

**MH6241 - Time Series Analysis**

**Group Project Report**

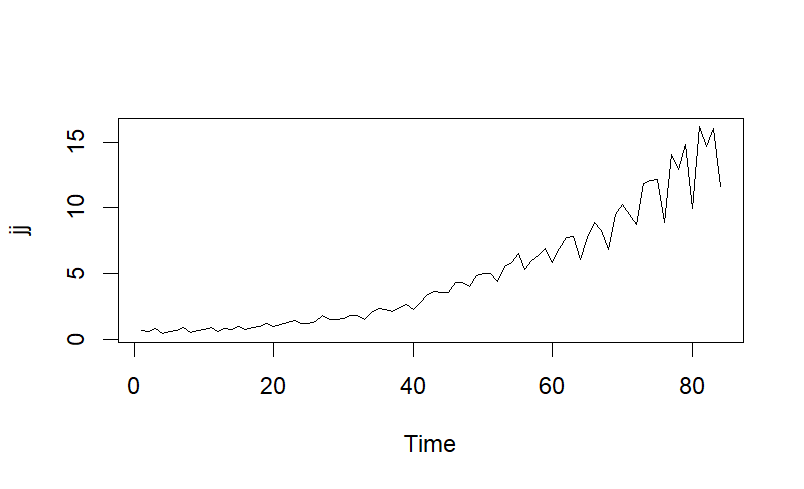
|  |  |
| --- | --- |
| **NAME** | **MATRICULATION NUMBER** |
| Khoo Chun Yun | G2302847F |
| Goel Ishita | G2302904C |
| Du Siyi | G2303045K |
| Zhao Jiayi | G2302767L |
| Ang Shu Wei | G2302794C |
| Nong Minh Hieu | G2302707A |
| Zhang Danrui | G2302855K |

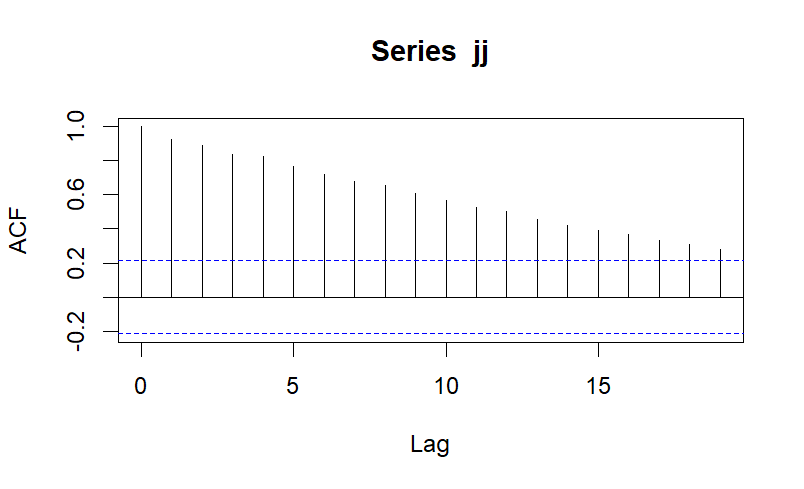
**Introduction**

A dataset named “JJ.dat” was imported into R. The dataset is about the quarterly earnings per share from 1960 to 1980 for the company of Johnson & Johnson. In the report we demonstrate fitting appropriate ARIMA models to the real world data.

**Methodology**

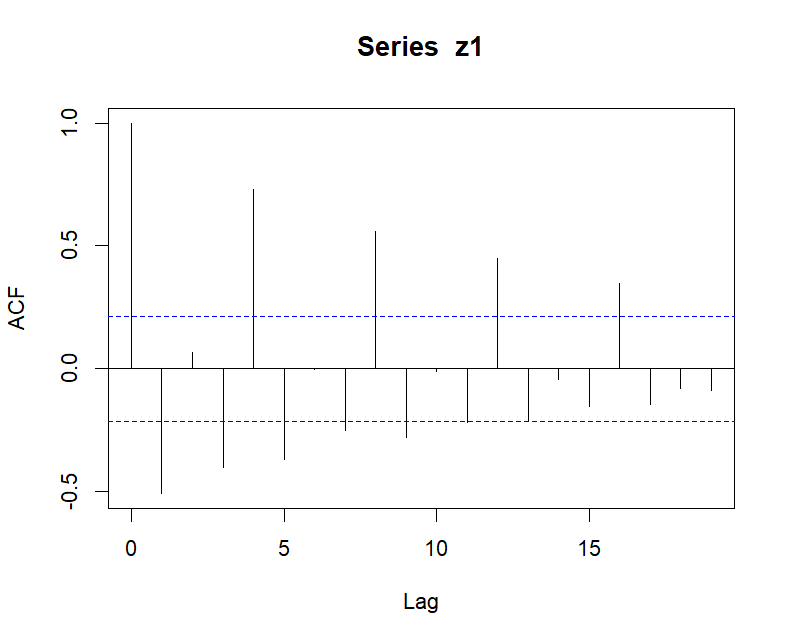
Firstly, we will use both ADF test and visualisation from time plot (Figure 1) and the ACF (Figure 2) to ascertain whether the data is stationary. First, the time plot is analysed. It is observed that there is an upward trend, which suggests the mean increases over time. Thus, there is non-stationary in the mean. It is also observed that it is non-stationary in the variance, given that the spread of values increases over time.

**Figure 1: Time Plot of the data**

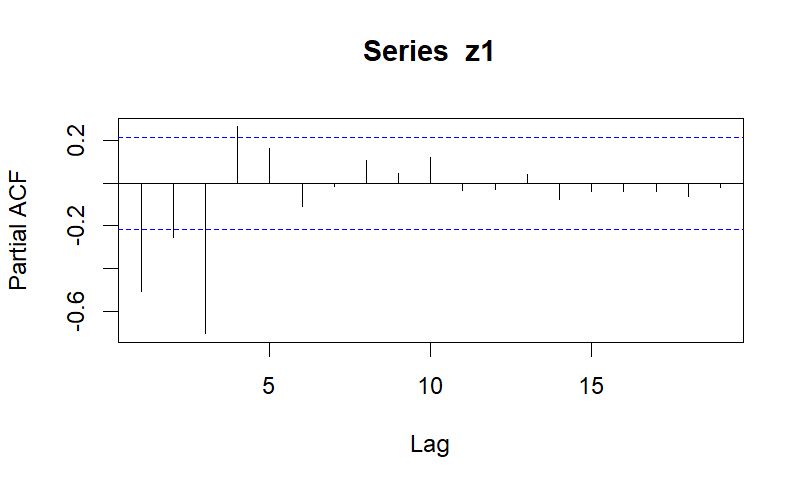
Also, the ACF decays very slowly in Figure 2, which suggests there is non-stationarity. The ADF test corroborates our interpretation of the time plot as the null hypothesis cannot be rejected. Therefore, we will apply log transformation and non-seasonal differencing. 

**Figure 2: ACF of the original (non-transformed, non-differenced) data**

Next, the ACF (Figure 3) of the transformed and differenced data is analysed. There is a pattern observed across lags in the multiple of 4, which suggests that we should apply one-time seasonal differencing.



**Figure 3: ACF of the log-transformed and differenced data**

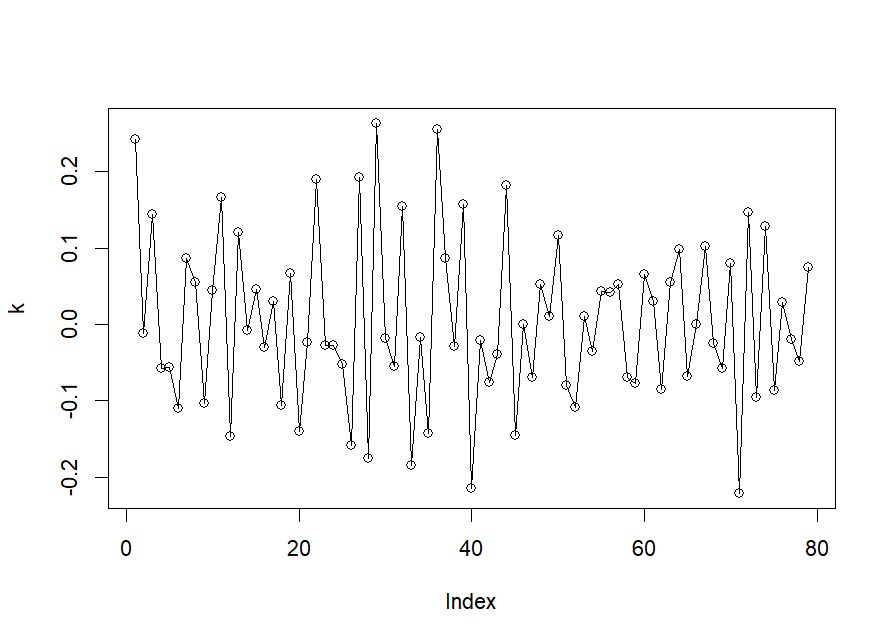


**Figure 4: PACF of the log-transformed and differenced data**

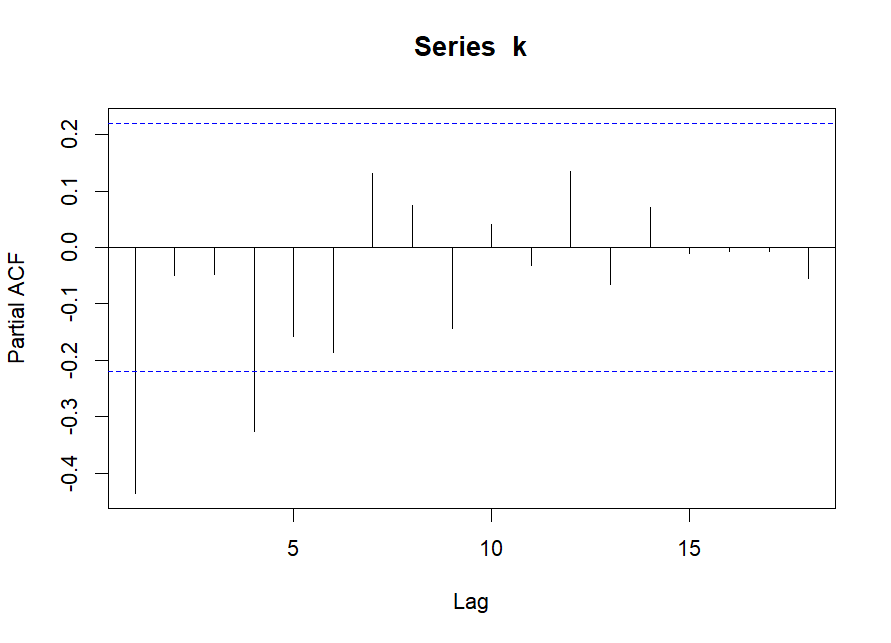
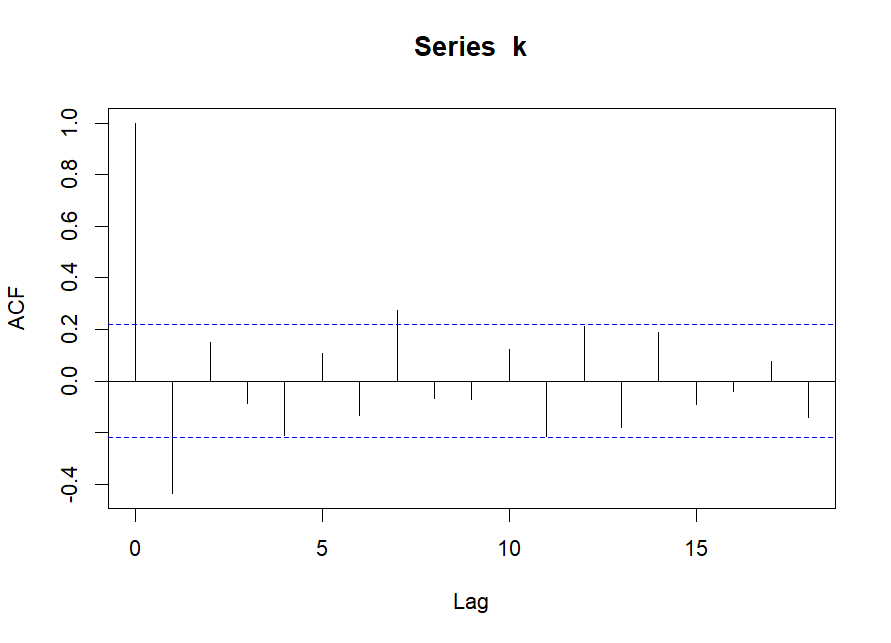
We apply one time seasonal differencing and use the adf test to check whether the seasonal differenced data is stationary. However, it depicts that the time series still has some non-stationary components. Thus, non-seasonal differencing is applied to the seasonal-differenced data.

Therefore, d=1, D=1, and s = 4 in SARIMA(p, d, q, P, D, Q, s).

After the transformation and differencing, based on the time plot(Figure 5), the differenced data looks stationary.



**Figure 5: Time plot of the transformed and differenced (seasonal and non-seasonal) data**



**Figure 6: ACF and PACF of the transformed and differenced (seasonal and non-seasonal) data**

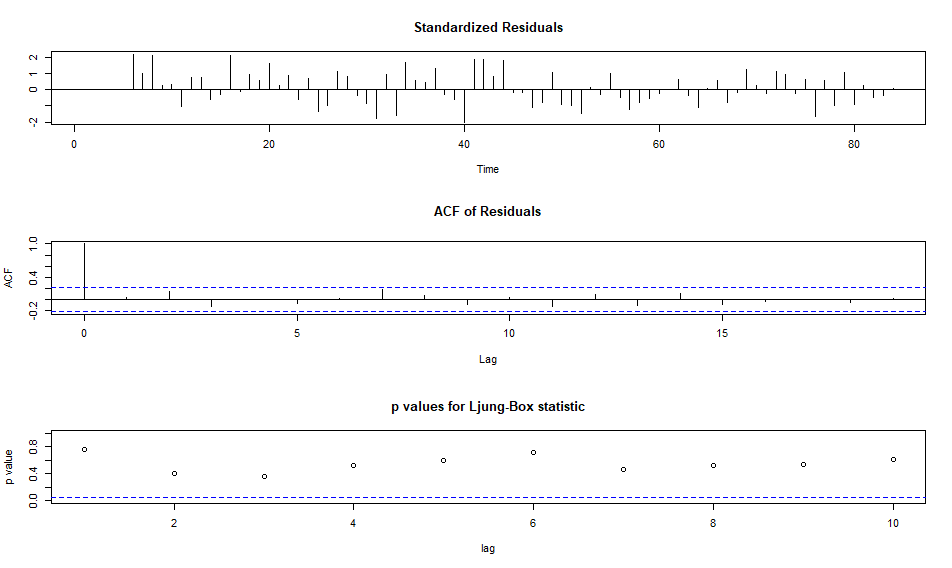
The initial models are derived by first looking at ACF and PACF (Figure 6), as elaborated in Model 1 and Model 2. Then, to look at possible improvements to the initial models, we will look at 2 other alternative combinations derived from SARIMA parameters from model 1 and 2, which will be labelled as model 3 and 4.

|  |  |  |
| --- | --- | --- |
| SARIMA Model | Parameters:  p,d,q, P, D,Q,s | AIC |
| 1 | **0,1,1,0,1,1,4** | **-150.75** |
| 2 | 1,1,0,1,1,0,4 | -146.03 |
| 3 | **0,1,1,1,1,0,4** | **-150.91** |
| 4 | 1,1,0,0,1,1,4 | -146.02 |

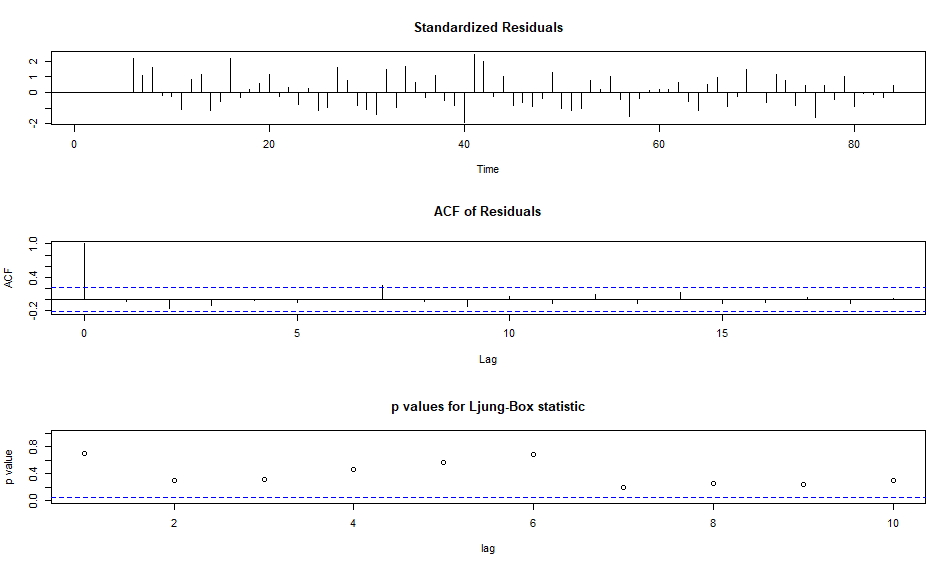
**Table 1: Models**

Looking at the diagnostic plots of the model we find that all these models are adequate. All of them have p-values greater than 0.05 in Ljung-Box Statistic Test. In addition, for all the four models have the acf of residual indicating that they are white noise.[Figure 7-Figure 10]

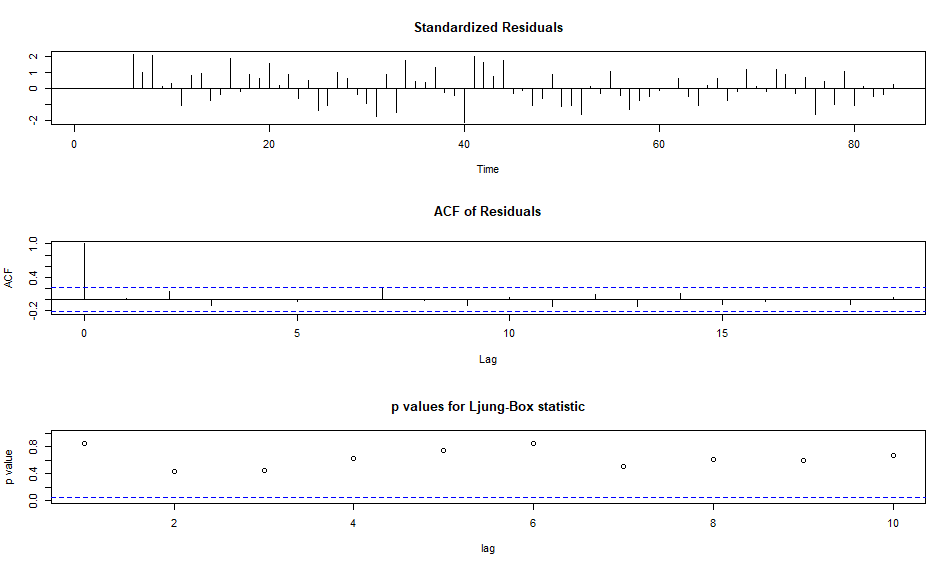
Since all the models are adequate we make use of AIC to find the best fit (Table 1). We observe that the model 3 and model 1 have the best AIC values (lowest AIC values).



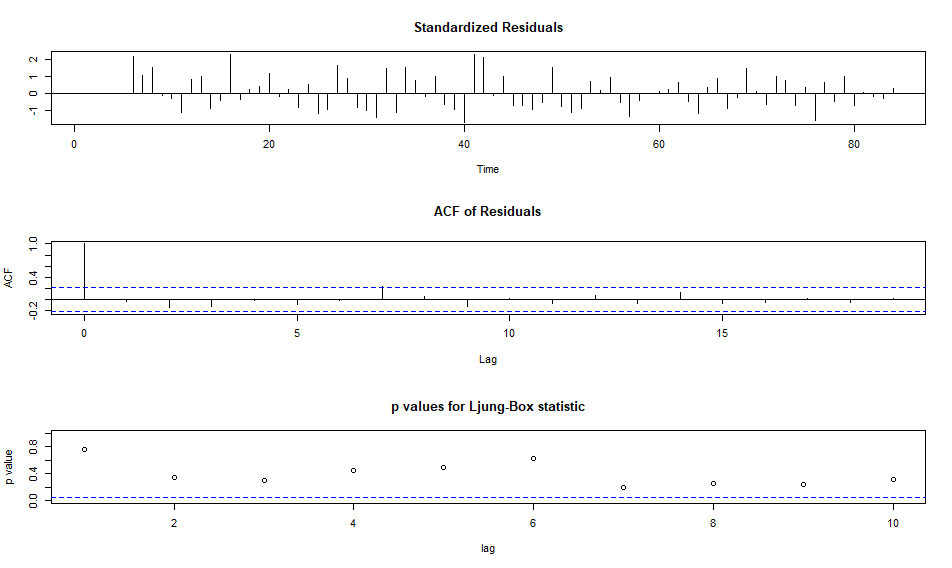
**Figure 7: Diagnostic Plots for Model 1 Fit**



**Figure 8: Diagnostic Plots for Model 2 Fit**



**Figure 9: Diagnostic Plots for Model 3 Fit**



**Figure 10: Diagnostic Plots for Model 4 Fit**

**Model 1: SARIMA (p, d, q, P, D, Q, s) = SARIMA(0, 1, 1, 0, 1, 1, 4)**

The ACF of the differenced data is observed in multiples of 4, e.g. at lags 4, 8, 12… to identify orders of Q. There is a slightly significant value at lag 4, after which it cuts off. Therefore, Q = 1, P = 0. Also, since s = 4, the ACF of the differenced data is observed at lags 1 to 3 to determine small q. There is a significant value at lag 1, after which it cuts off. Thus, ACF likely cuts off after lag 1. Therefore, q = 1, p = 0.

**Model 3: SARIMA(p, d, q, P, D, Q, s) = SARIMA(0, 1, 1, 1, 1, 0, 4)**

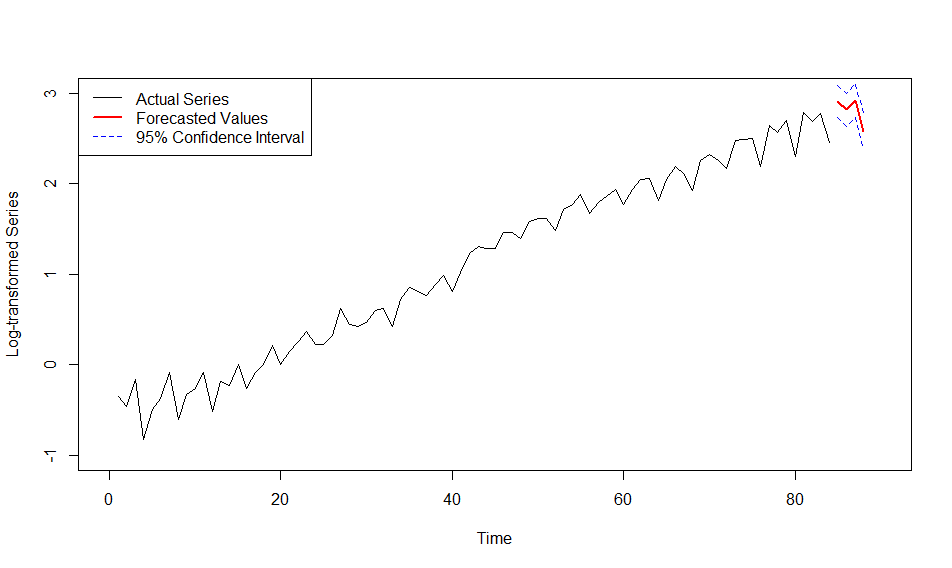
The orders of p and q are obtained from Model 1(as explained above), while the orders of P and Q are obtained from model 2.

In model 2, the PACF of the differenced data is observed in multiples of 4, e.g. at lags 4, 8, 12… to identify orders of P. There is a significant value at lag 4, after which it cuts off. PACF likely cuts off after lag 1.Therefore, P = 1, Q = 0.

**Forecast:**

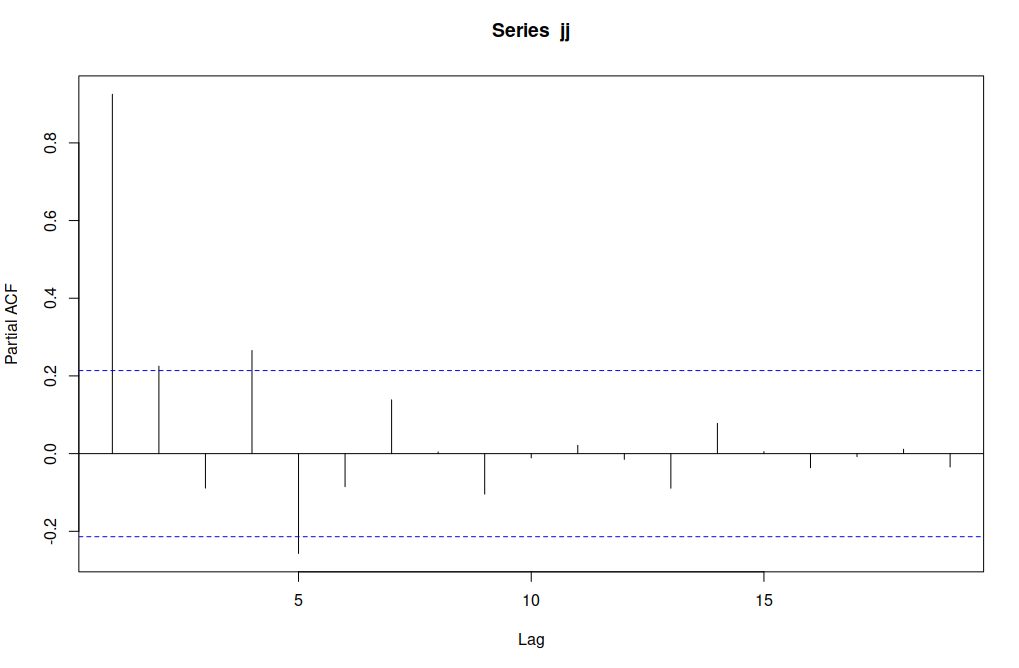
For forecasting the next 4 values (the next 4 quarters) we use the best fitted model inferred from the previous section, that is the Model 3.

The predicted values are represented in red and the 95% confidence interval accounting for errors is represented as the dotted lines.

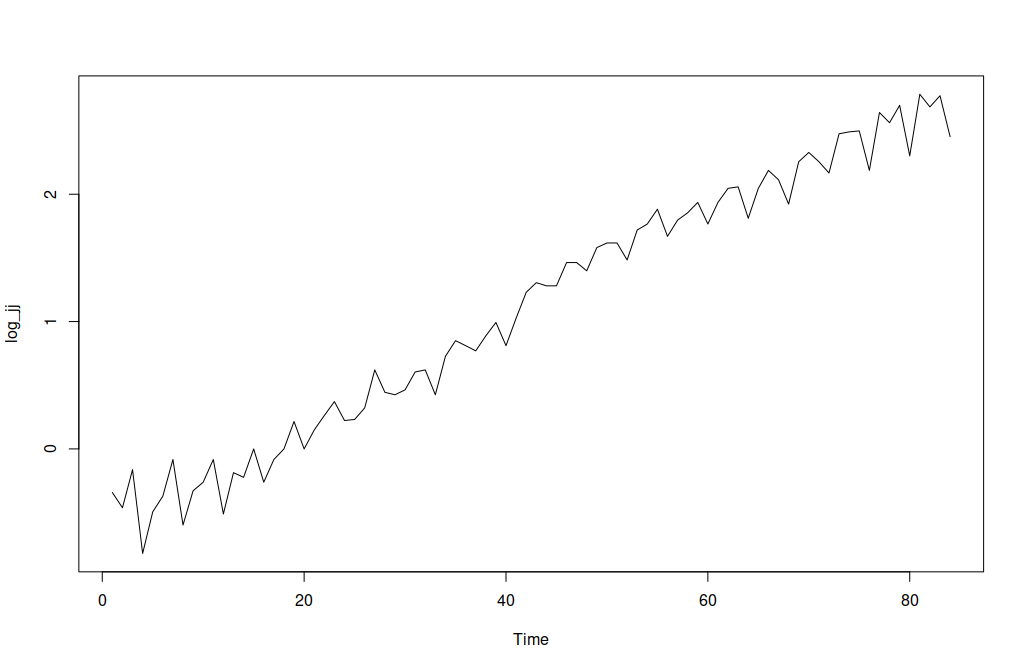


**Figure 11: Forecasted Values and Confidence Intervals from Model 3**

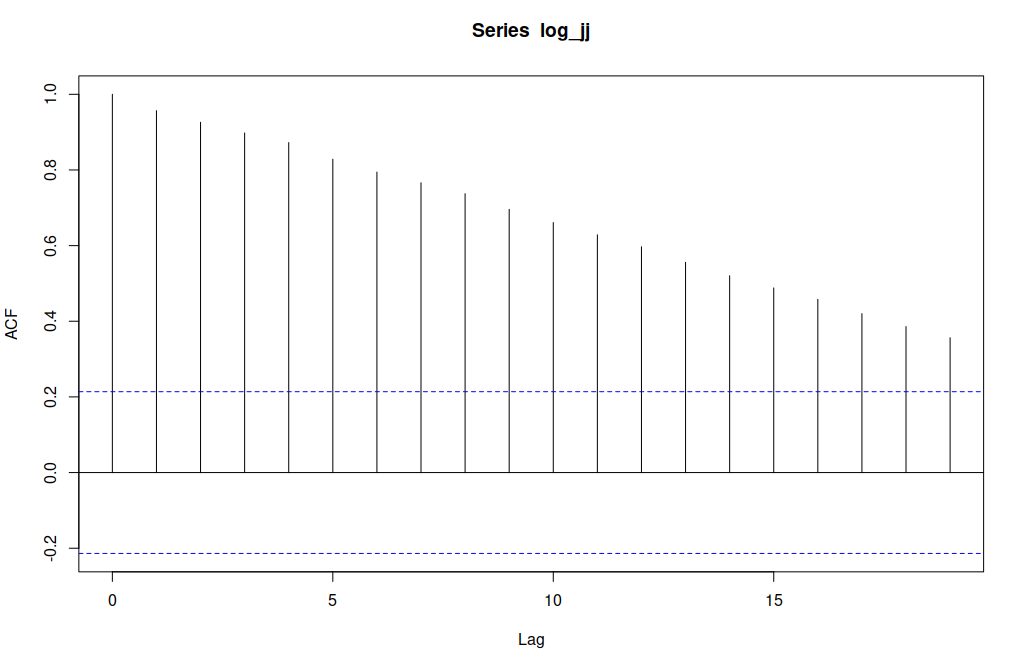
**Appendix - Additional visualisations**

****

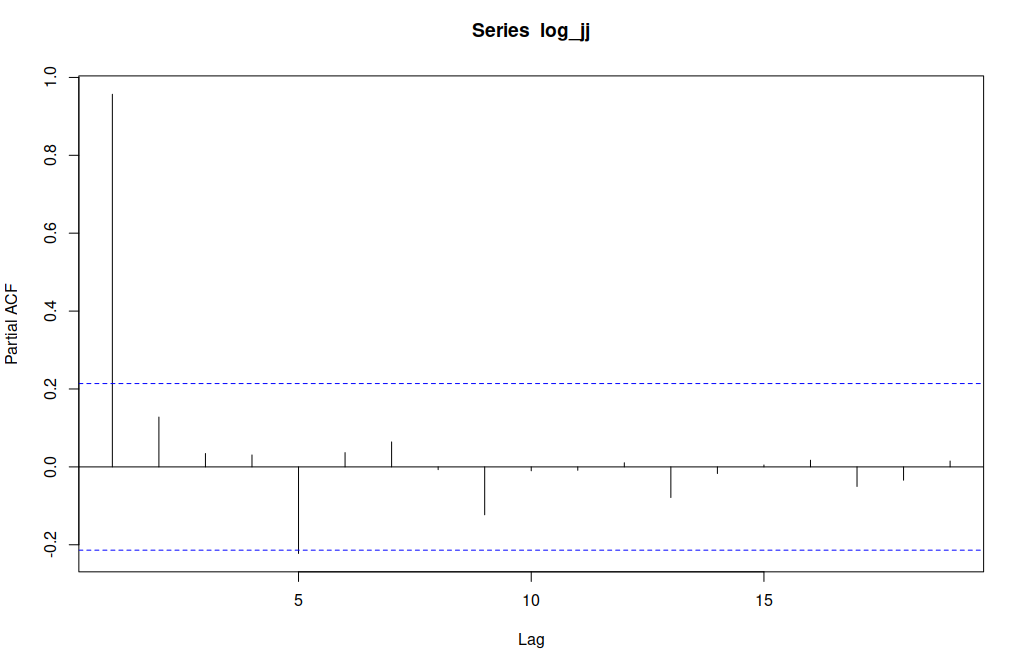
**Figure A.1: PACF of the original (non-transformed, non-differenced) data.**

****

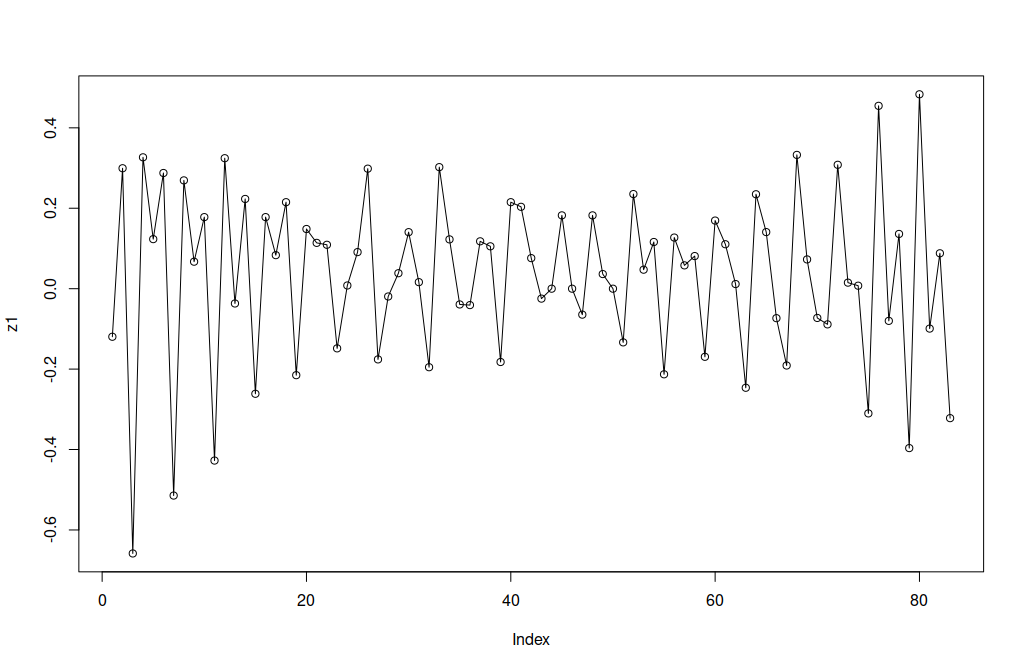
**Figure A.2: The log-transformed data time plot.**

****

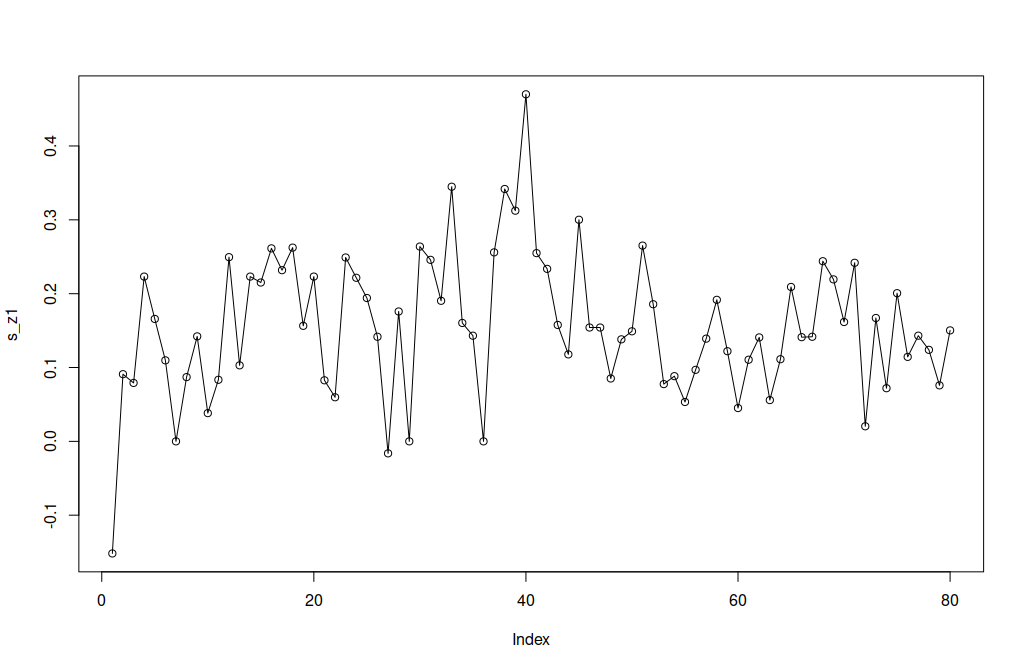
**Figure A.3: The ACF of the log-transformed data.**

****

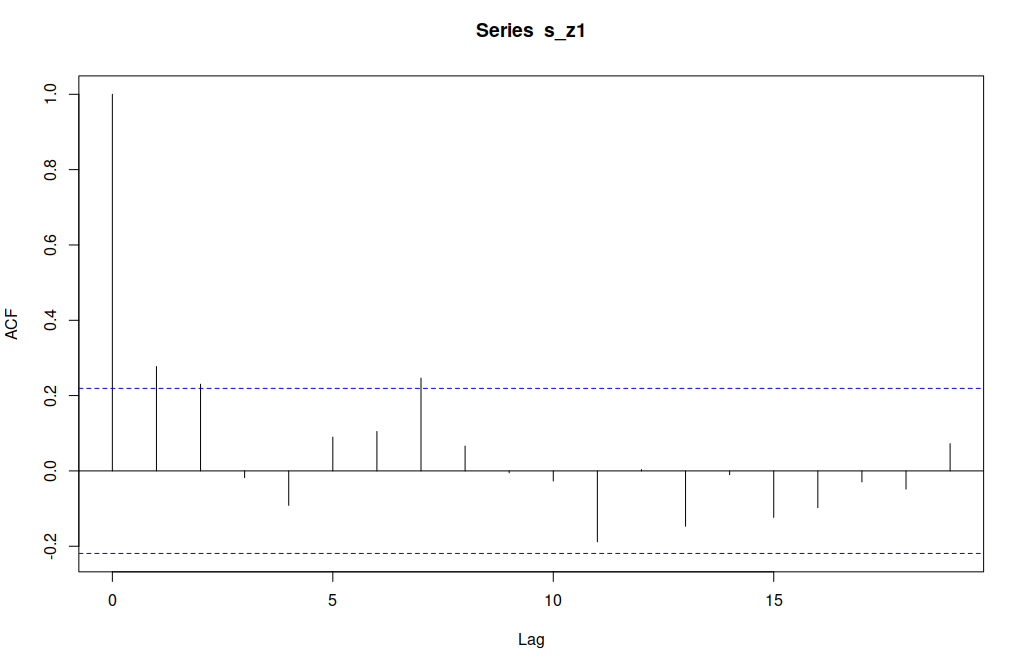
**Figure A.3: The PACF of the log-transformed data.**

****

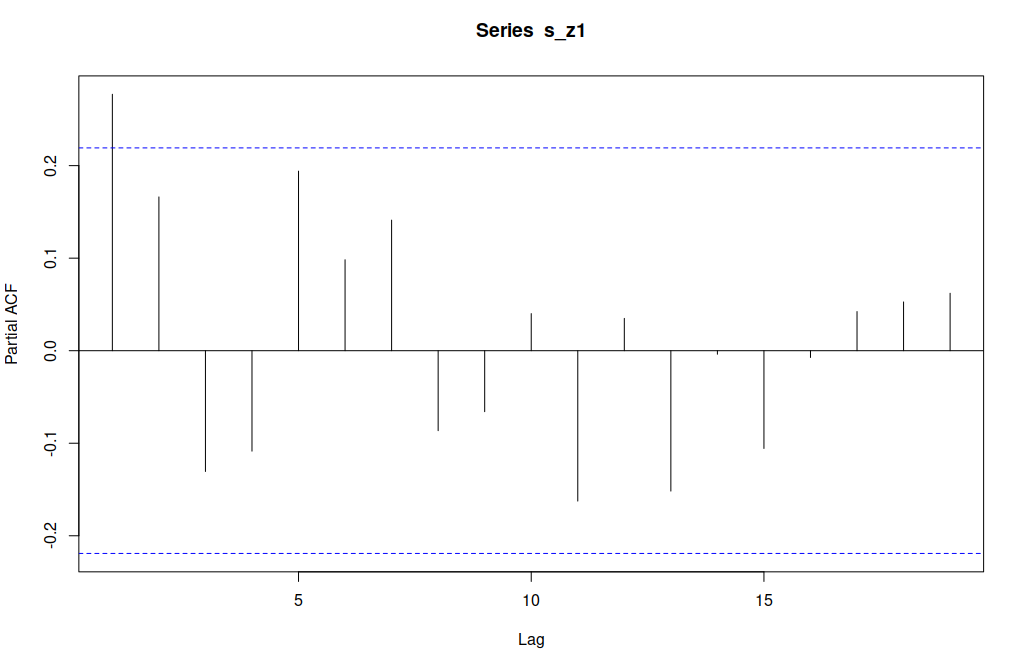
**Figure A.4: The time plot of the data after log-transform and first-order differencing.**

****

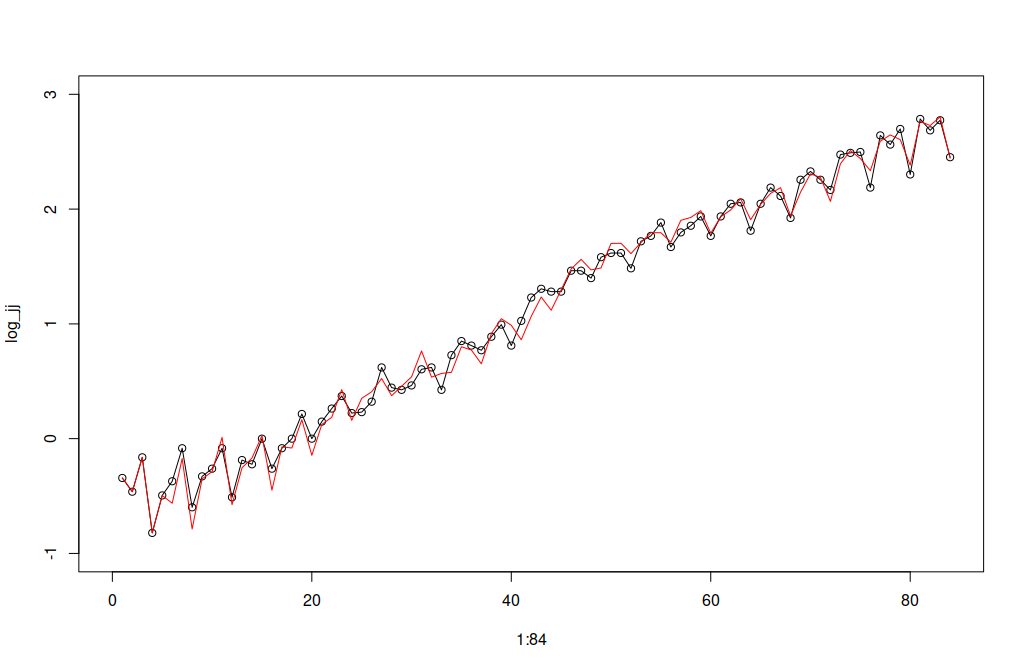
**Figure A.5: The time plot of the data after log-transform and seasonal differencing.**

****

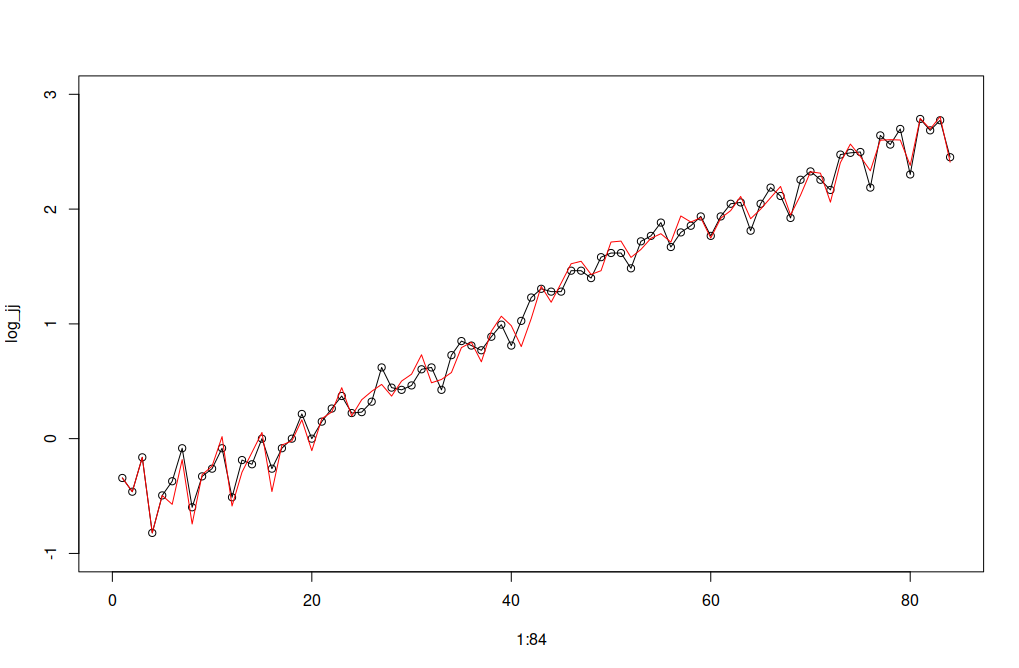
**Figure A.6: The ACF of the data after log-transform and seasonal differencing.**

****

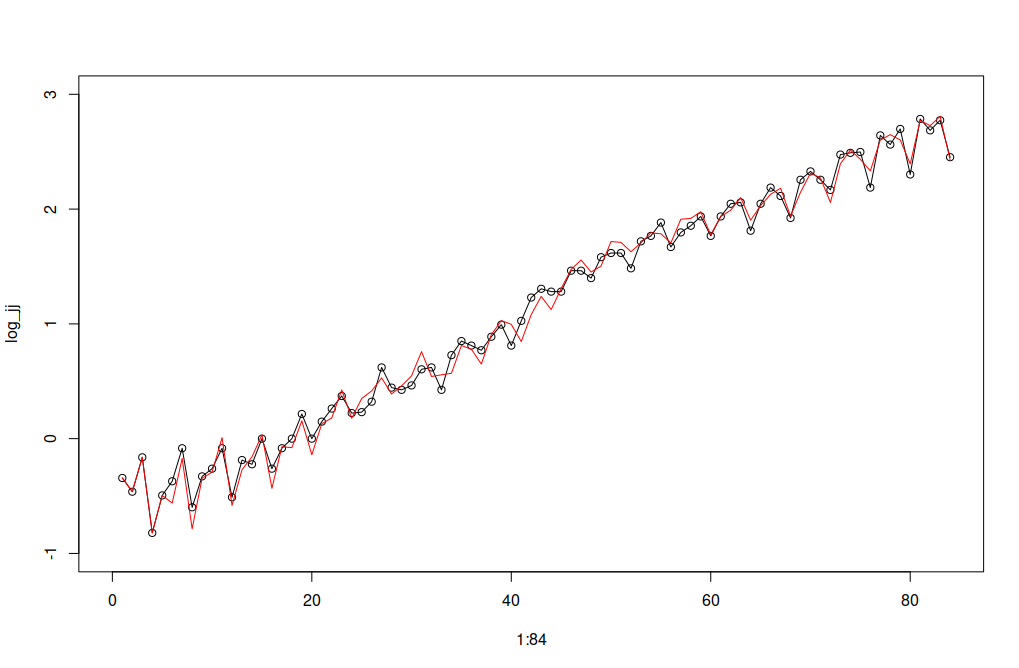
**Figure A.7: The PACF of the data after log-transform and seasonal differencing.**

****

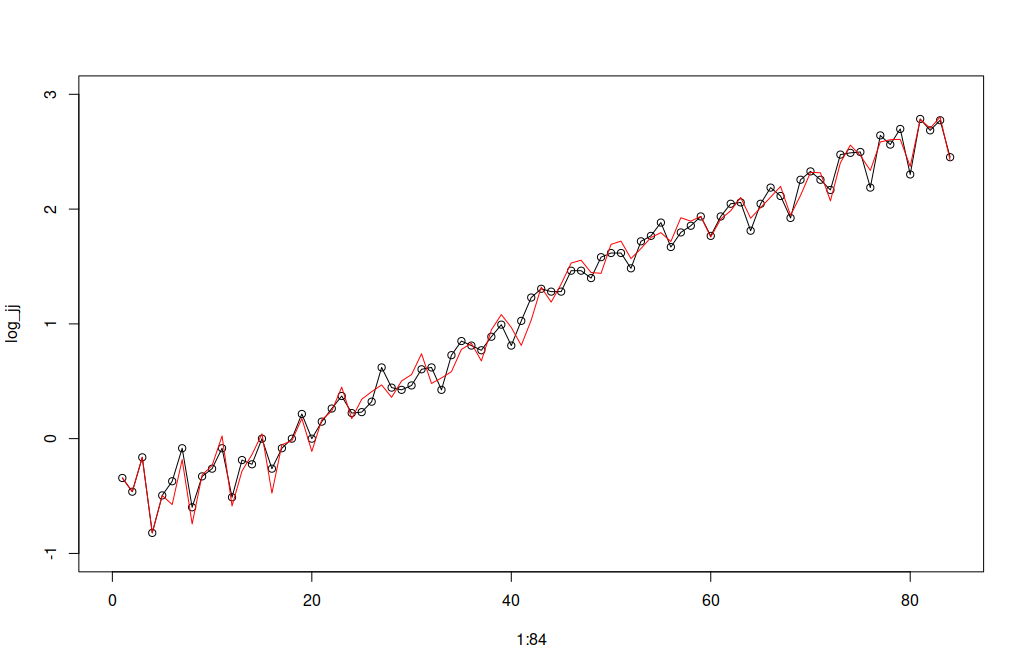
**Figure A.8: Time plot of the log-transformed data against log-transformed data minus the residuals of model#1.**

****

**Figure A.9: Time plot of the log-transformed data against log-transformed data minus the residuals of model#2.**

****

**Figure A.10: Time plot of the log-transformed data against log-transformed data minus the residuals of model#3.**

****

**Figure A.11: Time plot of the log-transformed data against log-transformed data minus the residuals of model#4.**