Exam1

Avery Loftin

10/15/2018

```
library(TSA)
```

```
##
## Attaching package: 'TSA'
```

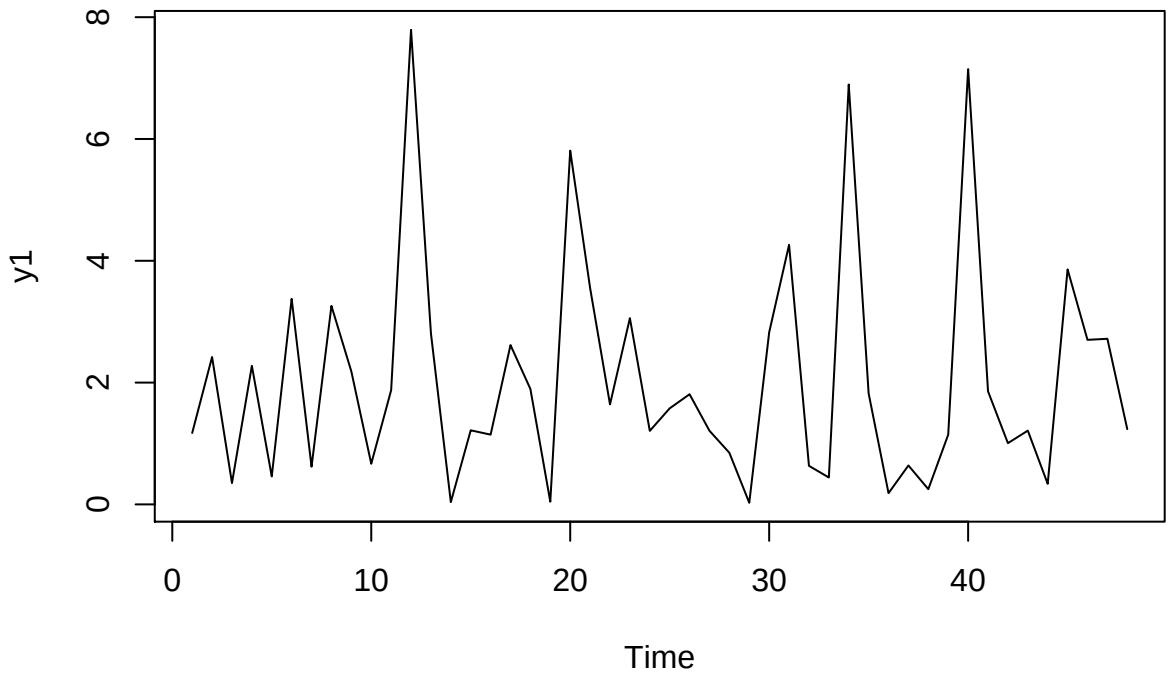
```
## The following objects are masked from 'package:stats':
##
##     acf, arima
```

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

1. Simulate a completely random process of length 48 with independent, chi-square distributed values, each with 2 degrees of freedom. Display the time series plot. Does it look random and non normal?

```
y1 <- rchisq(n = 48, df = 2)

plot.ts(y1)
```



This random process of iid chi-square distributed values looks random, as no distinct patterns can be seen. It also looks nonnormal, since there are no negative values. A process drawn from a normal distribution would have points randomly scattered about zero with variance of σ^2.

2. Let e_t be a sequence of independent normal random variables, each with mean 0 and variance σ^2, and a, b, and c be constants. Which, if any, of the following processes are stationary? For each stationary process specify their mean and autocovariance function.

a) $Y_t = a + be_t + ce_{t-1}$

This process is stationary because mean, variance, and autocovariance are all independent of time.

Expected Value: $E[Y_t] = E[a] + bE[e_t] + cE[e_{t-1}] = a + 0 + 0 = a$ Variance: $Var(Y_t) = Var(a) + b^2Var(e_t) + c^2Var(e_{t-1}) = \sigma^2(b^2 + c^2)$ Autocovariance: If k=1: $Cov(Y_t, Y_t - 1) = Cov(be_t + ce_{t-1}, be_{t-1} + ce_{t-2}) = cbCov(e_{t-1}, e_{t-1}) = cb(b^2\sigma^2 + c^2\sigma^2)$ If k>1: $Cov(Y_t, Y_{t-k}) = 0$

b) $Yt = e_te_{t-1}$

This process is stationary because mean, variance, and autocovariance are all independent of time.

Expected Value: $E[Y_t] = E[e_te_{t-1}] = Cov(e_t, e_{t-1}) + E[e_t]E[e_{t-1}] = 0$ Variance: $Var(Y_t) = Var[e_te_{t-1}] = [E(e_t)]^2Var(e_{t-1}) + [E(e_{t-1})]^2Var(e_t) + 2E[e_t]E[e_{t-1}]Cov(e_t, e_{t-1}) + Var(e_t)Var(e_{t-1}) + Cov(e_t, e_{t-1})^2 = Var(e_t)Var(e_{t-1}) = \sigma^4$ Autocovariance: $\gamma(k) = 0$ for all k

c) $Y_t = e_tcos(ct) + e_{t-1}sin(ct)$

This process is not stationary because variance depends on time

Expected Value: $E[Y_t] = E[e_tcos(ct) + e_{t-1}sin(ct)] = E[e_t]E[cos(ct)] + E[e_{t-1}]E[sin(ct)] = 0$ Variance: $Var(Y_t) = Var[e_tcos(ct) + e_{t-1}sin(ct)] = cos(ct)^2Var(e_t) + sin(ct)^2Var[e_{t-1}] = \sigma^2(cos(ct)^2 + sin(ct)^2)$ Autocovariance: if k=1: $Cov(Y_t, Y_{t-1]}) = Cov(e_{t-1}sin(ct), e_{t-1}cos(ct)) = sin(ct)cos(ct)\sigma^2$ if k>0: $Cov(Y_t, Y_{t-k]}) = 0$

3. What does autocovariance measure?

Autocovariance measures the covariance between data points of the same dataset separated by various lags.

4. Suppose, you are a data scientist at Analytics Vidhya. And you observed the views on the articles increases during the month of JanMar. Whereas the views during NovDec decreases. Does the above statement represent seasonality? Explain.

Yes, the values of data points from Jan-Mar are correlated with the data points of previous or future Jan toMar. The same applies with Nov to Dec datapoints. If the data points are monthly values, a lag of 6 would be negatively correlated, and a lag of 12 would be positively correlated.

5. Looking at the below ACF plot, would you suggest to apply AR or MA in ARIMA modeling technique?

MA(1) because the ACF shows one significant correlation with the 1st lag, and none with any others. If AR were a well suited model, the ACF would show exponenitally decaying correlation with sequential lags.

6. The second order moving average model, denoted by MA(2) is $Y_t = \mu + e_t + \theta_1e_{t-1} + \theta_2e_{t-2}$

a) Calculate mean and variance of a MA(2).

Mean: $E[Y_t] = E[\mu] + E[e_t] + \theta_1E[e_{t-1}] + \theta_2E[e_{t-2}] = \mu$ Variance: $Var(Y_t) = Var(\mu) + Var(e_t) + \theta_1^2Var(e_{t-1}) + \theta_2^2Var(e_{t-2}) = \sigma^2(1 + \theta_1^2 + \theta_2^2)$

b) Calculate the autocorrelation function and sketch the ACF for this model.

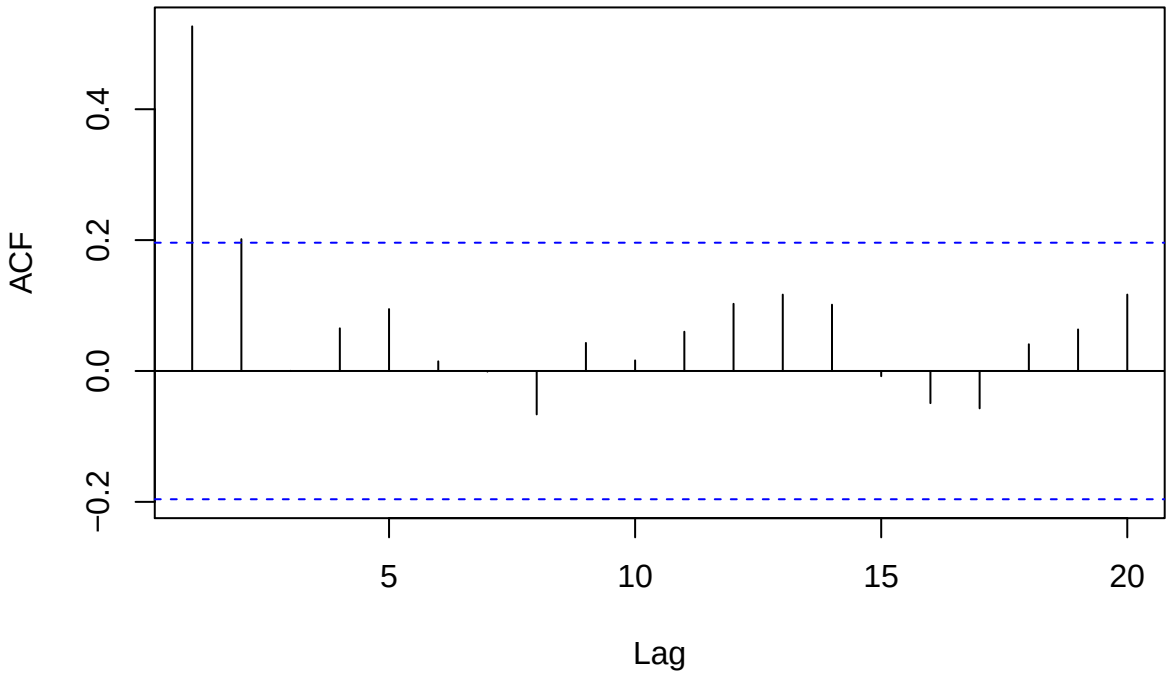
k=1: $Cov(Y_t, Y_{t-1}) = Cov(\theta_1e_{t-1} + \theta_2e_{t-2}, e_{t-1} + \theta_1e_{t-2}) = \theta_1\sigma^2 + \theta_2\sigma^2$ $Corr(Y_t, Y_{t-1}) = (\theta_1 + \theta_2)/(1 + \theta_1 + \theta_2)$

k=2: $Cov(Y_t, Y_{t-2}) = Cov(\theta_2e_{t-2}, e_{t-2}) = \theta_2\sigma^2$ $Corr(Y_t, Y_{t-2}) = \theta_2/(1 + \theta_1 + \theta_2)$

k>2: $Corr(Y_t, Y_{t-k}) = 0$

```
noise <- rnorm(100)
MA2 <- c()
MA2[1] <- noise[1]
MA2[2] <- noise[2]

for (t in 3:100){
    MA2[t] = noise[t] + .6*noise[t-1] + .4*noise[t-2]
}
acf(MA2)
```



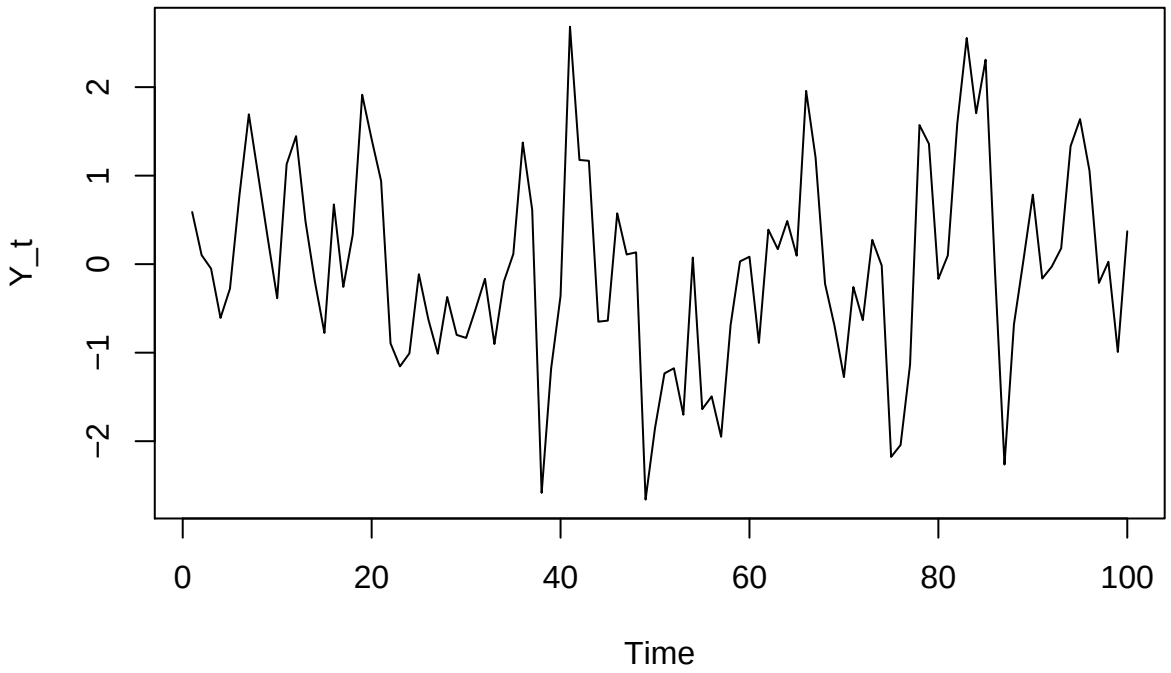
I don't know how to put images in here, so here's a simulation that illustrate what the acf looks like. There are 2 significant correlations with lag 1 and lag 2, the rest are insignificant.

c) Simulate 100 observations for MA(2) and plot it as a time series. Is it stationary?

```
theta1 <- .6
theta2 <- .4
mu <- 0
noise <- rnorm(100)
Y_t <- c()
Y_t[1] <- noise[1]
Y_t[2] <- noise[2]

for (i in 3:100) {
  Y_t[i]= mu+noise[i]+theta1*noise[i-1]+theta2*noise[i-2]
}

plot.ts(Y_t)
```



The MA(2) process is stationary, as mean, variance, and autocovariance are all independent of t.

d) For what values of θ is the MA(2) stationary?

Finite MA models, such as MA(2), are stationary for any values of theta.

7. Below is the qq plot of residuals of the linear regression fit. Does the normality assumption of the model met? Explain.

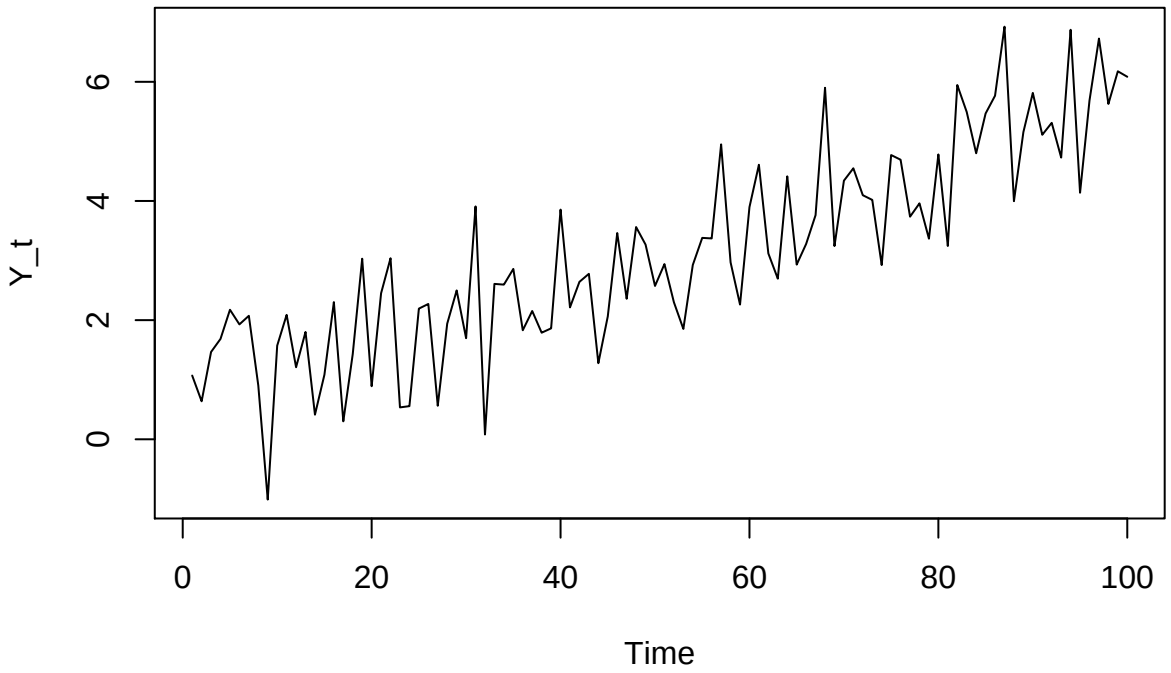
The normality assumption is not met, as the QQ Plot shows a lot of concavity, especially towards the tails of the plot.

8. Suppose $Y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + e_t$, where e_t is a zero mean stationary series with autocovariance function γ_k and β_0, β_1, and β_2 are constants.

a) Show that Y_t is not stationary and conform your answer using a simulation of 100 observations for the model. Take time ranges from 0 to 1.

Y_t is not stationary, as $E[Y_t] = \beta_0 + \beta_1 t + \beta_2 t^2$ which depends on time.

```
b0 <- 1
b1 <- 2
b2 <- 3
Y_t <- c()
noise <- rnorm(100)
for (i in 1:100){
  Y_t[i]=b0+b1*(i/100)+b2*(i/100)^2+noise[i]
}
plot.ts(Y_t)
```

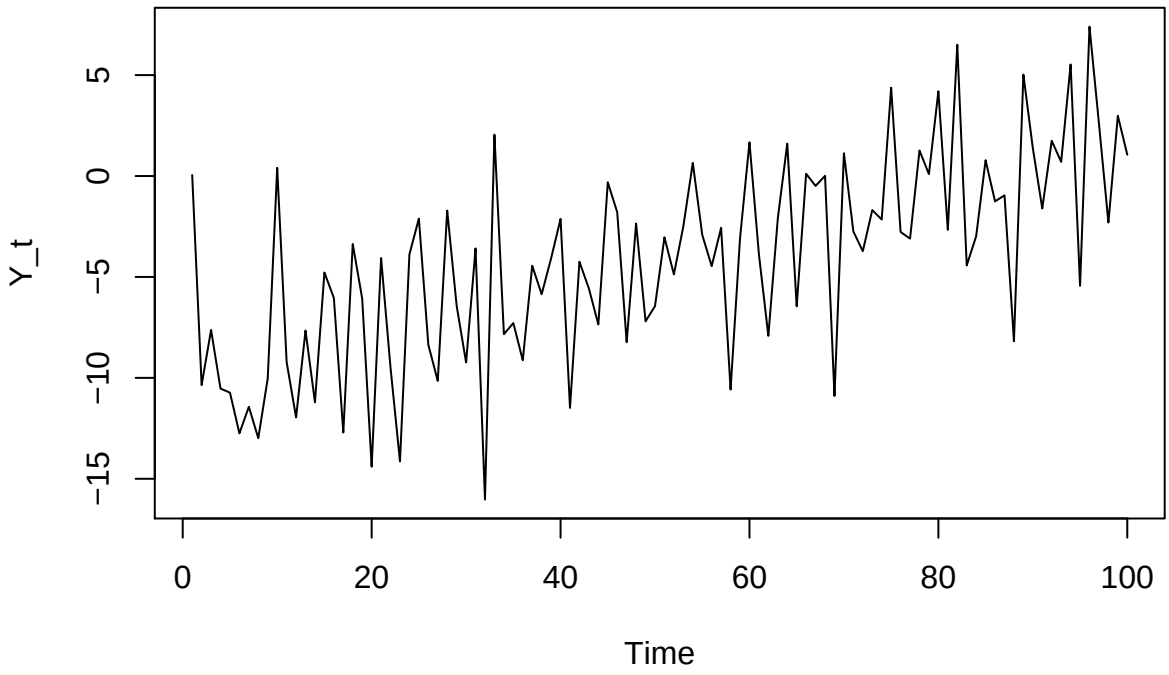


The plot confirms this finding, as it has an upward trend.

b) Suppose you take the difference $Z_t = \bigtriangledown Y_t = Y_t - Y_{t-1}$. Is it stationary? Find mean, variance and covariance if necessary to check stationarity. Plot Zt as a time series and explain if there is any trend.

It is still not stationary, as $E[Z_t] = E[\beta 1 + 4\beta_2 t - 4\beta_3 + e_t - e_{t-1}] = \beta 1 + 4\beta_2 t - 4\beta_3$

```
Y_t <- c()
Y_t[1] <- noise[1]
for (i in 2:100){
  Y_t[i]=b1+4*b2*(i/100)-4*b2+noise[i]-4*noise[i-1]
}
plot.ts(Y_t)
```

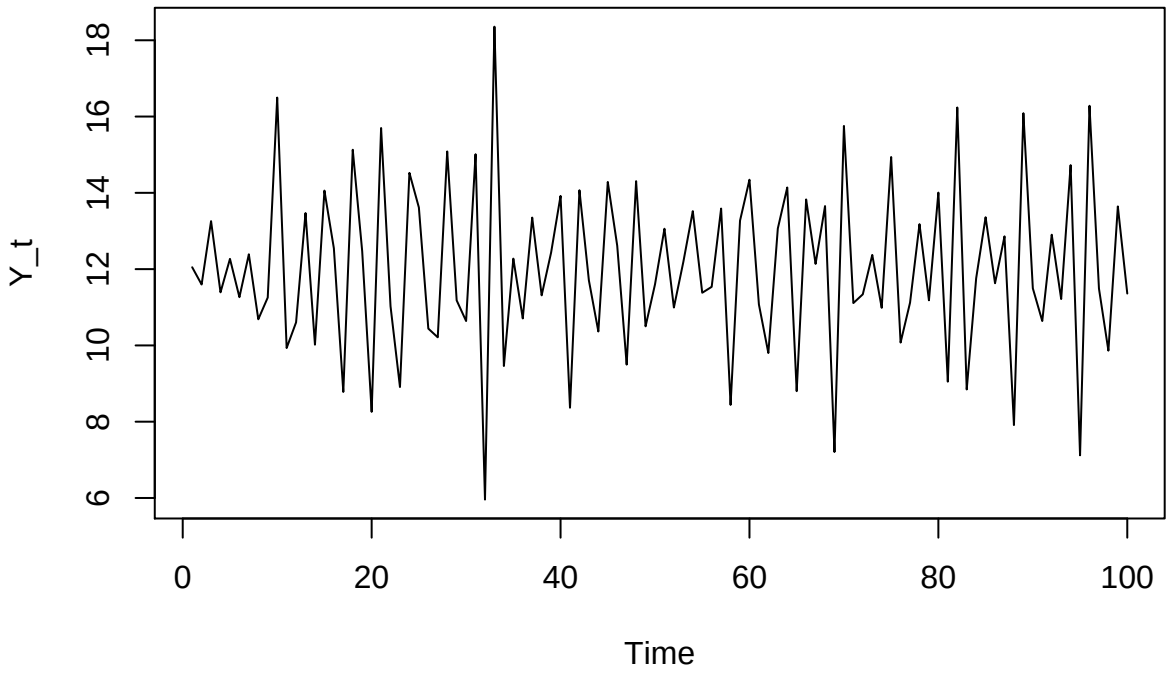


The plot confirms the upward trend.

c) Now take the difference of Z_t, that is $W_t = \bigtriangledown Z_t = \bigtriangledown^2 Y_t$. Is W_t stationarity? Find mean, variance, and covariance of W_t and plot it as a time series. Discuss if there is any trend in the model.

The data is now stationary, as mean, variant, and covariance are all independent of time. $W_t = 4\beta_1 + e_t - 2e_{t-1} + e_{t-2}$

```
Y_t <- c()
Y_t[1] <- noise[1]+12
Y_t[2] <- noise[2]+12
for (i in 3:100){
  Y_t[i]=4*b2+noise[i]-2*noise[i-1]+noise[(i-2)]
}
plot.ts(Y_t)
```

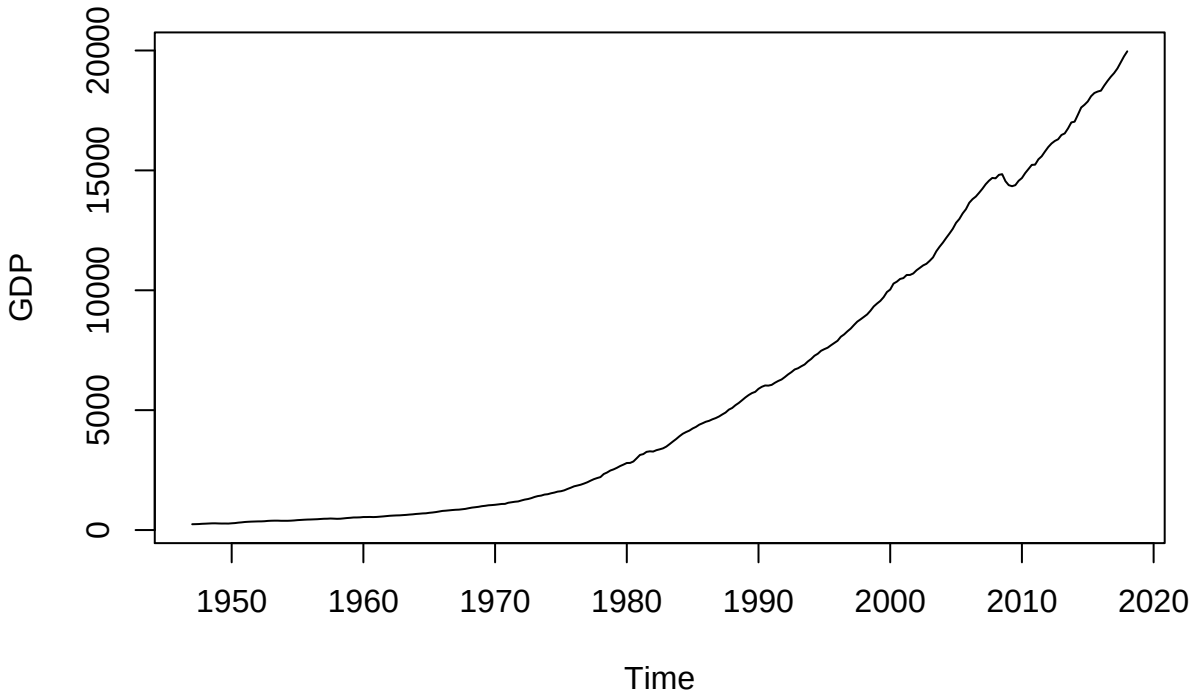


9. The data file gdp.txt contains quarterly GDP of the USA from 1947 to 2018. Use the data set to answer the following questions.

```
GDP <- read.table("gdp.txt", header = T)
GDP <- ts(GDP$GDP, start=c(1947, 1), frequency=4)
```

a) Display and interpret the time series plot for these data.

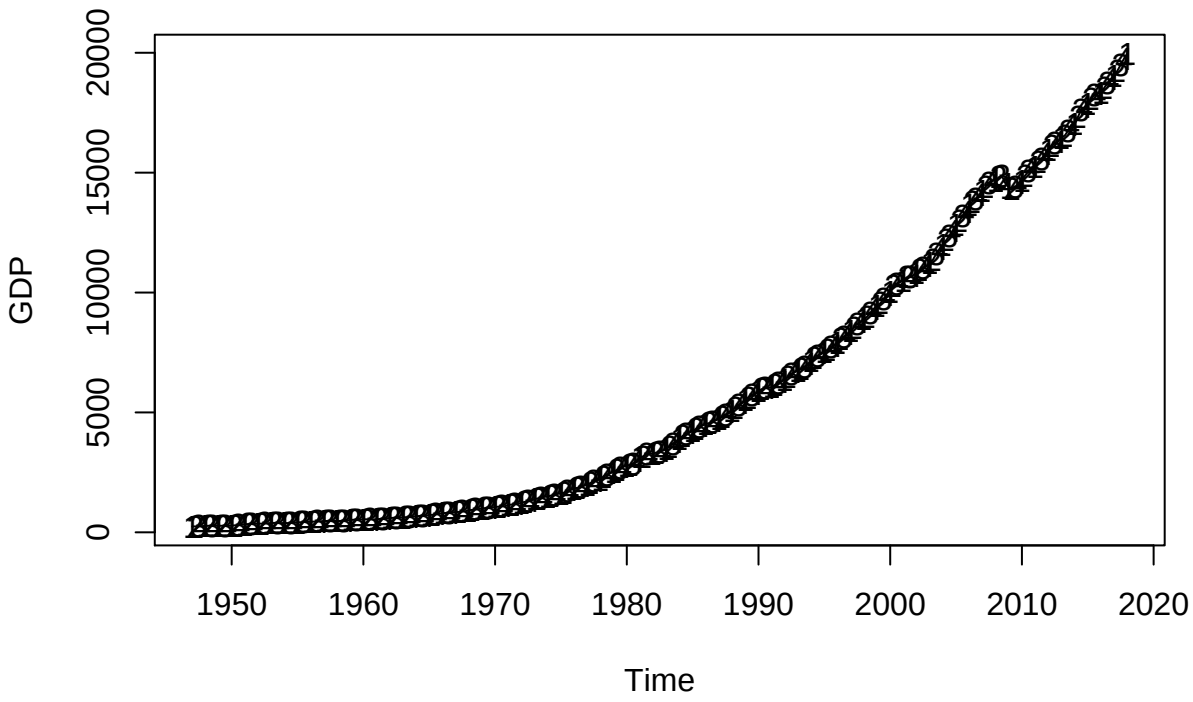
```
plot.ts(GDP)
```



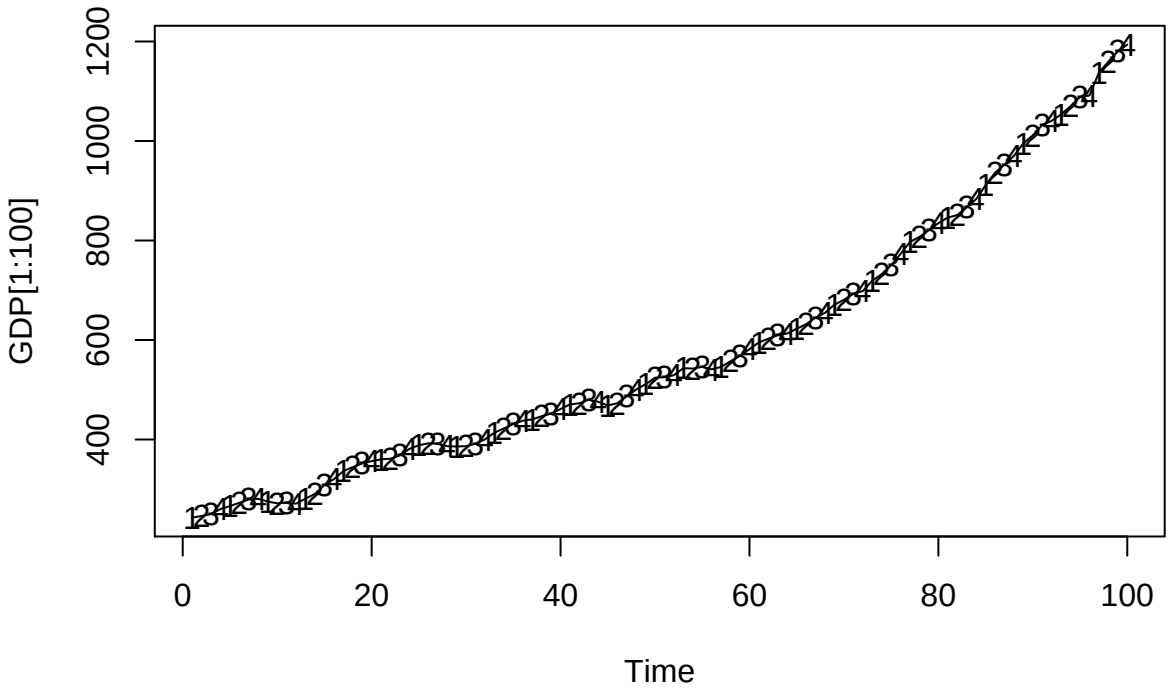
GDP has an upward trend with constant variance and no apparent seasonality.

b) Now construct a time series plot that uses separate plotting symbols for the various quarters. Does your interpretation change from that in part (a)?

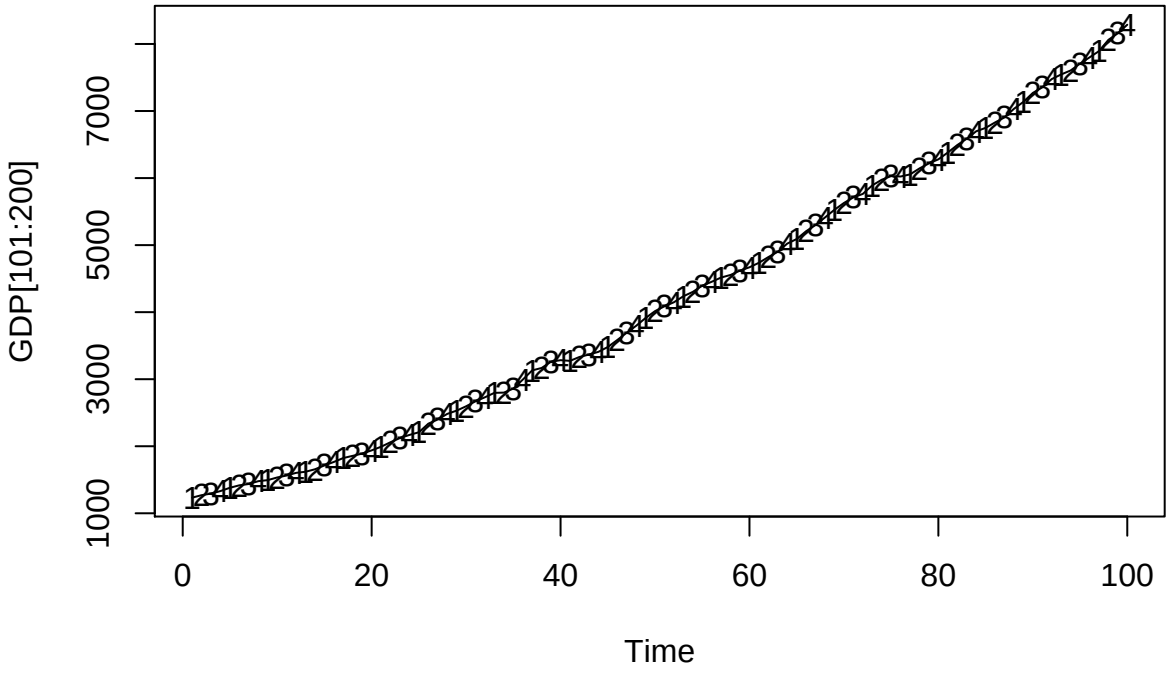
```
label <- rep(c("1","2","3","4"), 71)
label <- c(label, "1")
plot.ts(GDP)
points(x=time(GDP), y=GDP, pch=as.vector(label))
```




```
label2 <-rep(c("1","2","3","4"), 25)
plot.ts(GDP[1:100])
points(x=time(GDP[1:100]), y=GDP[1:100], pch=as.vector(label2))
```




```
plot.ts(GDP[101:200])
points(x=time(GDP[101:200]), y=GDP[101:200], pch=as.vector(label2))
```

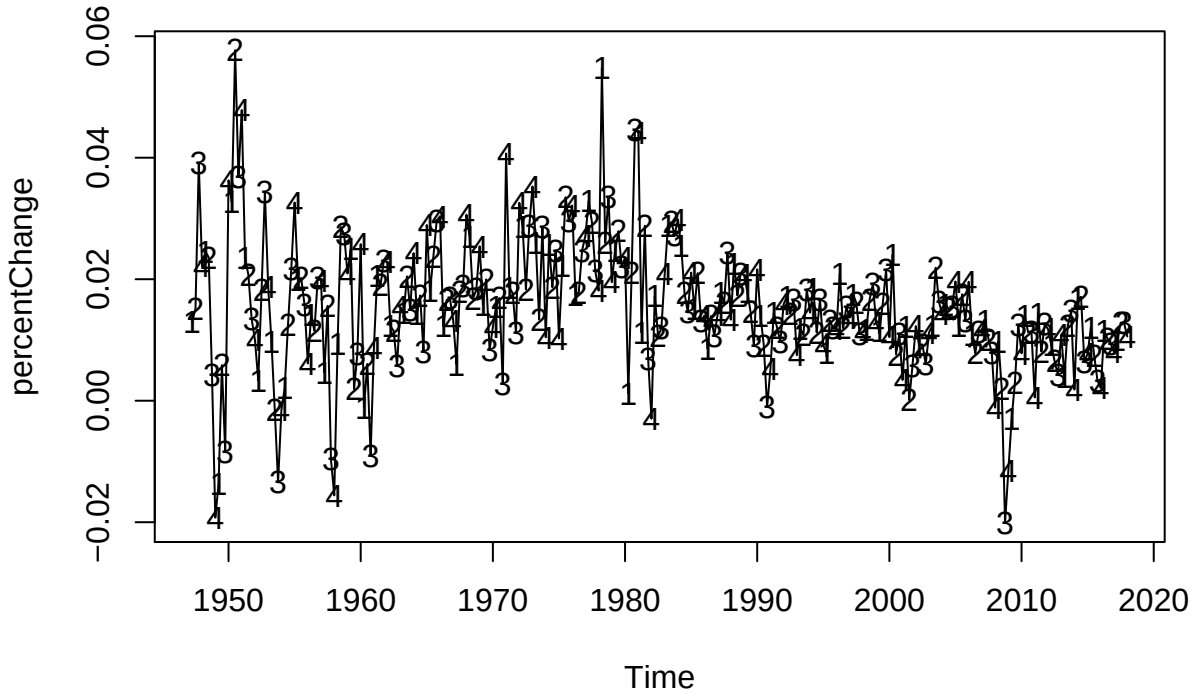


Unfortunately, the labels overshadow each other when viewing the entire dataset. looking at the first 100 observations, as well as the second 100 observations in turn, no seasonality jumps out at me. Especially since these data have very little concavity, and nearly monotonically increase.

c) Calculate and plot the sequence of quarter-quarter percentage changes in the GDP. Again, use plotting symbols that permit your to look for seasonality.

```
percentChange <- diff(GDP)/GDP
plot.ts(percentChange)
```

```
points(x=time(percentChange), y=percentChange, pch=as.vector(label))
```



This plot of the percent change quarter to quarter suggests that the percent changes for Q3 and Q4 tend to be the most significant, either having much larger or smaller values than the average, especially prior to 1985.

d) Use least squares to fit for both linear and quadratic trend to these data. Interpret the regression output and save the standardized residuals for further analysis.

```
lm1 <- lm(GDP ~ time(GDP))
cat("Linear Model:")
```

```
## Linear Model:
```

```
summary(lm1)
```

```
##
## Call:
## lm(formula = GDP ~ time(GDP))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2490.1 -1979.6  -314.4  1783.2  4610.8
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.260e+05  1.179e+04   -44.6   <2e-16 ***
## time(GDP)    2.683e+02  5.948e+00    45.1   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2065 on 283 degrees of freedom
## Multiple R-squared:  0.8779, Adjusted R-squared:  0.8774
## F-statistic:  2034 on 1 and 283 DF,  p-value: < 2.2e-16
```

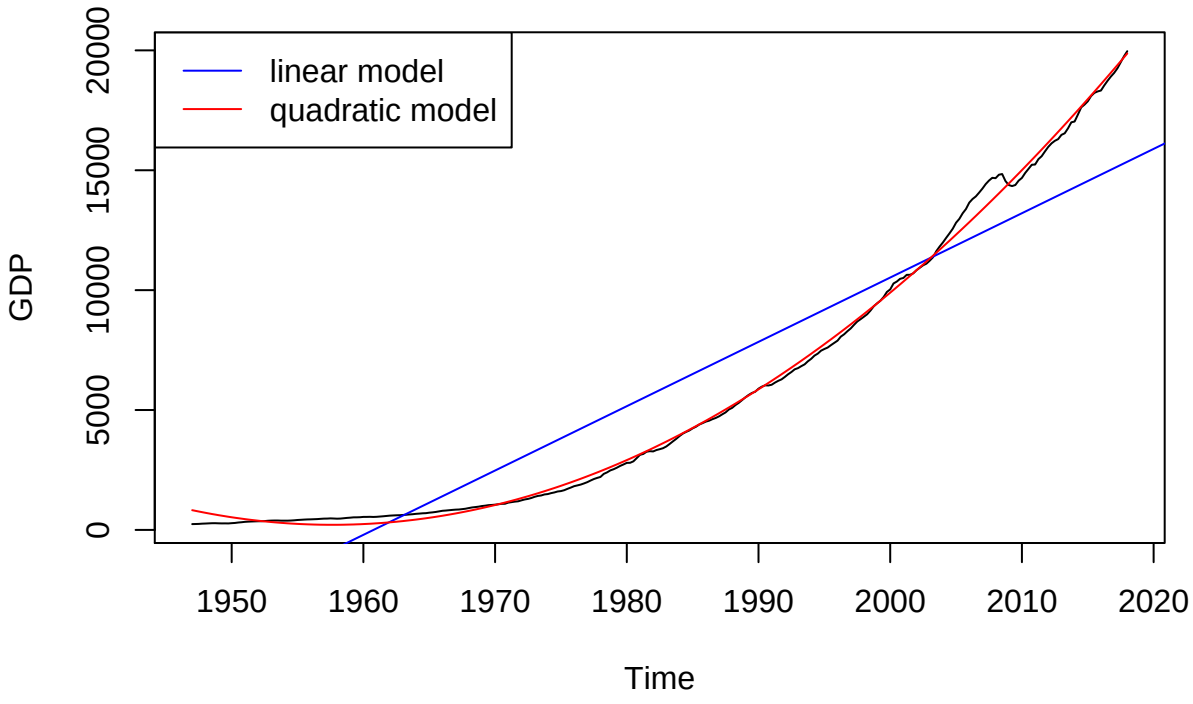
```r
cat("Quadratic Model:")
```

```
## Quadratic Model:
```

```r
sqrTerm <- time(GDP)^2
lm2 <- lm(GDP ~ time(GDP)+sqrTerm)
summary(lm2)
```

```
##
## Call:
## lm(formula = GDP ~ time(GDP) + sqrTerm)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -580.49 -181.73  -54.02  162.90  942.56
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.067e+07  1.629e+05   126.9   <2e-16 ***
## time(GDP)   -2.112e+04  1.644e+02  -128.5   <2e-16 ***
## sqrTerm      5.395e+00  4.145e-02   130.1   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 264.8 on 282 degrees of freedom
## Multiple R-squared:  0.998,  Adjusted R-squared:  0.998
## F-statistic: 7.035e+04 on 2 and 282 DF,  p-value: < 2.2e-16
```

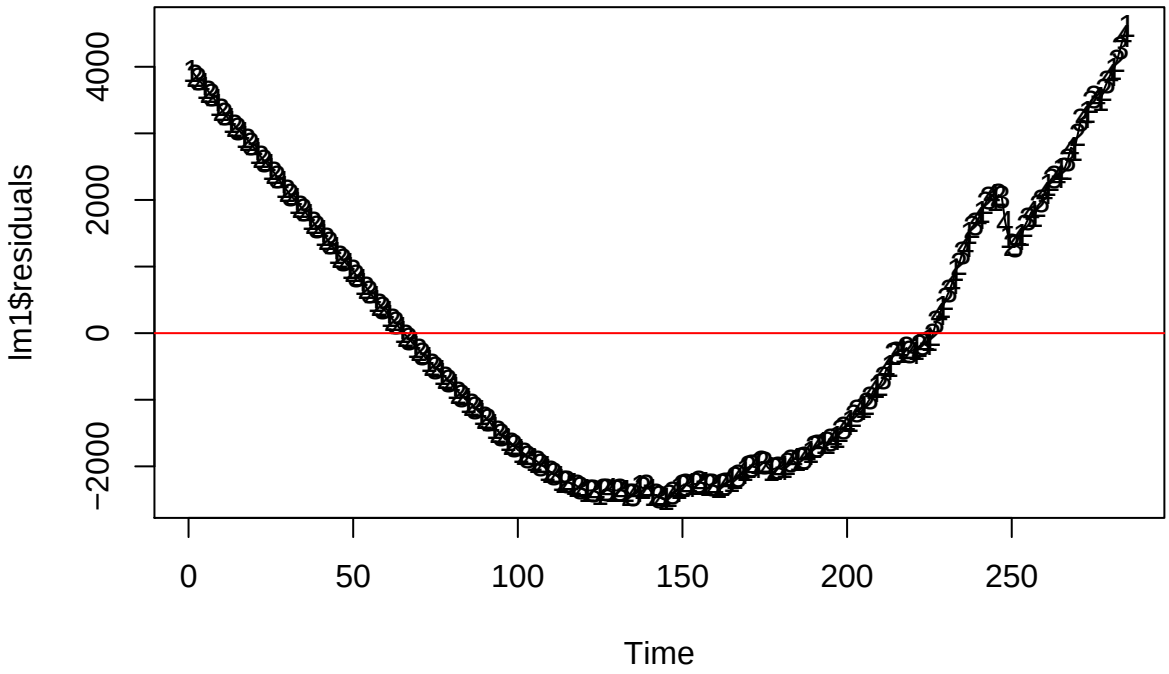
```r
plot.ts(GDP)
abline(lm1, col="blue")
lines(y=predict(lm2), x=time(GDP), type="l", col="red")
legend("topleft",legend=c("linear model", "quadratic model"), col = c("blue", "red"), lty=c(1,1))
```



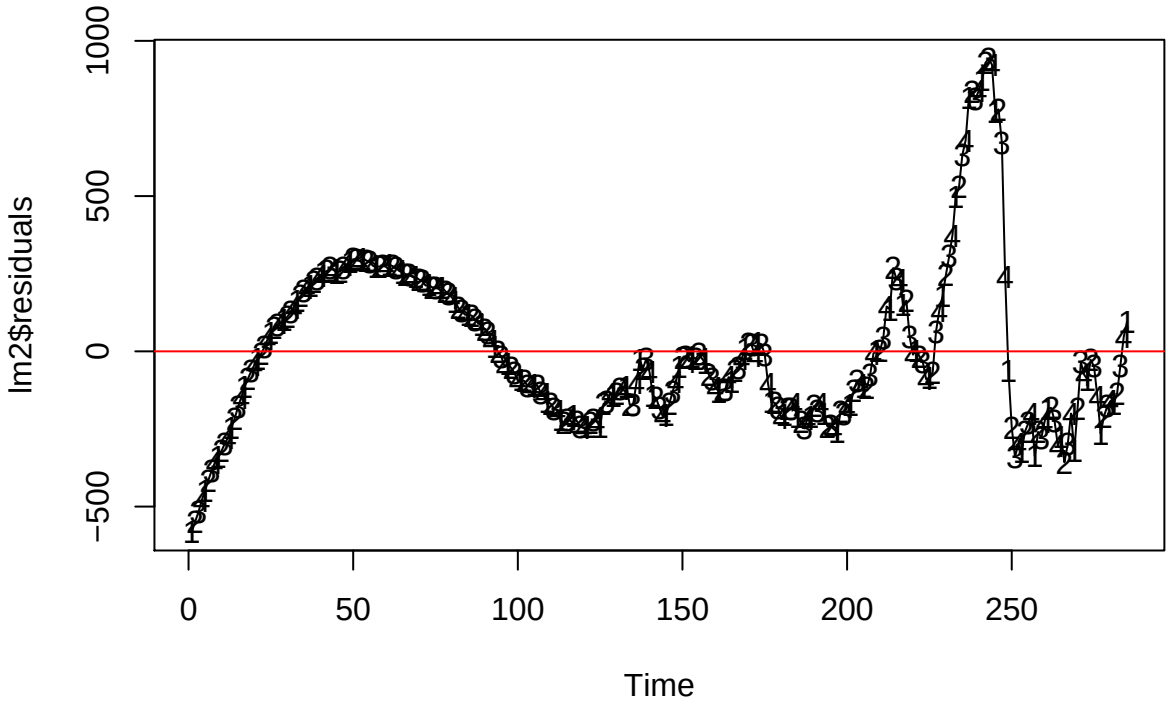
The linear and quadratic models have very statistically significant parameters and p values for the F-Test. The linear model has a mediocre R-squared value of .8779, while the quadratic model has an incredible R-squared value of .998. The quadratic model appears to be a much better fit.

e) Display a sequence plot of the standardized residuals and interpret. Use monthly plotting symbols so that possible seasonality may be readily identified.

```
plot.ts(lm1$residuals)
points(x=1:285, y=lm1$residuals, pch=as.vector(label))
abline(h=0, col="red")
```



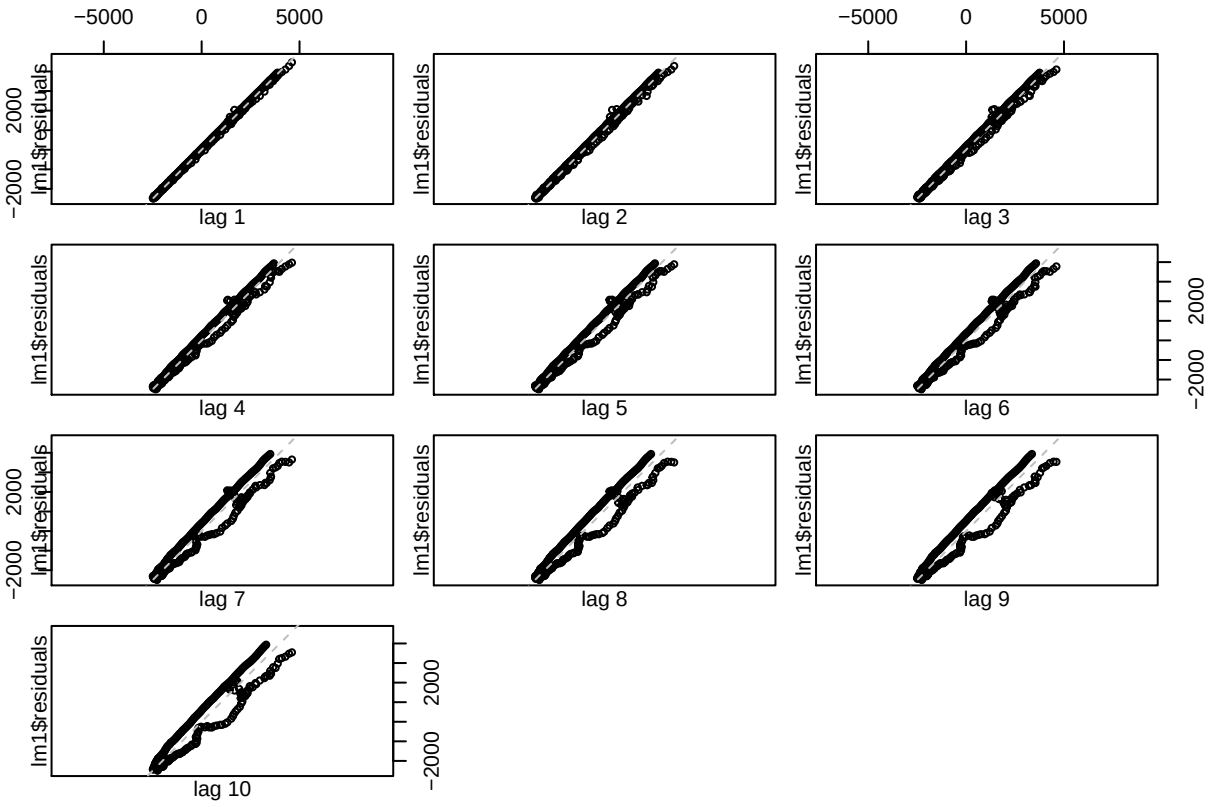

```
plot.ts(lm2$residuals)
points(x=1:285, y=lm2$residuals, pch=as.vector(label))
abline(h=0, col="red")
```



Both the linear model and quadratic model fail to meet the normality assumption of the error term. The residual values have very distinct patterns and are not evenly scattered about zero.

f) Perform the Runs test of the standardized residuals and interpret the results.

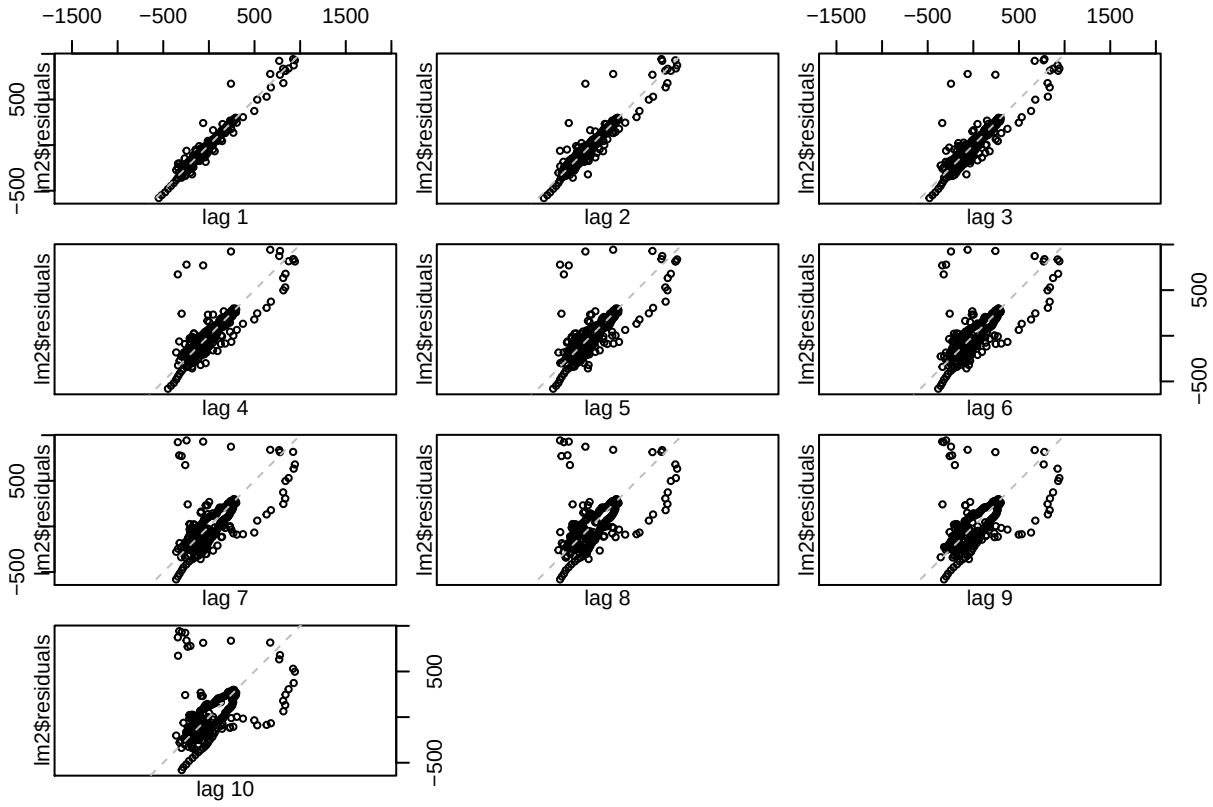
```
library(tseries)
lag.plot(lm1$residuals, lags = 10, do.lines = F, labels = F)
```



```
x <- factor(sign(lm1$residuals))
runs.test(x)
```

```
##
##  Runs Test
##
## data:  x
## Standard Normal = -16.671, p-value < 2.2e-16
## alternative hypothesis: two.sided
```

```
lag.plot(lm2$residuals, lags = 10, do.lines = F, labels = F)
```



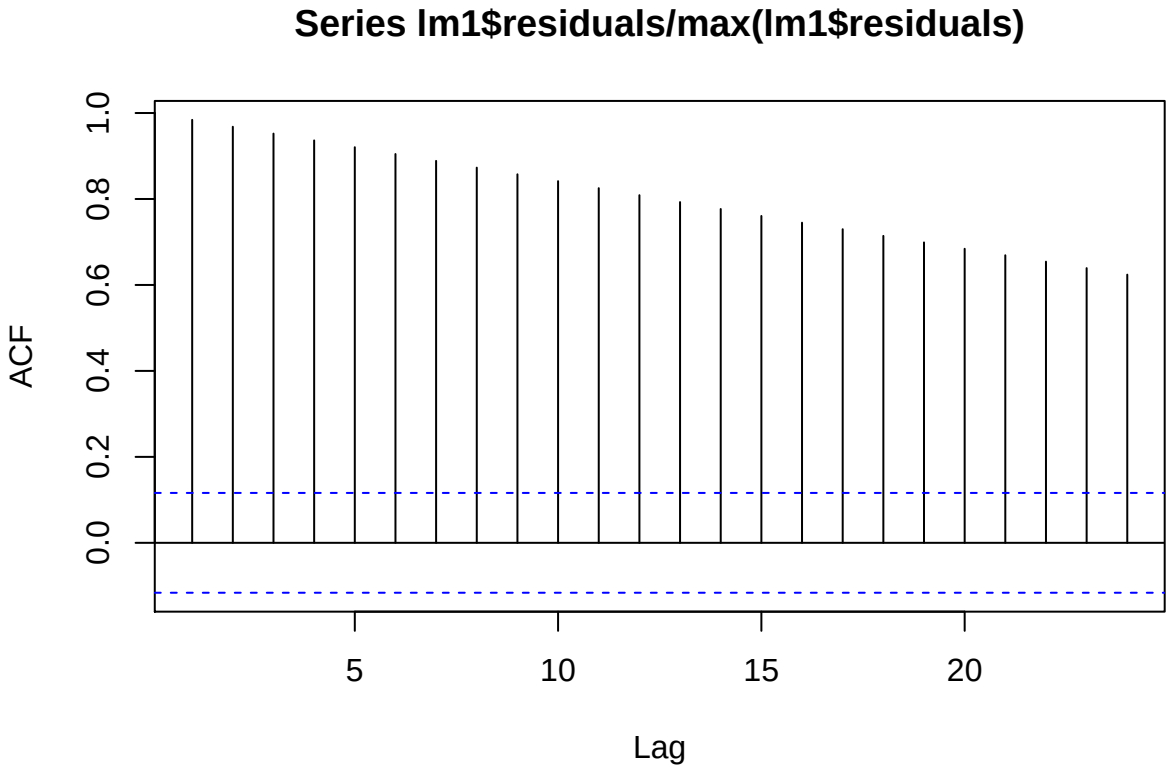

```
x <- factor(sign(lm2$residuals))
runs.test(x)
```

```
##
##  Runs Test
##
## data:  x
## Standard Normal = -15.297, p-value < 2.2e-16
## alternative hypothesis: two.sided
```

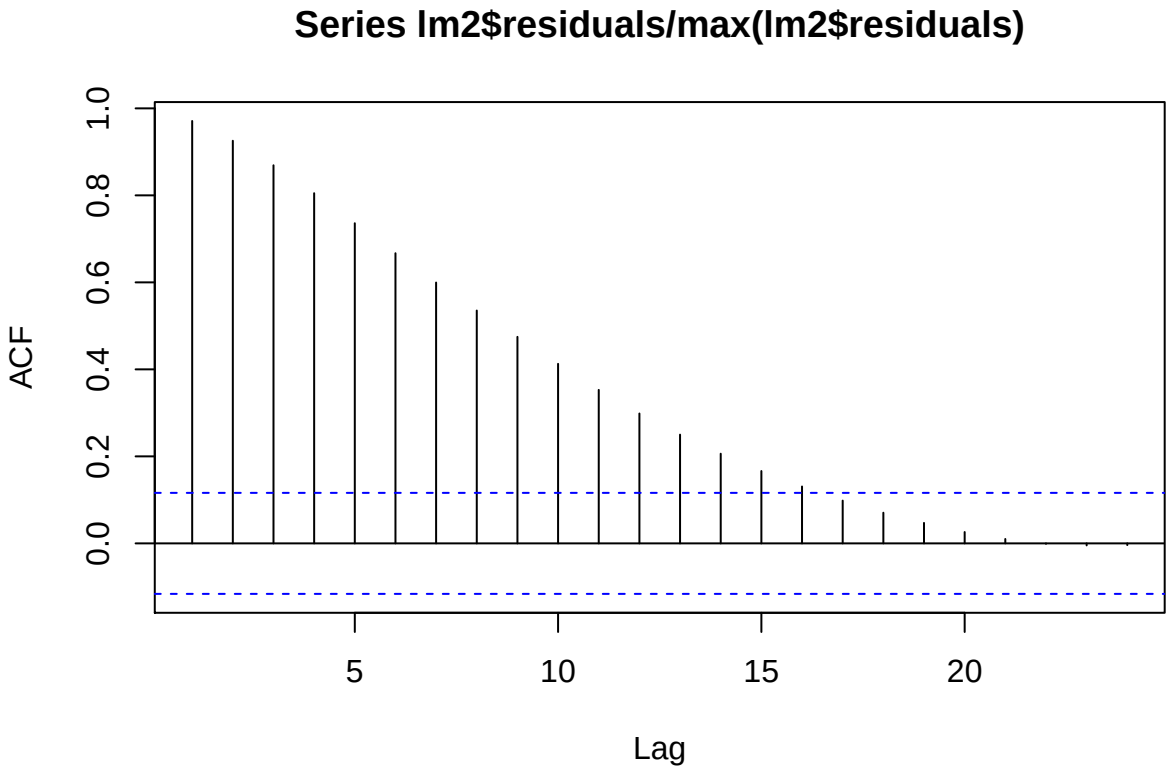
The runs test shows we confidently reject the null hypothese that the residuals are random for both the linear and quadratic models.

g) Calculate and interpret the sample autocorrelations for the standardized residuals.

```r
acf(lm1$residuals/max(lm1$residuals))
```




```r
acf(lm2$residuals/max(lm2$residuals))
```

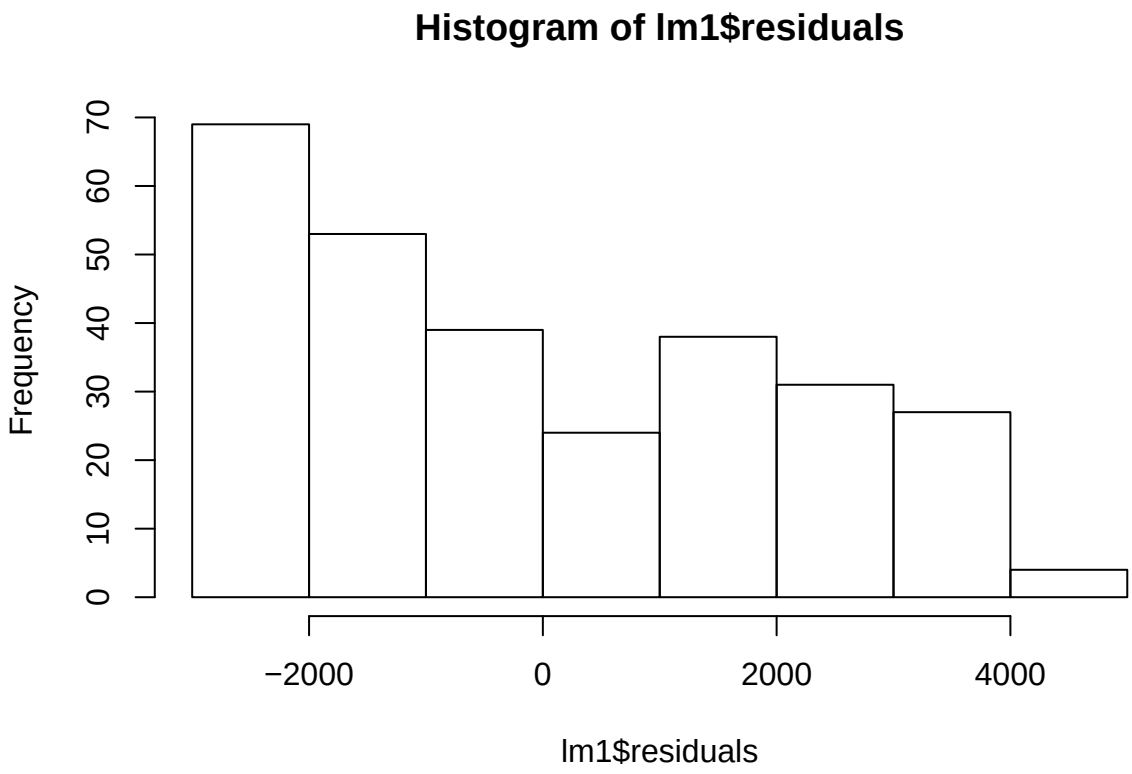


The residuals for both models have significant correlation with successive lags, futher suggesting the error

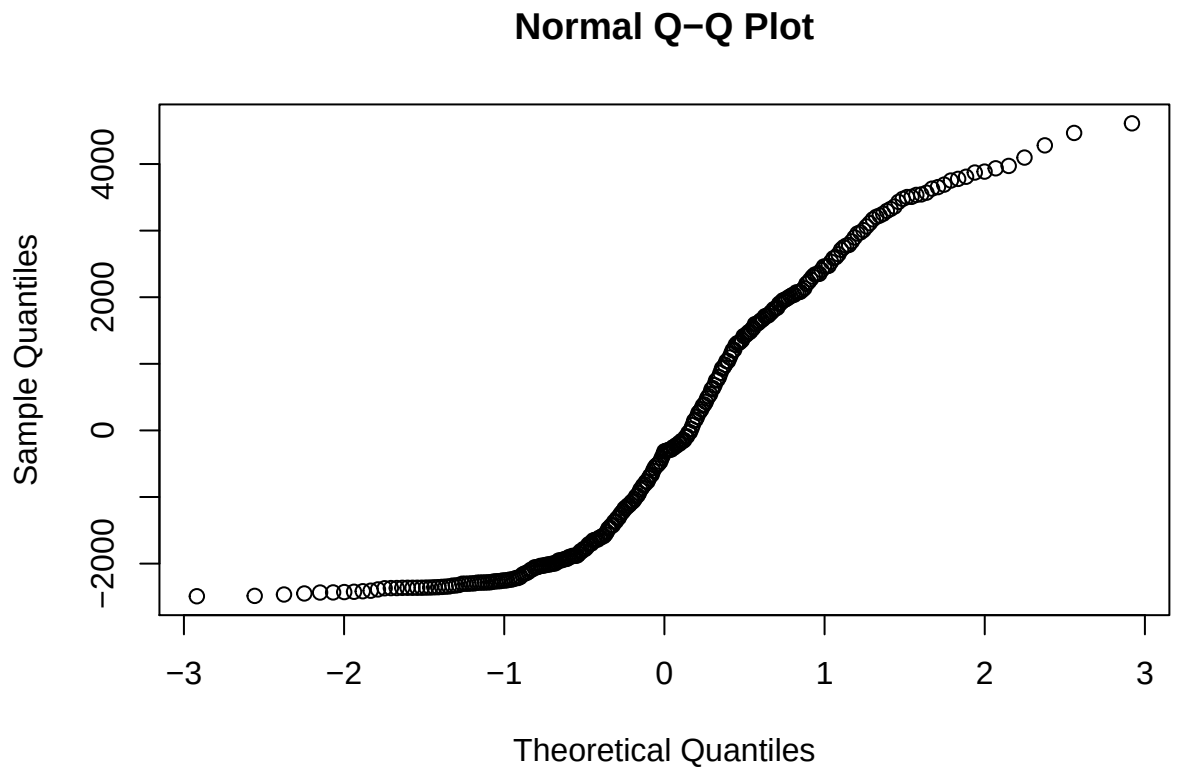
term fails to be normal.

h) Calculate and interpret the sample autocorrelations for the standardized residuals. Investigate the normality of the standardized residuals (error terms). Consider histograms and normal probability plots. Interpret the plots.

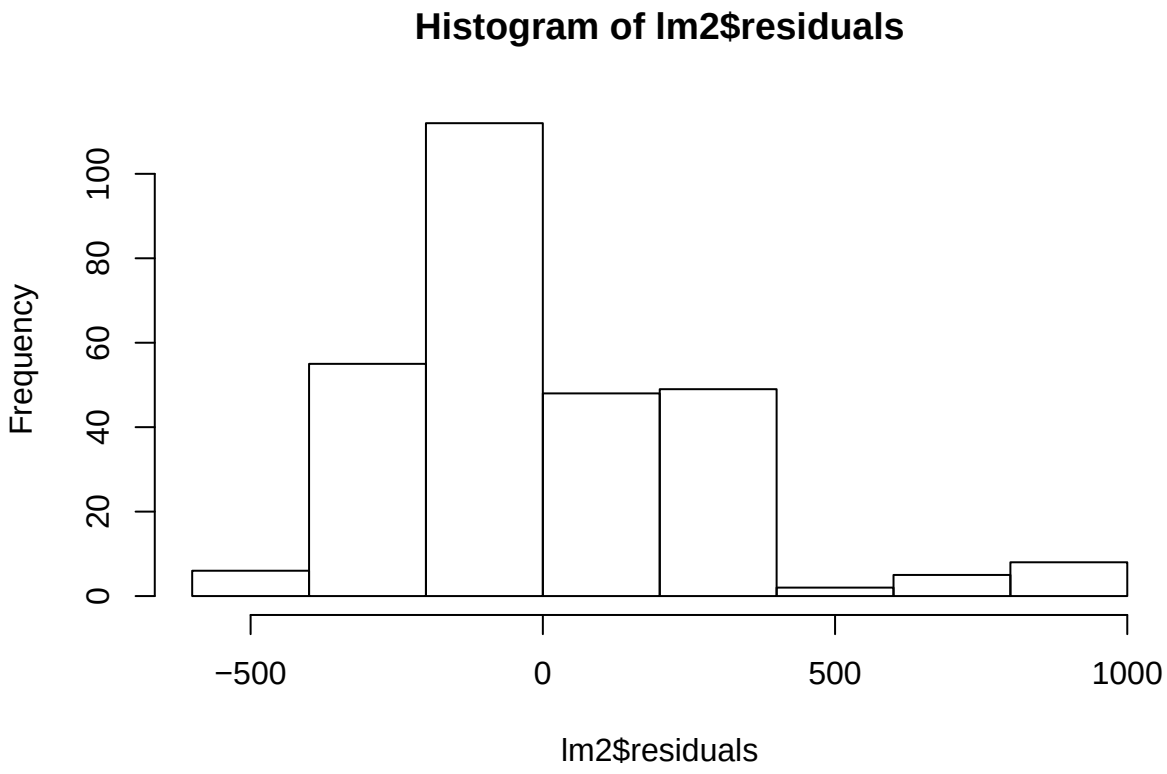
```
hist(lm1$residuals)
```



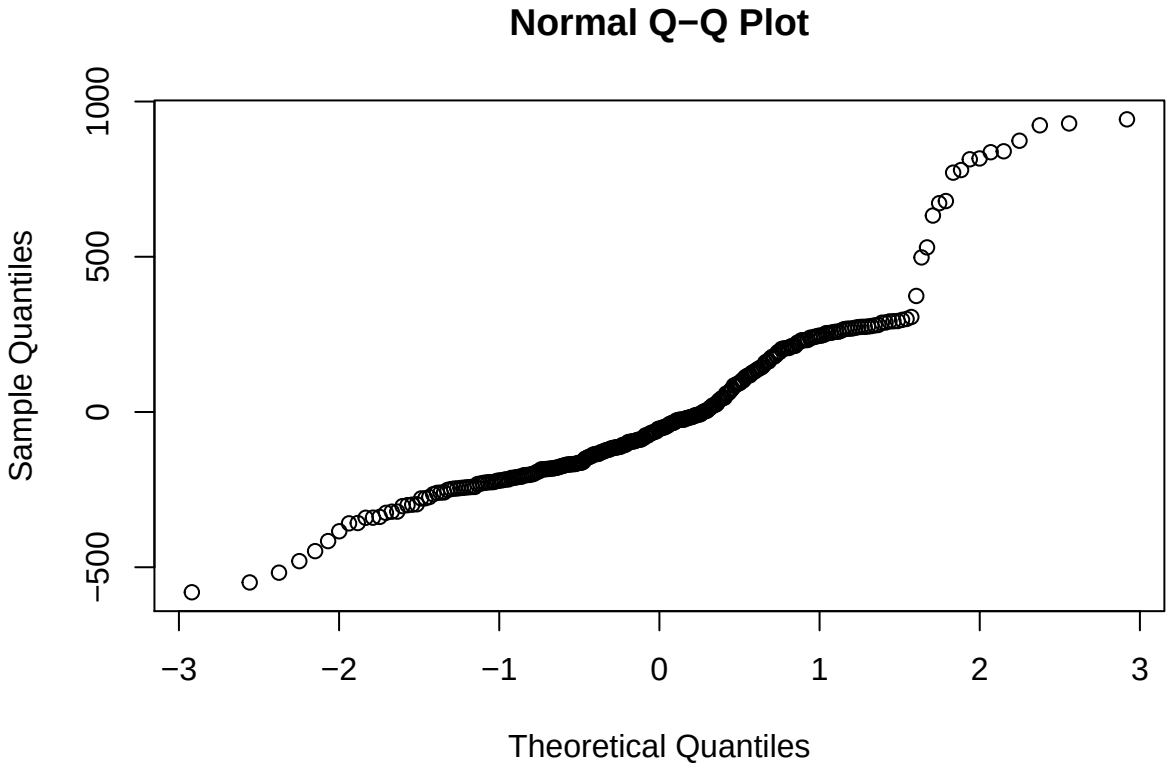

```
qqnorm(lm1$residuals)
```




```
hist(lm2$residuals)
```




```
qqnorm(lm2$residuals)
```



The histograms for the linear and quadratic models show the error term is not normally distributed. The histogram for the quadratic model is heavly skewed to the right. The QQ Plots for both also show extreme concavity.