stat443 assignment2

Yiming Shen 20891774

13/02/2023

Problem 1

```
library(TSA)
```

```
##
## Attaching package: 'TSA'

## The following objects are masked from 'package:stats':
##
##     acf, arima

## The following object is masked from 'package:utils':
##
##     tar
```

```
data("hours")
```

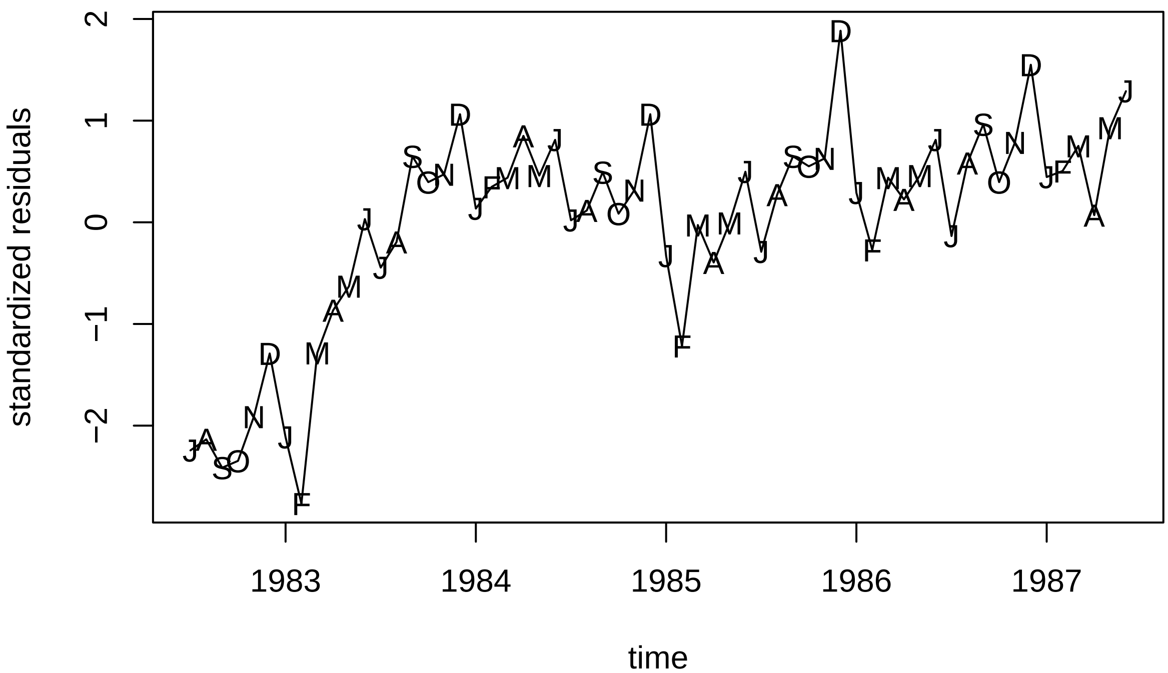
(a)

```
har. <- harmonic(hours)
hours.lm1 <- lm(hours ~ har.)
summary(hours.lm1)
```

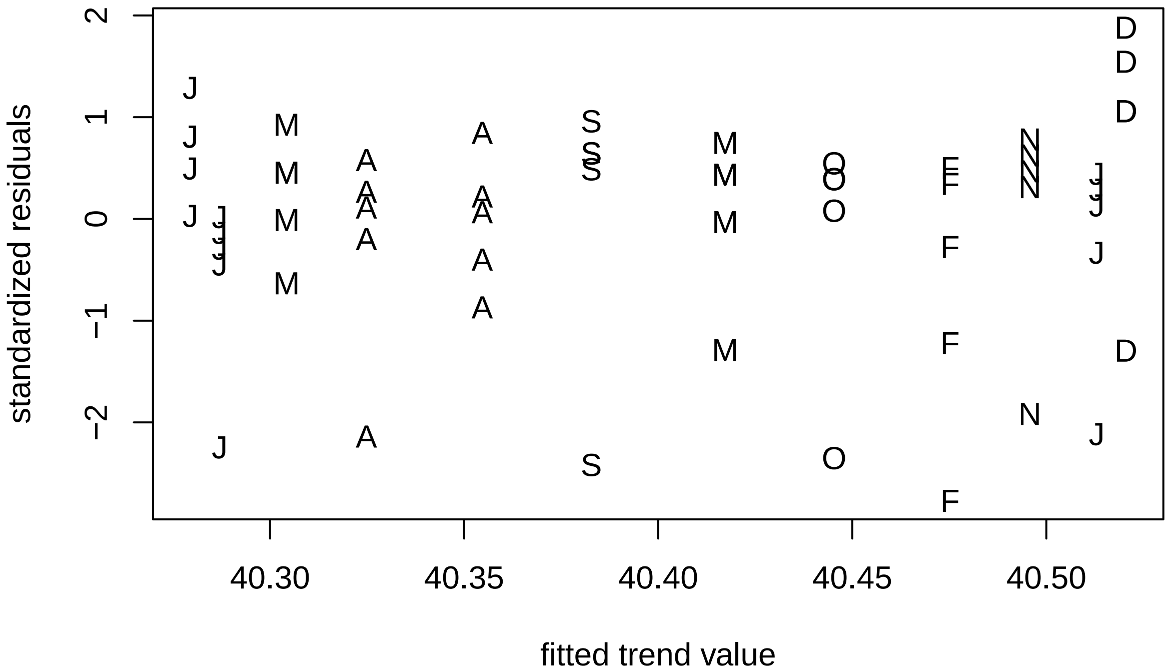
```
##
## Call:
## lm(formula = hours ~ har.)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.6752 -0.1935  0.1956  0.3825  1.1795
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     40.40000    0.08473 476.783   <2e-16 ***
## har.cos(2*pi*t)  0.11297    0.11983   0.943    0.350
## har.sin(2*pi*t) -0.04533    0.11983  -0.378    0.707
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6564 on 57 degrees of freedom
## Multiple R-squared:  0.01778,    Adjusted R-squared:  -0.01668
## F-statistic: 0.5159 on 2 and 57 DF,  p-value: 0.5997
```

(b)

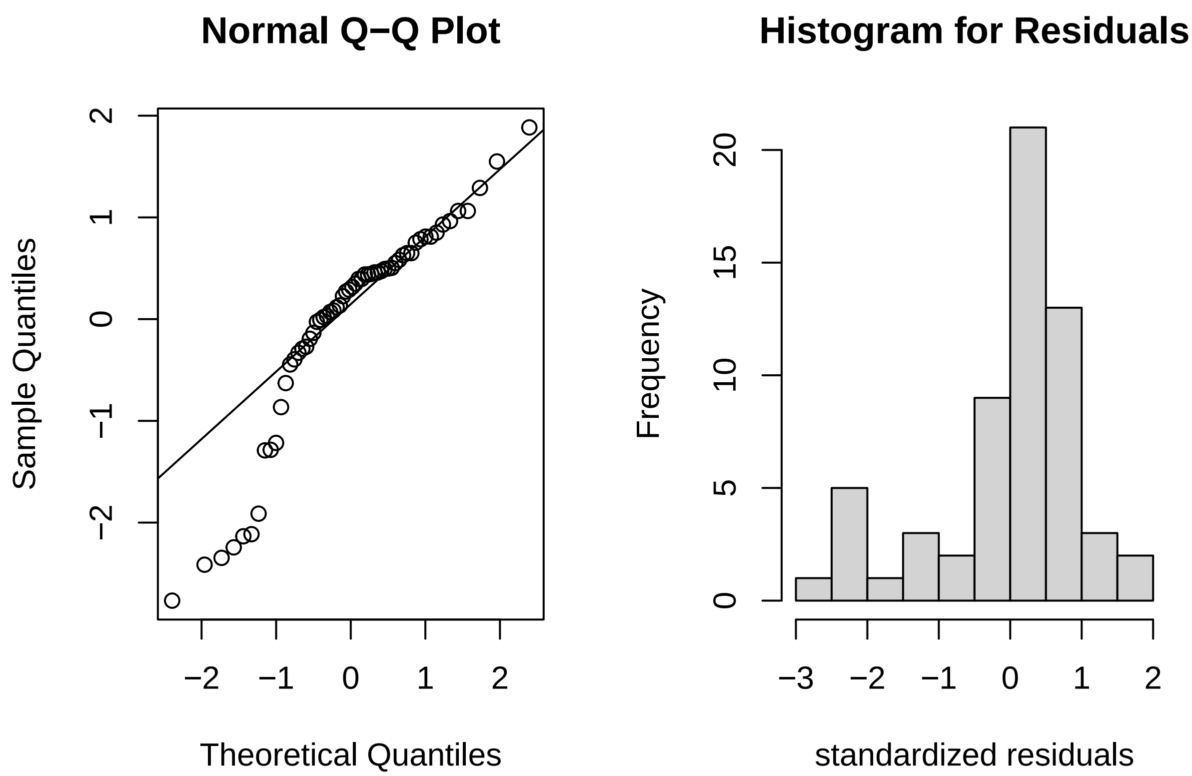
```r
# residual vs. time
plot(y=rstudent(hours.lm1), x=as.vector(time(hours)), type='l',
     xlab='time', ylab='standardized residuals')
points(y=rstudent(hours.lm1), x=as.vector(time(hours)),
       pch=as.vector(season(hours)))
```



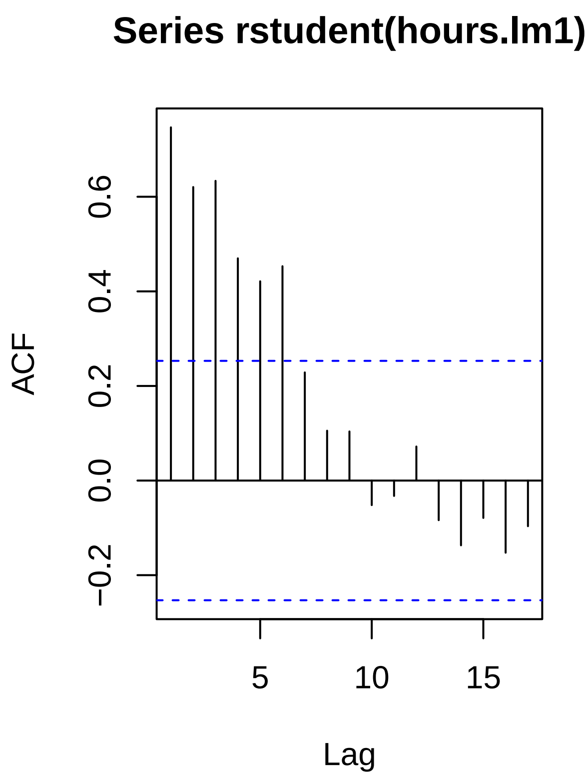

```r
# comments: I observed that there is a quadratic pattern in the residual time
# series but there is no obvious seasonal pattern.

# residual vs. estimated fit
plot(y=rstudent(hours.lm1), x=as.vector(fitted(hours.lm1)),
     xlab='fitted trend value', ylab='standardized residuals', type='n')
points(y=rstudent(hours.lm1), x=as.vector(fitted(hours.lm1)),
       pch=as.vector(season(hours)))
```




```
# comments: It seems to be a stable constant in the variability at different
# estimated levels.

# normality of residuals
par(mfrow=c(1,2))
qqnorm(rstudent(hours.lm1))
qqline(rstudent(hours.lm1))
hist(rstudent(hours.lm1), xlab='standardized residuals',
     main='Histogram for Residuals')
```




```
# comments: There are deviations on lower end of QQ-plot and the residuals are
# not normally distributed.
# The histogram also indicates a left-skewed distribution, which means that
# there is a heavy left tail.

# ACF plot
acf(rstudent(hours.lm1))
# comments: from the ACF plot, since many spikes go beyond the confidence
# band, so a strong correlation among residual values can be observed.
```




```
# test for normality
shapiro.test(rstudent(hours.lm1))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  rstudent(hours.lm1)
## W = 0.88612, p-value = 4.259e-05
```

```
# comments: since p-value = 4.259e-05 < 0.05,
# so we have strong evidence against our null hypothesis: data come from a
# normal distribution.


# test for constant variance
seg <- as.vector(season(hours))
fligner.test(rstudent(hours.lm1), seg)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  rstudent(hours.lm1) and seg
## Fligner-Killeen:med chi-squared = 2.7252, df = 11, p-value = 0.9939
```

```
# comments: since p-value = 0.9939 > 0.05,
# so we have no evidence against our null hypothesis: data have
# constant variance.

# test for randomness
library("randtests")
library("lawstat")
```

```
##
## Attaching package: 'lawstat'

## The following object is masked from 'package:randtests':
##
##     runs.test
```

```r
difference.sign.test(rstudent(hours.lm1))
```

```
##
##  Difference Sign Test
##
## data:  rstudent(hours.lm1)
## statistic = 3.77, n = 60, p-value = 0.0001632
## alternative hypothesis: nonrandomness
```

```r
# comments: since p-value = 0.0001632 < 0.05,
# so we have strong evidence against our null hypothesis: the data is random.
runs.test(rstudent(hours.lm1))
```

```
##
##  Runs Test - Two sided
##
## data:  rstudent(hours.lm1)
## Standardized Runs Statistic = -2.8646, p-value = 0.004176
```

```r
# comments: same as difference sign test, since p-value = 0.004176,
# so we have strong evidence against our null hypothesis: the data is random.
```

(c)

The cosine trend model is not appropriate for the trend estimation. Because from residual vs. time, we observed a quadratic trend. From QQ-plot and histogram, we found that the residuals are not normally distributed, which violated the normality assumption. From ACF plot, a strong correlation among residual values can be observed, which violated the uncorrelation assumption.

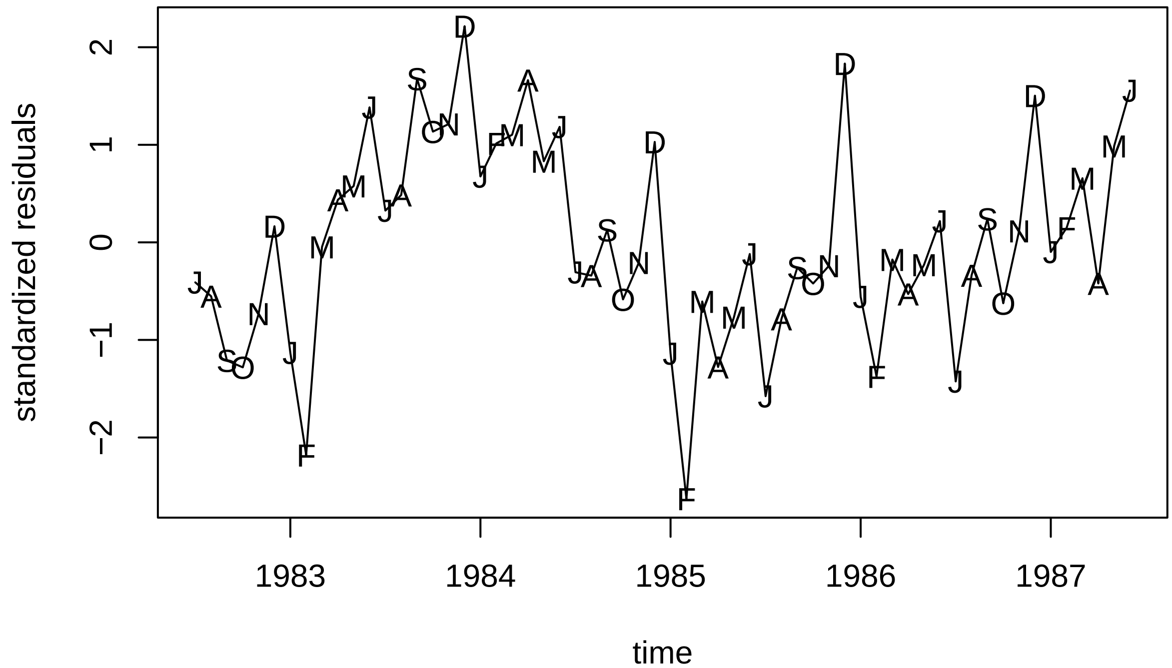
(d)

```r
# fit a combination of cosine trend model and quadratic model
hours.lm2 <- lm(hours~har.+time(hours)+I(time(hours)^2))
summary(hours.lm2)
```

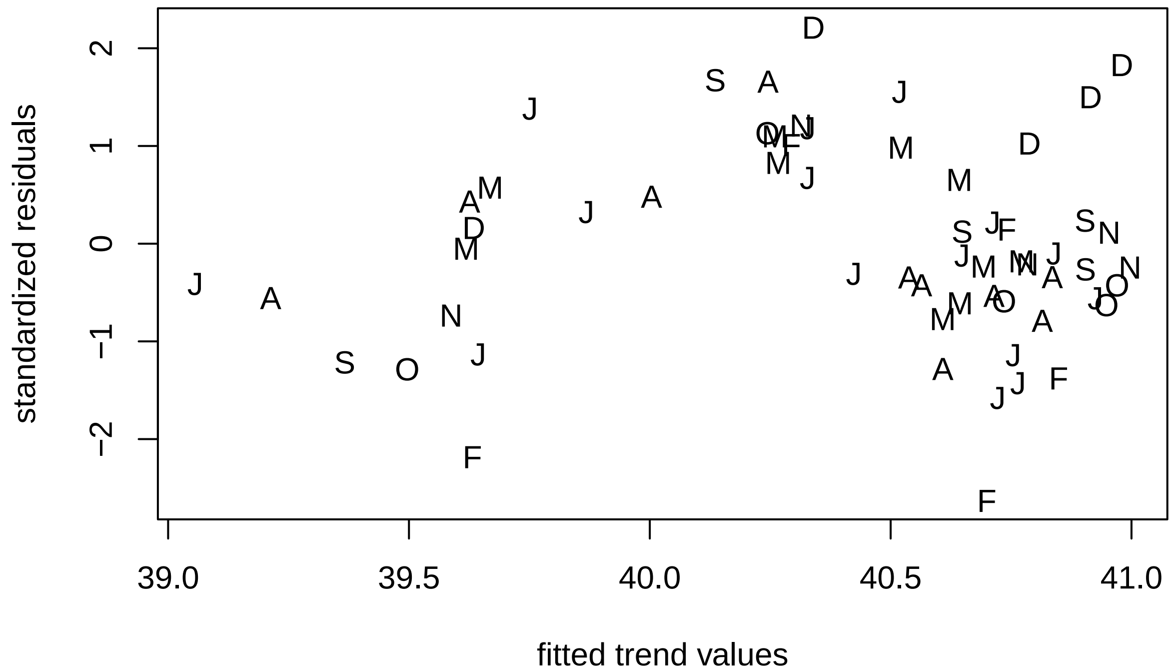
```
##
## Call:
## lm(formula = hours ~ har. + time(hours) + I(time(hours)^2))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.99955 -0.23753 -0.05985  0.28763  0.86045
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -5.061e+05  1.134e+05  -4.462 4.06e-05 ***
## har.cos(2*pi*t)   7.478e-02  7.586e-02   0.986   0.3286
## har.sin(2*pi*t)  -1.382e-01  7.665e-02  -1.804   0.0768 .
## time(hours)       5.097e+02  1.143e+02   4.460 4.09e-05 ***
## I(time(hours)^2) -1.283e-01  2.879e-02  -4.457 4.13e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4149 on 55 degrees of freedom
## Multiple R-squared:  0.6213, Adjusted R-squared:  0.5938
## F-statistic: 22.56 on 4 and 55 DF,  p-value: 4.573e-11
```

(e)

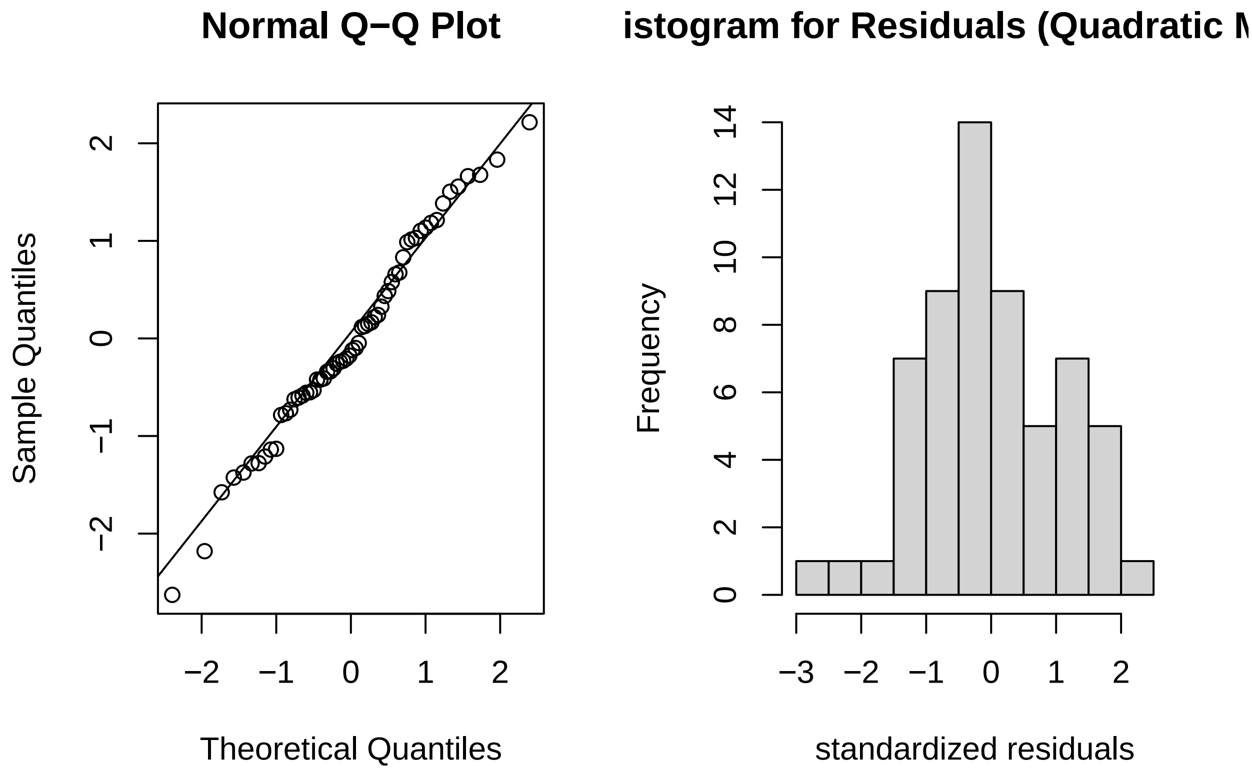
```r
# residual vs. time
plot(y=rstudent(hours.lm2), x=as.vector(time(hours)),
     type='l', xlab='time', ylab='standardized residuals')
points(y=rstudent(hours.lm2), x=as.vector(time(hours)),
       pch=as.vector(season(hours)))
```



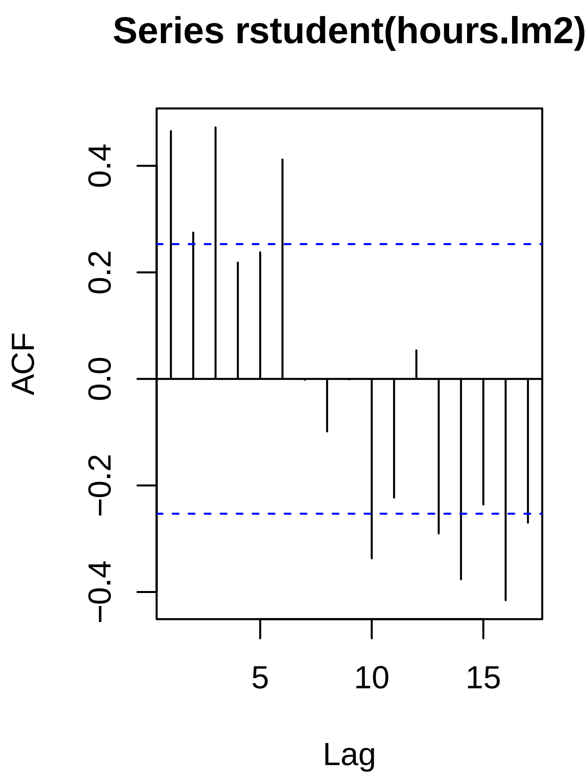

```r
# comments: there is no obvious pattern observed in the residual time series
# It seems that the combined fit significantly improved the quality of fit.


# residuals vs. estimated fit
plot(y=rstudent(hours.lm2),x=as.vector(fitted(hours.lm2)),
                xlab='fitted trend values',
                ylab='standardized residuals',
                type='n')
points(y=rstudent(hours.lm2),
                  x=as.vector(fitted(hours.lm2)),
                  pch=as.vector(season(hours)))
```




```r
# comments: It seems that there still exists pattern, since the residual decreases
# as the value of estimated fit increases.

# normality of residuals
par(mfrow=c(1,2))
qqnorm(rstudent(hours.lm2))
qqline(rstudent(hours.lm2))
hist(rstudent(hours.lm2), xlab='standardized residuals',
     main='Histogram for Residuals (Quadratic Model)')
```




```
# comments: The residuals follow normal distribution strictly.

# ACF plot
acf(rstudent(hours.lm2))
# comments: from the ACF plot, since many spikes go beyond the confidence
# band, so a strong correlation among residual values can be observed.
```




```r
# test for normality
shapiro.test(rstudent(hours.lm2))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  rstudent(hours.lm2)
## W = 0.98671, p-value = 0.7581
```

```r
# comments: since p-value = 0.7581 > 0.05,
# so we have no evidence against our null hypothesis: data come from a
# normal distribution.


# test for constant variance
seg <- as.vector(season(hours))
fligner.test(rstudent(hours.lm2), seg)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  rstudent(hours.lm2) and seg
## Fligner-Killeen:med chi-squared = 6.6642, df = 11, p-value = 0.8256
```

```r
# comments: since p-value = 0.8256 > 0.05,
# so we have no evidence against our null hypothesis: data have
# constant variance.

# test for randomness
library("randtests")
library("lawstat")
```

```r
difference.sign.test(rstudent(hours.lm2))
```

```
##
##  Difference Sign Test
##
## data:  rstudent(hours.lm2)
## statistic = 2.4394, n = 60, p-value = 0.01471
## alternative hypothesis: nonrandomness
```

```r
# comments: since p-value = 0.01471 < 0.05,
# so we have evidence against our null hypothesis: the data is random.
runs.test(rstudent(hours.lm2))
```

```
##
##  Runs Test - Two sided
##
## data:  rstudent(hours.lm2)
## Standardized Runs Statistic = -2.8646, p-value = 0.004176
```

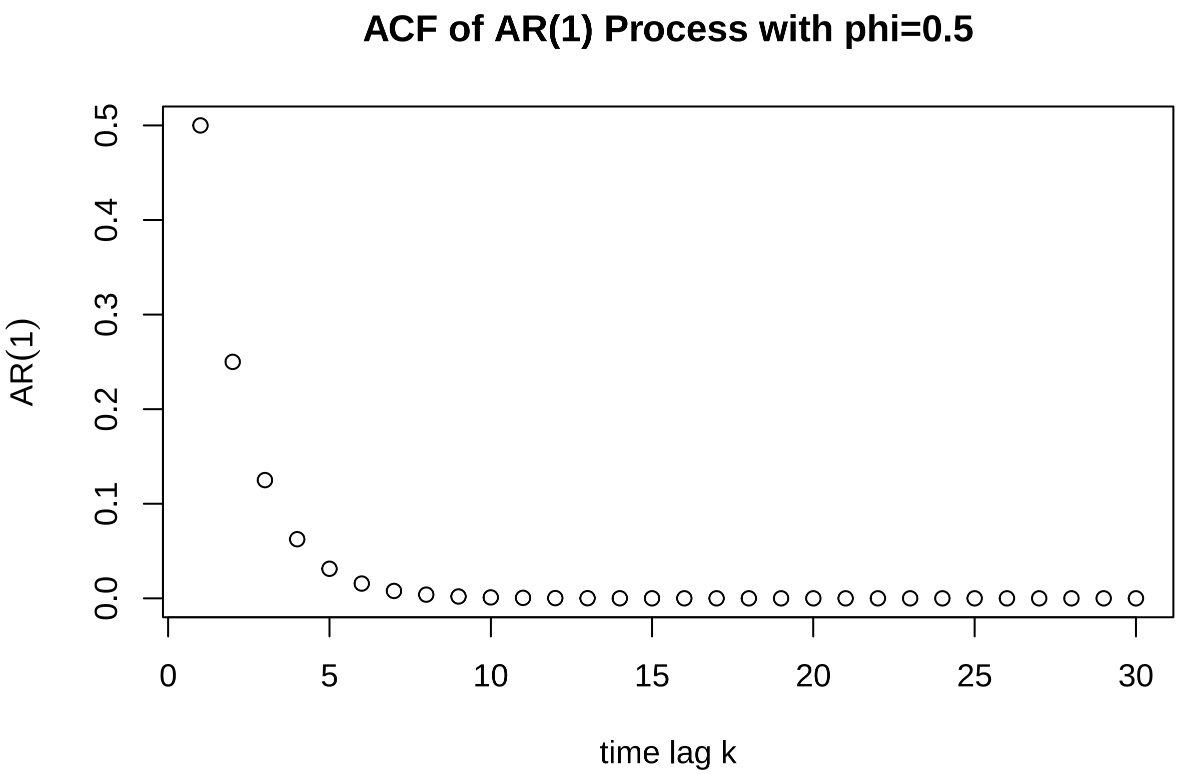
```r
# comments: same as difference sign test, since p-value = 0.004176 < 0.05
# so we have strong evidence against our null hypothesis: the data is random.
```

(f)

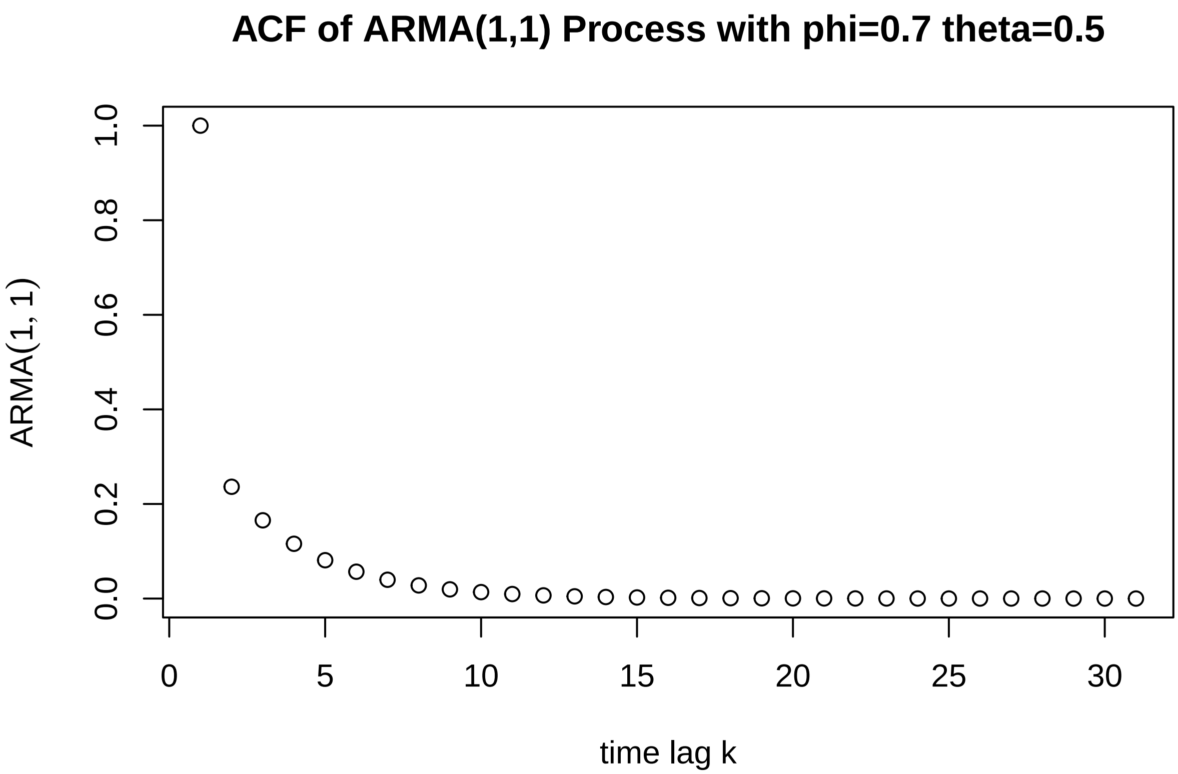
The combined model is relative appropriate for the trend estimation. Because from residual vs. time, we observed no obvious pattern anymore. From QQ-plot and histogram, we found that the residuals are normally distributed. From residual vs. estimated value, there is no obvious change in the variability at different estimated levels. But for ACF plot, a correlation still exists among residual values. I think that the combined model significantly improved the quality of fit.

Problem 2

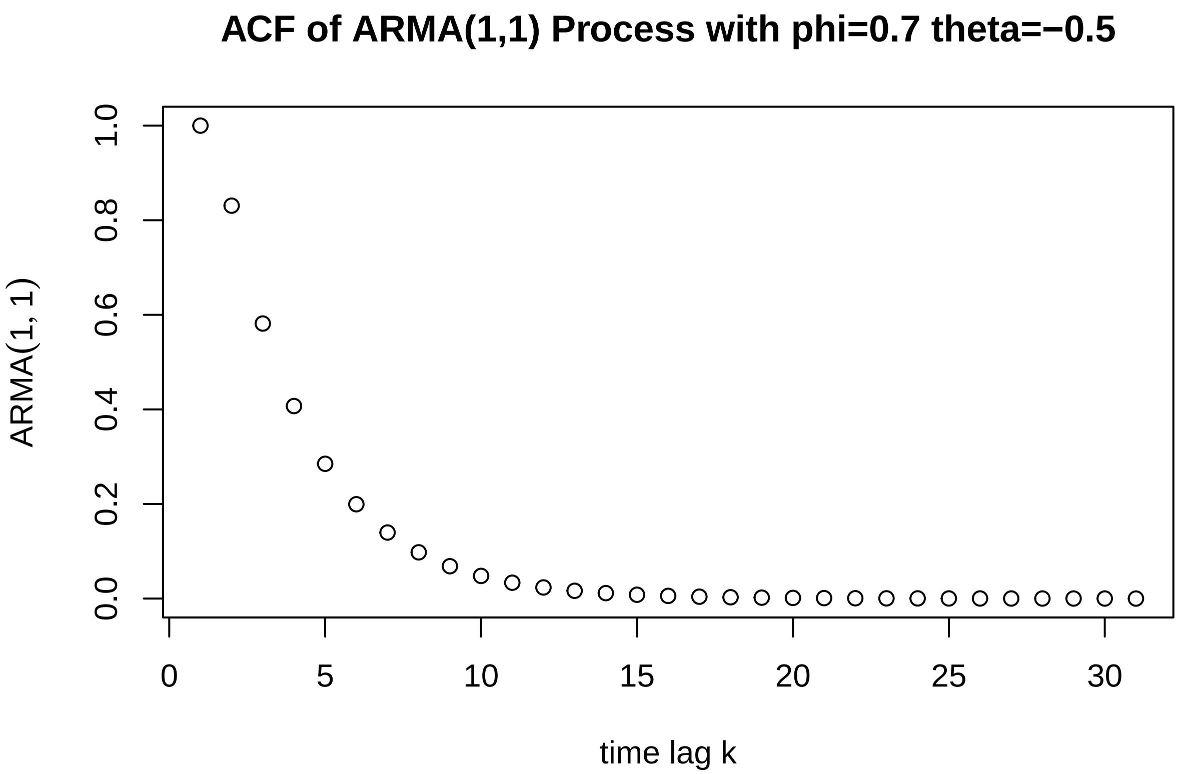
```
### (a)
phi <- 0.5
# we know ACF of AR(1) is phi^k with k = 0,1,2,3...
k <- 1:30
ACF_AR1 <- phi^k
plot(ACF_AR1, xlab="time lag k", ylab=expression(AR(1)),
     main="ACF of AR(1) Process with phi=0.5")
```



```r
### (b)
phi <- 0.7
theta <- 0.5
# we know ACF of ARMA(1,1) is when k>= 1:
# (phi-theta)*(1-theta*phi)/(1-2*theta*phi+theta^2)*phi^(k-1)
k <- 1:30
ACF_ARMA11 <- (phi-theta)*(1-theta*phi)/(1-2*theta*phi+theta^2)*phi^(k-1)
plot(c(1,ACF_ARMA11), xlab = "time lag k", ylab=expression(ARMA(1,1)),
     main="ACF of ARMA(1,1) Process with phi=0.7 theta=0.5")
```

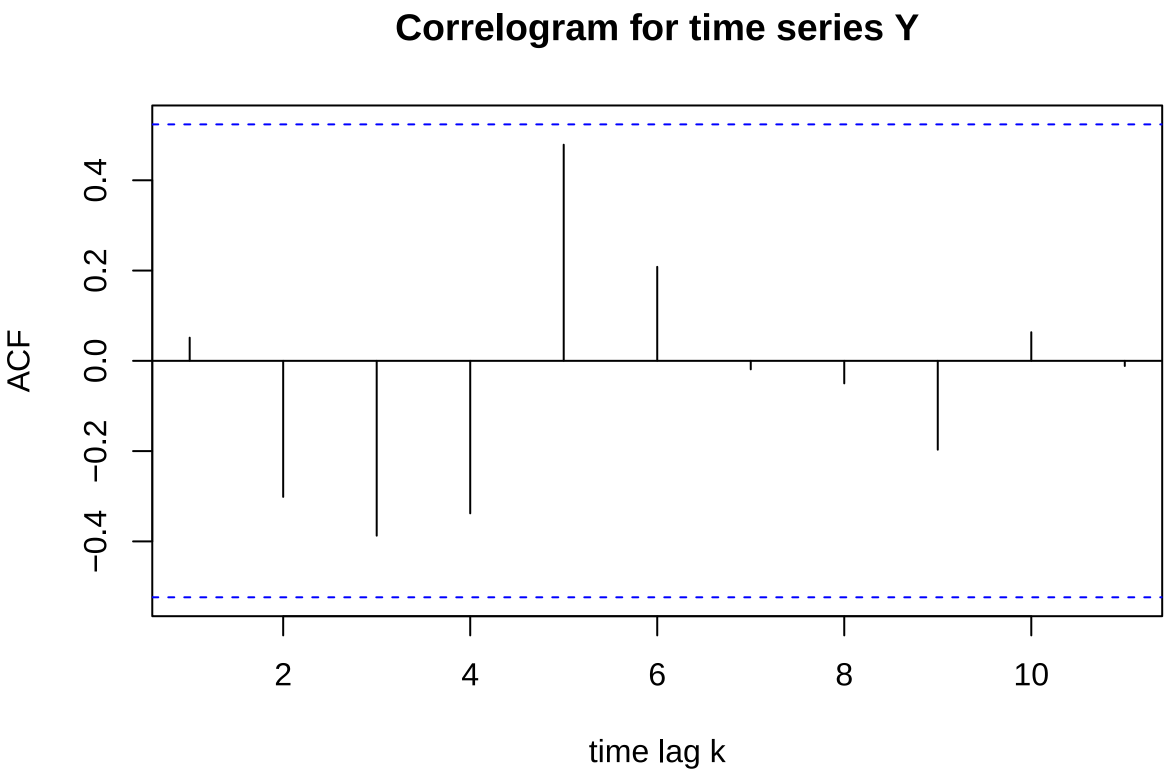


```r
### (c)
phi <- 0.7
theta <- -0.5
# we know ACF of ARMA(1,1) is when k>= 1:
# (phi-theta)*(1-theta*phi)/(1-2*theta*phi+theta^2)*phi^(k-1)
k <- 1:30
ACF_ARMA11 <- (phi-theta)*(1-theta*phi)/(1-2*theta*phi+theta^2)*phi^(k-1)
plot(c(1,ACF_ARMA11), xlab = "time lag k", ylab=expression(ARMA(1,1)),
     main="ACF of ARMA(1,1) Process with phi=0.7 theta=-0.5")
```



Problem 4

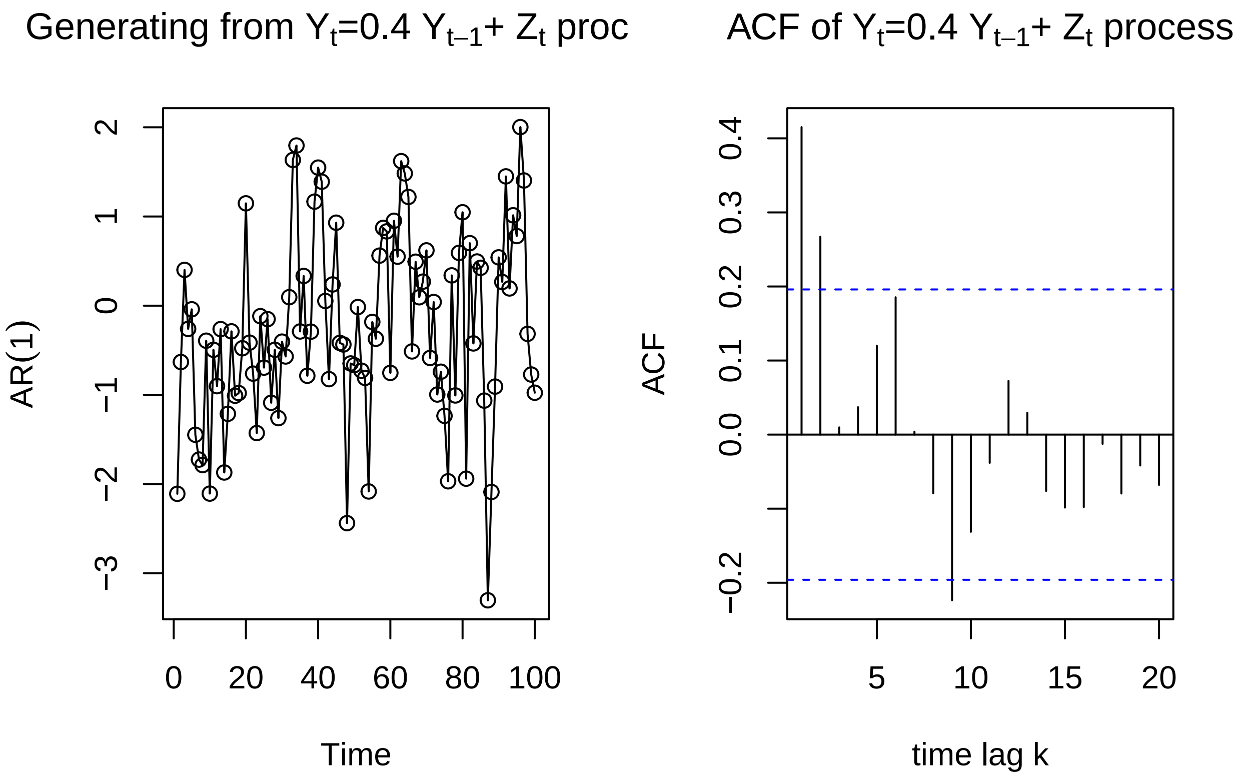
```r
## (c)
Y = c(18.25, 16.06, 7.81, 15.26, 16.61, 20.21, 22.03, 9.81, 12.58, 15.54,
      16.63, 21.20, 14.43, 17.71)

# plot correlogram in R
acf(Y, xlab='time lag k', main='Correlogram for time series Y')
```




```r
# comments: from the ACF, we observed that there is no spike go beyond
# confidence band, so no strong correlation is observed.
```

Problem 5

```r
## (c)
par(mfrow=c(1,2))
AR1 <- arima.sim(model = list(ar=0.4), n=100)
plot(AR1, type='o', ylab=expression(AR(1)),
     main = expression(paste("Generating from Y"[t], "=0.4 Y"[t-1], "+ Z"[t],
                             " process")))
acf(AR1, xlab = "time lag k",
    main=expression(paste("ACF of Y"[t], "=0.4 Y"[t-1], "+ Z"[t],
                            " process")))
```




```r
# comments on ACF: from the ACF, we observe that the magnitude of the ACF
# decays exponentially as the time lag k increases.
```