PSTAT274_HW5

Qifei Cui

2023-11-08

Problem 1: Glossary of R-commands for time series.

It should contain all commands that you learned so far in the labs, doing homework, and reviewing posted lecture slides.

Purpose	R Command
Define working directory	`setwd()`

Read and plot data	`read.table`
	`read.csv()`
	`ts()`
	`ts.plt()`

Simulate and plot ARMA models	`arima.sim()` `plot()`
Add trend/mean line to the original time series plot	`abline(lm(x~as.numeric(1:length(x))))` `abline(h=mean(x))`
Calculate/plot theoretical acf/pacf for ARMA models	`ARMAacf()` `plot()`
Calculate/plot sample acf/pacf	`acf()` `pacf()`

Check whether a particular model is causal/invertible	# Import R scripts from canvus
	`plot.roots()`
	# Using property of models
	`polyroot(c())`

Purpose	R Command
Perform Box-Cox transforms	`boxcox()`
Perform differencing data at lags 1 and 12	`diff(diff(x,1),12)`

Perform Yule-Walker estimation and find std of the estimates	`x.ywest=ar(x, aic=TRUE, method= "yule-walker")`
	`sqrt(diag(x.ywest$asy.var.coef))`
	`ar.yw(x,order=k)`
	`sqrt(yw$var.pred)`

Perform MLE and check AICC associated with the model	`x_fit = arima(x,order=c(p,d,q), method = "ML")` `AICc(x_fit)`

Problem 2: Dataset chosen and analysis

Choose a dataset that you will be interested to analyze for your class final project. URLs of time series libraries are posted on Canvas. Provide the following information about the project:

Dataset Description:

The dataset I plan to use is the "Individual Household Electric Power Consumption" dataset from the UCI Machine Learning Repository. It contains over 2 million instances of minute-averaged electric power consumption measurements from a single household near Paris, France, collected over almost four years (December 2006 - November 2010). This time-series dataset has nine features, including active and reactive power, voltage, current intensity, and three types of sub-metering data.

The following code will download, unzip and import the dataset with saving it in the ~/data directory of the root of the project.

```
download.file(url, file, method="curl")
unzip(file, exdir = "~/data")
data_file <- "~/data/household_power_consumption.txt"
data <- read.table(data_file, sep=";", header=TRUE)
```

Motivation and objective:

This dataset is significant because it represents a real-world scenario of energy consumption that can provide insights into usage patterns and efficiency. The motivation for using this dataset is to explore the potential of time-series forecasting in energy consumption, which is crucial for energy management and sustainability.

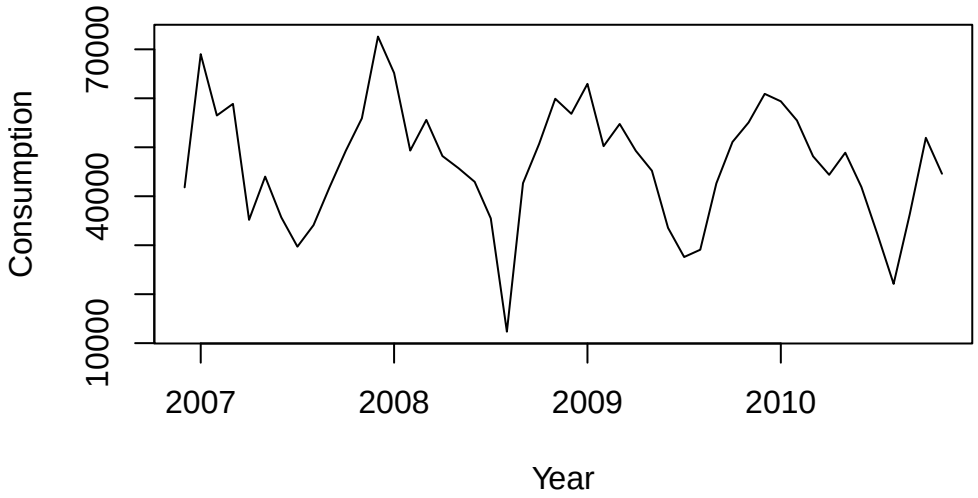
In the forthcoming time series analysis project, my primary focus will be on the "global_active_power" variable from the dataset. This variable records the household's total minute-averaged active power consumption, measured in kilowatts. This continuous variable is suited for time series forecasting, and is expected to provide a pattern of the household's energy consumption behavior.

Main features

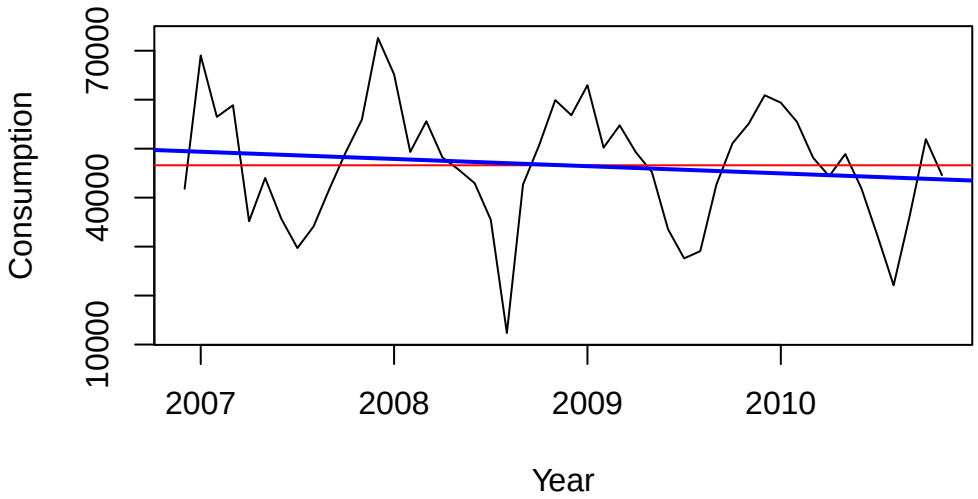
Plot and examine the main features of the graph, checking in particular whether there is (i) a trend; (ii) a seasonal component, (iii) any apparent sharp changes in behavior.

The dataset was prepared for time series analysis through a sequence of data processing steps. The `date` column was converted to the Date format suitable for R, and rows with missing values were omitted to ensure data quality. The `Global_active_power` variable was cast to numeric and daily power consumption was aggregated from the cleaned data.

After cleaning the data, a time series object was constructed which represent daily usage starting from December 2006.

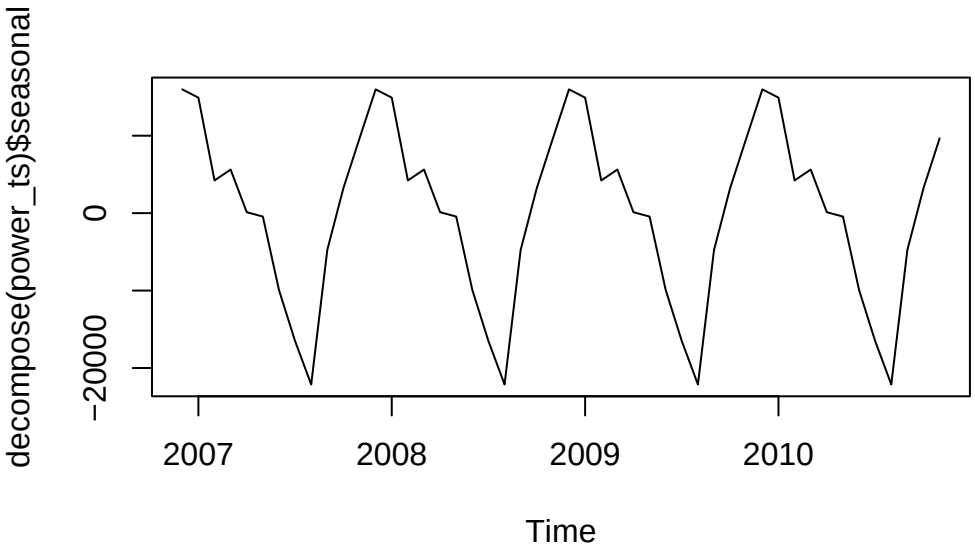


Trend: This data appears to fluctuate quite significantly over time, and trend in this graph is not significant.



To fit a trend line, I use the `lm()` function to perform a linear regression and added a regression line and a mean line to the plot. From the above plot, we could find that the time series data has a slight downward trend.

Seasonal component: There is a significant regular up and down pattern that repeats approximately yearly, which suggest a seasonal component. For further investigation, I employed the `decompose()` function.

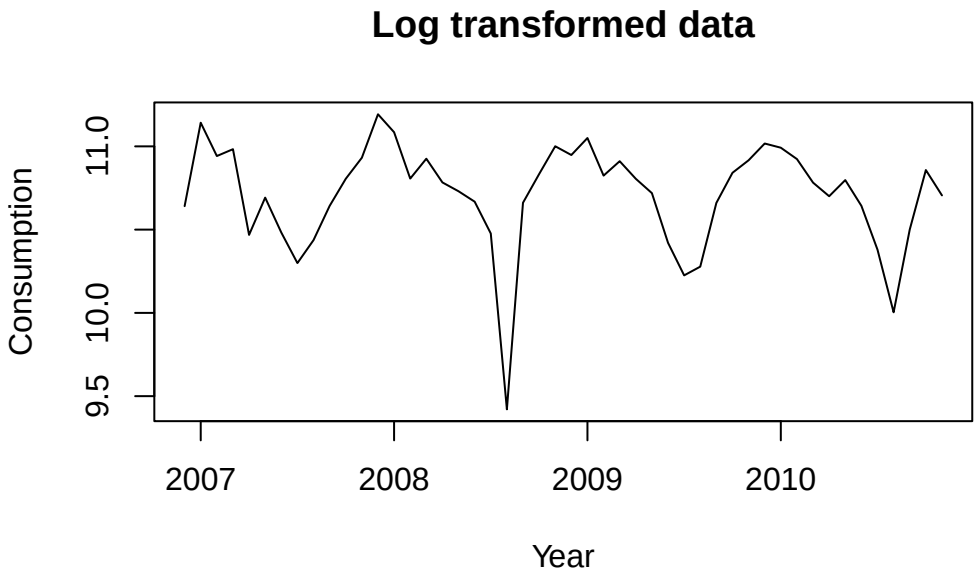


From the plot, we notice that there's a clear seasonal pattern given by the seasonal part of the decomposition plot, where the observations regularly go up and down.

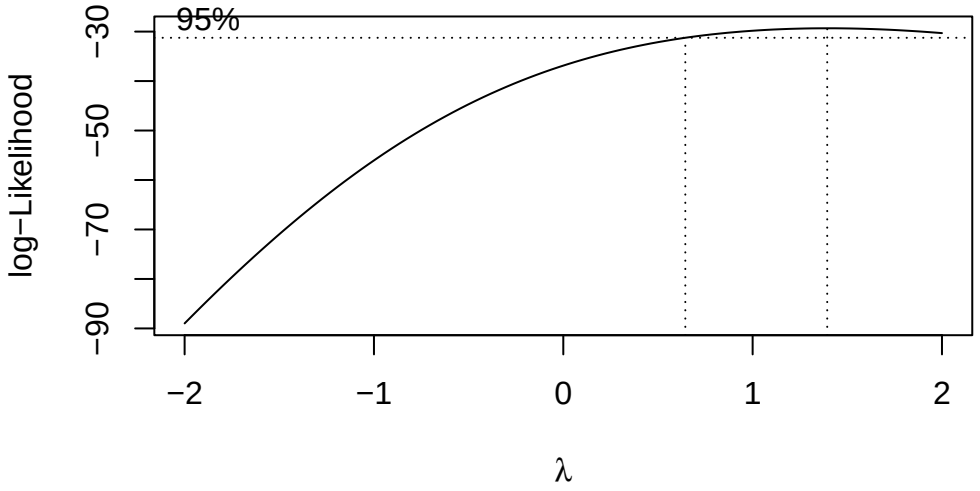
Apparent sharp change: there's no significant sharp change in the time series data.

Transformation

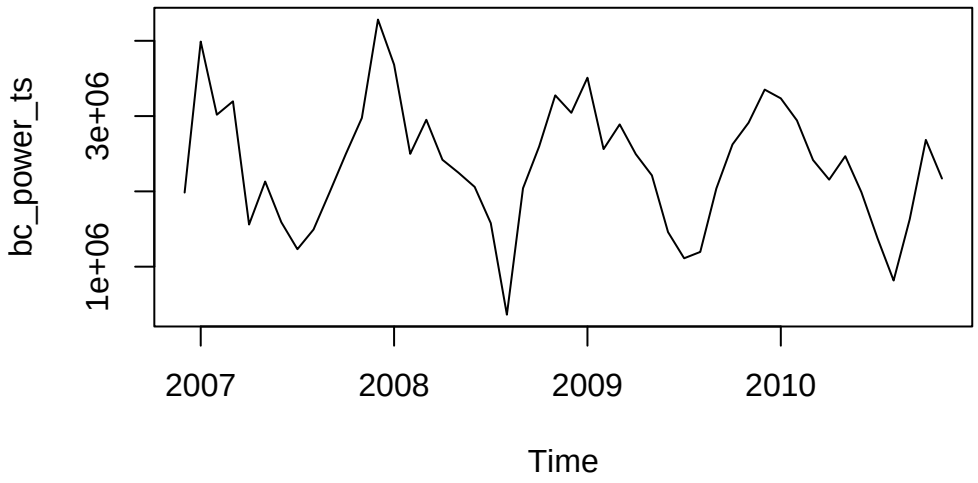
Notice that this series has non-constant extremely large variance $\sigma_1^2 = 368230.1$, I employed a log transformation to stabilize the variance, and reduced the variance to $\sigma_{log}^2 = 0.23935$.



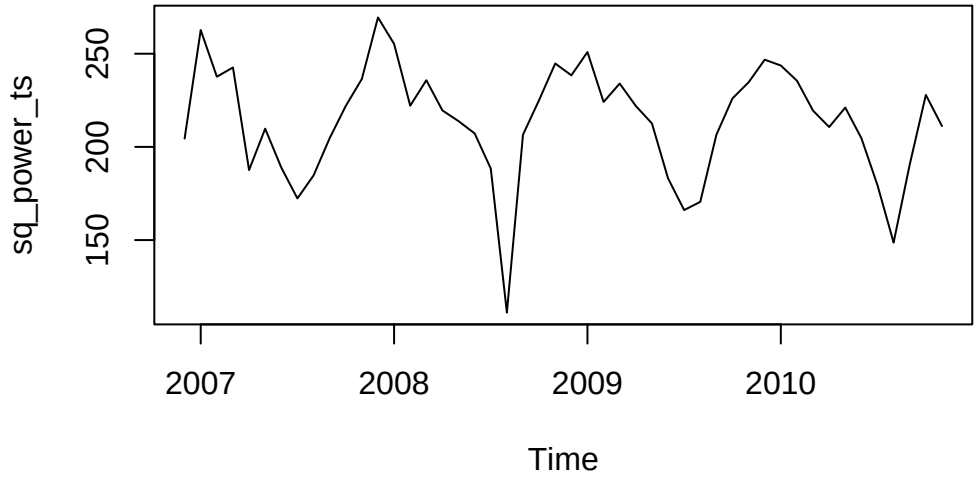
Another approach is using the Box-cox transformation. We wish to apply box-cox transformation to stabilize the variance in a time series. First we find the optimized λ.



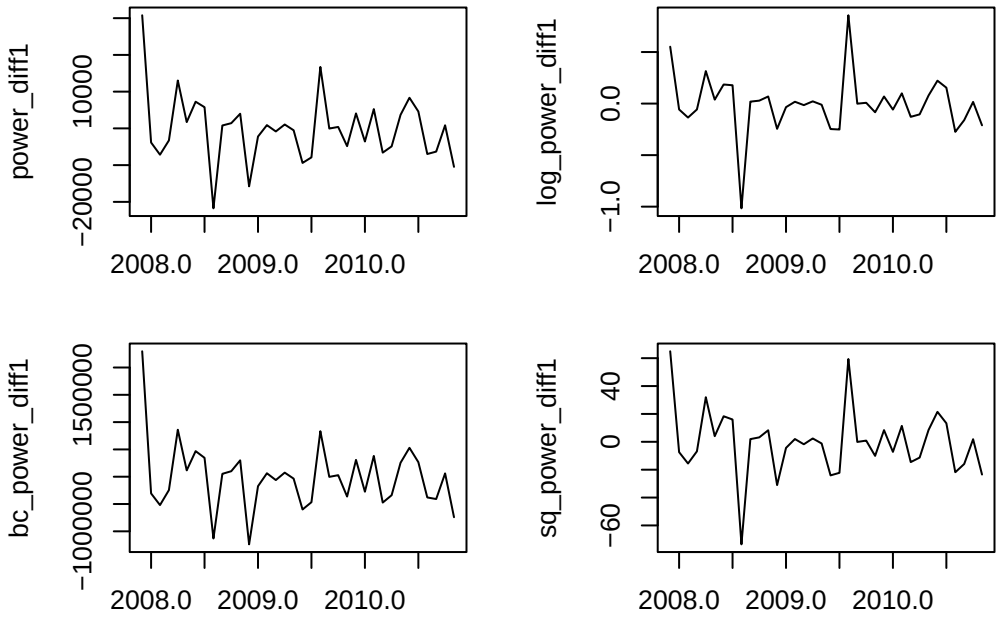
Then we applied box-cox transformation on data.



Also we can use the Square root transform to reduce the variance>



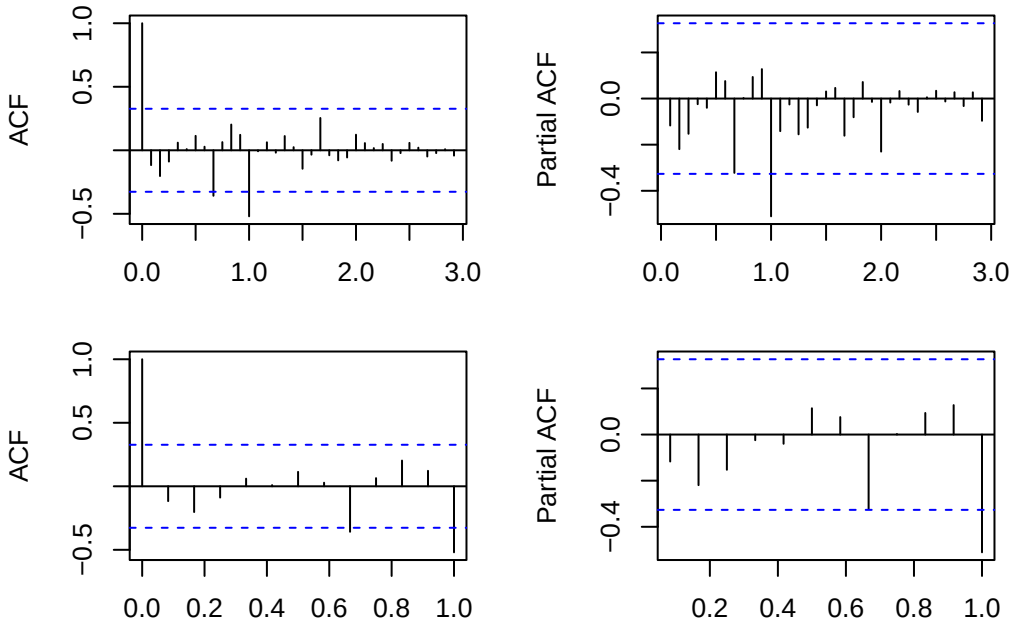
We need to eliminate the trend and the seasonal components in the `power_ts` time-series. Since this time series has no significant trend but a significant yearly patterns. We difference it at a lag of 12 (the yearly period) to eliminate the seasonal patterns in the data.



Notice that `log transformation` performance well and gives the smallest variance of data. Hence we choose `log transformation` with difference at a lag of 12 for further investigation.

ACF and PACF

The acf and pacf plots of the processed data are as follows:



Modeling the seasonal part by the following:

- Given $Y_t = (1 - B^1 2)X_t$, since we have applied one seasonal difference and hence $D = 1$ and $s = 12$.
- Notice that the ACF shows a strong peak at $h = 1s$, a good choice for the SMA part could be $Q = 1$.
- The PACF shows one strong peaks at $h = 1s$. A good choice for the SAR part could be $P = 1$.

Modeling the unseasonal part by the following:

- Given $Y_t = (1 - B^1 2)X_t$, since we haven't applied any unseasonal difference hence $d = 0$.
- The ACF cuts off at lag $k = 8$ hence for the MA part could be $q = 8$.
- The $PACF$ cuts off at lag $k = 7$. Since at $k = 7$ the pacf is slightly significant hence for the AR part could be $p = 0$ or $p = 7$.

Base on the information I get, I assume that the corresponding preliminary identified models is $SARIMA(1, 1, 1) \times (0, 0, 8)_{12}$ and $SARIMA(1, 1, 1) \times (7, 0, 8)_{12}$.

Problem 3: Forecast the value for Quarter 4 of 2023.

Give full explanation on how you arrived to your answer. Show calculations.

Response: based on the estimated coefficients table, the fitted $ARMA(3,0)$ model is as follows:

$$X_t - 2.637 = 0.252(X_{t-1} - 2.637) + 0.061(X_{t-2} - 2.637) - 0.202(X_{t-3} - 2.637) + Z_t$$

To forecast the value of Quarter 4 of 2023, consider it as X_t. By plug the value of Quarter of 2023 into the model, we get the following:

$$\begin{aligned}
X_t &= 2.637 + 0.252(X_{t-1} - 2.637) + 0.061(X_{t-2} - 2.637) - 0.202(X_{t-3} - 2.637) + Z_t \\
\hat{X}_4 &= \mathbb{E}\{X_4 | X_1, \dots, X_3\} \\
&= \mathbb{E}[2.637 + 0.252(X_3 - 2.637) + 0.061(X_2 - 2.637) - 0.202(X_1 - 2.637) + Z_t \\
&= 2.637 + 0.252(2.93 - 2.637) + 0.061(4.62 - 2.637) - 0.202(2.12 - 2.637) + 0 \\
&= 2.936233
\end{aligned}$$

Hence we get the expected value of Quarter 4 of 2023 is 2.936233, which is less than 3.

Problem 4: Calculate the forecasted value of X_{T+1}

A researcher uses an $AR(1)$ model with mean 0 to forecast X_{T+1} when the last known data point is $X_T = -0.3$. It is known that the model has acf $\rho_X(3) = -0.125$. Calculate the forecasted value of X_{T+1}.

By the property of $AR(1)$ model which, $\rho(k) = \phi_1^k$, and $\mu_X = \mathbb{E}[X_t] = \mu$. Take $\rho(3) = -0.125$ back to equation we get $\phi_1^3 = -0.125 \Rightarrow \phi_1 = \sqrt[3]{-0.125} = -0.5$. Which implies that this $AR(1)$ model is given by

$$X_t + 0.5X_{t-1} = Z_t$$

Hence we could find the forecasted(expected) value of X_{T+1} as,

$$\begin{aligned}
\hat{X}_{T+1} &= \mathbb{E}\{X_{T+1}|X_T\} \\
&= \mathbb{E}[-0.5X_T + Z_T] \\
&= \mathbb{E}[-0.5X_T] + \mathbb{E}[Z_t] \\
&= -0.5 \times -0.3 + 0 \\
&= 0.15
\end{aligned}$$

Problem 5: Selection of model

Working on her project, Mary fitted eight models to her data. The models are ranked using the Akaike Information Criterion corrected for bias (AICc) and are listed below. Mary would like to select the three models that exhibit the lowest AICc values for further analysis.

The three model with the lowest AICc values are: $SARIMA(0,1,0) \times (1,1,1)_{12}$ with corresponding $AICc = -14.72075$, $SARIMA(0,1,0) \times (2,1,1)_{12}$ with corresponding $AICc = -14.71839$, and $SARIMA(5,1,0) \times (2,1,1)_{12}$ with corresponding $AICc = -14.70548$.

The general from of the three models can be written as

$SARIMA(0,1,0)\times(1,1,1)_{12}$, with seasonal components $P = 1$, $D = 1$, $Q = 1$, and unseasonal components $p = 0$, $d = 1$, $q = 0$, with a seasonal cycle $s = 12$ is given by:

$$(1 - \Phi_1 B^{12})(1 - B)(1 - B^{12})X_t = (1 + \Theta_1 B^{12})Z_t$$

$SARIMA(0,1,0)\times(2,1,1)_{12}$, with seasonal components $P = 2$, $D = 1$, $Q = 1$, and unseasonal components $p = 0$, $d = 1$, $q = 0$, with a seasonal cycle $s = 12$ is given by:

$$(1 - \Phi_1 B^{12} - \Phi_2 B^{24})(1 - B)(1 - B^{12})X_t = (1 + \Theta_1 B^{12})Z_t$$

$SARIMA(5,1,0)\times(2,1,1)_{12}$, with seasonal components $P = 2$, $D = 1$, $Q = 1$, and unseasonal components $p = 5$, $d = 1$, $q = 0$, with a seasonal cycle $s = 12$ is given by:

$$(1-\phi_1 B-\phi_2 B^2-\phi_3 B^3-\phi_4 B^4-\phi_5 B^5)(1-\Phi_1 B^{12}-\Phi_2 B^{24})(1-B)(1-B^{12})X_t = (1+\Theta_1 B^{12})Z_t$$

Problem G1: Confidence interval of $\rho(1)$ and $\rho(2)$

Suppose that in a sample of size 100, we obtain $\rho(1) = 0.438$ and $\rho(2) = 0.145$. Assuming that the data were generated from an $MA(1)$ model, construct approximate 95% confidence intervals for both $\rho(1)$ and $\rho(2)$. Based on these two confidence intervals, are the data consistent with an $MA(1)$ model with $\rho = 0.6$?

Response: For $MA(1)$ process, the general formula is given by $X_t = Z_t + \theta Z_{t-1}$ with ACF at and only at lags 0 and 1 are non-zero. By *10.3*, for a large sample size n here $n = 100$, the distribution of $\hat{\rho}(h)$ is close to Gaussian.

Now given the sample acf $\underline{\hat{\rho}}_h$ is apporximately $\mathcal{N}(\underline{\rho}_h, n^{-1}W)$. Then the 95% confidence intervals for $\underline{\hat{\rho}}_h$ is given by $\underline{\hat{\rho}}_h \pm 1.96\sqrt{Var(\hat{\rho}(h))} = \underline{\hat{\rho}}_h \pm 1.96 \times \sqrt{n^{-1}W_{ij}}$.

Notice that $W = w_{i,j}{}_{i,j}$ is following by

$$w_{i,j} = \sum_{k=1}^{\infty}\{\rho(k+i) + \rho(k-i) - 2\rho(i)\rho(k)\} \times \{rho(k+j) + \rho(k-j) - 2\rho(j)\rho(k)\}$$

Hence we can caluclate the corresponding $w_{1,1}$ and $w_{2,2}$ to find the confidence intervals of $\rho(1)$ and $\rho(2)$.

$$\begin{aligned}
w_{1,1} &= \sum_{k=1}^{\infty}\{\rho(k+1) + \rho(k-1) - 2\rho(1)\rho(k)\} \times \{\rho(k+1) + \rho(k-1) - 2\rho(1)\rho(k)\} \\
&= \sum_{k=1}^{\infty}\{\rho(k+1) + \rho(k-1) - 2\rho(1)\rho(k)\}^2 \\
&= [\rho(2) + \rho(0) - 2\rho(1)^2]^2 + [\rho(3) + \rho(1) - 2\rho(1)\rho(2)]^2 + ... \\
&= [\rho(0) - 2\rho(1)^2]^2 + \rho(1)^2 \\
&= \rho(0)^2 - 4\rho(0)\rho(1)^2 + 4\rho(1)^4 + \rho(1)^2 \\
&= 1 - 3\rho(1)^2 + 4\rho(1)^4
\end{aligned}$$

Given that the data is generated by $MA(1)$ model with $\theta = 0.6$. The theoretical $\rho(1) = \frac{\theta_1}{1+\theta_1^2} = 0.4411765$. Hence plug in $\rho_1 = 0.4411765$, we could find that $w_{1,1} = 1 - 3(0.44118)^2 + 4(0.44118)^4 = 0.5676237$. Then the confidence level of $\hat{\rho}(1) = 0.44118 \pm 1.96(\sqrt{\frac{0.571685}{100}}) = (0.2929847, 0.5893753)$.

Similarily the confidence interval of $\rho(2)$ can be calculated by:

$$
\begin{aligned}
w_{2,2} &= \sum_{k=1}^{\infty}\{\rho(k+2)+\rho(k-2)-2\rho(2)\rho(k)\}\times\{\rho(k+2)+\rho(k-2)-2\rho(2)\rho(k)\} \\
&= \sum_{k=1}^{\infty}\{\rho(k+2)+\rho(k-2)-2\rho(2)\rho(k)\}^2 \\
&= \sum_{k=1}^{\infty}\{\rho(k+2)+\rho(k-2)\}^2 \\
&= [\rho(3)+\rho(-1)]^2+[\rho(4)+\rho(0)]^2+[\rho(5)+\rho(1)]^2+... \\
&= 2\rho(1)^2+\rho(0)^2+0 \\
&= 1+2\rho(1)^2
\end{aligned}
$$

Hence plug in $\rho_1 = 0.4411765$, we could find that $w_{2,2} = 1+2\rho(1)^2 = 1.389273$. The confidence level of $\widehat{\rho}(2) = 0 \pm 1.96 \times \sqrt{\frac{1.389273}{100}} = (-0.23102, 0.23102)$.

Notice that sample acf $\widehat{\rho}(1)$ falls in the confidence level of $\rho(1)$. Also we could find that $|\widehat{\rho}(h)| = 0.145 < 1.96 \times \sqrt{\frac{1.389}{100}} = 0.2310202$. We could assume that $\rho(2)$ is almost 0 (significantly different from zero), which consists with our assumption that this model is an $MA(1)$ model. We could conclude that the data consistent with the $MA(1)$ model with $\theta = 0.6$.