STAT 153 Homework 1

Donggyun Kim

12/15/2018

Computer exercise:

1. Download the google trends time series dataset for the query microsoft.

   (a) Estimate the trend function in the data by fitting a parametric curve to the data. Provide a plot of the original data along with the corresponding trend estimate. Also provide a time plot and correlogram of the residuals. Comment on each of these plots.

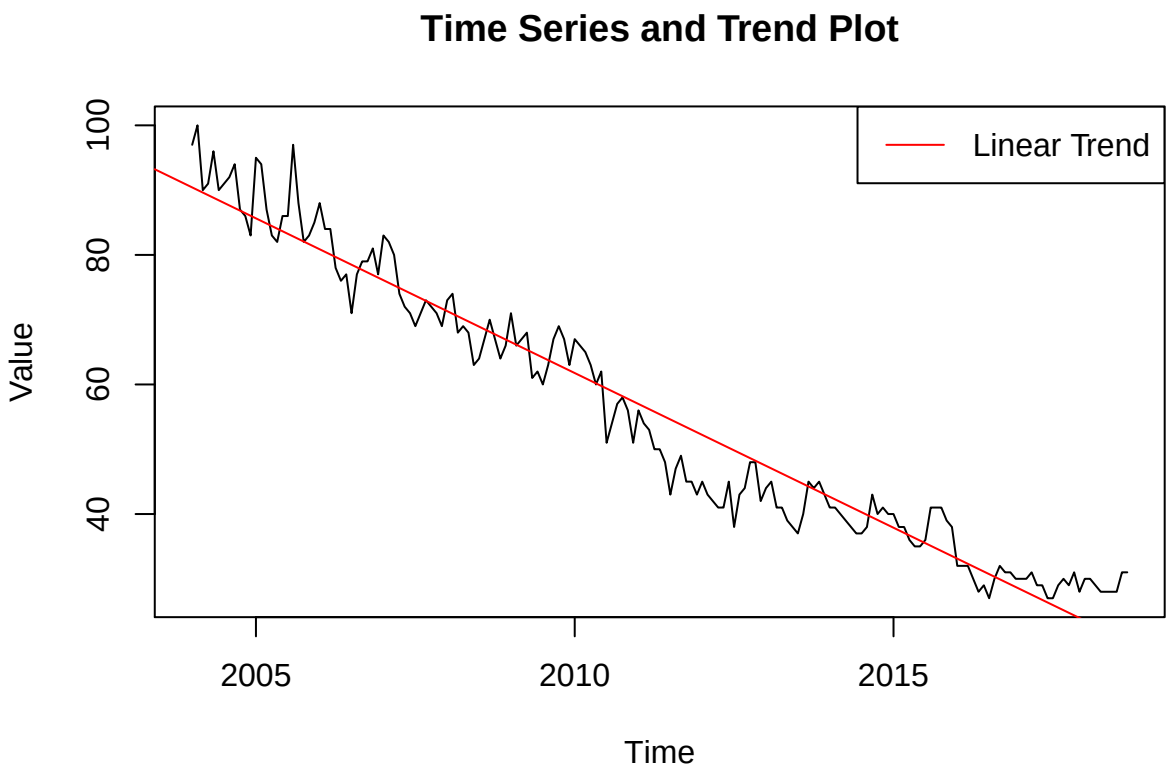
```
# import data
dat <- read.csv("microsoft.csv", stringsAsFactors = FALSE)

# data cleaning
dat <- as.numeric(dat[-1, ])
my_ts <- ts(dat, start = c(2004, 1), end = c(2018, 9), frequency = 12)
my_ts
```

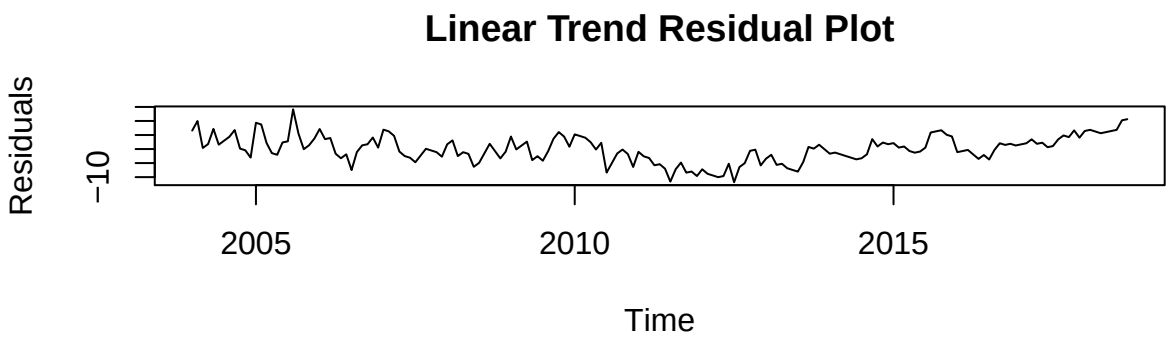
```
##      Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
## 2004  97 100  90  91  96  90  91  92  94  87  86  83
## 2005  95  94  87  83  82  86  86  97  88  82  83  85
## 2006  88  84  84  78  76  77  71  77  79  79  81  77
## 2007  83  82  80  74  72  71  69  71  73  72  71  69
## 2008  73  74  68  69  68  63  64  67  70  67  64  66
## 2009  71  66  67  68  61  62  60  63  67  69  67  63
## 2010  67  66  65  63  60  62  51  54  57  58  56  51
## 2011  56  54  53  50  50  48  43  47  49  45  45  43
## 2012  45  43  42  41  41  45  38  43  44  48  48  42
## 2013  44  45  41  41  39  38  37  40  45  44  45  43
## 2014  41  41  40  39  38  37  37  38  43  40  41  40
## 2015  40  38  38  36  35  35  36  41  41  41  39  38
## 2016  32  32  32  30  28  29  27  30  32  31  31  30
## 2017  30  30  31  29  29  27  27  29  30  29  31  28
## 2018  30  30  29  28  28  28  28  31  31
```

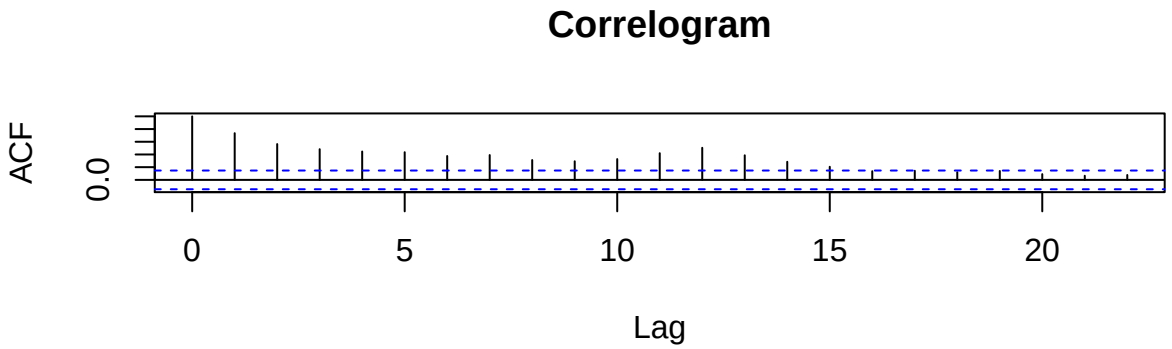
```
# trend estimate
t <- time(my_ts)
lm_trend <- lm(dat ~ t)

# plot
plot(my_ts, main = "Time Series and Trend Plot", ylab = "Value")
abline(lm_trend, col = "red")
legend("topright", legend = "Linear Trend", col = "red", lty = 1)
```




```
# residual plot and correlogram
r <- lm_trend$residuals
par(mfrow=c(2,1))
plot(x = t, y = r, type = "l", ylab = "Residuals", xlab = "Time",
     main = "Linear Trend Residual Plot")
acf(r, main = "Correlogram")
```



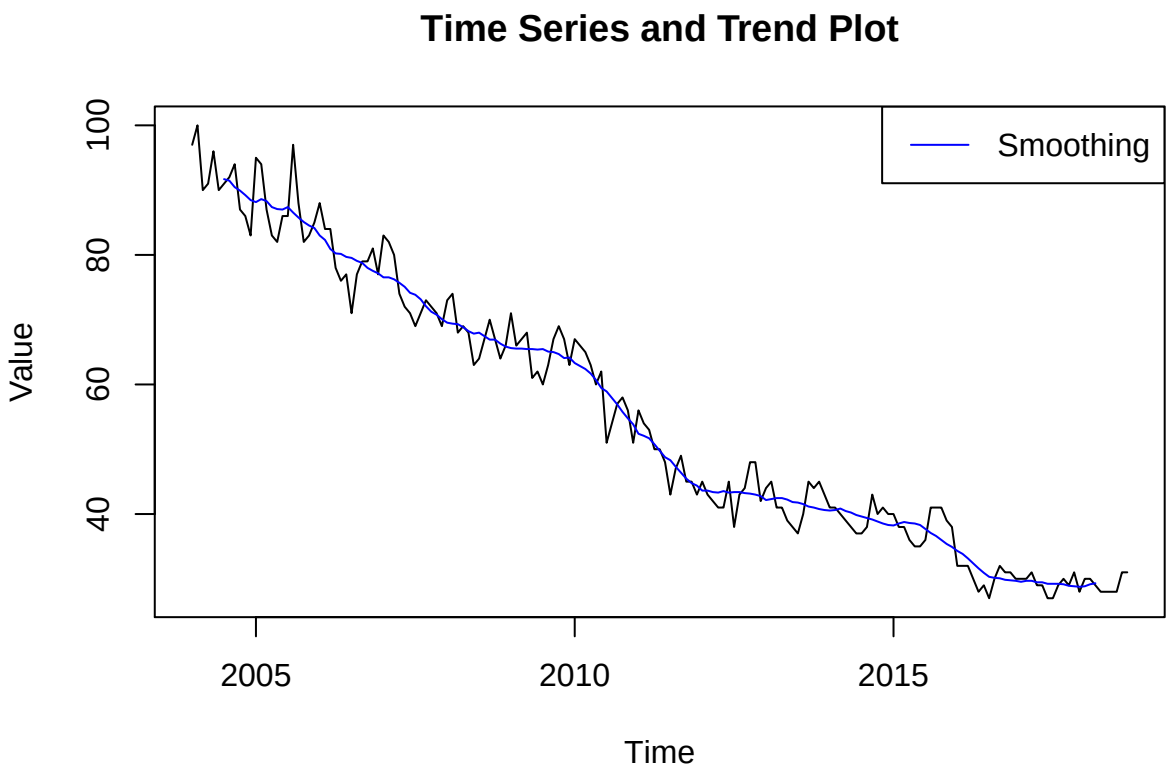


Origianl data plot shows that there is a trend in the series. The trend can be estimated by linear regreesion model. To see if it is good fit of the trend, find residuals of model and plot a correlogram. The correlogram above shows that the residuals do not act like white noise. Therefore, linear trend is not a good fit.

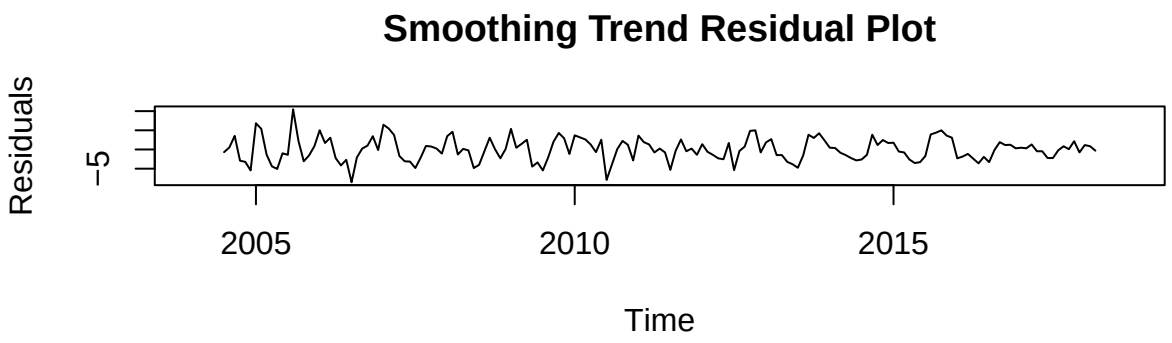
(b) Estimate the trend by smoothing. Explain reasons behind your choice of the smoothing parameter. Once again, provide a plot of the original data along with the corresponding trend estimate. Also provide a time plot and correlogram of the residuals. Comment on each of these plots.

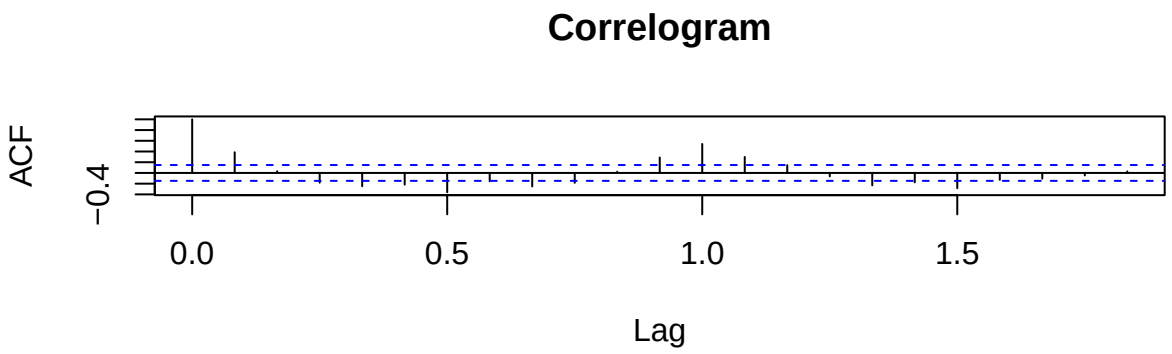
```r
# smoothing with parameter q = 6
filter_trend <- filter(my_ts, filter = rep(1/(2*6 + 1), (2*6 + 1)))

# plot
plot(my_ts, main = "Time Series and Trend Plot", ylab = "Value")
lines(filter_trend, col = "blue")
legend("topright", legend = "Smoothing", col = "blue", lty = 1)
```




```
# residual plot and correlogram
r <- my_ts - filter_trend
par(mfrow=c(2, 1))
plot(x = t, y = r, type = "l", ylab = "Residuals", xlab = "Time",
     main = "Smoothing Trend Residual Plot")
acf(r, main = "Correlogram", na.action = na.pass)
```



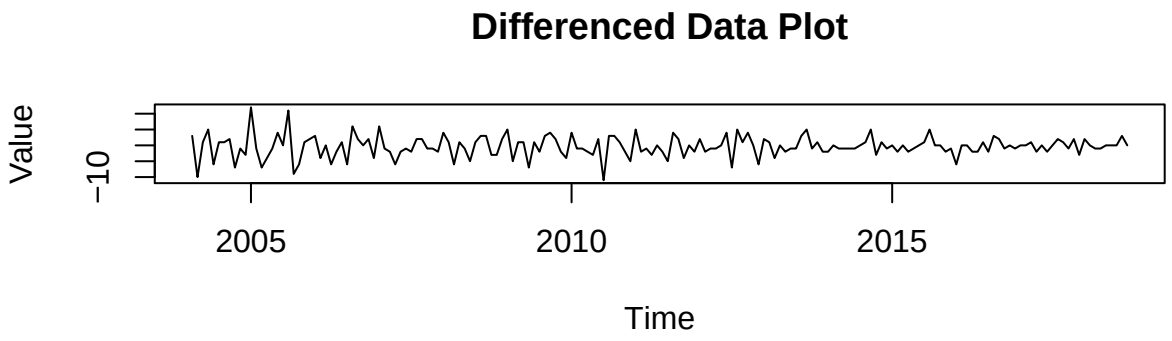


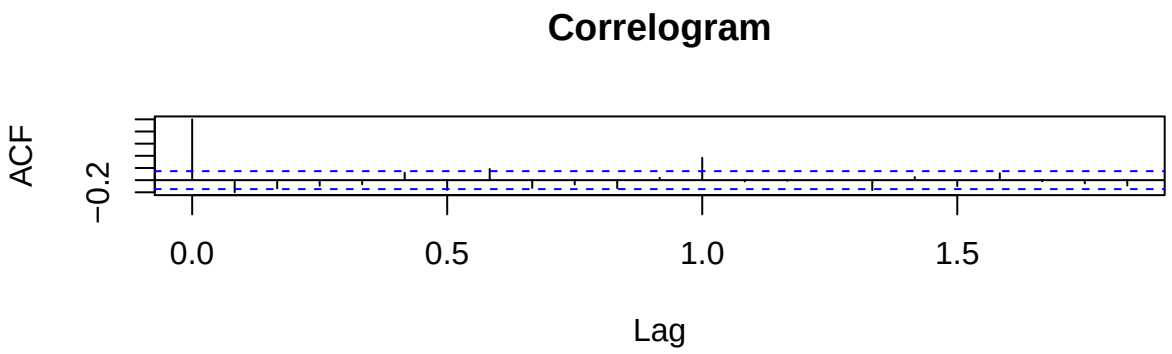
There is no correct answer about choosing smoothing parameter q. Large q yields low variance but high bias. Small q yields high variance but low bias. The correlogram shows that residuals do not act like white noise and there is a seasonality in residuals.

(c) Difference the data and provide a plot of the differenced data. Is there any trend in the differenced data? Also plot the correlogram of the differenced data and comment on it.

```
# differencing
ts_diff <- diff(my_ts)

# plot and correlogram
par(mfrow=c(2,1))
plot(ts_diff, main = "Differenced Data Plot", ylab = "Value")
acf(ts_diff, main = "Correlogram")
```



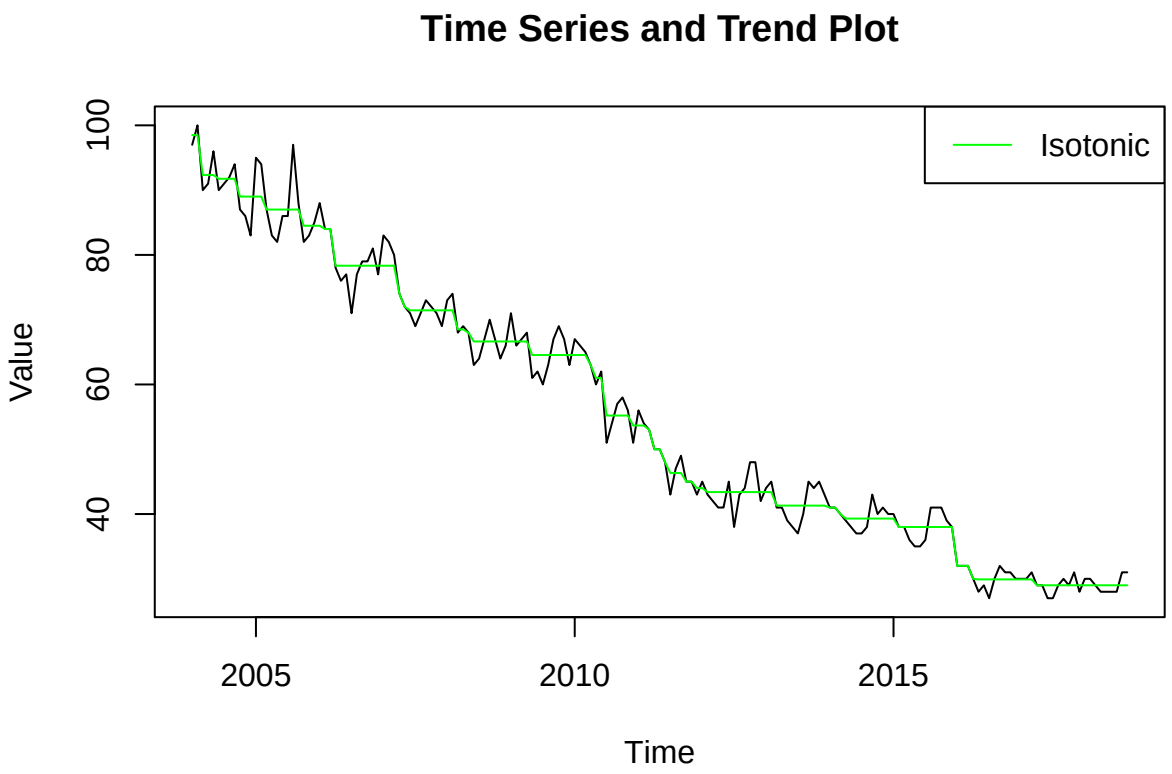


Diffencing method is much better than above two methods about removing a trend. Correlogram shows residuals act like white noise and there is a positive autocorrelation at lag 1.

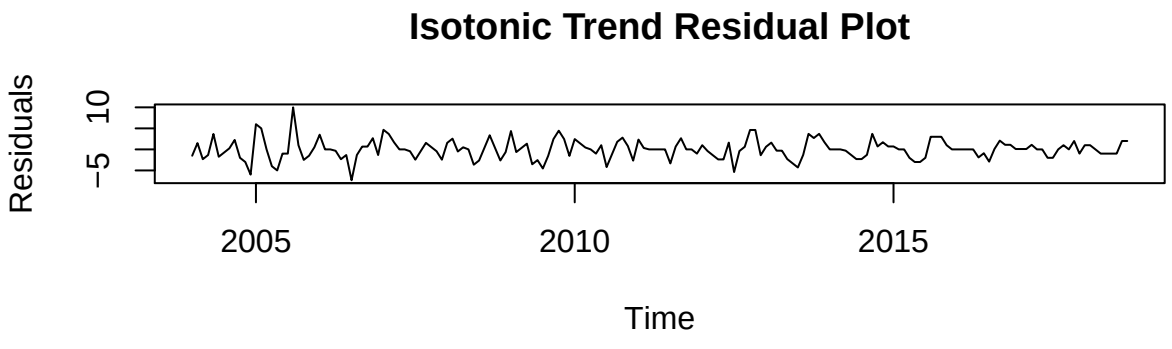
(d) Estimate the trend function in the data using isotonic estimation. Provide a plot of the original data along with the corresponding trend estimate. Also provide a time plot and correlogram of the residuals. Comment on each of the plots.

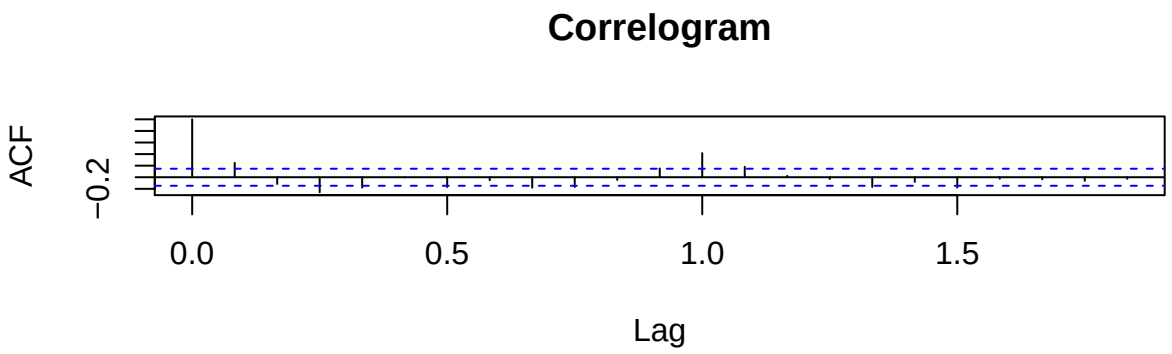
```r
# isotonic estimation
isoreg_trend <- isoreg(-my_ts)
fv <- -isoreg_trend$yf
fv_ts <- ts(fv, start = c(2004, 1), end = c(2018, 9), frequency = 12)

# plot
plot(my_ts, ylab = "Value", main = "Time Series and Trend Plot")
lines(fv_ts, col = "green")
legend("topright", legend = "Isotonic", col = "green", lty = 1)
```




```
# residual plot and correlogram
r <- my_ts - fv_ts
par(mfrow=c(2,1))
plot(x = t, y = r, type = "l", ylab = "Residuals", xlab = "Time",
     main = "Isotonic Trend Residual Plot")
acf(r, main = "Correlogram")
```





Isotonic trend estimation works well for this series. The correlogram shows residuals act like white noise and there is a positive autocorrelation at lag 1.

Theoretical exercise:

2. While analyzing their annual sales data (number of car sales per year) for the past 30 years, a car company found that after taking three successive differences, the resulting data had a mean of 2051 and looked like white noise. If the actual data for the past three years were 2017 - 61214, 2016 - 52574, and 2015 - 39381. What would be a reasonable forecast for sales in 2018? Explain.

   $\nabla^3 X_t = \nabla^2(X_t - X_{t-1}) = \nabla(X_t - 2X_{t-1} + X_{t-2}) = X_t - 3X_{t-1} + 3X_{t-2} - X_{t-3} = 2051 + Z_t$
   $\implies X_t = 3X_{t-1} - 3X_{t-2} + X_{t-3} + 2051 + Z_t$
   $X_{2018} = 3X_{2017} - 3X_{2016} + X_{2015} + 2051 + \mathbb{E}[Z_t]$
   $X_{2018} = 3 * 61214 - 3 * 52574 + 39381 + 2051 = 67352$

3. For white noise $X_1, \ldots, Xn$ with n sufficiently large, what is (approximately) the probability that if you plot in R the correlogram for the first 100 lags $r_1, \ldots, r_{100}$ at least 3 of the r_k's lie outside of the blue 5% confidence band?

   By theorem from lecture note, sample autocorrelations of white noise are approximately normally distributed with mean 0 and variance $\frac{1}{n}$ as $n \to \infty$.
   $1 - \binom{100}{0}(0.95)^{100} - \binom{100}{1}(0.05)(0.95)^{99} - \binom{100}{2}(0.05)^2(0.95)^{98}$

4. Consider the stochastic trend model

$$X_t = m_t + Z_t, \quad t = 1, \ldots, n$$

where Z_t is a white noise process (with variance σ_Z^2) and m_t is a stochastic trend which follows a random walk with drift model

$$m_t = \delta + m_{t-1} + W_t$$

where W_t is another white noise process (with variance σ_W^2), independent of Z_t. Let $m_0 = 0$.

(a) Find the mean function $\mu(t) = \mathbb{E}(X_t)$, autocovariance function $\gamma(s,t) = Cov(X_s, X_t)$, and autocorrelation function $\rho(s,t) = \frac{\gamma(s,t)}{\sqrt{\gamma(s,s)\gamma(t,t)}}$ of $\{X_t\}$.

$X_t = m_t + Z_t = \delta + m_{t-1} + W_t + Z_t = \cdots = t\delta + \sum_{k=1}^{t} W_k + Z_t$

$\gamma(s,t) = Cov(X_s, X_t) = Cov(s\delta + \sum_{k=1}^{s} W_k + Z_s, t\delta + \sum_{j=1}^{t} W_j + Z_t) = \begin{cases} \min(s,t)\sigma_W^2 & \text{if } s \neq t, \\ t\sigma_W^2 + \sigma_Z^2 & \text{if } s = t \end{cases}$

$\rho(s,t) = \frac{\min(s,t)\sigma_W^2}{\sqrt{s\sigma_W^2+\sigma_Z^2}\sqrt{t\sigma_W^2+\sigma_Z^2}}$

(b) A time series process is denoted as weakly stationary if the mean function $\mu(t)$ is constant (i.e., does not depend on t) and the autocovariance function $\gamma(s,t)$ depends on s and t only through their difference $|s-t|$. Is the process X_t defined above weakly stationary? Explain.
No. $\mathbb{E}[X_t] = \mathbb{E}[t\delta + \sum_{j=1}^{t} W_j + Z_t] = t\delta$ does depend on t. Therefore, X_t is not weakly stationary.

(c) Show that $\rho(t-1,t) \to 1$ as $t \to \infty$. What is the implication of this result?
$\lim_{t\to\infty} \rho(t-1,t) = \lim_{t\to\infty} \frac{(t-1)\sigma_W^2}{\sqrt{((t-1)\sigma_W^2+\sigma_Z^2)(t\sigma_W^2+\sigma_Z^2)}} = 1$
In the distant future, autocorrelation between two successive times will be 1.

(d) Suggest a transformation to make the series X_t weakly stationary and prove that the transformed series is weakly stationary.
$Y_t = X_t - X_{t-1} = t\delta + \sum_{k=1}^{t} W_k + Z_t - ((t-1)\delta + \sum_{k=1}^{t-1} W_k + Z_{t-1}) = \delta + W_t + Z_t - Z_{t-1}$
$\mathbb{E}[Y_t] = \mathbb{E}[\delta + W_t + Z_t - Z_{t-1}] = \delta \implies$ does not depend on t.

$Cov(Y_s, Y_t) = Cov(\delta + W_s + Z_s - Z_{s-1}, \delta + W_t + Z_t - Z_{t-1}) = \begin{cases} \sigma_W^2 + 2\sigma_Z^2 & \text{if } s = t \\ \sigma_Z^2 & \text{if } |s-t| = 1 \\ 0 & \text{if } |s-t| > 1 \end{cases}$

5. Consider a monthly time series dataset for which we believe that the model $X_t = (at+b)s_t + Z_t$ is appropriate where s_t is a seasonal function of known period d i.e., $s_{t+d} = s_t$ and Z_t is white noise. What would be a way to difference the data in order to eliminate both trend and seasonality?
$Y_t = X_t - X_{t-d} = (at+b)s_t + Z_t - ((a(t-d)+b)s_{t-d} + Z_{t-d}) = ads_{t-d} + Z_t - Z_{t-d}$
$Y_t - Y_{t-d} = ads_{t-d} + Z_t - Z_{t-d} - (ads_{t-2d} + Z_{t-d} - Z_{t-2d}) = Z_t - 2Z_{t-d} + Z_{t-2d}$
$Y_t - Y_{t-d}$ is a MA process.