Econometrics III HW - part 1

A. Schmidt and P. Assunção

March, 2020.

Assignment 1

See the source code if interested in all functions (chunks were ommited unless relevant for the assignment). Click here to access the code. In particular, we are using our own modified functions for ACF/PACF, histograms and summary statistics that are not being shown in this report.

Introduction

Let us go back in time to the first quarter of 2009. The world economy has just been hit by a major financial crisis. In just one year, the Dutch quarterly GDP growth rate has fallen from 1.4%, in the first quarter of 2008, to -2.7%, in the first quarter of 2009. In the first quarter of 2009, at the peak of the economic recession, suppose that government officials ask you to describe the dynamics of the Dutch GDP quarterly growth rate and deliver a forecast for the two years ahead. The available sample of observed GDP growth rates spans from the second quarter of 1987 to the first quarter of 2009.

Importing and checking data

Import using read.csv2() function (remember to change the directory name - in VU computers you can only read from the downloads folder).

```r
urlRemote  <- "https://raw.githubusercontent.com/aishameriane"
pathGithub <- "/Mphil/master/EconIII/data_assign_p1.csv"
token      <- "?token=AAVGJTXCFWWNBCCXWYDJQM26RLSEY"

url      <- paste0(urlRemote, pathGithub, token)
dfData01 <- read.csv2(url, sep = ",", dec = ".", header = TRUE)
```

Check if everything is ok with the dataset: header and tail and summary statistics to check for missing data/outliers. We can see from the head and tail that the Data set indeed goes until the second quarter of 2009 and at least those observations seems to be completely filled with adequate ranges.

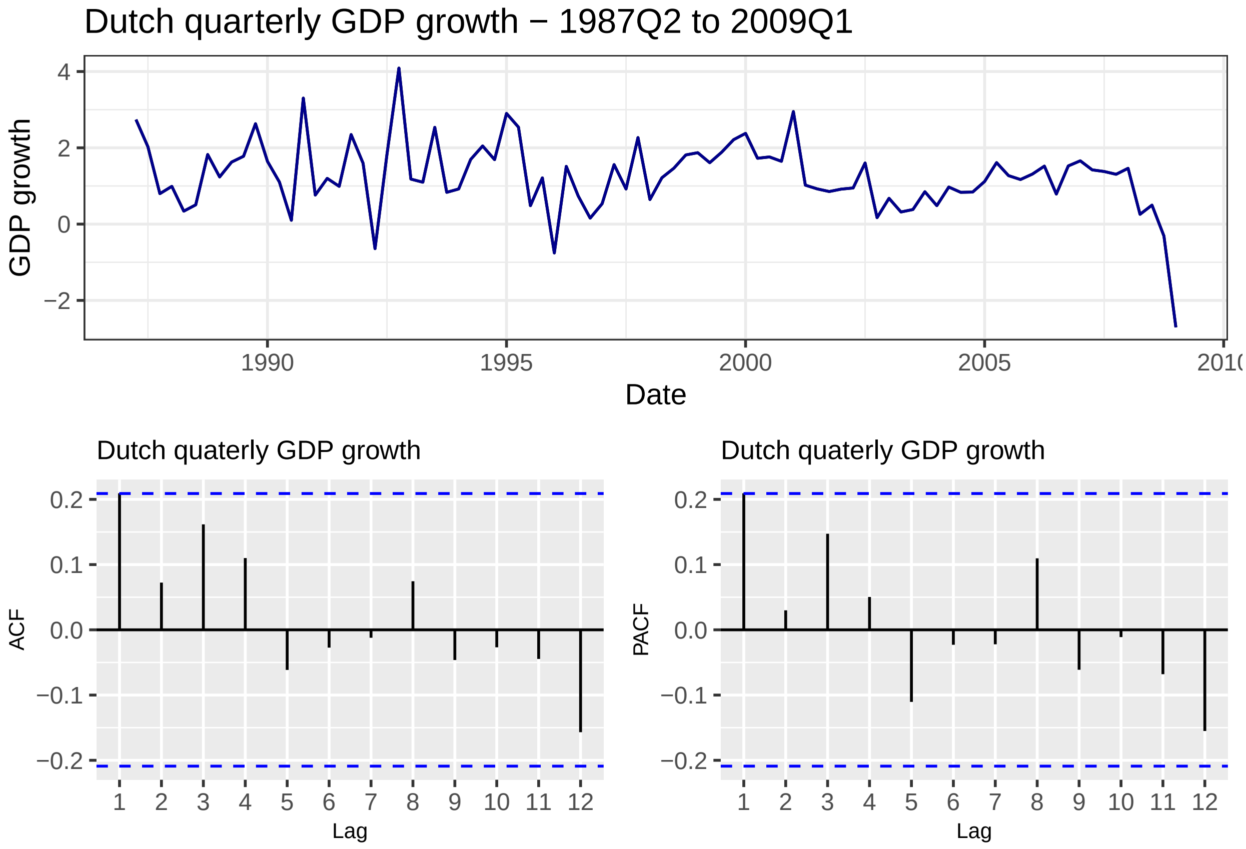
obs	GDP_QGR	obs	GDP_QGR
1987Q2	2.7404816	2007Q4	1.3058706
1987Q3	2.0275355	2008Q1	1.4630896
1987Q4	0.7983716	2008Q2	0.2568760
1988Q1	0.9903613	2008Q3	0.4978456
1988Q2	0.3396408	2008Q4	-0.3114342
1988Q3	0.5052618	2009Q1	-2.7050852

The next table has the descriptive statistics for the column that contains the values of GDP Growth. We can see that indeed we don't have any missing information and all values are numeric (there are no problems of formatting).

	descriptives
Observations	88.0000
Minimum	-2.7051
1st quartile	0.8245
Mean	1.2632
Median	1.2259
3rd quartile	1.7025
Maximum	4.0908
Desv. Pad.	0.9254

Question 1

Plot the sample of Dutch GDP quarterly growth rates that you have at your disposal. Report the 12-period sample ACF and PACF functions and comment on their shape. What does the sample ACF tells you about the dynamic properties of GDP quarterly growth rates?



Comments on the series: From the graph (top graph in the picture above), the series of the Dutch GDP quarterly growth visually does not seem to be stationary, since it looks like the volatility in the 90s is smaller than the volatility in the 2000s. Also, there is an apparent strutural break in 2008, which most likely is associated with the financial crisis.

Comments on the ACF/PACF graphs: Since the ACF is within the confidence interval (represented by the horizontal blue lines), there is no evidence of autocorrelation between the GDP growth from time $t = 0$ and $t = h$, for $h = 2, \ldots, 12$ (where 12 lags represents 3 years for quaterly data), i.e., the ACF is statistically insignificant (for a 5% significance level) from lag $h = 2$ onward, when considering a bandwitch of 12. Note, however, that autocorrelation in the first lag is very close to the upper bound of the confidence interval, which offers evidence for the existance of some relevant autocorrelation between the current quartely GDP growth and the growth in the period just before. The same applies for the PACF.

To further investigate the apparent difference in the volatility behavior, we have below the comparison for the descriptive statistics of the GDP Growth from Q1 1990 to Q4 1996 in the first column and the same descriptives for the series from Q1 2000 to Q4 2006. Notice that we are not including the period where change in volatility occured, so both columns should have near similar behaviors. However, we observe that indeed there is a discrepancy on the standard deviations, which is consistent to the visual inspection in the graph.

We will not in here make a model considering that the series could possibly be non-stationary, but it is important to have in mind that some of the techniques we saw in class would only apply if assuming stability of the AR coefficients.

	X90.96	X00.06
Observations	28.0000	28.0000
Minimum	-0.7562	0.1684
1st quartile	0.8150	0.8389
Mean	1.3970	1.1632
Median	1.2061	0.9964
3rd quartile	1.8853	1.5448
Maximum	4.0908	2.9509
Desv. Pad.	1.0934	0.6102

Question 2

Estimate an AR(p) model for the same time-series. Please use the general-to-specific modeling approach by starting with a total $p = 4$ lags and removing insignificant lags sequentially. Report the final estimated AR(p) model, working at a 5% significance level. Comment on the estimated coefficients. What do these coefficients tell you about the dynamic properties of the GDP quarterly growth rate?

Comments about the estimation procedure: We opted by a model including the intercept because, from the graphic in the previous item, it is clear that the series is not centered at zero.

We used the function `Arima()` from the `forecast` package (source code available at: https://www.rdocumentation.org/packages/forecast/versions/8.11/source), which uses the same estimation procedure as the `arima()` function that comes with the `stats` package (source code available here: https://svn.r-project.org/R/trunk/src/library/stats/R/arima.R). The estimation procedure, roughly speaking, is based on the maximum likelihood method. Since this is a numerical procedure (i.e., it requires a numerical optmization), there is a need for an initial value. The package uses the conditional sum of squares to initialize the algorithm. We tested the estimation routine with and without this initialization method and the results were the same, so we opted for having the initial condition for better performance in terms of computational time (although for this very simple model this doesn't make any significant difference).

As for the maximum likelihood procedure, it is done by filtering. More specifically, the model is treated as in its state-space representation and the Kalman filter is applied. Again, very roughly speaking, the Kalman filter is used for linear and gaussian latent models where the observation today is used to "predict" the observation tomorrow. This is made sequentially for the entire series using the bayes rule (in some sense this can be seen as a bayesian update procedure). Further details on the package procedure can be found at https://rdrr.io/r/stats/arima.html and the Kalman filter details can be found in Durbin and Koopman (2012).

First model: AR(4)

```
mAR4  <- Arima(tsData01, order = c(4,0,0))

AR4coef           <- tidy(coeftest(mAR4), stringsAsFactors = FALSE)
AR4coef           <- cbind(AR4coef[, 1], round(AR4coef[, 2:5], digits = 2))
colnames(AR4coef) <- c('Variable', 'Estimate', 'Std. Error', 't-statistic', 'P-Value')
AR4coef[,1] <- c("Lag 1", "Lag 2", "Lag 3", "Lag 4", "Intercept")

AR4coef %>%
kable("latex") %>%
kable_styling(bootstrap_options = c("striped", "hover"))
```

Variable	Estimate	Std. Error	t-statistic	P-Value
Lag 1	0.24	0.12	2.05	0.04
Lag 2	0.03	0.12	0.25	0.80
Lag 3	0.19	0.12	1.58	0.11
Lag 4	0.09	0.12	0.72	0.47
Intercept	1.21	0.20	5.96	0.00

Second model: AR(3)

```
mAR3              <- Arima(tsData01, order = c(3,0,0))

AR3coef           <- tidy(coeftest(mAR3), stringsAsFactors = FALSE)
AR3coef           <- cbind(AR3coef[, 1], round(AR3coef[, 2:5], digits = 2))
colnames(AR3coef) <- c('Variable', 'Estimate', 'Std. Error', 't-statistic', 'P-Value')
AR3coef[,1]       <- c("Lag 1", "Lag 2", "Lag 3", "Intercept")

AR3coef %>%
kable("latex") %>%
kable_styling(bootstrap_options = c("striped", "hover"))
```

Variable	Estimate	Std. Error	t-statistic	P-Value
Lag 1	0.26	0.12	2.19	0.03
Lag 2	0.03	0.12	0.25	0.80
Lag 3	0.20	0.12	1.68	0.09
Intercept	1.23	0.18	6.83	0.00

Third model: AR(2)

```
mAR2              <- Arima(tsData01, order = c(2,0,0))

AR2coef           <- tidy(coeftest(mAR2), stringsAsFactors = FALSE)
AR2coef           <- cbind(AR2coef[, 1], round(AR2coef[, 2:5], digits = 2))
colnames(AR2coef) <- c('Variable', 'Estimate', 'Std. Error', 't-statistic', 'P-Value')
AR2coef[,1]       <- c("Lag 1", "Lag 2", "Intercept")
```

```
AR2coef %>%
kable("latex") %>%
kable_styling(bootstrap_options = c("striped", "hover"))
```

Variable	Estimate	Std. Error	t-statistic	P-Value
Lag 1	0.26	0.12	2.21	0.03
Lag 2	0.06	0.12	0.48	0.63
Intercept	1.25	0.14	8.95	0.00

Fourth model: AR(1)

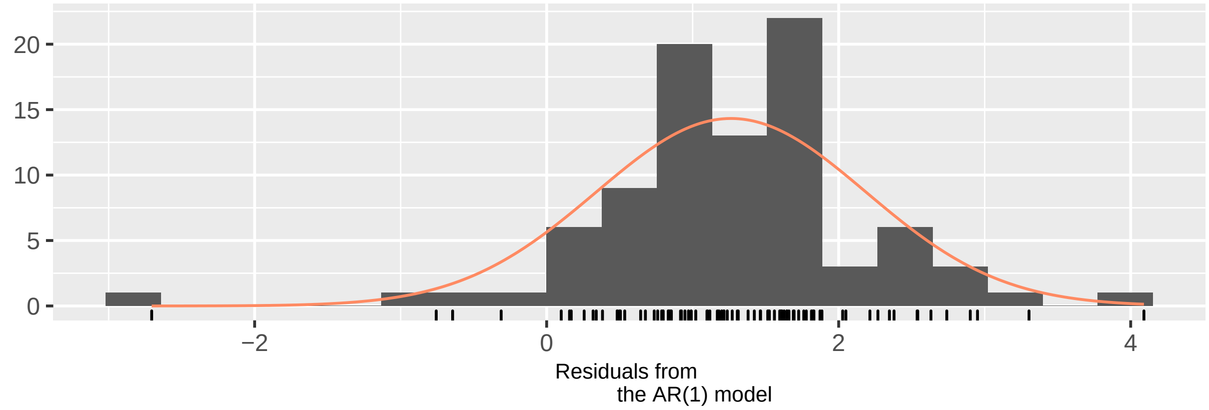
```
mAR1               <- Arima(tsData01, order = c(1,0,0))

AR1coef            <- tidy(coeftest(mAR1), stringsAsFactors = FALSE)
AR1coef            <- cbind(AR1coef[, 1], round(AR1coef[, 2:5], digits = 2))
colnames(AR1coef) <- c('Variable', 'Estimate', 'Std. Error', 't-statistic', 'P-Value')
AR1coef[,1]        <- c("Lag 1", "Intercept")

AR1coefpvalue      <- AR1coef$`P-Value`[1]
AR1coefsderror     <- AR1coef$`Std. Error`[1]

AR1coef %>%
kable("latex") %>%
kable_styling(bootstrap_options = c("striped", "hover"))
```

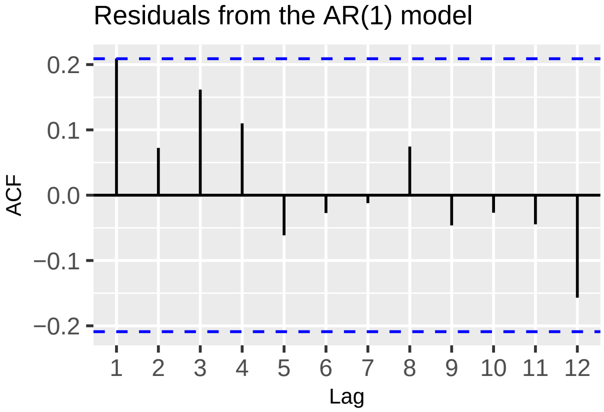
Variable	Estimate	Std. Error	t-statistic	P-Value
Lag 1	0.27	0.12	2.31	0.02
Intercept	1.25	0.13	9.61	0.00

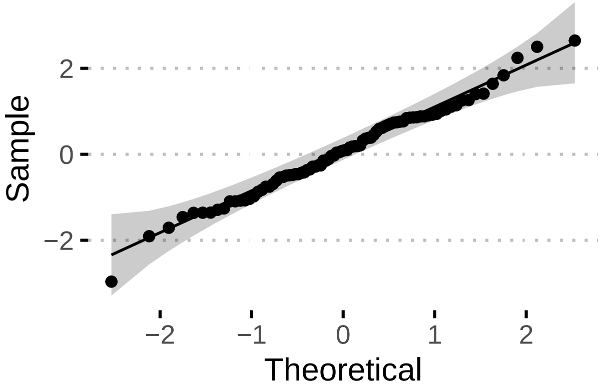
Comments on the results: Keeping only the first lag in the model is in line with the results from the ACF/PACF. Also, a relationship between previous and current periods in economic series was reported in Nelson and Plosser (1982) for the US Economy. The coefficient suggests a positive effect of 0.27 (se = 0.12; p-value= 0.02) of GDP growth in previous quarter on the growth of current quarter. Thus, shocks that occur in the previous quarter tend to persist until the current quarter. Note, however, that the sample ends around the financial crisis of 2008, which might lead to some inconsistency over the results, since we observe a longer period of sequential decline in the economic activity by the end of the sample, i.e., it seems that the persistency of the shocks in the AR model changes its behavior around 2008. A possibility would be re-estimating the model considering maybe the first half of the sample and another estimation considering only the second half of the sample or includind a dummy variable for the period pre/post crisis (which might not be ideal given the frequency of the data and low availability) or estimate a dynamic model with time-varying coefficients (i.e., include a stochastic variation in either the coefficients or the volatility).

Question 3

Check the regression residuals of the estimated AR(p) model for autocorrelation by plotting the estimated residual ACF function. Does the model seem well specified?







Comments on the results: The ACF graph show that all lags are whithin the 5% confidence interval, as we expected. As sanity check, we ran both a KS and a Jarque-Bera test for normality to check the residuals. The hypothesis of both tests are:

- *H0 :* The data comes from a normal distribution.
- *H1 :* The data does not comes from a normal distribution.

Since the p-value for the KS test is higher than 5% (KS statistic = 0.0947; p-value = 0.3847), we cannot reject the hypothesis of normality of the residuals, consistent with the QQ plot, where all sample points are inside the confidence bands.

However, when looking at the histogram of the residuals the series exhibits behavior compatible to the presence of heavy tails. The Jarque Bera test rejects the null hypothesis of normality (JB statistic = 26.298; p-value $\approx 2 \times 10^{-6}$). We opted by not follow the trail of the non-normality of the residuals because the JB test tends to be overly conservative for small sample sizes (even with a sample size of 88, which is our case).

Question 4

Make use of your estimated AR model to produce a 2-year (8 quarters) forecast for the Dutch GDP quarterly growth rate that spans until the first quarter of 2011. Report the values you obtained and explain how you derived them.

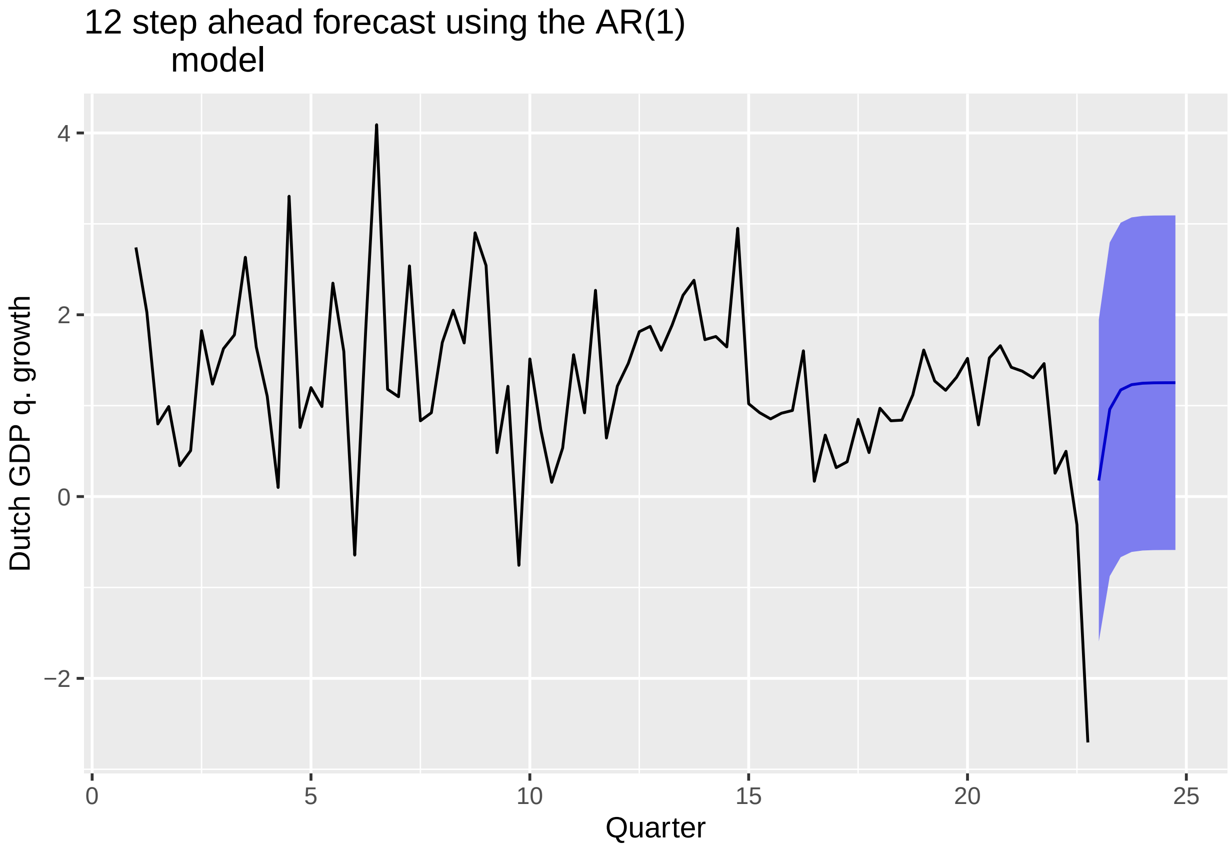
Comments about the estimation procedure: For this part of the exercise, we are using the `forecast()` function of the package `forecast` (Hyndman and Khandakar 2008). This is a general package (that is suitable for a large range of models, not just ARMA), so the function itself calls another functions depending on the type of object used as argument. The forecast is done via exponential smoothing, using the function `ets()` (source code: https://www.rdocumentation.org/packages/forecast/versions/8.11/source), whenever the number of lags is larger than 3. Broadly speaking, the one period ahead forecast is written in terms of a level, a seasonal and a trend component, where the parameters for each component are estimated internally by the method. Similarly to what is done in the estimation part, the model is written in the state space form and in the case of a linear model (such as ours), the algorithm runs iteratively for each step.

This is a more general procedure than the one studied in our classes, but given the fact that we only studied methods assuming that the true GDP is known (i.e., we only incorporated the uncertainty regarding the error term, not the uncertainty regarding the correct specification of the model neither the uncertainty about the parameter's estimates), using the functions of the package seemed the correct modeling choice.

```r
h = 8
AR1forecast  <- forecast(mAR1, h, level = 95)
vdNewDate    <- c("2009Q2", "2009Q3", "2009Q4", "2010Q1", "2010Q2", "2010Q3", "2010Q4",
                  "2011Q1")
AR1forecast  <- cbind(vdNewDate, data.frame(AR1forecast))

colnames(AR1forecast) <- c("Date", "Forecast", "L95", "H95")

AR1forecast %>%
  kable("latex") %>%
  kable_styling(bootstrap_options = c("striped", "hover"))
```

	Date	Forecast	L95	H95
23 Q1	2009Q2	0.1765306	-1.5942839	1.947345
23 Q2	2009Q3	0.9600700	-0.8750397	2.795180
23 Q3	2009Q4	1.1731221	-0.6666522	3.012896
23 Q4	2010Q1	1.2310530	-0.6090657	3.071172
24 Q1	2010Q2	1.2468050	-0.5933391	3.086949
24 Q2	2010Q3	1.2510881	-0.5890579	3.091234
24 Q3	2010Q4	1.2522527	-0.5878934	3.092399
24 Q4	2011Q1	1.2525694	-0.5875768	3.092716

```r
autoplot(forecast(mAR1, h, level = 95), main = "12 step ahead forecast using the AR(1)
         model", xlab = "Quarter", ylab = "Dutch GDP q. growth")
```



Question 5

Suppose that the innovations in your AR model are iid Gaussian. Produce 95% confidence intervals for your 2-year forecast. Furthermore, comment on the following statement issued by government officials: "Given the available GDP data, we believe that there is a low probability that the Dutch GDP growth rate will remain negative in the second quarter of 2009."

Comments on the question: Given the 95% confidence interval for our estimate of the Dutch GDP in 2009 Q2 (CI= (-1.5943;1.9473)), it is hard to make assertions on the probability of the GDP growth remaining negative in the second quarter of 2009. Basically what the confidence interval says is that we can have either negative or positive values for the GDP. However, the point estimate is equal to 0.1765, which might have mislead the policy maker. Another possibility is that the officials mistakenly interpreted the confidence interval: since it is shifted to the right (with respect to zero), they might have thought that this was a sign favouring positive growth. But this is not the correct interpretation. The interval is for the point estimate.

In fact, our point estimate is not significanly different from zero: the corresponding hypothesis test to check if the 0.1765 is equal to zero with 5% significance would point towards not rejecting the null hypothesis.

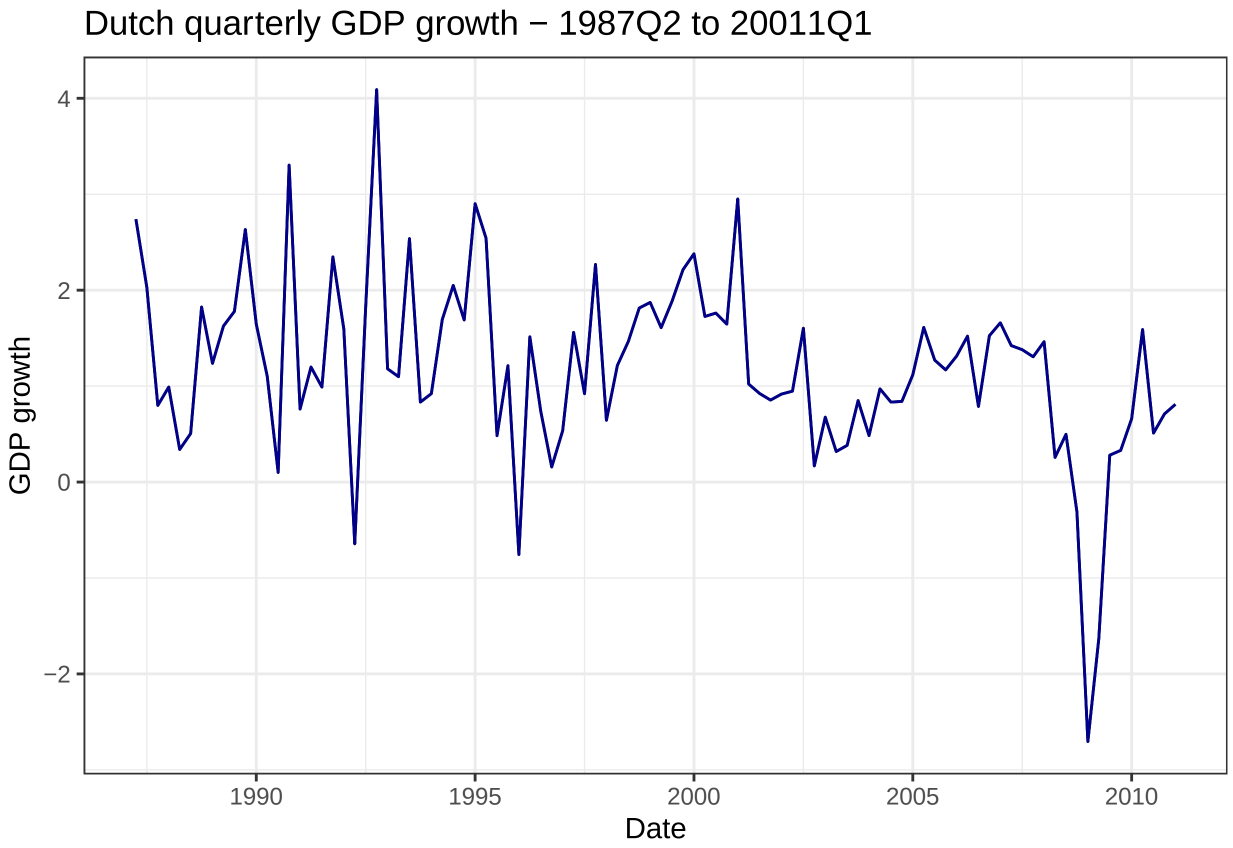
Question 6

Do you find the assumption of iid Gaussian innovations reasonable? How does this affect your answer to the previous question?

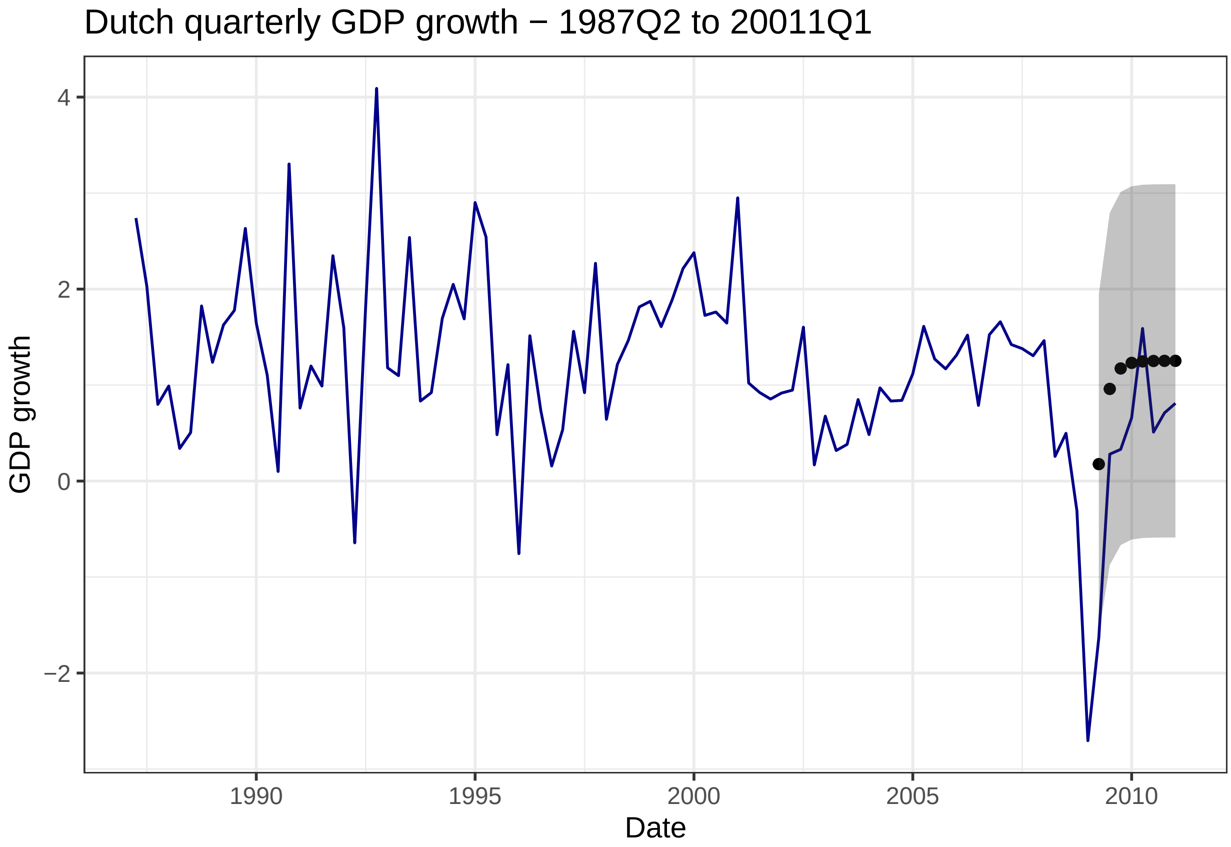
Comment: Given the ACF behaviour and the KS-test performed in the previous items, it seems reasonable to accept the hypothesis of iid Gaussian innovations. However, if this is not true (and could be false, for example, under the presence of heavy tails), the confidence intervals for our predictions would be higher and the uncertainty around the estimate would increase.

Question 7

Suppose that 2 years have passed since you delivered your forecasts to the government, in the first quarter of 2009. Compare your point forecasts and confidence bounds with the following actual observed values for the 12 quarters from 2009q1 to 2011q1. Please comment on the accuracy of your forecasts.



	obs	GDP_QGR	Forecast	L95	H95
23 Q1	2009Q2	-1.63	0.1765306	-1.5942839	1.947345
23 Q2	2009Q3	0.28	0.9600700	-0.8750397	2.795180
23 Q3	2009Q4	0.33	1.1731221	-0.6666522	3.012896
23 Q4	2010Q1	0.66	1.2310530	-0.6090657	3.071172
24 Q1	2010Q2	1.59	1.2468050	-0.5933391	3.086949
24 Q2	2010Q3	0.51	1.2510881	-0.5890579	3.091234
24 Q3	2010Q4	0.71	1.2522527	-0.5878934	3.092399
24 Q4	2011Q1	0.81	1.2525694	-0.5875768	3.092716



Comments: The point forecasts, with exception of 2009 Q2 and 2010 Q2, are above the true realizations. However, the realizations are within the 95% confidence bands, which shows that there is no significant evidence of the forecasts and the true realizations being different. However, this result must be taken with cautious, because as mentioned before, the confidence intervals for the predictions are quite large and range from positive to negative values. Also, if the iid Gaussian assumption for the innovations does not hold, the intervals my lead us to a worng conclusiong about the forecast results credibility.

Question 8

Repeat question 2 above, but this time using a 10% significance level for the general-to-specific modeling approach. Use the newly estimated model to produce a 2-year (8 quarters) point forecast of the Dutch GDP quarterly growth rate. Comment on the accuracy of the forecast generated by the newly estimated AR model. Is it better than the model you estimated before?

Comments: In this case, we would keep the model with three lags, removing the intermediate lag. More specifically, since the coefficient for the 3rd lag is positive and significative, we observe a medium run effect on the current GDP. As for the second lag, it is not significative at a 10% significance level and was removed. The first lag continues significant in this model and could be seen as a short run effect.

Overall, by using a larger significance level in the model we allow the existence of medium and short run effects on the current GDP growth.

```
mAR3                 <- Arima(tsData01, order = c(3,0,0), fixed = c(NA, 0, NA, NA))

AR3coef              <- tidy(coeftest(mAR3), stringsAsFactors = FALSE)
AR3coef              <- cbind(AR3coef[, 1], round(AR3coef[, 2:5], digits = 2))
```

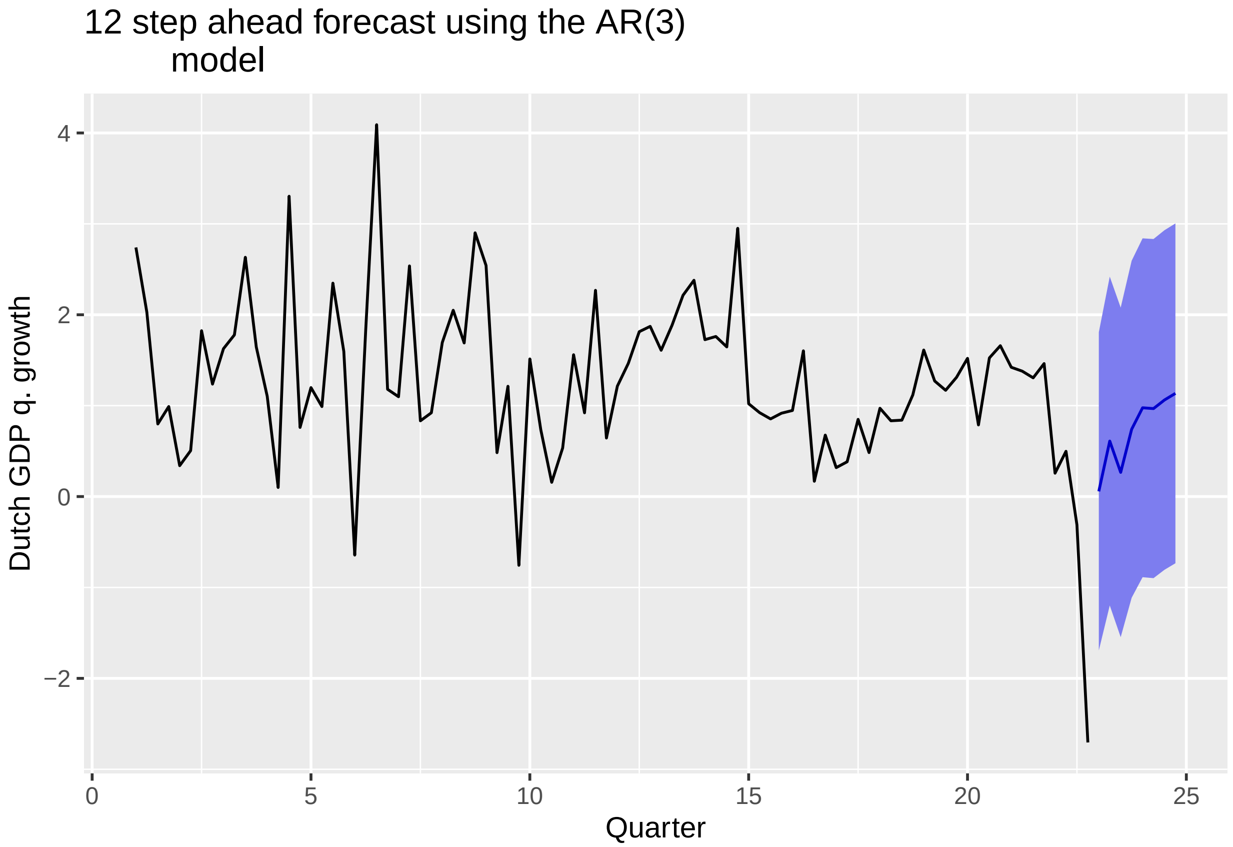
```
  colnames(AR3coef) <- c('Variable', 'Estimate', 'Std. Error', 't-statistic', 'P-Value')
  AR3coef[,1]       <- c("Lag 1", "Lag 3", "Intercept")

  AR3coef %>%
  kable("latex") %>%
  kable_styling(bootstrap_options = c("striped", "hover"))
```

Variable	Estimate	Std. Error	t-statistic	P-Value
Lag 1	0.26	0.12	2.25	0.02
Lag 3	0.20	0.12	1.73	0.08
Intercept	1.23	0.17	7.13	0.00

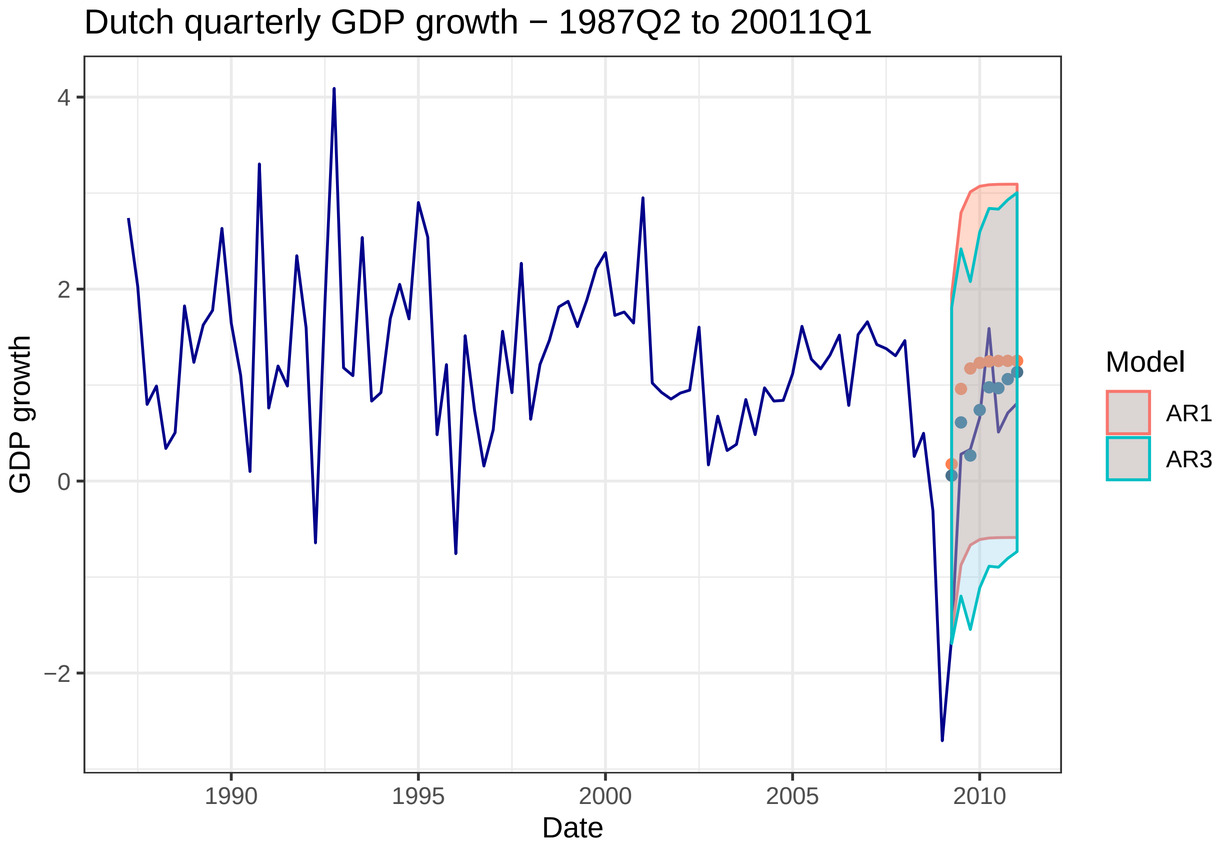
Forecasting using AR(3) model

```
h = 8
AR3forecast  <- forecast(mAR3, h, level = 95)
```

	Date	Forecast	L95	H95
23 Q1	2009Q2	0.0573419	-1.6930458	1.807729
23 Q2	2009Q3	0.6107510	-1.1978975	2.419399
23 Q3	2009Q4	0.2660075	-1.5465162	2.078531
23 Q4	2010Q1	0.7403280	-1.1133001	2.593956
24 Q1	2010Q2	0.9767110	-0.8870371	2.840459
24 Q2	2010Q3	0.9678193	-0.8974229	2.833062
24 Q3	2010Q4	1.0623482	-0.8055024	2.930199
24 Q4	2011Q1	1.1352020	-0.7337886	3.004193



	obs	GDP_QGR	ForecastAR1	L95AR1	H95AR1	ForecastAR3	L95AR3	H95AR3
23 Q1	2009Q2	-1.63	0.1765306	-1.5942839	1.947345	0.0573419	-1.6930458	1.807729
23 Q2	2009Q3	0.28	0.9600700	-0.8750397	2.795180	0.6107510	-1.1978975	2.419399
23 Q3	2009Q4	0.33	1.1731221	-0.6666522	3.012896	0.2660075	-1.5465162	2.078531
23 Q4	2010Q1	0.66	1.2310530	-0.6090657	3.071172	0.7403280	-1.1133001	2.593956
24 Q1	2010Q2	1.59	1.2468050	-0.5933391	3.086949	0.9767110	-0.8870371	2.840459
24 Q2	2010Q3	0.51	1.2510881	-0.5890579	3.091234	0.9678193	-0.8974229	2.833062
24 Q3	2010Q4	0.71	1.2522527	-0.5878934	3.092399	1.0623482	-0.8055024	2.930199
24 Q4	2011Q1	0.81	1.2525694	-0.5875768	3.092716	1.1352020	-0.7337886	3.004193



Comments on the forecast results: As mentioned before, by allowing for further lags, we incorporate a medium run aspect to the model, in addition to the short run component (first lag). As a result, the forecasted points are able to better behave in an ascending trajectory, instead acommodating in a certain average level as the forecasts with the AR1 model. That is, the GDP quartely growth shows a more gradual recovery from the 2008 crisis, which resembles more the observed data from the second quarter of 2009 onwards. In terms of the size of confidence intervals there were no visible changes, as can be seen in the graph.

References

Durbin, James, and Siem Jan Koopman. 2012. *Time Series Analysis by State Space Methods*. Oxford university press.

Hyndman, Rob J, and Yeasmin Khandakar. 2008. "Automatic Time Series Forecasting: The Forecast Package for R." *Journal of Statistical Software* 27 (3).

Nelson, Charles R, and Charles R Plosser. 1982. "Trends and Random Walks in Macroeconmic Time Series: Some Evidence and Implications." *Journal of Monetary Economics* 10 (2): 139–62.

Econometrics III HW - part 2

A. Schmidt and P. Assunção

March, 2020

Assignment 2

See the source code if interested in all functions (chunks were ommited unless relevant for the assignment). Click here to access the code.

Introduction

By the beginning of 2014, in the face of fast rising unemployment, the Dutch Ministry of Social Affairs is concerned that the provisions for future government expenditure with social pensions may be severely underestimated. This is particularly true if the economy is hit again by a large negative shock. Suppose that you have been asked to analyze alternative unemployment scenarios.

Importing and checking data

```r
urlRemote  <- "https://raw.githubusercontent.com/aishameriane"
pathGithub <- "/Mphil/master/EconIII/data_assign_p2.csv"
token      <- "?token=AAVGJTRVJIKVBISR37WEY3K6RIY4E"

url      <- paste0(urlRemote, pathGithub, token)
dfData21 <- read.csv2(url, sep = ",", dec = ".", header = TRUE)
```

As in Part I, we make an initial check on the dataset. The head and tail of the imported dataset indeed show the quarterly data ranging from from the second quarter of 1987 until the first quarter of 2014.

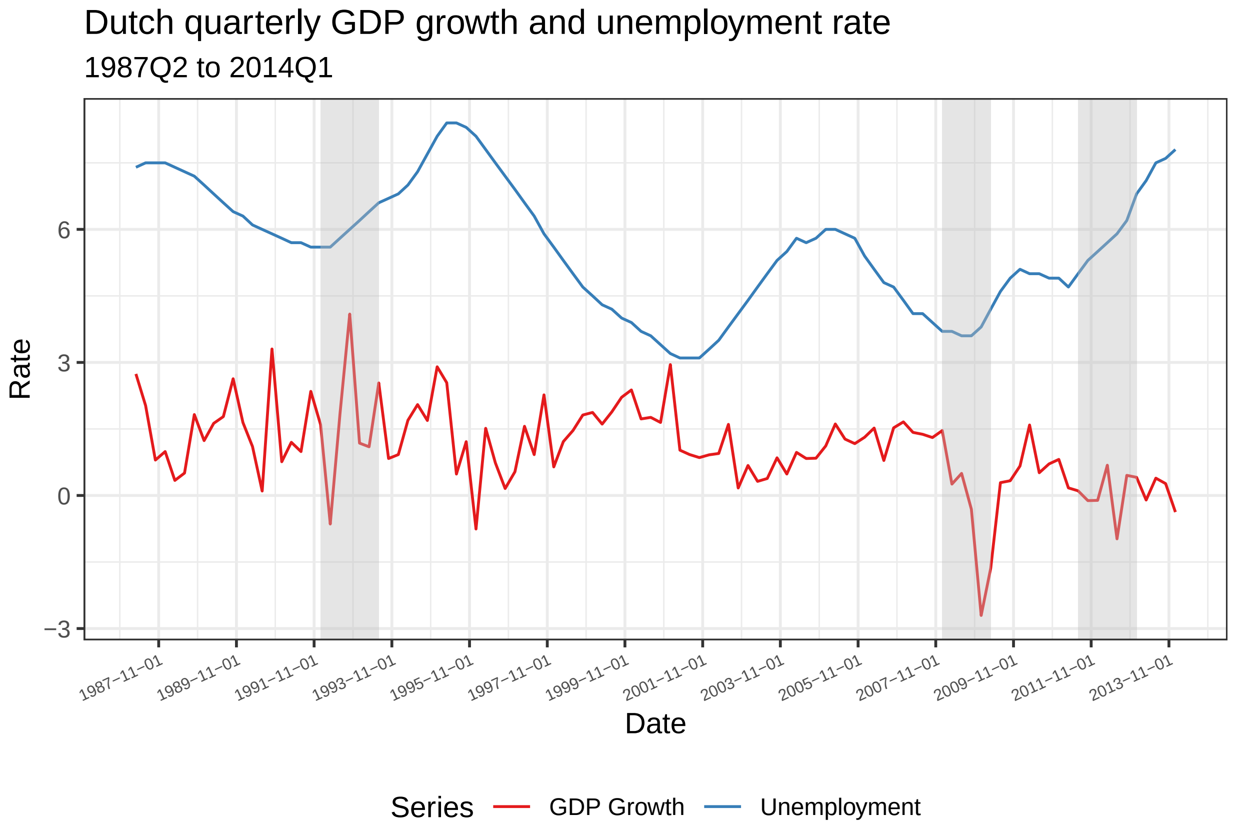
obs	GDP_QGR	UN_RATE	obs	GDP_QGR	UN_RATE
1987Q2	2.7404816	7.4	2012Q4	0.4536453	6.2
1987Q3	2.0275355	7.5	2013Q1	0.4088997	6.8
1987Q4	0.7983716	7.5	2013Q2	-0.1027331	7.1
1988Q1	0.9903613	7.5	2013Q3	0.3907661	7.5
1988Q2	0.3396408	7.4	2013Q4	0.2676572	7.6
1988Q3	0.5052618	7.3	2014Q1	-0.3744039	7.8

As for the descriptive statistics for GDP and unemployment, we see there is no missing information and all values are numeric (there are no problems of formatting). Furthermore, the statistics show values reasonable with the variables pourpose. That is, we do not observe unreasonable high/low levels of GDP growth and unemployment.

	Quart. GDP	Un. Rate
Observations	108.0000	108.0000
Minimum	-2.7051	3.1000
1st quartile	0.4944	4.5750
Mean	1.0670	5.6185
Median	1.0058	5.7000
3rd quartile	1.6310	6.7250
Maximum	4.0908	8.4000
Desv. Pad.	0.9739	1.4074

Question 1

Plot the sample of Dutch quarterly unemployment rates and Dutch GDP quarterly growth rates that you have at your disposal. Estimate an AR model for the GDP growth rate and an ADL model for the unemployment rate using the GDP growth rate as an exogenous explanatory variable. Please adopt a general-to-specific methodology for both models by eliminating insignificant lags. For the AR model start with four lags of GDP. For the ADL model start with four lags of each variable. Report the final estimated AR and ADL models, working at a 5% significance level. Comment on the estimated coefficients. What do these coefficients tell you about the dynamic properties of the unemployment rate and the GDP growth rate?



Comments on the plots: We start by analysing the GDP series. As discussed in Part I, there seems to be some change in the time series behavior after the turn of the millenium, with a large drop during the 2008 financial crisis. Now, with data up to 2014, we observe a recovery of the GDP growth to levels pre crisis around 2 years after the crisis. However, after the 2010s, GDP growth seems to be oscilating around zero (or very low) growth, with a volatily smaller than the one observed in the 90s.

As for the unemployment rate, we observe some sort of cycical behaviour that seems to last for (roughly) 5 years. One exception to this behaviour is around the financial crisis period. Although unemployment was indeed smaller in 2010 than in 2005, one would expect that the continuous drop from 2005 until 2010. That is, that the favorable cenario for umemployment rate, where the economoy is experiencing diminishing unemployment, would last for at least a couple more years. Instead, we observe that such drop occurs up to 2008, where the unemployment trend is reversed and the economy faces more unemployment as the crisis goes on. After the crises, we observe a mostly increasing unemployment trend until the end of the data sample, with a short drop after 2010.

The gray areas in the graph represents the recessions in the Euro Area for the period (link for the source: Euro Area Business Cycle Network: https://eabcn.org/dc/chronology-euro-area-business-cycles). Notice that indeed the Dutch GDP growth rate was in decline during the last two periods flagges as recessions, but it was not case for the first one (from 1992Q1 to 1993Q3), but this is not an error: Netherlands only adopted the Euro in 1999.

PALOMA: nao sei mais o que adicionar aqui, ja que nem é solicitado explicitamente uma análise desses gráficos. *AISHA* Acho que está bom. Eu fiz um gráfico onde as duas séries ficam juntas e adicionei as recessões com uma explicação no final.

AR(p) model for GDP - AR(4)

Comments about the estimation procedure: As in Part I, we opted by a model including the intercept and we used the function `Arima()` from the `forecast` package.

To avoid too many code in the document, only the final model code is appearing below and the remaining is omited — the original code is available in the link provided at the begining of this document.

Variable	Estimate	Std. Error	t-statistic	P-Value
Lag 1	0.33	0.10	3.42	0.00
Lag 2	0.06	0.10	0.56	0.57
Lag 3	0.21	0.10	2.08	0.04
Lag 4	0.05	0.10	0.47	0.64
Intercept	1.06	0.22	4.83	0.00

AR(p) model for GDP - AR(3)

With all lags

Variable	Estimate	Std. Error	t-statistic	P-Value
Lag 1	0.34	0.09	3.61	0.00
Lag 2	0.06	0.10	0.59	0.55
Lag 3	0.22	0.10	2.34	0.02
Intercept	1.07	0.21	5.05	0.00

Removing the second lag

```r
mAR3               <- Arima(tsData21[,1], order = c(3,0,0), fixed = c(NA,0,NA, NA))

AR3coef            <- tidy(coeftest(mAR3), stringsAsFactors = FALSE)
AR3coef            <- cbind(AR3coef[, 1], round(AR3coef[, 2:5], digits = 2))
colnames(AR3coef) <- c('Variable', 'Estimate', 'Std. Error', 't-statistic', 'P-Value')
AR3coef[,1]        <- c("Lag 1", "Lag 3", "Intercept")
```

```r
AR3coef %>%
kable("latex") %>%
kable_styling(bootstrap_options = c("striped", "hover"))
```

Variable	Estimate	Std. Error	t-statistic	P-Value
Lag 1	0.36	0.09	4.05	0.00
Lag 3	0.24	0.09	2.69	0.01
Intercept	1.07	0.20	5.31	0.00

Comments on the results: We started by estimating and AR(4) and we use a significance level of 5% in our analysis. Since the coefficient for the fourth lag is statistically insignificant, we droped the fourth lag and re-estimated the model as an AR(3). From the results of this second model, the coefficient for the fourth lag is significant. However, the second lag p-value shows evidence of non-sisgnificance. Therefore, we reestimated the AR(3) with a zero coefficient for the second lag and obtained the final model:

$$GDP\_QGR_t = 1.07 + 0.36GDP\_QGR_{t-1} + 0.24GDP\_QGR_{t-3} + \epsilon_t,$$
$$\text{with} \quad \{\epsilon_t\} \sim WN(0, \sigma_\epsilon^2).$$

According to this model, we have a persistence of the past GDP growth levels on the current growth. Also, since the coefficients are positive, we observe a "positive persistence". That is, past GDP growth tends to influence growth today in the same direction: growth would be related to some more growth in the future, and decline would be related with more future decline. Such persistence would last for about three quarters.

This lag persistence is consistent with the definition from the Euro Area Business Cycle Dating Committee (EABCDC) for a recession. Accordingly to the EABCDC, a recession is characterized as *"a significant decline in the level of economic activity, spread across the economy of the euro area, usually visible in* ***two or more*** *consecutive quarters of negative growth in GDP, employment and other measures of aggregate economic activity for the euro area as a whole."* (link for the source: https://eabcn.org/dc/methodology).

ADL(p,q) model for unemployment (GDP growth as exogenous regressor) - ADL(4,4)

Comments about the estimation procedure:

To estimate the model, we used the function `dynlm()` from the package `dynlm`. This is a wrapper for the class of `lm` functions in R, that offers some advantages by allowing to easily add lags from the series (by using the operator `L`). The same result would have been achieved by using the `lag()` function, but without the convenience of being able to preserve the objects as time series objects. If the model had seasonal or other components, it would be more easy to incorporate as well. The estimation procedure is the same as the one used in `lm()`, which can be consulted here.

Now, we proceed with the estimation of an ADL model for quarterly unemployment rate, using quarterly GDP growth as regressor. At first, we attempted to follow the general-to-specific method for selecting the ADL model with best fit. However, by sequentially removing insignificant lags (respecting a chronological ordering), we ended up with a model with three lags for unemployment an no lags and no current value of GDP. Since we are interesting in investigating some dynamics between unemployment and economic activity, we opted for alternative ways of selecting the best fit. The procedure we opted for is to eliminate the most insignificant lags (now we are not following a chronological ordering).

Sequentially, those were the models estimated and the lags removed (based on the p-values):

```
* We started with an ADL(4,4) model and realized that the 4th lag of GDP growth is
the one witht the highest p-value. Thus, this lag was removed.
```

```
* Next, we estimated an ADL(4,3) and decided to eliminate the 2nd lag of GDP
growth, and the the current GDP growth.
* After estimating this model, we decided to eliminate the fourth lag of unemployment.
By doing this, we were left with an ADL(3,3) with zero coefficient for the coefficient
of current and second lag of GDP growth.
* Finally, we proceed by eliminating the third lag of GDP growth and, since we are
considering a significance level of 5%, we also eliminate the second lag of unemployment rate.
```

As before, to avoid too many code in the document, only the final model code is appearing below and the remaining is omited — the original code is available in the link provided at the begining of this document.

Variable	Estimate	Std. Error	t-statistic	P-Value
Intercept	0.17	0.06	2.90	0.00
UN_RATE(-1)	1.60	0.10	15.57	0.00
UN_RATE(-2)	-0.30	0.19	-1.60	0.11
UN_RATE(-3)	-0.44	0.19	-2.24	0.03
UN_RATE(-4)	0.11	0.11	1.05	0.29
GDP_QGR	-0.01	0.01	-0.94	0.35
GDP_QGR(-1)	-0.02	0.02	-1.33	0.19
GDP_QGR(-2)	-0.01	0.02	-0.56	0.58
GDP_QGR(-3)	0.02	0.02	1.34	0.18
GDP_QGR(-4)	-0.01	0.01	-0.47	0.64

ADL(4,3)

Variable	Estimate	Std. Error	t-statistic	P-Value
Intercept	0.16	0.06	2.89	0.00
UN_RATE(-1)	1.60	0.10	15.69	0.00
UN_RATE(-2)	-0.29	0.19	-1.55	0.12
UN_RATE(-3)	-0.44	0.19	-2.26	0.03
UN_RATE(-4)	0.11	0.10	1.02	0.31
GDP_QGR	-0.01	0.01	-0.95	0.35
GDP_QGR(-1)	-0.02	0.01	-1.45	0.15
GDP_QGR(-2)	-0.01	0.02	-0.59	0.56
GDP_QGR(-3)	0.02	0.01	1.26	0.21

ADL(4,3) with zero coefficient for the second lag of GDP growth

Variable	Estimate	Std. Error	t-statistic P-Value	NA
Intercept	0.16	0.06	2.85	0.01
UN_RATE(-1)	1.61	0.10	16.01	0.00
UN_RATE(-2)	-0.30	0.19	-1.61	0.11
UN_RATE(-3)	-0.44	0.19	-2.28	0.02
UN_RATE(-4)	0.11	0.10	1.06	0.29
GDP_QGR	-0.01	0.01	-1.00	0.32
GDP_QGR(-1)	-0.02	0.01	-1.68	0.10
GDP_QGR(-3)	0.02	0.01	1.15	0.25

ADL(4,3) with zero coefficient for the current and second lag of GDP growth

Variable	Estimate	Std. Error	t-statistic	P-Value
Intercept	0.1621085	0.0562086	2.8840491	0.0048366
UN_RATE(-1)	1.5991858	0.1000723	15.9803049	0.0000000
UN_RATE(-2)	-0.3015861	0.1872108	-1.6109438	0.1104414
UN_RATE(-3)	-0.4120201	0.1906222	-2.1614486	0.0331218
UN_RATE(-4)	0.0885453	0.1017021	0.8706338	0.3861037
GDP_QGR(-1)	-0.0287103	0.0133990	-2.1427213	0.0346380
GDP_QGR(-3)	0.0127485	0.0137022	0.9304018	0.3544728

ADL(3,3) with zero coefficient for the current and second lag of GDP growth

Variable	Estimate	Std. Error	t-statistic	P-Value
Intercept	0.18	0.05	3.33	0.00
UN_RATE(-1)	1.58	0.10	16.35	0.00
UN_RATE(-2)	-0.34	0.18	-1.86	0.07
UN_RATE(-3)	-0.27	0.10	-2.80	0.01
GDP_QGR(-1)	-0.03	0.01	-2.12	0.04
GDP_QGR(-3)	0.01	0.01	0.90	0.37

ADL(3,1) with coefficient zero for current GDP

Variable	Estimate	Std. Error	t-statistic	P-Value
Intercept	0.18	0.05	3.54	0.00
UN_RATE(-1)	1.57	0.10	16.34	0.00
UN_RATE(-2)	-0.34	0.18	-1.88	0.06
UN_RATE(-3)	-0.26	0.10	-2.73	0.01
GDP_QGR(-1)	-0.03	0.01	-1.98	0.05

ADL(3,1) with coefficient zero for current GDP and second lag of unemployment

```r
mADL31_0 <- dynlm(ts(tsData21[,2]) ~ L(ts(tsData21[,2])) +L(ts(tsData21[,2]),3)
                  + L(ts(tsData21[,1])))

ADL31_0coef           <- tidy(coeftest(mADL31_0), stringsAsFactors = FALSE)
ADL31_0coef           <- cbind(ADL31_0coef[, 1], round(ADL31_0coef[, 2:5], digits = 2))
colnames(ADL31_0coef) <- c('Variable', 'Estimate', 'Std. Error', 't-statistic',
                           'P-Value')
ADL31_0coef[,1]       <- c("Intercept", "UN_RATE(-1)", "UN_RATE(-3)", "GDP_QGR(-1)")

ADL31_0coef %>%
kable("latex") %>%
kable_styling(bootstrap_options = c("striped", "hover"))
```

Variable	Estimate	Std. Error	t-statistic	P-Value
Intercept	0.21	0.05	4.04	0.00
UN_RATE(-1)	1.40	0.03	48.71	0.00
UN_RATE(-3)	-0.43	0.03	-14.88	0.00
GDP_QGR(-1)	-0.03	0.01	-2.07	0.04

Comments on the results: The resulting model from this procedure for model selection is an ADL(3,1) with no secong lag for unemployment and no current GDP growth:

$$UN\_RATE_t = 0.21 + 1.40UN\_RATE_{t-1} - 0.43UN\_RATE_{t-3} - 0.03GDP\_QGR_{t-1} + \varepsilon_t,$$
$$\text{with} \quad \{\varepsilon_t\} \sim WN(0, \sigma_\varepsilon^2).$$

From the coefficients, we note a very small, but signifficant, negative effect of previous period GDP growth on current unemployment. This effect makes sense since it is expected that, as the economic activity declines, employers would find it more difficult to keep employing the same amount of workers. There is also a second effect, which is more observed in recessions: during uncertain times, people who otherwise would be outside the labor force (for example studying or as domestic partner), starts to look for jobs, increasing the unemployment rate.

As for the coefficients of unemployment lags, they show a positive effect of the unemployment rate from the period immediately before, but a negative effect of unemployment three periods before. This means that the trend for unemployment will most likely persist until next period, but then it will start changing a few quarters from now. This behavior is somewhat consistent with the cycles observed in the graph analysis.

Question 2

Use the estimated ADL model to calculate and interpret the short-run multiplier, the 2-step-ahead multiplier, and the long-run multiplier. Finally, report and interpret the long-run relation between the unemployment rate and the GDP growth rate.

Let the unemployment rate be u_t, the gdp quartely growth be gdp_t. The final model will be given by:

$$u_t = \alpha + \phi_1 u_{t-1} + \phi_2 u_{t-3} + \beta gdp_{t-1} + \varepsilon_t$$
$$\{\varepsilon_t\} \sim WN(0, \sigma_\varepsilon^2)$$

The estimated model will be:

$$u_t = 0.21 + 1.40u_{t-1} - 0.43u_{t-3} - 0.03gdp_{t-1} + \varepsilon_t$$
$$\{\varepsilon_t\} \sim WN(0, \sigma_\varepsilon^2)$$

The short-run multiplier tells us how the current value of the exogenous variable (quarterly GDP growth in our case) explains the endogenous variable (quarterly unemployment). Since we do not have gdp_t as an exogenous regressor in the final model, the short-run multiplier will be zero (percentual increases over GDP lead to no changes in current unemployment. The derivation is as follows:

$$u_t^* = \alpha + \phi_1 u_{t-1} + \phi_2 u_{t-3} + \beta gdp_{t-1} + \varepsilon_t = u_t$$

The 2-step-ahead multiplier will give us the impact on unemployment two-step ahead of a one unit temporary increase in current GDP growth. This multiplier will be equal to $\phi_1\beta = 1.4 \cdot (-0.03) = -0.042$. That is, a percentage increase in GDP is expected to result in a fall of 0.042% in unemployment. The derivation is as follows:

$$\begin{aligned}
u^*_{t+1} &= \alpha + \phi_1 u_t + \phi_2 u_{t-2} + \beta(gdp_t + 1) + \varepsilon_{t+1} \\
&= \alpha + \phi_1 u_t + \phi_2 u_{t-2} + \beta gdp_t + \beta + \varepsilon_{t+1} \\
&= u_{t+1} + \beta \\
u^*_{t+2} &= \alpha + \phi_1 u_{t+1} + \phi_2 u_{t-1} + \beta gdp_{t+1} + \varepsilon_{t+2} \\
&= \alpha + \phi_1 (u_{t+1} + \beta) + \phi_2 u_{t-1} + \beta gdp_{t+1} + \varepsilon_{t+2} \\
&= u_{t+2} + \phi_1 \beta
\end{aligned}$$

The long-run multiplier will give us the effect over unemployment of a permanent change in GDP growth, and it is obtained from the long run relationship between these two variables. Such relatioship is obtained from the fact that, in the long-run, variables should be equal (or very close to) their historical averages. Thus, the long-run relatioship between unemployment and GDP growth will be:

$$\begin{aligned}
\overline{u} &= \alpha + \phi_1 \overline{u} + \phi_2 \overline{u} + \beta \overline{gdp} \\
\overline{u} &= \frac{\alpha}{1 - \phi_1 - \phi_2} + \frac{\beta}{1 - \phi_1 - \phi_2} \overline{gdp}.
\end{aligned}$$

Thus, the long-run multiplier will be $\frac{\beta}{1-\phi_1-\phi_2} = \frac{-0.03}{1-1.40+0.43} = -1$: 1% increase in GDP is expected to result in 1% fall on unemployment.

Question 3

Please provide a detailed comment on the following statement: "An increase in the GDP growth rate causes a reduction in the unemployment rate."

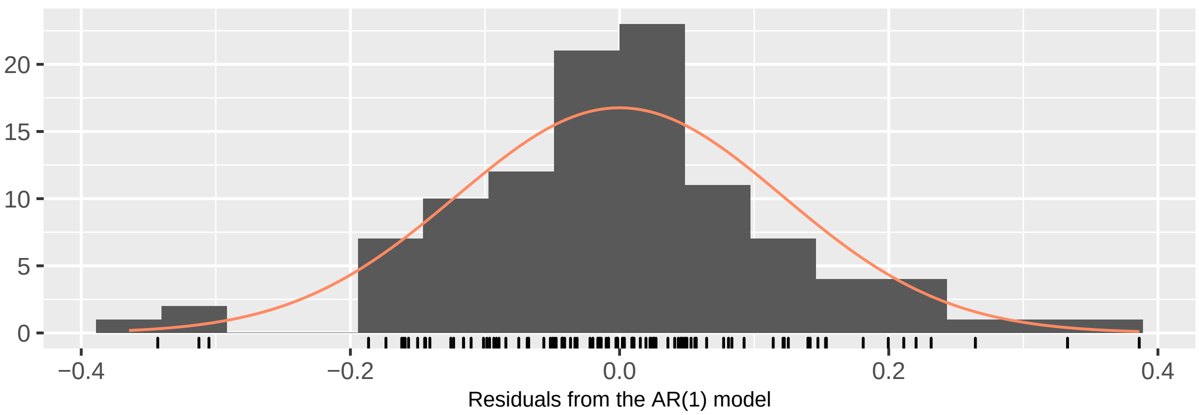
From the multipliers in the previous question, we can breakdown the effects of a change in GDP growth over unemployment rate by analyzing the multipliers for the short and long run. Since the short-run multiplier β_0 (the coefficient of GPD at time t in the equation of unemployment) is zero, there is no immediate effect of the GDP growth rate over unemployment.

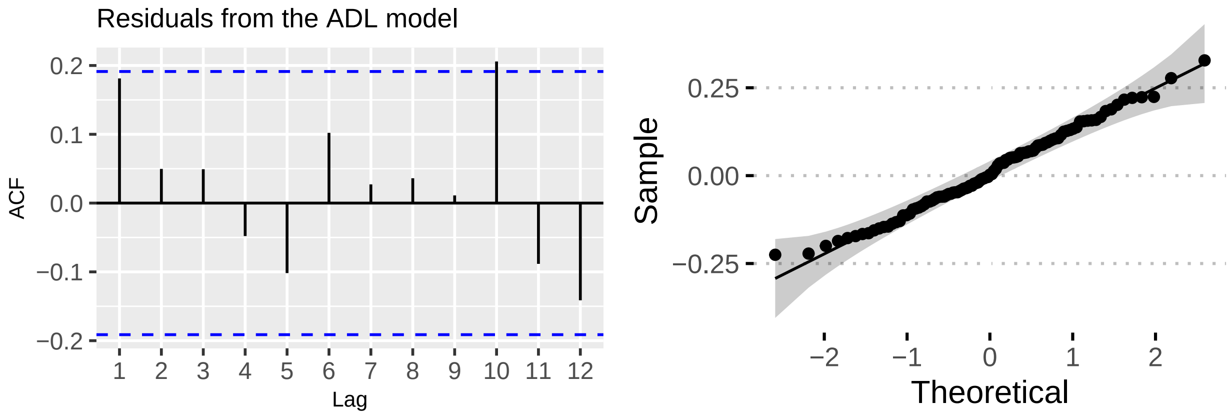
As time goes by, the effect of GPD over unemployment tends to get stronger and more negative. In the previous question, we found that there is no immediate effect of the GDP growth rate over unemployment (zero short-run multiplier). As part of the solution for the 2-step multiplier, however, we found the 1-step multiplier, which equals $\beta_1 = -0.03$. Thus, we have that changes in GDP start to have an effect over unemployment only after 1 period. The 2-step ahead multiplier, which is given by $\phi\beta_1$ = -0.042, tells us that the effect of a positive GDP shock today will lead to even more decrease in unemployment two steps ahead, which indicates that the effect tends to become stronger (i.e. more negative) as time goes by. That is, there is some sort of persistence of temporary GDP shocks over unemployment rate that would lead to smaller and smaller unemployment. This result extends to the effects of a permanent shock on GDP growth, which according to the long-run multiplier, is of a proportional decrease in unemployment.

Note that these numbers tell us something about how the dynamics of GDP growth would affect unemployment. This is not a causal analysis. Basically, what we found is that, for example, increasing GDP growth tends to be associated with decreasing unemployment, but we cannot tell yet whether increasing GDP causes decreasing unemployment. To be able to draw conclusions about the causality, one needs to make a deeper investigation using not only statistical methods but also more economic theory to sustain these conclusions. In the latter, one could discuss the relationship between Okun's Law and the model results.

Question 4

Suppose that the innovations are iid Gaussian. What is the probability of the unemployment rate rising above 7.8% in the second quarter of 2014? What is the probability that it drops below 7.8%? Do you trust the iid Gaussian assumption?





To compute these probabilities, we first find the one step ahead estimator for UN_RATE. Again, let the unemployment rate be u_t, the gdp quartely growth be gdp_t and the one step ahead forecast be $\hat{u}_{T+1}$. Since we are assuming $\{\varepsilon_t\} \sim iid(0, \sigma_\varepsilon^2)$, we have that:

$$\begin{aligned}
\hat{u}_{T+1} &= E[u_{T+1}|D_T = (u_1, \ldots, u_T, gdp_1, \ldots, gdp_T)] \\
&= E[\alpha + \phi_1 u_T + \phi_2 u_{T-2} + \beta gdp_T + \varepsilon_{T+1}|D_T] \\
&= \alpha + \phi_1 u_T + \phi_2 u_{T-2} + \beta gdp_T + E[\varepsilon_{T+1}|D_T] \\
&= \alpha + \phi_1 u_T + \phi_2 u_{T-2} + \beta gdp_T \\
&= u_{T+1} - \varepsilon_{T+1}
\end{aligned}$$

The last step follows from the innovations being iid Gaussian distributed.

The forecast error term will be given by

$$e_{T+1} = u_{t+1} - \hat{u}_{T+1} = \varepsilon_{T+1}.$$

Therefore, we have that $e_T \sim NID(0, \sigma_\varepsilon^2)$.

From this result, it follows that the probability that unemployment in the second quarter of 2014 will be

$$\begin{aligned}
P(u_{T+1} > 7.8) &= P(u_{T+1} - \hat{u}_{T+1} > 7.8 - \hat{u}_{T+1}) \\
&= P(e_{T+1} > 7.8 - \hat{u}_{T+1}) \\
&= P\left(\frac{e_{T+1}}{\sigma_\varepsilon} > \frac{7.8 - \hat{u}_{T+1}}{\sigma_\varepsilon}\right)
\end{aligned}$$

From the historical values and model estimates for the parameters and the residual standard error (0.1233), we get:

$$\begin{aligned}\hat{u}_{T+1} &= 0.21 + 1.40 \cdot 7.8 - 0.43 \cdot 7.5 - 0.03 \cdot (-0.37) = 7.9161 \\ \hat{\sigma}_{\varepsilon} &= 0.1233\end{aligned}$$

Thus, the probability that the unemployment exeeceds 7.8% in the second quarter of 2014 is

$$P(u_{T+1} > 7.8) = P\left(\frac{e_{T+1}}{\hat{\sigma}_{\varepsilon}} > \frac{7.8 - 7.9161}{0.1233}\right) = P\left(\frac{e_{T+1}}{\hat{\sigma}_{\varepsilon}} > -0.94\right) = 0.8264 = 82.64\%$$

and the probability that unemployment will be smaller than 7.8% is $1 - 0.8264 = 17.36\%$.

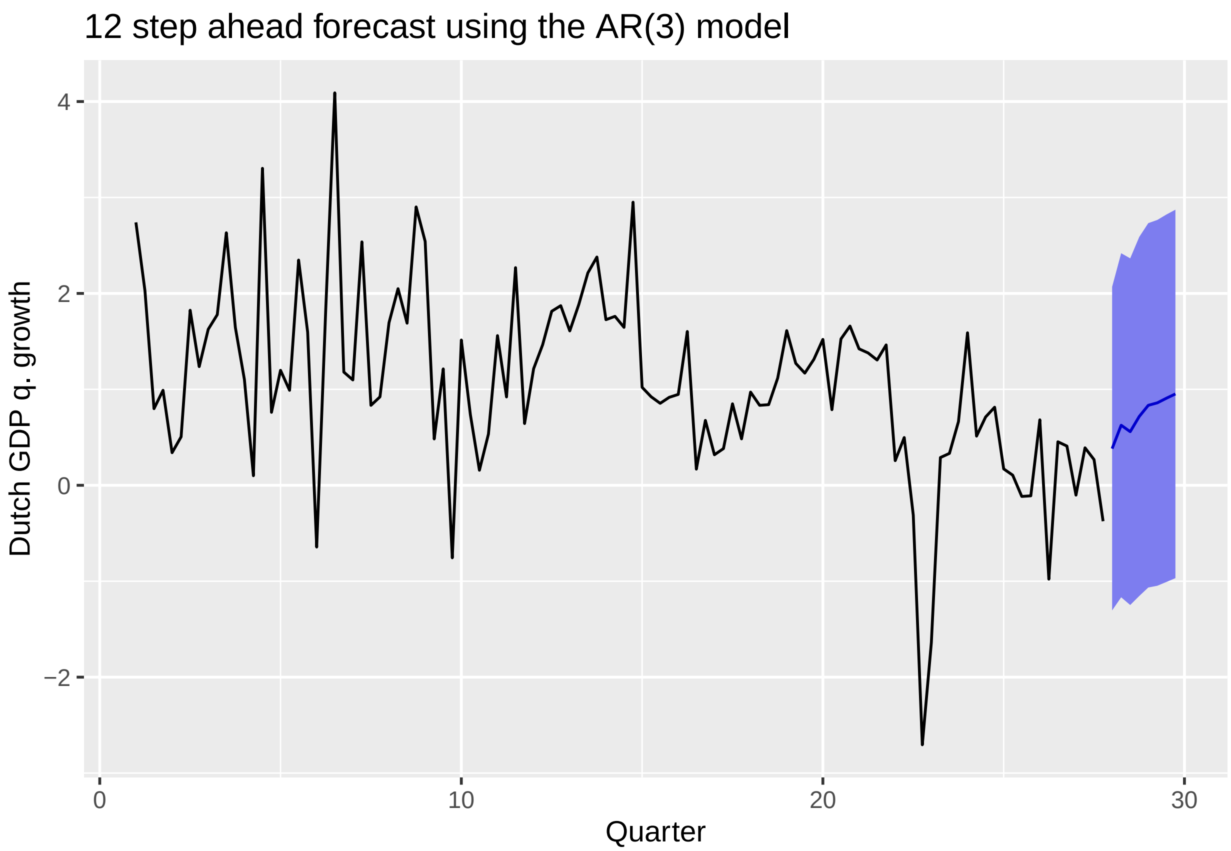
As for the last question, we look at the histogram and the p-value of Kolmogorov-Smirnov test. The null hypothesis of such test is that the data comes from a normal distribution, with mean 0 and standard deviation 0 (as remark, the mean of the residuals is virtually zero, as we would expect). Since the p-value for the KS test is higher than 5 (KS statistic = 0.103; p-value = 0.2154), we have evidence to not reject the null hypothesis. That is, we have evidence that the assumption of iid Gaussian distrubution for the innovations does seem reasonable. From the histogram of the residuals, we see that their distribution indeed seems to be bell shaped, but with somewhat heavy tails, speccialy on the left of the mean, but this was not significant in terms of the test and also doesn't show in the qqplot.

Question 5

Make use of your estimated AR and ADL models to produce a 2-year (8 quarter) forecast for the Unemployment rate that spans until the first quarter of 2016. Report the obtained values.

First, we estimate values for quartely GDP growth. We used the same procedure as the one used for the forecasts in part 1, the `forecast()` function from the package `forecast`.

```r
h = 8
AR3forecast  <- forecast::forecast(mAR3, h, level = 95)
vdNewDate    <- c("2014Q2", "2014Q3", "2014Q4", "2015Q1", "2015Q2", "2015Q3", "2015Q4",
                  "2016Q1")
AR3forecast  <- cbind(vdNewDate, data.frame(AR3forecast))

colnames(AR3forecast) <- c("Date", "Forecast", "L95", "H95")

autoplot(forecast::forecast(mAR3, h, level = 95),
         main = "12 step ahead forecast using the AR(3) model",
         xlab = "Quarter", ylab = "Dutch GDP q. growth")
```



```r
AR3forecast %>%
  kable("latex") %>%
  kable_styling(bootstrap_options = c("striped", "hover"))
```

	Date	Forecast	L95	H95
28 Q1	2014Q2	0.3820861	-1.3043609	2.068533
28 Q2	2014Q3	0.6258468	-1.1674324	2.419126
28 Q3	2014Q4	0.5588832	-1.2478931	2.365659
28 Q4	2015Q1	0.7174058	-1.1538686	2.588680
29 Q1	2015Q2	0.8335985	-1.0654103	2.732607
29 Q2	2015Q3	0.8594311	-1.0471856	2.766048
29 Q3	2015Q4	0.9070621	-1.0079544	2.822079
29 Q4	2016Q1	0.9523491	-0.9679906	2.872689

From the forecasted results, we see that the model expects, on average, a gradual increase in GDP growth for the next 2 years. However, it is important to notice that the 95% confidence region goes below the last observed point, covering negative values for the forecasted results.

Now, we use these forecasted values for GDP growth in the unemployment forecast. We used the approach to compute the forecasts “step-by-step”, i.e., we used the estimated model from question (1) to predict what will be the Unemployment for 2014 Q2 using the forecasted value for the GPD. Next, we use this forecasted Unemployment to compute a forecast for 2014 Q3 and so on until 2016 Q1. We couldn’t find any reliable source on how to compute the confidence intervals – more specifically, given that we have the uncertainty from the forecasts for GDP, we were unsure how to compute the confidence bands for this model. The next chunk of code contains the steps used to assemble the data and make the forecast.

```
# Setting up the data frame with GDP growth forecasts
vdNewDate <- c("2014Q2", "2014Q3", "2014Q4", "2015Q1",
               "2015Q2", "2015Q3", "2015Q4", "2016Q1")

dfNewData <- data.frame(vdNewDate, AR3forecast[,2])
 names(dfNewData) <- names(dfData21[,1:2])
 dfData22 <- rbind(dfData21[,1:2], dfNewData)
 dfData22 <- merge(dfData22, dfData21[,3], by = "row.names", all = TRUE)

 # This will result in two columns, where the last 8 values
 # for Unemployment are NA.
 tsData22 <- xts(dfData22[,3:4], order.by = as.yearqtr(dfData22[,2]), frequency = 4)
 names(tsData22) <- c("GDP_QGR", "UN_RATE")

# We are doing the forecast step by step
# Recovers the index for the row where the first forecast of GDP is located
indexts          <- which(index(tsData22) == "2014 Q2")
ADL22_forecast   <- tsData22

for (i in indexts:nrow(tsData22)){
  ADL22_forecast[i,2] <- coef(mADL31_0) %*% c(1, ADL22_forecast[(i-1), 2],
                                              ADL22_forecast[(i-3), 2],
                                              ADL22_forecast[(i-1), 1])
}
```

The last two observed periods (2013 Q3 and 2014 Q1) as well the 8 forecast results are displayed in the table below.

	GDP_QGR	UN_RATE
2013 Q4	0.2676572	7.600000
2014 Q1	-0.3744039	7.800000
2014 Q2	0.3820861	7.899095
2014 Q3	0.6258468	7.974145
2014 Q4	0.5588832	7.986278
2015 Q1	0.7174058	7.962313
2015 Q2	0.8335985	7.892082
2015 Q3	0.8594311	7.785393
2015 Q4	0.9070621	7.645696
2016 Q1	0.9523491	7.479175

Question 6

Use impulse response functions (IRFs) to analyze two different scenarios for the Dutch un-employment rate:

(a) In the 'good scenario' the GDP quarterly growth rate is hit by a positive shock of 2%.

(b) In the 'bad scenario' the GDP quarterly growth rate suffers a negative shock of 2%.

Please use the last observed value of the unemployment rate and gdp growth rate as the origin of your IRFs. In particular, set the origin to -0.37% for GDP and 7.8% for the unemployment rate.

We start by computing the IRF for the AR(3) model. In general terms, when we have the following model

$$X_t = \alpha + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \phi_3 X_{t-3} + \varepsilon_t, \text{ with } \varepsilon_t \sim \text{i.i.d.}(0, \sigma^2),$$

to evaluate the IRF, we are interested in determine the above system:

$$\begin{aligned}
\tilde{x}_t &= x, \quad t < s \\
\tilde{x}_t &= x + \epsilon, \quad t = s \\
\tilde{x}_t &= x + \frac{\partial X_t}{\partial \varepsilon_s}\epsilon, \quad \forall t > s
\end{aligned}$$

Therefore, are interested in knowing $\frac{\partial X_t}{\partial \varepsilon_s}$.

We know that for an AR(3) model, we have the following:

$$\begin{aligned}
\frac{\partial X_{s+1}}{\partial \varepsilon_s} &= \phi_1 \frac{\partial X_s}{\partial \varepsilon_s} \\
\frac{\partial X_{s+2}}{\partial \varepsilon_s} &= \phi_1 \frac{\partial X_{s+1}}{\partial \varepsilon_s} + \phi_2 \frac{\partial X_s}{\partial \varepsilon_s} = (\phi_1^2 + \phi_2)\frac{\partial X_s}{\partial \varepsilon_s} \\
\frac{\partial X_{s+3}}{\partial \varepsilon_s} &= \phi_1 \frac{\partial X_{s+2}}{\partial \varepsilon_s} + \phi_2 \frac{\partial X_{s+1}}{\partial \varepsilon_s} + \phi_3 \frac{\partial X_s}{\partial \varepsilon_s} = (\phi_1^3 + 2\phi_1\phi_2 + \phi_3)\frac{\partial X_s}{\partial \varepsilon_s} \\
\frac{\partial X_{s+4}}{\partial \varepsilon_s} &= \phi_1 \frac{\partial X_{s+3}}{\partial \varepsilon_s} + \phi_2 \frac{\partial X_{s+2}}{\partial \varepsilon_s} + \phi_3 \frac{\partial X_{s+1}}{\partial \varepsilon_s} = (\phi_1^4 + \phi_2^2 + 3\phi_1^2\phi_2 + 2\phi_1\phi_3)\frac{\partial X_s}{\partial \varepsilon_s} \\
\frac{\partial X_{s+5}}{\partial \varepsilon_s} &= \phi_1 \frac{\partial X_{s+4}}{\partial \varepsilon_s} + \phi_2 \frac{\partial X_{s+3}}{\partial \varepsilon_s} + \phi_3 \frac{\partial X_{s+2}}{\partial \varepsilon_s}, \\
&\text{etc}
\end{aligned}$$

To implement this, we only need to compute separately the first three components. From there, we can use a recursion to obtain a vector with $\frac{\partial X_{s+i}}{\partial \varepsilon_s}$, $\forall\ i > 3$.

Now that we have a recursion for the AR part, we need to write the recursion for the ADL part. Using a general structure for an arbitrary ADL(3,1)-AR(3) system given by

$$\begin{aligned}
Y_t &= \gamma + \psi_1 Y_{t-1} + \psi_2 Y_{t-2} + \psi_3 Y_{t-3} + \beta X_{t-1} + \varepsilon_t, \\
X_t &= \alpha + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \phi_3 X_{t-3} + v_t
\end{aligned}$$

we have the same equations above for the IRF of the x (which represents the AR(3) part) plus the following:

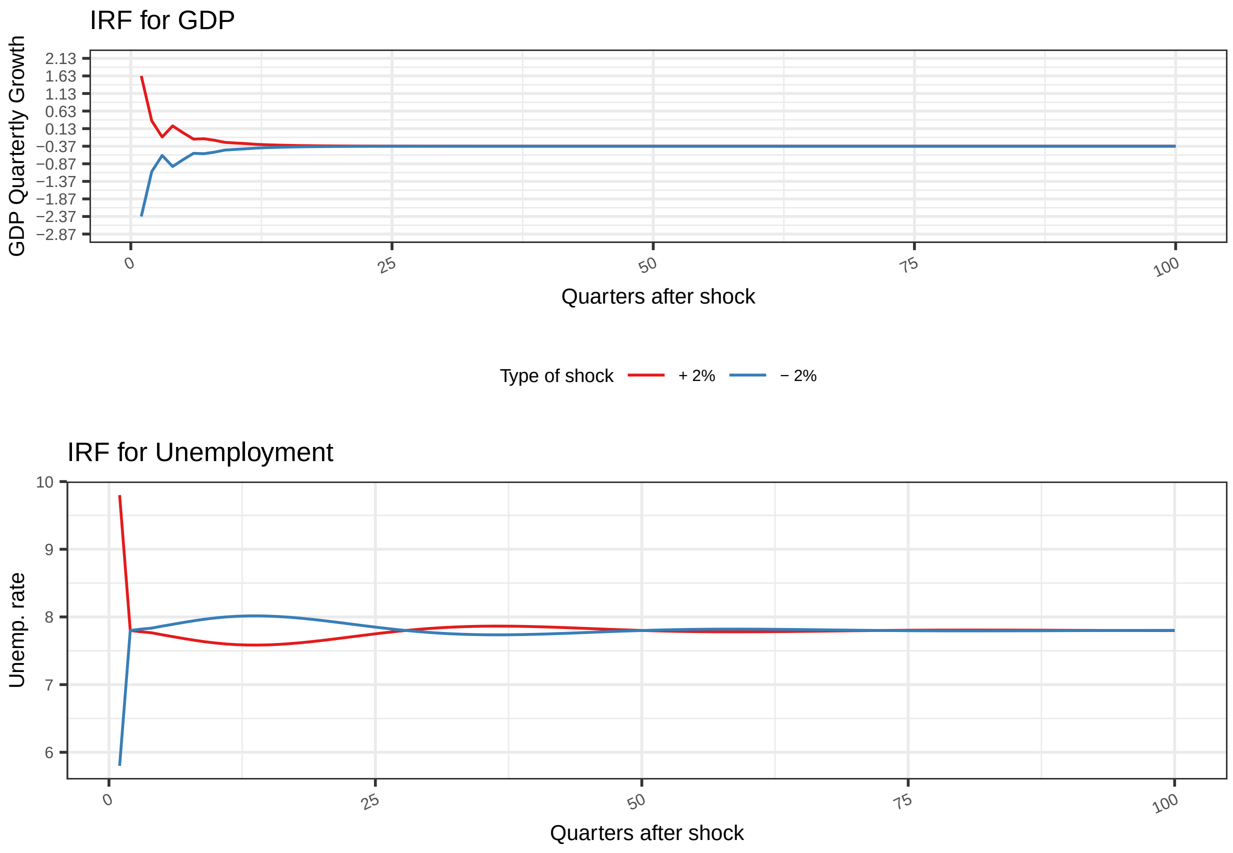
$$\begin{aligned}
\tilde{y}_{s-1} &= y, \\
\tilde{y}_s &= y + 0, \\
\tilde{y}_{s+1} &= y + \beta \frac{\partial X_s}{\partial v_s} \\
\tilde{y}_{s+2} &= y + \psi_1 \frac{\partial Y_{s+1}}{\partial v_s} + \beta \frac{\partial X_{s+1}}{\partial v_s} \\
\tilde{y}_{s+3} &= y + \psi_1 \frac{\partial Y_{s+2}}{\partial v_s} + \psi_2 \frac{\partial Y_{s+1}}{\partial v_s} + \beta \frac{\partial X_{s+2}}{\partial v_s} \\
\tilde{y}_{s+j} &= y + \psi_1 \frac{\partial Y_{s+j-1}}{\partial v_s} + \psi_2 \frac{\partial Y_{s+j-2}}{\partial v_s} + \psi_3 \frac{\partial Y_{s+j-3}}{\partial v_s} + \beta \frac{\partial X_{s+j-1}}{\partial v_s}, \ j > 3
\end{aligned}$$

Given this, we can write a function that will do the recursions:

```r
# The function below computes the IRF for an ADL(3,1)-AR(3) system
fIRFADL31 <- function(param, shock = 0.02*100, h = 20,
                      lastgdp = -0.0037*100, lastun = 0.078*100){

  # To avoid errors
  if (length(param) != 7){
    param <- rep(0,7)
  }

  # To look pretty (because the first value is at the moment of the shock)
  h <- h -1

  # Extract parameters
  phi1 <- param[1]
  phi2 <- param[2]
  phi3 <- param[3]

  psi1 <- param[4]
  psi2 <- param[5]
  psi3 <- param[6]

  beta <- param[7]

  # Create empty vectors
  derivativesx <- c(shock, rep(NA, h))
  derivativesy <- c(shock, rep(NA, h))

  # Make the recursion
  for (i in 1:h){
    if (i == 1){
      derivativesx[i+1] <- phi1 * shock
      derivativesy[i+1] <- 0
    } else if (i == 2) {
      derivativesx[i+1] <- phi1 * derivativesx[i] + phi2 * shock
      #derivativesy[i+1] <- psi1 * derivativesy[i] * shock + beta * derivativesx[i] * shock
      derivativesy[i+1] <- psi1 * derivativesy[i] + beta * derivativesx[i]
    } else if (i == 3){
      derivativesx[i+1] <- phi1 * derivativesx[i] + phi2 * derivativesx[i-1] + phi3 * shock
      #derivativesy[i+1] <- psi1 * derivativesy[i] * shock + psi2 * derivativesy[i-1] * shock
      #+ beta * derivativesx[i] * shock
      derivativesy[i+1] <- psi1 * derivativesy[i] + psi2 * derivativesy[i-1] + beta * derivativesx[i]
    } else if (i > 3) {
      derivativesx[i+1] <- phi1 * derivativesx[i] + phi2 * derivativesx[i-1] + phi3 * derivativesx[i-2]
      #derivativesy[i+1] <- psi1 * derivativesy[i] * shock + psi2 * derivativesy[i-1] * shock +
      #psi3 * derivativesy[i-2] * shock + beta * derivativesx[i] * shock
      derivativesy[i+1] <- psi1 * derivativesy[i] + psi2 * derivativesy[i-1] + psi3 * derivativesy[i-2]
    }
  }

  irfx <- lastgdp + derivativesx
  irfy <- lastun + derivativesy

  dfIRF <- data.frame(1:(h+1), irfx, irfy)
  names(dfIRF) <- c("Lag", "GDP_QGR", "UN_RATE")

  return(dfIRF)
}
```

And we can compute the IRF to plot the graphs.

```r
# Get the coefficients
phi1  <- coef(mAR3)[1]
phi2  <- coef(mAR3)[2]
phi3  <- coef(mAR3)[3]
psi1  <- coef(mADL31_0)[2]
psi2  <- 0
psi3  <- coef(mADL31_0)[3]
beta  <- coef(mADL31_0)[4]

# Create a vector with parameters
parameters <- c(phi1, phi2, phi3, psi1, psi2, psi3, beta)
janela = 100

# If you want, may also create variables for shock, last observation, etc.
dfIRF1 <- fIRFADL31(parameters, shock = 0.02*100, h = janela,
                    lastgdp = -0.0037*100, lastun = 0.078*100)
dfIRF2 <- fIRFADL31(parameters, shock = -0.02*100, h = janela,
                    lastgdp = -0.0037*100, lastun = 0.078*100)
```



As discussed before, the AR(3) model provides some sort of persistence in the GDP series. Thus, the IRF for GDP growth is in line with what is expected from this model. That is, we see that a positive shock tends to gradually stabilize over the quarters. Regarding unemployment, we observer a slower movement back to the initial levels. This might be a result of the cyclical behaviour of unemployment discussed previously, which ended up being reflecting on how it responds to shocks over the economic activity.

However, those interpretations must be taken with cautious, given that the confidence intervals for the IRF are not available and we do not know exactly in which lags they are significantly different from zero.

Econometrics III HW - part 3

A. Schmidt and P. Assunção

March, 2020

Assignment 3

See the source code if interested in all functions (chunks were ommited unless relevant for the assignment). Click here to access the code.

Introduction

Suppose that you are asked to analyze investment opportunities in the stock market by studying the dynamic behavior of 10 major stock prices. In particular, you are asked to study the stocks of Apple, Intel, Microsoft, Ford, General Electrics, Netflix, Nokia, Exxon Mobil, and Yahoo, as well as the S&P500 stock market index.

You can find the entire sample of daily data at your disposal in the Eviews workfile labeled data assign p3.wf1. This data set contains the daily stock prices of all the companies mentioned above and spans from the 14th of February 2007 to the 28th of January of 2013.

Importing and checking data

Import using read.csv2() function - the data is in an online directory and there is no need to change this address.

```
urlRemote  <- "https://raw.githubusercontent.com/aishameriane"
pathGithub <- "/Mphil/master/EconIII/data_assign_p3.csv"
token      <- "?token=AAVGJTT32POPYVK67LDJURK6QMGQG"

url      <- paste0(urlRemote, pathGithub, token)
dfData01 <- read.csv2(url, sep = ",", dec = ".", header = TRUE)
```

Check if everything is ok with the dataset: header and tail and summary statistics to check for missing data/outliers. We can see from the head and tail that the dataset that at least those observations seems to be completely filled with adequate ranges. To avoid breaking the margin spaces, we transposed the table, so the columns are in the rows and the original row names (which are the row positions) are appearing as the column names in the table below.

	1	2	3	1497	1498	1499
obs	1.00	2.00	3.00	1497.00	1498.00	1499.00
APPLE	84.63	85.44	85.25	460.00	451.69	437.83
DATE	732720.00	732721.00	732722.00	734891.00	734892.00	734895.00
EXXON_MOBIL	75.19	75.26	74.90	91.57	91.67	91.21
FORD	8.45	8.55	8.54	13.82	13.83	13.49
GEN_ELECTRIC	35.93	36.36	36.07	21.96	22.28	22.44
INTEL	20.96	21.21	21.26	21.10	21.04	21.01
MICROSOFT	29.17	29.58	28.91	27.70	27.58	28.01
NETFLIX	22.90	22.48	22.52	143.99	145.67	172.51
NOKIA	22.91	22.95	23.03	4.21	4.18	4.26
SP500	1443.91	1455.15	1456.77	1494.81	1494.82	1502.96
YAHOO	29.69	30.82	31.00	20.08	20.43	20.50

The column with the dates (DATE) is in numeric format and not in a nice position, so we will change to y-m-d and also switch places with the APPLE column.

```r
# To discover the origin, first we computed ymd("2007-02-14")-732720.
dfData01$DATE <- as.Date(dfData01$DATE, origin="0001-01-01")
dfData01      <- dfData01[,c(1,3,2,4,5,6,7,8,9,10,11,12)]
```

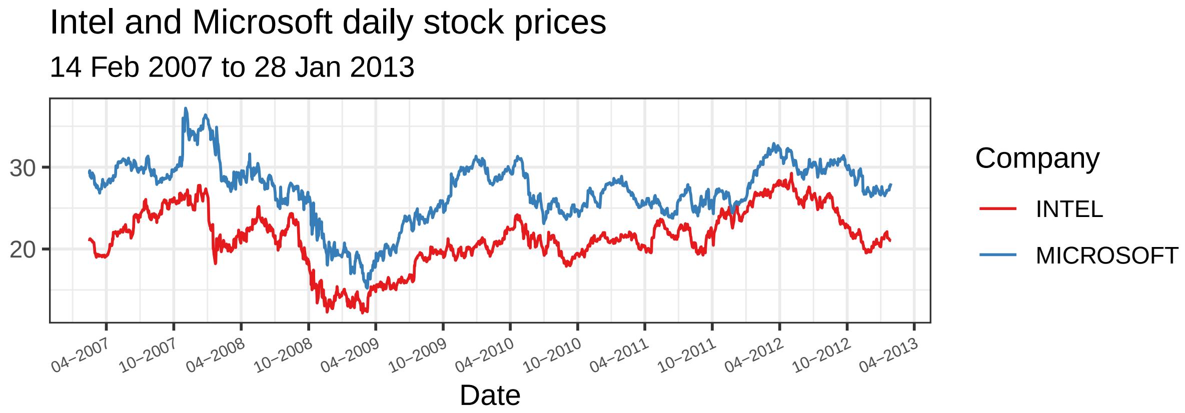
The next table has the descriptive statistics for the columns with information regarding the stocks. We can see that indeed we don't have any missing information and all values are numeric (there are no problems of formatting).

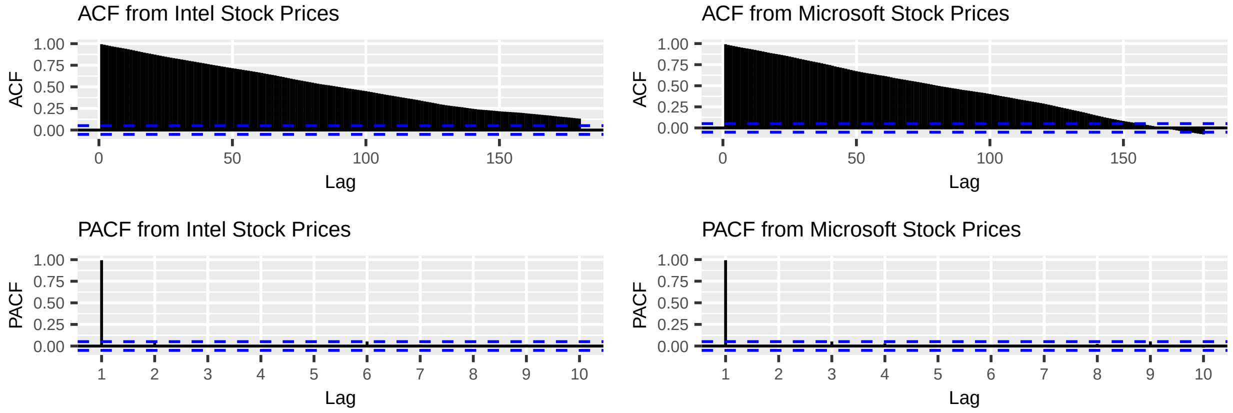
	Minimum	1st quartile	Mean	Median	3rd quartile	Maximum	Desv. Pad.
APPLE	79.39	139.670	275.6362	207.87	369.410	702.41	166.3191
EXXON_MOBIL	56.85	70.130	78.1742	79.69	85.935	95.08	9.2136
FORD	1.31	6.890	9.4389	9.46	12.085	18.81	3.7610
GEN_ELECTRIC	6.75	16.075	22.0685	19.08	28.135	42.03	8.6233
INTEL	12.17	19.630	21.3759	21.39	23.715	29.26	3.5034
MICROSOFT	15.20	25.240	27.1031	27.54	29.640	37.22	3.7298
NETFLIX	16.58	30.255	82.6573	57.45	109.910	300.70	68.9289
NOKIA	1.66	6.280	14.8695	12.55	22.480	41.70	10.3756
SP500	679.28	1104.265	1237.3728	1280.21	1392.240	1564.98	197.8307
YAHOO	9.10	15.010	18.2945	16.17	20.085	34.07	5.2936

The total number of observations is 1499.

Question 1

Provide plots for two stock market time-series at your choice and report 12-period ACF and PACF functions for those two time series. What does the sample ACF tell you about the dynamic properties of these stocks?





Comments on the graphs: Both stocks seems to move closely (top graph), which is not surprising given that they are in the same sector. However, Microsoft' series exhibits a higher level (average) than Intel's, which can be observed by the gap between the two series. In general, the gap seems constant, with exceptions in te beginning of 2009 and the end of 2011, where they it is smaller (series are close to each other). Also, we notice some periods when the gap is wider: in the second semester of 2007, both series were in an ascending escalade, but Microsoft value went higher and the gap was also wider; during 2009, Microsoft showed higher increase over its stock prices, culminating in a wider gap during the transition to 2010.

As for the ACF, we notice that there is a persistent behavior in both series (bottom graphs). That is, there is a slow autocorrelation decay. Almost all the lags up to 180 (6 months, approximately) are significant and there is no indication of a "cut" in some lag, as would be the case in an ARMA model with the MA coefficients different from zero. Moreover, the PACF, which is obtained by computing the autocorrelation that is remaining after considering the previous lags, is highly persistent for the first lag, as for the others up until lag 10 it seems that they all fall inside the confidence intervals (we also did for higher lags and observed the same behavior). So, at least for these two series, it seems that we are dealing with an AR(1) with an unit root (i.e., a random walk).

Question 2

Perform an ADF unit-root test for all the 10 time series using the general-to-specific approach based on the Schwarz Information Criterion (SIC). Report the values of the ADF test statistics. Is the unit-root hypothesis rejected for any time-series at the 90% confidence level? Did you expect to reject the unit-root hypothesis for some time-series at this confidence level? Justify your answer carefully.

From the lecture slides, we have that the Augmented Dickey Fuller (ADF) test for an AR(p) model can be obtained as follows.

Start from the AR(p) equation:

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_p X_{t-p} + \epsilon_t \quad \Rightarrow \quad \phi(L) X_t = \epsilon_t \tag{1}$$

where $\phi(L) = 1 - \phi_1 L - \cdots - \phi_p L^p$. There is unit root if $\phi_1 L + \cdots + \phi_p L^p = 1$. An easier way to test if this is true is by writing the model above in terms of first differences in the AR(p) model above.

$$\Delta X_t = (\phi_1 L + \cdots + \phi_p L^p - 1) X_{t-1} + \phi_2^* \Delta X_{t-1} + \phi_3^* \Delta X_{t-2} + \cdots + \phi_p^* \Delta X_{t-p+1} + \epsilon_t \tag{2}$$
$$\Delta X_t = \beta X_{t-1} + \phi_2^* \Delta X_{t-1} + \phi_3^* \Delta X_{t-2} + \cdots + \phi_p^* \Delta X_{t-p+1} + \epsilon_t \tag{3}$$

where $\phi_j^* = -\sum_{i=j}^{p} \phi_i$ for $j = 2, \cdots, p$. Therefore, since zero *beta* would imply unit roots, the ADF test will be testing the null hypothesis $H_0 : \beta = 0$ being equal to zero versus the alternative $H_A : -2 < \beta < 0$, which implies stationarity. The test statistic of such test is $ADF = \frac{\hat{\beta}}{SE(\hat{\beta})}$, which follows a Dickey-Fuller distribution.

Our procedure then will be:

- For each serie, estimate an AR(10) model with drift using the same function used in part 1 (`Arima()`);
- In sequence, remove the lags, and store the SIC value for each combination, until reach an AR(1);
- Recover which model had the smallest SIC (BIC is the same, see: https://en.wikipedia.org/wiki/Bayesian_information_criterion);
- Perform the ADF test considering the model with best BIC.

We opted by not estimating an `ARMA(p,q)` model to avoid parameter redundancy (as discussed in Chapter 7 of Box et al. (2015)). In particular, since we know in advance (from previous courses) that it is not possible to predict stock prices (or returns) and the usual procedure in finance is to model the variance, we already expected beforehand that any `ARMA(p,q)` would not be correctly specified anyway and focused in just making the exercise using the `AR(p)` model. But to make things more interesting, we challenged ourselves in developing the routine to pick the best `AR(p)` model from all possible lag combinations (including the removal of intermediate lags).

```
# This function will take the series and test all combinations of lags
# up until lag max to fit an AR model
# Default method is MLE, in some cases will
# change the estimation method to quasi likelihood (therefore no AIC)
# It returns the model with the minimum BIC

fEstAR <- function(data, series, lagmax){

    # Extract the correct series
    tsData <- xts(data[,series], order.by = data[,2])

    # Prepare the combinations of the lags
    matriz <- list(matrix(rep(NA,1), ncol = 1))

    if (lagmax > 1){
        for (i in 2:lagmax){
          matriz[[i]] <- matrix(rep(NA,(2^(i-1))*i), ncol = i)
        }
    }

        # Assemble a list with all possible combinations of lags up until an AR(15)
        for (j in 1:lagmax){
          # The idea is that our objects are of this type
          #matriz[[j]][,] <- matrix(rep(0,(2^(j-1))*j), ncol = j, nrow = (2^(j-1)))
          # Now we fill with the bynary representation of the numbers,
```

```r
  #this gives all the possible combinations
  for (i in (2^(j-1)):(2^(j)-1)){
      # Feeds with the bynary combinations
      matriz[[j]][i-(2^(j-1)-1),] <- as.integer(intToBits(i)[1:j])
  }
}

# Now we just reorganize to use each line in the Arima() function
for (j in 1:lagmax){
  # Converts to what we need for the Arima() fix argument
  matriz[[j]] <- ifelse(matriz[[j]] == 1, NA, 0)
  # Reverses the order to make consistent with the first column being higher lag
  matriz[[j]] <- matriz[[j]][ , ncol(matriz[[j]]):1]
}

# Get the number of models

nModels <- 1
for (j in 2:length(matriz)){
  nModels <- nModels + nrow(matriz[[j]])
}
dfModels <- data.frame(rep(NA, nModels), rep(NA, nModels))
names(dfModels) <- c("BIC", "Model")

#tic("Total time:") ## Use for debugging
for (j in 1:length(matriz)){
    #tic(paste(c("AR ", j, ":"))) ## Use for debugging
    order <- j
    matriz2 <- matriz[[j]]

    for (i in 1:max(1, nrow(matriz2))) {
      coef  <- i

      if (j == 1){
        model <- Arima(tsData, order = c(order,0,0), fixed = c(matriz2[coef], NA))
      } else {
        model <- try2(Arima(tsData, order = c(order,0,0),
                            fixed = c(matriz2[coef, ], NA), method="CSS-ML",
                            optim.method = "BFGS"),
                      Arima(tsData, order = c(order,0,0),
                            fixed = c(matriz2[coef, ], NA), method="CSS"))
      }

      dBIC  <- model$bic
      if (j == 1){
        dfModels[((i-1)+2^(j-1)),] <- c(dBIC,
                                        paste("AR(", as.character(order),
                                              "), with coef (", NA, ")", sep=""))
      } else {
        dfModels[((i-1)+2^(j-1)),] <- c(dBIC,
                                        paste("AR(", as.character(order),
                                              "), with coef (", paste(matriz2[coef, ],
                                               collapse=", "), ")", sep=""))
```

```
              }

            }
        #toc() ## Use for debugging
        }
        #toc() ## Use for debugging

        selModel <- dfModels[which(dfModels[,1] == min(dfModels[,1], na.rm=TRUE)),]

    return(selModel)
}
```

Now we are ready to estimate the best (using BIC criteria) AR model for all the stocks.

The results from the best models for each company stock are summarized below.

```
## [1] "Best AR model for APPLE"
```

Variable	Estimate	Std. Error	t-statistic	P-Value
Lag 1	0.88	0.03	34.43	0.00
Lag 2	0.12	0.03	4.48	0.00
Intercept	275.44	143.85	1.91	0.06

```
## [1] "Best AR model for EXXON_MOBIL"
```

Variable	Estimate	Std. Error	t-statistic	P-Value
Lag 1	0.89	0.02	47.00	0
Lag 2	0.10	0.02	5.28	0
Intercept	78.99	3.84	20.57	0

```
## [1] "Best AR model for FORD"
```

Variable	Estimate	Std. Error	t-statistic	P-Value
Lag 1	1.00	0.00	612.13	0
Intercept	9.45	2.11	4.47	0

```
## [1] "Best AR model for GEN_ELECTRIC"
```

Variable	Estimate	Std. Error	t-statistic	P-Value
Lag 1	0.92	0.03	35.75	0
Lag 2	0.08	0.03	3.04	0
Intercept	24.05	6.24	3.85	0

```
## [1] "Best AR model for INTEL"
```

Variable	Estimate	Std. Error	t-statistic	P-Value
Lag 1	0.99	0.00	346.85	0
Intercept	21.37	1.32	16.21	0

```
## [1] "Best AR model for MICROSOFT"
```

Variable	Estimate	Std. Error	t-statistic	P-Value
Lag 1	0.99	0.00	320.25	0
Intercept	27.11	1.35	20.13	0

```
## [1] "Best AR model for NETFLIX"
```

Variable	Estimate	Std. Error	t-statistic	P-Value
Lag 1	1.0	0.00	743.43	0.00
Intercept	97.5	47.83	2.04	0.04

```
## [1] "Best AR model for NOKIA"
```

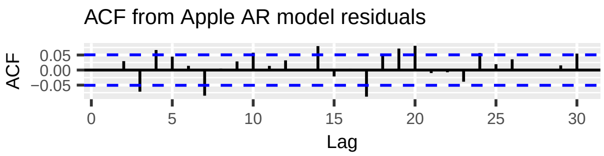
Variable	Estimate	Std. Error	t-statistic	P-Value
Lag 1	1.00	0.00	927.61	0.00
Intercept	14.85	7.01	2.12	0.03

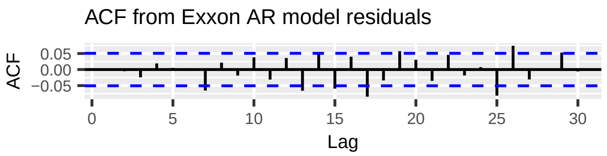
```
## [1] "Best AR model for SP500"
```

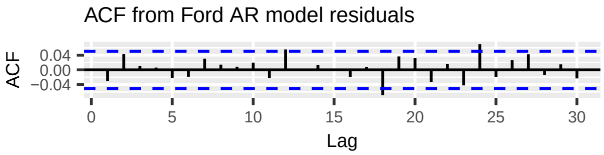
Variable	Estimate	Std. Error	t-statistic	P-Value
Lag 1	0.94	0.02	49.47	0
Lag 2	0.06	0.02	3.25	0
Intercept	1281.50	116.02	11.05	0

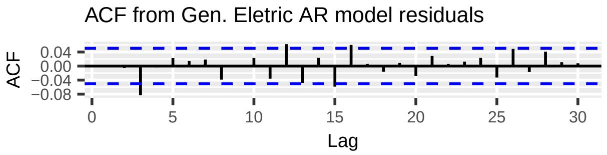
```
## [1] "Best AR model for YAHOO"
```

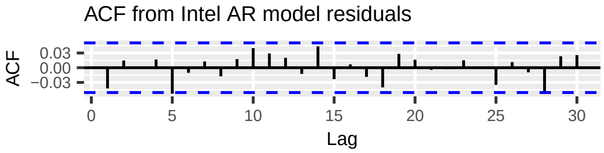
Variable	Estimate	Std. Error	t-statistic	P-Value
Lag 1	0.96	0.01	73.83	0
Lag 2	0.13	0.03	4.60	0
Lag 3	-0.09	0.03	-3.53	0
Intercept	19.32	2.86	6.76	0

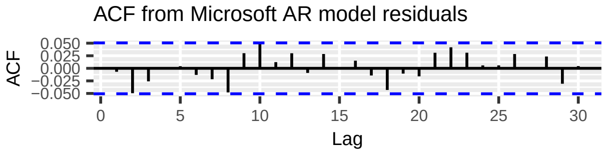

ACF from Apple AR model residuals

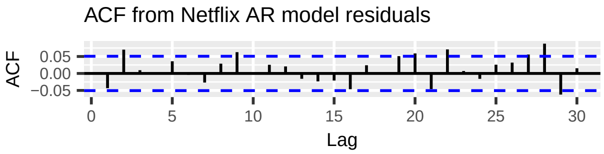

ACF from Exxon AR model residuals

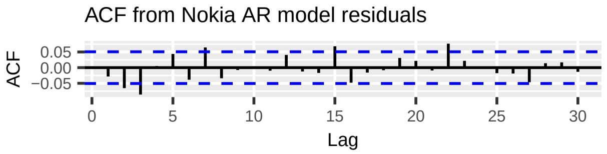

ACF from Ford AR model residuals

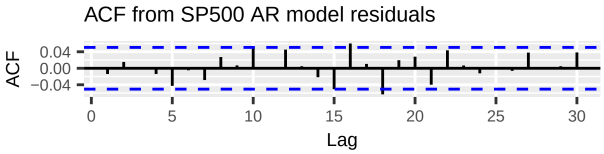

ACF from Gen. Eletric AR model residuals

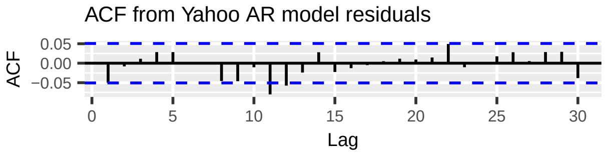

ACF from Intel AR model residuals


ACF from Microsoft AR model residuals

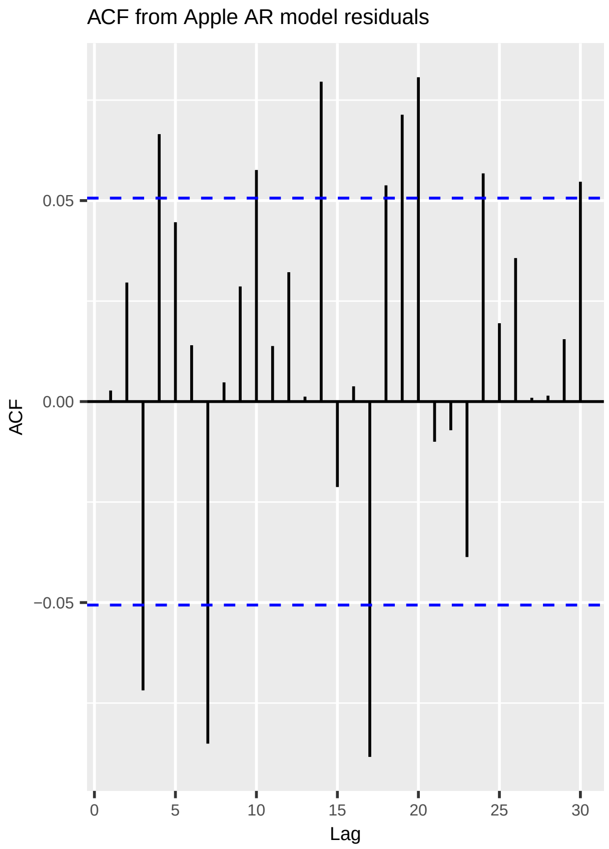

ACF from Netflix AR model residuals

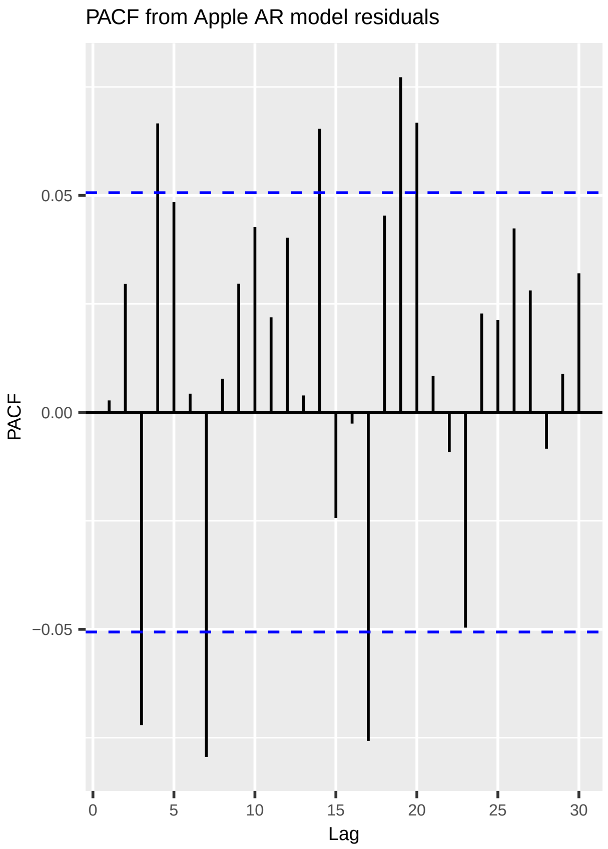

ACF from Nokia AR model residuals


ACF from SP500 AR model residuals


ACF from Yahoo AR model residuals

Comments: Although the models indeed show only significant coefficients, the ACF graphs of the residuals shows that some models might not be correctly specified, since not all lags are inside of the confidence intervals. For example, the ACF and PACF of the residuals from the AR(2) model estimated for Apple stocks are represented below. Clearly they do not resemble the expected graphs from a white noise (which we would have in the case of correct specification).


ACF from Apple AR model residuals


PACF from Apple AR model residuals

To further investigate this, we used an automatic model selection function, `auto.arima()`, which fits an ARIMA(p,d,q) model to the data. According to the function, the best model would be an ARIMA(0,1,0), i.e., no lag for the AR neither for the MA part, only a differencing term. However, since the goal of this exercise is not to find the best fit but work with the unit roots, we decided to not continue in this direction.

Now that we have our final models, we can compute the ADF statistic.

Dickey Fuller test

Following the lecture slides, we know that for an $AR(p)$ model, we are interested in computing $\beta = \phi_1 + \phi_2 + \ldots + \phi_p - 1$ and our test is $H_0 : \beta = 0$ versus $H_1 : -2 < \beta < 0$. The test statistic is $ADF = \frac{\hat{\beta}}{SE(\hat{\beta})}$. Although our "best" models are considering combinations of lags (with ommited intermediate lags), the computation of the test statistic would be cumbersome. **Due to restrictions of time, we will use just all the lags instead removing the intermediate ones that might not be significant.**

The function below is based on the `adf.test()` function from the package `tseries`. We changed slightly to take our series more easily from the dataframe and to organize the output for better visualization.

```
adfTest <- function (data, models, series, alternative = c("stationary", "explosive")) {

        model <- models[[series]]
        x     <- xts(data[,series+2], order.by = data[,2])
        k     <- length(model$coef)-1

        if ((NCOL(x) > 1) || is.data.frame(x))
          stop("x is not a vector or univariate time series")
```

```
    if (any(is.na(x)))
      stop("NAs in x")
    if (k < 0)
      stop("k negative")
    alternative <- match.arg(alternative)
    #DNAME <- deparse(substitute(x))
    DNAME  <- colnames(data[series+2])
    k <- k + 1
    x <- as.vector(x, mode = "double")
    # Takes the first difference, i.e, y = \Delta x_t
    y <- diff(x)
    n <- length(y)
    # creates a length(y)-k \times k matrix with lags of the
    #first difference series
    z <- embed(y, k)
    # This is the 1st lag of the first difference series
    yt <- z[, 1]
    # This is just to make size compatible
    xt1 <- x[k:n]
    # This is a sequence of numbers from k to n
    tt <- k:n
    if (k > 1) {
      # For more than one lag in the original model,
      #you need this adittional guys, they are lags of y
      yt1 <- z[, 2:k]
      # To understand the regression, use head(cbind(yt, xt1, x))
      # The lag of the first difference, yt,
      #is being regressed against the lag of the original series,
      #similar to the slides
      res <- lm(yt ~ xt1 + 1 + tt + yt1)  # This is for AR(p), p>1
    }
    else res <- lm(yt ~ xt1 + 1 + tt)     # This is for AR(1)
    res.sum <- summary(res)
    # Compute the test statistic
    STAT <- res.sum$coefficients[2, 1]/res.sum$coefficients[2,2]
    table <- cbind(c(4.38, 4.15, 4.04, 3.99, 3.98, 3.96), # Critical values
                   c(3.95, 3.8, 3.73, 3.69, 3.68, 3.66),
                   c(3.6, 3.5, 3.45, 3.43, 3.42, 3.41),
                   c(3.24, 3.18, 3.15, 3.13, 3.13, 3.12),
                   c(1.14, 1.19, 1.22, 1.23, 1.24, 1.25),
                   c(0.8, 0.87, 0.9, 0.92, 0.93, 0.94),
                   c(0.5, 0.58, 0.62, 0.64, 0.65, 0.66),
                   c(0.15, 0.24, 0.28, 0.31, 0.32, 0.33))
    table <- -table
    tablen <- dim(table)[2]
    tableT <- c(25, 50, 100, 250, 500, 1e+05)
    tablep <- c(0.01, 0.025, 0.05, 0.1, 0.9, 0.95, 0.975, 0.99)
    tableipl <- numeric(tablen)
    for (i in (1:tablen)) tableipl[i] <- approx(tableT, table[,i], n, rule = 2)$y
    # The next line locates the statistic in
    #terms of the critical values and gives the corresponding p-value
    interpol <- approx(tableipl, tablep, STAT, rule = 2)$y
    if (!is.na(STAT) && is.na(approx(tableipl, tablep, STAT,rule = 1)$y))
```

```
            if (interpol == min(tablep))
              warning("p-value smaller than printed p-value")
            else warning("p-value greater than printed p-value")
          if (alternative == "stationary")
            # If the test is H1 = stationary, then a p-value
            #above 0.1 will show evidence of an unit root (we are using this test)
            PVAL <- interpol
          else if (alternative == "explosive")
            # If the test is H1 = explosive, then a p-value
            #below 0.1 will show evidence of an unit root
            PVAL <- 1 - interpol
          else stop("irregular alternative")
          PARAMETER <- k - 1
          METHOD <- "Augmented Dickey-Fuller Test"
          names(STAT) <- "Dickey-Fuller"
          names(PARAMETER) <- "Lag order"
          return(data.frame(statistic = STAT, parameter = PARAMETER,
            alternative = alternative, p.value = PVAL, method = METHOD,
            data.name = DNAME))
          #structure(list(statistic = STAT, parameter = PARAMETER,
          #  alternative = alternative, p.value = PVAL, method = METHOD,
          #  data.name = DNAME), class = "htest")
}
```

```
## [1] "Results from the ADF test"
```

Company	p of AR(p)	Test-statistic	p-value
APPLE	2	-1.5281	0.7781
EXXON_MOBIL	8	-1.8534	0.6404
FORD	1	-1.7744	0.6738
GEN_ELECTRIC	2	-0.9946	0.9396
INTEL	1	-2.1926	0.4968
MICROSOFT	1	-2.4319	0.3955
NETFLIX	1	-1.3892	0.8369
NOKIA	1	-2.2834	0.4584
SP500	3	-1.3469	0.8548
YAHOO	7	-2.4886	0.3715

Comments: From the table with the the ADF test, using as null-hypothesis that the series has a unit root, we cannot reject H_0 for any ot the companies using a significance level of 10%. This result means that none of the stock prices series were generated by a stationary process. However, we give a word of caution: as mentioned in the other parts of this assignment, there is evidence that the AR models estimated were not correctly specified (look back at the residual analysis, for example). Overall, the identification of the presence of unit roots is a first step towards investigating issues related to spurious regression.

From the economic point of view, since the series are prices, not returns, it is not surprising that they all exibit unit roots using the test. Prices usually have a long dependency through the price level, which is known to be non-stationary because it is the accumulation of past prices. This is a similar problem we would have faced in part one when using GDP if we used the level GDP and not the growth rate.

Question 3

Assume that both the stocks of Apple and Microsoft follow a random walk process. Produce a 5-day forecast for the stocks of Apple and Microsoft. Add 95% confidence bounds to your forecasts under the assumption of Gaussian innovations. Is there any investment advice you can give on these stocks? Is their value expected to increase or decrease?

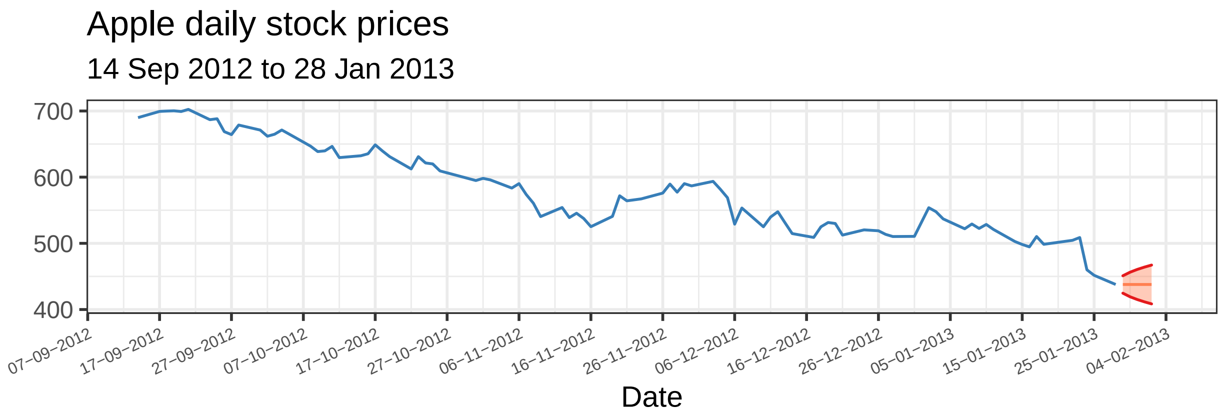
The random walk equation is given by:

$$X_t = X_{t-1} + \varepsilon_t$$

where $\varepsilon_t \sim \text{WN}(0, \sigma^2)$. As discussed in tutorial 4, the forecast to h period ahead is given by:

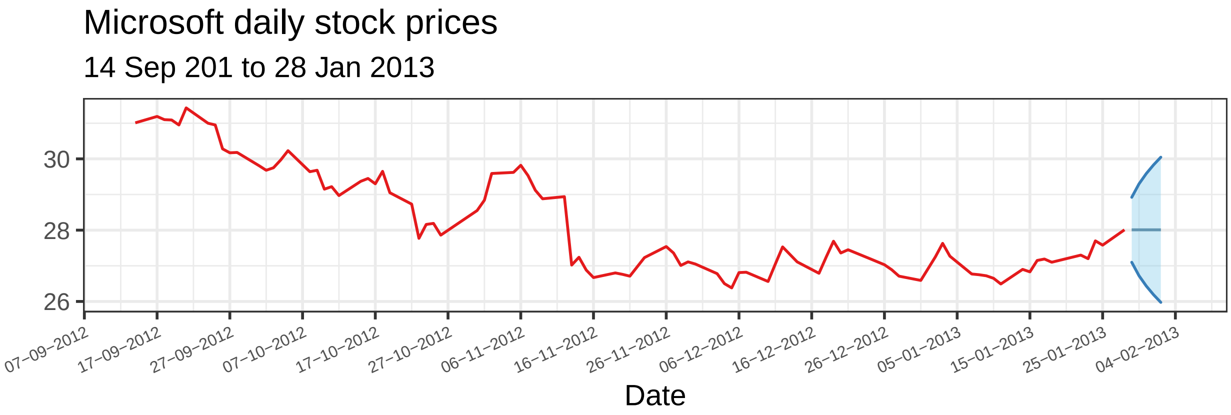
$$\hat{X}_{T+h} = 1^h \cdot \sigma^2 = \sigma^2$$

and the variance of the prediction error will be given by $h \cdot \sigma^2$. To compute things properly, we need an estimate for σ^2. Our procedure was to fit an AR(1) model (using the standard `lm()` function from R, since the `Arima()` function complains about the unit root) to obtain the estimate for σ^2 and use the formulas to obtain the forecasts.



Apple daily stock prices

14 Sep 2012 to 28 Jan 2013



Microsoft daily stock prices

14 Sep 201 to 28 Jan 2013

Comments: "Plots were made in different graphs due to scale differences and to be able to better visualize the forecast region we only ploted the last 90 days of each series. As expected, the forecasts are not informative about the prices because, as stated above, they are just the standard deviation of the innovations, which means that forecasts basically look at the innovations to compute the next period price. That is, random walk model applied in a finance context would imply that prices tomorrow cannot be predicted from historical trend, since prices today will not tell you what will happen with the prices tomorrow. With this in mind, we would not give any investment advice based on this kind of model, given it does not help when it comes to predicting stock prices. In other words, it is not possible, based on this model, to say if we expected the price to go up or down, since the forecast is always the same last value

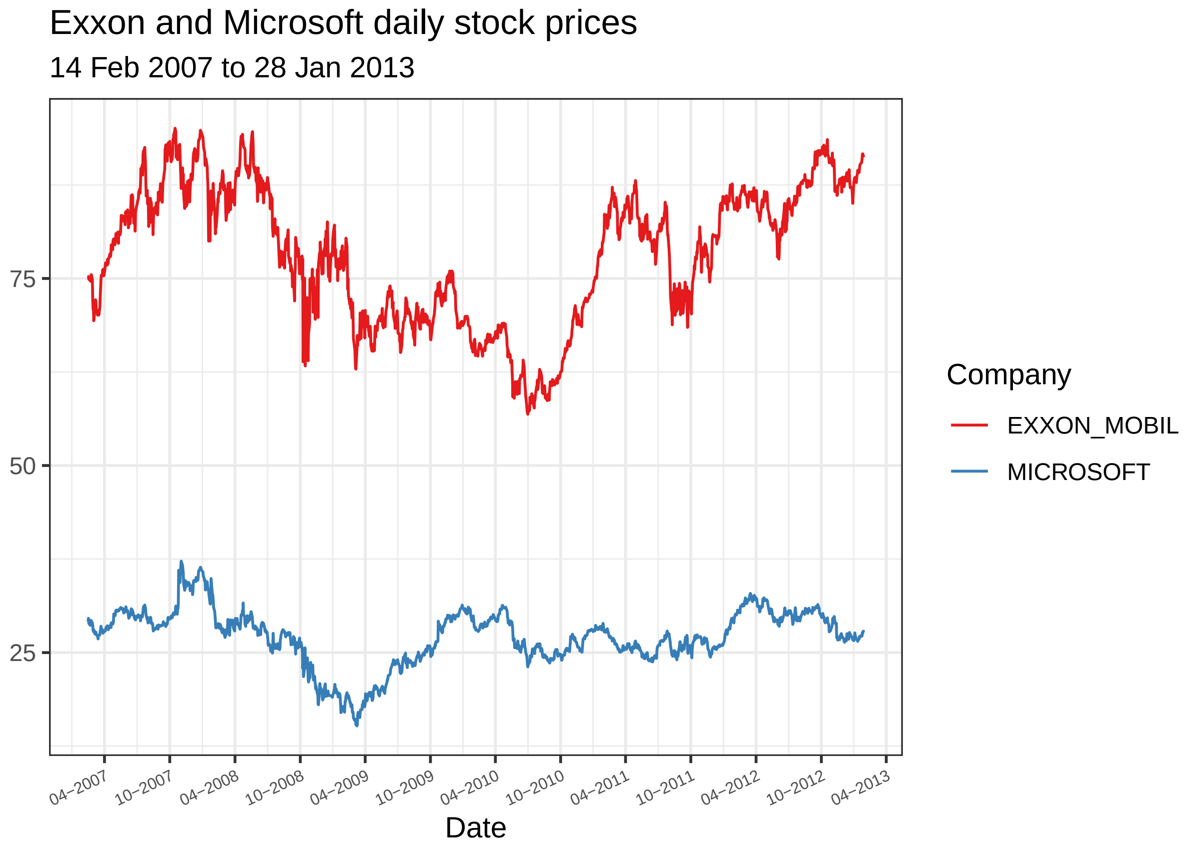
Question 4

Please investigate the following claim:

"Financial analysts have found that changes in the price of Microsoft stocks can be largely explained by fluctuations in the market value of Exxon Mobile. According to these analysts, this shows the extent to which the Microsoft Corporation is currently exposed to the market performance of the oil and gas industry."

Do you find a statistically significant contemporaneous relation between Microsoft and Exxon Mobile stock prices? Do you agree that changes in Microsoft stock prices are largely explained by fluctuations in the stock price of Exxon Mobile? Justify your answer.

To better answer this question, first we shall make the visual inspection in both series.



Comments: Given the different scale in the stock prices, the visual inspection is not clear that the series move together, as it was for Intel and Microsoft prices. However, we can notice some common downwards and upward trends in the series, which might suggest that they are both $I(1)$ and candidate to be cointegrated.

So, we proceed to analyze if in fact the series are cointegrated to rule out spurious regression. As we discussed in the lectures, if two series Y_t and X_t are cointegrated, then $Y_t - \lambda X_t$ is stationary. Since we do not have the value for λ, it must first be estimated.

A simple way to do this is first compute the regression $Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t$ to get an estimate for β_1 using OLS. Then, we compute the residuals using $\hat{\varepsilon}_t = Y_t - \hat{\beta}_0 - \hat{\beta}_1 X_t$ and verify the stationarity using the ADF test routine that we used before in previous tests (since now we have a single series and not several as before, it was best to use the package function directly and not our modified version).

Using the null hypothesis that the residuals are not stationary, we obtained a test statistic of -2.2186, which corresponds to a p-value of 0.4858, so the conclusion is that we cannot reject H_0 for a significance level of 10%.

Therefore, the analysis leads us to conclude that both series are not cointegrated and it is not correct to use Exxon prices to explain Microsoft prices. Apart from the statistical analysis, even without this results, we think it would be hard to sell an economic story on how the oil and gas prices could affect Microsoft shares. It is true that oil prices have an impact in production of goods and services in the real world, because they affect directly the transportation sector. However, Microsoft is a software industry, mostly, therefore the delivery of its products should not be so sensible to gas and oil, even if the demand for its products rise as a consequence of the oil price fluctuation. And, of course, in case of increase in demand, it could be that Microsoft needs to increase the inputs, but this means hiring more workers, which also would not be affected by gas prices.

References

Box, George EP, Gwilym M Jenkins, Gregory C Reinsel, and Greta M Ljung. 2015. *Time Series Analysis: Forecasting and Control*. John Wiley & Sons.