Spring Semester 2023–2024

MATH3026/4022: Time Series Analysis/Time Series and Forecasting
Computer Practical 3 – SOLUTIONS

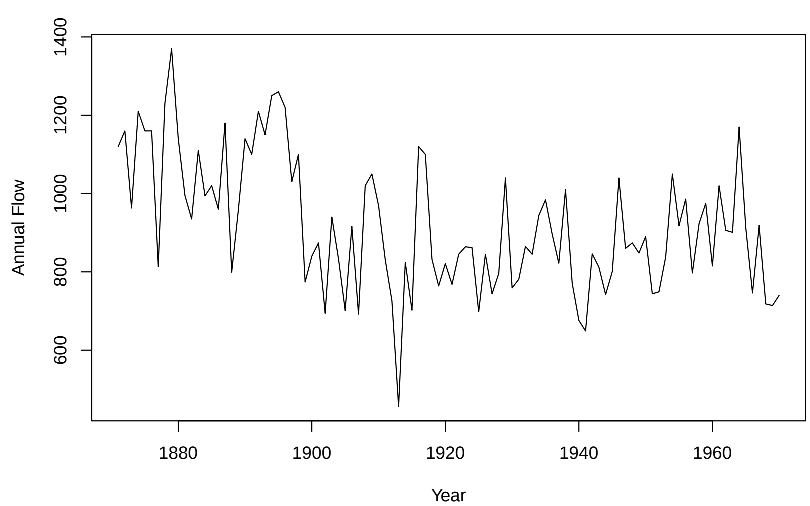
During this practical, we will use R by loading RStudio and running commands. When writing R commands, it is good practice to write comments so that you understand your code.

Before beginning the practical, it is recommended that you create a folder named `MATH3026` or `MATH4022` within your `Documents` folder on your University workspace. Then, within your `MATH3026` or `MATH4022` folder, create a new folder called `Practical3` in which you can save all of the material covered during this practical session.

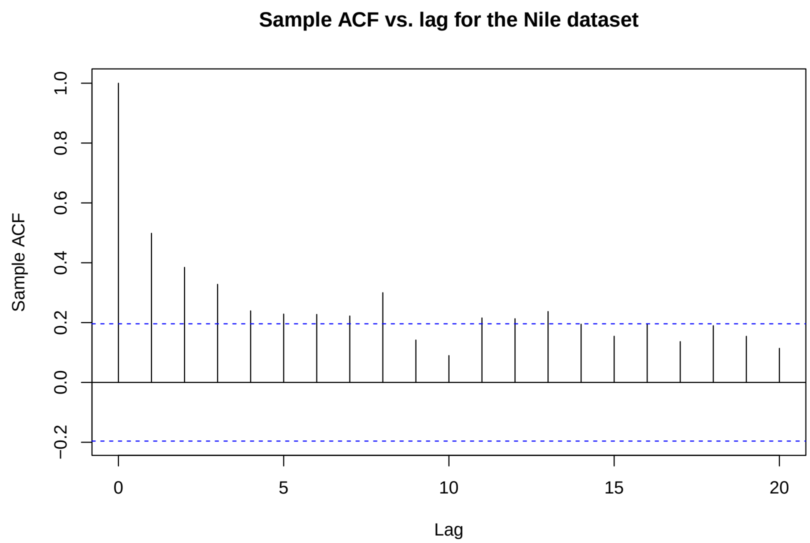
You should set your working directory to be the `Practical3` folder that you've created. In RStudio, this can be done by selecting the `Session` menu and then `Set Working Directory` and `Choose Directory`, navigating to the correct folder. You can then easily save your R code and plots in this directory.

1. (b) **SOLUTION:**

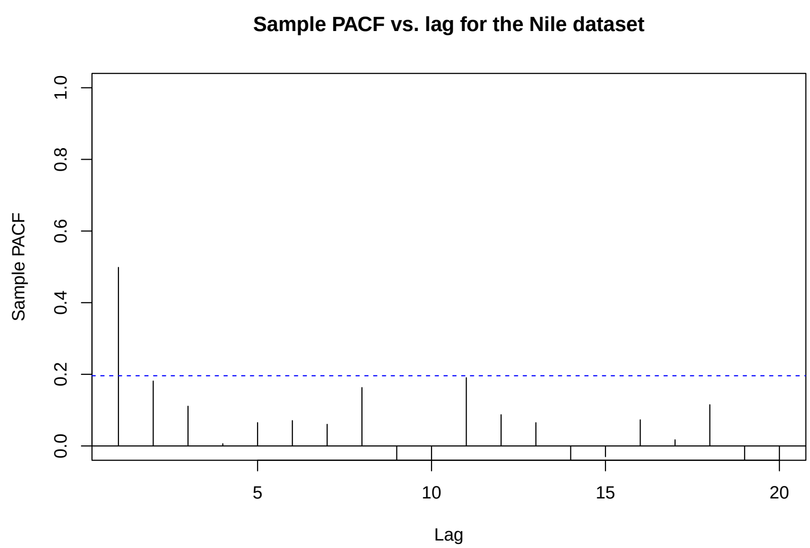
A time plot of the `Nile` dataset is shown below.



A plot of the sample ACF against the lag is shown below.



A plot of the sample PACF against the lag is shown below.



Looking at the time plot, the mean of the series appears higher between 1871–1895 compared to later time periods (for example 1920–1945). Although the plot of the sample ACF against the lag decreases initially as the lag increases, the decline to zero is not very rapid (for example, the ACF peak height at lag 12 is similar to that at lag 4) and so it is not certain that the series is stationary. The sample PACF plot provides less information about whether or not the series is stationary compared to the time plot and sample ACF plot.

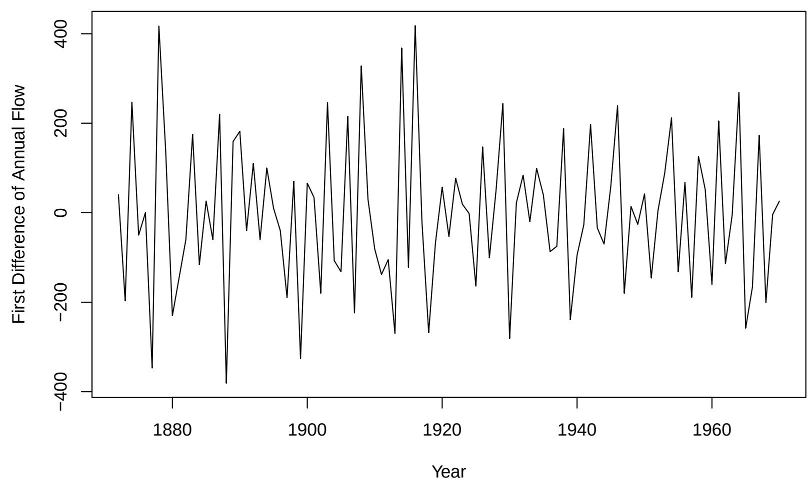
(c) Now run the following command to define the time series `Nile_diff` as the first difference of the time series `Nile`:

```
Nile_diff<-diff(Nile)
```

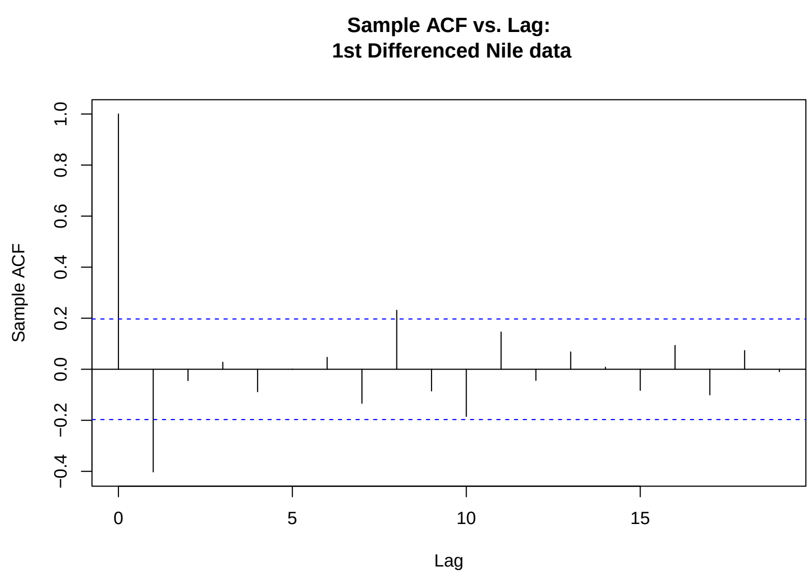
Produce a time plot and plots of the sample ACF and PACF against the lag for the `Nile_diff` time series. Does the time series appear to be stationary?

SOLUTION:

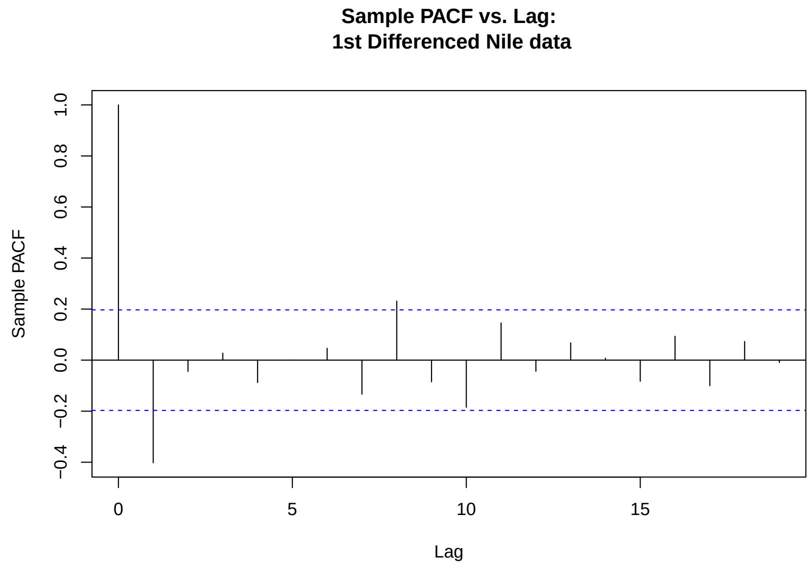
A time plot of the first-differenced `Nile` dataset is shown below.



A plot of the sample ACF against the lag is shown below.



A plot of the sample PACF against the lag is shown below.



In contrast to the plots in 1(b), these plots appear to show that the first-difference of the data is (weakly) stationary. The time plot has a constant mean (= 0) and appears to show constant variability over time. The sample ACF declines rapidly to zero as the lag increases (although there are a couple of spikes close to $\pm 2/\sqrt{n}$, there is not systematic departure from zero). Likewise, the sample PACF declines fairly rapidly to zero as the lag increases.

(d) Now we will consider fitting an ARIMA time series model to the `Nile` time series. The following command can be run to fit an ARIMA model to these data using a likelihood-based fit (we'll call this model `model1`, note that you can name your models as you wish).

```
#Code to fit an ARIMA model to the Nile dataset
#Here p = order of the AR part of the model
#q = order of the MA part of the model
#d = order of differencing
#p, d and q should be replaced by integers of your choice
 model1<-arima(Nile,order=c(p,d,q),method="ML")

#For example, to fit an ARMA(1,1) model we'd run the command
 model1<-arima(Nile,order=c(1,0,1),method="ML")

#To view the parameter estimates, standard error estimates and
#other model features, run the command
 model1
```

As an example, we begin by fitting an $AR(1)$ model to the `Nile` dataset (note - not necessarily the most desirable model). To do this, run the command:

```
model.AR1<-arima(Nile,order=c(1,0,0),method="ML")
```

What are the estimates of ϕ_1, μ and σ_z^2?

SOLUTION:

The R output is shown below.

```
Call:
arima(x = Nile, order = c(1, 0, 0), method = "ML")

Coefficients:
         ar1  intercept
      0.5063   919.5499
s.e.  0.0867    29.1419

sigma^2 estimated as 21125:  log likelihood = -639.95,  aic = 1285.9
```

The parameter estimates are: $\hat{\mu} = 919.5$, $\hat{\phi} = 0.51$ and $\hat{\sigma}^2 = 21125$.

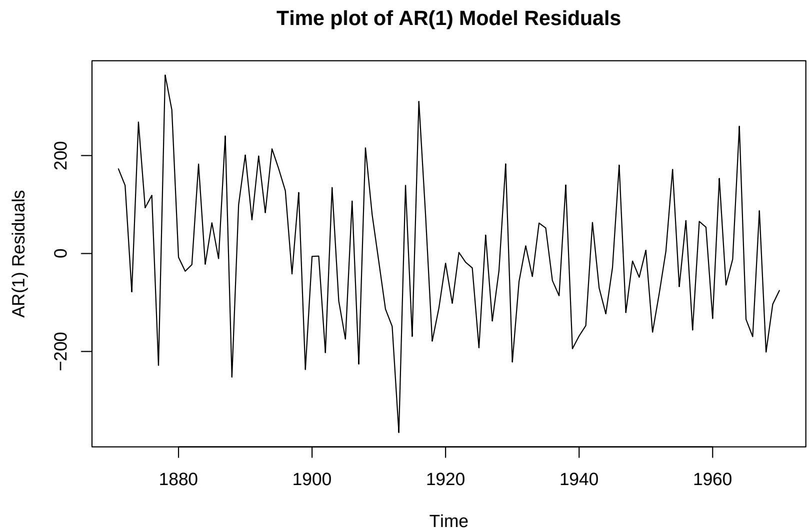
(e) Run the following command to extract the model residuals and store these as `resid.AR1`.

```
resid.AR1<-residuals(model.AR1)
```

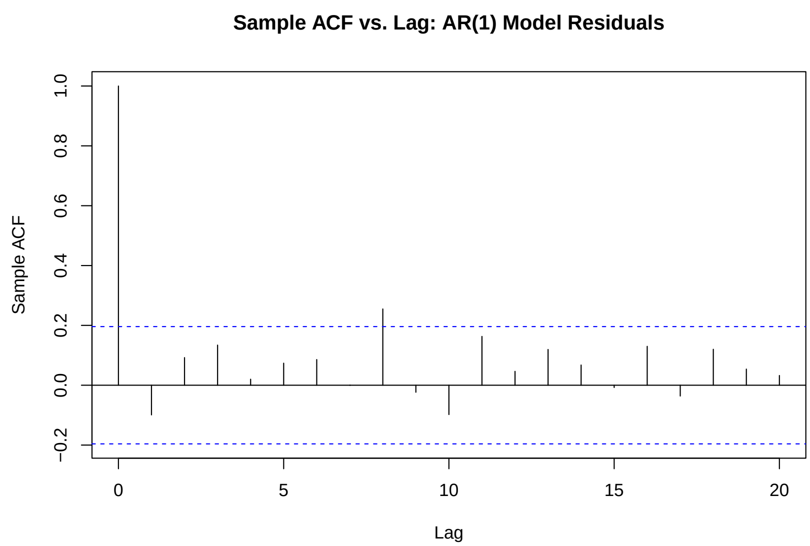
Produce a time plot and a plot of the sample ACF against the lag for these residuals. Do the residuals look like white noise?

SOLUTION:

A time plot of the AR(1) model residuals is shown below.



A plot of the sample ACF against the lag for the AR(1) model residuals is shown below.



Although many of the spikes in the sample ACF plot appear to be close to zero for lags > 0, the time plot suggests that the residuals are not white noise. The mean of the residuals appears to be higher for the years 1871-1895 than for later years.

(f) Finally, we'll examine the Ljung-Box test P-values with respect to the model residuals extracted in (e). Calculation of the Ljung-Box test P-values requires use of the bespoke function `LB_test`. The R code for this function can be found in the file `Ljung-Box-test.R`, available from the course moodle page.

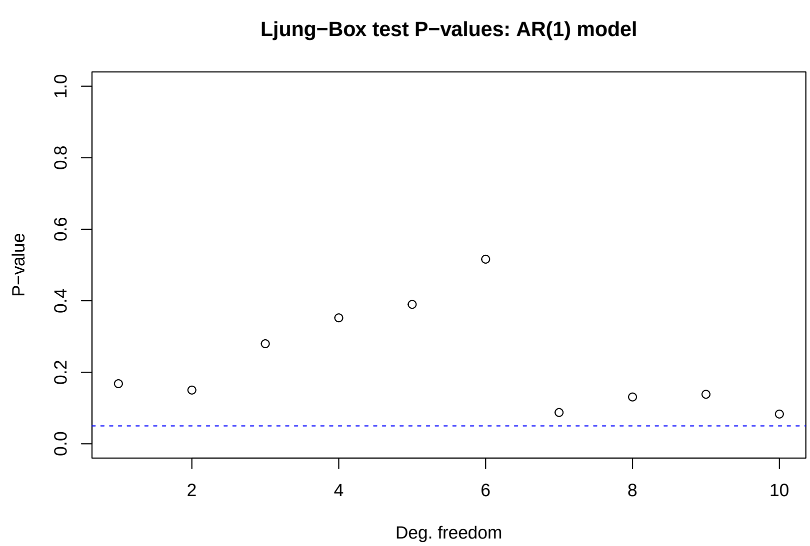
Download the file `Ljung-Box-test.R` and open this in RStudio. Run the commands in this file, then run the commands below to produce a plot of the first ten Ljung-Box test P-values against the degrees of freedom.

```
#Since p+q=1, we run the following command to perform the first ten
#Ljung-Box tests for the model residuals (max.k=11)
AR1.LB<-LB_test(resid.AR1,max.k=11,p=1,q=0)

#To see the table of P-values, type
AR1.LB

#To produce a plot of the P-values against the degrees of freedom and
#add a blue dashed line at 0.05, we run the commands
plot(AR1.LB$deg_freedom,AR1.LB$LB_p_value,xlab="Degrees of freedom",ylab="P-
value",main="Ljung-Box test P-values",ylim=c(0,1))
abline(h=0.05,col="blue",lty=2)
```

What do the P-values of the Ljung-Box test suggest about the fit of the AR(1) model?

SOLUTION:

A plot of the first ten Ljung-Box test P-values for the AR(1) model is shown below. Although all P-values are >0.05 (and hence non-significant at the 5% level) some



P-values are fairly small and - with the time plot of the residuals in mind - we may wish to consider fitting an alternative model to these data.

(g) By adapting the code from (d)–(f), find and fit an appropriate model to the time series `Nile` and report your model's parameter estimates and associated standard error estimates. Check the fit of your models by producing:

(i) A time plot of the model residuals.

(ii) A plot of the sample ACF of the model residuals against the lag.

(iii) A plot of the first ten P-values for the Ljung-Box test.

You should consider how to obtain your final model by using appropriate hypothesis tests and/or information criteria to be sure that your final model is parsimonious.

HINT: Consider whether or not differencing might be appropriate for this time series (see part (c)).

SOLUTION:

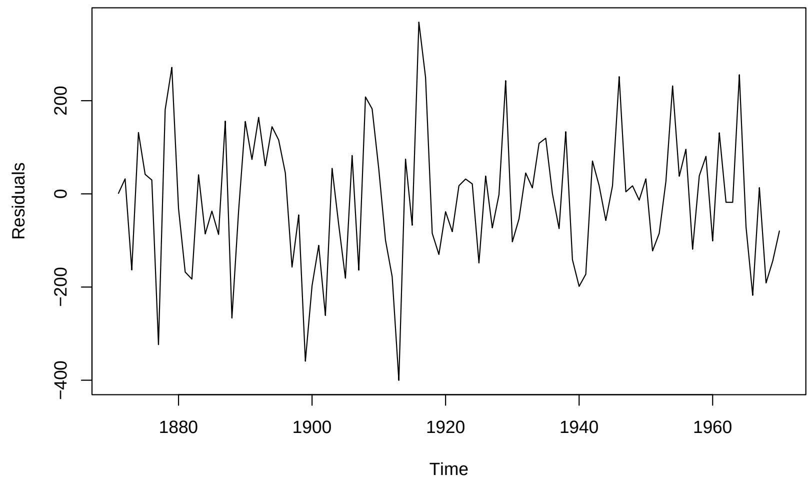
Looking at the hint above and the answer to part (c), it seems appropriate to consider fitting an $ARIMA(p, 1, q)$ model to these data for some values p and q that we will determine.

From part (c), noting the downward spike at lag 1 in the plot of the sample ACF, we might begin by fitting an $ARIMA(0, 1, 1)$ model, with the fitted model output shown below.

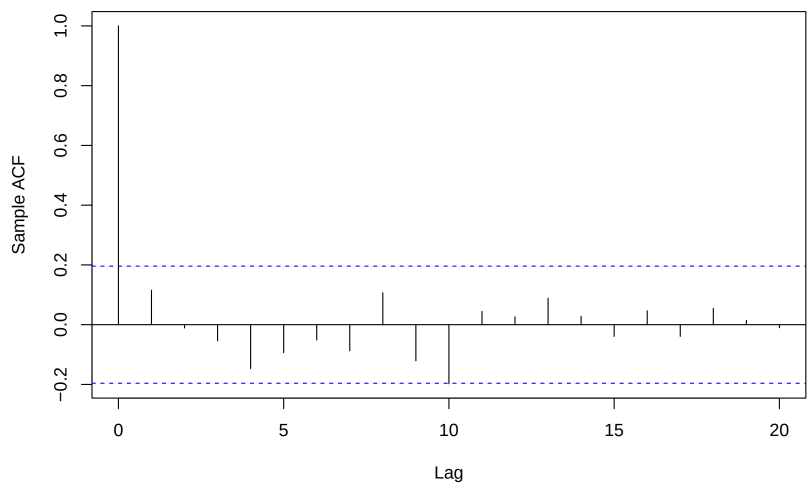
```
Call:
arima(x = Nile, order = c(0, 1, 1), method = "ML")

Coefficients:
          ma1
      -0.7329
s.e.   0.1143

sigma^2 estimated as 20600:  log likelihood = -632.55,  aic = 1269.09
```

We extract the model residuals and, using these, produce a time plot and a plot of the sample ACF against the lag.

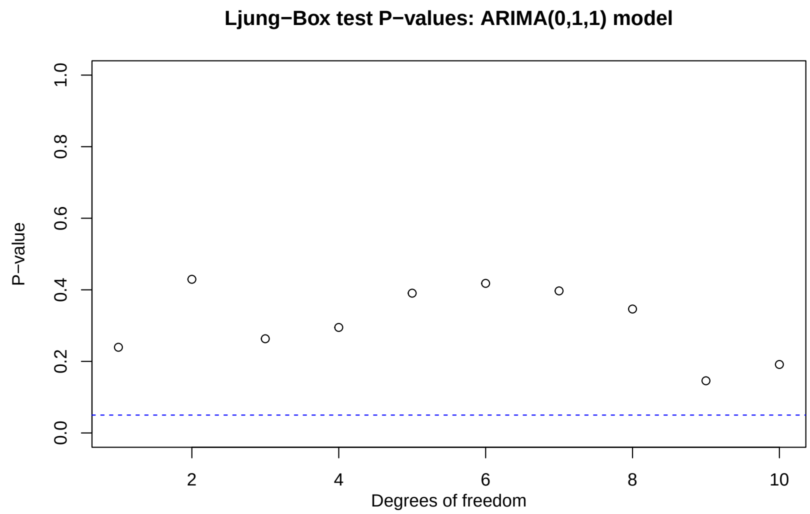
Time plot of ARIMA(0,1,1) Model Residuals



Sample ACF vs. Lag: ARIMA(0,1,1) Model Residuals



The time plot of the residuals appears to look similar to that for white noise. In addition, the sample ACF seems to be zero (or close to zero) at all lags > 0. This suggests that the residuals are independent. Next we produce a plot of the Ljung-Box test P-values for some lags, shown below.



All P-values are greater than 0.05 and this suggests that the ARIMA(0,1,1) provides a good fit to the data.

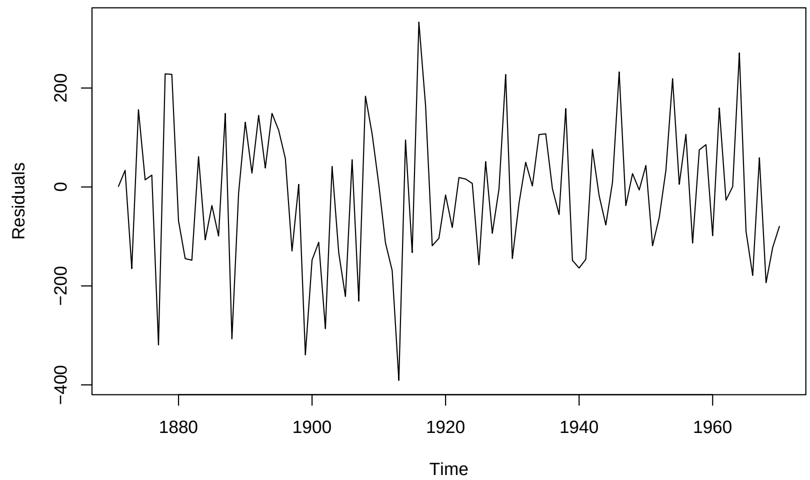
Having fitted an ARIMA(0,1,1) model we should check whether the addition of further parameters improves the model fit. As a result, we consider adding an AR component and fit an ARIMA(1,1,1) model, with model output shown below.

```
Call:
arima(x = Nile, order = c(1, 1, 1), method = "ML")

Coefficients:
         ar1      ma1
      0.2544  -0.8741
s.e.  0.1194   0.0605

sigma^2 estimated as 19769:  log likelihood = -630.63,  aic = 1267.25
```

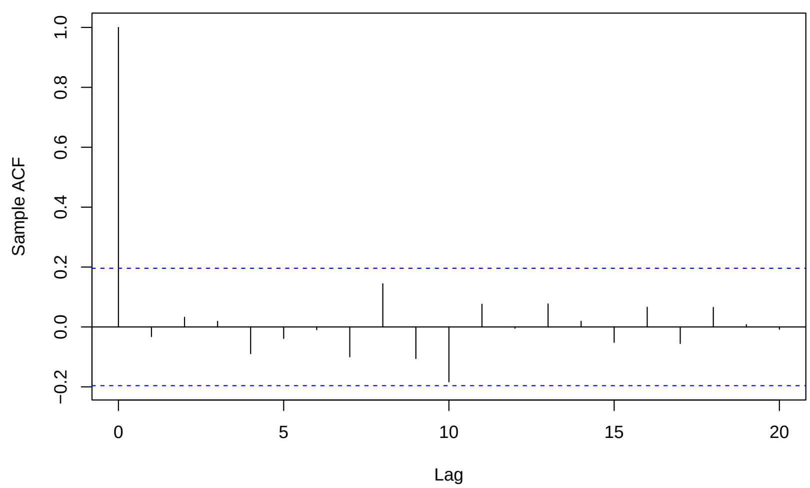
First we see that the AIC for this model is lower than that for the ARIMA(1,1,1) model (1267.25 vs. 1269.09) which may mean that we should prefer the ARIMA(1,1,1) model. In addition, the test statistic for a test of the hypotheses: $\mathrm{H}_0 : \phi = 0$ versus $\mathrm{H}_1 : \phi \neq 0$ is $0.2544/0.1194 = 2.13$ which is greater than 2. Hence we reject H_0 and conclude that the AR term should be included in the model.

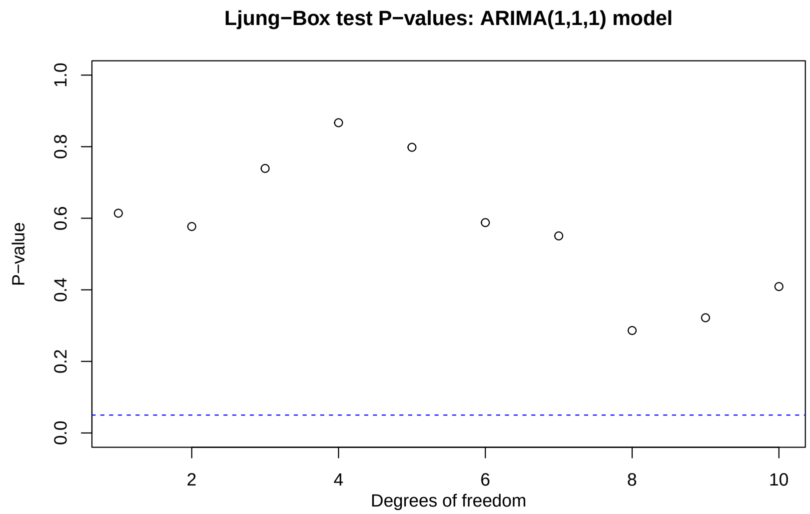
We should check the fit of the ARIMA(1,1,1) model by examining plots of the model residuals (shown below).

Time plot of ARIMA(1,1,1) Model Residuals



Sample ACF vs. Lag: ARIMA(1,1,1) Model Residuals





These plots suggest that the ARIMA(1,1,1) model fits the data well.

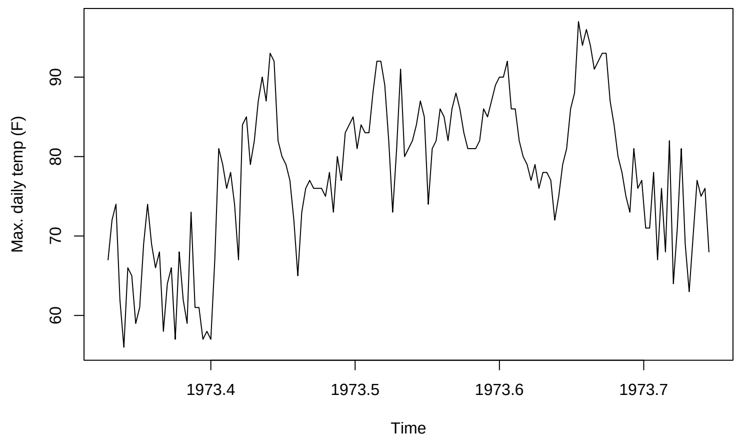
As a further check we might consider fitting an ARIMA(2,1,1) and an ARIMA(1,1,2) model to see if the addition of further AR or MA terms might improve the model fit. Each of these models has an AIC value of 1268.9 (and therefore higher than the ARIMA(1,1,1) model). Moreover, hypothesis tests to check whether or not additional terms should be included result in large P-values, thereby suggesting that the ARIMA(1,1,1) model is preferable.

At this point, we would choose the ARIMA(1,1,1) model as the most appropriate for the Nile dataset.

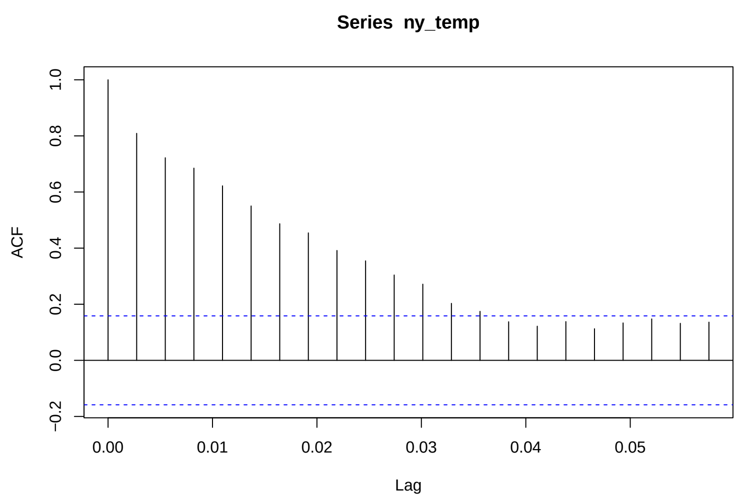
2. (b) Produce a time plot, plot of the sample ACF against the lag and a plot of the sample PACF against the lag for the `ny_temp` data. Does this time series appear to be stationary?

SOLUTION:

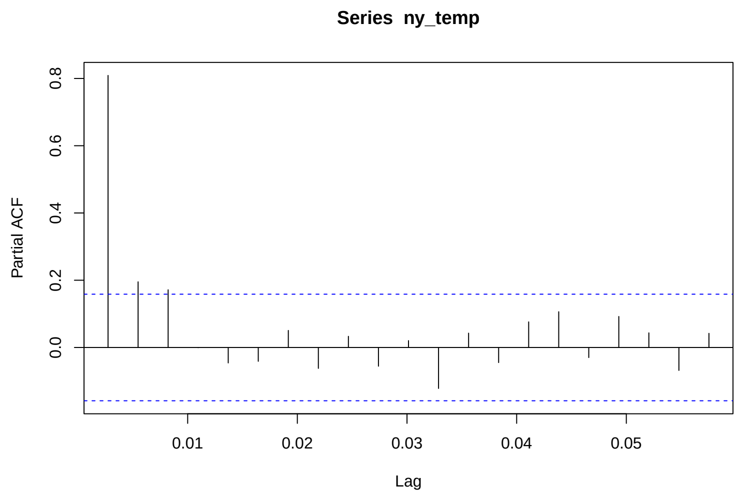
Time plot:



Plot of the sample ACF against the lag:



Plot of the sample PACF against the lag:



Examining these plots it is clear that the process does not appear to be stationary. The time plot shows variation over time and the sample ACF values do not decline sharply as the lag increases. As such, we'll explore models fitted to the differenced data in part (c).

(c) Now take the first difference of the `ny_temp` dataset, by running the command:

```
temp_diff<-diff(ny_temp)
```

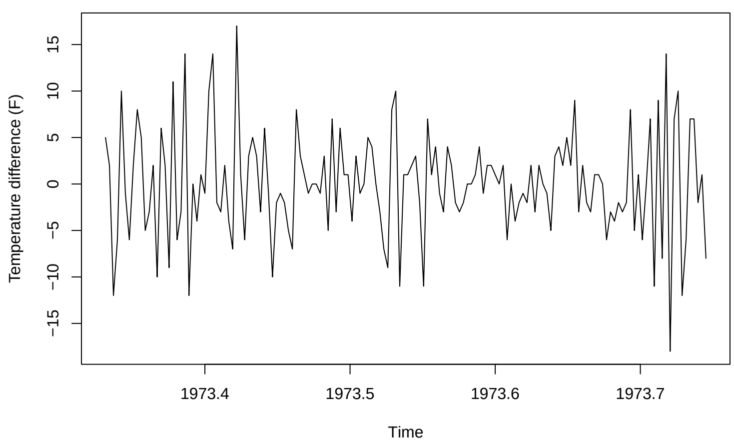
Using the methods from question 1. choose an appropriate ARMA model to fit to the differenced New York temperature data. You should justify your choice of model and assess its fit.

Note: this process should involve fitting more than one model and then choosing the most appropriate (and parsimonious) model to fit to the data.

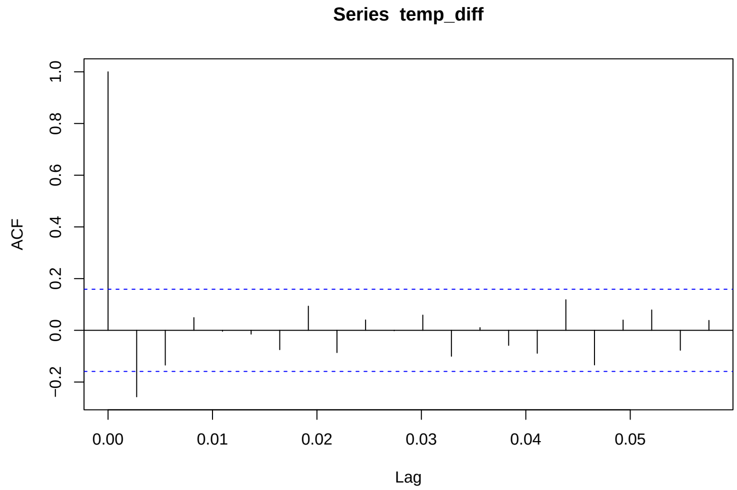
SOLUTION:

We begin by producing a time plot of the differenced data and also plots of the sample ACF and sample PACF against the lag, shown below.

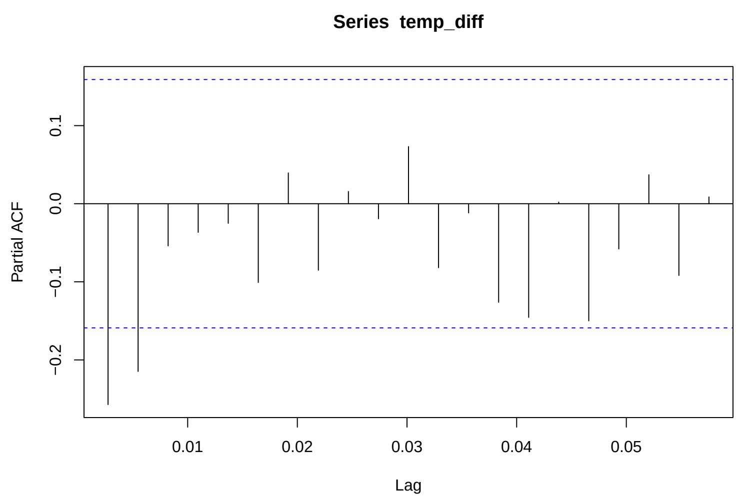
Time plot:



Plot of the sample ACF against the lag:



Plot of the sample PACF against the lag:

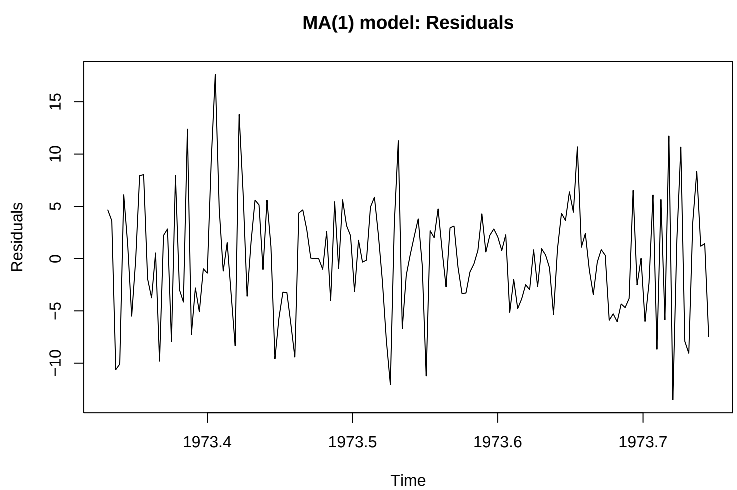


Examining these plots, the time plot suggests that the process is stationary. The plot of the sample ACF may cut off to zero after lag 1, which could suggest that an MA(1) model may be appropriate. There is no obvious peak in the sample PACF which may imply that there is no AR component in the process.

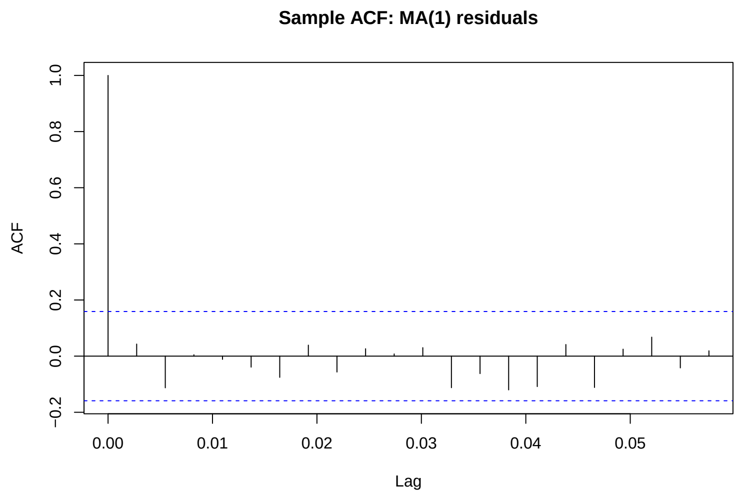
As a result, we begin by fitting an MA(1) model to the data, with the model output shown below.

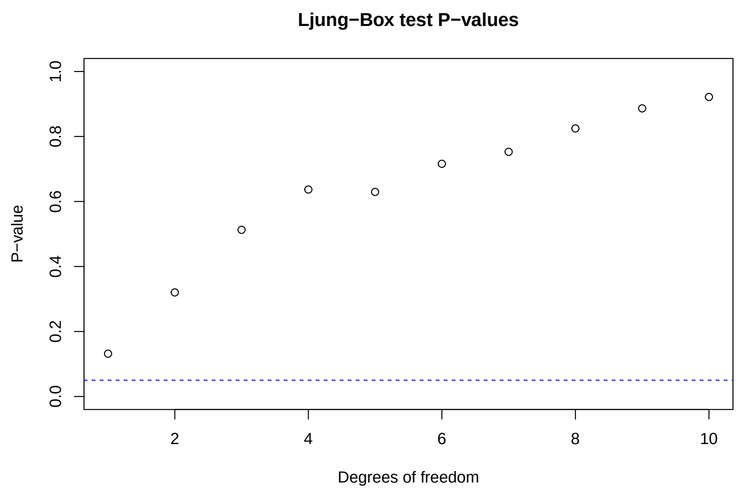
```
Call:
arima(x = temp_diff, order = c(0, 0, 1), method = "ML")

Coefficients:
          ma1  intercept
      -0.3841     0.0156
s.e.   0.0847     0.2722

sigma^2 estimated as 29.45:  log likelihood = -472.85,  aic = 951.7
```

A time plot, plot of the sample ACF against the lag and plot of the Ljung-Box test P-values for the model residuals are shown below.



Examining these plots, at first sight, it seems that the residuals may look like a white noise process. The time plot would suggest this and, in addition, the sample ACF is close to zero for all lags and the P-values for the Ljung-Box test are all non-significant.





The next step would be to consider fitting an $MA(2)$ model which we'll now do. The fitted model output for the $MA(2)$ model is shown below.

```
Call:
arima(x = temp_diff, order = c(0, 0, 2), method = "ML")

Coefficients:
          ma1      ma2  intercept
      -0.3297  -0.1350     0.0208
s.e.   0.0822   0.0871     0.2355

sigma^2 estimated as 28.99:  log likelihood = -471.67,  aic = 951.34
```

First, note the very slight reduction in AIC for this model (AIC $=951.34$) compared to the $MA(1)$ model (AIC $=951.7$). Perhaps this might cause us to favour the $MA(2)$ model over the $MA(2)$ model. To examine this further, we perform a test of the hypotheses

$$H_0 : \theta_2 = 0 \text{ versus } H_1 : \theta_2 \neq 0.$$

The test statistic is $|-0.1350/0.0871| = 1.55$. Clearly, 1.55 is less than 2, so we would retain H_0 at the 5% level and conclude that we'd prefer the $MA(1)$ model over the $MA(2)$ model. The MA(1) model has fewer parameters so, despite the tiny increase in AIC, we'd probably choose to fit an MA(1) model here.

As a final check, let us consider fitting an $ARMA(1,1)$ model, just in case this provides a better fit to the data. The fitted model output for the $ARMA(1,1)$ model is shown below.

```
Call:
arima(x = temp_diff, order = c(1, 0, 1), method = "ML")

Coefficients:
         ar1      ma1  intercept
      0.3004  -0.6414     0.0229
s.e.  0.2202   0.1826     0.2263

sigma^2 estimated as 29.06:  log likelihood = -471.85,  aic = 951.71
```

Firstly, the AIC for this model is approximately equal to that for the $MA(1)$ model (but the MA(1) has one less parameter, hence we *might* prefer the MA(1) model). To check, let us test the hypotheses

$$H_0 : \phi_1 = 0 \text{ versus } H_1 : \phi_1 \neq 0.$$

The test statistic is $0.3004/0.2202 = 1.36 < 2$. Hence we retain the null hypothesis (at the 5% level) and conclude that we should prefer the $MA(1)$ model over the $ARMA(1,1)$. Overall, we would fit an $MA(1)$ model to these data.