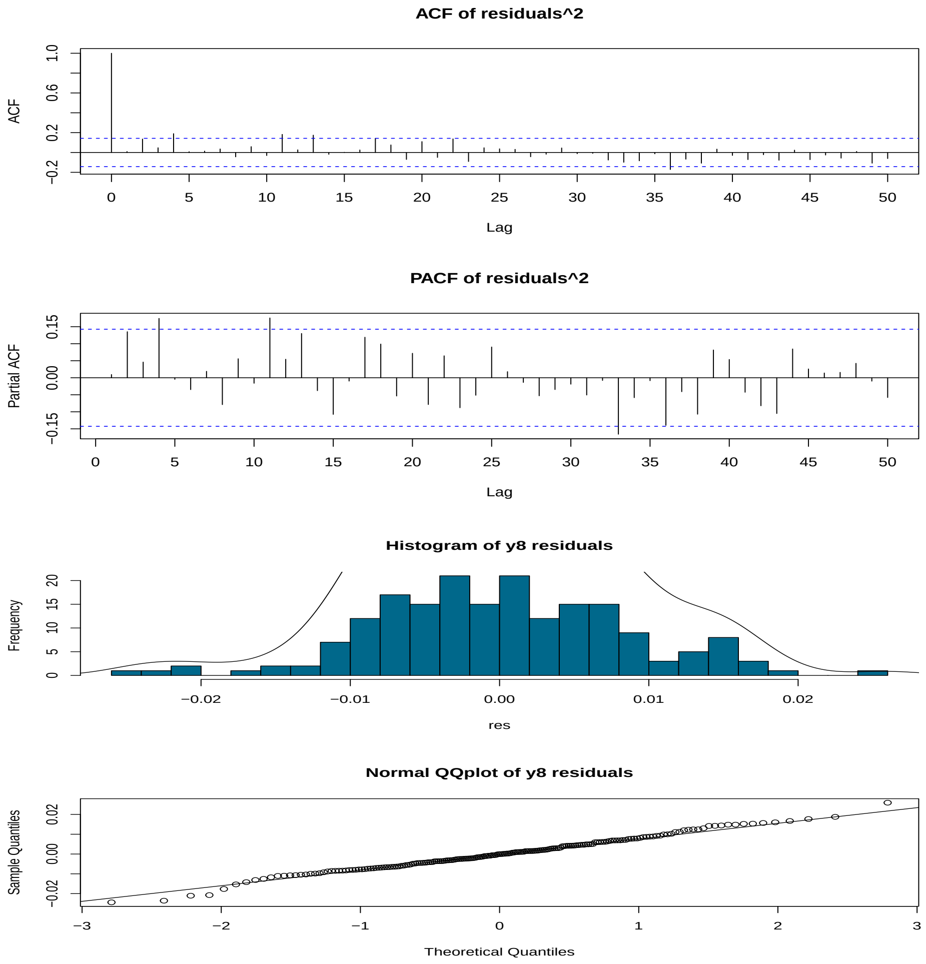


Figure 35: y8 squared residuals with AR(6)+GARCH

As we see, using a student's t distribution also fixed the normality problem that we had with the fat-tails. We can also make forecasts for 8 months ahead using the new model :

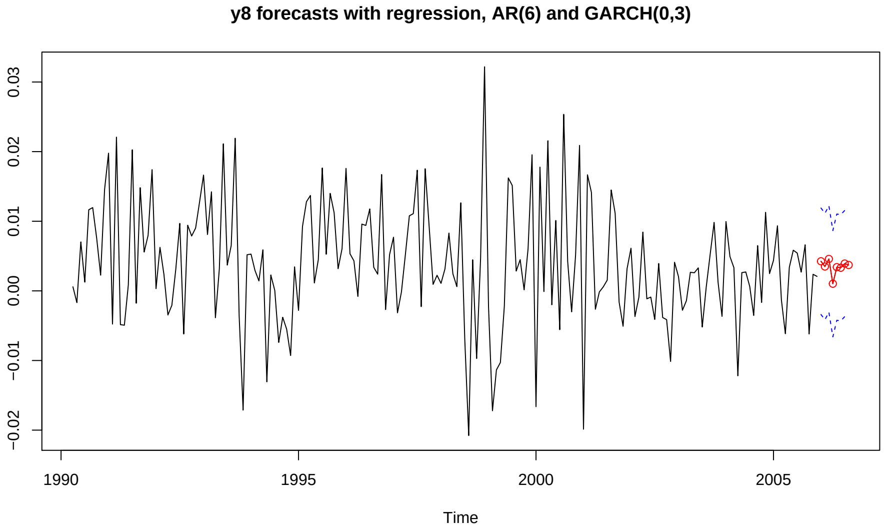


Figure 36: y8 predictions with AR(6)+GARCH

Exercise 4

Model	AIC	BIC	variable
AR(26)	-1272	-1256	y8
Regression	-1305	-1282	y8
Regression+AR(6)	-1310	-1284	y8
AR(6)+ARCH(6)	-1262	-1239	y8
AR(49)	-998	-984	y5
Regression	-1234	-1202	y5
Regression+MA(46)	-1246	-1201	y5
Regression+AR(46)+ARCH(46)	300	359	y5

- **y5** In terms of **AIC** the model with the best explainability was Regression with MA(46). In terms of **BIC** the model with the best explainability was the simple Regression.

- **y8** In terms of **AIC** the model with the best explainability was Regression with AR(6). In terms of **BIC** the model with the best explainability was again Regression with AR(6).

The process of criterion (AIC, BIC) calculation at the **rugarch** package is different. The numbers are scaled with the sample size. We transformed those numbers using our sample size in order to be in the same scale as the other measurements.

The volatility model for $y8$ had good scores although not better than the other models. The model for $y5$ had very bad AIC & BIC . This might be due to the fact that when we treated the autocorrelation problem we didn't have any heteroscedasticity problems to begin with. Thus, our ARCH modeling was incorrect.

Nevertheless, the AIC, BIC are in-sample criteria and they give us only an indication of how the model interprets the sample data. When we make predictions another metrics should be used, like MSE.