Time Series 732A62

Lab 2

Eric Herwin
erihe068

Albin Västerlund
albva223

Contents

```
## Warning: package 'forecast' was built under R version 3.4.2
```

Assignment 1. Computations with simulated data

a)

In this assignment we will simulate 1000 observations from a AR(3) process with $\phi_1 = 0.8$, $\phi_2 = -0.2$ and $\phi_3 = 0.1$. We will then use the definition of PACF to compute ϕ_{33}. We will compare the ϕ_{33} computed with regressions with ϕ_{33} from the function `pacf()`.

To compute ϕ_{33} we use this formula:

$\hat{x}_t = \beta_1 x_{t+1} + \beta_2 x_{t+2}$

$\hat{x}_{t+3} = \beta_1 x_{t+2} + \beta_2 x_{t+1}$

$\phi_{33} = corr(x_t - \hat{x}_t, x_{t+3} - \hat{x}_{t+3})$

```
cor(res_xt,res_xt3)# Computed by with regressions
```

```
## [1] 0.1233879
```

```
pacf_phi_33 # Computed from pacf()
```

```
## [1] 0.1049312
```

The ϕ_{33} computed with regressions is very close to the ϕ_{33} computed by `pacf()`.

b)

In this assignment we will simulate 100 observations from a AR(2) process with $\phi_1 = 0.8$ and $\phi_2 = 0.1$. We will then fit three diffrent models to estimate the parameters (ϕ_1 and ϕ_2). The three diffrent models are computed by Yule-Walker equations, conditional least squares and maximum likelihood (ML). We will then compare the models to se which one who seems to be the best model. We will also compute a confidence interval for for ML estimate of ϕ_2 to look if the theoretical ϕ_2 is inside the interval.

```
set.seed(12345)
dataB <- arima.sim(model = list(ar=c(0.8, 0.1)), n = 100)
```

```
#Yule-Walker equations
ar.yw(x = dataB, order.max = 2,aic = FALSE)
```

```
##
## Call:
## ar.yw.default(x = dataB, aic = FALSE, order.max = 2)
##
## Coefficients:
##      1       2
## 0.8029  0.1037
##
## Order selected 2  sigma^2 estimated as  1.267
```

```
#conditional least squares
arima(x = dataB, order = c(2,0,0),method = "CSS",include.mean = FALSE)
```

```
##
## Call:
```

```
## arima(x = dataB, order = c(2, 0, 0), include.mean = FALSE, method = "CSS")
##
## Coefficients:
##          ar1     ar2
##       0.8092  0.1254
## s.e.  0.0983  0.0986
##
## sigma^2 estimated as 1.13:  part log likelihood = -148.01
```

```
#Maximum likelihood (ML)
ML_arima <- arima(x = dataB, order = c(2,0,0), method = "ML", include.mean = FALSE)
ML_arima
```

```
##
## Call:
## arima(x = dataB, order = c(2, 0, 0), include.mean = FALSE, method = "ML")
##
## Coefficients:
##          ar1     ar2
##       0.8013  0.1255
## s.e.  0.0991  0.0995
##
## sigma^2 estimated as 1.13:  log likelihood = -148.93,  aic = 301.86
```

The estimate of the parameters from the three models are very similar. The larger n becomes the more similar will the parameters will get.

```
lower<-ML_arima$coef[2] - 1.96*0.0995
upper<- ML_arima$coef[2] + 1.96*0.0995
c(lower,upper)
```

```
##         ar2         ar2
## -0.06949667  0.32054333
```

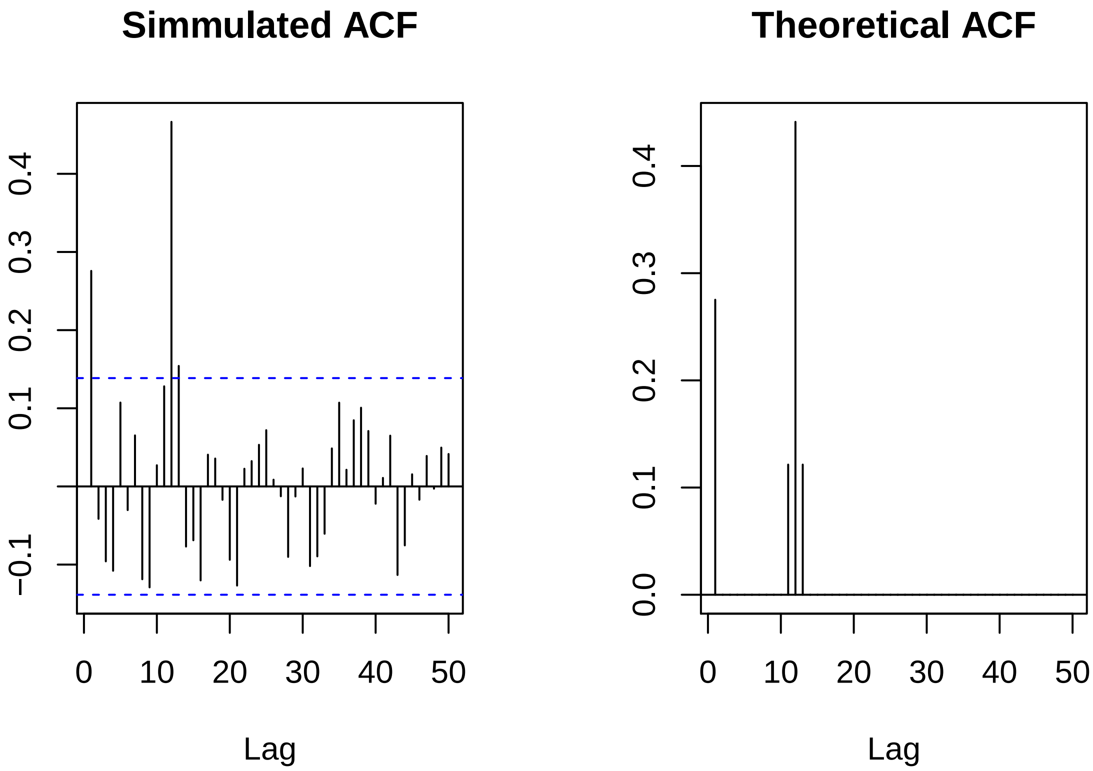
The interval include the theoretical value of ϕ_2, which is 0.1

c)

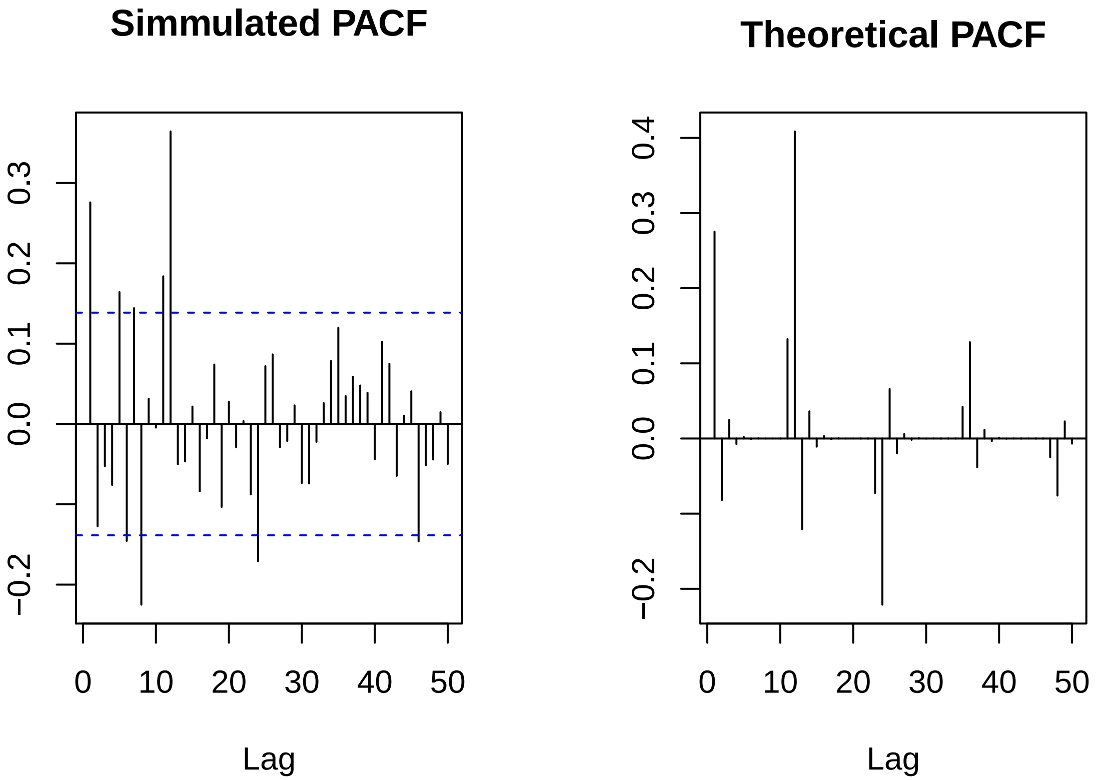
In this assignment we will simulate 200 observations from a SARIMA(0,0,1)*(0,0,1)$_{12}$ process with $\theta = 0.3$ and $\Theta = 0.6$. Will will then compare the theoretical and sample ACF and PACF.

$x_t = \Theta(B^{12})\theta(B)w_t = (1 + \Theta B^{12})(1 + \theta B)w_t = w_t + \theta w_{t-1} + \Theta w_{t-12} + \theta\Theta w_{t-13}$

$x_t = w_t + 0.3w_{t-1} + 0.6w_{t-12} + 0.3 * 0.6w_{t-13}$



The ACF for the simmulated data and the theoretical ACF seems to have the same pattern. If we would have a larger sample the simmulated ACF and theoretical ACF would turn out to be equal.



Like the ACF, the simmulated PACF have the same pattern as the theoretical PACF. If an infinity large sample would have been taken the graphs would be equal.

d)

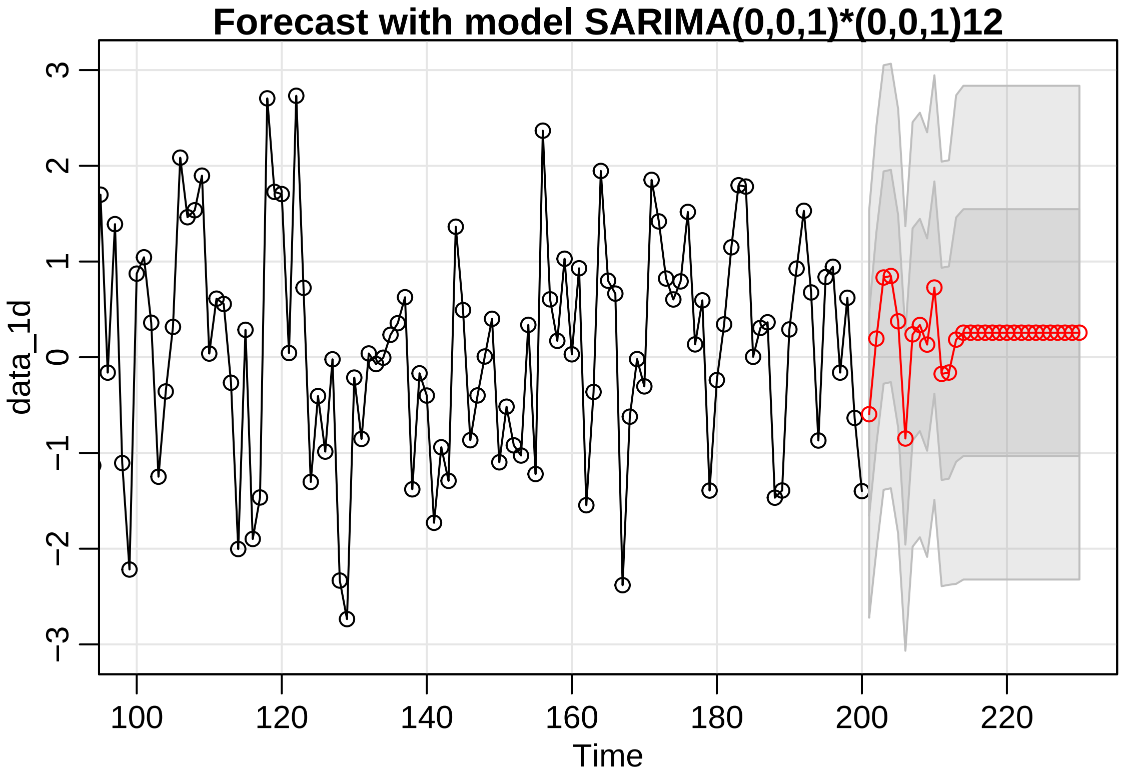
In this assignment we will use the data from 1c and fit a SARIMA(0,0,1)*(0,0,1)$_{12}$ to the data. We will then compute forecasts and a prediction band 30 points ahead and plot the original data and the forecast with the prediction band. We will also fit a model by using the function `gausspr` in r and plot it's fitted values and the orginal data in the same graph. In the end we will compare the two graphs and make conclusions.

We start with fit a model:

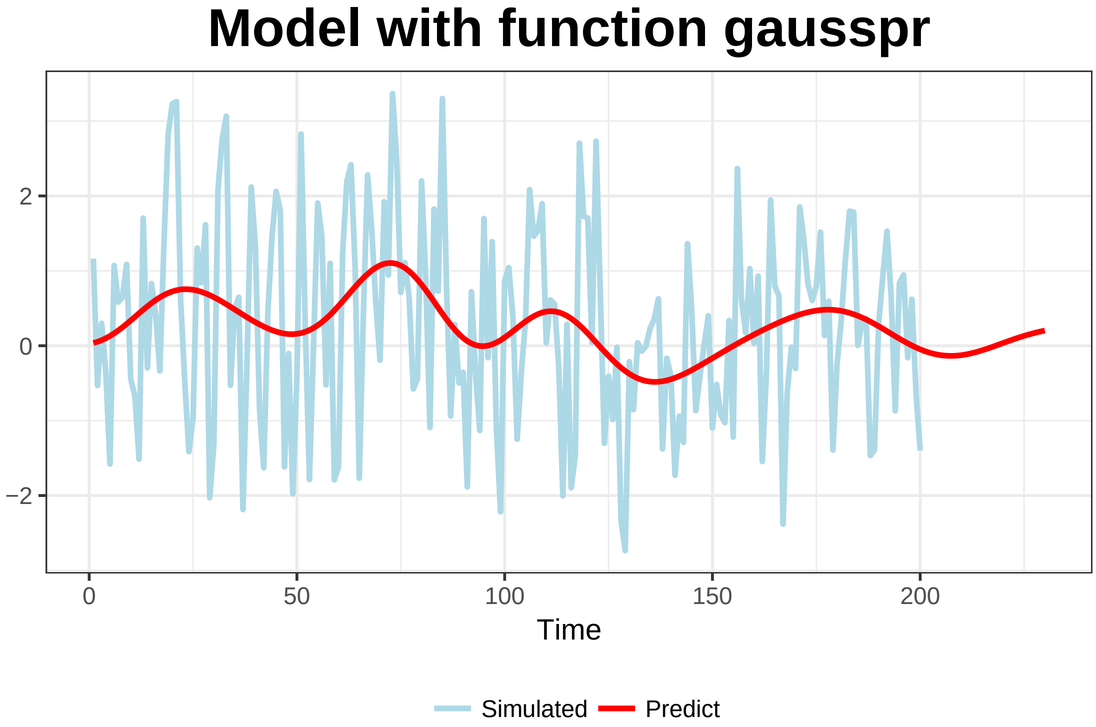
```
arima(data_1d,order = c(0,0,1),seasonal = list(order = c(0, 0, 1), period = 12))
```

```
##
## Call:
## arima(x = data_1d, order = c(0, 0, 1), seasonal = list(order = c(0, 0, 1), period = 12))
##
## Coefficients:
##          ma1    sma1  intercept
##       0.2966  0.5938     0.2568
## s.e.  0.0709  0.0615     0.1519
##
## sigma^2 estimated as 1.13:  log likelihood = -298.67,  aic = 603.34
```

The standard error is low compared to the parameter estimation of θ and Θ. Which means that the parameters indicates to be significant. We can also se that the estimate of θ and Θ are very close to the true value of the parameters ($\theta = 0.3$, $\Theta = 0.6$).



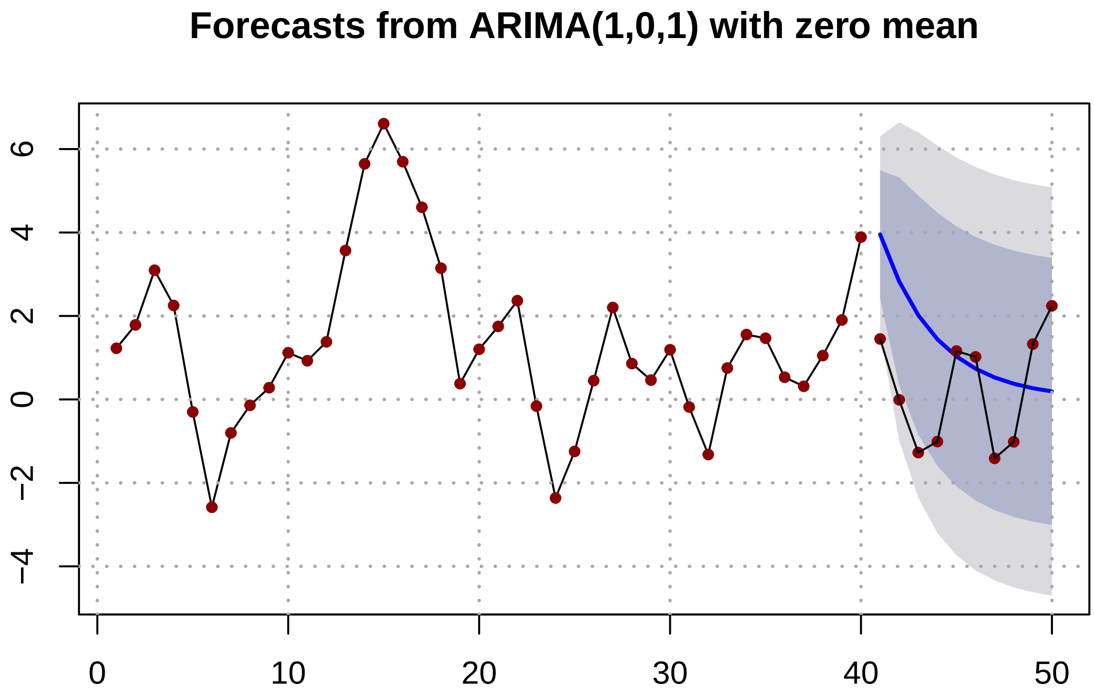
```
## Using automatic sigma estimation (sigest) for RBF or laplace kernel
```



The big diffrent between `gausspr` function and SARIMA is that `gausspr` fit a much smoother prediction than SARIMA. But they are similar in the way that both of their prediction ending up in the mean of the simulated data. We are not sure how `gausspr` estimate it's parameters and make the prediction. But it seems like the model have a hard time picking up the SMA(1) part in the data. One example would be to include a seasonal vector in to the `gausspr` model to get the season.

e)

In this assignment we simulate 50 observations from a ARIMA(1,0,1) process with $\theta = 0.5$ and $\phi = 0.7$. We will then fit a ARIMA(1,0,1) to the first 40 observations and use that model to fit the observations between 40 and 50. After that we plot the simulated data with the prediction interval to se if the last 10 observation is in the prediction interval or outside.



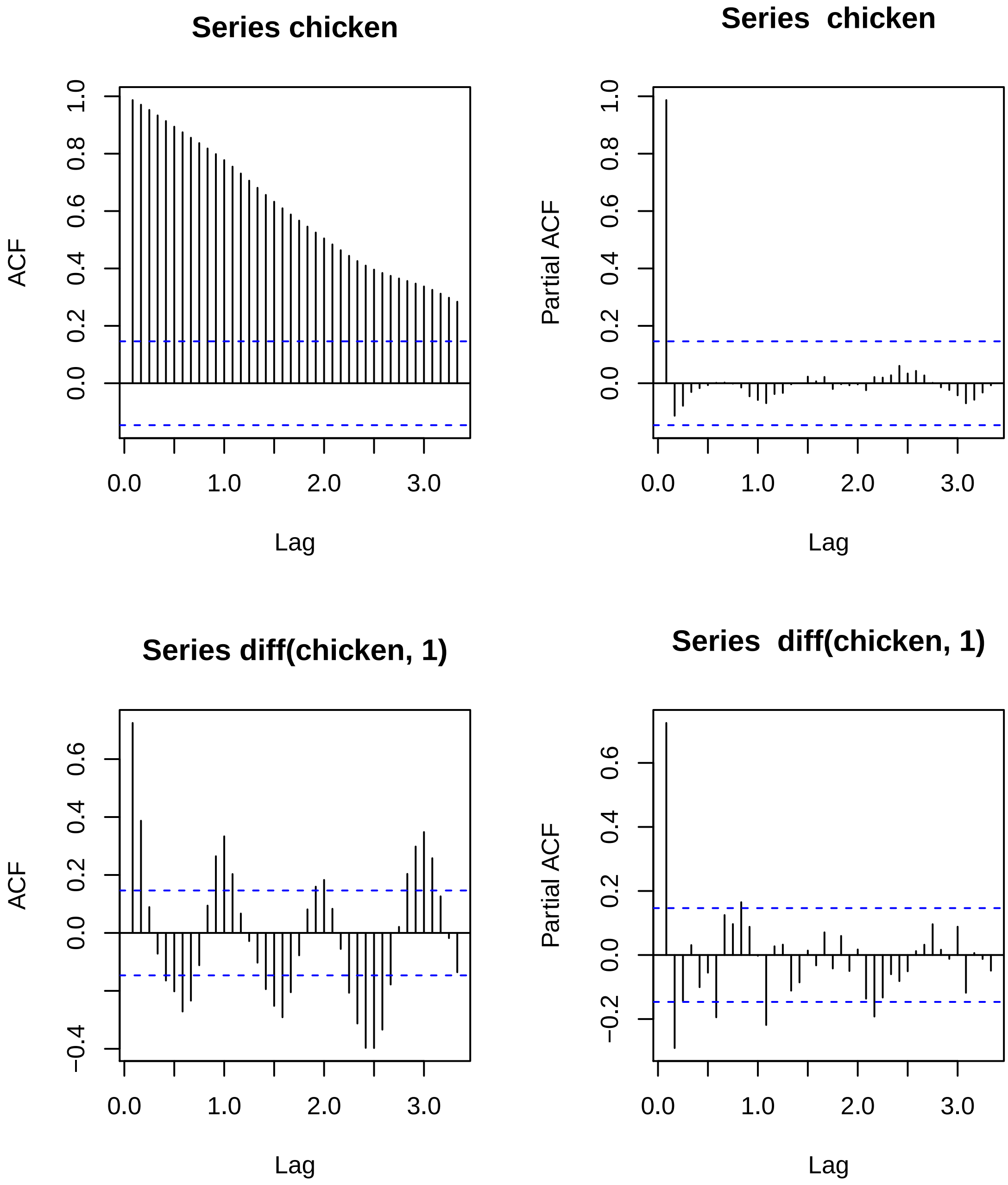
The graph show the simulated data and the fitted values for the last 10 observation with 80% and 95% prediction interval. All "true" last observation are within the 95% prediction interval except the first which is just outside interval. The meaning of a 95% prediction interval is that if we would calculate infinity prediction forward 95% of them should be within the interval.

Assignment 2. ACF and PACF diagnostics

a)

In this assignment we will study the chicken data which is in the astsa package. Four different graphs will be made, two ACF and two PACF.

$ACF(x_t)$, $PACF(x_t)$, $ACF(\nabla x_t)$, $PACF(\nabla x_t)$

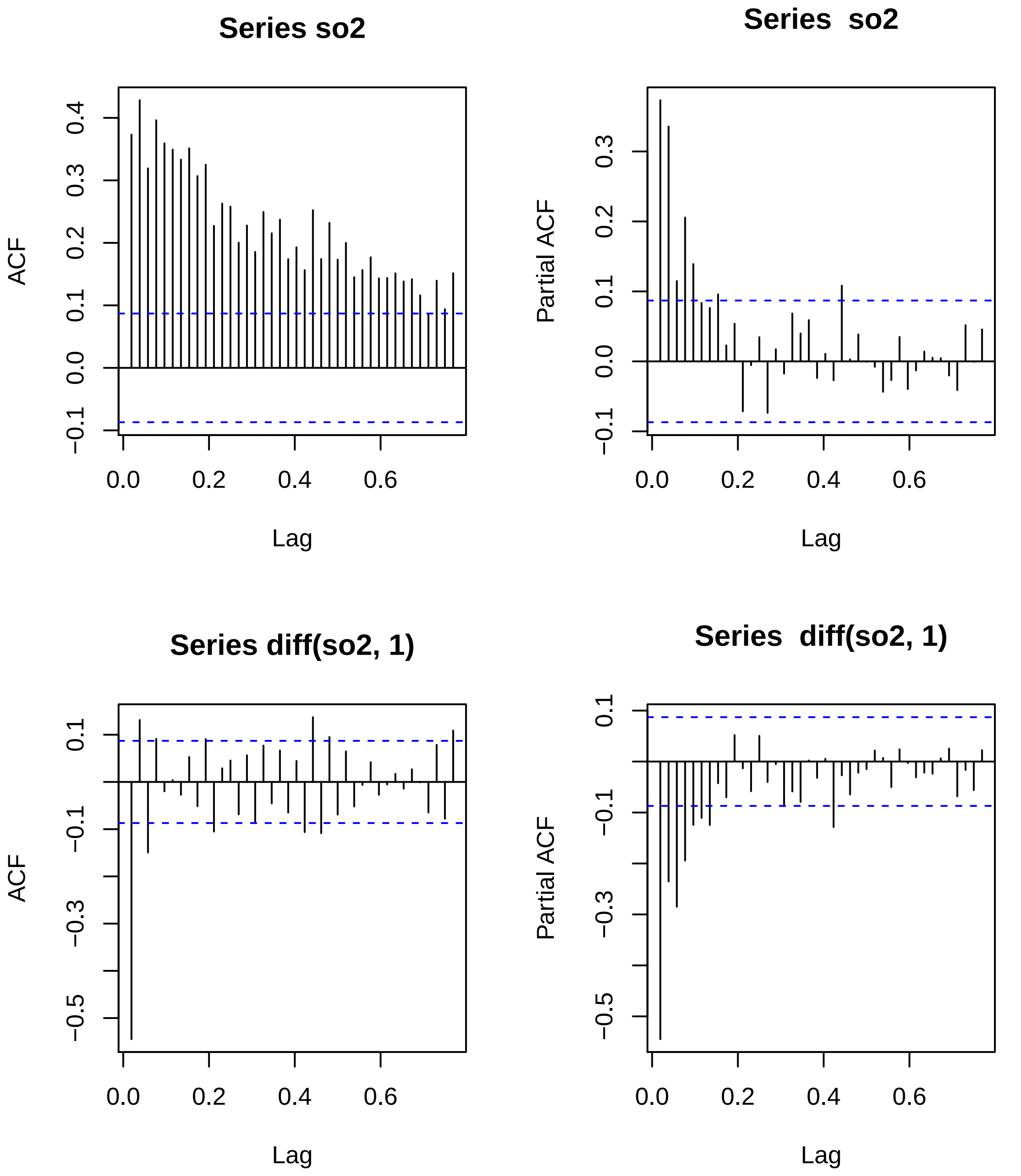


Without differentiation the serie is not stationary (the ACF decreaseing to slowly). So that serie is not optimal to do a model on. If we do the first differentiation the serie seems to get stationary. We se a decreasing pattern in the ACF and two spikes in the PACF. We would suggest the model ARIMA(2,1,0).

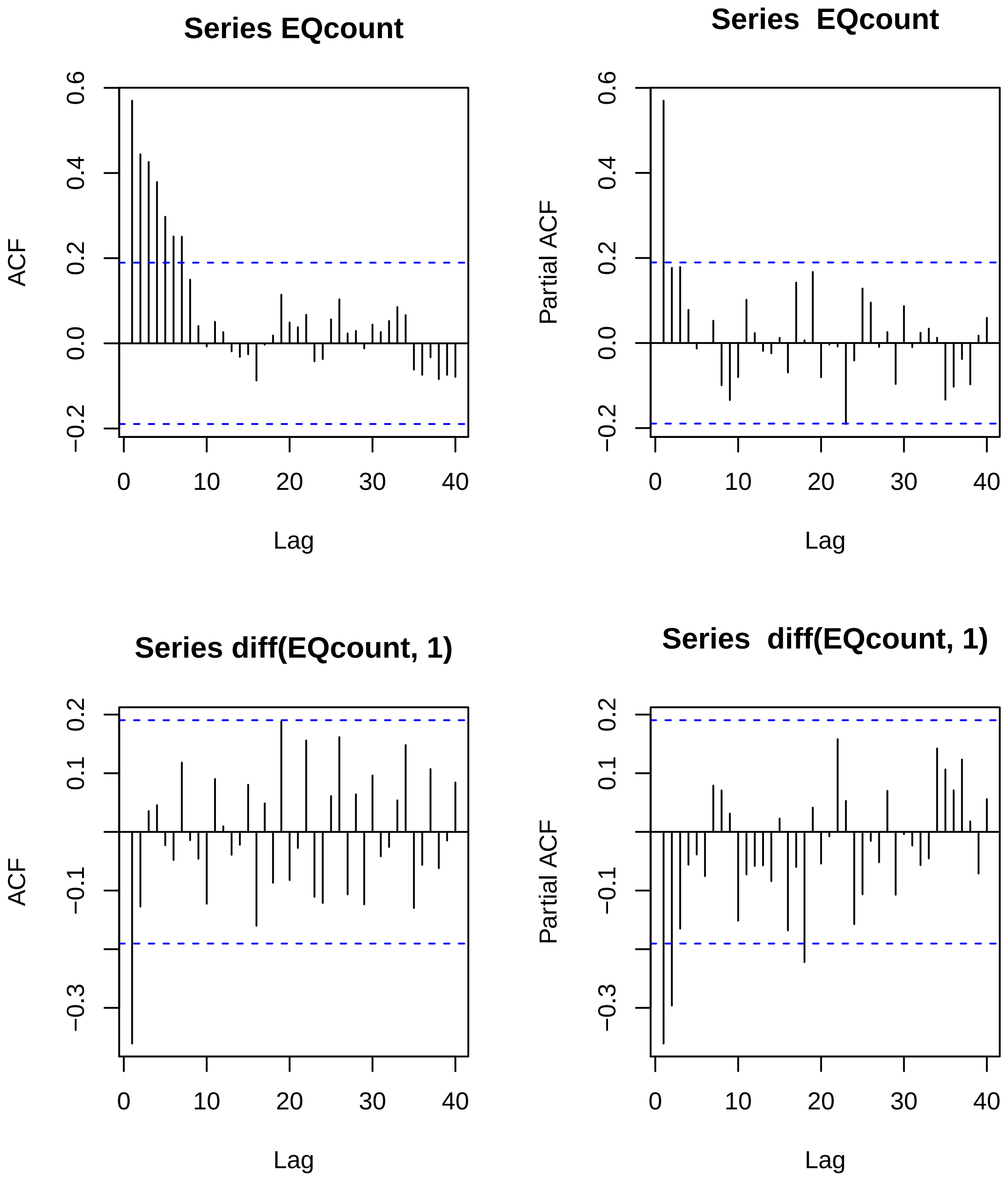
b)

In this assignment we study three diffrent datasets, so2, EQcount and HCT.

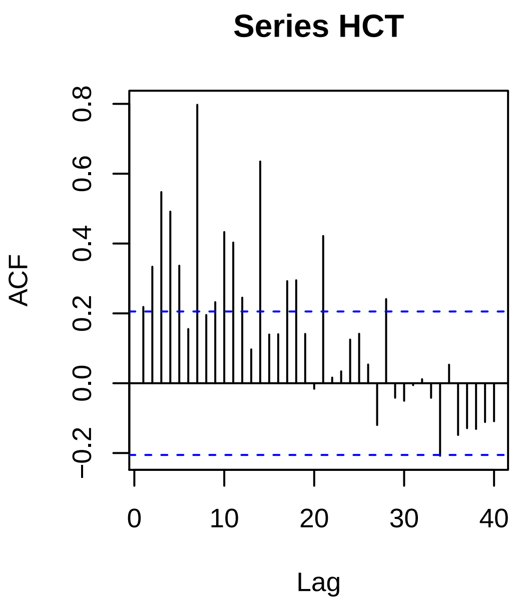
We start to analyse the dataset so2.

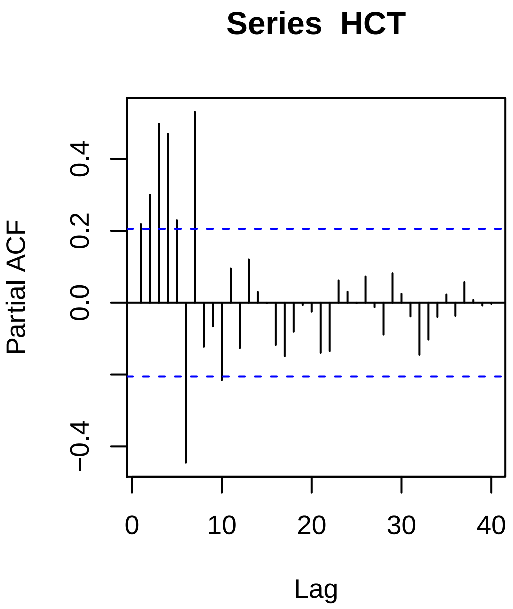


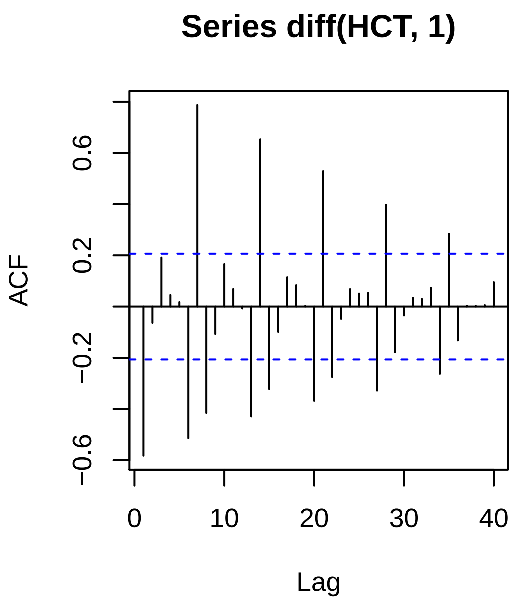
The x_t serie seems not to be stationary. We use ∇x_t. By looking at the ACF and PACF we don't se any season in the data, which means that $P = 0$, $Q = 0$ and $S = 0$. To decide $p = 0$ and $q = 0$ we look at the first spikes at the ACF and PACF. This part for this dataset is a little tricky. It may be so that a ARIMA(0,1,1) would be appropriate because the ACF got a decreasing pattern and the PACF got a large spike at lag 1. But the thing is that lag 2 and 3 are also significant.

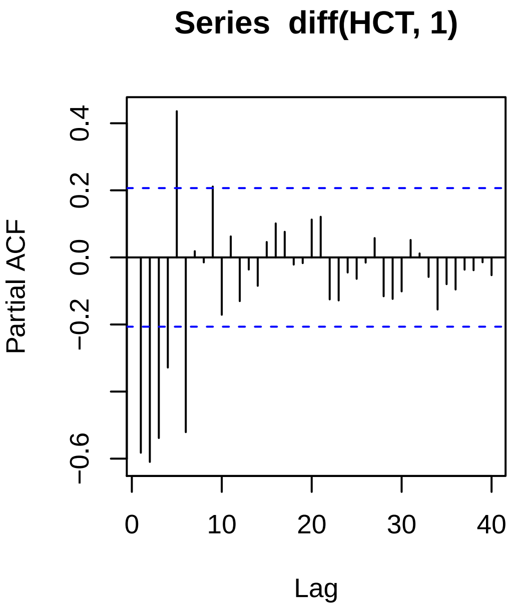


Just by looking at the ACF and PACF of the dataset of x_t it's hard to decide if the dataset is stationary or not. The first 8 spikes in the ACF deacreasing slowly and then just disappear. So a look at the dataset over time would be appropriate to decide if it's a good ide to make a model on this dataset. But say it's appropriate to fit a model to the data, then an ARIMA(1,0,0) probably would be a good model for the data. If we would come up to that the x_t is not stationary by looking at the data over time we would need to differentiate the series and use ∇x_t. It is a decreaseing pattern in the PACF and one big spike at lag 1. Which means that a ARIMA(0,1,1) would be a appropriate model.









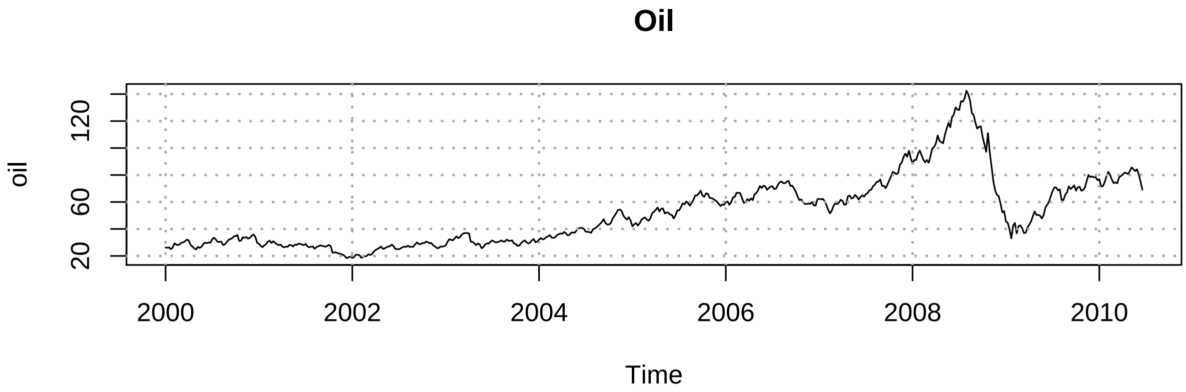
The x_t series dont have the pattern in ACF and PACF that we looking for, so we assume that the x_t is not stationary. We use ∇x_t insted and it is clear se that the dataset contains a season pattern by looking at the ACF where $S = 7$. But if we look at the PACF we don't se any spike at lag 7 which we should if we got a Θ in or model. This indicates that something probably is not right. We would recommend to plot the data over time to see that the time series looks stationary. But by just looking at the ACF and PACF we would recommend to start by trying the model SARIMA(0,1,2)(0,0,1)$_7$.

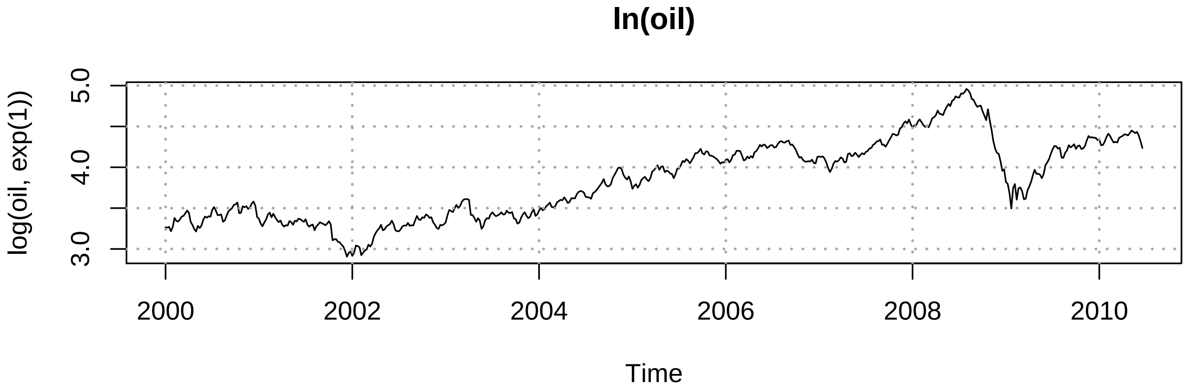
Assignment 3. ARIMA modeling cycle

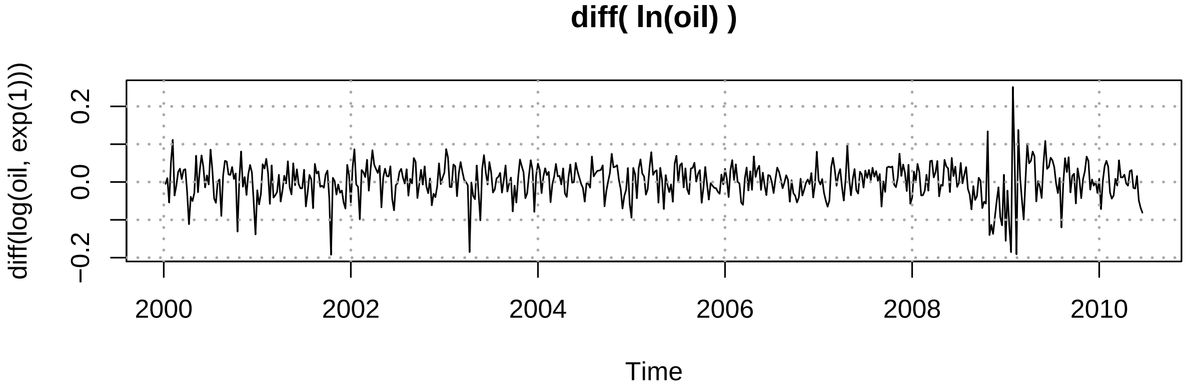
a)

In this assignment we will analys and build a model on the dataset oil in the package `astsa`.

We start by modelling the series to get it stationary.







To get rid of the the inconstant variance over time (first period have low vainace and the last period have higher variance) we do `ln` on the serie. After that we do the first difference $z_t = \nabla x_t = x_t - x_{t-1}$.

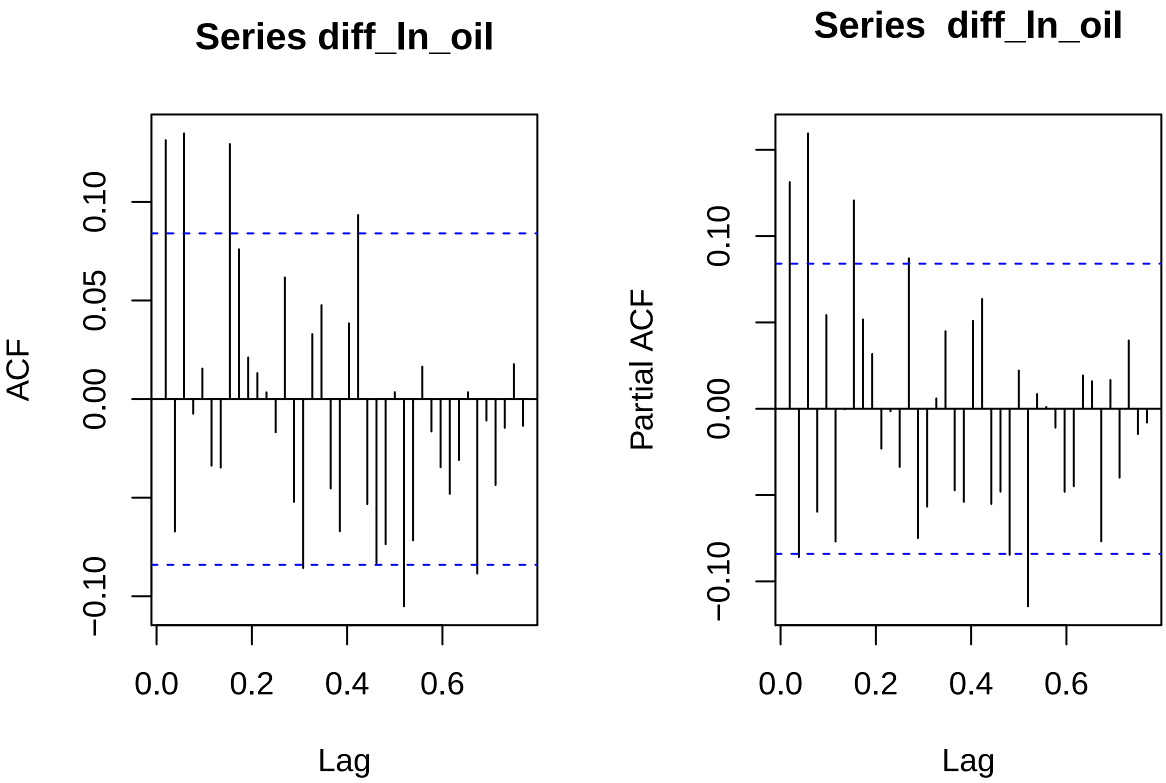
We also want to do the Dickey–Fuller Test to test if the z_t as a unit root.

```
adf.test(diff_ln_oil)
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  diff_ln_oil
## Dickey-Fuller = -6.3708, Lag order = 8, p-value = 0.01
## alternative hypothesis: stationary
```

The Dickey–Fuller Test get significant which indicates that the series (z_t) is stationary.

Now then we are satisfied with the z_t serie we want to determine which model that would fit the data best. To determine the model we look at ACF, PACF and EACF on the data.



```
## AR/MA
##   0 1 2 3 4 5 6 7 8 9 10 11 12 13
## 0 x o x o o o o x o o o  o  o  o
## 1 x o x o o o o x o o o  o  o  o
## 2 x x x o o o o x o o o  o  o  o
## 3 x x x o o o o x o o o  o  o  o
## 4 x o x o o o o x o o o  o  o  o
## 5 x x x o x o o x o o o  o  o  o
## 6 o x x o x x o x o o o  o  o  x
## 7 o x x x x x x x o x o  o  o  o
```

All the spikes is reasonably small in both the ACF and PACF. The biggest spikes in both graph is around 0.15, but it's still significant so we should apply a model on the data to get the best results.

We dont se any clear decreasing pattern in either the ACF or the PACF. So for determine the best model we look at the EACF. In the EACF we look for the first value of a triangele. We dont get a perfect triangle, but it's seems lika a ARIMA(0,1,1) or ARIMA(1,1,1) would fit our data best.

We will now fit these two models and check what model seems to be the best one by doing a residual analysis, looking at parameter significance, AIC and BIC.

ARIMA(0,1,1)

```
##          Estimate     SE t.value p.value
## ma1        0.1701 0.0499  3.4064  0.0007
## constant   0.0018 0.0023  0.7578  0.4489
```

```
##       AIC
## -5.131614
```

```
##       BIC
## -6.115831
```

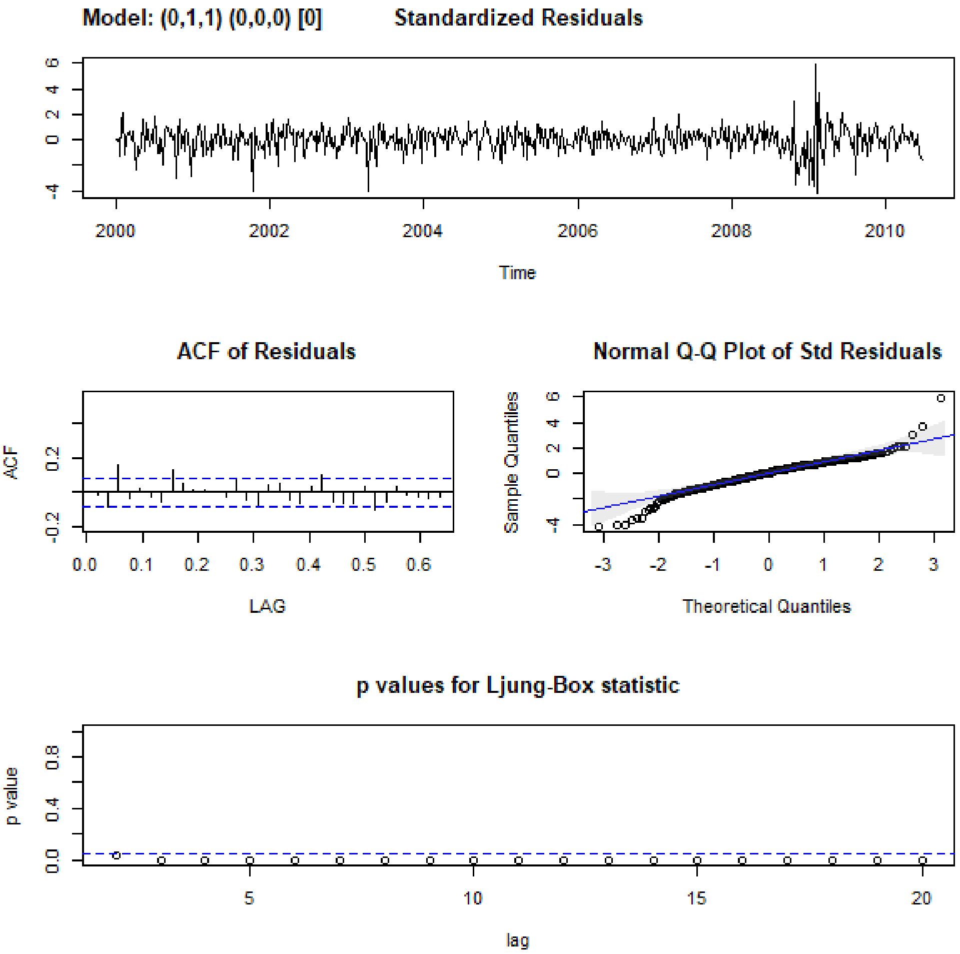


Figure 1:

ARIMA(1,1,1)

```
##          Estimate     SE t.value p.value
## ar1       -0.5264 0.0871 -6.0422  0.0000
## ma1        0.7146 0.0683 10.4699  0.0000
## constant   0.0018 0.0022  0.7981  0.4252
```

```
##       AIC
## -5.153858
```

```
##       BIC
## -6.130184
```

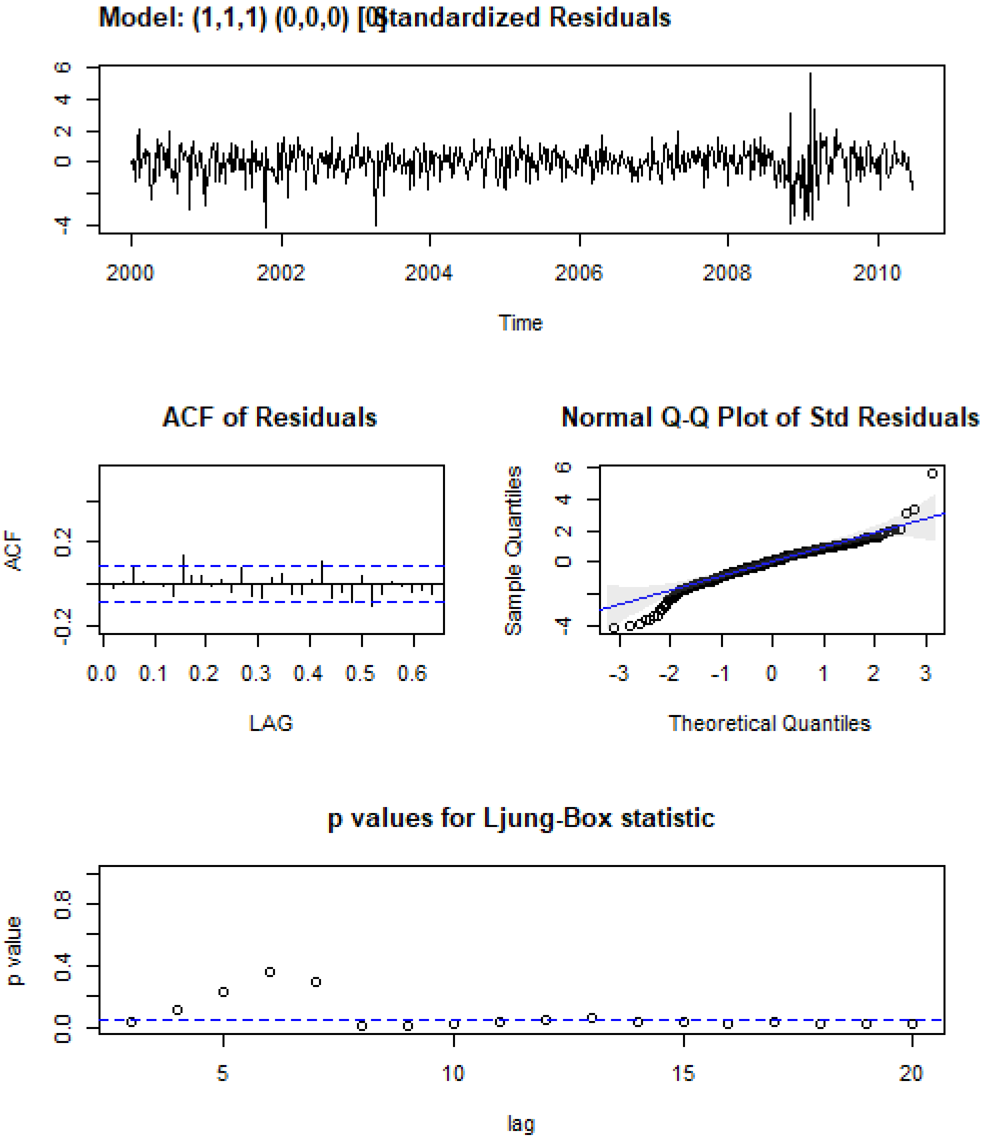


Figure 2:

The first thing we look at is that if the models have significant AR and MA parameters, which they have. Then we look if there is any autocorrelation in the residuals by lokking at the ACF. In the model ARIMA(0,1,1) we can se a samll spike at lag 3 that is significant. This make the Ljung-Box test significant on all lags which we can see in the graph and it means that ther still is some information left in the residuals. For the model ARIMA(1,1,1) is the Ljung-Box not significant up to lag 7. The Q-Q plot from both models looks similar. Both tails in the plot seems to tail of, but it is expected to have some outliers in data.

If we compare the models AIC and BIC with each other we see that the model ARIMA(1,1,1) got the smallest values, but the difference is not huge.

From the analysis, the model ARIMA(1,1,1) seems to be the most fitting one for the data.

The model ARIMA(1,1,1):

$\phi(B)(z_t - \mu_z) = \theta(B)w_t$

$(1 - \phi B)(z_t - \mu_z) = (1 + \theta B)w_t$

$y_t = z_t - \mu_z$

$y_t - \phi y_{t-1} = w_t + \theta w_{t-1}$

$y_t = \phi y_{t-1} + w_t + \theta w_{t-1}$

$\phi = -0.53$ and $\theta = -0.71$

$y_t = -0.53 y_{t-1} + w_t + -0.71 w_{t-1}$

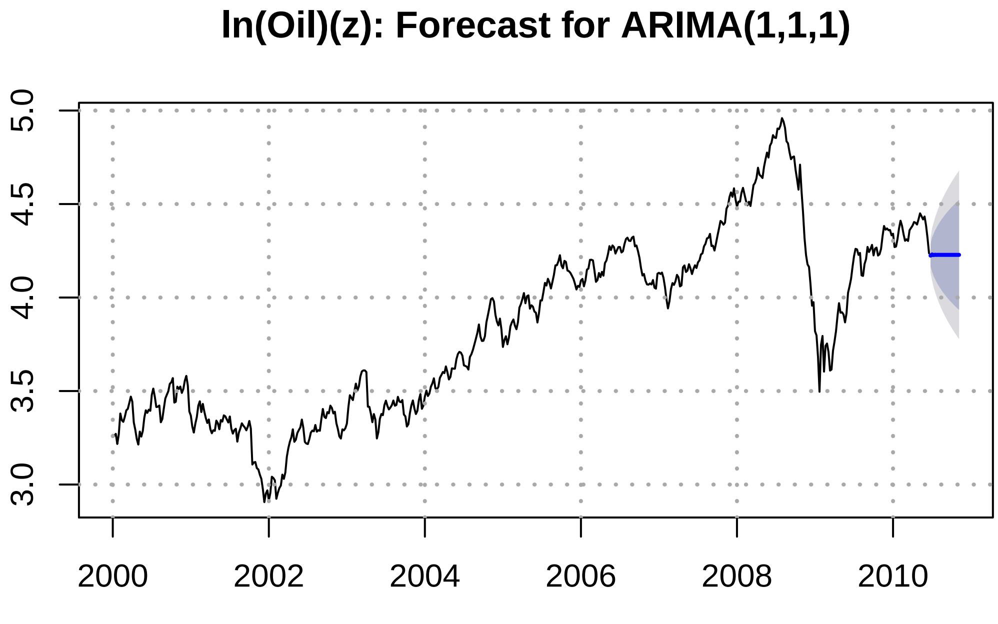
$z_t - \mu_z = -0.53(z_{t-1} - \mu_z) + w_t + -0.71 w_{t-1}$

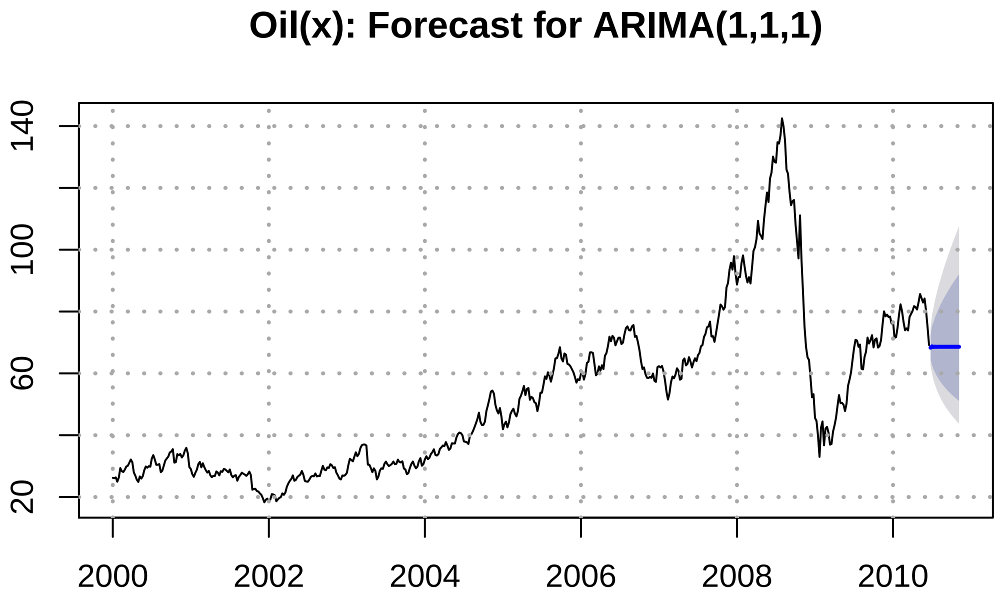
$z_t = \mu_z + -0.53(z_{t-1} - \mu_z) + w_t + -0.71 w_{t-1}$

$z_t = ln(x_t)$

$x_t = e^{\mu_z + -0.53(z_{t-1} - \mu_z) + w_t + -0.71 w_{t-1}}$

Now we also want to do a forcast 20 observations ahead.



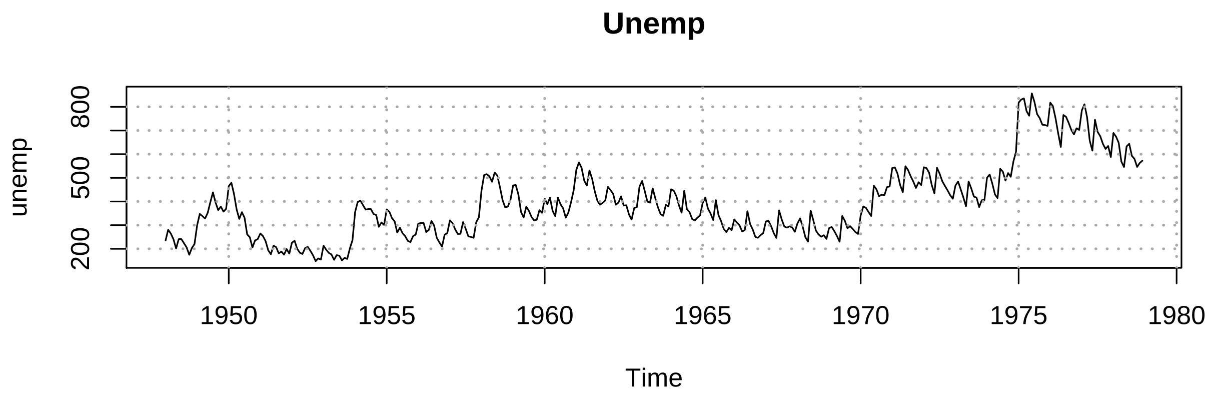


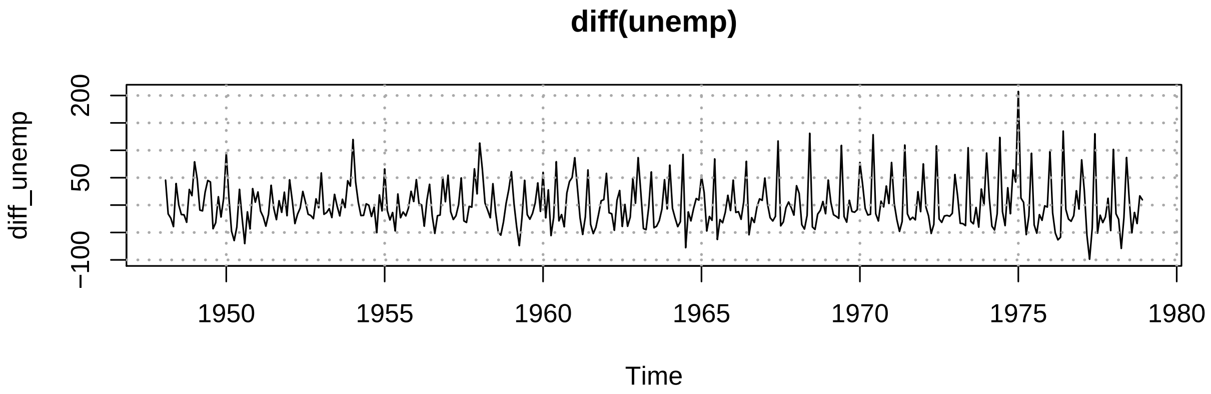
In the graph we see forcast for Oil 20 observations ahead.

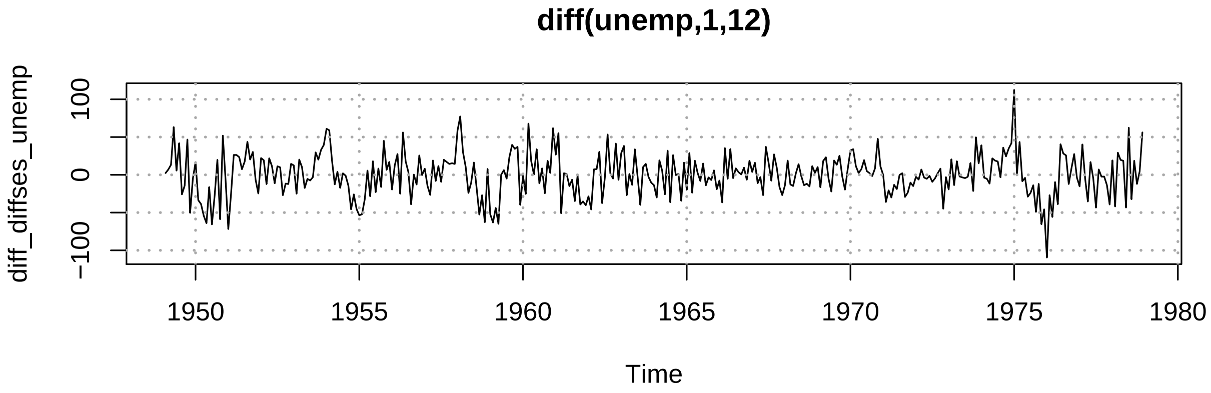
b)

In this assignment we will analys and build a model on the dataset `unemp` in the package astsa.

We start by modeling the serie to get it stationary.







To get the time series stationary we do the first difference of the time series. It's still a season pattern with a slowly decreasing pattern i the PACF. So we need to difference the time series by lag 1 and 12. Then we get the time series:

$z_t = x_t - x_{t-1} - x_{t-12} + x_{t-13} = (1 - B - B^{12} + B^{12})x_t = (1 - B^{12})(1 - B)x_t = \nabla_{12}\nabla_{x_t}$

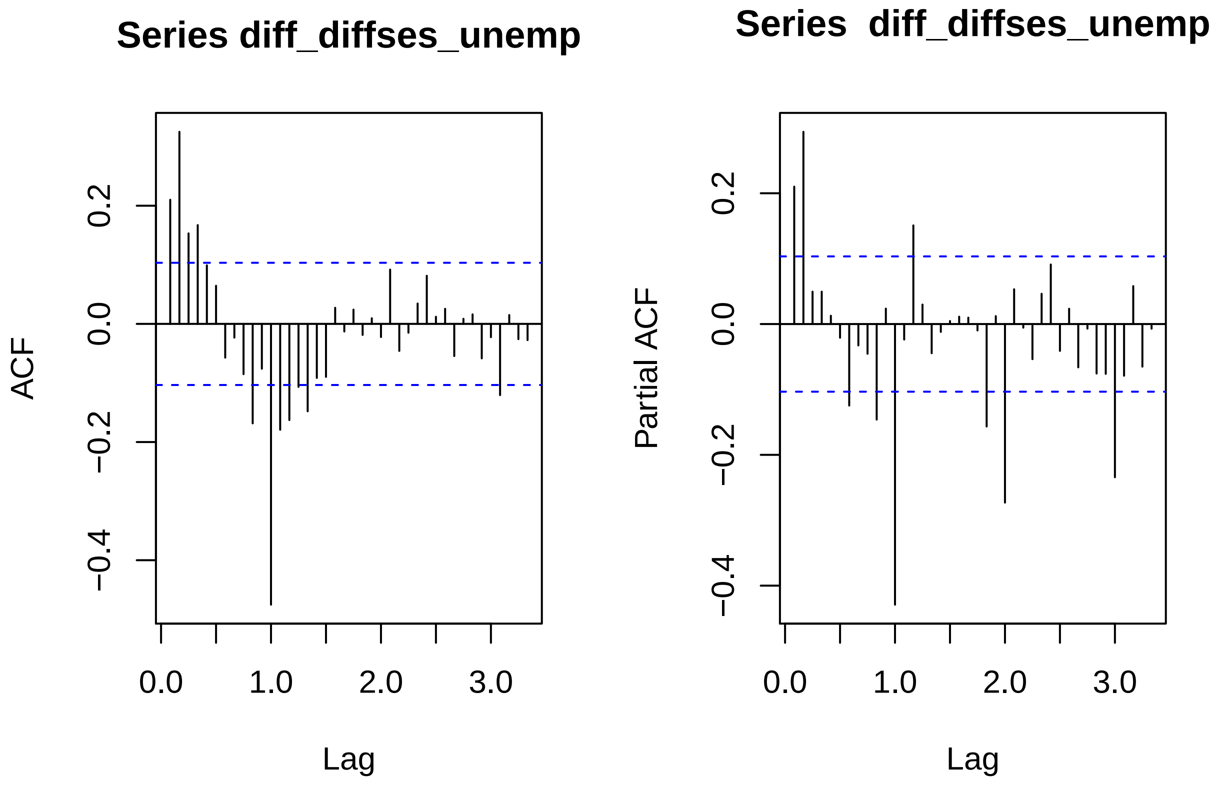
We also want to do the Dickey–Fuller Test to test if the z_t as a unit root.

```
adf.test(diff_diffses_unemp)
```

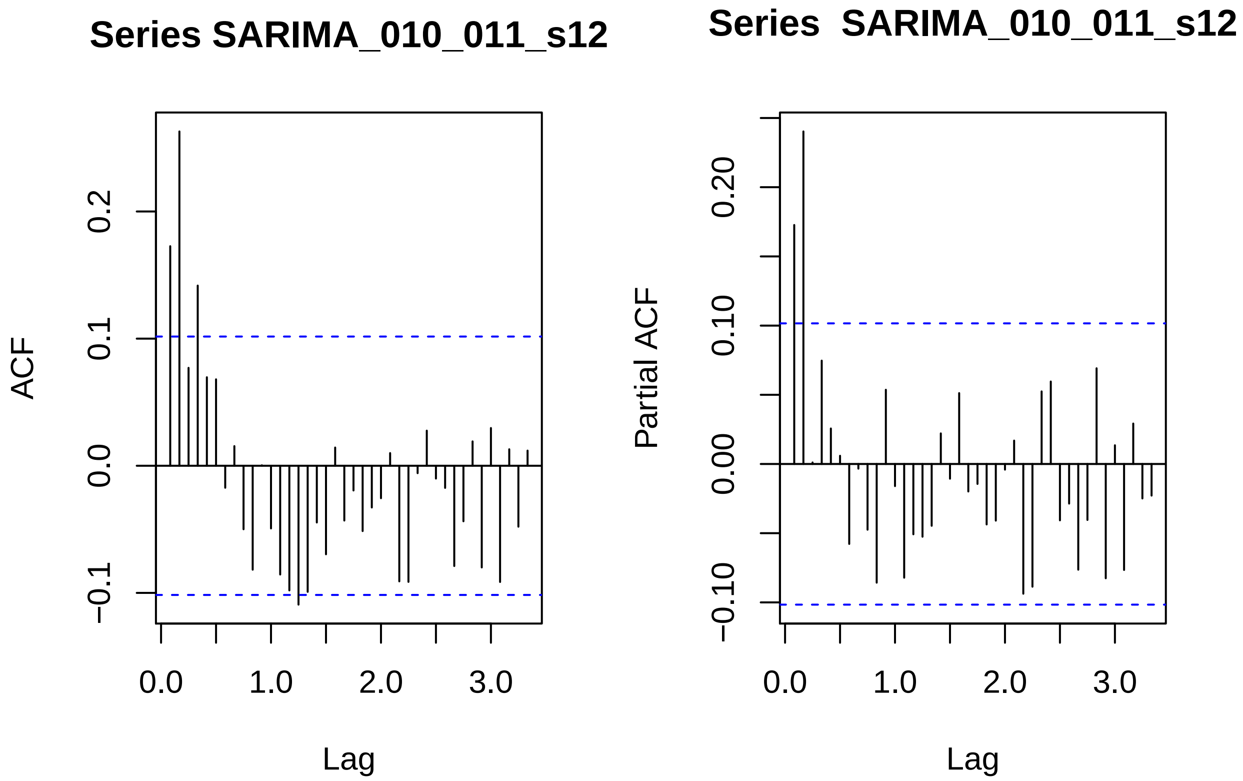
```
##
##  Augmented Dickey-Fuller Test
##
## data:  diff_diffses_unemp
## Dickey-Fuller = -6.171, Lag order = 7, p-value = 0.01
## alternative hypothesis: stationary
```

The Dickey–Fuller Test get significant which indicants that the series (z_t) is stationary.

Now then we are satisfied with the z_t series we want to determine which model that would fit the data best. To determine the model we look at ACF and PACF.



We see a decreasing seasonal pattern in PACF and that the 12:th lag have a spike. That means that we can start to fit a SARIMA(0,1,0)(0,1,1)$_{12}$ and then look at a ACF, PACF and EACF of the residuals for determine p and q in the model.



```
## AR/MA
##   0 1 2 3 4 5 6 7 8 9 10 11 12 13
## 0 x x o x o o o o o o o  o  o  o
## 1 x x x o o o o o o o o  o  o  o
## 2 o x o o o o o o o o o  o  o  o
## 3 o x x o o o o o o o o  o  o  o
## 4 x o o x o o o o o o o  o  o  o
## 5 x o x o x o o o o o o  o  o  o
## 6 o x x o x o o o o o o  o  o  o
## 7 o x x x x o o o o o o  o  o  o
```

The ACF got a decreasing pattern and at lag 1 and 2 in the PACF is it spikes. That means that $p = 2$ and $q = 0$ which gives us the model SARIMA(2,1,0)(0,1,1)$_{12}$.

But we also want to check if SARIMA(0,1,2)(0,1,1)$_{12}$ would be a good model because we can see a triangle pattern in the EACF.

We will now fit these two models and check what model seems to be the best one by doing a residual analysis, looking at parameter significans, AIC and BIC.

SARIMA(2,1,0)(0,1,1)$_{12}$

```
##      Estimate     SE  t.value p.value
## ar1    0.1351 0.0513   2.6326  0.0088
## ar2    0.2464 0.0515   4.7795  0.0000
## sma1  -0.6953 0.0381 -18.2362  0.0000
```

```
##     AIC
## 7.12457
##       BIC
## 6.156174
```

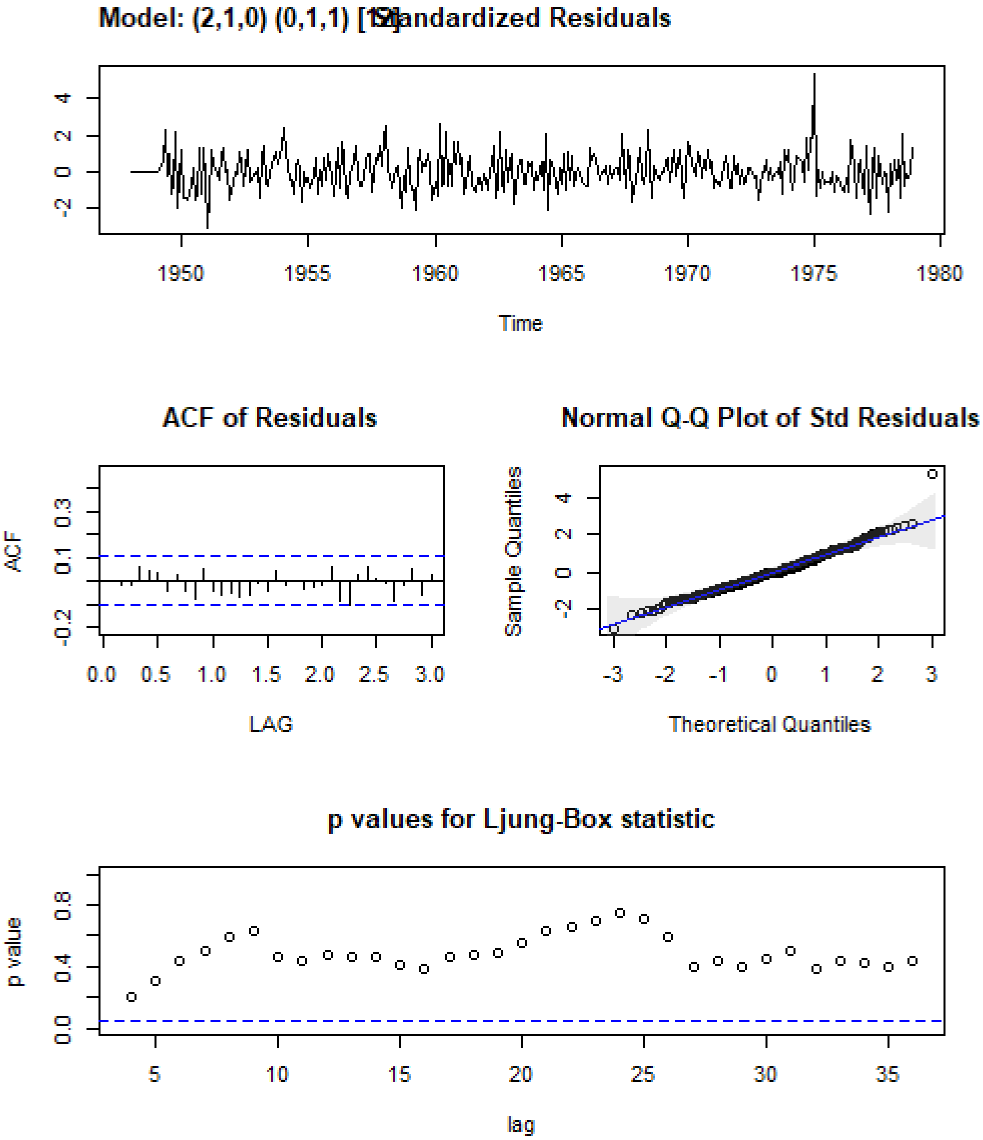


Figure 3:

SARIMA(0,1,2)(0,1,1)$_{12}$

```
##      Estimate     SE  t.value p.value
## ma1    0.1381 0.0525   2.6294  0.0089
## ma2    0.2257 0.0477   4.7369  0.0000
## sma1  -0.7054 0.0373 -18.9020  0.0000
```

```
##      AIC
## 7.136909
```

```
##      BIC
## 6.168513
```

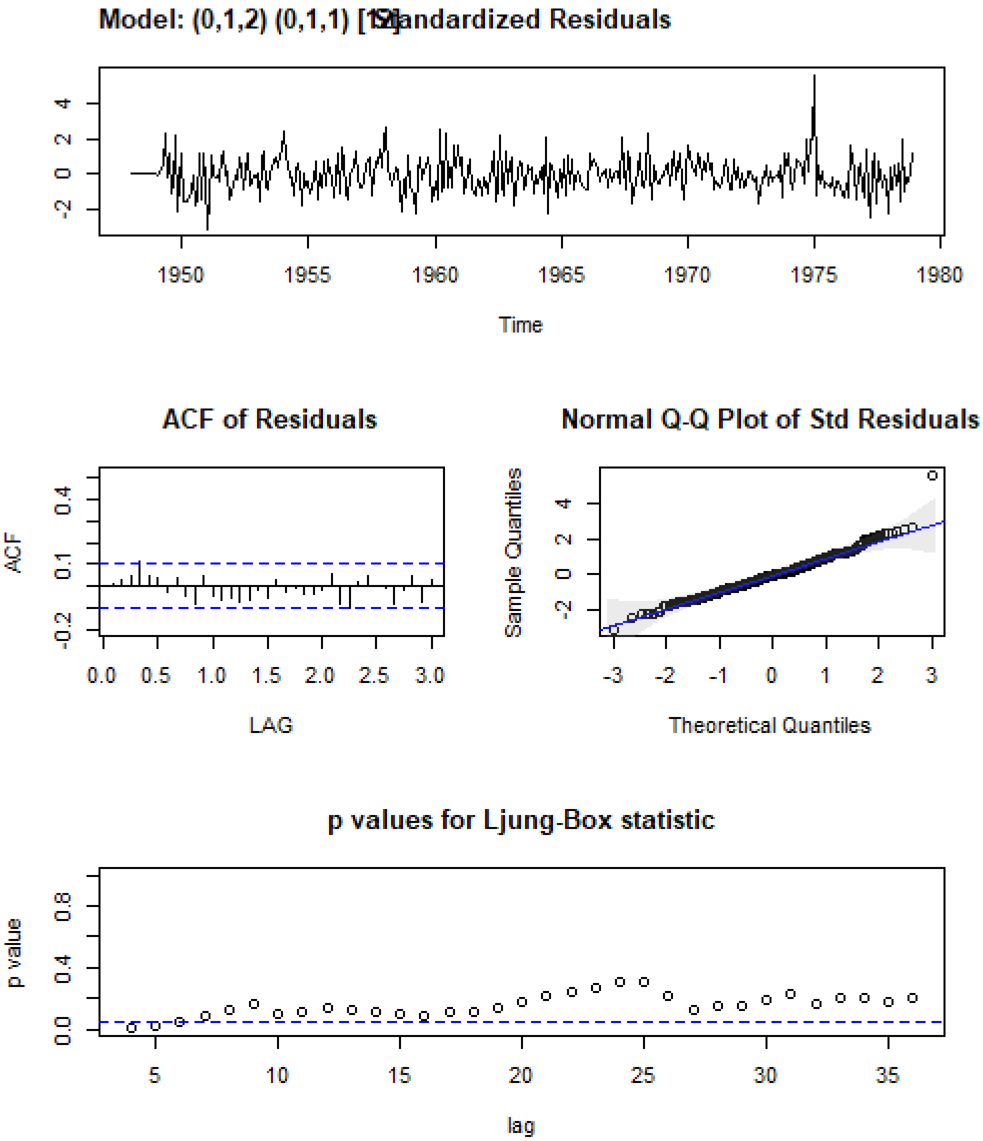


Figure 4:

The first thing we look at is that if the models have significant AR and MA parameters, which they have. Then we look if there is any autocorrelation in the residuals by looking at the ACF and the Ljung-Box test.

For the model SARIMA(0,1,2)(0,1,1)$_{12}$ lag 3 in the ACF got a significant spike. The Q-Q plot from both models looks similar and good. We can see that we maybe have 2 outliers.

If we compare the models AIC and BIC with each other we see that the model SARIMA(2,1,0)(0,1,1)$_{12}$ got the smallets values. But the difference is not much.

With the analysis the selected model is SARIMA(2,1,0)(0,1,1)$_{12}$.

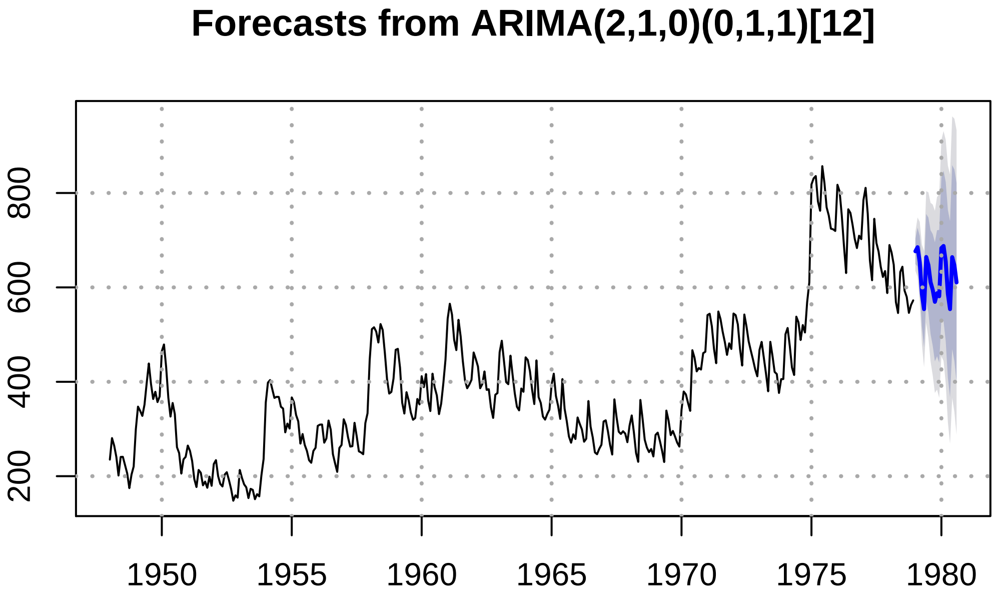
The model SARIMA(2,1,0)(0,1,1)$_{12}$:

$\phi(B)\nabla_s^1\nabla^1(x_t - \mu_x) = \Theta(B^s)w_t$

$(1 - \phi_1 B - \phi_2 B^2)(1 - B^{12})(1 - B)(x_t - \mu_x) = (1 + \Theta B^{12})w_t$

$(1 - 0.14B - 0.25B^2)(1 - B^{12})(1 - B)(x_t - \mu_x) = (1 - 0.70B^{12})w_t$

Now we also want to do a forcast 20 observations ahead.



In the graph we see forcast for `unemp` 20 observations ahead.