Homework-3_Mostafa_5061

Mostafa

2/17/2021

1.

The data file winnebago in the TSA package contains monthly unit sales of recreational vehicles from Winnebago, Inc. from November 1966 through February 1972.

```
rm(list = ls())

library (TSA) # Load the TSA package
```

```
## Warning: package 'TSA' was built under R version 4.3.1

##
## Attaching package: 'TSA'

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

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

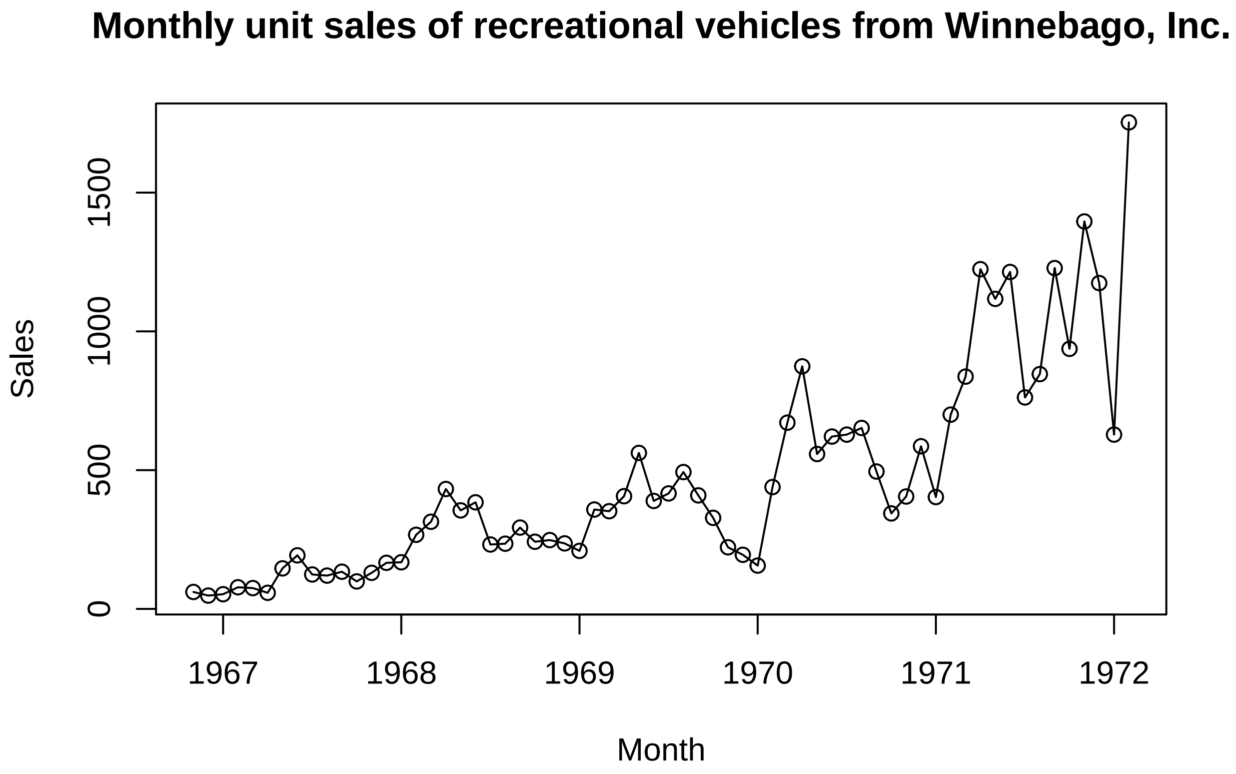
```
data(winnebago)
ts_data = winnebago
str(ts_data)
```

```
##  Time-Series [1:64] from 1967 to 1972: 61 48 53 78 75 58 146 193 124 120 ...
```

1(a)

Display and interpret the time series plot for this data.
Answer: There is an increasing monthly trend in the TS data. Also we can interpret that as times goes, this increasing trend is increasing too. I mean maybe a non-linear model may interpret this data better than a linear one. Of course, my interpretation is just based on this plot, after more investigation (for example in part c), the interpretation may change slightly.

```
plot(ts_data, ylab="Sales",xlab="Month",type="o"
            ,main="Monthly unit sales of recreational vehicles from Winnebago, Inc.")
```



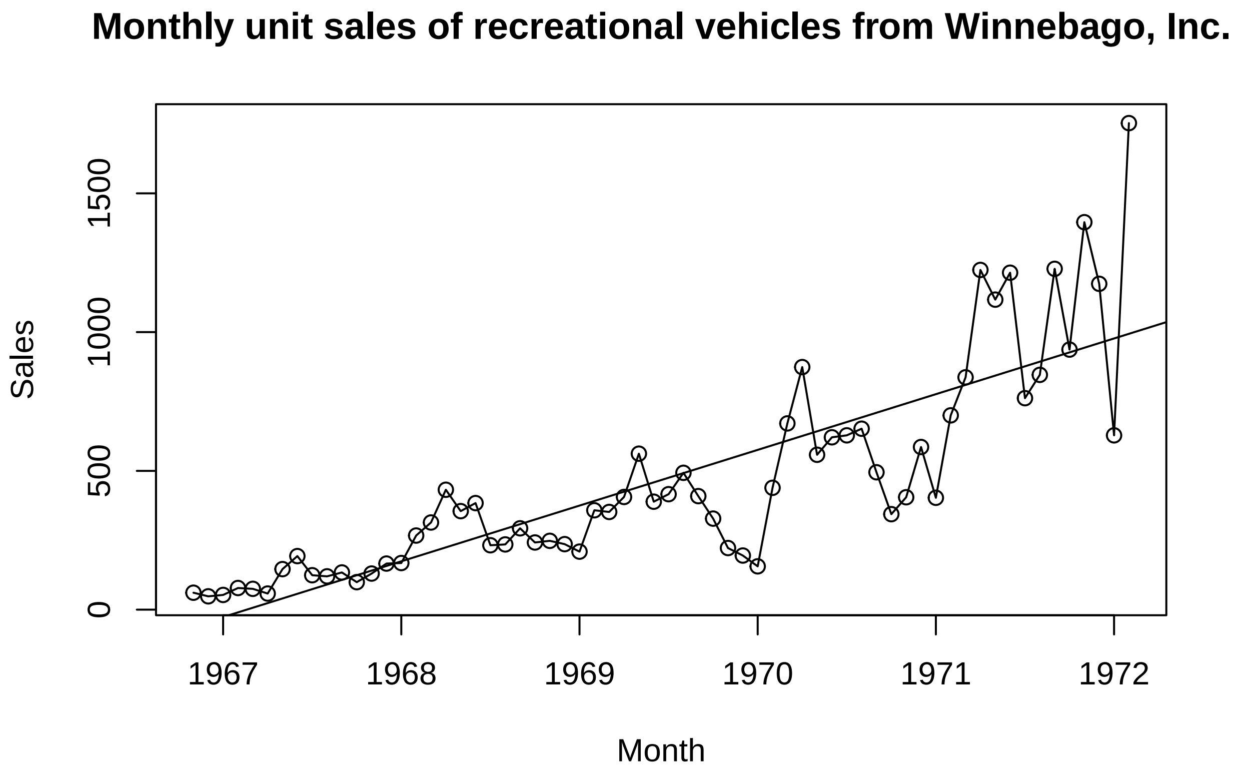
1(b)

Use least squares to fit a linear time trend to this time series and write down your fitted model.
Answer: $Sales = -394885.68 + 200.74 * time(ts\_data)$

```
# fitting using simple linear regression
fit <- lm(ts_data~time(ts_data))
summary(fit)
```

```
##
## Call:
## lm(formula = ts_data ~ time(ts_data))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -419.58  -93.13  -12.78   94.96  759.21
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -394885.68   33539.77  -11.77   <2e-16 ***
## time(ts_data)     200.74      17.03   11.79   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 209.7 on 62 degrees of freedom
## Multiple R-squared:  0.6915, Adjusted R-squared:  0.6865
## F-statistic: 138.9 on 1 and 62 DF,  p-value: < 2.2e-16
```

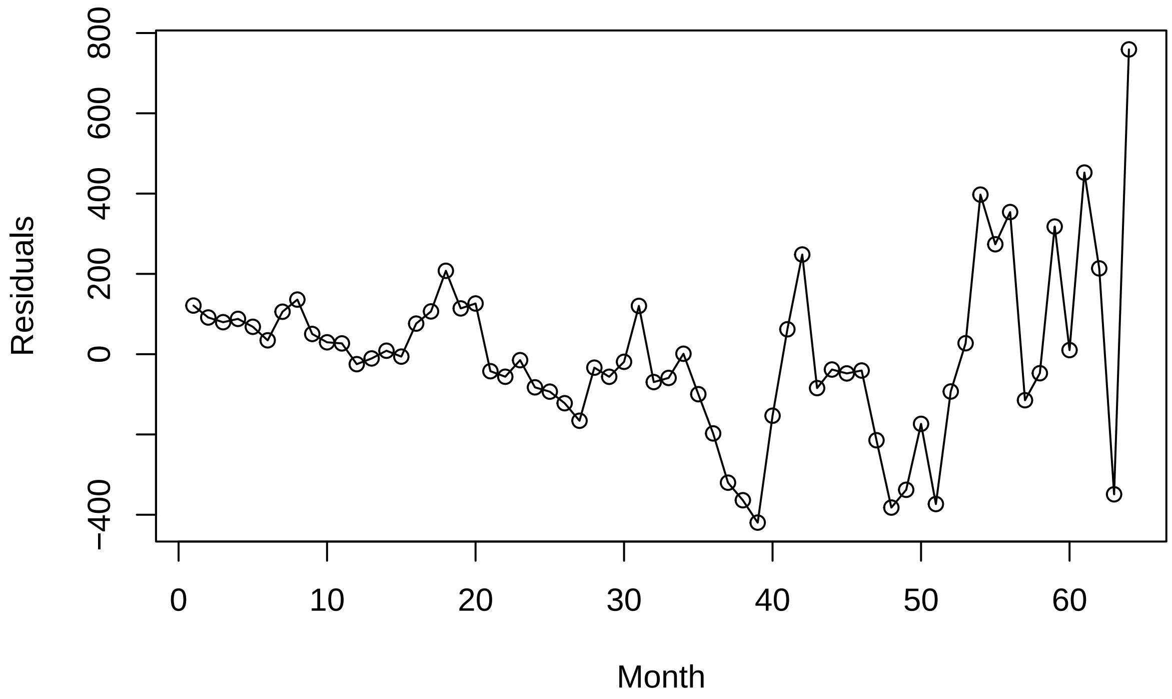
```
plot(ts_data, ylab="Sales",xlab="Month",type="o"
            ,main="Monthly unit sales of recreational vehicles from Winnebago, Inc.")
abline(fit)
```



1(c)

Construct and interpret the time series plot for the residuals obtained from part (b).
Answer: The residual plot helped me to figure out that Although there is an uptrend, there is also a fluctuation around the fitted line. And as times goes, in general, the residuals become bigger and bigger. Of course again, this interpretation is based on the these plots and it may slightly change in part d.

```
# Residuals from straight line model fit
fit <- lm(ts_data~time(ts_data))
plot(resid(fit),ylab="Residuals",xlab="Month",type="o")
```

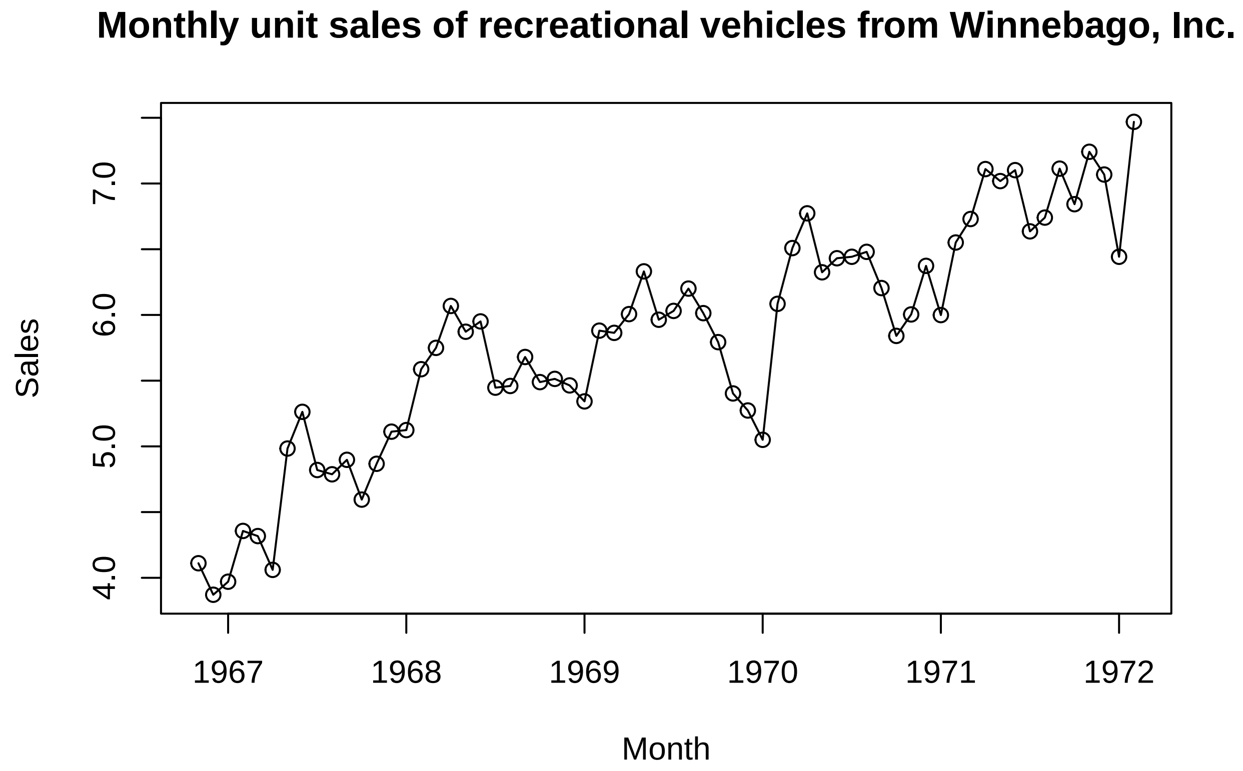


1(d)

Now take natural logarithms of the monthly sales figures and display and interpret the time series plot of the transformed values.

Answer: Now, I have a better idea about the data. They have a logarithmic nature. Because the logarithmic transformed data has a uptrend and explains why as time goes, the residuals become biger and biger.

```
log_ts_data = log(ts_data)
#str(log_ts_data)
plot(log_ts_data, ylab="Sales",xlab="Month",type="o"
            ,main="Monthly unit sales of recreational vehicles from Winnebago, Inc.")
```



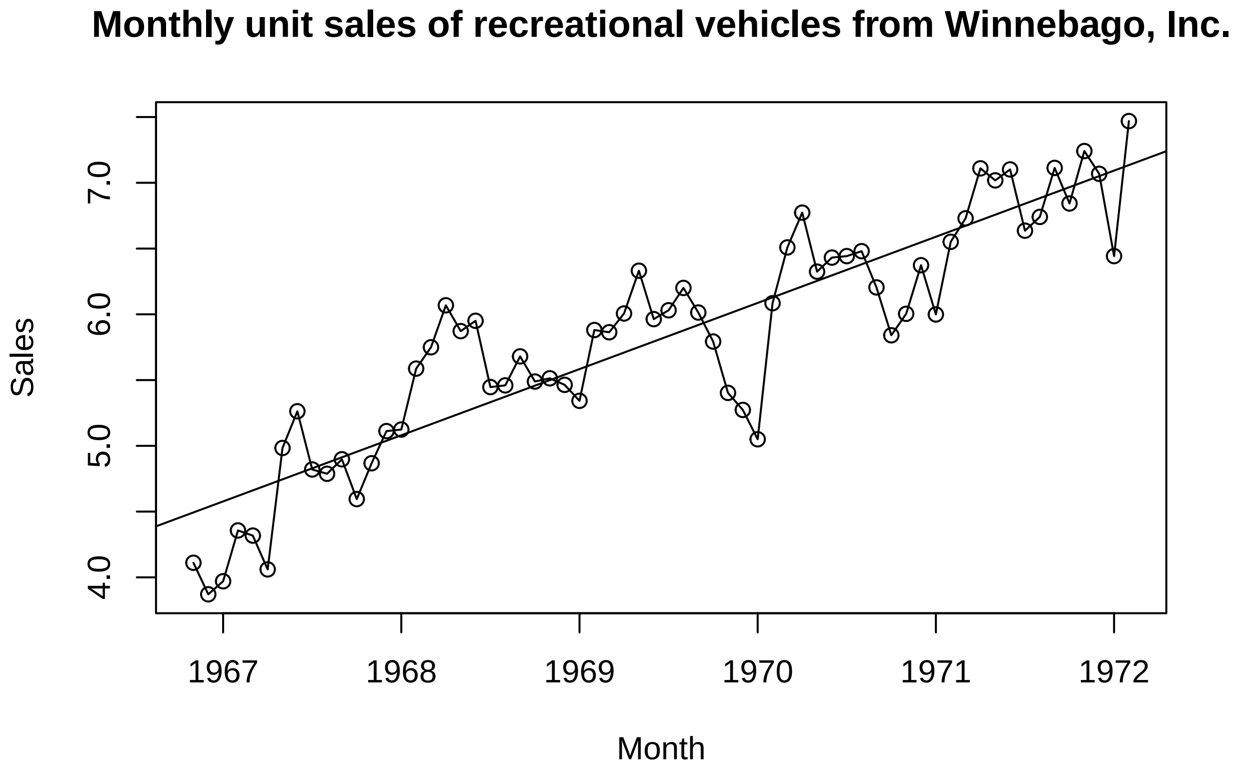
1(e)

Use least squares to fit a line to the logged data and write down your fitted model.
Answer: $log(Sales) = -984.93878 + 0.50306 * time(ts\_data)$

```
# fitting using simple linear regression
fit <- lm(log_ts_data~time(ts_data))
summary(fit)
```

```
##
## Call:
## lm(formula = log_ts_data ~ time(ts_data))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.03669 -0.20823  0.04995  0.25662  0.86223
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -984.93878   62.99472  -15.63   <2e-16 ***
## time(ts_data)    0.50306    0.03199   15.73   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3939 on 62 degrees of freedom
## Multiple R-squared:  0.7996, Adjusted R-squared:  0.7964
## F-statistic: 247.4 on 1 and 62 DF,  p-value: < 2.2e-16
```

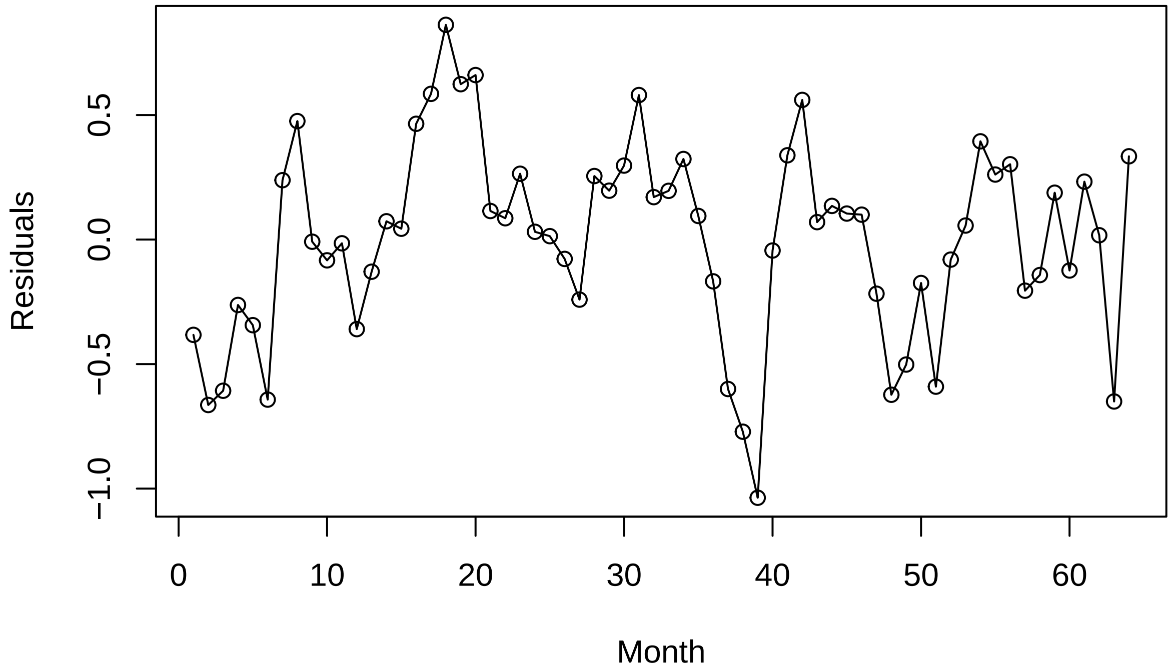
```
plot(log_ts_data, ylab="Sales",xlab="Month",type="o"
            ,main="Monthly unit sales of recreational vehicles from Winnebago, Inc.")
abline(fit)
```



1(f)

Construct and interpret the time series plot for residuals from part (e).
Answer: The points evenly distributed around zero. So, we can interpret that the sales data has a logarithmic nature and a logarithmic transform can explain the behavior of this data.

```
# Residuals from straight line model fit
fit <- lm(log_ts_data~time(ts_data))
plot(resid(fit),ylab="Residuals",xlab="Month",type="o")
```



2.

Note: In problem 2, although the data do show some seasonality as well as linear decreasing pattern in general, you don't need to consider seasonality for this problem. I have added the specific form (linear trend) for the model in problem 2 a).

Tuberculosis, commonly known as TB, is a bacterial infection that can spread through the lymph nodes and bloodstream to any organ in your body (it is most often found in the lungs). Most people who are exposed to TB never develop symptoms, because the bacteria can live in an inactive form in the body. But if the immune system weakens, such as in people with HIV or in elderly adults, TB bacteria can become active and fatal if untreated. The numbers of TB cases (per month) in the United States from January 2000 to December 2009 are catalogued in the data file "tb".

Hint: You may use the following codes to load the data.

```
library(TSA)
##
## Attaching package: 'TSA'
## The following objects are masked from 'package:stats':
##
## acf, arima
## The following object is masked from 'package:utils':
##
## tar
tb = ts(read.table(file = "tb.txt"), freq=12, start=c(2000,1))
tb = ts(read.table(file = "tb.txt"), freq=12, start=c(2000,1))
ts_data = tb
ts_data
```

```
##       Jan  Feb  Mar  Apr  May  Jun  Jul  Aug  Sep  Oct  Nov  Dec
## 2000 1100 1324 1482 1259 1412 1406 1218 1335 1176 1141  941  535
## 2001 1106 1159 1376 1353 1447 1324 1153 1359 1100 1159  900  553
## 2002 1053 1129 1318 1353 1329 1253 1206 1182 1047 1018  812  518
## 2003 1106 1065 1241 1306 1306 1188 1206 1129 1065 1094  818  529
## 2004 1059 1082 1335 1212 1229 1241 1182 1182 1018  994  788  565
## 2005 1059 1024 1259 1171 1265 1235 1065 1106 1000  965  794  524
## 2006 1018 1041 1200 1153 1218 1118 1024 1171  935  941  776  547
## 2007  971  965 1165 1141 1124 1176 1088 1112  959  853  788  500
## 2008 1000  971 1071 1159 1106 1041 1129  994  929  929  700  541
## 2009  776  888 1006 1035  976 1012  965  906  818  771  688  535
```

2(a)

Use the methods in Chapter 3 to fit a linear trend model of the form $Y = b_0 + b_1 t + X_t$ where $E(X_t) = 0$.

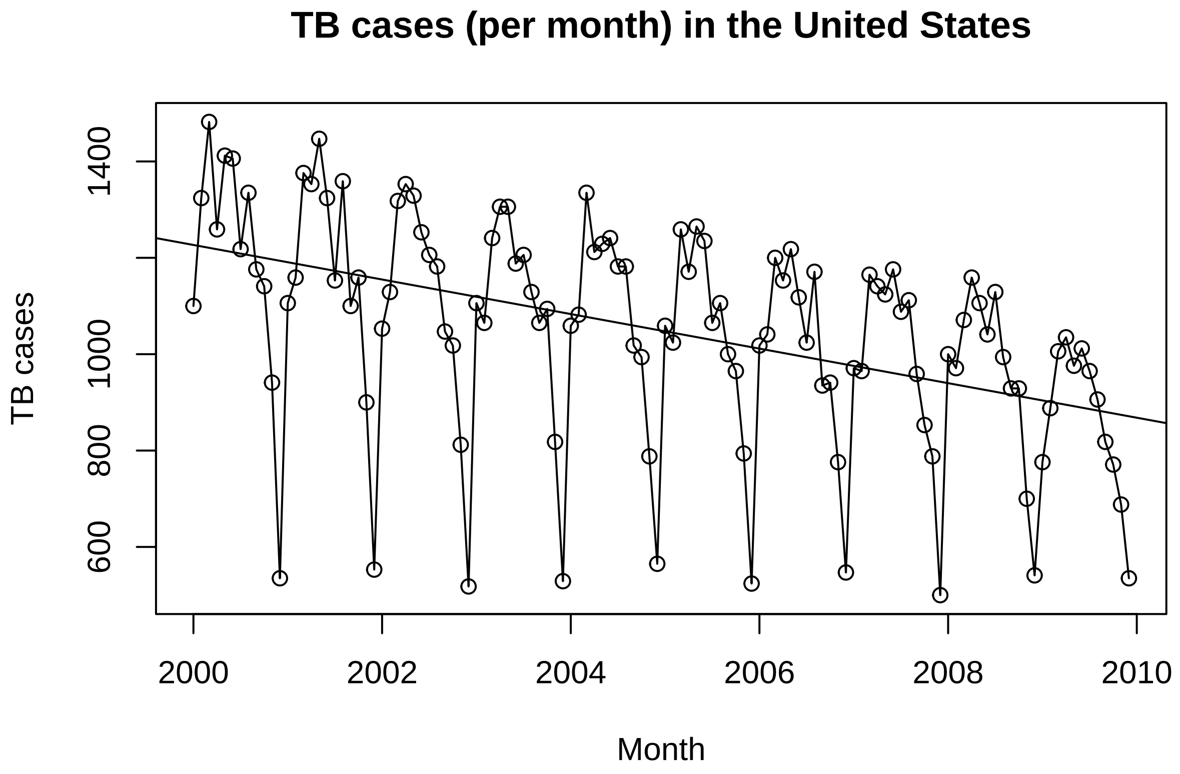
```
# fitting using simple linear regression
fit <- lm(ts_data~time(ts_data))
summary(fit)
```

```
##
## Call:
## lm(formula = ts_data ~ time(ts_data))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -658.84  -62.21   30.55  148.66  268.10
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    72929.27   12770.83   5.711 8.56e-08 ***
## time(ts_data)    -35.85       6.37  -5.628 1.25e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 201.4 on 118 degrees of freedom
## Multiple R-squared:  0.2117, Adjusted R-squared:  0.205
## F-statistic: 31.68 on 1 and 118 DF,  p-value: 1.248e-07
```

2(b)

Produce a plot that displays the time series data with your fitted line.

```
plot(ts_data, ylab="TB cases",xlab="Month",type="o"
            ,main="TB cases (per month) in the United States")
abline(fit)
```



(c)

Examine the standardized residuals $\hat{X}_t^*$ from your fitted model for normality and independence. What are your conclusions? Do the standardized residuals look to resemble a normal, zero mean white noise process?
Answer:
Histogram:
The histogram seems left-skewed. So, based on that I can say it is a left-skewed.

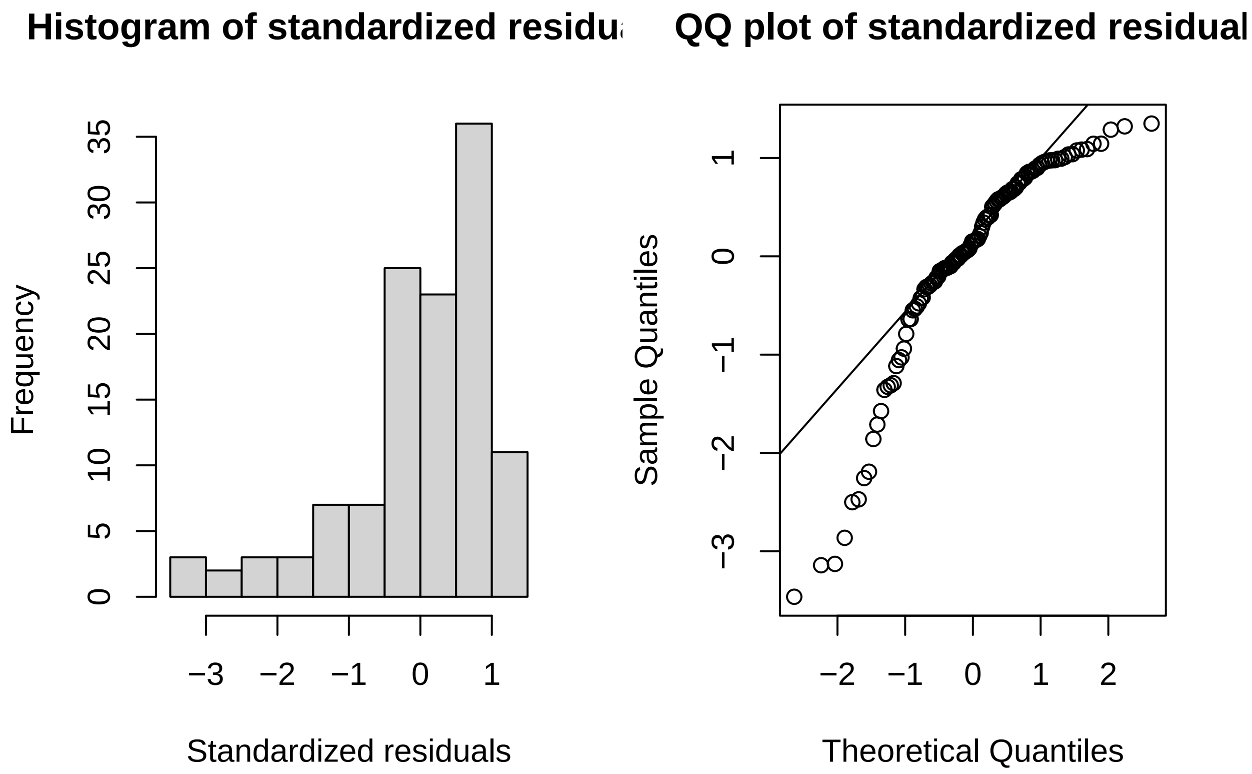
QQ plot:
As most of the data have been centered and they are on the reference line. However, the tails somehow do not follow reference line. And the tails' points are below the reference line. So, it is a left-skewed.

Residuals plot:
Although most of the points are above the reference line, those who are below the reference line have bigger residuals. And I would say they almost evenly distributed around the reference line. However, based on this plot I can not make sure about the normality.

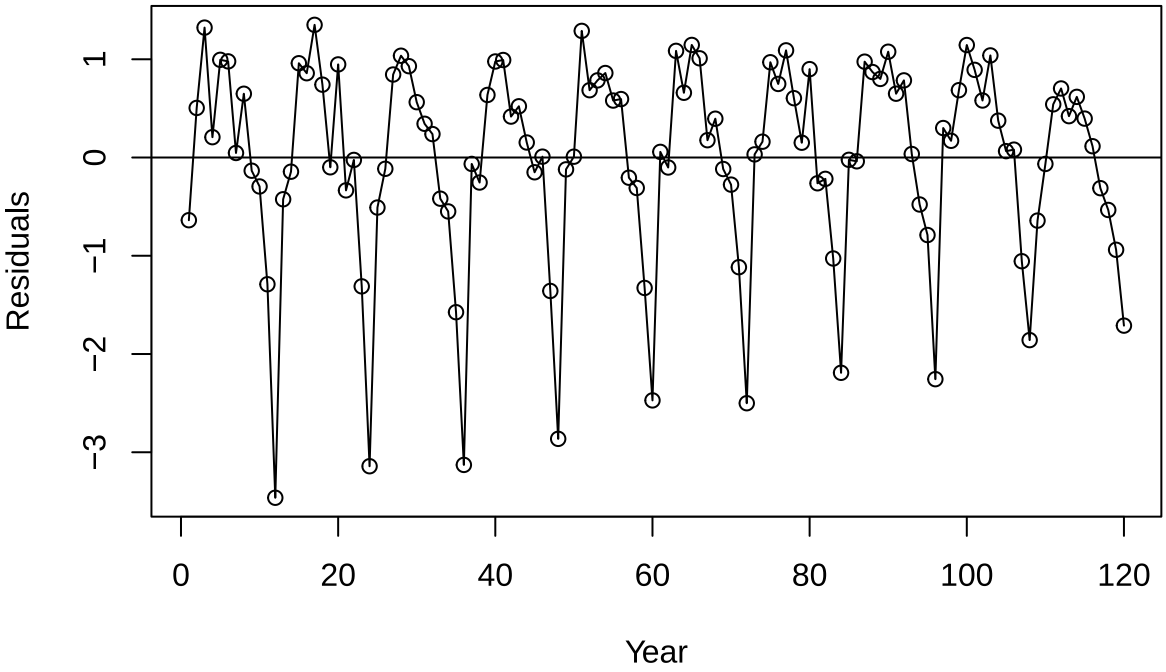
Test:
"The p-value for the test is extremely small, so we would reject H0. The evidence points to the standardized residuals being not independent. The R output also produces the expected number of runs (computed under the assumption of independence). The observed number of runs is too much lower than the expected number to support independence." Lecture note, page 24.

```
# Standardized residuals from straight line model fit using a Histogram and a Q-Q plot
par(mfrow=c(1,2))
hist(rstudent(fit),  main="Histogram of standardized residuals",xlab="Standardized residuals")
```

```
qqnorm(rstudent(fit),main="QQ plot of standardized residuals")
qqline(rstudent(fit),main="QQ plot of standardized residuals")
```



```
### Part 2: Assessing Independence
# Standardized residuals from straight line model fit Horizontal line added at 0
plot(rstudent(fit),ylab="Residuals",xlab="Year",type="o")
abline(h=0)
```



```
# Runs test for independence on standardized residuals
runs(rstudent(fit))
```

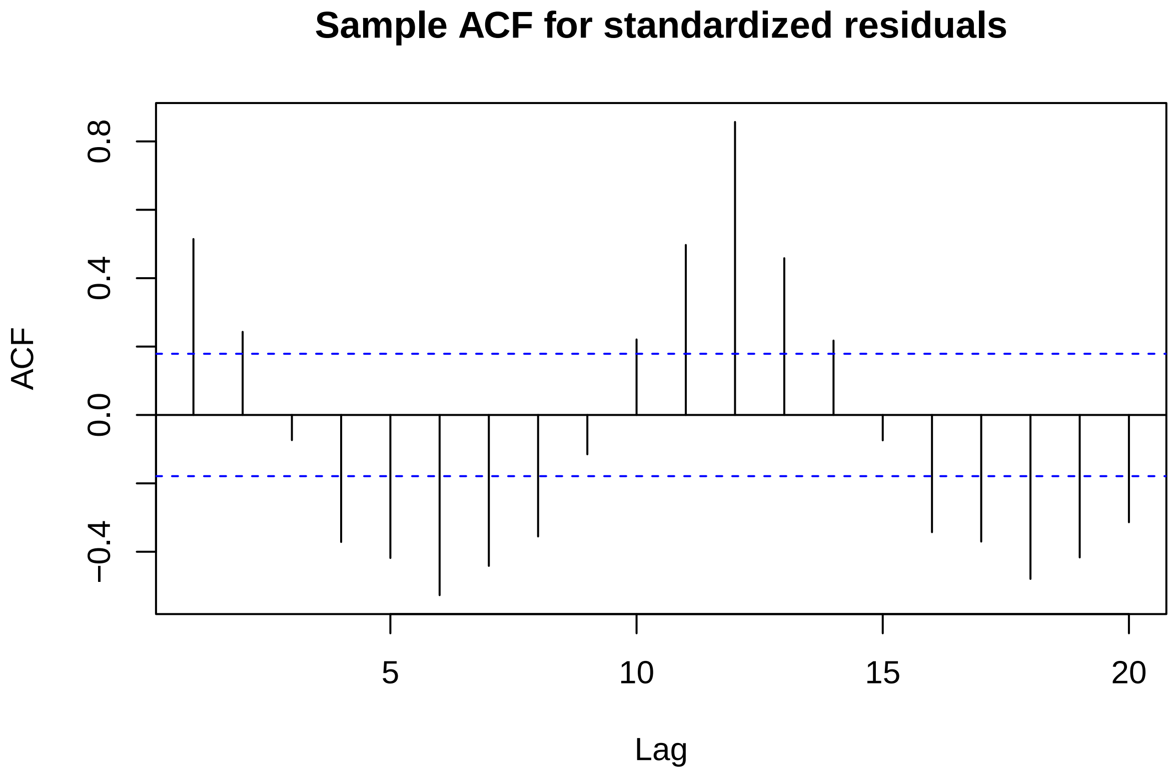
```
## $pvalue
## [1] 8.79e-10
##
## $observed.runs
## [1] 27
##
## $expected.runs
## [1] 59.33333
##
## $n1
## [1] 50
##
## $n2
## [1] 70
##
## $k
## [1] 0
```

(d)

Display the sample ACF for the standardized residual in part (c). What's your conclusion?
Answer: As there are some values that have exceeded the blue dash lines, I can interpret that there is a

clear trend in the data. Also, at least one of the lags is much bigger (>0.6) than the blue dash line and I could consider "one" as too many. And it is not a white noise.

```
### Part 3: Sample autocorrelation function
# Calculate sample ACF for standardized residuals
acf(rstudent(fit),main="Sample ACF for standardized residuals")
```


Sample ACF for standardized residuals
ACF
Lag