STA457-Assignment 2[*]

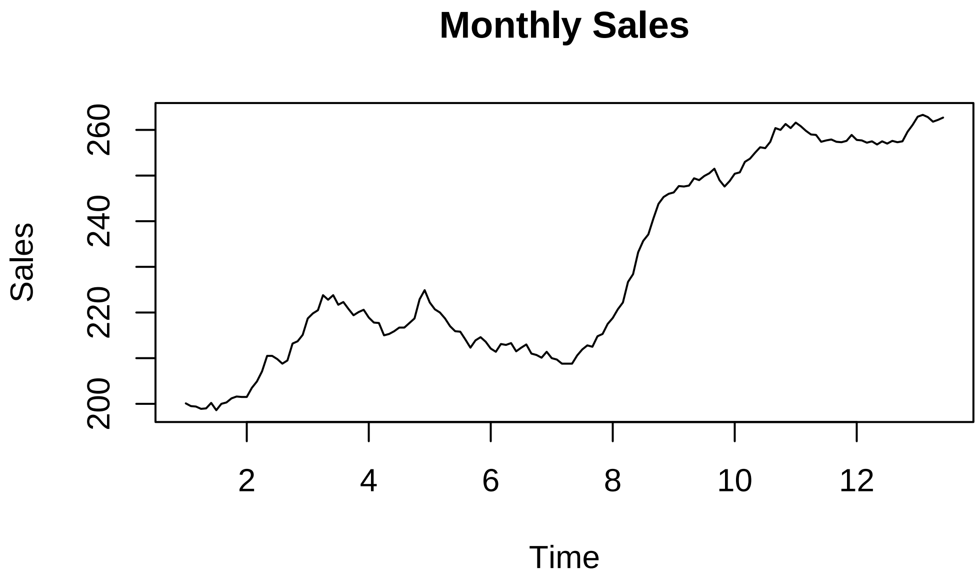
Hyuk Jang

```
# Code Space Setup
library(ggplot2)
library(tidyverse)
library(astsa)
library(forecast)
library(knitr)
data(sales)
data(lead)
data(salt)
data(saltemp)
```

Q1(i)

```
#Q1(i)
# plot the time series(sales)
sales_ts <- ts(sales, frequency = 12)  # monthly data, frequency = 12
plot(sales_ts, main="Monthly Sales", ylab="Sales", xlab="Time")
```

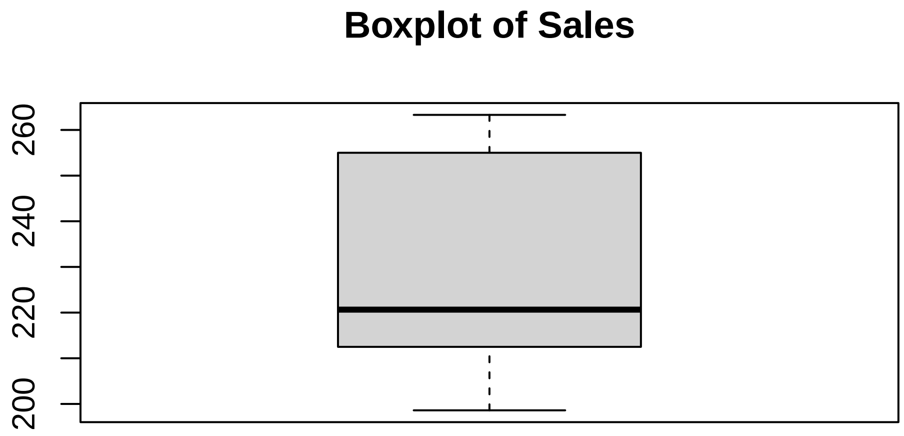
[*] If there are any issues please check: https://github.com/anggimude/Time-series-ARIMA.git



```
#Q1(i)
# summary of time series(sales)
summary(sales_ts)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  198.6   212.6   220.7   230.0   254.7   263.3 
```

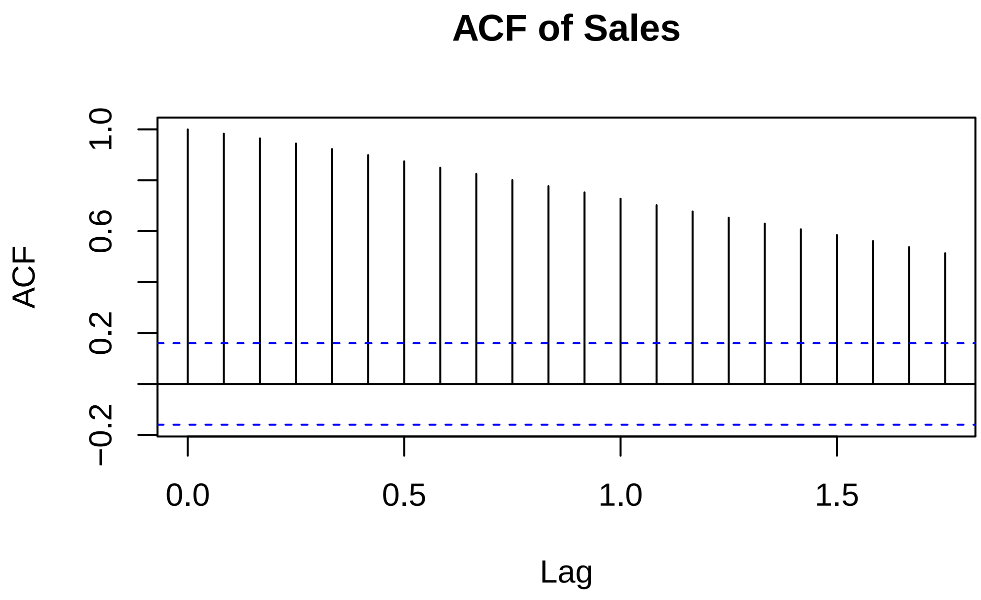
```
#Q1(i)
# boxplot of sales
boxplot(sales_ts, main="Boxplot of Sales")
```



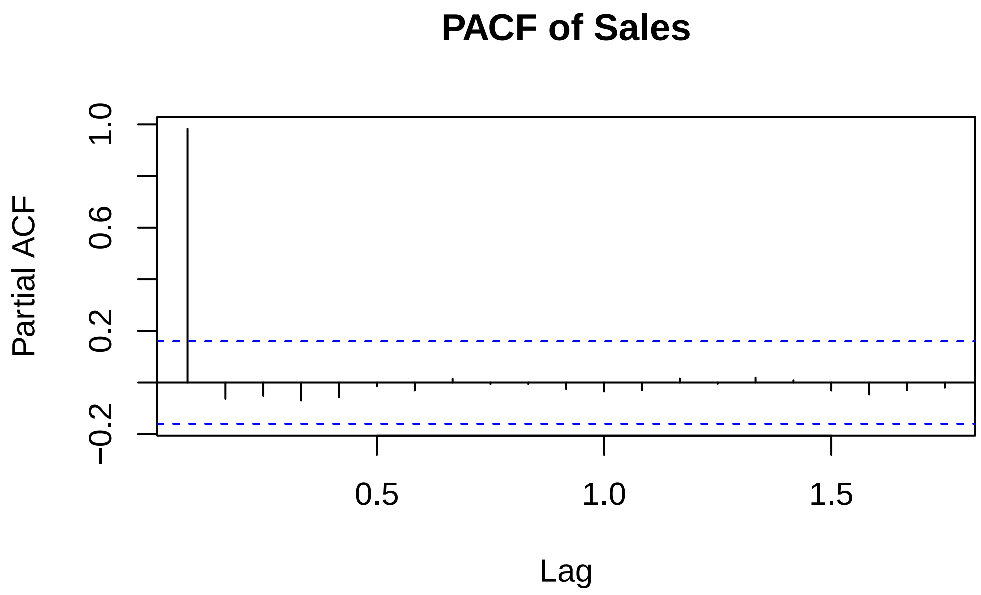
The plot of the monthly sales data (S_t) shows a gradual upward trend with no clear seasonality elements. The data has a mean at 230, with a maximum at 263.3 and minimum at 198.6. At t = 50, there is a steep increase sales. The boxplot suggests the data is slightly skewed towards the maximum but does not exhibit extreme outliers. Strong seasonality is not obvious and the trend recommends that it may not be stationary.

Q1(ii)

```
#Q1(ii)
# ACF of sales
acf(sales_ts, main="ACF of Sales")
```



```r
#Q1(ii)
# pacf of sales
pacf(sales_ts, main="PACF of Sales")
```

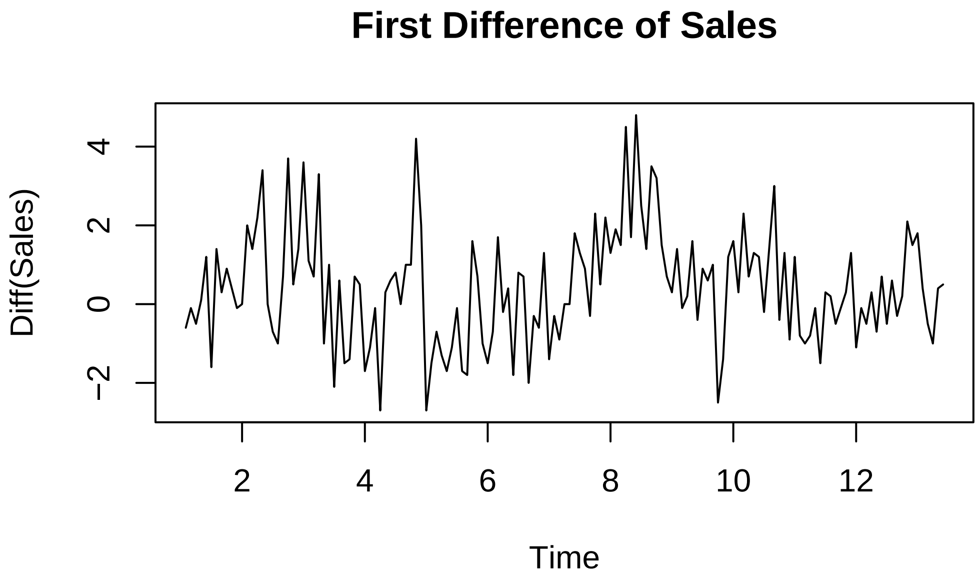


```
#Q1(ii)
# Dickey-Fuller test of sales
library(tseries)
adf.test(sales_ts)
```

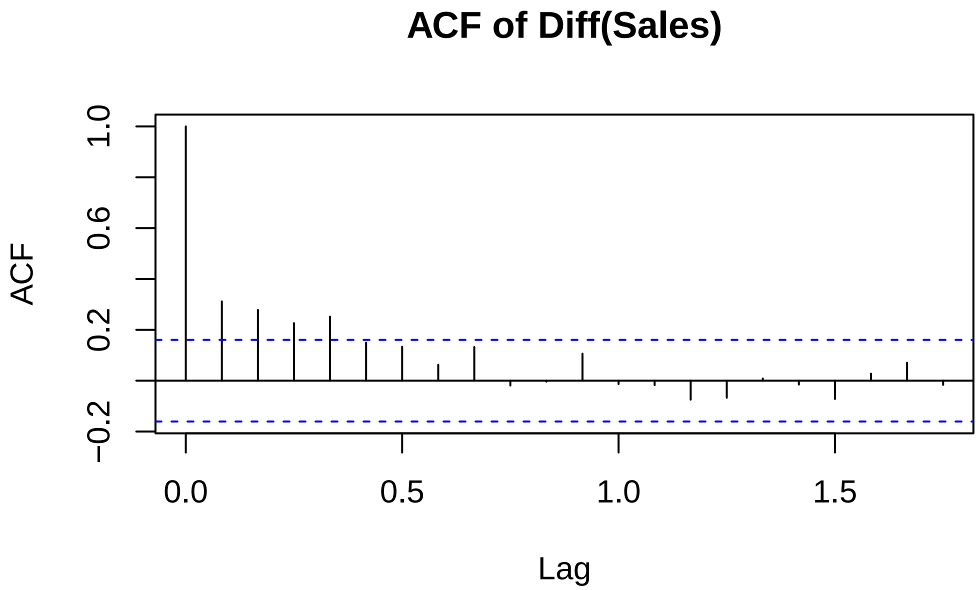
```

    Augmented Dickey-Fuller Test

data:  sales_ts
Dickey-Fuller = -2.1109, Lag order = 5, p-value = 0.5302
alternative hypothesis: stationary
```

The ACF and the PACF plot of the sales time series data suggests that it is potentially non-stationary as the ACF decays slowly and the p-value of the unit root test is significantly higher than 0.05 implying sales_ts must be non-stationary. Thus, we difference the data before fitting a model.

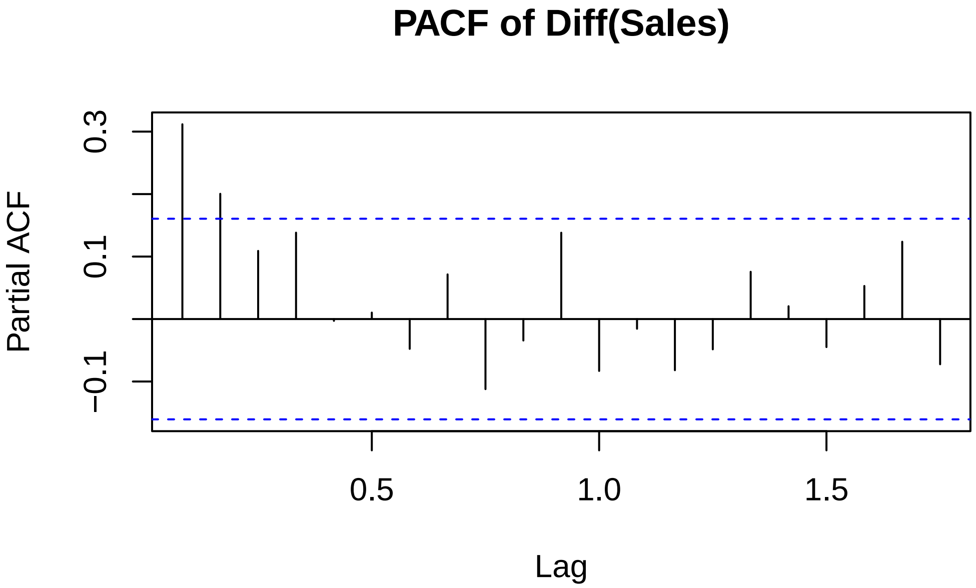
```
#Q1(ii)
# plot first differences of sales
sales_diff <- diff(sales_ts)
plot(sales_diff, main="First Difference of Sales", ylab="Diff(Sales)")
```



```
#Q1(ii)
# acf of differenced sales
acf(sales_diff, main="ACF of Diff(Sales)")
```



```
#Q1(ii)
# pacf of differenced sales
pacf(sales_diff, main="PACF of Diff(Sales)")
```

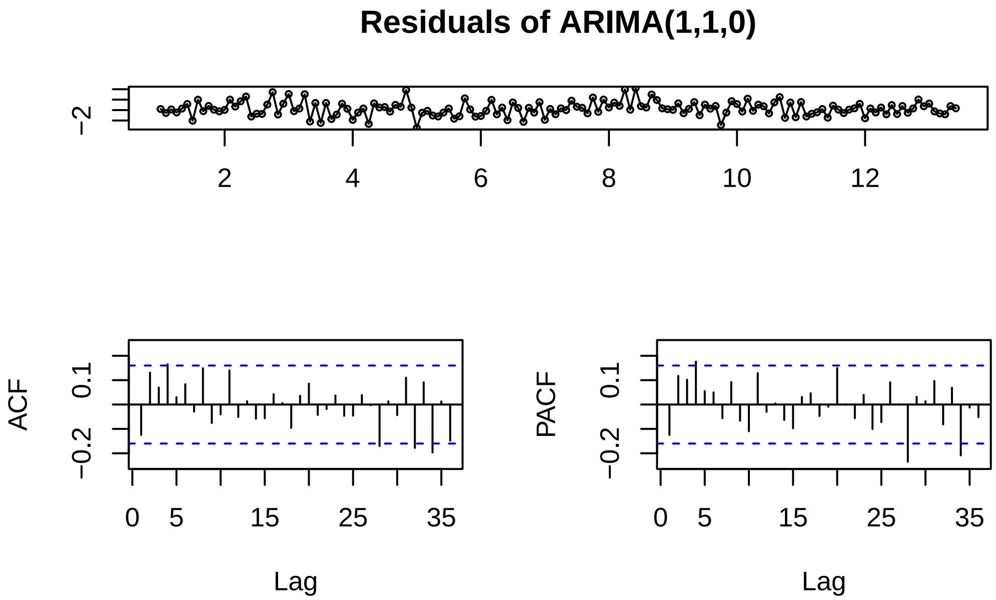


The ACF spikes at lag 1, then is mostly within the bounds. The PACF of the diff(sales) spikes at lag 1 as well, then oscillates around zero. This implies an ARIMA(1,1,0), ARIMA(0,1,1) are possible choices to fit the data

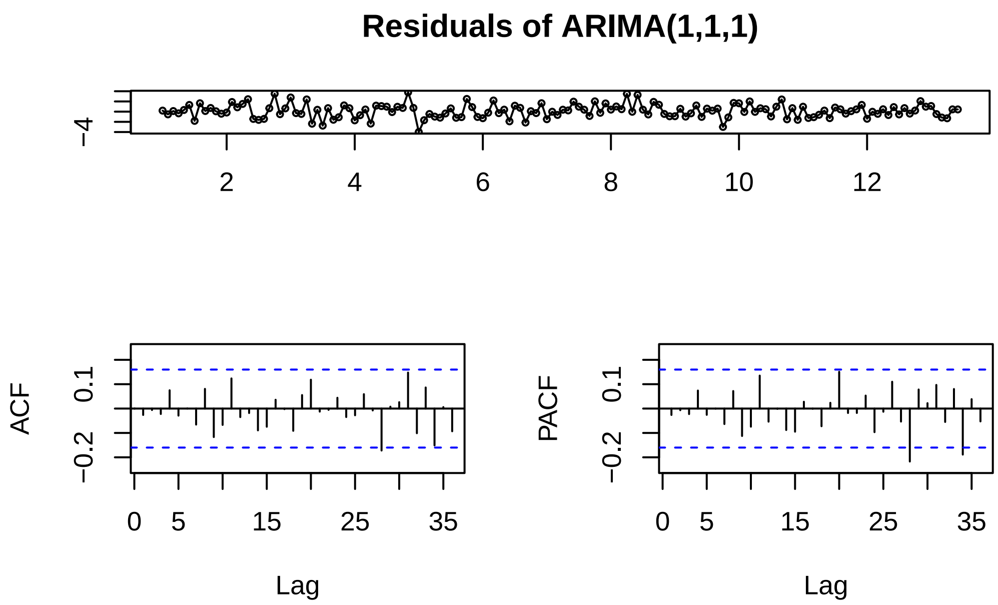
Q2(iii) The two models selected are ARIMA(1,1,0) and ARIMA(1,1,1)

```
#Q2(iii)
# calculae the different arima models
arima110 <- arima(sales_ts, order = c(1, 1, 0))
arima111 <- arima(sales_ts, order = c(1, 1, 1))
arima011 <- arima(sales_ts, order = c(0, 1, 1))
```

```
# plot of residuals of arima(1,1,0)
tsdisplay(residuals(arima110), main="Residuals of ARIMA(1,1,0)")
```



```
#Q2(iii)
# arima(1,1,1) residual plots
tsdisplay(residuals(arima111), main="Residuals of ARIMA(1,1,1)")
```



```r
#Q2(iii)
# create table for AIC values
model_comparison <- data.frame(
  Model = c("ARIMA(1,1,0)", "ARIMA(1,1,1)", "ARIMA(0,1,1)"),
  AIC   = c(AIC(arima110), AIC(arima111), AIC(arima011))
)
print(model_comparison)
```

```
         Model      AIC
1 ARIMA(1,1,0) 526.1264
2 ARIMA(1,1,1) 514.7360
3 ARIMA(0,1,1) 533.2657
```

The AIC values suggest that ARIMA(1,1,1) is the best model because it has the lowest values.

Q2(iv)

$(1 - \phi_1\ B)(1 - B)(S_t - \mu) = (1 + \theta_1\ \mathrm{B})e_t, \quad e_t \sim \text{Normal}(0, \sigma^2),$
$\Delta S_t = \mu + \phi_1\ \Delta S_{t-1} + e_t + \theta_1 e_{t-1},$

Backshift notation: The backshift operator B is defined by: $B\ S_t = S_{t-1}$.

First difference: The first difference of the series is given by $\Delta S_t = S_t - S_{t-1}$.

AR component: ϕ_1 is the autoregressive (AR) parameter applied to ΔS_{t-1}, the first-differenced series.

MA component: θ_1 is the moving average (MA) parameter applied to the previous error term e_{t-1},

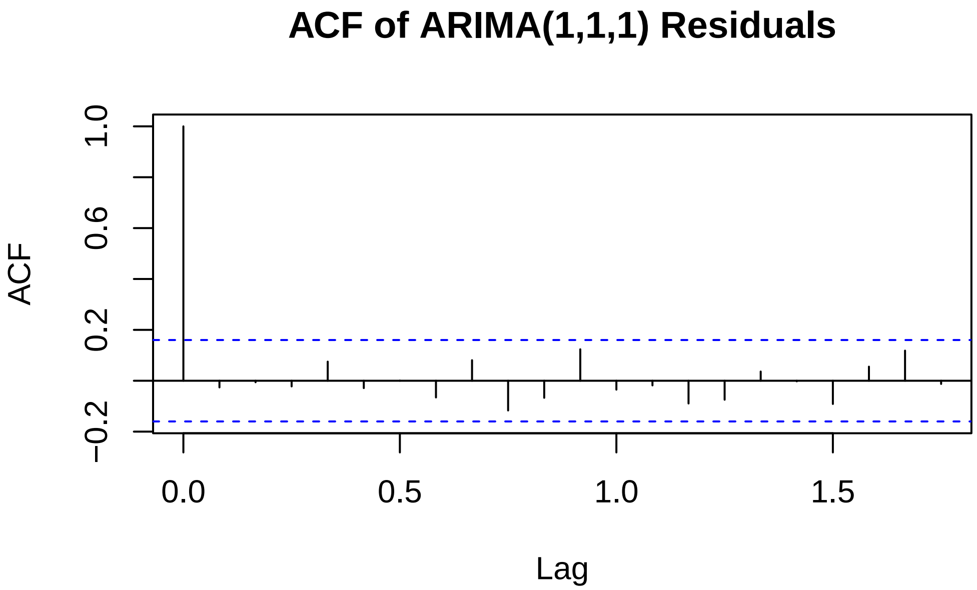
White Noise: The error term e_t is assumed to be white noise, that is, $e_t \sim \text{Normal}(0, \sigma^2)$.

Q3(v)

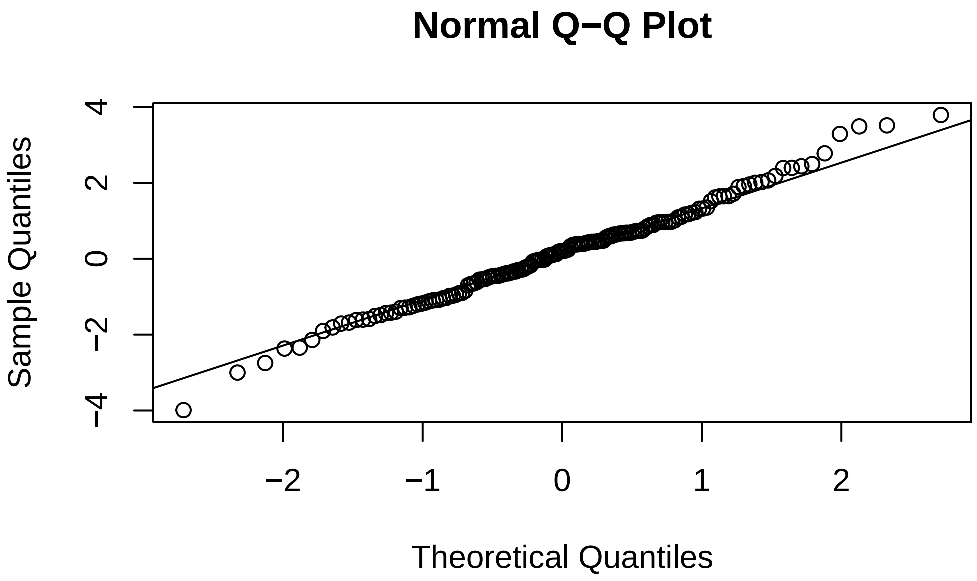
One-step-ahead forecast is obtained by taking expectations at time t, noting that e_{t+1} has mean 0) is: $\hat{S}_{t+1|t} = S_t + \mu + \phi_1(S_t - St - 1) + \theta_1 e_t$.

Q3(vi)

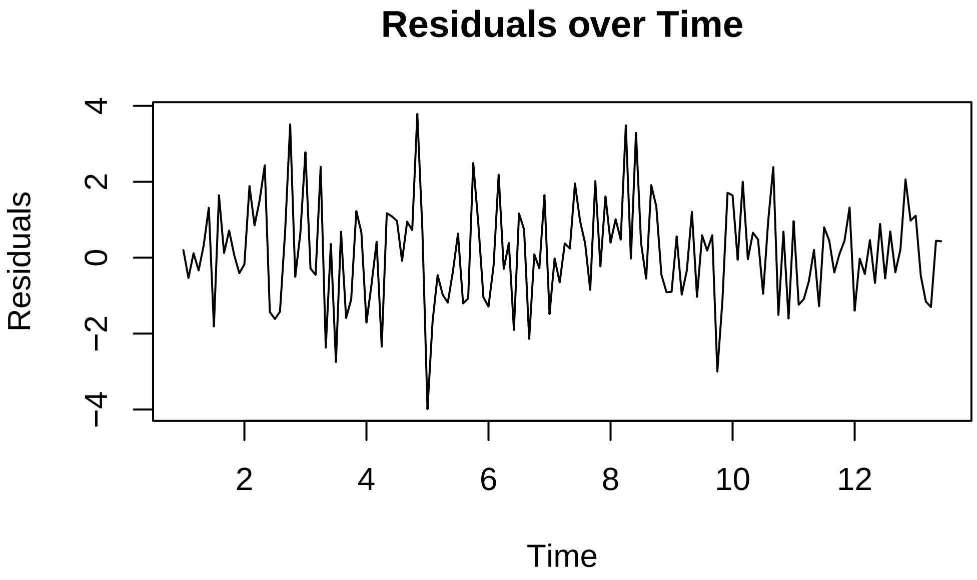
```
#Q3(vi)
# ACF of residuals of arima(1,1,1)
acf(residuals(arima111), main="ACF of ARIMA(1,1,1) Residuals")
```



```
#Q3(vi)
# Q-Q plot of residuals arima(1,1,1)
qqnorm(residuals(arima111))
qqline(residuals(arima111))
```



```
#Q3(vi)
# residual plot of arima(1,1,1)
plot(residuals(arima111), main="Residuals over Time", ylab="Residuals")
```



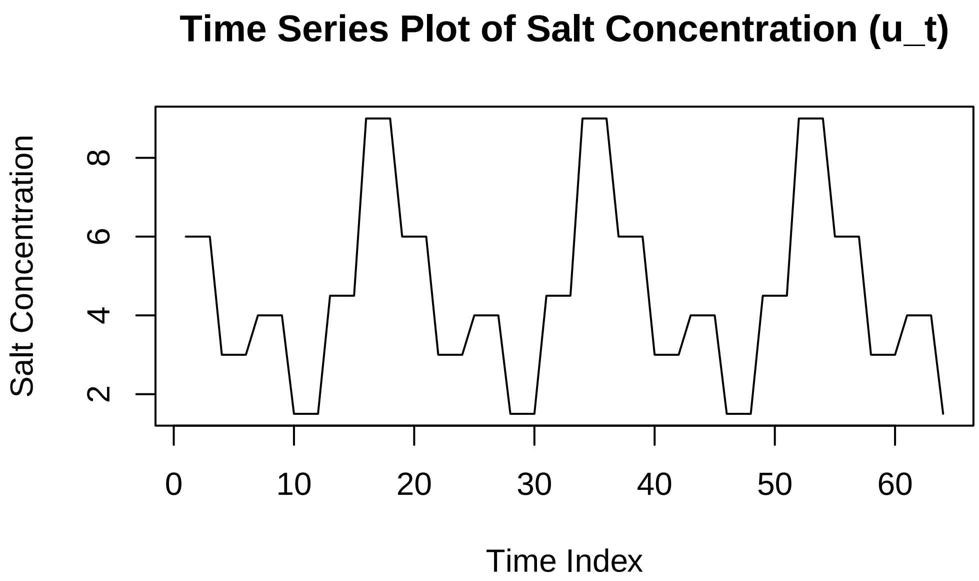
```
#Q3(vi)
# Ljung box test of arima(1,1,1)
Box.test(residuals(arima111), lag=10, type="Ljung")
```

```

    Box-Ljung test

data:  residuals(arima111)
X-squared = 5.8814, df = 10, p-value = 0.8251
```

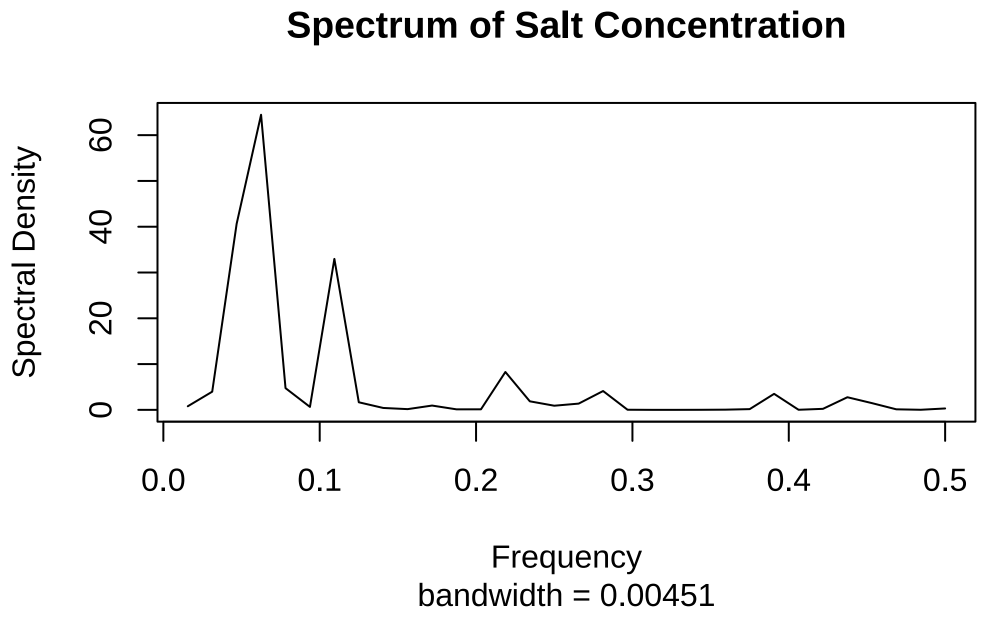
The ACF of the residuals do not show any significant spikes throughout the lags, Q-Q plot is almost normal, and the p-value from the Box-Ljung test doesn't reject the null hypothesis (no autocorrelation) implying ARIMA(1,1,1) must be an appropriate model.

Q4(a)

```
#Q4(a)
# plot the time series of salt
plot.ts(salt,
        main = "Time Series Plot of Salt Concentration (u_t)",
        ylab = "Salt Concentration",
        xlab = "Time Index")
```



```
#Q4(a)
# spectrum of salt
salt.spec <- spectrum(salt,
                      main = "Spectrum of Salt Concentration",
                      xlab = "Frequency",
                      ylab = "Spectral Density",
                      log = "no")  # 'log="no"' to avoid log-scaling on the y-axis
```



```
#Q4(a)
# find predominant frequency
peak_index <- which.max(salt.spec$spec)
omega_hat  <- salt.spec$freq[peak_index]
spec_hat   <- salt.spec$spec[peak_index]

cat("Predominant frequency (omega_hat) =", omega_hat, "\n")
```

```
Predominant frequency (omega_hat) = 0.0625
```

Thus, predominant frequency $w_u = 0.0625$

Q4(b)

```
#Q4(b)
# Degrees of freedom of spectrum of salt
df_val <- salt.spec$df


# Compute lower and upper 95% bounds
lower_95 <- spec_hat * df_val / qchisq(0.975, df_val)
upper_95 <- spec_hat * df_val / qchisq(0.025, df_val)
```

```r
cat("Degrees of freedom (df) =", df_val, "\n")
```

```
Degrees of freedom (df) = 1.79159
```

```r
cat("95% CI for the spectrum at omega_hat: [", lower_95, ",", upper_95, "]\n")
```

```
95% CI for the spectrum at omega_hat: [ 16.66115 , 3680.941 ]
```

The spectral analysis peaks around the frequency of 0.0625 which corresponds to around 16 time units. The confidence interval of [16.661, 36880.941] is a large because the degree of freedom is low. This indicates high uncertainty of the exact magnitude of the peak but since a peak exists, it suggests that there is a meaningful cycle at 16 time units.