panders2_stat626_hw03

Philip Anderson; panders2@tamu.edu

6/6/2018

```r
library("tidyverse")
library("gmodels") # for full contingency table
library("astsa")
```

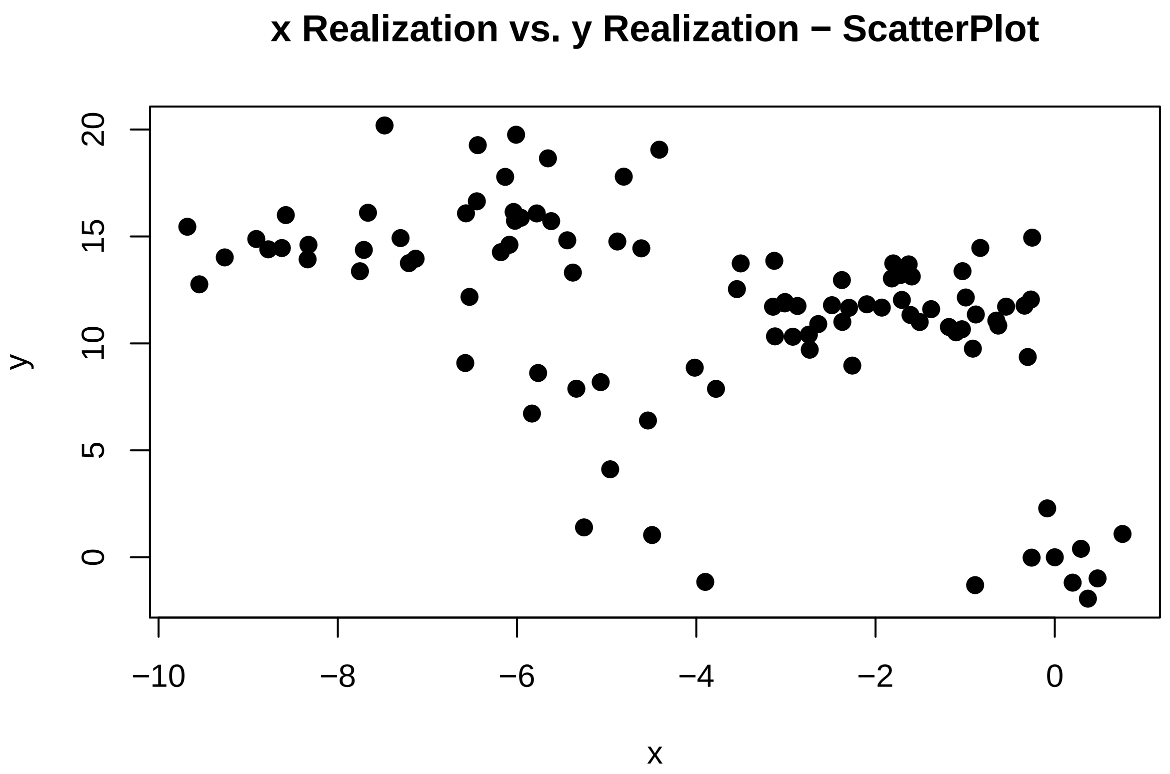
Question I

Question I) a)

```r
# simulate two random walks with initial values equal to zero
# rw1
set.seed(7)
y <- cumsum(c(0, rnorm(99, 0, 1)))
# rw2
set.seed(8)
x <- cumsum(c(0, rnorm(99, 0, 1)))
```

Question I) a) i)

```r
# plot y_t vs x_t in scatterplot form
plot(x, y, pch=16, cex=1.25
     , main="x Realization vs. y Realization - ScatterPlot"
     )
```



From the above, it appears that there is a non-linear negative relationship between y and x. The correlation coefficient (below) corroborates this.

```
# nonparametric correlation coefficient
cor(x, y, method="spearman")
```

```
## [1] -0.5853465
```

Question I) a) ii)

I expect the hypothesis to be rejected based on the data realizations I have, but probably not if repeated numerous times. There is a clear (to me) negative relationship between the variables, and we have a non-trivial amount of data.

Question I) a) iii)

```
lin_mod <- lm(y ~ x)
summary(lin_mod)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.547  -1.591   1.302   2.731   7.183
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   7.7748     0.7483   10.39  < 2e-16 ***
## x            -0.9287     0.1574   -5.90 5.21e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.414 on 98 degrees of freedom
## Multiple R-squared:  0.2621, Adjusted R-squared:  0.2546
## F-statistic: 34.81 on 1 and 98 DF,  p-value: 5.209e-08
```

From the summary printout, it is clear the beta_1 coefficient is negative, and we are soundly rejecting the null hypothesis that there is no relationship between x and y. Because of the random nature of the random walk, I would not necessarily expect this relationship to hold under repeated trials.

Question I) b)

```
# repeat the experiment 1000 times

beta_val_vec <- numeric(1000)
p_val_vec <- numeric(1000)

for (i in 1:1000) {

      set.seed(10 + i)
      y <- cumsum(c(0, rnorm(99, 0, 1)))
      set.seed(5000 + i)
      x <- cumsum(c(0, rnorm(99, 0, 1)))

      lin <- lm(y ~ x)
      lin_summ <- summary(lin)
      beta_val <- lin_summ$coefficients[2,1]
      p_val <- lin_summ$coefficients[2,4]
      beta_val_vec[i] <- beta_val
      p_val_vec[i] <- p_val

}

outcome <- data.frame(cbind(beta_val_vec, p_val_vec))
```

```
# label experiment outcomes.  Positive/Negative, Significant/Non-Significant
outcome$beta_pos <- ifelse(outcome$beta_val_vec >= 0, "pos", "neg")
outcome$signif <- ifelse(outcome$p_val_vec < 0.05, "sig", "not_sig")
```

```
# generate a full-featured contigency table
gmodels::CrossTable(outcome$beta_pos, outcome$signif)
```

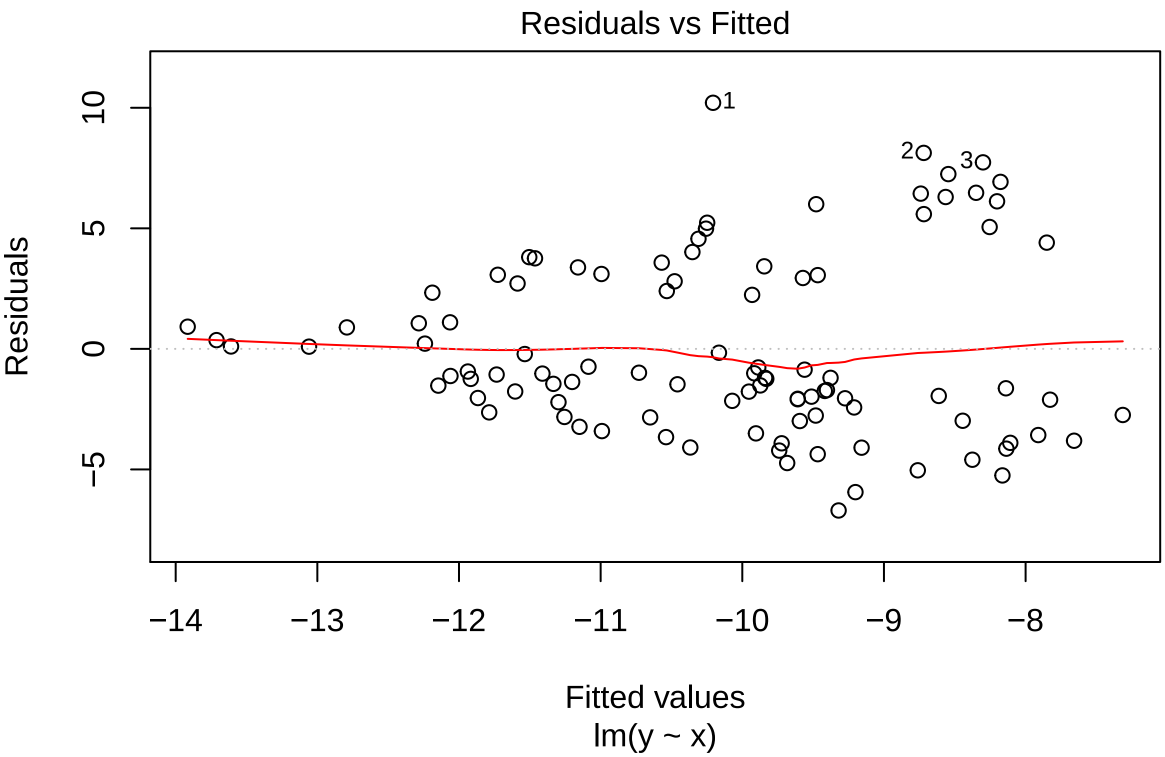
```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  1000
##
##
##                  | outcome$signif
## outcome$beta_pos |   not_sig |       sig | Row Total |
## -----------------|-----------|-----------|-----------|
##              neg |       121 |       388 |       509 |
##                  |     0.177 |     0.058 |           |
##                  |     0.238 |     0.762 |     0.509 |
##                  |     0.490 |     0.515 |           |
##                  |     0.121 |     0.388 |           |
## -----------------|-----------|-----------|-----------|
##              pos |       126 |       365 |       491 |
##                  |     0.184 |     0.060 |           |
##                  |     0.257 |     0.743 |     0.491 |
##                  |     0.510 |     0.485 |           |
##                  |     0.126 |     0.365 |           |
## -----------------|-----------|-----------|-----------|
##     Column Total |       247 |       753 |      1000 |
##                  |     0.247 |     0.753 |           |
## -----------------|-----------|-----------|-----------|
##
##
```

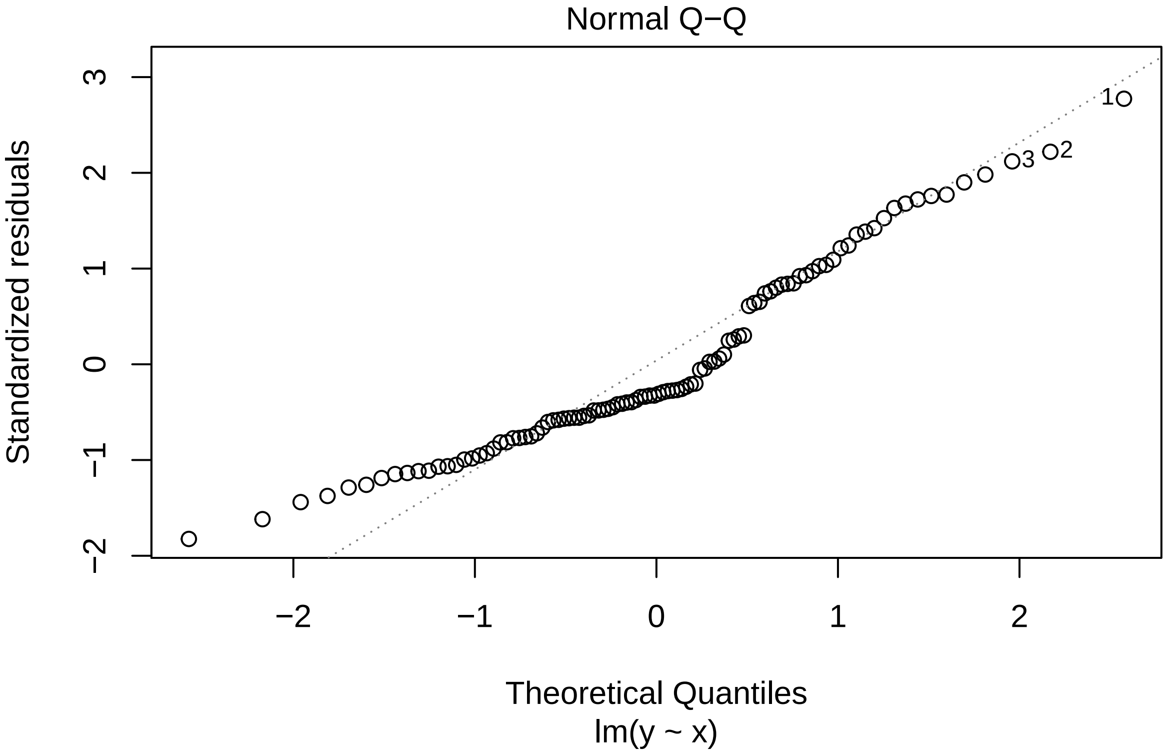
From the above cross-tabulation, my hypothesis that repeated sampling will show that the relationship between the variables is not always negative appears to hold. I am, however, surprised that so many of the *beta* coefficients are significant. Given the randomness of the value generation, I would have expected around 5% of these to be significant, not over 75%.

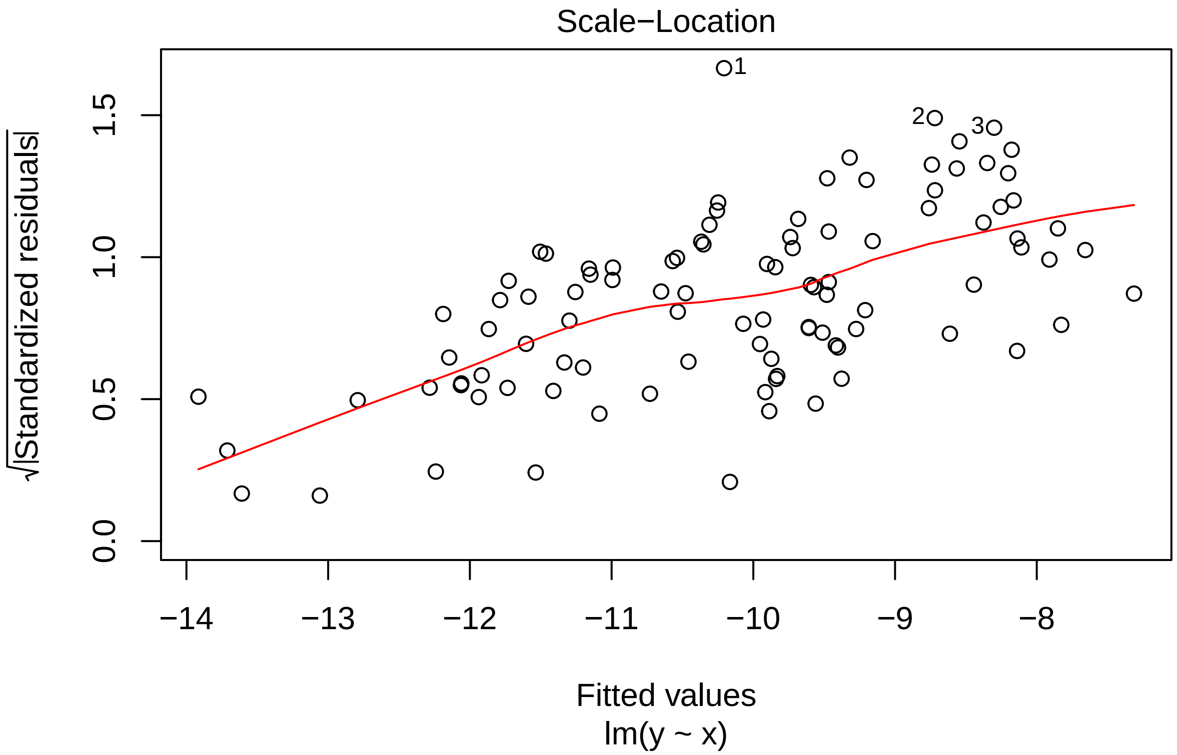
I will check the regression assumptions for one of the simulations.

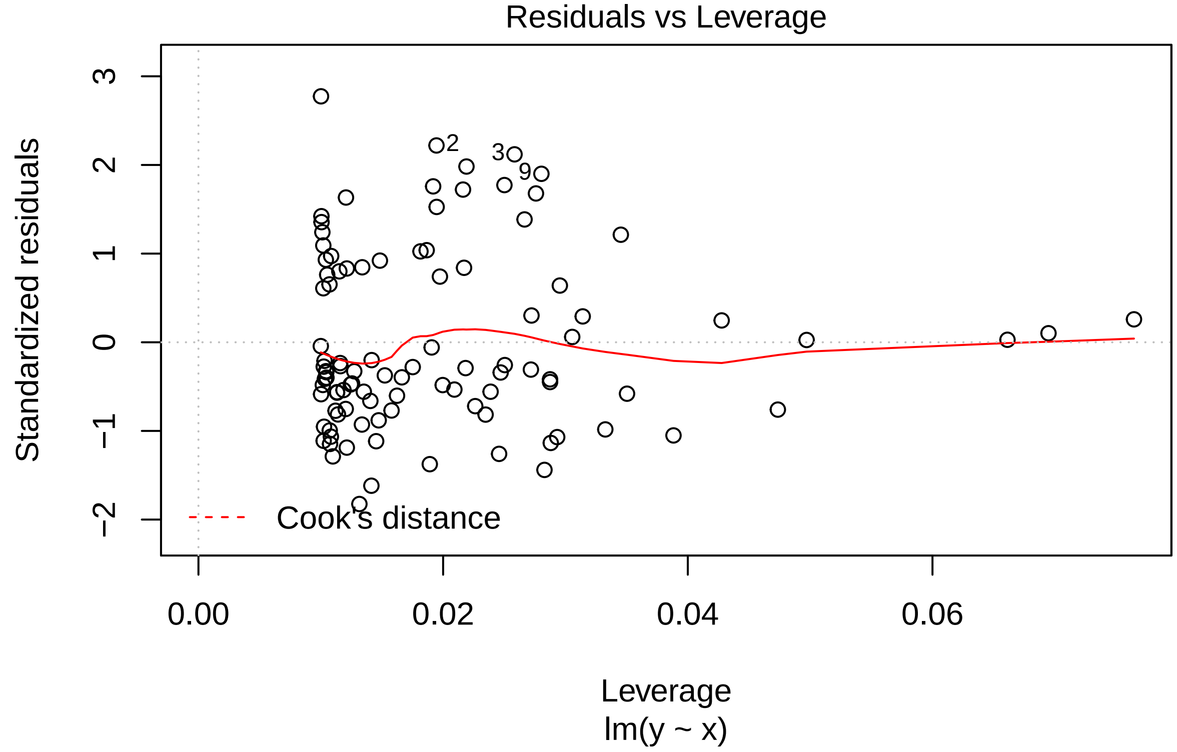
```r
i <- 1
set.seed(10 + i)
y <- cumsum(c(0, rnorm(99, 0, 1)))
set.seed(5000 + i)
x <- cumsum(c(0, rnorm(99, 0, 1)))

lin <- lm(y ~ x)
lin_summ <- summary(lin)
plot(lin)
```


Residuals vs Fitted
Residuals
10
5
0
−5
1
2
3
−14
−13
−12
−11
−10
−9
−8
Fitted values
lm(y ~ x)


Normal Q−Q
Standardized residuals
3
2
1
0
−1
−2
1
3
2
−2
−1
0
1
2
Theoretical Quantiles
lm(y ~ x)





From the above diagnostic plots, it appears that the regression assumptions are violated. The Residuals vs. Fitted Values plot shows non-constant variance, or heteroscedacticity. The Normal Q-Q plot shows that we have substantial deviation from normality. These types of assumption violations are likely occurring throughout the simulations, and are likely responsible for the surprising number of significant results.

Question 2.6

Question 2.6.A

```r
# split the varve data set into 2
mid_point <- length(varve) / 2
varve1 <- varve[1:mid_point]
varve2 <- varve[(mid_point+1):length(varve)]

# take the sample variance of each
paste0("First-half variance: ", var(varve1))
```

```
## [1] "First-half variance: 133.457415667053"
```

```r
paste0("Second-half variance: ", var(varve2))
```

```
## [1] "Second-half variance: 594.490438823224"
```

From the above, it is clear that the second half of the varve series displays higher variance than the first half.

```r
# take log transformation of each half of the series and check variance
paste0("First-half log-transformed variance: "  , var(log(varve1)))
```

```
## [1] "First-half log-transformed variance: 0.270721652653357"
```

```r
paste0("Second-half log-transformed variance: "  , var(log(varve2)))
```

```
## [1] "Second-half log-transformed variance: 0.451371011716303"
```

```r
# take the ratios of the second half of the varve series to the first half for
# the non-transformed and transformed data
print(594 / 133)
```

```
## [1] 4.466165
```

```r
print(0.45 / 0.27)
```

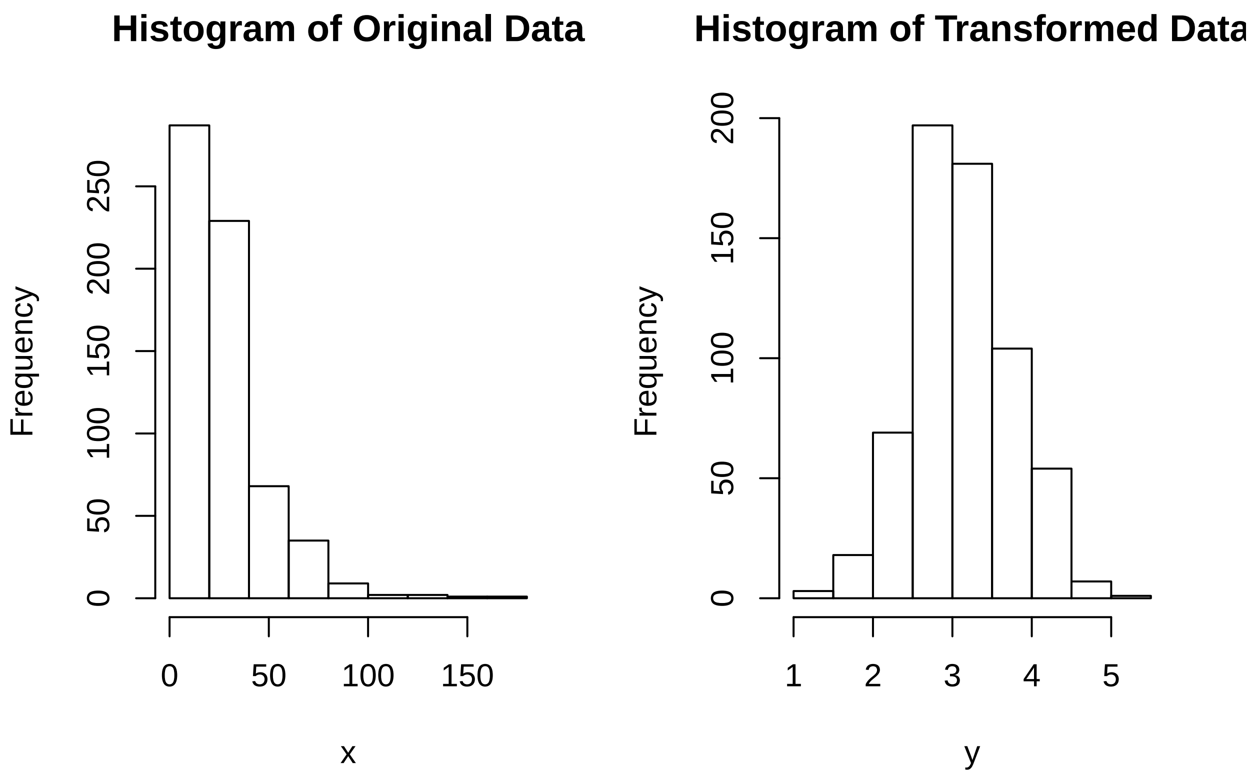
```
## [1] 1.666667
```

```r
# percent change in variance
paste0("Percent change in variance ratio: ", ((594 / 133) - (0.45 / 0.27)) / (594 / 133))
```

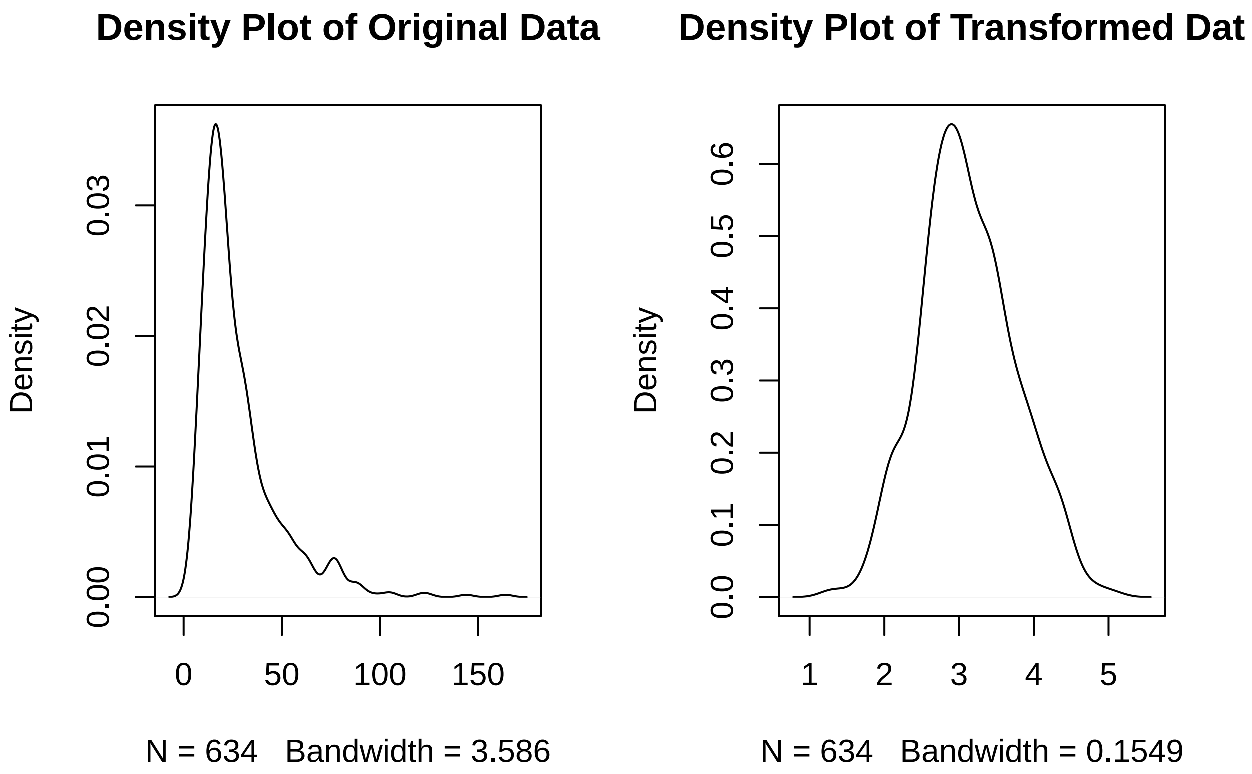
```
## [1] "Percent change in variance ratio: 0.62682379349046"
```

When we examine the ratio of the variance of the second half of the varve time series to the second, we can see that taking the log decreases that ratio by over 60%.

```r
# create density estimates of data.  histogram + density plot
x <- varve
y <- log(varve)
par(mfrow=c(1,2))
hist(x, main="Histogram of Original Data")
hist(y, main="Histogram of Transformed Data")
```

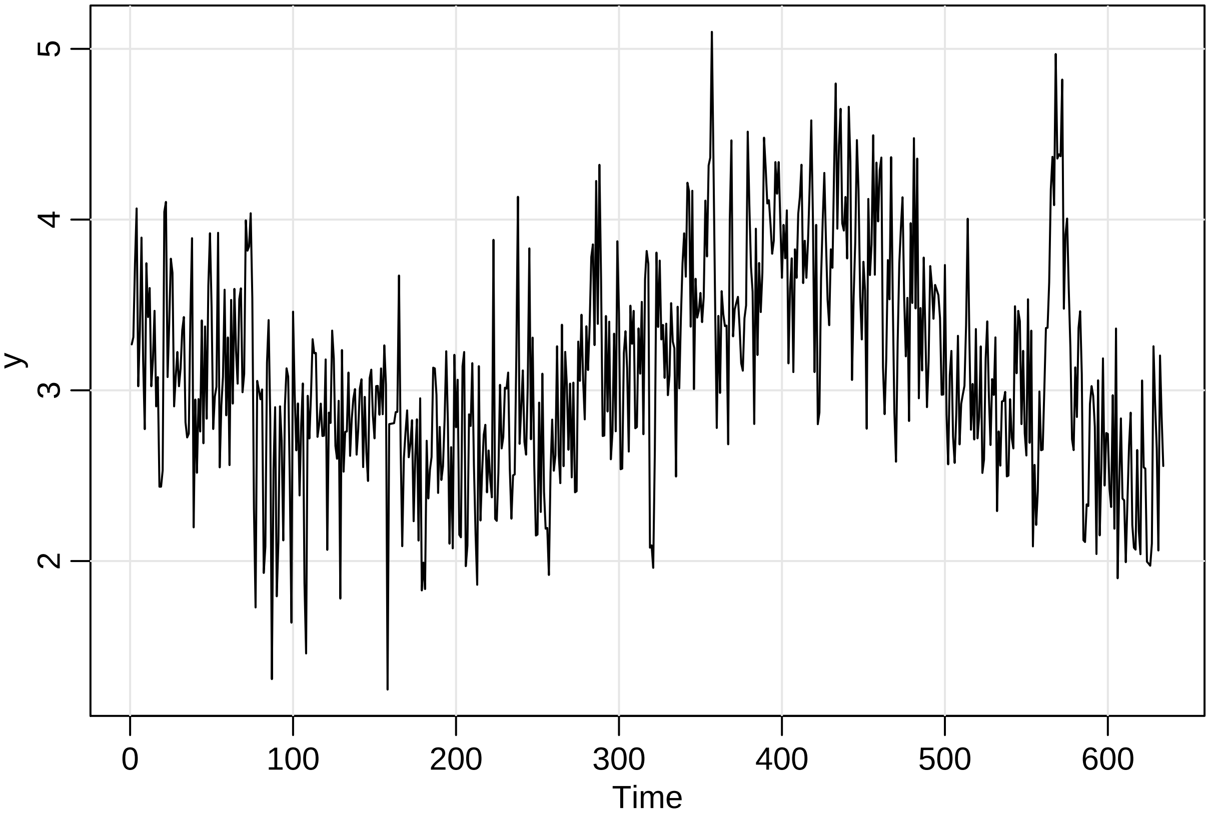


```
plot(density(x), main="Density Plot of Original Data")
plot(density(y), main="Density Plot of Transformed Data")
```



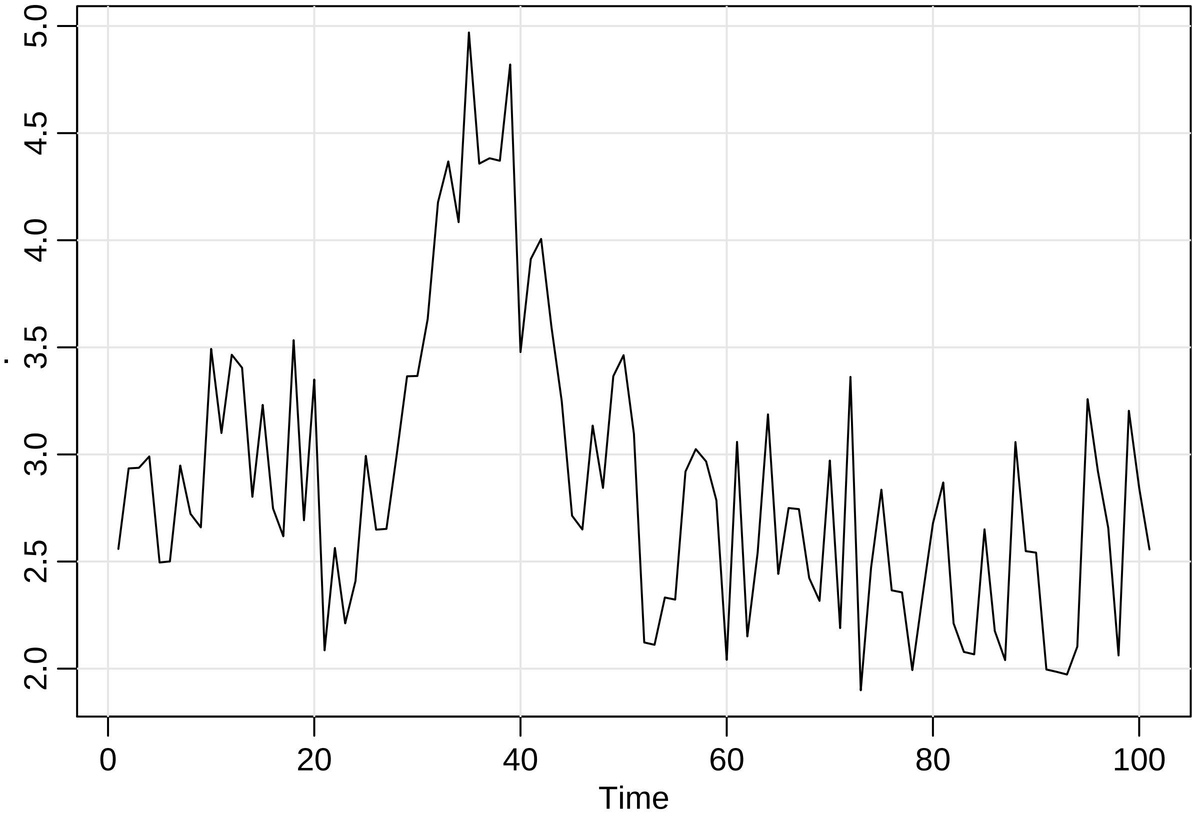
Question 2.6.B

```r
# plot the series y_t
astsa::tsplot(y)
```



The Figure 1.3 shows Global Average Temperature Deviations for approximately the last 100 years. In the above graph, it appears that there is a spike in varve value between 500 and 600 years ago (the most recent 100 years). I limit to the most recent 100 years and re-plot.

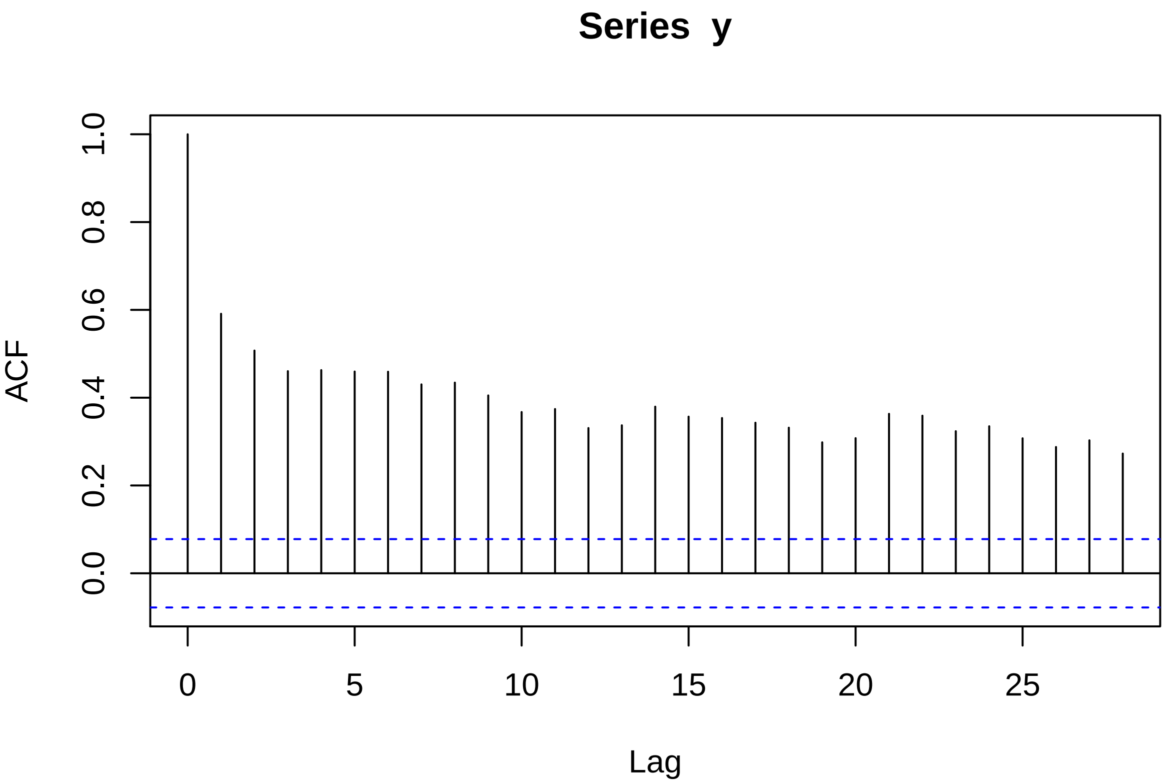
```r
y[(length(y)-100):length(y)] %>%
  astsa::tsplot()
```



The spike between ~25 and ~45 corresponds to the spike in temperature deviation.

Question 2.6.C

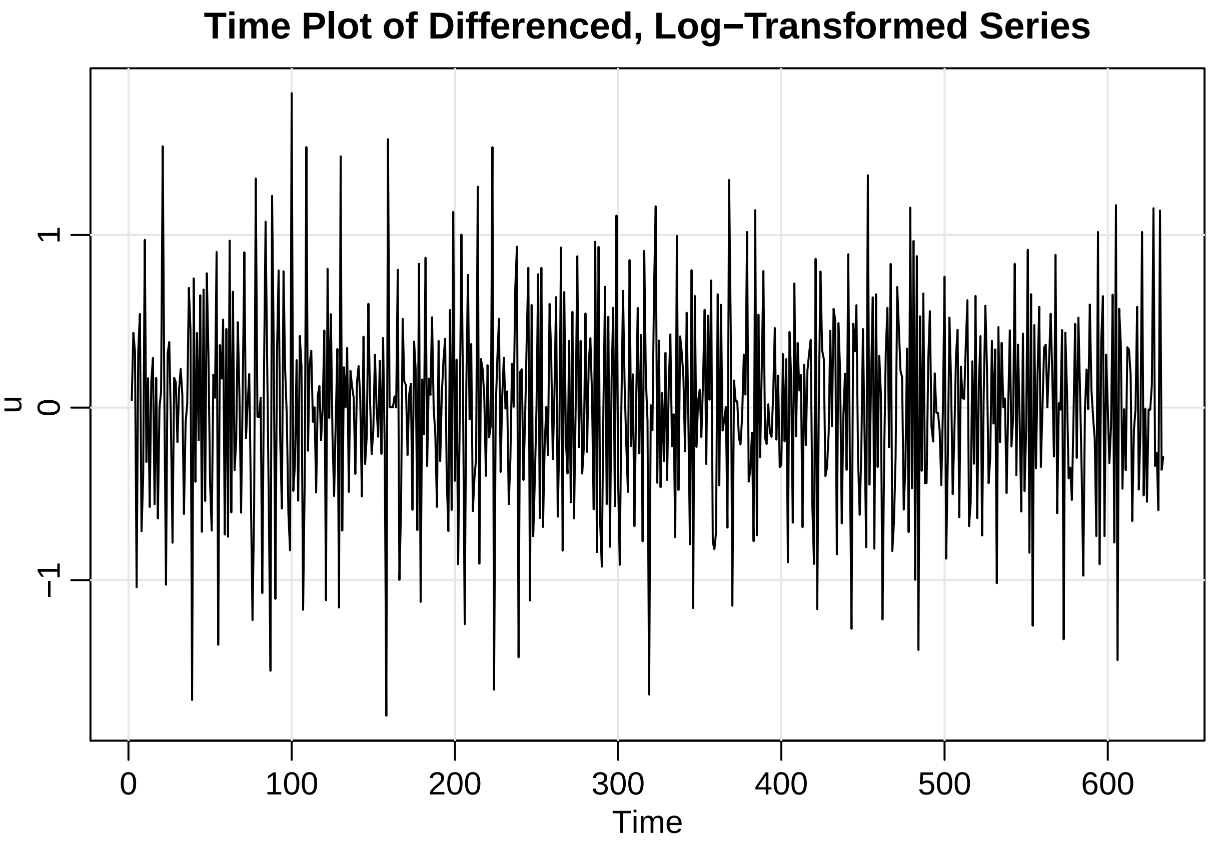
```
# Examine the ACF of y.
acf(y)
```



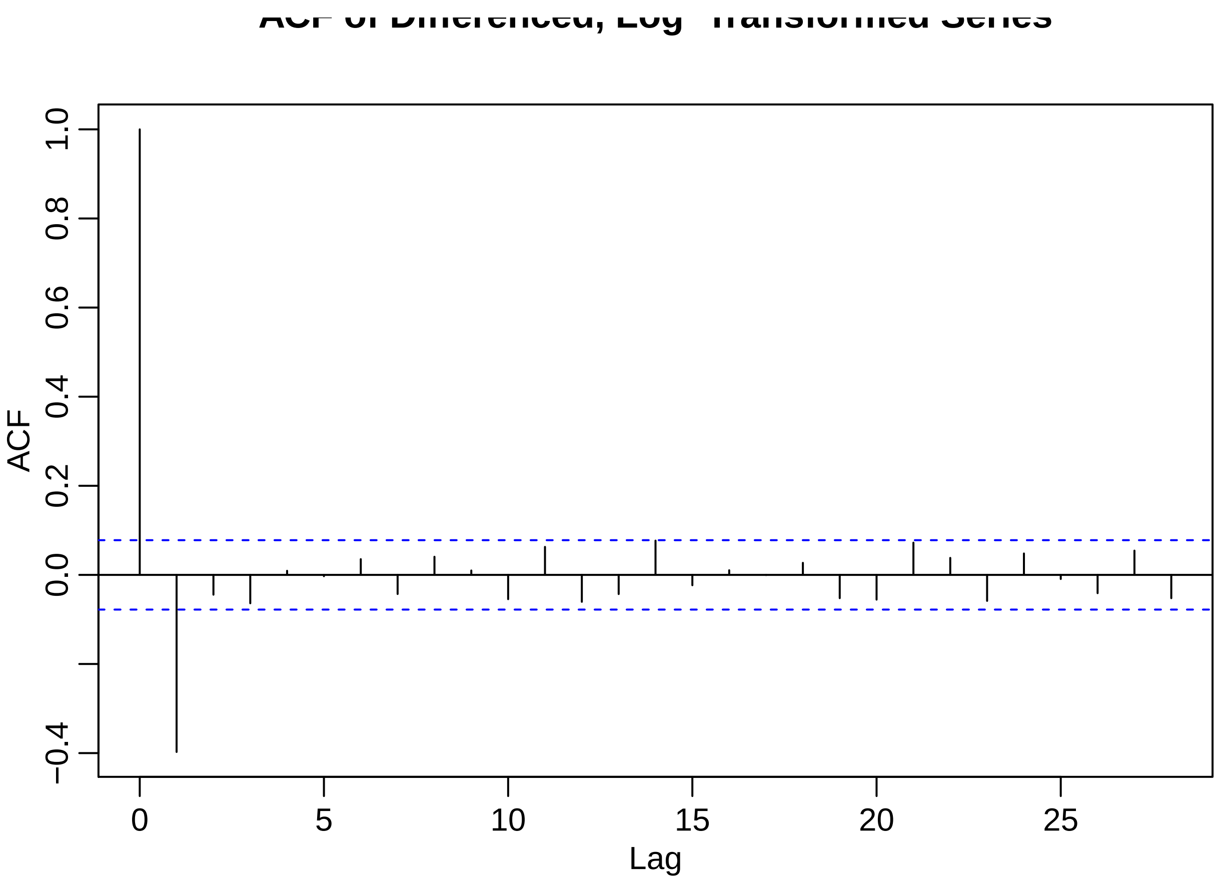
It appears that we have significant lag in the data, that is, the data points that we see are significantly related to those that came before. This is consistent with a random walk.

Question 2.6.D

```
# compute the first difference of the series
u <- diff(y)
# produce its time plot
astsa::tsplot(u, main="Time Plot of Differenced, Log-Transformed Series")
```



```
acf(u, main="ACF of Differenced, Log-Transformed Series")
```

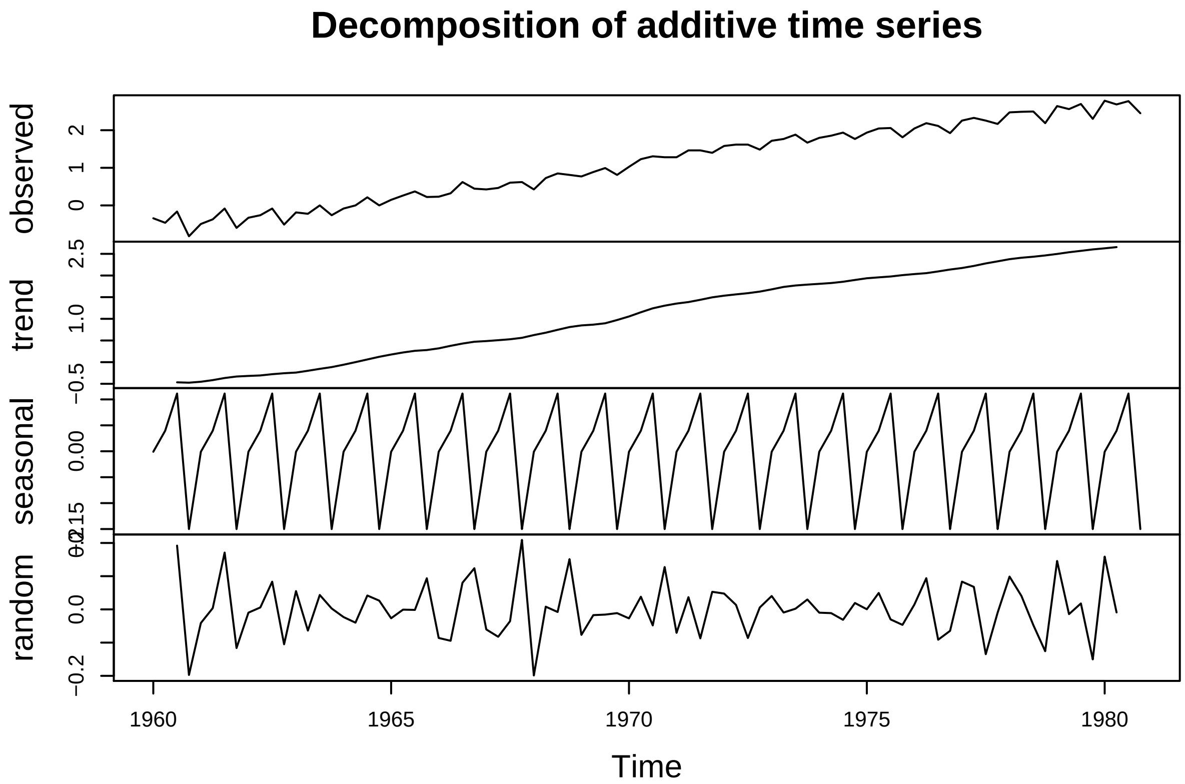


A practical interpretation for the above time series is that we are reducing the variation by taking the natural log, and then detrending it by taking the first difference.

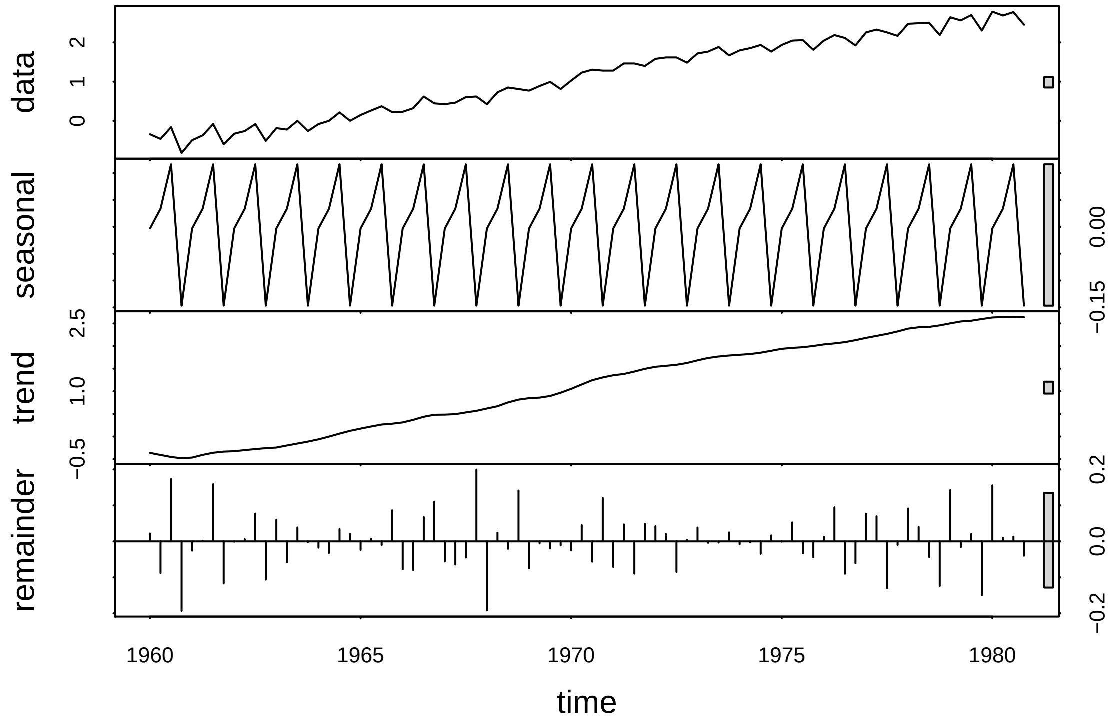
Question 2.9

I will use the *decompose* and *stl* functions with various parameters to generate structural models for the log transform of the Johnson&Johnson data.

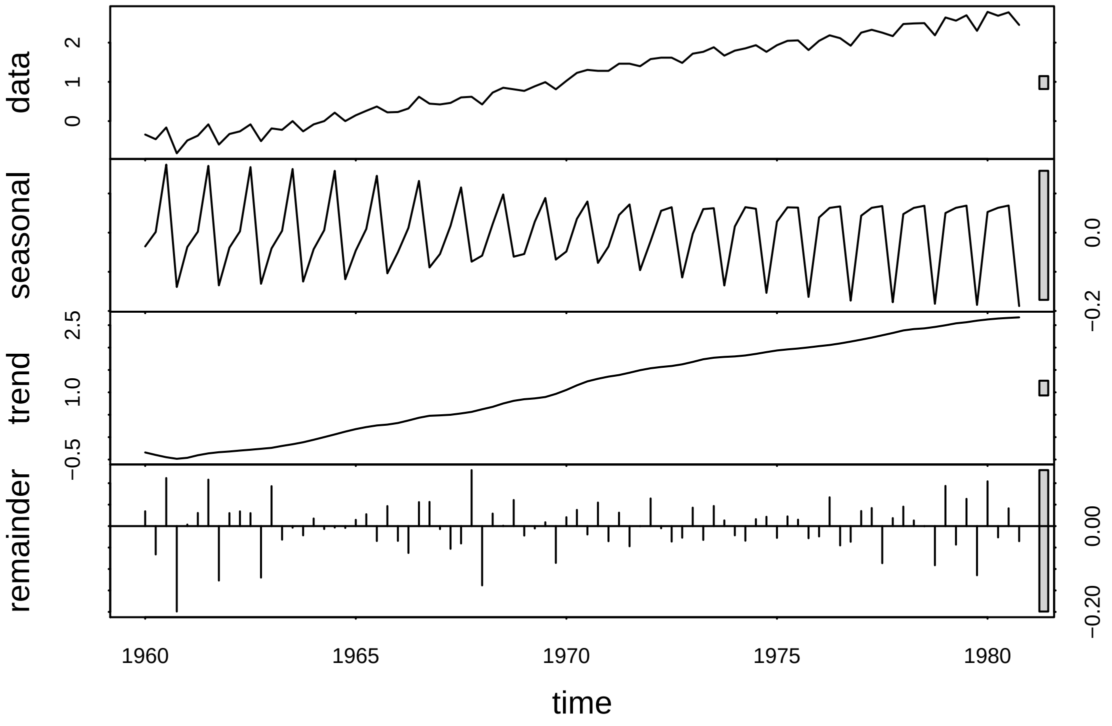
```
x <- log(jj)
plot(decompose(x))
```



```
plot(stl(x, s.window="periodic"))
```



```
plot(stl(x, s.window=15))
```



In all decompositions, it is clear that we have a linear trend throughout the duration of the sample. There appears to be a cyclical component recycling every year. To me, the most interesting difference amongst these models is the residual display in the fourth panel of each window. We have a more difficult time explaining the variation in the beginning and end of the time series, with the same issue occurring around 1967. There also appears to be some pattern in the residuals, in that we appear to have some sort of periodicity occurring throughout.