STAT 153 Homework 2

Donggyun Kim

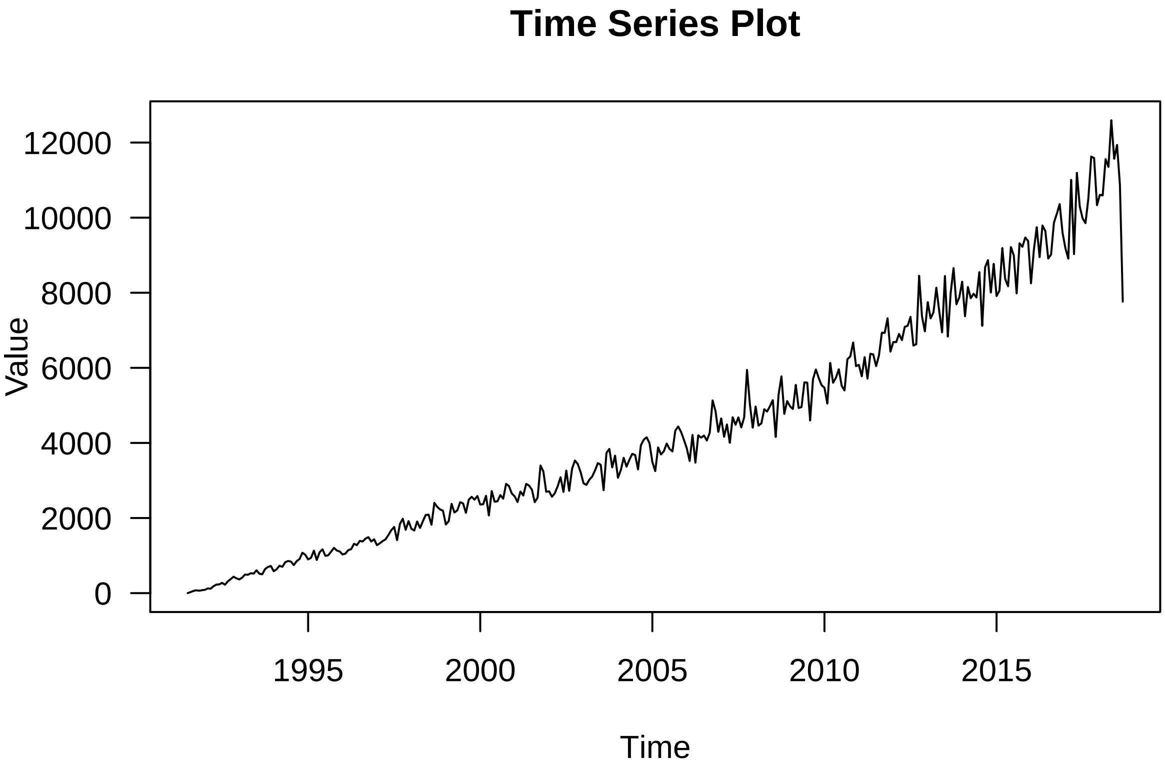
12/16/2018

Computer exercise:

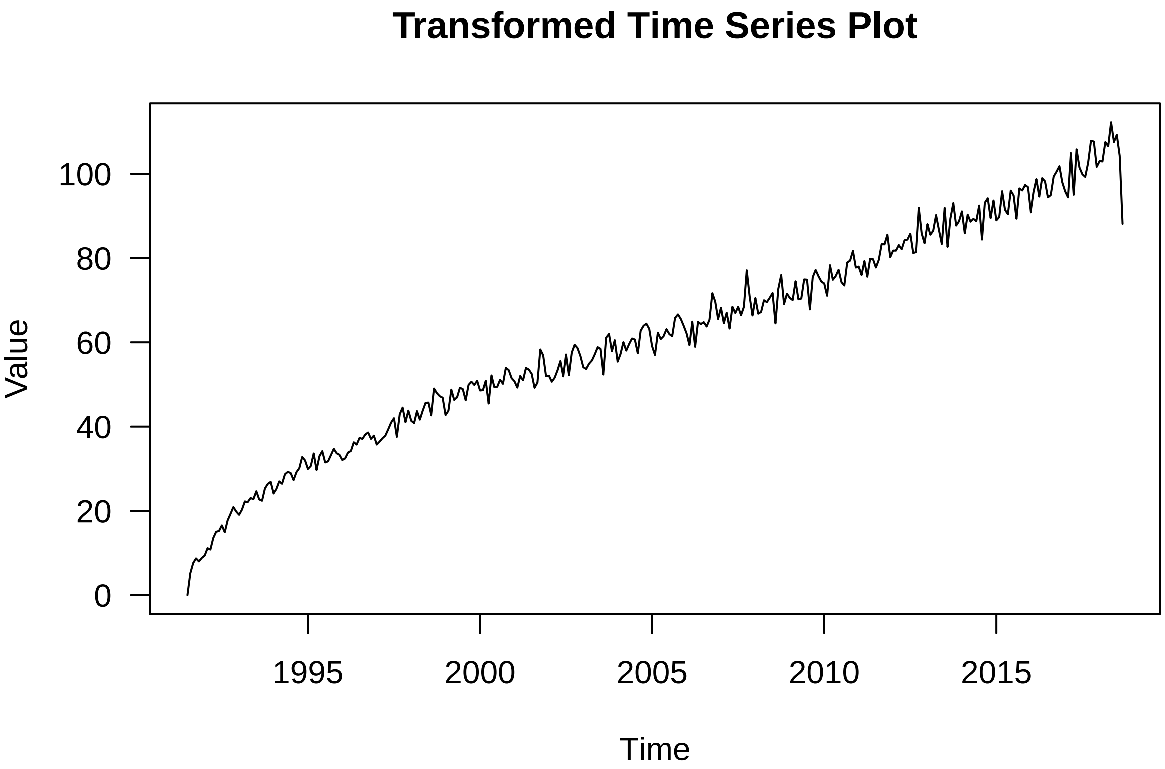
1. Go to https://arxiv.org/stats/monthly_submissions and download the data for number of new arXiv sumbmissions received during each month since August 1991.

(a) Argue whether or not you would expect such data to have a constant variance over time. How can you transform the data such that the variance is approximately constant over time (homogeneous)?

```r
# import data
dat <- read.csv("arXiv.csv", stringsAsFactors = FALSE)

# data cleaning
my_ts <- ts(dat[, 2], start = c(1991, 7), end = c(2018, 9), frequency = 12)

# series plot
plot(my_ts, main = "Time Series Plot", ylab = "Value", las = 1)
```



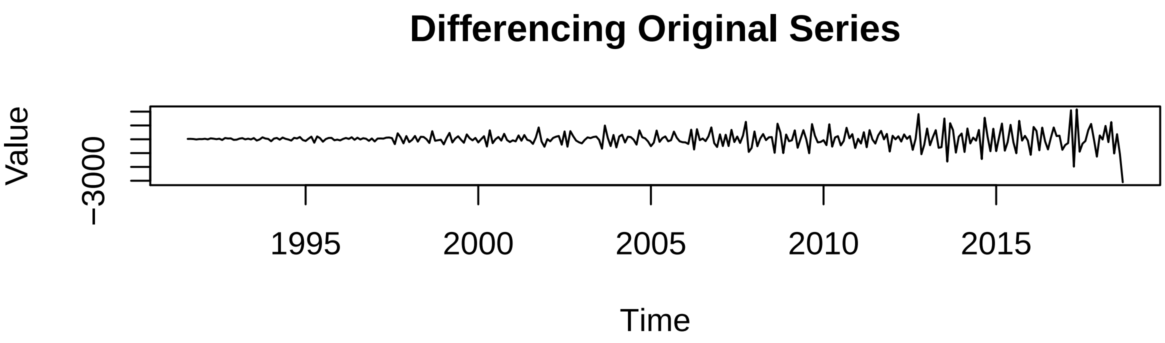
```r
# transformation
trans_ts <- ts(c(0, my_ts[-1] / sqrt(my_ts[-1])), start = c(1991, 7),
               end = c(2018, 9), frequency = 12)

# transformed series plot
plot(trans_ts, main = "Transformed Time Series Plot", ylab = "Value", las = 1)
```

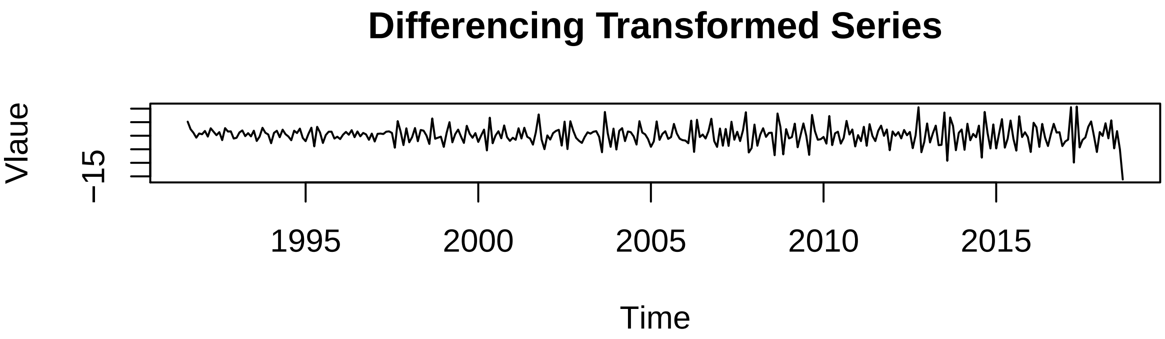


Since count data approximately follows Poisson distribution whose variance euqlas its mean that is linear function, dividing the series by its squared values for each time makes it has constant variance.

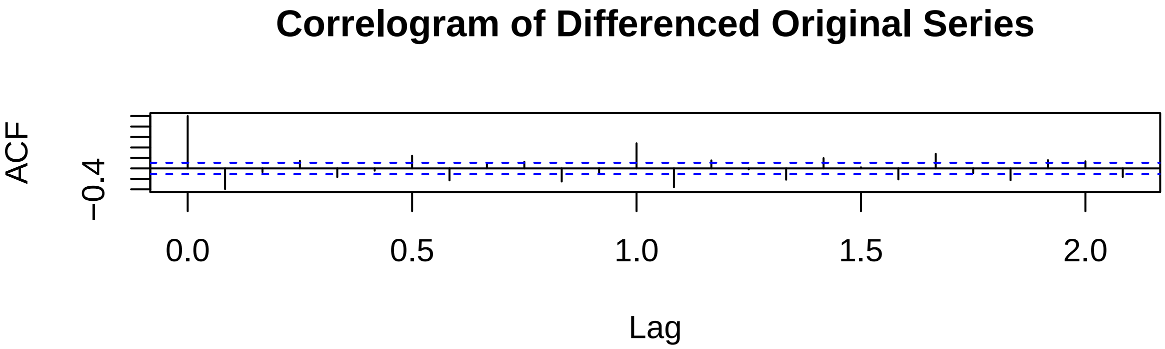
(b) Difference both, the original and the transformed data. Which one looks more like white noise?

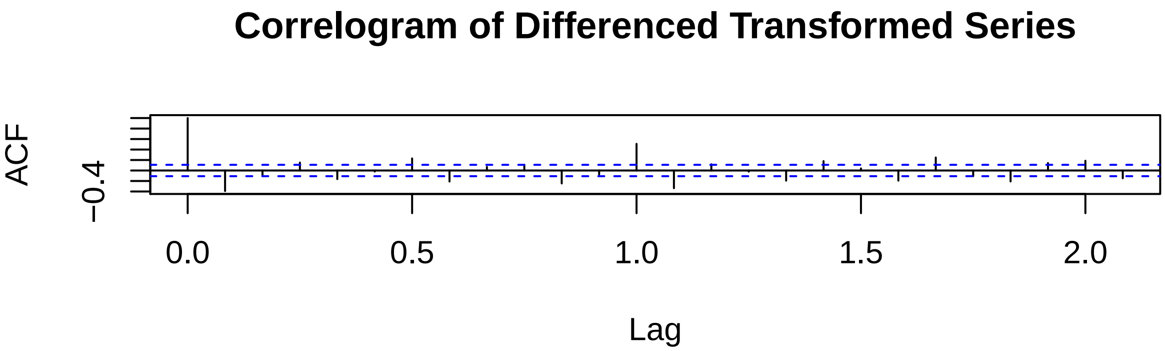
```
# differencing and plot
diff_my_ts <- diff(my_ts)
diff_trans_ts <- diff(trans_ts)
par(mfrow=c(2,1))
plot(diff_my_ts, main = "Differencing Original Series", ylab = "Value")
plot(diff_trans_ts, main = "Differencing Transformed Series", ylab = "Vlaue")
```





```r
acf(diff_my_ts, main = "Correlogram of Differenced Original Series")
acf(diff_trans_ts, main = "Correlogram of Differenced Transformed Series")
```





Time series plots show that differenced transformed series looks more like white noise. However, correlograms show there is no big difference between two of them.

(c) Based on the differenced data, provide a forecast for the number of new arXiv submissions for September (Remark: Ignore the data for the partial submissions in September).

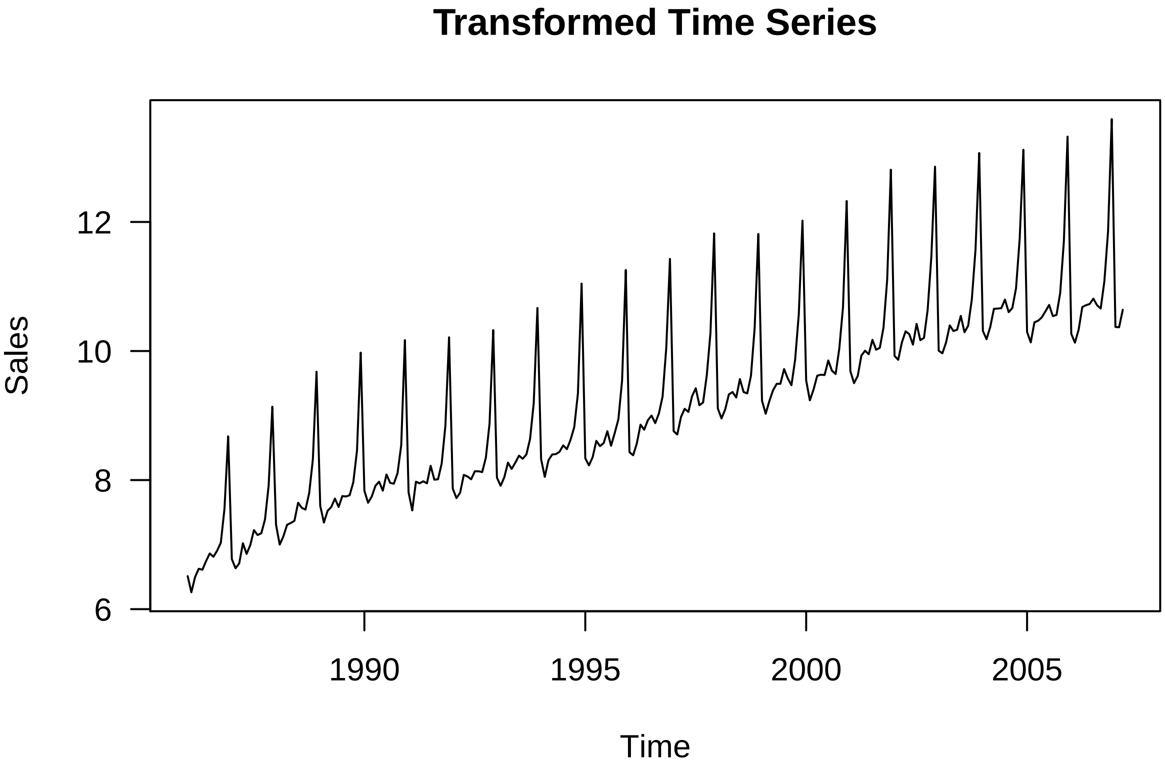
```r
last <- length(trans_ts)
diff_last <- length(diff_trans_ts)
(trans_ts[last - 1] + mean(diff_trans_ts[-diff_last]))^2
```

```
## [1] 10937
```

2. The data file retail in the R package TSA lists total U.K. (United Kingdom) retail sales (in billions of pounds) from January 1986 through March 2007. For the purpose of this exercise, take the square root of the observations and work with the transformed data. We will call the transformed data (y_t).

(a) Make a time series plot of (y_t). Is there any trend or seasonality?

```r
library(TSA)
data("retail")
yt <- sqrt(retail)
plot(yt, main = "Transformed Time Series", las = 1)
```



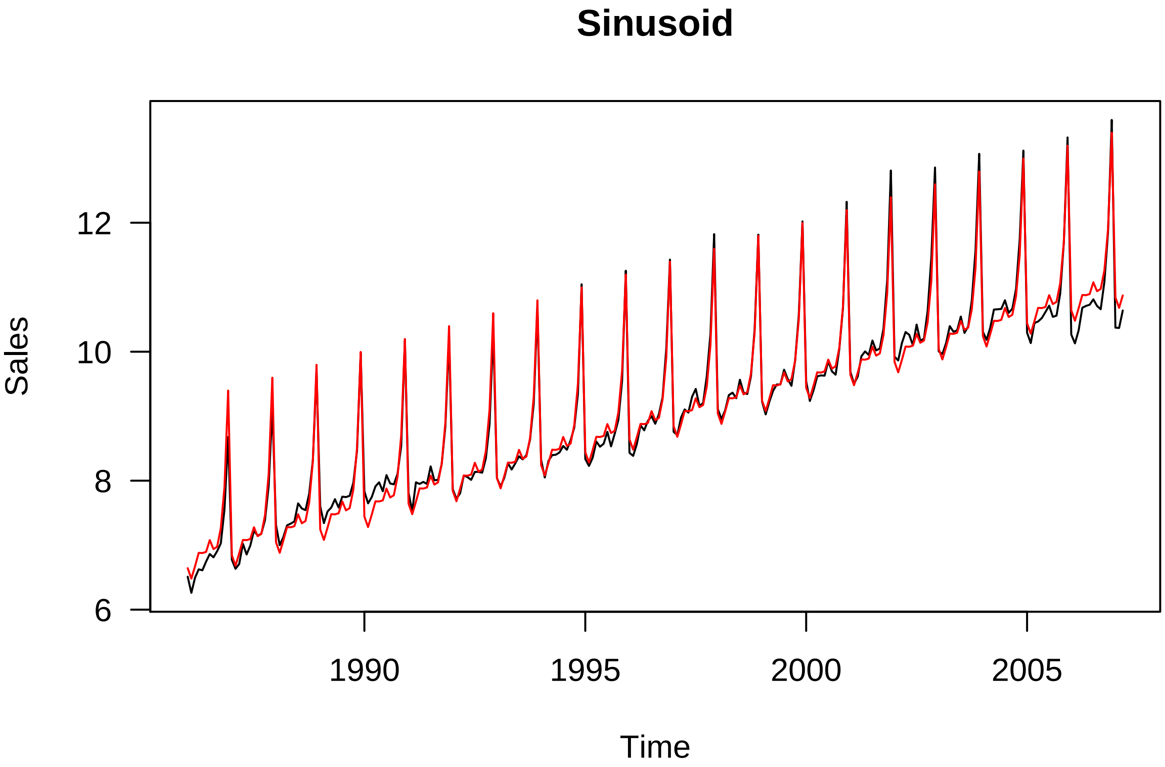
The plot shows there are both trend and seasonality in the series.

(b) Use least squares to fit the following model

$$y_t = \beta_0 + \beta_1 t + \sum_{j=1}^{6} \left[\beta_{2j} \cos(\frac{2\pi jt}{12}) + \beta_{2j+1} \sin(\frac{2\pi jt}{12}) \right] + w_t$$

where (w_t) is zero mean white noise. Note that $\sin(2\pi(6)/12) = 0$, so you have to only estimate $(\beta_0, \dots, \beta_{12})$. Plot (y_t) and the fitted values on the same graph.

```r
t <- 1:nrow(yt)
sinusoid <- matrix(0, nrow = nrow(yt), ncol = 12)
for (i in 1:6){
  sinusoid[, 2*i - 1] <- cos(2*pi*i*t/12)
  sinusoid[, 2*i] <- sin(2*pi*i*t/12)
}
colnames(sinusoid) <- paste0(rep(c("cos", "sin"), 6), rep(1:6, each = 2))
lm_sinusoid <- lm(yt ~ t + sinusoid[, -12])
fv_sinusoid <- ts(lm_sinusoid$fitted.values, start = c(1986, 1), end = c(2007, 3),
         frequency = 12)
plot(yt, main = "Sinusoid", las = 1)
lines(fv_sinusoid, col = "red")
```

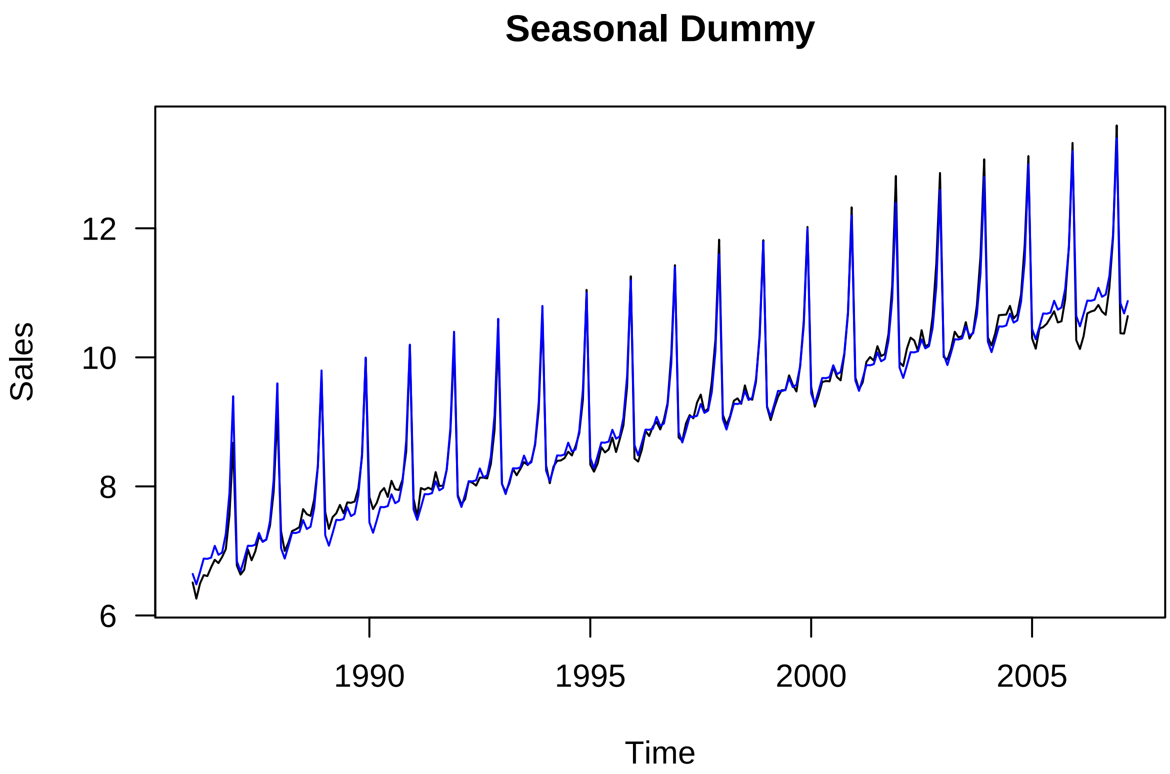


(c) Use least squares to fit the following model

$$y_t = \beta_0 + \beta_1 t + \beta_2 I(\text{t is January}) + ... + \beta_{12} I(\text{t is November}) + w_t$$

where (w_t) is zero mean white noise. The indicator variable $I(A)$ takes values 1 if A and 0 otherwise. Plot (y_t) and the fitted values on the same graph. (Hint: You may find the function seasonaldummy() in the forecast package useful.)

```
library(forecast)
dummies <- seasonaldummy(yt)
lm_dummy <- lm(yt ~ t + dummies)
fv_dummy <- ts(lm_dummy$fitted.values, start = c(1986, 1), end = c(2007, 3),
         frequency = 12)
plot(yt, main = "Seasonal Dummy", las = 1)
lines(fv_dummy, col = "blue")
```



(d) Compare the fitted values from (b) and those from (c). What do you notice?

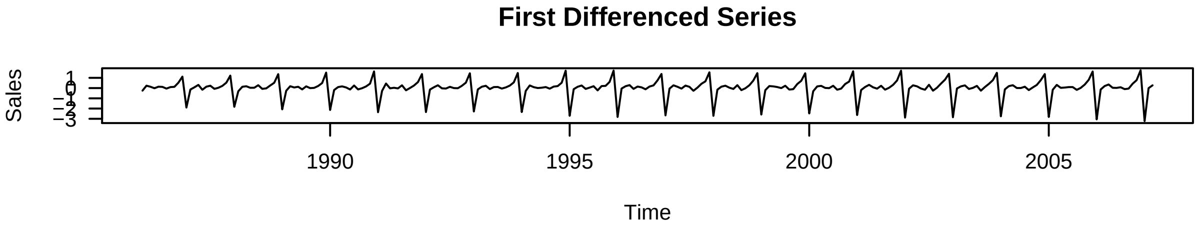
```
are_they_equal <- function(a, b, tolerance=10^(-10)){
  difference <- a - b
  if (sum(abs(difference)) < tolerance) {
    TRUE
  } else {
    FALSE
  }
}
are_they_equal(fv_dummy, fv_sinusoid)
```

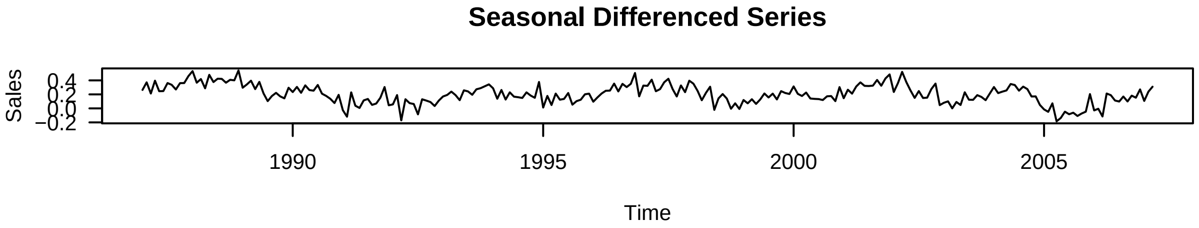
```
## [1] TRUE
```

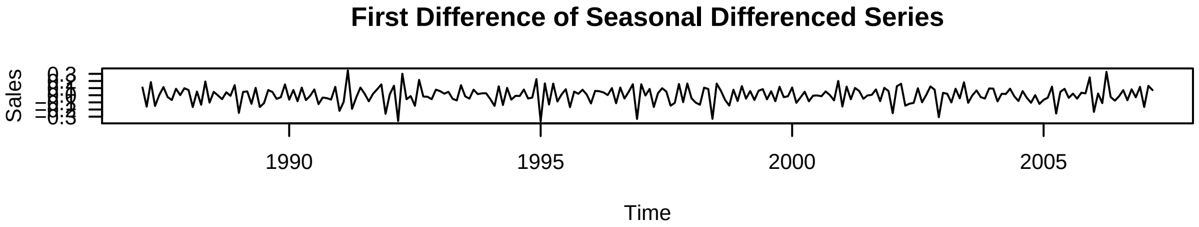
(e) Plot the following three versions of (y_t):

- (i) (∇y_t): the first difference of the data
- (ii) $(\nabla_{12} y_t) = (y_t - y_{t-12})$: the seasonal difference of the data (with 12 months being a season)
- (iii) $(\nabla \nabla_{12} y_t)$: the first difference of the seasonal difference of the data (with 12 months being a season) which one look more like white noise?

```
diff_yt <- diff(yt)
seas_yt <- diff(yt, lag = 12)
diff_seas_yt <- diff(diff(yt, lag = 12))
par(mfrow=c(3,1))
plot(diff_yt, main = "First Differenced Series", las = 1)
plot(seas_yt, main = "Seasonal Differenced Series", las = 1)
plot(diff_seas_yt, main = "First Difference of Seasonal Differenced Series",
    las = 1)
```


First Differenced Series


Seasonal Differenced Series


First Difference of Seasonal Differenced Series

The third one looks more like white noise.

Theoretical exercises:

3. (Stationarity does not imply strict stationarity) Consider (wt) i.i.d. Normal(0,1). Define

$$x_t = \begin{cases} w_t & \text{if } t = 1, 3, 5, 7, \dots \\ \frac{1}{\sqrt{2}}(w_{t-1}^2 - 1) & \text{if } t = 2, 4, 6, 8, \dots \end{cases}$$

(a) Compute the mean function for x_t.
$\mathbb{E}[x_t] = \begin{cases} E[w_t] = 0 & \text{if } t = 1, 3, 5, 7, \dots \\ \mathbb{E}[\frac{1}{\sqrt{2}}(w_{t-1}^2 - 1)] = 0 & \text{if } t = 2, 4, 6, 8, \dots \end{cases}$
$\implies \mathbb{E}[x_t] = 0$

(b) Compute the autocovariance function for x_t.
For $t = 1, 3, 5, 7, \dots,$
$Cov(x_t, x_{t+1}) = Cov(w_t, \frac{1}{\sqrt{2}}(w_t^2 - 1)) = \frac{1}{\sqrt{2}}Cov(w_t, w_t^2) = 0$
$Cov(x_t, x_t) = Cov(w_t, w_t) = 1$
$Cov(x_t, x_s) = 0$ otherwise.
For $t = 2, 4, 6, 8, \dots,$
$Cov(x_t, x_{t-1}) = Cov(\frac{1}{\sqrt{2}}(w_{t-1}^2 - 1), w_{t-1}) = \frac{1}{\sqrt{2}}Cov(w_{t-1}^2, w_{t-1}) = 0$
$Cov(x_t, x_t) = Cov(\frac{1}{\sqrt{2}}(w_{t-1}^2 - 1), \frac{1}{\sqrt{2}}(w_{t-1}^2 - 1)) = \frac{1}{2}Cov(w_{t-1}^2, w_{t-1}^2) = 1$
$Cov(x_t, x_s) = 0$ otherwise.

(c) Explain why (x_t) is zero mean white noise with variance 1, and hence weakly stationary.
Autocorrelation of x_t is 1 at lag 0 and 0 at the other lags. This implies x_t is a white noise. Since its mean and autocovariance do not depend on t, it is weakly stationary.

(d) Are x_1 and x_2 identically distributed? If not, find thier corresponding density functions.
They are not identically distributed. $w_t \sim N(0,1)$ and $w_t^2 \sim \chi_1^2$.

(e) Is (x_t) strictly stationary? Why or why not?
It is not strictly stationary because x_1 and x_2 have different distributions.

4. (Check invertibility of an MA process) For each of the following MA process, check if it is invertible.

(a) $x_t = w_t - w_{t-1} + \frac{3}{16}w_{t-2}$
$\theta(z) = 1 - z + \frac{3}{16}z^2 = (1 - \frac{1}{4}z)(1 - \frac{3}{4}z) = 0 \implies z_1 = 4 > 0, z_2 = \frac{4}{3} > 0$
It is invertible.

(b) $x_t = 2w_{t-2} + 0.4w_{t-1} + w_t$
$\theta(z) = 2z^2 + 0.4z + 1 = 0 \implies z = \frac{-0.4 \pm \sqrt{0.16-8}}{4} = -0.1 \pm 0.7i$
$|z| = \sqrt{0.01 + 0.49} < 1 \implies$ it is not invertible.

(c) $x_t = (1 - \frac{1}{2}B)(1 - \frac{2}{3}B)(1 - \frac{5}{2}B)w_t$
It is not invertible.

5. (Linear combinations of observations from a weakly stationary process) Let (X_t) be a weakly stationary process with mean μ and autocovariance function $\gamma(k) = Cov(X_t, X_{t+k})$. Consider a derived series (Y_t) defined as

$$Y_t = \sum_{j=a}^{b} c_j X_{t+j} \qquad (4)$$

where a and b are integers with a $\leq$ b, and $(c_a, ..., c_b)$ are all fixed real numbers.

(a) Show that $\{Y_t\}$ is a weakly stationary series.
$Cov(Y_t, Y_{t+h}) = Cov(\sum_{j=a}^{b} c_j X_{t+j}, \sum_{k=a}^{b} c_k X_{t+h+k}) = \sum_{j=a}^{b} c_j \sum_{k=a}^{b} c_k \gamma(h + k - j)$
Since X_t is weakly stationary, its autocovariance function does not depend on t. Autocovariance of Y_t does not depend on t as well.

(b) Show that order k (k $\geq$ 1) differencing (that is, $\nabla^k X_t$) can be put in the form of (4). Identify the corresponding a, b, and $(c_a, ..., c_b)$.
$\nabla^k X_t = \sum_{i=0}^{k} \binom{k}{i}(-1)^i X_{t-i}$
$b = 0,\ a = -k,\ c_j = \binom{k}{|j|}(-1)^{|j|}$

(c) Recall that smoothing via simple averaging with parameter q corresponds to computing for each t

$$\frac{1}{2q+1} \sum_{j=-q}^{q} X_{t+j}$$

Show that smoothing via simple averaging with parameter q can be put in the form of (4). Identify the corresponding a, b, and $(c_a, ..., c_b)$.
$b = q,\ a = -q,\ c_j = \frac{1}{2q+1}$

(d) Is the kth differenced version of a weakly stationary process always weakly stationary? Is the smoothed (via simple averaging) version of a weakly stationary process always weakly stationary?
Since both can be put in the form of (4), they are weakly stationary.