STK4060 Vår 2020

Oblig 1

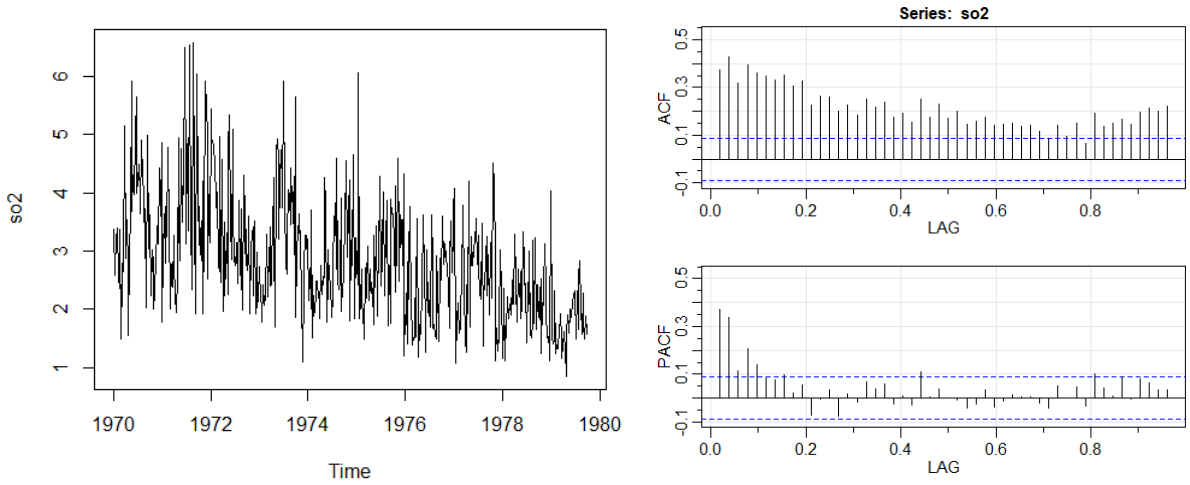
He Gu

Oppgave a)

We could simply plot the figures by running the following codes. It is natural to set the maximum lag as 50, and although there are many other choices, the difference between them shall not change our conclusion.

```
require(astsa)
plot(so2)
acf2(so2,50)
```

And then, the result should be



Obviously, according to the first plot, we see that there exists a "trend" inside the plot, and therefore, the time series does not seem to be stationary. Besides, recall that the (weak) stationary means that

$$\begin{cases} \mu \text{ is a constant} \\ \text{covariance function } \gamma(s,t) \text{ depends on } s,t \text{ only through the difference } |s-t| \end{cases}$$

And the plot does not satisfy the "μ *is a constant*" condition.

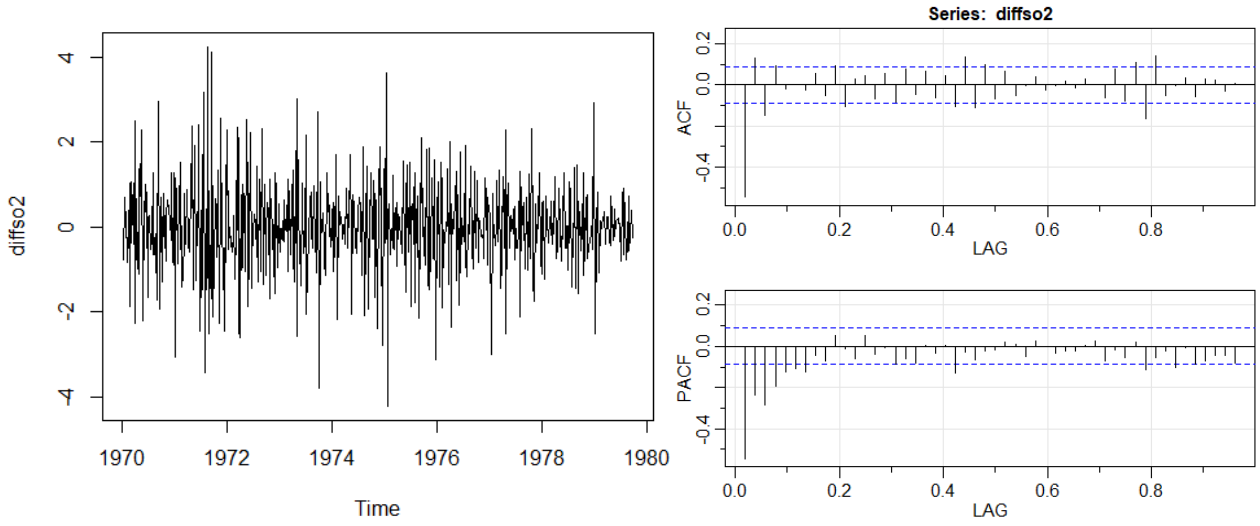
Besides, if we take a look at the ACF plot, the ACF value does not seem to be converging to 0, which also indicates that the time series is NOT stationary.

Oppgave b)

By using a similar codes from above,

```
1. diffso2 = diff(so2)
2. plot(diffso2)
3. acf2(diffso2,50)
```

we see that



Obviously, the "trend" is eliminated to some degree if compared with the plot before. And the ACF plot now seems to be converging to 0. Recall the definition of (weak) stationary, the difference of the series seems to be stationary.

ARGUE THAT DIFFERENCING IS REASONABLE:

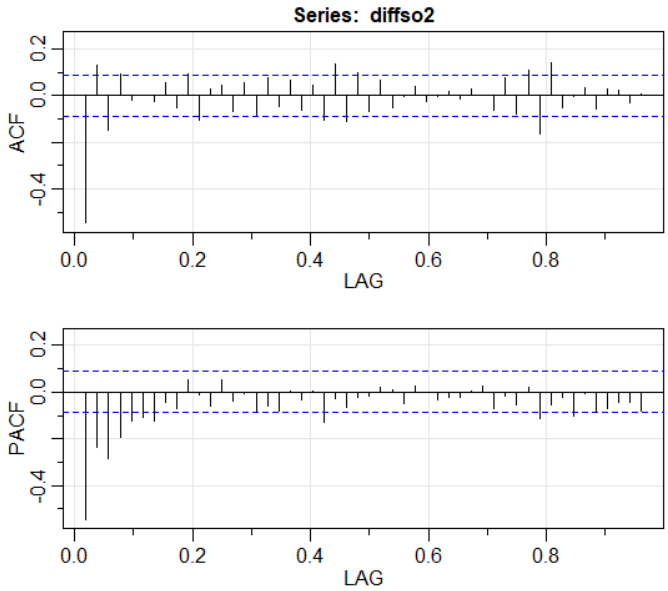
We want to let the original unstationary time series be stationary, which means that we aim at eliminating the "trend" of the original series. And therefore, in such case, a differencing is natural.

Besides, we see that part of the "trend" is linear, and thus, we are using the first-order differencing, i.e. $y_t = so2_t - so2_{t-1}$.

The plots above show that the first differences seems to be stationary. But if the first differences would not be stationary, we could still try a second-order differencing or an even higher one.

Oppgave c)

Recall the ACF and PACF plots we have produced before.



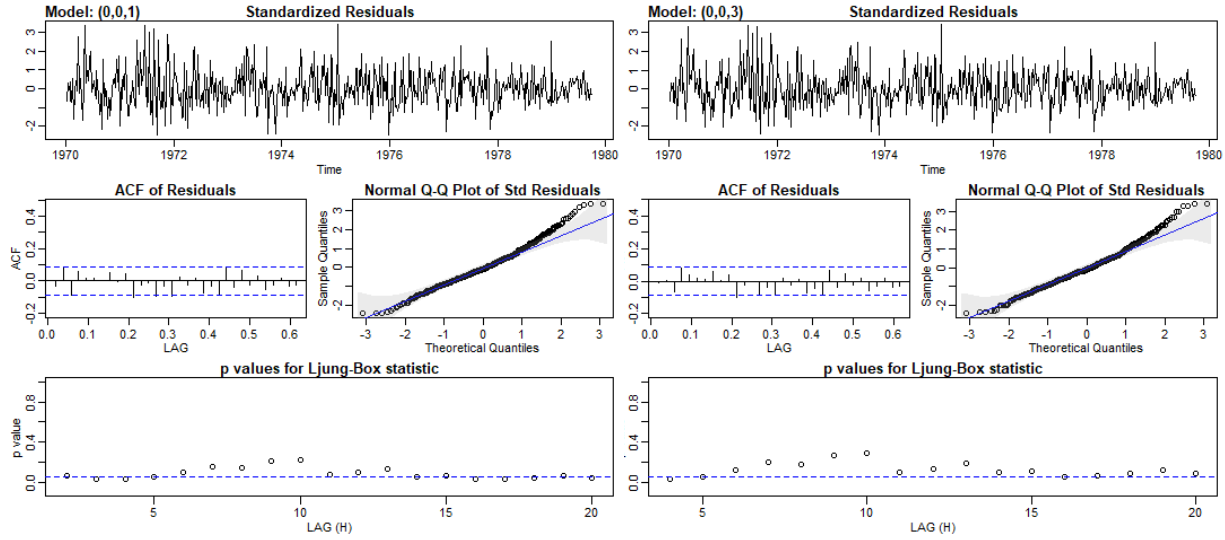
To be honest, we cannot precisely justify whether they are “tailing off” or “cutting off”, but obviously the plots ARE ABLE TO BE regarded as that the ACF cuts off after lag 1 or maybe lag 3, while the PACF tails off. However, at the same time, we may also consider that the PACF plot cuts off after lag 7 while ACF tails off. We could simply try all possible hypotheses and just compare their performance, e.g. AIC, BIC and etc. But since this question does not ask us to do so, we are not going to really accomplish this method.

Anyway, we could conclude that it is reasonable to consider the plots as that the ACF cuts off after lag 1 or maybe lag 3, while the PACF tails off. And thus, the corresponding models, i.e. $MA(1)$ and $MA(3)$, are plausible.

We could fit the model by running following codes

```
1.  fit1 = sarima(diffso2, 0, 0, 1)
2.  fit2 = sarima(diffso2, 0, 0, 3)
```

and get the following results(diagnostics).



COMMENTS:

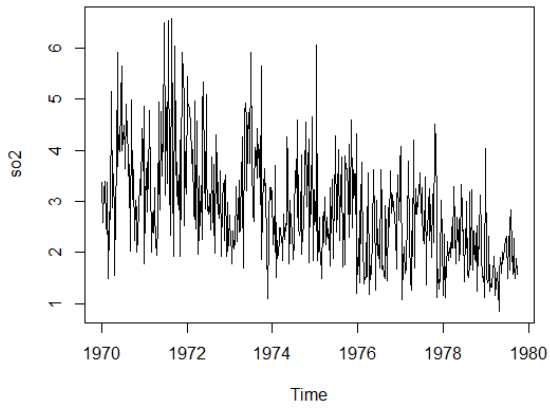
First of all, we see that the plots of **standardized residuals** have no obvious pattern and only a few points exceed the 3 standard deviations, which means that there is no obvious departure from the assumptions of our models.

Both **ACF** plots perform like a white noise, and therefore, the independent assumption seems to be suitable. And besides, both **Normal Q-Q plots** shows that the distribution of residuals is likely to be normal, since there are only a few extreme points in the tails.

But as for the ***p*-values of Ljung-Box statistic**, it seems that both plots have several LAGs which have a small p-value and may reject the null hypothesis. Recall that the null hypothesis of Ljung-Box test is "the model does not exhibit lack of fit", therefore, the plots indicate lack of fit for both models, although $MA(3)$ seems better from such aspect.

Oppgave d)

Recall the original plot of $so2$.



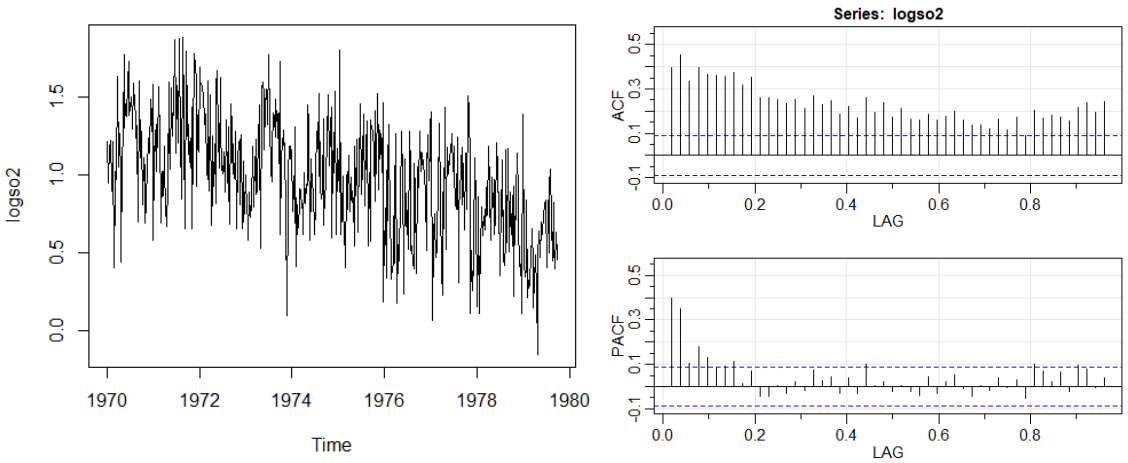
We could easily find out that the variation seems to decrease over time, which indicates that the nonstationarity may not simply be a "linear trend". Thus, in order to remove such a nonstationarity, it is natural to use the log transformation, since the log transformation shall not only transfer an exponential trend into a linear trend, but would also be useful when we are aiming at eliminating the nonstationarity as a function of time. And besides, the log transformation could make the time series more "stationary" without changing the statistical properties between variables, which could be helpful for further study.

REPEAT OF OPPGAVE A)

With running the following codes,

```
1. logso2 = log(so2)
2. plot(logso2)
3. acf2(logso2,50)
```

we see that



Using a similar thought as in oppgave a), this time series does not seem to be stationary. First of all, there is a "trend" inside the plot, and besides, the ACF plot seems not to be converging to 0. And in addition, according to the definition of (weak) stationary,

$$\begin{cases} \mu \ is \ a \ constant \\ covariance \ function \ \gamma(s,t) \ depends \ on \ s,t \ only \ through \ the \ difference \ |s-t| \end{cases}$$

we see that, obviously, the series above does not satisfy the first condition, i.e. $\mu \ is \ a \ constant$, and therefore is **NOT** stationary.

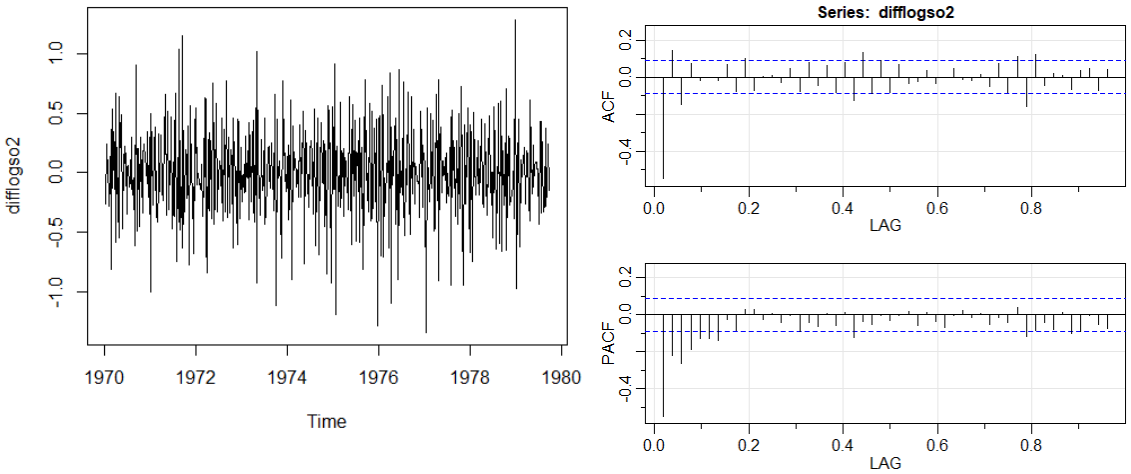
But if we compare this plot with the original $so2$ plot, it is clearly improved with respect to stationarity. Although there still exists a "trend", but the "thicknesses" of the left side and the right side is already "balanced" to some degree.

REPEAT OF OPPGAVE B)

Running the following codes,

```
1. difflogso2 = diff(log(so2))
2. plot(difflogso2)
3. acf2(difflogso2,50)
```

we see that



Compared with the $\log(so2)$ plot before, it seems that the "trend" has been removed to some degree. And if we compare it with the $diff(so2)$ plot, it is clear that there is some improvement, since the $diff(\log(so2))$ seems to be more "balanced" with respect to the "thickness" of the whole plot. And besides, as for the ACF plot, now it seems to be converging to 0.

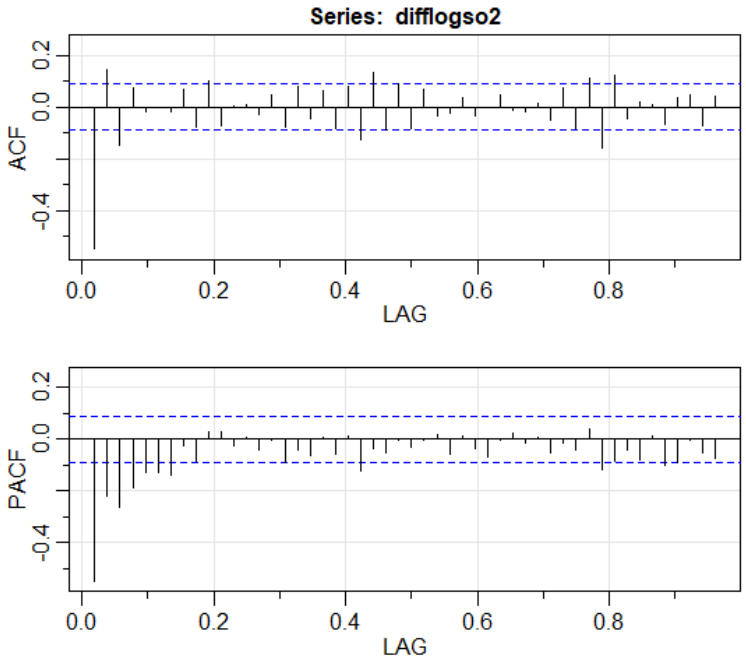
Thus, according to the definition of (weak) stationary, the difference of the series seems to be stationary.

ARGUE THAT DIFFERENCING IS REASONABLE:

Anyway, we want to let the unstationary time series, i.e. $\log(so2)$, be stationary, which means that, in such case, we are going to eliminate the "trend" of $\log(so2)$. And therefore, it is natural to try the differencing. Besides, we see that the "trend" of $\log(so2)$ seems to be "linear", and thus, we are using the first-order differencing, i.e. $y_t = \log(so2_t) - \log(so2_{t-1})$.

REPEAT OF OPPGAVE C)

We know that the ACF plot and the PACF plot of $\log(so2)$ should be



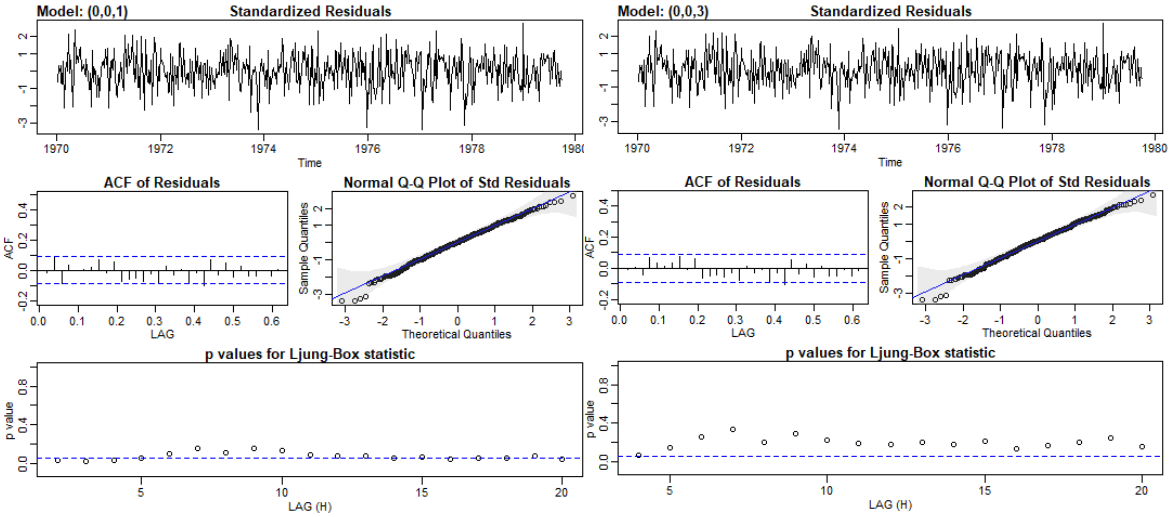
In such case, it is hard to be 100% confident about whether they are tailing off or cutting off, but obviously, it is reasonable to regard the situation as that ACF plot cuts off after lag 1 or maybe lag 3, while PACF plot tails off. To be honest, there could be other types of hypothesis, e.g. both plots cut off after several lag. In reality, we could just try all possible models and choose the best according to AIC, BIC, and etc. But this time, the assignment does not ask us to do so, and therefore we are not going to really achieve it.

Anyway, according to the plot above, it is reasonable to consider the situation as that ACF cuts off after lag 1 or 3, while PACF tails off, and thus, the corresponding models, i.e. $MA(1)$ and $MA(3)$, are plausible.

Then, by running the following codes,

```
1. fit3 = sarima(difflogso2, 0, 0, 1)
2. fit4 = sarima(difflogso2, 0, 0, 3)
```

we see that



COMMENTS:

At first, we shall take a look at the **standardized residuals**. Apparently, none of these two plots shows any departure from our model assumption, since only a few of points exceed the 3 standard deviations, i.e. -3 and +3.

As for the **ACF** plots, we see that the autocorrelation of all the LAGs are located inside the confidence interval, which means that both residual series perform like a white noise, and therefore, the independent assumption holds.

Besides, recall the results we got based on $diff(so2)$, we will notice that the Normal Q-Q plots based on $\log(diff(so2))$ perform better than the plots based on $diff(so2)$. Although the Normal Q-Q plot based on $diff(so2)$ also indicates a normality, the plots based on the log-transformed series just have much fewer extreme points in the tail, and are more likely to be normal.

In addition, according to the definition of Ljung-Box test, the null hypothesis shall be "the model does not exhibit lack of fit". Therefore, the LAGs with a significant p-value shall reject this

hypothesis and indicates that "the model exhibits lack of fit", and thus, comparing the number of LAGs which has a significant p-value, we may consider that $MA(1)$ exhibit lack of fit for pretty many LAGs, while $MA(3)$ performs pretty well and exhibit lack of fit only for one LAG. And recall the model we have shown before in oppgave c), we see that, from such aspect, $MA(1)$ based on $\log(diff(so2))$ performs worse than the model based on $diff(so2)$, while $MA(3)$ based on $\log(diff(so2))$ performs better than the model based on $diff(so2)$.

Oppgave e)

According to the plots we have produced in oppgave c), we see that there is no large difference between $MA(1)$ and $MA(3)$ based on $diff(so2)$. And according to oppgave d), we see that there is no large difference between $MA(1)$ and $MA(3)$ based on $\log(diff(so2))$, either.

But recall the comparison we have made in oppgave d), we see that the log transformation seems better than the original series, i.e. $diff(so2)$, since the log transformation deal with the "trend" in a better and more precise way. And thus, our choice shall be one of the models based on log transformation.

It seems that $MA(3)$ performs better than $MA(1)$ with respect to those statistics and plots we have shown and discussed before, but we know that the best-fitted model might not be the best model overall, since there might be some overfitting issues and so on, and therefore, I shall choose the $MA(1)$ model with log transformed seires, i.e. $\log(diff(so2))$.

In order to justify this choice, the best way might be an independent test dataset, but obviously it is not provided. And thus, we are going to use AIC and BIC to compare all the four models. The model with a lower AIC/BIC shall be considered as better according to their definition.

Running the following codes,

```
1. c(fit1$AIC,fit1$BIC)
2. c(fit2$AIC,fit2$BIC)
3. c(fit3$AIC,fit3$BIC)
4. c(fit4$AIC,fit4$BIC)
```

we see that

```
> c(fit1$AIC,fit1$BIC)
[1] 2.609604 2.634625
> c(fit2$AIC,fit2$BIC)
[1] 2.610519 2.652220
> c(fit3$AIC,fit3$BIC)
[1] 0.4989322 0.5239530
> c(fit4$AIC,fit4$BIC)
[1] 0.4949513 0.5366526
```

It indicates that $MA(1)$ and $MA(3)$ based on log transformation shall be considered as better than the models based on $diff(so2)$. But as for the comparison inside them, we see that $MA(1)$ has a higher AIC but a lower BIC than $MA(3)$, since BIC is more likely to choose a sparser model than AIC.

Now, it seems that any choice between $MA(1)$ and $MA(3)$ based on log transformation shall be plausible. And it means that if one wants a sparser model, then he/she shall choose $MA(1)$, and if one wants a best-fitted model, then he/she shall choose $MA(3)$.

As for me, I go to $MA(1)$ (based on log transformation) in such case.

Good

The report is thorough and correct. Just note what is tested by the Ljung-Box test. Another minor comment is that it would have been better to put the R code at the end of the report, and reported the results in the text or in tables instead.

_____end_____