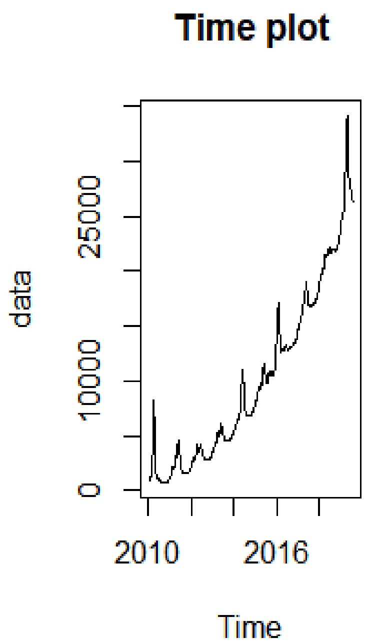
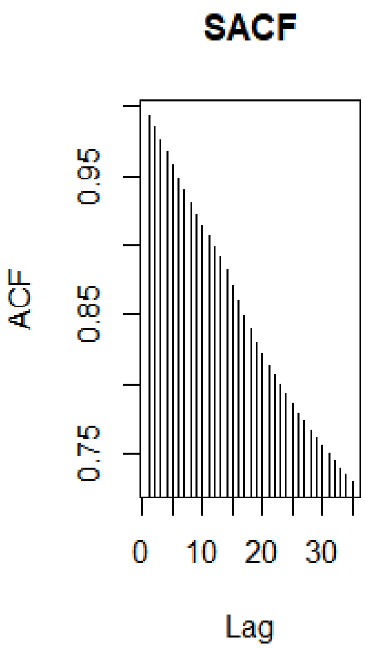
Data Analysis Exam 1

2017311974 진수정

(a) Time plot, correlograms (ACF) and discuss key features of the data.





1. (increasing Trend): Time plot shows that there exists (linear or quadratic) increasing trend. Also, slowly decaying and linearly decaying SACFs indicate the existence of trend.

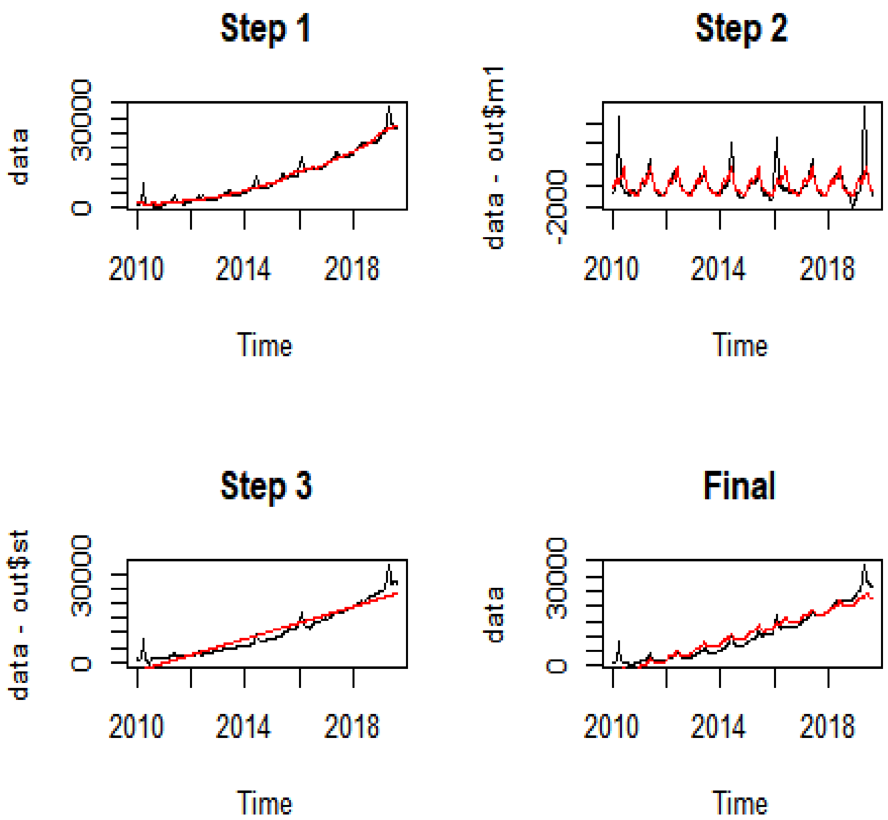
2. (Seasonality with period 52): From the fact that the given data is weekly data, it can be easily inferred that there would be seasonality with period 52 (since there are 52 weeks in a year). Some repeated pattern in time plot shows that there exists seasonality.

3. (Outliers): There are some outliers nearby t = 2010, 2014, 2016, 2019.

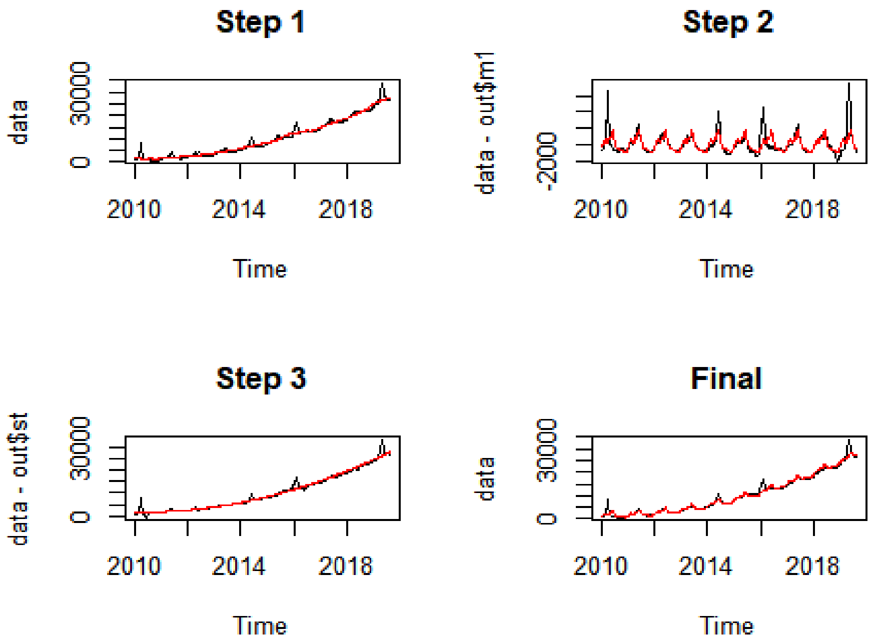
(b) Remove any trend or seasonality or both to make the series as stationary if necessary.

1) Smoothing based classical decomposition

To remove both trend and seasonality, we will try classical decomposition. Since there are linear or quadratic trend and lag 52 seasonality, assign d = 52 and order = 1 to the 'classical' function first.



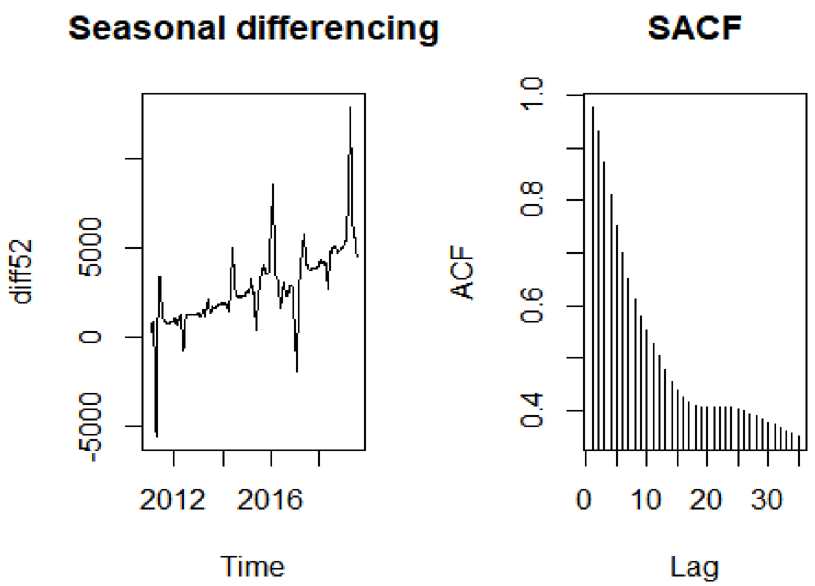
At step 3, linear regression model cannot fully explain the data. Thus, proceed with order = 2.



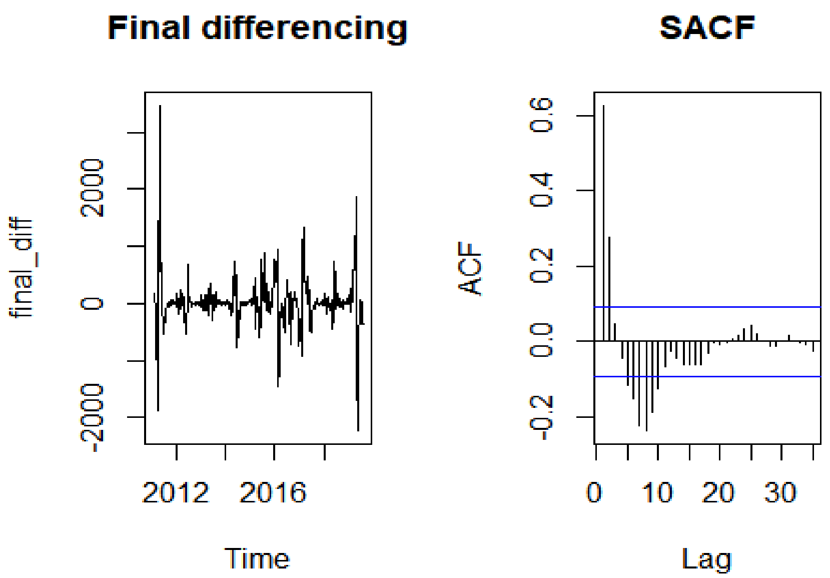
The above plots show that estimated line fits the data well. Thus, classical decomposition with d = 52, order = 2 works fine.

2) Differencing method

Next, we will try differencing method to detrend and deseasonalize. First, apply seasonal differencing with lag = 52.



There is linear trend left. Also, slowly decaying and almost linearly decaying SACFs indicate that there remains some trend. To remove trend, apply additional 1st differencing.



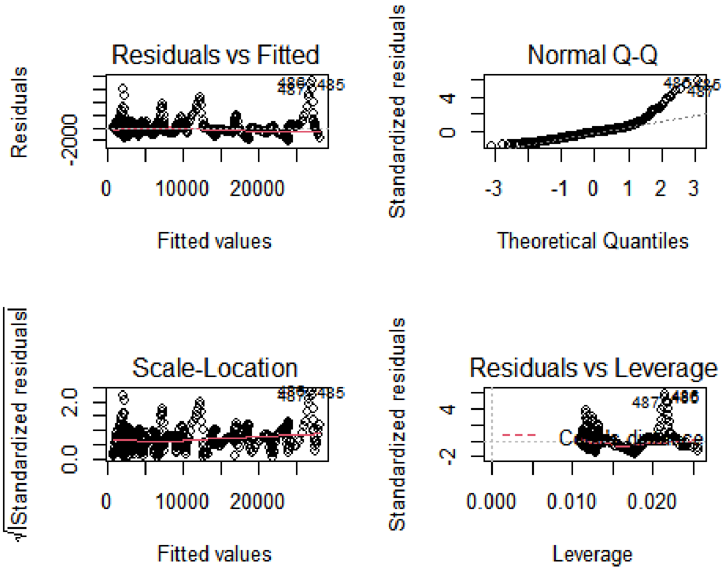
Since there is no clear trend left, both trend and seasonality are successfully removed. Thus, differencing method works fine.

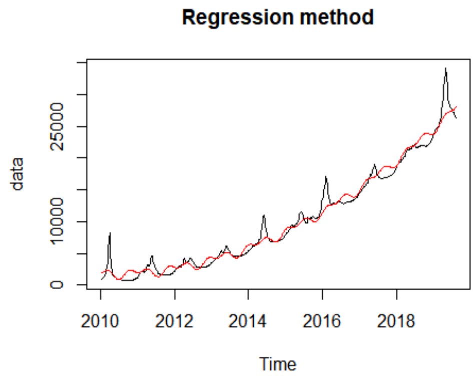
3) Regression method

Next, we will try regression method. To remove both trend and seasonality, apply polynomial regression for trend and harmonic regression for seasonal component simultaneously. Before modeling, we need to determine k in harmonic regression.

```
##
## Call:
## lm(formula = data ~ costerm1 + sinterm2 + costerm2 + x + I(x^2))
##
## Coefficients:
## (Intercept)     costerm1     sinterm2     costerm2            x       I(x^2)
##   1.233e+09    4.467e+02    2.032e+02   -4.033e+02   -1.227e+06    3.051e+02
```

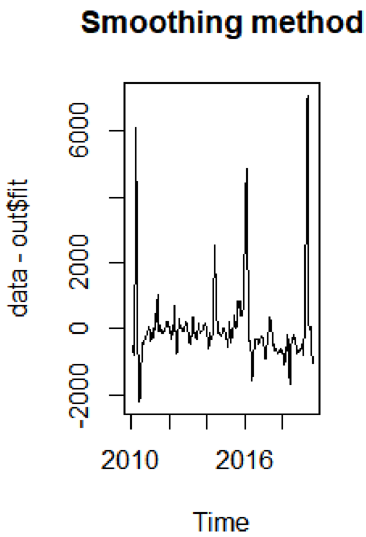
Based on the above stepwise selection result, we will use harmonic regression with k = 2, and polynomial regression with order = 2 simultaneously.

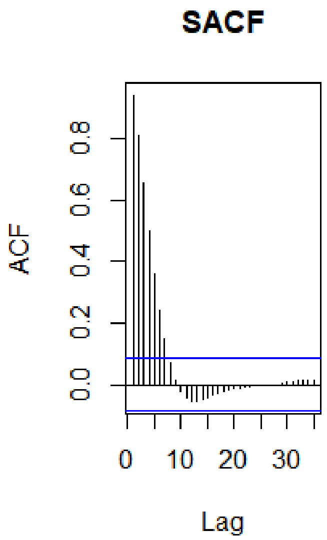


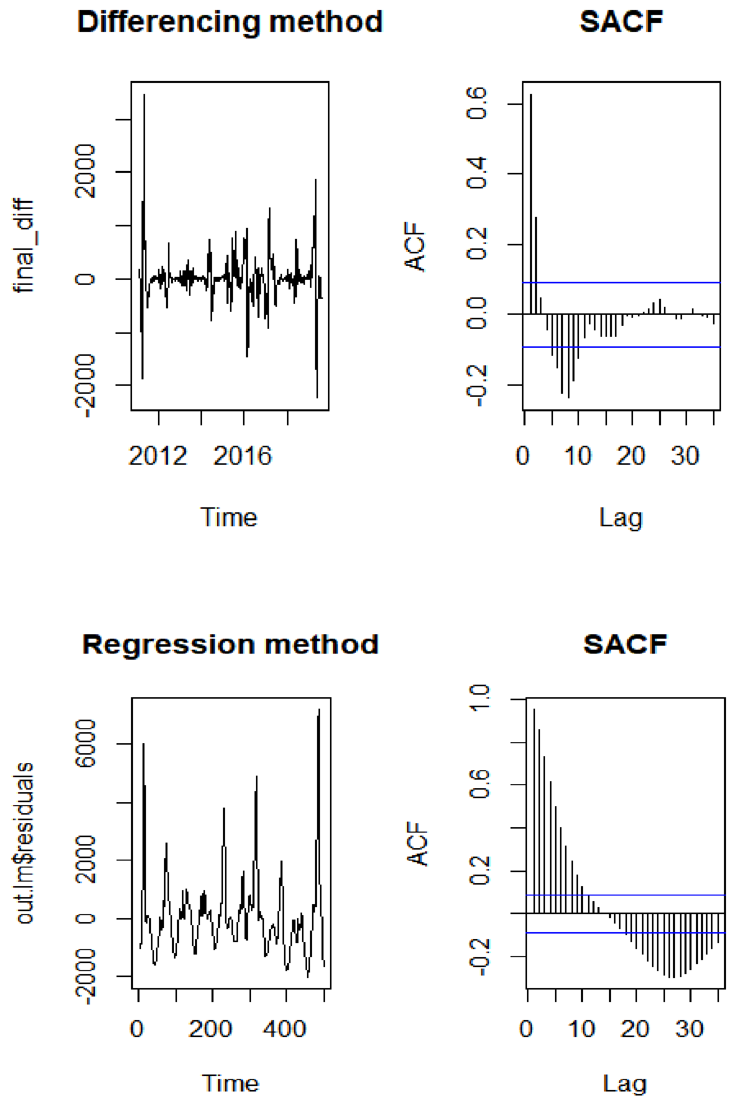


The above plots are results from the regression model. Residuals vs Fitted plot looks fine, and estimated line looks not bad.

Then, we need to select the final model among the models obtained by applying classical decomposition (smoothing), differencing, and regression method.



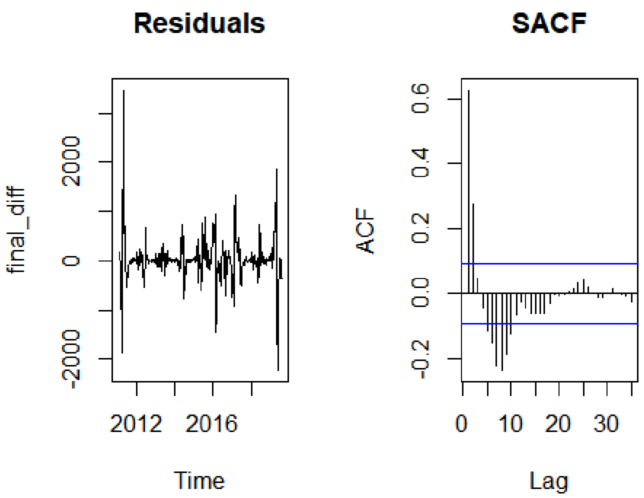




Since some clear pattern still remains in the residual plot of regression method, regression method might not be appropriate.

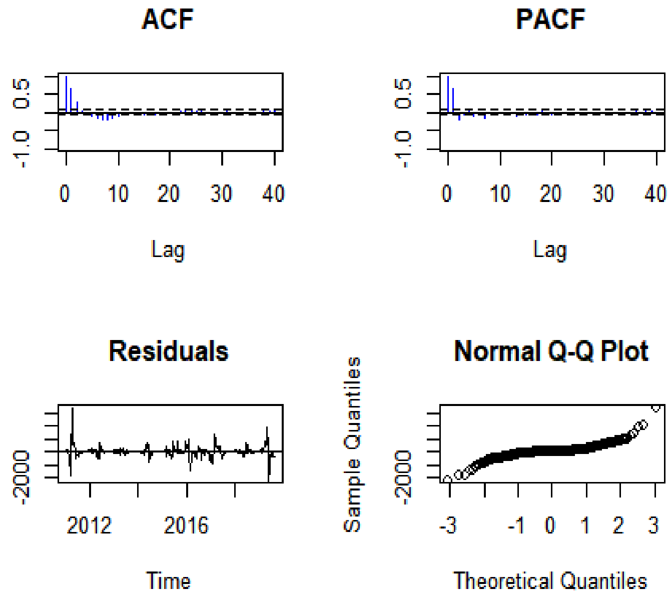
Among the residuals obtained by applying the other two methods, the residuals of differencing method seem to be more stable toward mean. Also, for classical decomposition, it is needed to follow multistage algorithm to estimate trend and seasonality. However, just simple seasonal/order differencing is enough for differencing method. Thus, for its stability and handy calculation, the model obtained by differencing method is selected as the final model.

(c) Include reasoning why the residuals from the selected model in (b) is stationary. Also, can you claim that the removed series is an IID sequence?



There is no clear pattern left in the residual plot, and slowly decaying SACFs in the original data is disappeared. Thus, the residuals from the selected model are stationary.

```
## Null hypothesis: Residuals are iid noise.
## Test                        Distribution Statistic   p-value
## Ljung-Box Q                Q ~ chisq(20)    311.27         0 *
## McLeod-Li Q                Q ~ chisq(20)     203.4         0 *
## Turning points T  (T-296.7)/8.9 ~ N(0,1)       239         0 *
## Diff signs S        (S-223)/6.1 ~ N(0,1)       216    0.2519
## Rank P       (P-49840.5)/1577.7 ~ N(0,1)     51003    0.4612
```



Also, the removed series is not iid. The reason is as follows.

1. All the first three tests (Ljung-Box Q, McLeod-Li Q, Turning points T) are rejected, so residuals are not iid.

2. Based on SACFs, there is strong positive correlation on small lags. Thus, residuals are correlated.

(d) Write one paragraph summary on your findings about (a)-(c).

First, there exists (linear or quadratic) increasing trend and period-52 seasonality in the given data, based on time plot and correlogram. To remove both trend and seasonality, classical decomposition, differencing method, and regression method were used. Among them, the model obtained by applying differencing method was selected as the final model. This is because differencing method has handy calculation, and the obtained residuals are more stable toward mean than the others. The residuals from the final model are stationary since it shows no clear pattern. However, based on tests of randomness and SACFs, the residuals are not iid. Thus, it is needed to model the residual structure in later steps.

(e) Attach R (or other softwares you used) code you have used in this analysis.

```
setwd("C:/Users/SJ/OneDrive/바탕 화면/시계열/시험")
rm(list = ls())
source("TS-library.R")
library(aTSA)

##
## Attaching package: 'aTSA'

## The following object is masked from 'package:itsmr':
##
##     forecast

## The following object is masked from 'package:graphics':
##
##     identify
```

```r
data = scan("2021exam1.txt")
data = ts(data,start = c(2010,2),freq = 52)
```

```r
# Time plot & Correlogram
par(mfrow = c(1,2))
plot.ts(data)
title("Time plot")
acf2(data)
title("SACF")
```

```r
n = length(data)
t = 1:n
x = as.vector(time(data))
```

```r
# Classical Decomposition (d = 52, order = 1)
out = classical(data,d = 52,order = 1)

par(mfrow = c(2,2))
plot.ts(data)
title("Step 1")
lines(x,out$m1,col = 'red')

plot.ts(data - out$m1)
title("Step 2")
lines(x,out$st,col = 'red')

plot.ts(data - out$st)
title("Step 3")
lines(x,out$m,col = 'red')

plot.ts(data)
title("Final")
lines(x,out$fit,col = 'red')
```

```r
# Classical Decomposition (d = 52, order = 2)
out = classical(data,d = 52,order = 2)

par(mfrow = c(2,2))
plot.ts(data)
title("Step 1")
lines(x,out$m1,col = 'red')

plot.ts(data - out$m1)
title("Step 2")
lines(x,out$st,col = 'red')
```

```r
plot.ts(data - out$st)
title("Step 3")
lines(x,out$m,col = 'red')

plot.ts(data)
title("Final")
lines(x,out$fit,col = 'red')
```

```r
# Seasonal differencing
diff52 = diff(data,lag = 52)

par(mfrow = c(1,2))
plot.ts(diff52)
title("Seasonal differencing")
acf2(diff52)
title("SACF")
```

```r
# 1st differencing
final_diff = diff(diff52,1)

par(mfrow = c(1,2))
plot.ts(final_diff)
title("Final differencing")
acf2(final_diff)
title("SACF")
```

```r
# Regression
m1 = floor(n/52)
m2 = 2*m1
m3 = 3*m1
m4 = 4*m1

sinterm1 = sin(m1*2*pi/n*t)
costerm1 = cos(m1*2*pi/n*t)
sinterm2 = sin(m2*2*pi/n*t)
costerm2 = cos(m2*2*pi/n*t)
sinterm3 = sin(m3*2*pi/n*t)
costerm3 = cos(m3*2*pi/n*t)
sinterm4 = sin(m4*2*pi/n*t)
costerm4 = cos(m4*2*pi/n*t)

step(lm(data ~ 1 + sinterm1 + costerm1 + sinterm2 + costerm2 +
         sinterm3 + costerm3 + sinterm4 + costerm4 +
         x + I(x^2)))
```

```
## Start:  AIC=7153.1
## data ~ 1 + sinterm1 + costerm1 + sinterm2 + costerm2 + sinterm3 +
##     costerm3 + sinterm4 + costerm4 + x + I(x^2)
##
##            Df  Sum of Sq        RSS    AIC
## - costerm3  1      92297  781658454 7151.2
## - sinterm3  1    1293527  782859684 7151.9
## - costerm4  1    1780636  783346793 7152.2
## - sinterm1  1    2381410  783947567 7152.6
## - sinterm4  1    2827685  784393842 7152.9
## <none>                    781566157 7153.1
## - sinterm2  1   10266723  791832880 7157.6
## - costerm2  1   40661274  822227431 7176.5
## - costerm1  1   49892783  831458940 7182.0
## - x         1 2200400580 2981966736 7820.6
## - I(x^2)    1 2210172925 2991739081 7822.3
##
## Step:  AIC=7151.16
## data ~ sinterm1 + costerm1 + sinterm2 + costerm2 + sinterm3 +
##     sinterm4 + costerm4 + x + I(x^2)
##
##            Df  Sum of Sq        RSS    AIC
## - sinterm3  1    1293625  782952079 7150.0
## - costerm4  1    1780616  783439069 7150.3
## - sinterm1  1    2381011  784039464 7150.7
## - sinterm4  1    2827578  784486031 7151.0
## <none>                    781658454 7151.2
## - sinterm2  1   10267138  791925592 7155.7
## - costerm2  1   40661383  822319837 7174.5
## - costerm1  1   49892613  831551067 7180.1
## - x         1 2200441021 2982099475 7818.6
## - I(x^2)    1 2210213569 2991872022 7820.3
##
## Step:  AIC=7149.99
## data ~ sinterm1 + costerm1 + sinterm2 + costerm2 + sinterm4 +
##     costerm4 + x + I(x^2)
##
##            Df  Sum of Sq        RSS    AIC
## - costerm4  1    1781045  784733123 7149.1
## - sinterm1  1    2389779  785341857 7149.5
## - sinterm4  1    2829932  785782010 7149.8
## <none>                    782952079 7150.0
## - sinterm2  1   10258079  793210158 7154.5
## - costerm2  1   40659340  823611418 7173.3
## - costerm1  1   49894846  832846925 7178.9
## - x         1 2200466698 2983418777 7816.9
## - I(x^2)    1 2210237381 2993189459 7818.5
##
## Step:  AIC=7149.12
## data ~ sinterm1 + costerm1 + sinterm2 + costerm2 + sinterm4 +
##     x + I(x^2)
##
##            Df  Sum of Sq        RSS    AIC
## - sinterm1  1    2388022  787121145 7148.6
```

```
## - sinterm4  1    2829460  787562583 7148.9
## <none>                   784733123 7149.1
## - sinterm2  1   10259901  794993024 7153.6
## - costerm2  1   40659788  825392912 7172.4
## - costerm1  1   49894230  834627353 7177.9
## - x        1 2200557156 2985290279 7815.2
## - I(x^2)    1 2210328434 2995061557 7816.8
##
## Step:  AIC=7148.64
## data ~ costerm1 + sinterm2 + costerm2 + sinterm4 + x + I(x^2)
##
##            Df  Sum of Sq        RSS    AIC
## - sinterm4  1    2819843  789940988 7148.4
## <none>                   787121145 7148.6
## - sinterm2  1   10297226  797418371 7153.1
## - costerm2  1   40668164  827789309 7171.8
## - costerm1  1   49885100  837006245 7177.4
## - x        1 2200453652 2987574797 7813.6
## - I(x^2)    1 2210232332 2997353477 7815.2
##
## Step:  AIC=7148.43
## data ~ costerm1 + sinterm2 + costerm2 + x + I(x^2)
##
##            Df  Sum of Sq        RSS    AIC
## <none>                   789940988 7148.4
## - sinterm2  1   10307145  800248132 7152.9
## - costerm2  1   40670392  830611380 7171.5
## - costerm1  1   49882667  839823655 7177.0
## - x        1 2200425859 2990366847 7812.0
## - I(x^2)    1 2210206543 3000147531 7813.7

##
## Call:
## lm(formula = data ~ costerm1 + sinterm2 + costerm2 + x + I(x^2))
##
## Coefficients:
## (Intercept)     costerm1     sinterm2     costerm2            x       I(x^2)
##   1.233e+09    4.467e+02    2.032e+02   -4.033e+02   -1.227e+06    3.051e+02
```

```r
out.lm = lm(data ~ 1 + x + I(x^2) + sinterm1 + costerm1 +
             sinterm2 + costerm2)
summary(out.lm)
```

```
##
## Call:
## lm(formula = data ~ 1 + x + I(x^2) + sinterm1 + costerm1 + sinterm2 +
##     costerm2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2043.5  -765.8  -102.9   377.4  7231.7
##
```

```
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.233e+09  3.330e+07  37.033  < 2e-16 ***
## x           -1.227e+06  3.306e+04 -37.115  < 2e-16 ***
## I(x^2)       3.051e+02  8.203e+00  37.197  < 2e-16 ***
## sinterm1    -9.791e+01  8.024e+01  -1.220   0.2230
## costerm1     4.468e+02  7.994e+01   5.589 3.80e-08 ***
## sinterm2     2.029e+02  8.001e+01   2.536   0.0115 *
## costerm2    -4.033e+02  7.994e+01  -5.045 6.38e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1264 on 493 degrees of freedom
## Multiple R-squared:  0.9751, Adjusted R-squared:  0.9748
## F-statistic:  3218 on 6 and 493 DF,  p-value: < 2.2e-16

par(mfrow = c(2,2))
plot(out.lm)
```

```
plot.ts(data)
title("Regression method")
lines(x,out.lm$fitted,col = 'red')
```

```
par(mfrow = c(1,2))
plot.ts(data - out$fit)
title("Smoothing method")

acf2(data - out$fit)
title("SACF")
```

```
par(mfrow = c(1,2))
plot.ts(final_diff)
title("Differencing method")

acf2(final_diff)
title("SACF")
```

```
par(mfrow = c(1,2))
plot.ts(out.lm$residuals)
title("Regression method")

acf2(out.lm$residuals)
title("SACF")
```

```r
# Stationarity
par(mfrow = c(1,2))
plot.ts(final_diff)
title("Residuals")

acf2(final_diff)
title("SACF")
```

```r
# whether IID sequence
test(final_diff)

## Null hypothesis: Residuals are iid noise.
## Test                        Distribution Statistic   p-value
## Ljung-Box Q                Q ~ chisq(20)    311.27         0 *
## McLeod-Li Q                Q ~ chisq(20)     203.4         0 *
## Turning points T  (T-296.7)/8.9 ~ N(0,1)       239         0 *
## Diff signs S        (S-223)/6.1 ~ N(0,1)       216    0.2519
## Rank P      (P-49840.5)/1577.7 ~ N(0,1)     51003    0.4612
```