MSDS 596 Regression & Time Series Fall 2018

Homework 4 Solution

Due: Tues 11/13/18 @ 6:40pm rutgers.instructure.com/courses/17597

Attention:

- You'll need to download two datasets for this homework: `hawaii-new.dat` and `lt.dat`. Both are available on Canvas.
- Be sure to save your `R` code for Problems 3 and 4. You will need them for the next homework.

Problem 1. Let U_1 and U_2 be independent random variables with zero means and $\text{Var}(U_1) = \text{Var}(U_2) = \sigma^2$. Consider the time series

$$x_t = U_1 \sin(2\pi\omega t) + U_2 \cos(2\pi\omega t),$$

where $\omega \in [0, 1)$ is a fixed constant.

(a) Show that this series is weakly stationary and find γ_h, its lag-h autocovariance function. [Hint: $\sin\alpha \sin\beta + \cos\alpha \cos\beta = \cos(\alpha - \beta)$.]

(b) Find the lag-h autocorrelation function.

Solution. (a). For $\{x_t\}$ to be stationary, we need to show that 1) its mean $\mu_t = E(x_t)$ is constant with respect to t, and 2) the autocovariance between x_t and x_{t+h} for any t is a function of the lag h only. We have that for all t,

$$\mu_t = E(x_t) = \sin(2\pi\omega t)E(U_1) + \cos(2\pi\omega t)E(U_2) = 0,$$

and

$$\begin{aligned} Cov(x_t, x_{t+h}) &= Cov(U_1 \sin(2\pi\omega t) + U_2 \cos(2\pi\omega t), U_1 \sin(2\pi\omega(t+h)) + U_2 \cos(2\pi\omega(t+h))) \\ &= \sin(2\pi\omega t)\sin(2\pi\omega(t+h))Var(U_1) + \cos(2\pi\omega t)\cos(2\pi\omega(t+h))Var(U_2) \\ &= \sigma^2(\sin(2\pi\omega t)\sin(2\pi\omega(t+h)) + \cos(2\pi\omega t)\cos(2\pi\omega(t+h))) \\ &= \sigma^2\cos(2\pi\omega h), \end{aligned}$$

which is a function of h only. Thus $\{x_t\}$ is weakly stationary as desired, and $\gamma_h = \sigma^2\cos(2\pi\omega h)$.

(b).

$$\rho_h = \gamma_h/\gamma_0 = \cos(2\pi\omega h).$$

Problem 2. Let w_t, $t \in \mathbb{Z}$ be a normal white noise (i.e. they are iid normal) with variance 1, and consider the time series

$$x_t = w_t w_{t-1}$$

(a) Find the mean, autocovariance, and autocorrelation functions of x_t.

(b) Simulate x_t of length $T = 500$. Give the time series plot, and the sample autocorrelations plot. Comment.

(c) Perform the Ljung-Box test on the simulated series x_t, using $m = 1, 2, 3, 4, 5, 6$.

Solution. (a). The mean

$$\mu_t = E(x_t) = E(w_t w_{t-1}) = Cov(w_t, w_{t-1}) = 0.$$

We also have that

$$\begin{aligned} Cov(x_t, x_t) &= 1, \\ \text{for } h \geq 1, \quad Cov(x_t, x_{t+h}) &= Cov\left(w_t w_{t-1}, w_{t+h} w_{t+(h-1)}\right) = 0 \end{aligned}$$

Thus the time series $\{x_t\}$ is weakly stationary, with autocovariance function

$$\gamma_h = \mathbf{1}\{h = 0\},$$

and autocorrelation function

$$\rho_h = \gamma_h / \gamma_0 = \mathbf{1}\{h = 0\}.$$

(b) and (c). See R output attached.

Problem 3. Tourism is one of the largest economic components of Hawaii. The data `hawaii-new.dat` contains monthly record of the number of tourists visited Hawaii from January, 1970 to December, 1995. Download the dataset to your R working directory and read the data using the following command:

```
hawaii <- read.table('hawaii-new.dat',
          col.names = c('year_month', 'total', 'west', 'east'))
```

In this dataset, the first column shows the year-month. The second column is the total number of monthly tourists. The third and fourth columns show the number of west-bound (mainly from US and Canada) and east bound (mainly from Japan and Australia) visitors. Perform the following analysis.

(a) Draw time series plots of the three series on the *same* graph. Comment on what you observe (trend, seasonality, variance, relationships, etc.).

(b) Perform a log transformation of the `total` series. Plot the time series and comment on it.

(c) If you are to use a polynomial trend model for the log transformed `total` series, which order of the polynomial (e.g. linear, quadratic, cubic etc) will you use? Use what we've learned about polynomial regression variable selection procedure to decide. Fit the trend model, plot the fitted line with the log-transformed time series and plot the de-trended series. What do you think about the results?

(d) Perform a 13-month moving average (with half weights for the first and last month compared to the rest eleven) of the `total` series. Plot the moving average with the original series, as well as the de-trended series. Comment on what you observe.

Solution. See attached.

Problem 4. Use the `scan()` function to read in the data from the file `lt.txt`, which is a vector of length 500. Consider a local trend model for the data

$$\begin{aligned} y_t &= s_t + e_t, \qquad e_t \sim N(0, 0.25); \\ s_{t+1} &= s_t + \eta_t, \qquad \eta_t \sim N(0, 0.01), \quad s_0 \sim N(0.2, 2.25). \end{aligned}$$

Adopt the same notation from the lecture: $s_{t|t-1} = E(s_t \mid F_{t-1})$, $\Sigma_{t|t-1} = \text{Var}(s_t \mid F_{t-1})$, $y_{t|t-1} = E(y_t \mid F_{t-1})$, $v_t = y_t - y_{t|t-1} = y_t - s_{t|t-1}$, and $V_t = \text{Var}(v_t \mid F_{t-1})$.

(a) Prove that $E(v_t) = 0$, and $\text{Cov}(v_t, y_j) = 0$ when $j < t$. [Hint: *Law of Iterated Expectation* says for random variables X and Y, $E(E(X \mid Y)) = E(X)$. Also, $E(XY \mid X) = X \cdot E(Y \mid X)$. The proofs should only be about two lines long.]

For the next three questions, write your own program to implement the Kalman filter, using $s_{1|0} = 0.2$, and $\Sigma_{1|0} = 2.26$ as the initial values.

(b) Plot the predicted state variables $s_{t|t-1}$ for $1 \leq t \leq T$, together with the 95% confidence intervals $s_{t|t-1} \pm 2\sqrt{\Sigma_{t|t-1}}$. Your plot should look similar to Figure 11.4 of the textbook.

(c) Plot the filtered state variables $s_{t|t}$ together with the 95% confidence intervals $s_{t|t} \pm 2\sqrt{\Sigma_{t|t}}$.

(d) Plot the smoothed state variables $s_{t|T}$ together with the 95% confidence intervals $s_{t|T} \pm 2\sqrt{\Sigma_{t|T}}$.

Solution. (a).

$$\begin{aligned} E(v_t) &= E(y_t - E(y_t \mid F_{t-1})) = E(y_t) - E(E(y_t \mid F_{t-1})) = 0. \\ Cov(v_t, y_j) &= E(v_t y_j) - E(v_t)E(y_j) = E(v_t y_j) = E(E(v_t y_j \mid F_{t-1})) \\ &= E(y_j \cdot E(v_t \mid F_{t-1})) = E\left(y_j \cdot \left(y_{t|t-1} - y_{t|t-1}\right)\right) = 0. \end{aligned}$$

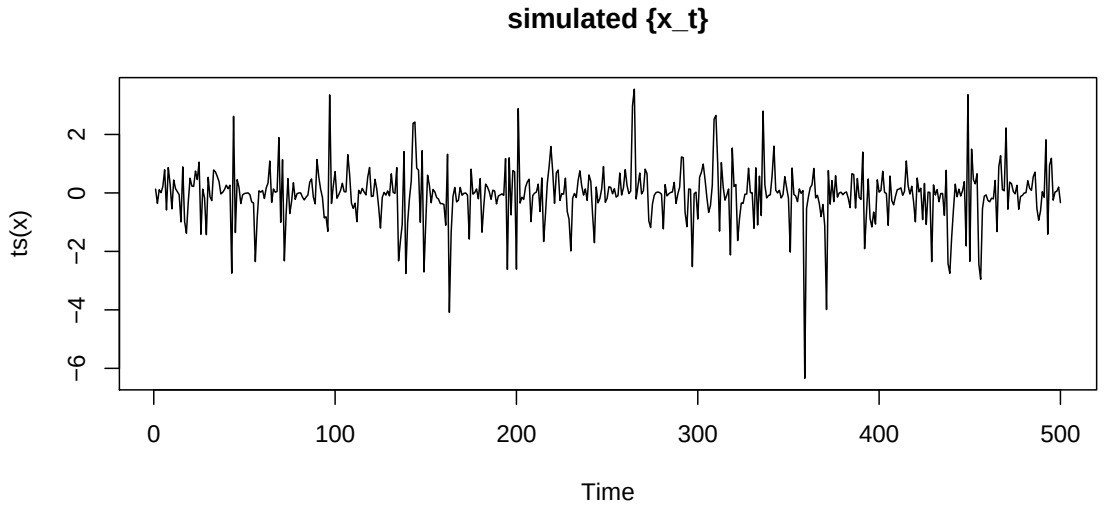
(b) - (d). See R output attached.

Problem 2.

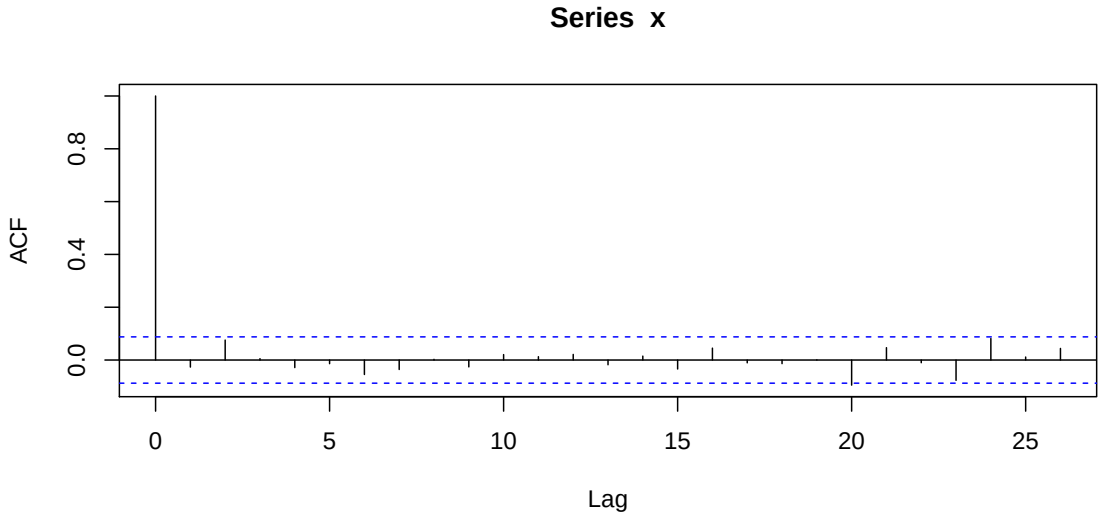
(b).

The simulated time series $\{x_t\}$ for $1 \leq t \leq 500$ and its ACF plot are as follows. The mean of the series appears to be stable with average around 0. The volatility of the time series isn't stable and shows clustering. The sample autocorrelation function at all non-zero lags appear to be within 95% of sample standard errors.

```
set.seed(123)
w <- rnorm(501, 0, 1)
x <- w[-1]*w[-501]
plot(ts(x), main = 'simulated {x_t}')
```




```
acf(x)
```



(c).

Ljung-Box test for autocorrelations up till lag 1 through lag 6 shows that none of the null hypotheses:

$$H_{0(m)} : \rho_1 = \cdots \rho_m = 0,$$

for $m = 1, ..., 6$, can be rejected at level $\alpha = 5\%$.

```
LB <- array(NA, dim = c(6, 3))
colnames(LB) <- c('X^2', 'df', 'p-value')
rownames(LB) <- paste('Lag', 1:6)
for (i in 1:6){
  lb <- Box.test(x, lag = i, type = 'Ljung-Box')
  LB[i, ] <- c(lb$statistic, lb$parameter, lb$p.value)
}
LB
```

```
##             X^2 df   p-value
## Lag 1 0.3582146  1 0.5494995
## Lag 2 3.2384293  2 0.1980542
## Lag 3 3.2524174  3 0.3543203
## Lag 4 3.6673378  4 0.4528943
## Lag 5 3.7648142  5 0.5837489
## Lag 6 5.2879206  6 0.5074504
```

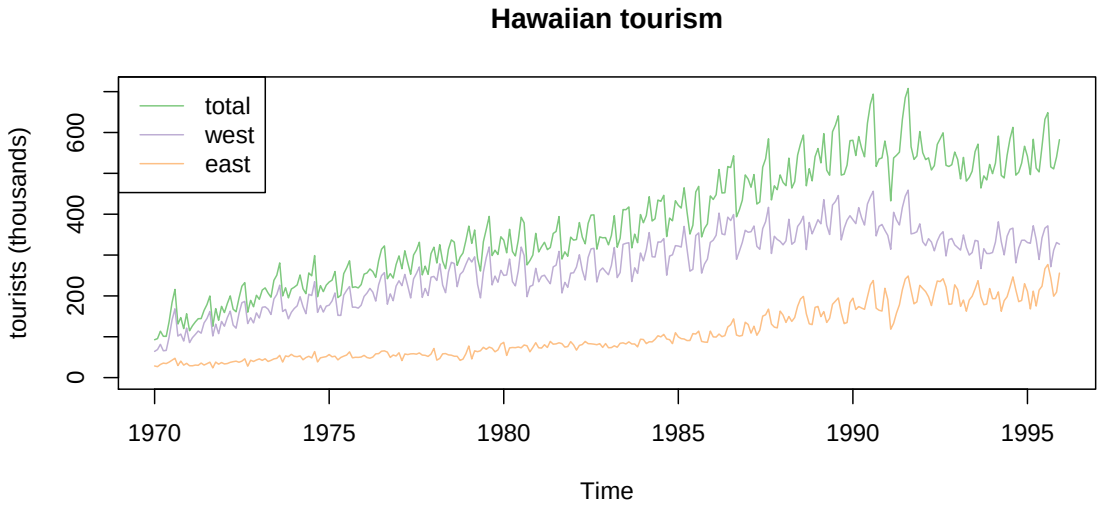
Problem 3.

(a).

The number of incoming tourists to Hawaii each month from year 1970 to 1995 are plotted as follows. Green represents the total number of tourists (in thousands) each month, whereas purple and orange are the west-bound and east-bound tourists respectively.

We can see that there exists an overal increasing trend among all three tourist time series up till around 1990. Since then, the west-bound tourists suffers a slight decrease which results in a corresponding decrease in the total number of tourists. In addition, there is increasing volatility over the years, and strong seasonality per year period.

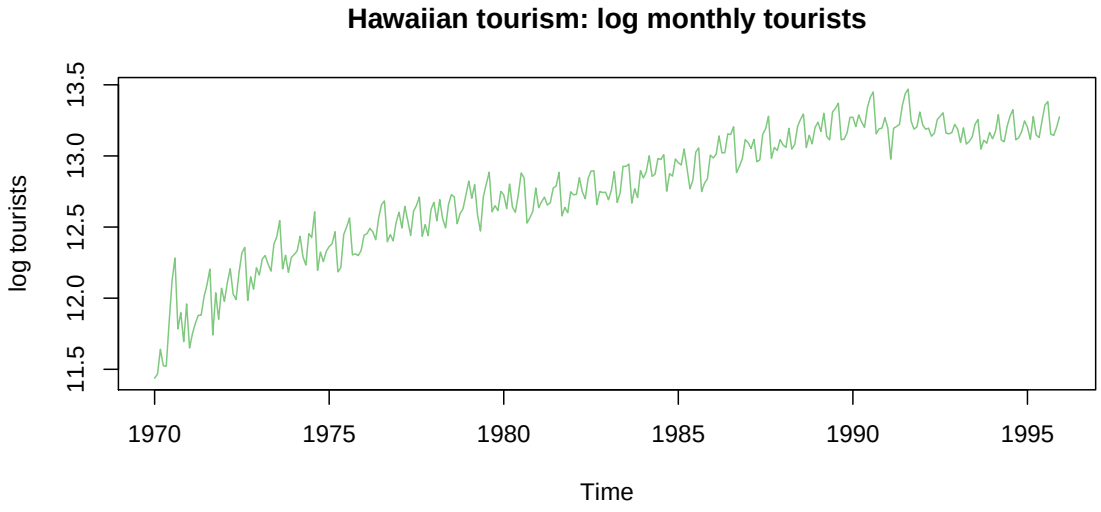
```
hawaii <- read.table('hawaii-new.dat',
                     col.names = c('year_month', 'total', 'west', 'east'))
total.ts <- ts(hawaii$total, start=1970, freq = 12)
west.ts <- ts(hawaii$west, start=1970, freq = 12)
east.ts <- ts(hawaii$east, start=1970, freq = 12)
pal <- brewer.pal(6, 'Accent')
plot.ts(total.ts/1e3, col = pal[1], ylim = c(0, max(total.ts/1e3)),
        ylab = 'tourists (thousands)', main = 'Hawaiian tourism')
lines(east.ts/1e3, col = pal[2])
lines(west.ts/1e3, col = pal[3])
legend('topleft', col = pal, lty = 1, legend = colnames(hawaii)[-1])
```



(b).

After logging the total time series we observe that the overall trend becomes closer to that of a quadratic, while volatility became stable over the years.

```
plot.ts(log(total.ts), col = pal[1], ylab = 'log tourists',
        main = 'Hawaiian tourism: log monthly tourists')
```



(c).

Using what we learned in polynomial regression, let's start with a higher polynomial order (3) and decrease if the linear model deems it excessive. We begin by fitting a cubic model

```
log(total) ~ year_month + year_month^2 + year_month^3
```

```
summary(lm(log(total)~year_month+I(year_month^2)+I(year_month^3), data = hawaii))
```

```
##
## Call:
```

```
## lm(formula = log(total) ~ year_month + I(year_month^2) + I(year_month^3),
##     data = hawaii)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.41854 -0.09007 -0.00418  0.08190  0.41845
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -2.182e+01  1.217e+01  -1.793   0.0740 .
## year_month       9.892e-03  4.458e-03   2.219   0.0272 *
## I(year_month^2) -9.309e-07  5.420e-07  -1.718   0.0869 .
## I(year_month^3)  2.928e-11  2.187e-11   1.339   0.1817
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.127 on 308 degrees of freedom
## Multiple R-squared:  0.9158, Adjusted R-squared:  0.915
## F-statistic:  1116 on 3 and 308 DF,  p-value: < 2.2e-16
```

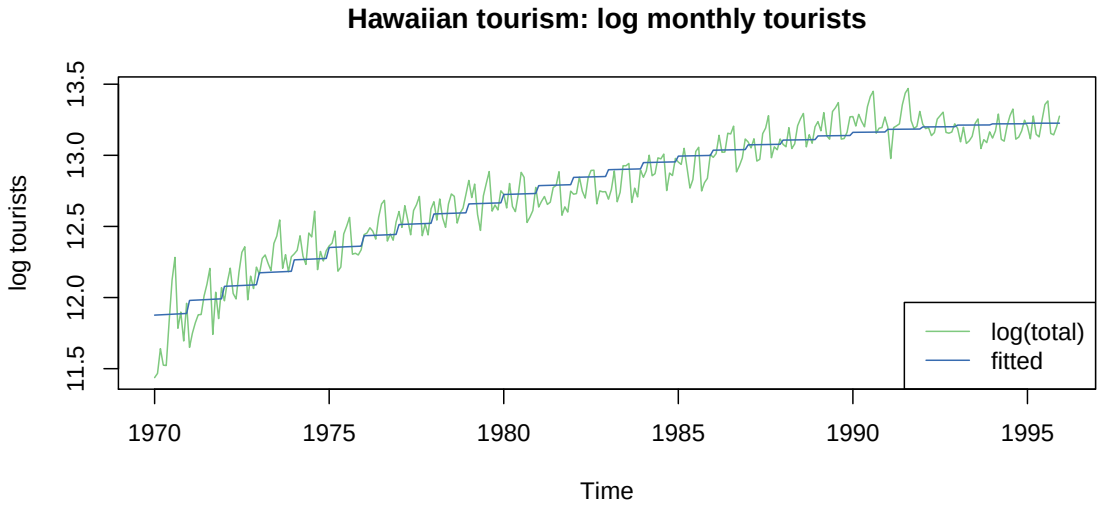
and realize that the highest order term is not significant. So, reduce to a quadratic model:

```
lmod <- lm(log(total)~year_month+I(year_month^2), data = hawaii)
summary(lmod)
```

```
##
## Call:
## lm(formula = log(total) ~ year_month + I(year_month^2), data = hawaii)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.43934 -0.08393  0.00028  0.08029  0.39840
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -5.580e+00  9.730e-01  -5.734 2.33e-08 ***
## year_month       3.933e-03  2.371e-04  16.591  < 2e-16 ***
## I(year_month^2) -2.056e-07  1.434e-08 -14.336  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1272 on 309 degrees of freedom
## Multiple R-squared:  0.9153, Adjusted R-squared:  0.9147
## F-statistic:  1669 on 2 and 309 DF,  p-value: < 2.2e-16
```

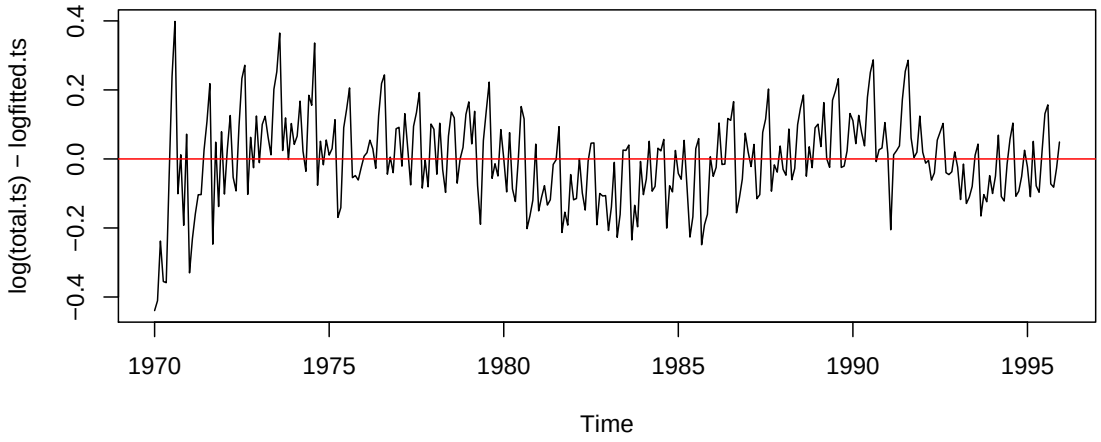
The quadratic term is highly significant, so let's go with that. The fitted values are plotted over the log-transformed time series:

```
logfitted.ts <- ts(lmod$fitted.values, start = 1970, freq = 12)
plot.ts(log(total.ts), col = pal[1], ylab = 'log tourists',
        main = 'Hawaiian tourism: log monthly tourists');
lines(logfitted.ts, col = pal[5])
legend('bottomright', col = pal[c(1, 5)], lty = 1, legend = c('log(total)', 'fitted'))
```



The de-trended series is as follows. While the de-trended series is overall centered around zero, there is clearly a rather long-term periodic trend (of about 15 years per period) that remains.

```
plot.ts(log(total.ts)-logfitted.ts)
abline(h = 0, col = 'red')
```



(d).

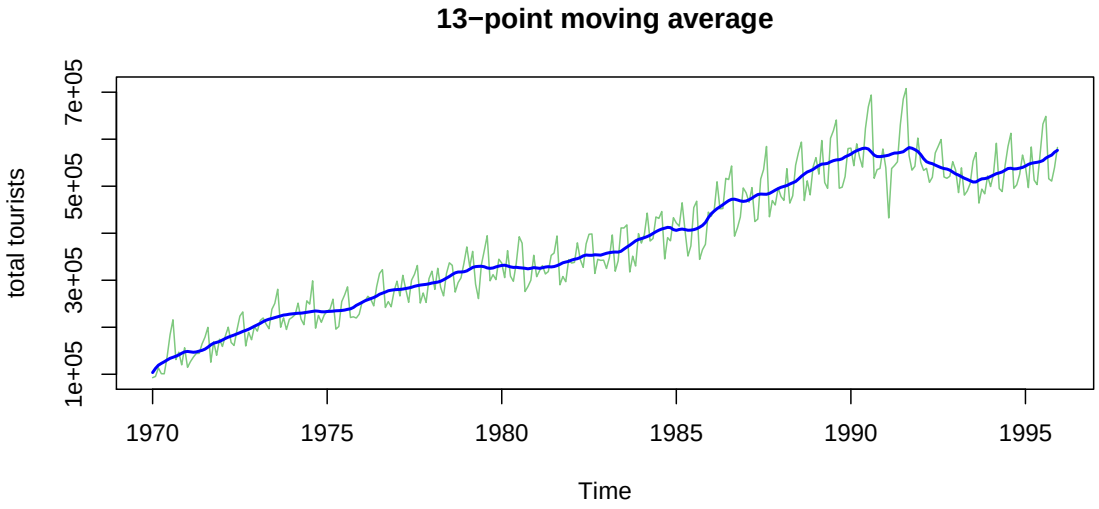
Apply the 13-point moving average with weights:

$$(\frac{1}{24}, \frac{1}{12}, \frac{1}{12}, \cdots, \frac{1}{12}, \frac{1}{24})$$

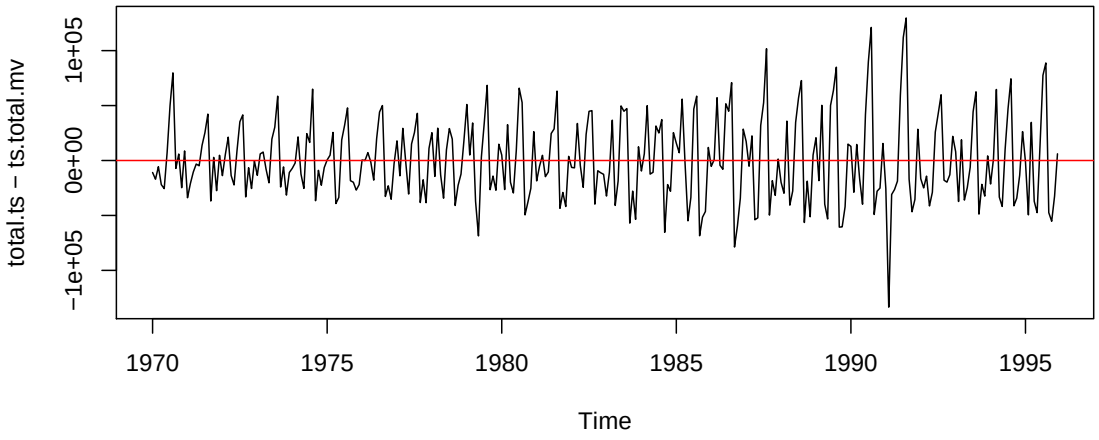
The first and last six boundary points are set to the first and last observations, respectively. Below are the moving average superimposed on the original series as well as the de-trended series. The 13-point moving average seems to be able to follow the time series quite closely. The de-trended series look a lot more uniform and pattern-less compared to that from the previous question (although they're subject to different transformations).

```
t0 = total.ts[1]; tn = rev(total.ts)[1]   # get the first and last data points
total.mv=filter(c(rep(t0, 6),total.ts,rep(tn,6)),
                sides=2, c(1/24, rep(1/12, 11), 1/24))[7:(length(total.ts)+6)]
ts.total.mv  <- ts(total.mv, start=1970, freq = 12)
```

```
plot(total.ts, type="l", col = pal[1], main = '13-point moving average', ylab = 'total tourists')
lines(ts.total.mv,col="blue", lwd = 2)
```




```
plot(total.ts - ts.total.mv); abline(h = 0, col = 'red')
```



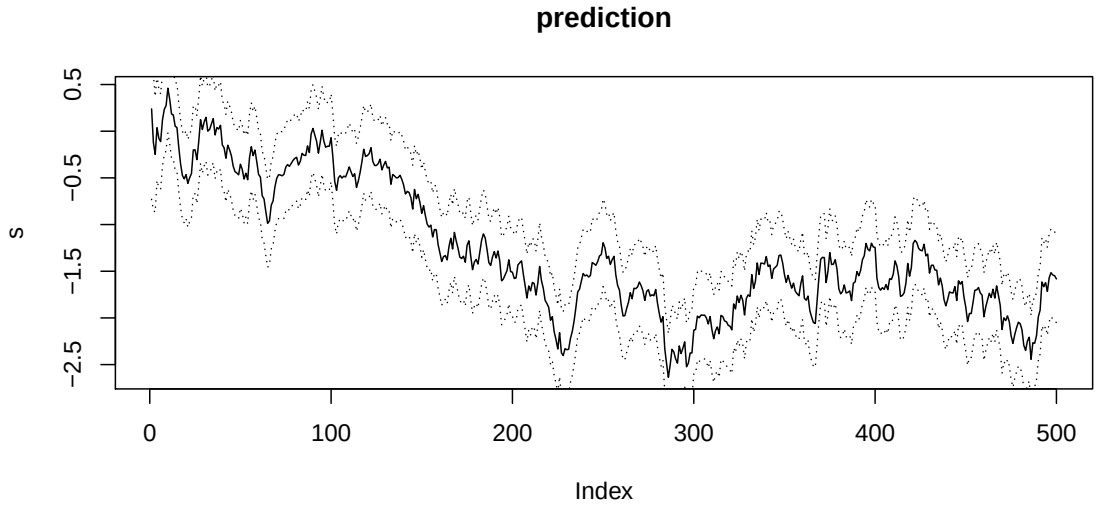
Problem 4.

(b)

Predicted state variables $\{s_{t|t-1}\}$ and 95% confidence intervals:

```
y <- scan('lt.txt')
# initialization
s = c(0.2, rep(NA, 500))
sigma = c(2.26, rep(NA, 500))
Se2 = 0.25; Seta2 = 0.01; Mu0 = 0.2; S02 = 2.25
v=V=K=sigma.f=rep(NA, 500)
# Kalman filter forward recursion
for (i in 1:500){
  v[i] = y[i] - s[i]
  V[i] = sigma[i] + Se2
```

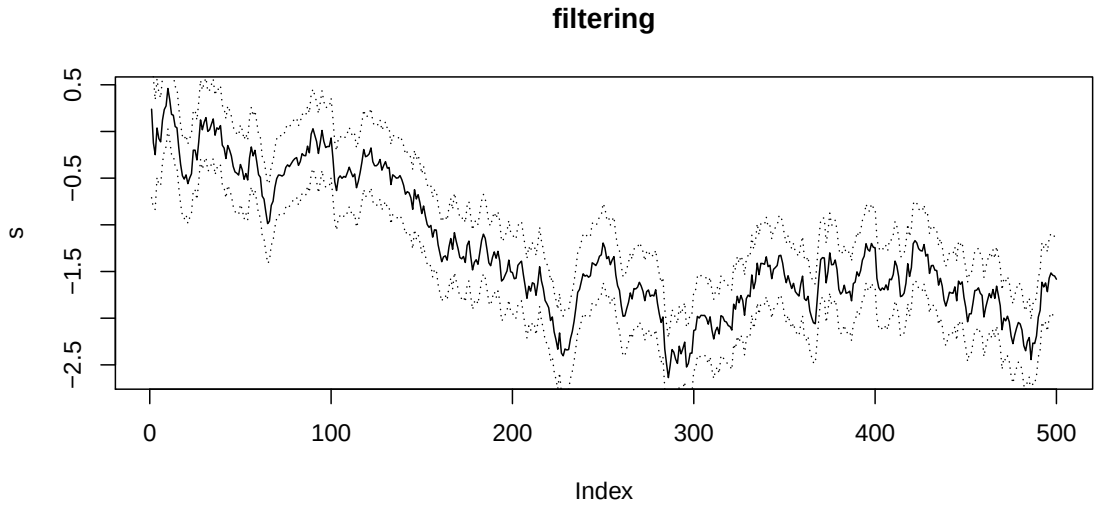
```
  K[i] = sigma[i]/V[i]
  s[i+1] = s[i] + K[i]*v[i]
  sigma.f[i] = (1 - K[i])*sigma[i]
  sigma[i+1] = sigma.f[i] +Seta2
}
s <- s[-1]          # get rid of the initial values
sigma <- sigma[-1]
# prediction
plot(s, type = 'l', main = 'prediction');
lines(s - 2*sqrt(sigma), lty = 3);
lines(s + 2*sqrt(sigma), lty = 3)
```



(c)

Filtered state variables $\{s_{t|t}\}$ and 95% confidence intervals:

```
# filtering
plot(s, type = 'l', main = 'filtering');
lines(s - 2*sqrt(sigma.f), lty = 3);
lines(s + 2*sqrt(sigma.f), lty = 3)
```



(d)

Smoothed state variables $\{s_{t|T}\}$ and 95% confidence intervals:

```
# smoothing
q = c(rep(NA, 500), 0)
M = c(rep(NA, 500), 0)
ss = ssigma = rep(NA, 500)
# backward recursion
for (i in 500:1){
  q[i] = v[i]/V[i] + (1 - K[i])*q[i+1]
  ss[i] = s[i] + sigma[i]*q[i+1]
  M[i] = 1/V[i] + (1 - K[i])^2*M[i+1]
  ssigma[i] = sigma[i] - sigma[i]^2*M[i]
}
plot(ss, type = 'l', main = 'smoothing');
lines(ss - 2*sqrt(ssigma), lty = 3);
lines(ss + 2*sqrt(ssigma), lty = 3)
```