HW4

Avery Loftin

10/7/2018

```
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. The data file wages contains monthly values of the average hourly wages (in dollars) for workers in the U.S. apparel and textile products industry for July 1981 through June 1987.

```
data(wages)
```

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

```
plot(wages)
```

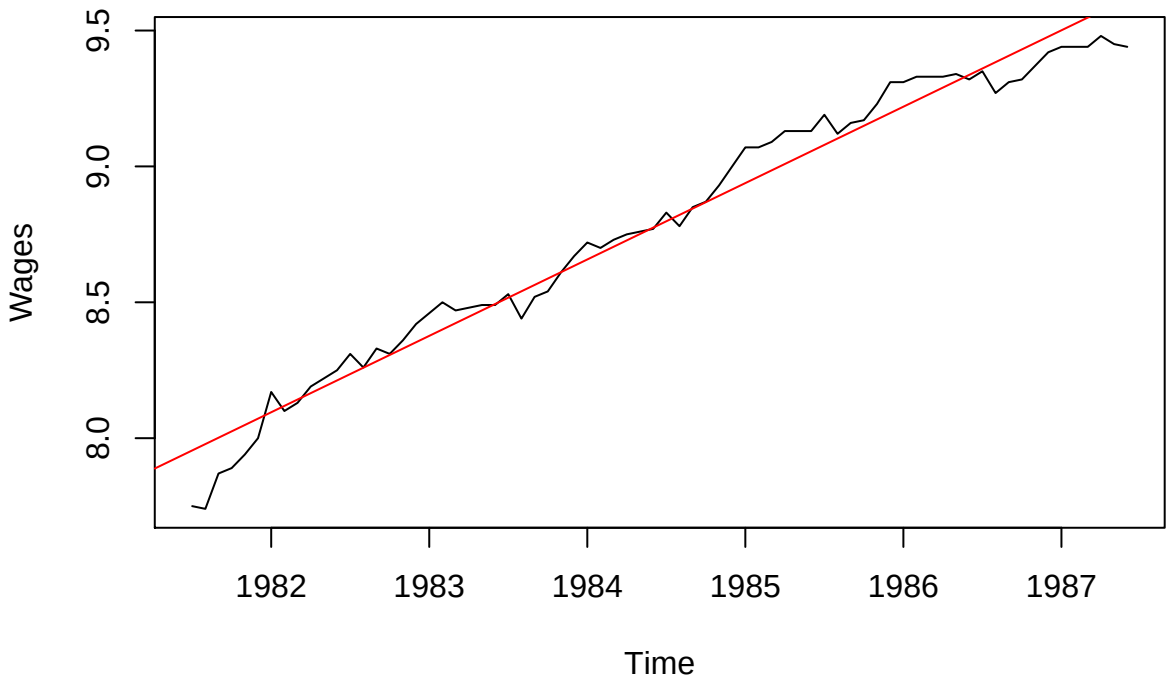
The average salary shows an upward trend with constant variance. These data also show signs of seasonality with positive correlation between each month of the years.

(b) Use least squares to fit a linear time trend to this time series. Interpret the regression output. Save the standardized residuals from the fit for further analysis.

```r
lm <- lm(wages ~ time(wages))
summary(lm)
```

```
##
## Call:
## lm(formula = wages ~ time(wages))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.23828 -0.04981  0.01942  0.05845  0.13136
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.490e+02  1.115e+01  -49.24   <2e-16 ***
## time(wages)  2.811e-01  5.618e-03   50.03   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08257 on 70 degrees of freedom
## Multiple R-squared:  0.9728, Adjusted R-squared:  0.9724
## F-statistic:  2503 on 1 and 70 DF,  p-value: < 2.2e-16
```

```r
plot(wages)
abline(lm, col="red")
```

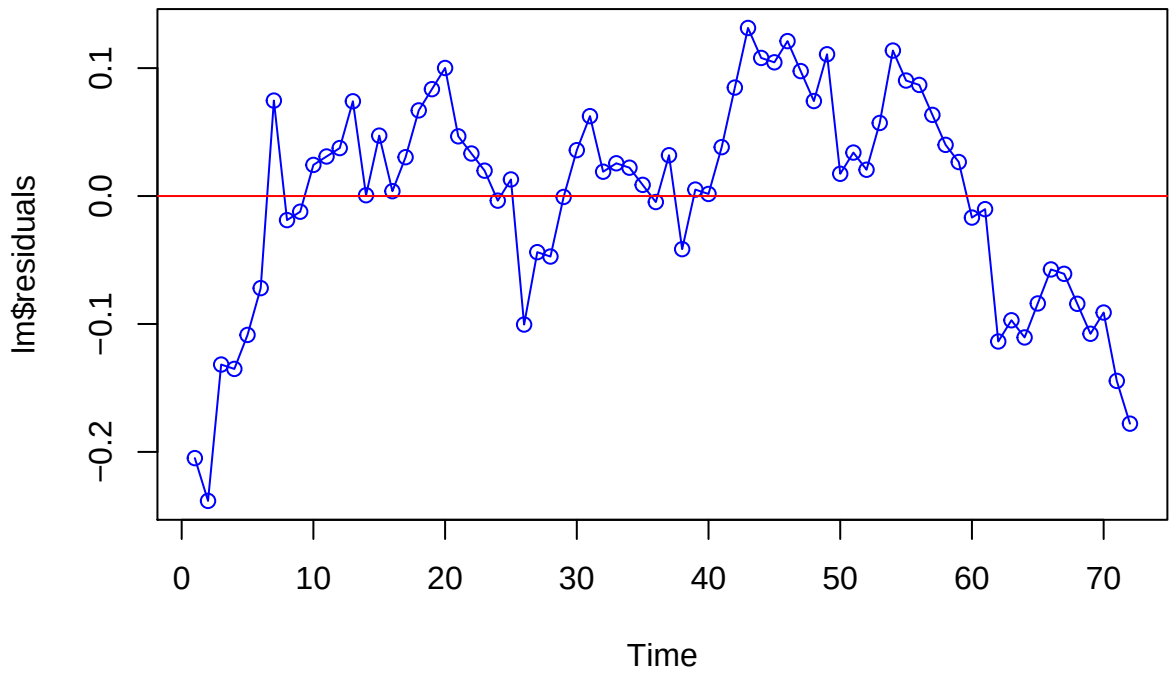


The summary of the linear model suggests that both the intercept and slope values are very statistically significant in modeling the change of wages, meaning they are different from zero. The F Test also has a very small p-value, meaning that both the parameters in the model are different from zero holding the other one constant.

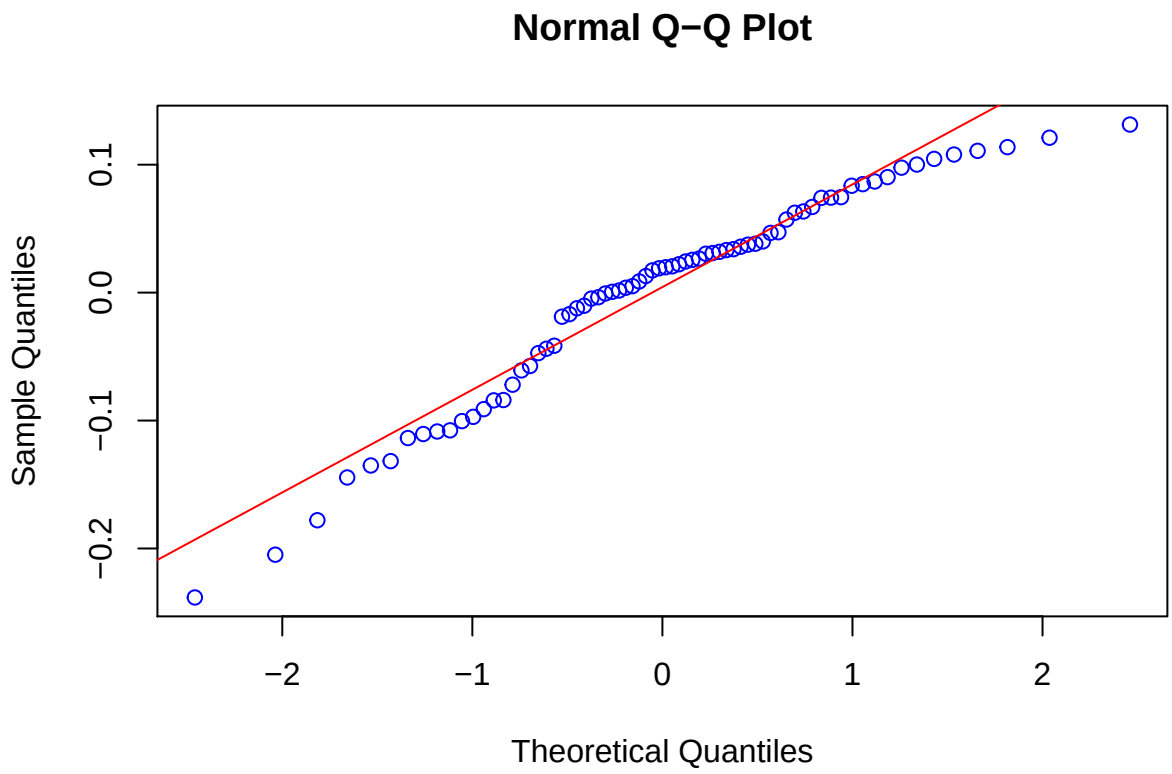
(c) Construct and interpret the time series plot of the standardized residuals from part (b).

```r
plot.ts(lm$residuals, col="blue")
points(lm$residuals, col="blue")
```

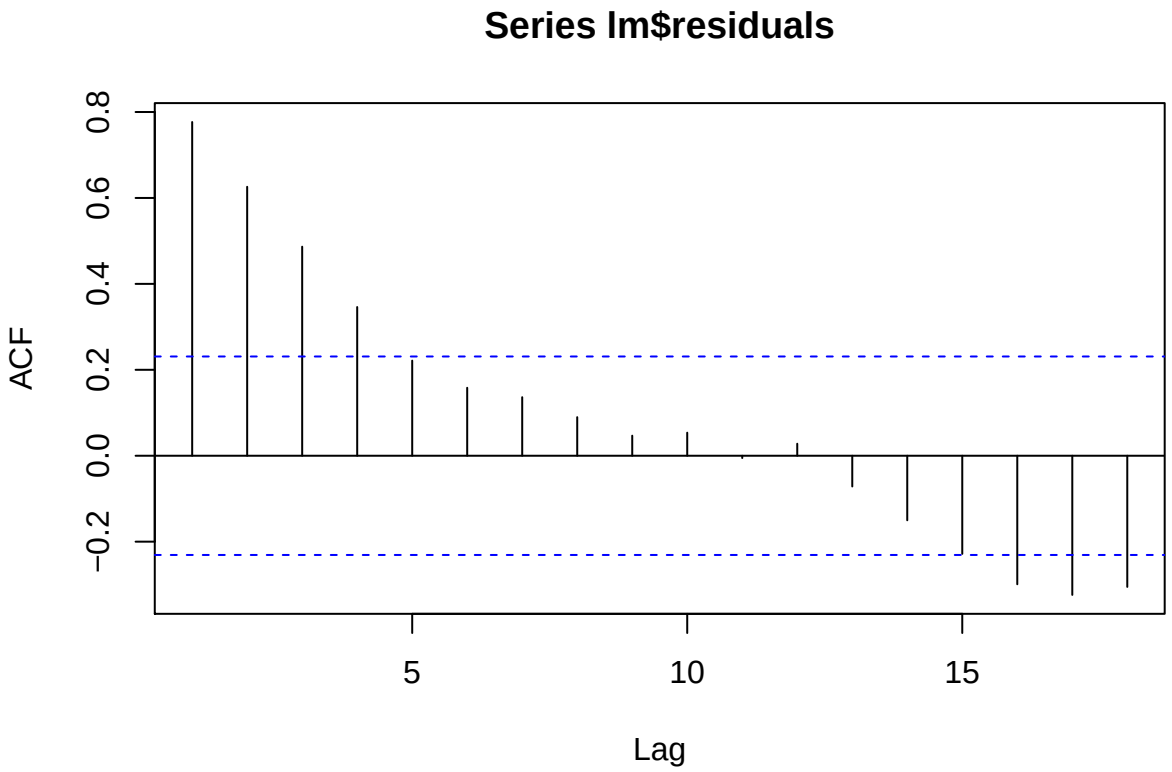
```r
abline(h=0, col="red")
```




```r
qqnorm(lm$residuals, col="blue")
qqline(lm$residuals, col="red")
```




```r
acf(lm$residuals)
```



Although the model summary suggests the linear model is appropriate here, it fails to satisfy the requirement that the residuals are normally distributed. The QQ Plot shows a lot of concavity and skewness toward the tails. The residual plot also has a distinct pattern rather than the points being evenly scattered about zero. Furthermore, the autocorrelation function of residuals shows that there are statistically significant correlations between points separated by various lags. Normally distributed points would not show any significant autocorrelation.

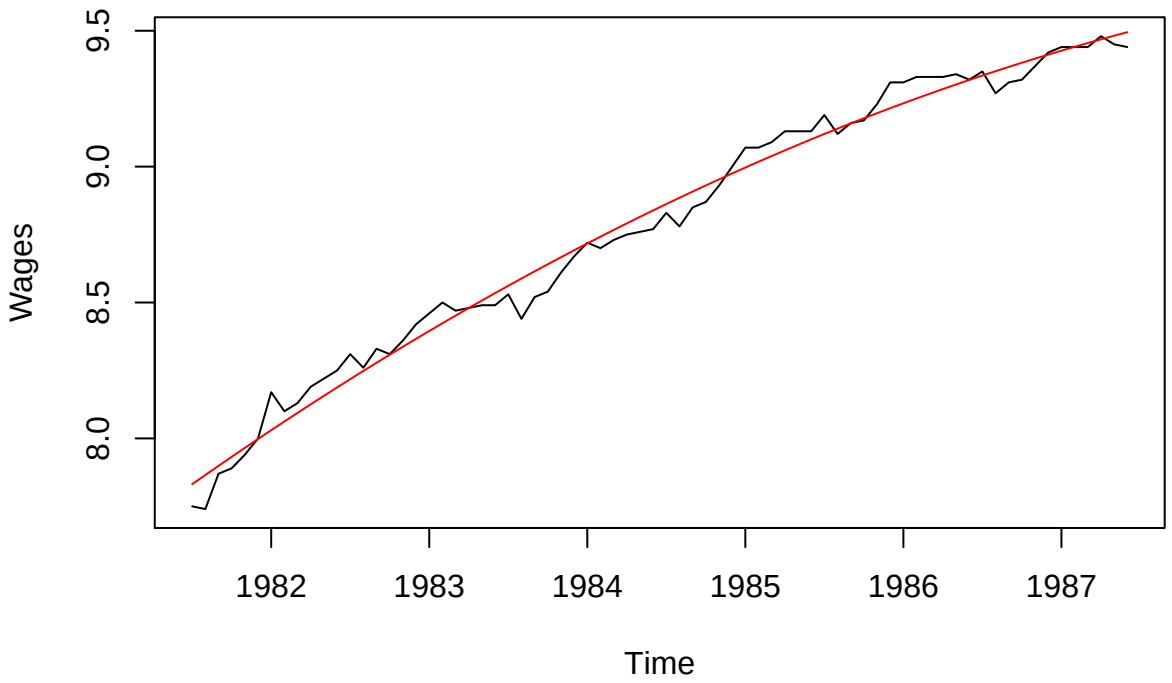
(d) Use least squares to fit a quadratic time trend to the wages time series (i.e yt $= \beta_0 + \beta_1 t + \beta_3 t^2 + e_t$). Interpret the regression output. Save the standardized residuals from the fit for further analysis.

```
squaredTerm <- time(wages)^2
nlm <- lm(formula = wages ~ time(wages) + squaredTerm)
summary(nlm)
```

```
##
## Call:
## lm(formula = wages ~ time(wages) + squaredTerm)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.148318 -0.041440  0.001563  0.050089  0.139839
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept) -8.495e+04  1.019e+04  -8.336 4.87e-12 ***
## time(wages)  8.534e+01  1.027e+01   8.309 5.44e-12 ***
## squaredTerm -2.143e-02  2.588e-03  -8.282 6.10e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05889 on 69 degrees of freedom
```

```
## Multiple R-squared:  0.9864, Adjusted R-squared:  0.986
## F-statistic:  2494 on 2 and 69 DF,  p-value: < 2.2e-16
```

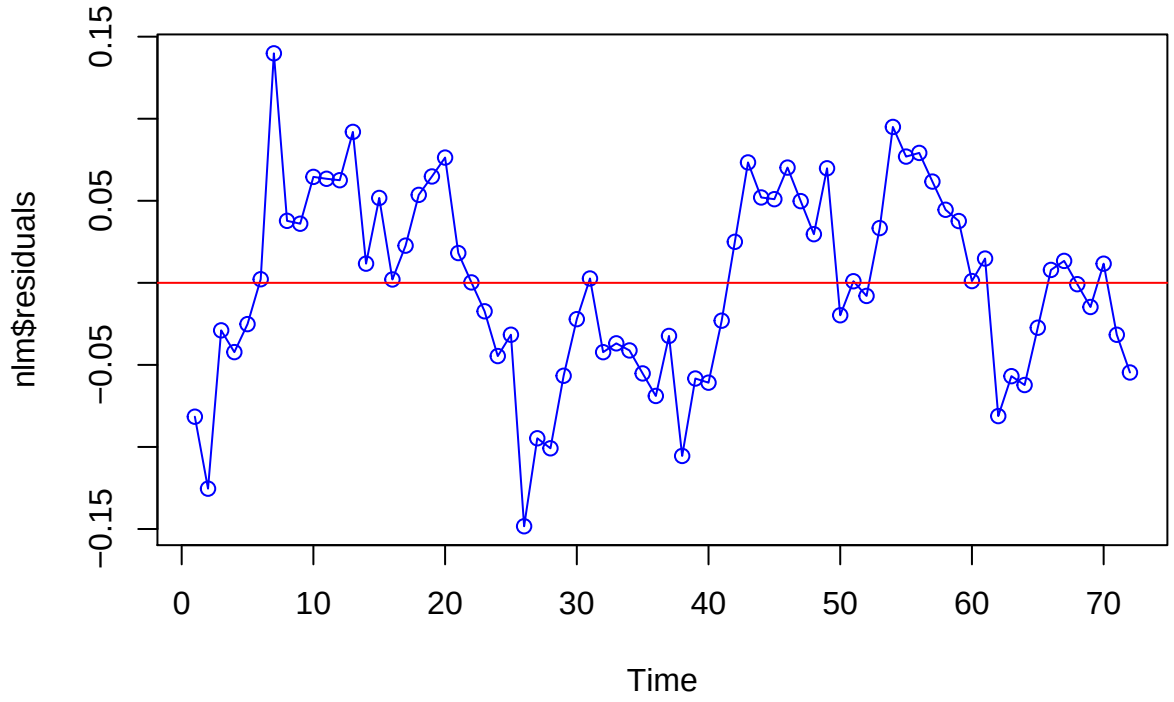
```
plot(wages)
lines(y=predict(nlm),x=time(wages), col="red", type="l")
```



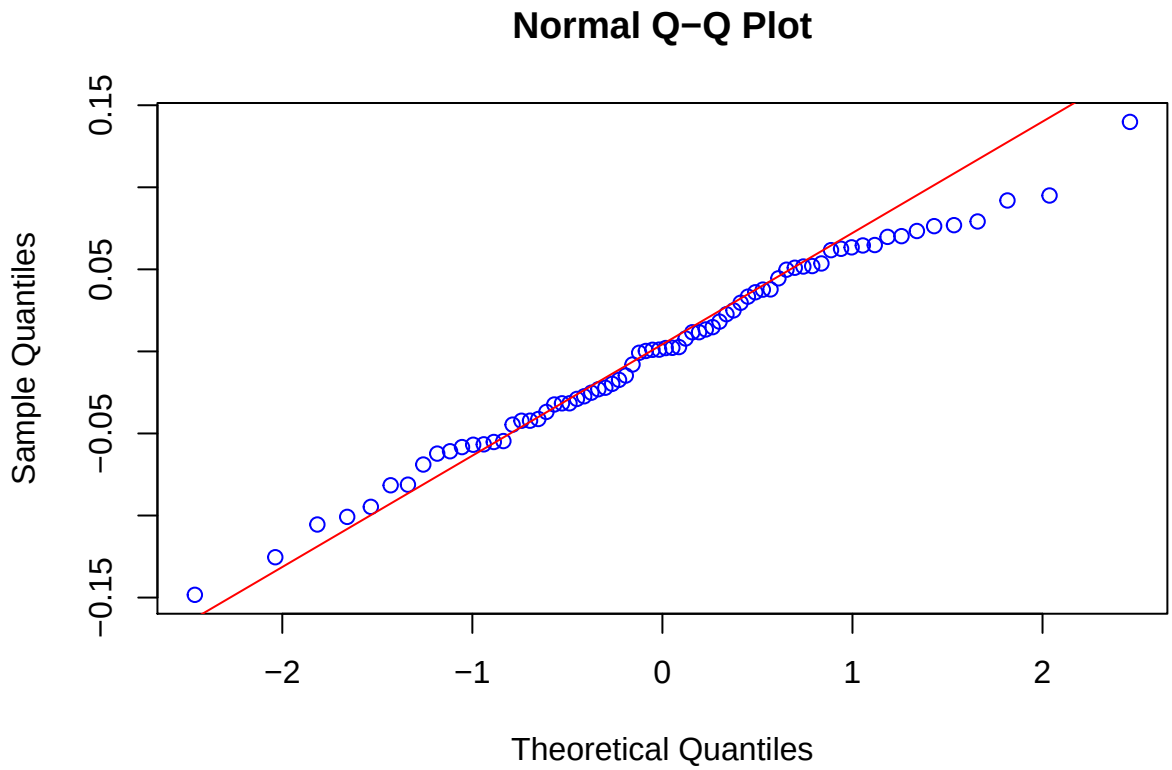
The summary suggests that all of the parameters are statistically significant, and therefore different from zero.

(e) Construct and interpret the time series plot of the standardized residuals from part (d).

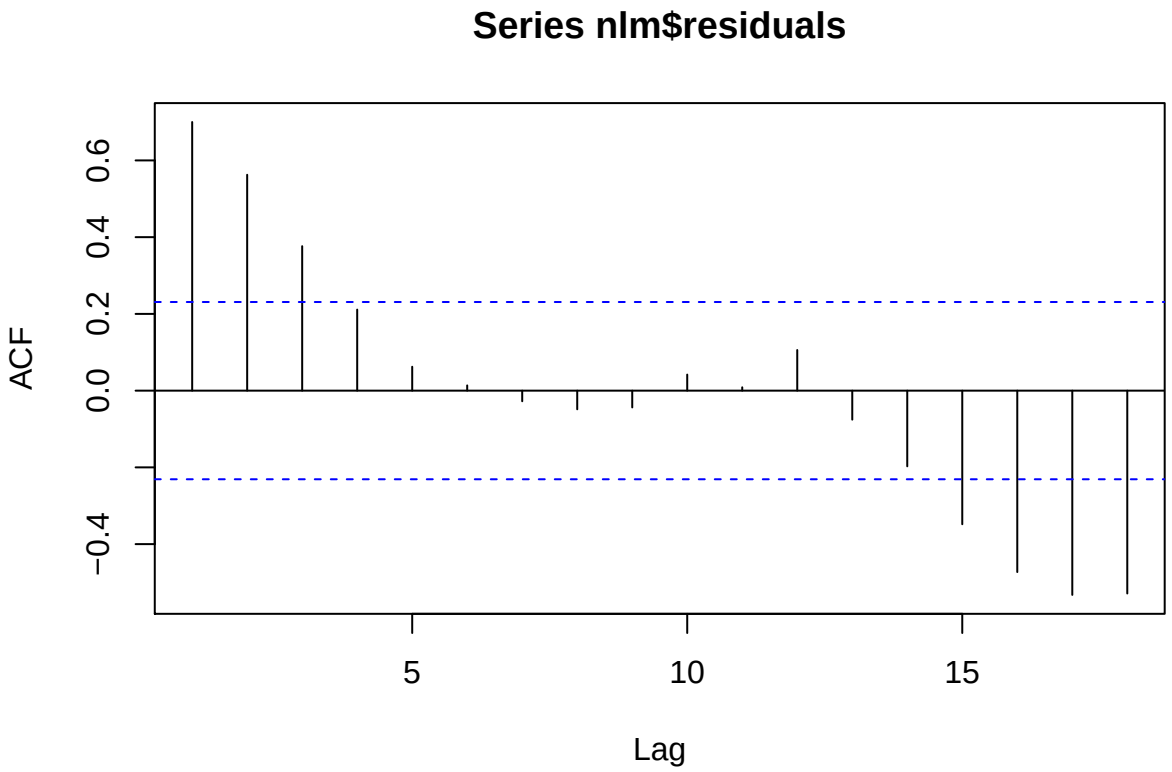
```
plot.ts(nlm$residuals, col="blue")
points(nlm$residuals, col="blue")
abline(h=0, col="red")
```




```
qqnorm(nlm$residuals, col="blue")
qqline(nlm$residuals, col="red")
```




```
acf(nlm$residuals)
```



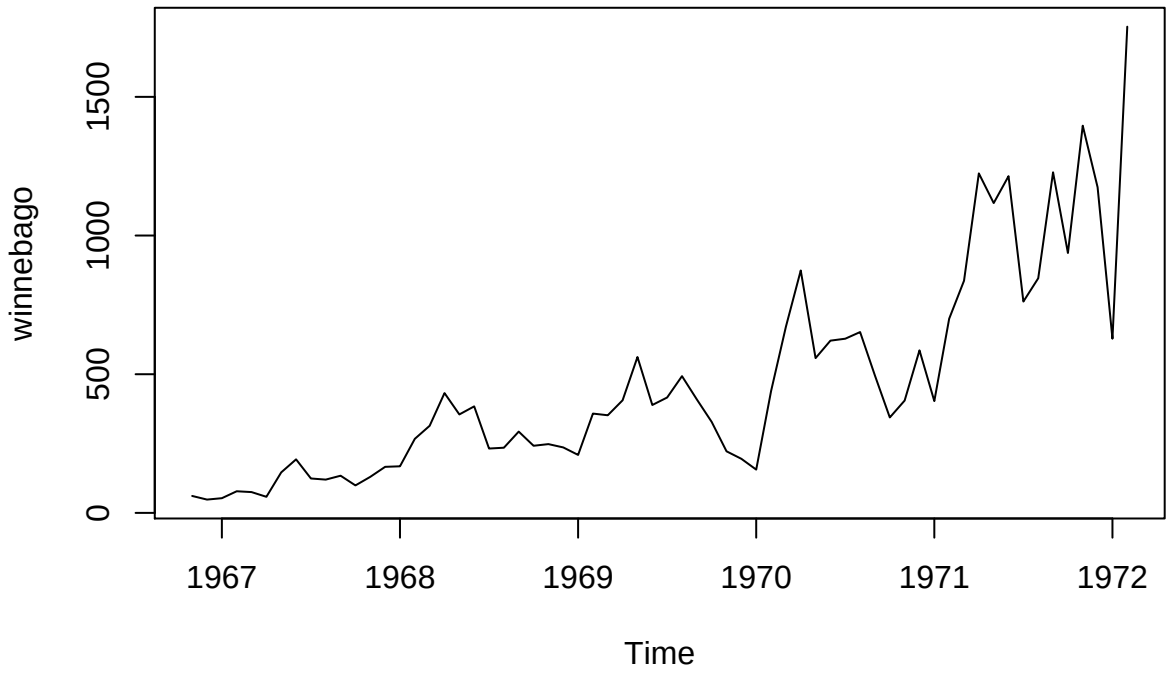
Although the model summary suggests the quadratic model is appropriate here, like the linear model, it fails to satisfy the requirement that the residuals are normally distributed. The QQ Plot shows a lot of concavity and skewness toward the tails. The residual plot also has a distinct pattern rather than the points being evenly scattered about zero. Lastly, the ACF shows that there is correlation between data points of various lags, also implying the residuals are not normally distributed.

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

```
data("winnebago")
```

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

```
plot(winnebago)
```



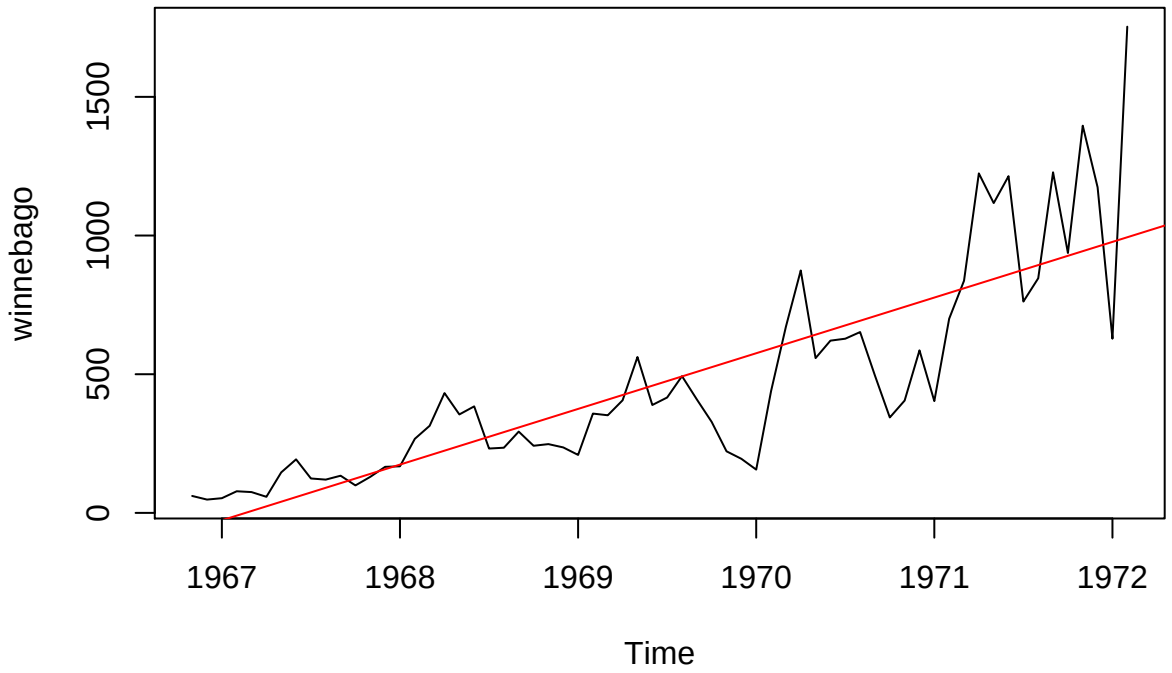
The monthly unit sales of recreational vehicles shows an upward trend with increasing variance from 1966 to 1972. There is also seasonality year to year.

(b) Use least squares to fit a line to these data. Interpret the regression output. Plot the standardized residuals from the fit as a time series. Interpret the plot.

```
lm <- lm(winnebago ~ time(winnebago))
summary(lm)
```

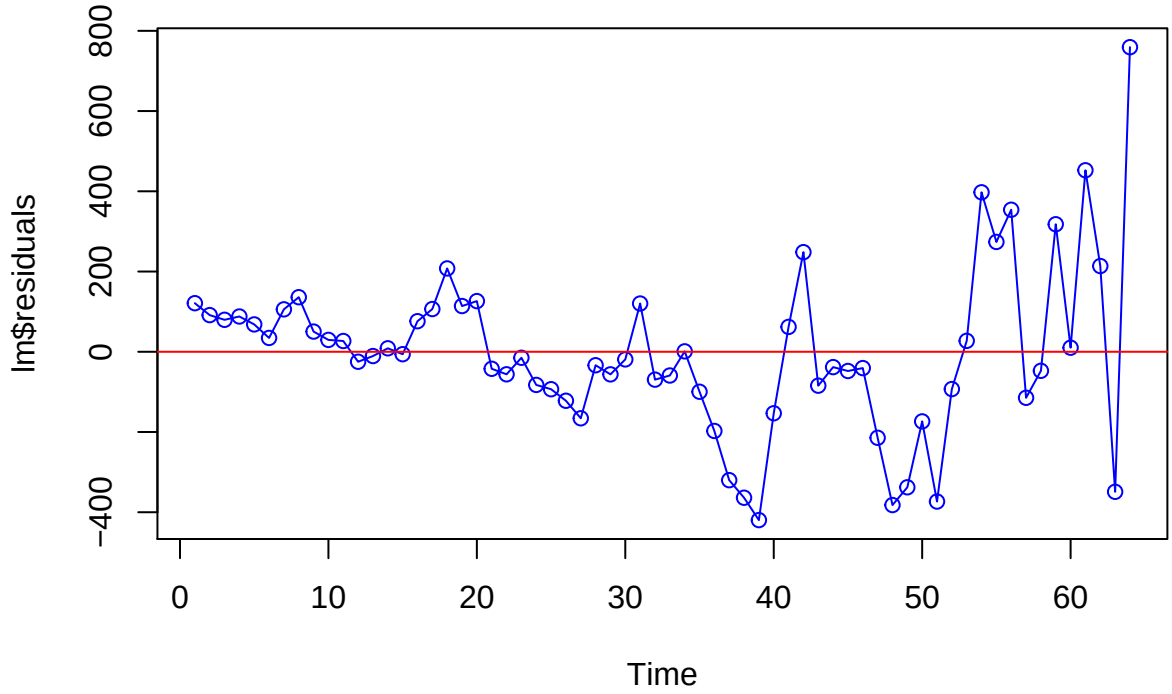
```
##
## Call:
## lm(formula = winnebago ~ time(winnebago))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -419.58  -93.13  -12.78   94.96  759.21
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -394885.68   33539.77  -11.77   <2e-16 ***
## time(winnebago)     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(winnebago)
abline(lm, col="red")
```

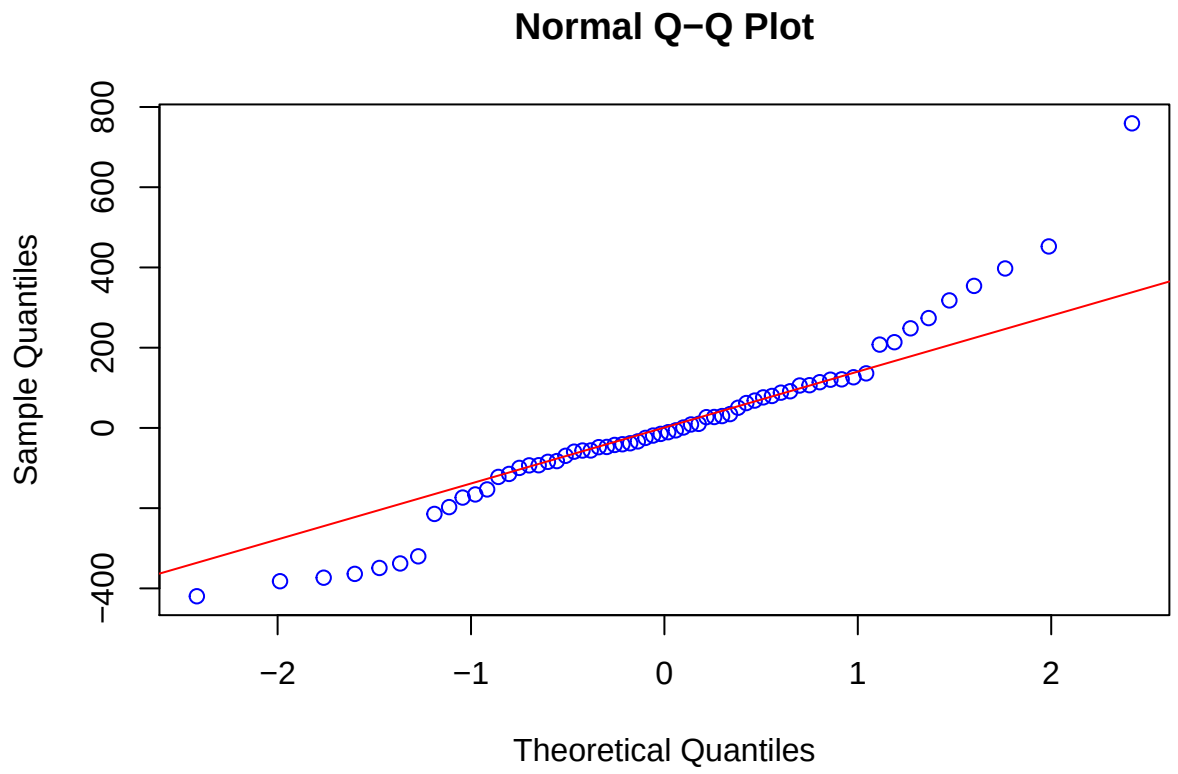


The summary of the linear model suggests that both the intercept and slope values are very statistically significant in modeling these data, meaning these parameters are different from zero. The F Test also has a very small p-value, meaning that both the parameters in the model are different from zero holding the other one constant.

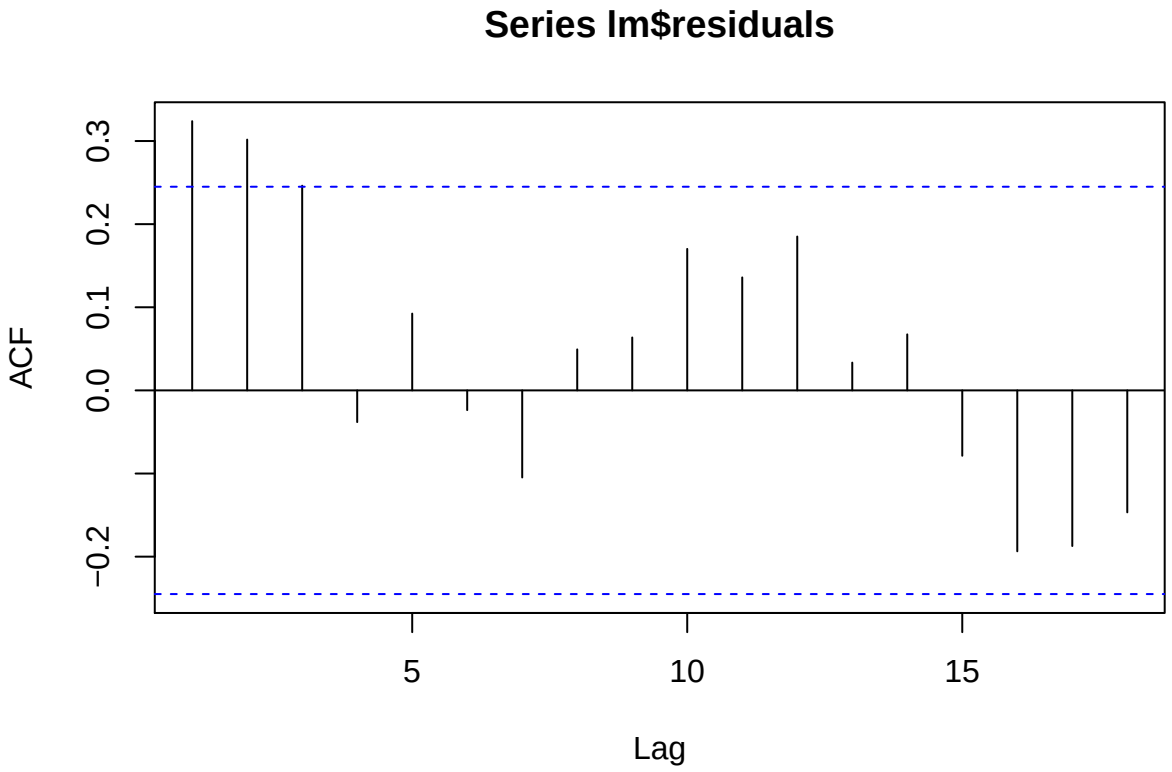
```
plot.ts(lm$residuals, col="blue")
points(lm$residuals, col="blue")
abline(h=0, col="red")
```




```
qqnorm(lm$residuals, col="blue")
qqline(lm$residuals, col="red")
```




```
acf(lm$residuals)
```



The residual plot of this linear model shows heteroscedasticity as the variance increases dramatically with time. This finding, along with the skewed tails on the QQ Plot suggest that the error term is not normally distributed, making this a poor model to use. The ACF is promising, however these data are highly correlated

for a lag of 1 and 2.

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

```
log_winnebago <- log(winnebago)
plot(log_winnebago)
```

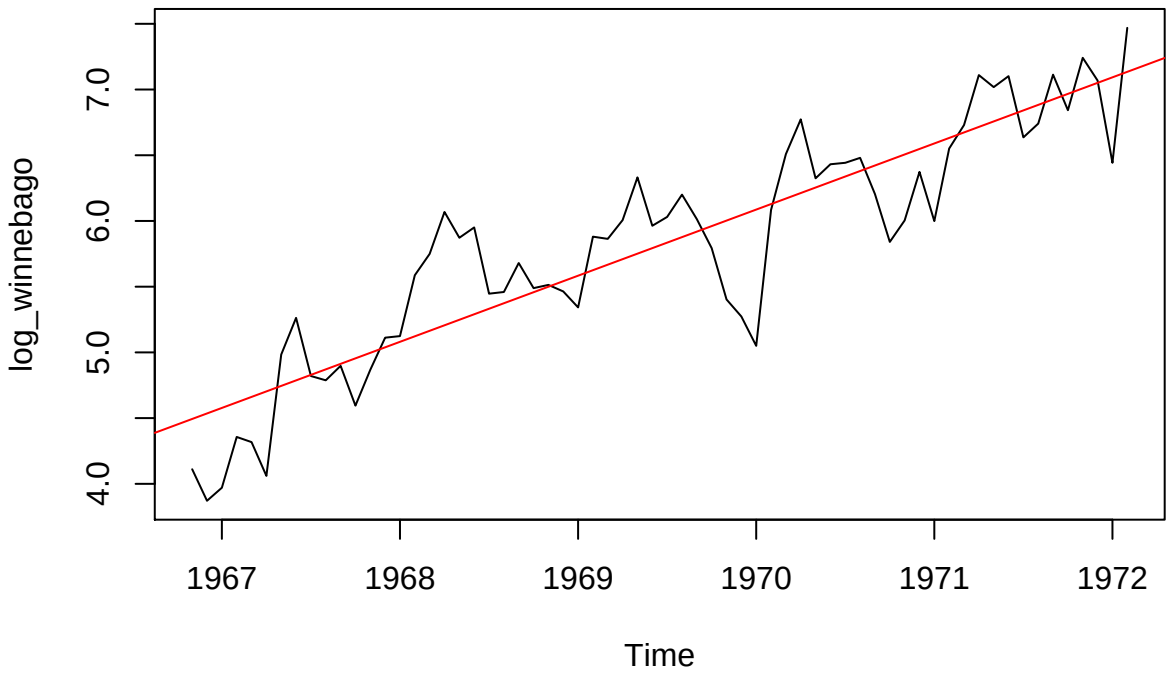
The log of the vehical unit sales shows an upward trend with seasonality, but now with constant variance.

(d) Use least squares to fit a line to the logged data. Display and interpret the time series plot of the standardized residuals from this fit.

```
lm <- lm(log_winnebago ~ time(winnebago))
summary(lm)
```

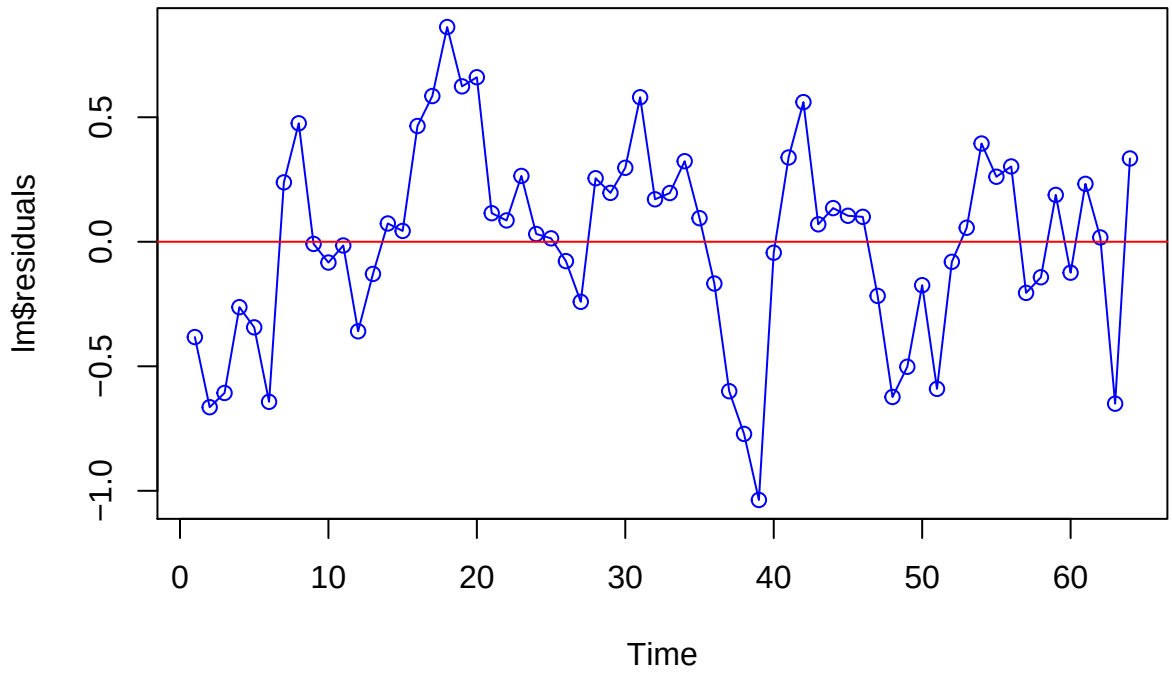
```
##
## Call:
## lm(formula = log_winnebago ~ time(winnebago))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.03669 -0.20823  0.04995  0.25662  0.86223
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -984.93878   62.99472  -15.63   <2e-16 ***
## time(winnebago)    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.ts(log_winnebago)
abline(lm, col="red")
```

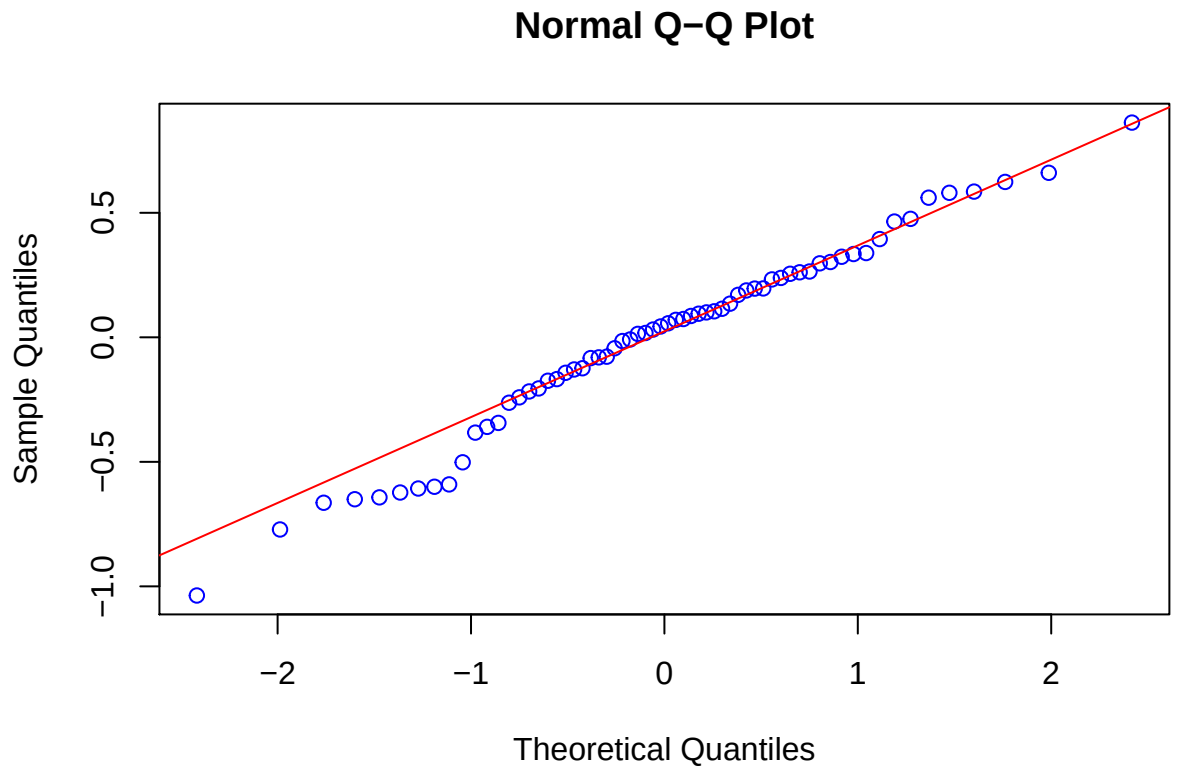


Again, the summary of the linear model suggests that both the model parameters are very statistically significant in modeling these data, meaning these parameters are different from zero. The F Test also has a very small p-value, meaning that both the parameters in the model are different from zero holding the other one constant.

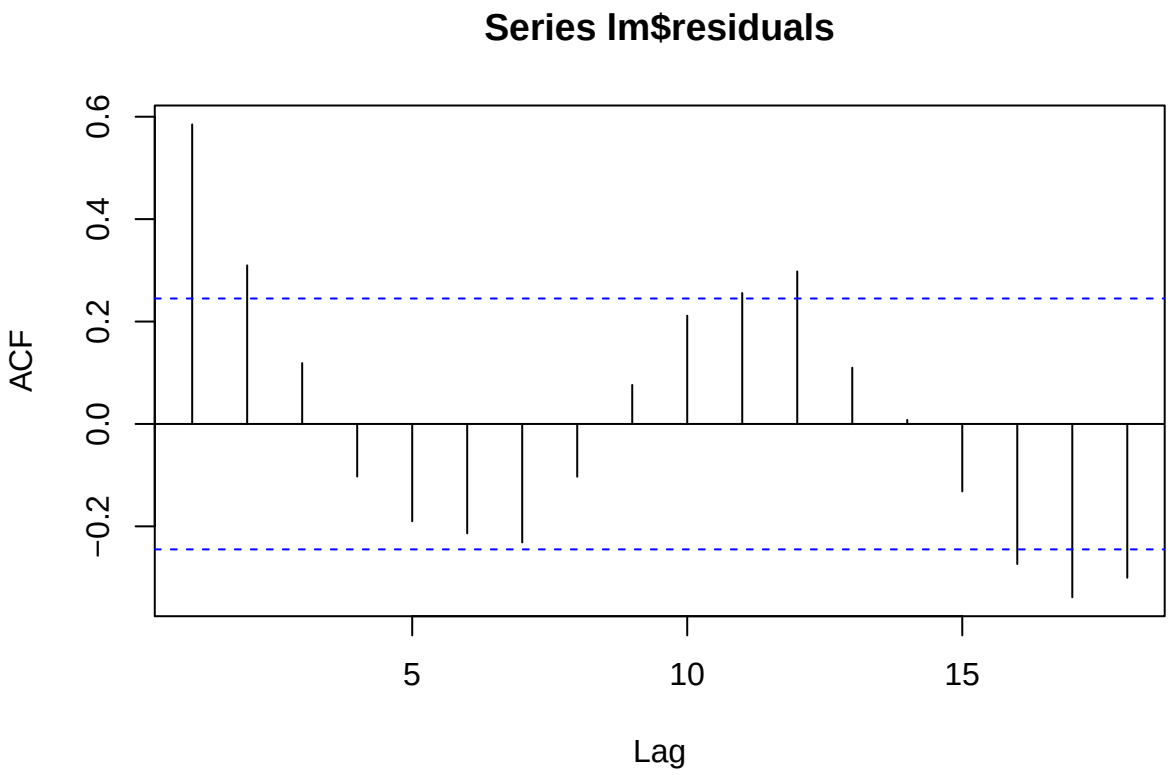
```
plot.ts(lm$residuals, col="blue")
points(lm$residuals, col="blue")
abline(h=0, col="red")
```



```r
qqnorm(lm$residuals, col="blue")
qqline(lm$residuals, col="red")
```




```r
acf(lm$residuals)
```



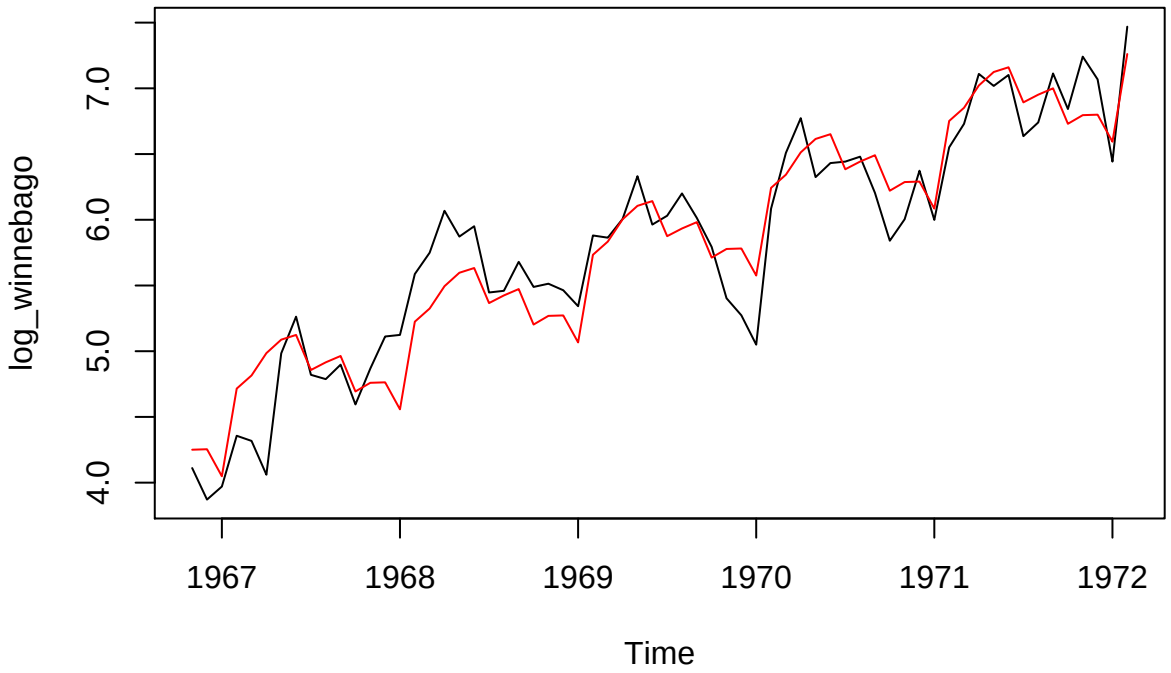
While the model summary suggests that the linear model is a good fit, the residual plot, while better than the residual plot for the untransformed data, still shows an uneven distribution about zero. Furthermore, the QQ Plot shows a lot of skewness on the left tail. For these reasons, a linear model is still not a good fit. Like the ACF for the untransformed data, there is high correlation for a lag of 1, also suggesting that the error term is not normally distributed.

(e) Now use least squares to fit a seasonal-means plus linear time trend to the logged sales time series and save the standardized residuals for further analysis. Check the statistical significance of each of the regression coefficients in the model.

```
nls <- lm(log_winnebago ~ time(log_winnebago) + season(log_winnebago))
summary(nls)
```

```
##
## Call:
## lm(formula = log_winnebago ~ time(log_winnebago) + season(log_winnebago))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.92501 -0.16328  0.03344  0.20757  0.57388
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   -997.33061   50.63995 -19.695  < 2e-16 ***
## time(log_winnebago)              0.50909    0.02571  19.800  < 2e-16 ***
## season(log_winnebago)February    0.62445    0.18182   3.434 0.001188 **
## season(log_winnebago)March       0.68220    0.19088   3.574 0.000779 ***
## season(log_winnebago)April       0.80959    0.19079   4.243 9.30e-05 ***
## season(log_winnebago)May         0.86953    0.19073   4.559 3.25e-05 ***
## season(log_winnebago)June        0.86309    0.19070   4.526 3.63e-05 ***
## season(log_winnebago)July        0.55392    0.19069   2.905 0.005420 **
## season(log_winnebago)August      0.56989    0.19070   2.988 0.004305 **
## season(log_winnebago)September   0.57572    0.19073   3.018 0.003960 **
## season(log_winnebago)October     0.26349    0.19079   1.381 0.173300
## season(log_winnebago)November    0.28682    0.18186   1.577 0.120946
## season(log_winnebago)December    0.24802    0.18182   1.364 0.178532
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3149 on 51 degrees of freedom
## Multiple R-squared:  0.8946, Adjusted R-squared:  0.8699
## F-statistic: 36.09 on 12 and 51 DF,  p-value: < 2.2e-16
```

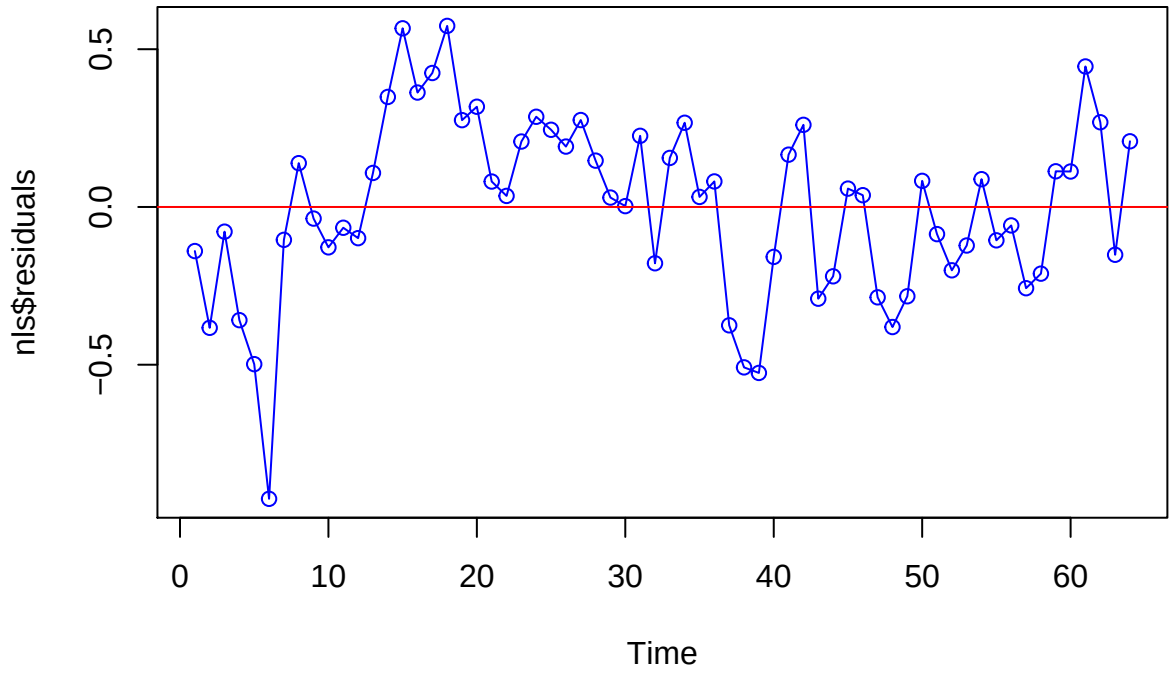
```
plot(log_winnebago)
lines(y=predict(nls), x=time(log_winnebago), type="l", col="red")
```



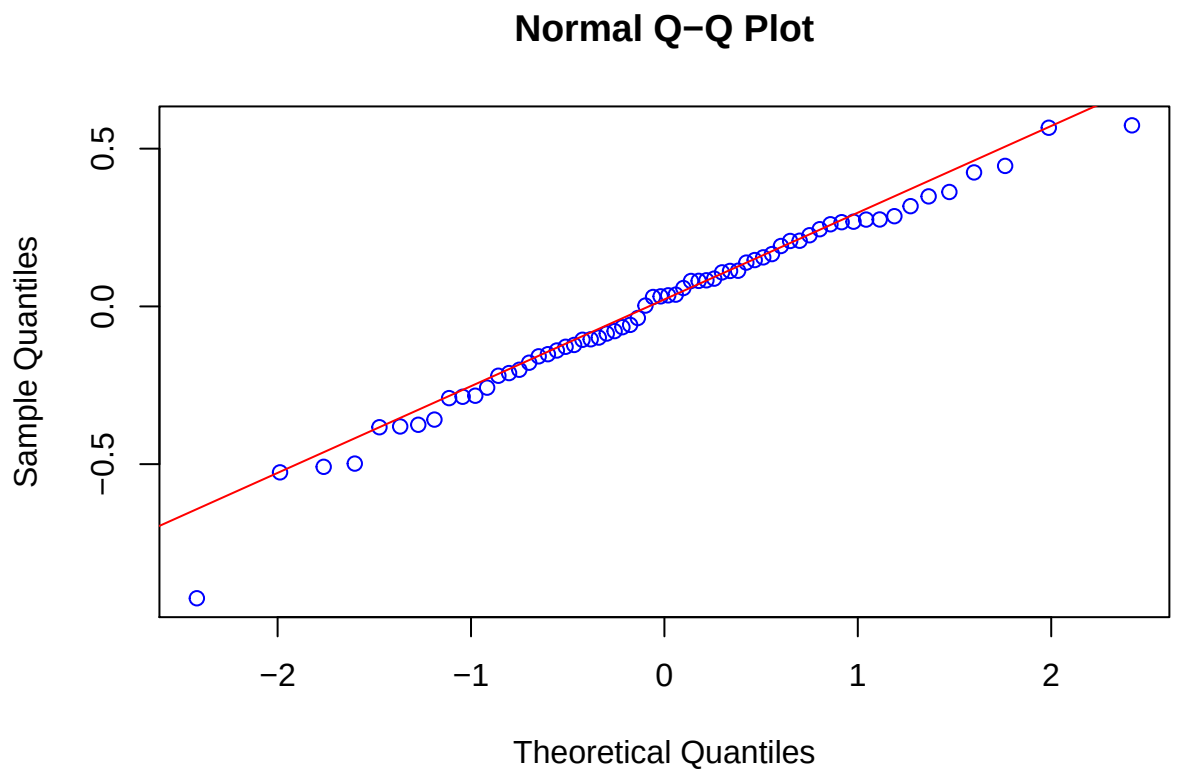
All of the parameters in the seasonal-means plus linear time trend model are very statistically significant, meaning they are different from zero.

(f) Display the time series plot of the standardized residuals obtained in part (e). Interpret the plot.

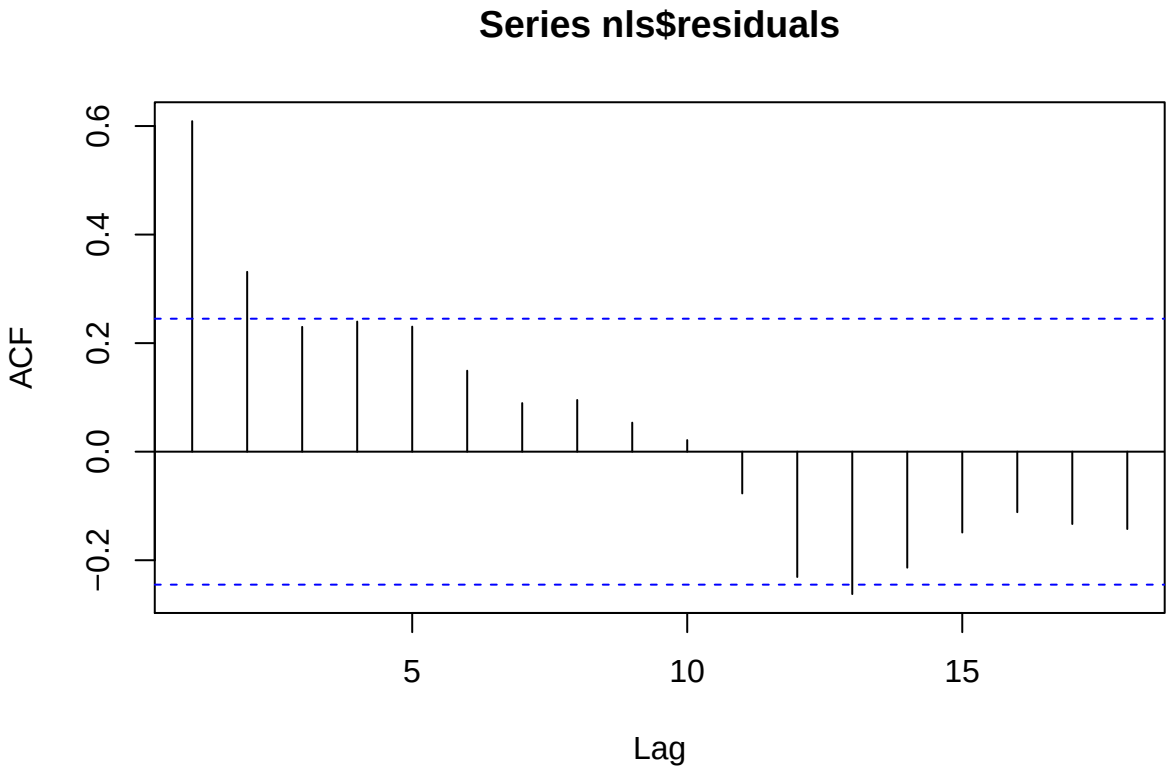
```
plot.ts(nls$residuals, col="blue")
points(nls$residuals, col="blue")
abline(h=0, col = "red")
```




```
qqnorm(nls$residuals, col="blue")
qqline(nls$residuals, col="red")
```




```
acf(nls$residuals)
```



The residual plot is not evenly distributed about zero, the qqplot shows some concavity towards the tails, and the acf shows correlation with the first two lags. All of these findings suggest the error term is not normally distributed, making this a poor model to use.

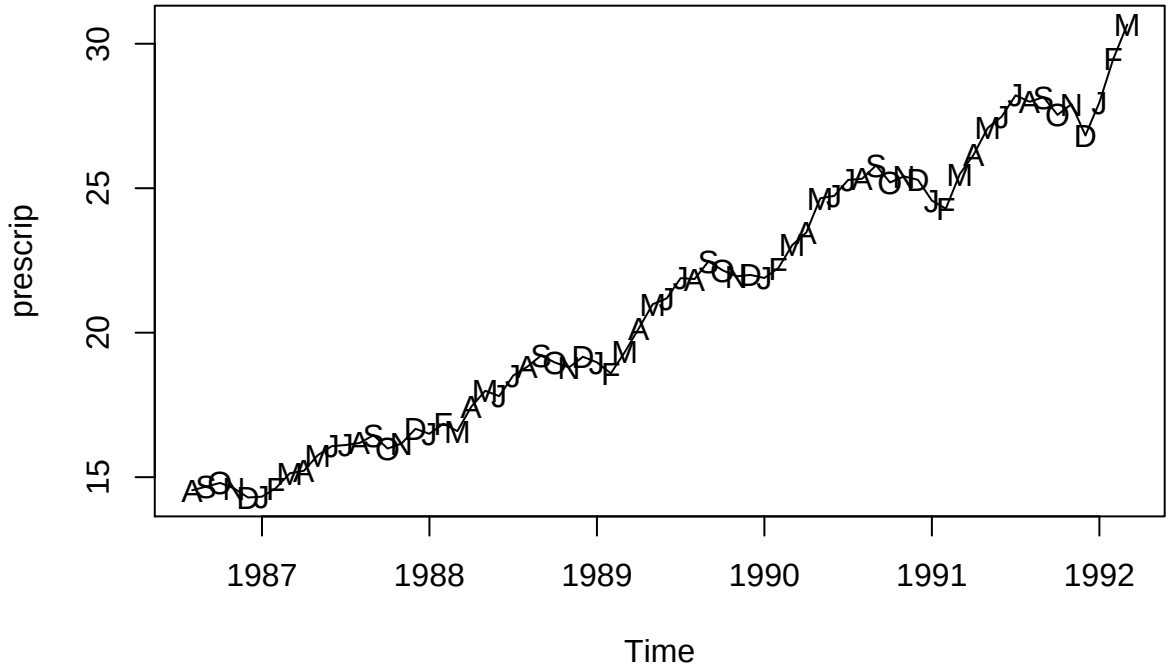
3)

The data file prescrip gives monthly U.S. prescription costs for the months August 1986 to March 1992. These data are from the State of New Jerseys Prescription Drug Program and are the cost per prescription claim.

```
data("prescrip")
```

(a) Display and interpret the time series plot for these data. Use plotting symbols that permit you to look for seasonality.

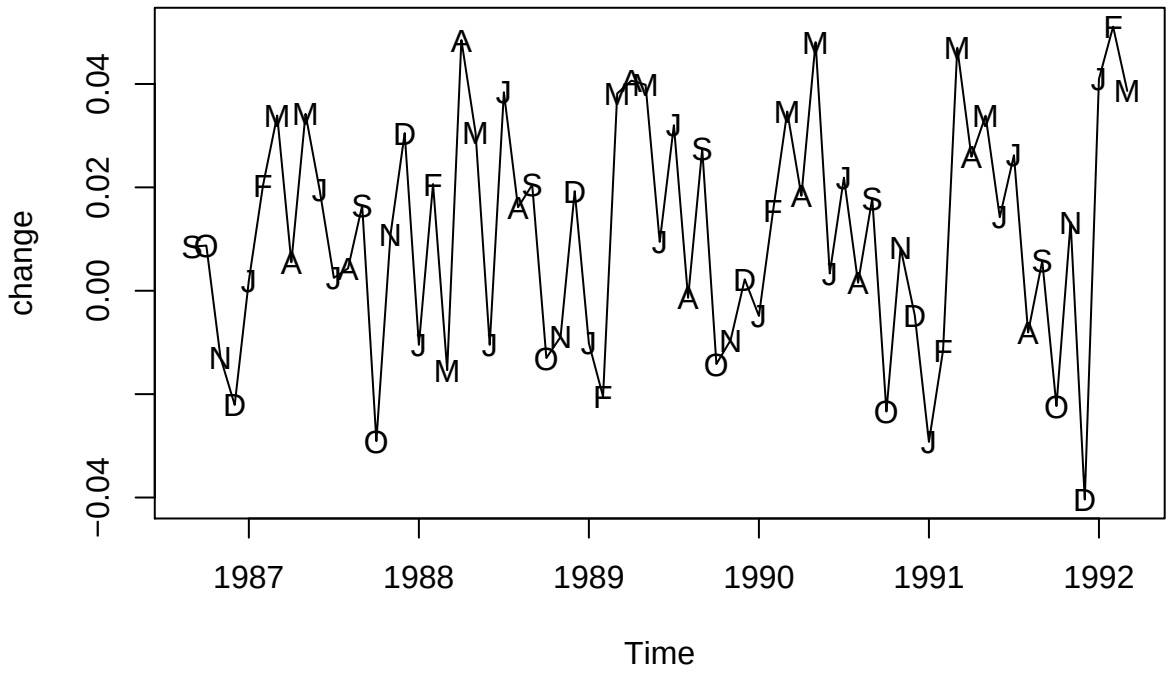
```
plot(prescrip)
points(y=prescrip, x=time(prescrip), pch=as.vector(season(prescrip)))
```



The monthly U.S. perscription costs shows an upward trend with increasing variance from 1986 to 1992. There is also seasonality year to year.

(b) Calculate and plot the sequence of month-to-month percentage changes in the prescription costs. Again, use plotting symbols that permit you to look for seasonality.

```
change <- diff(prescrip)/prescrip
plot(change)
points(y=change, x=time(change), pch=as.vector(season(change)))
```



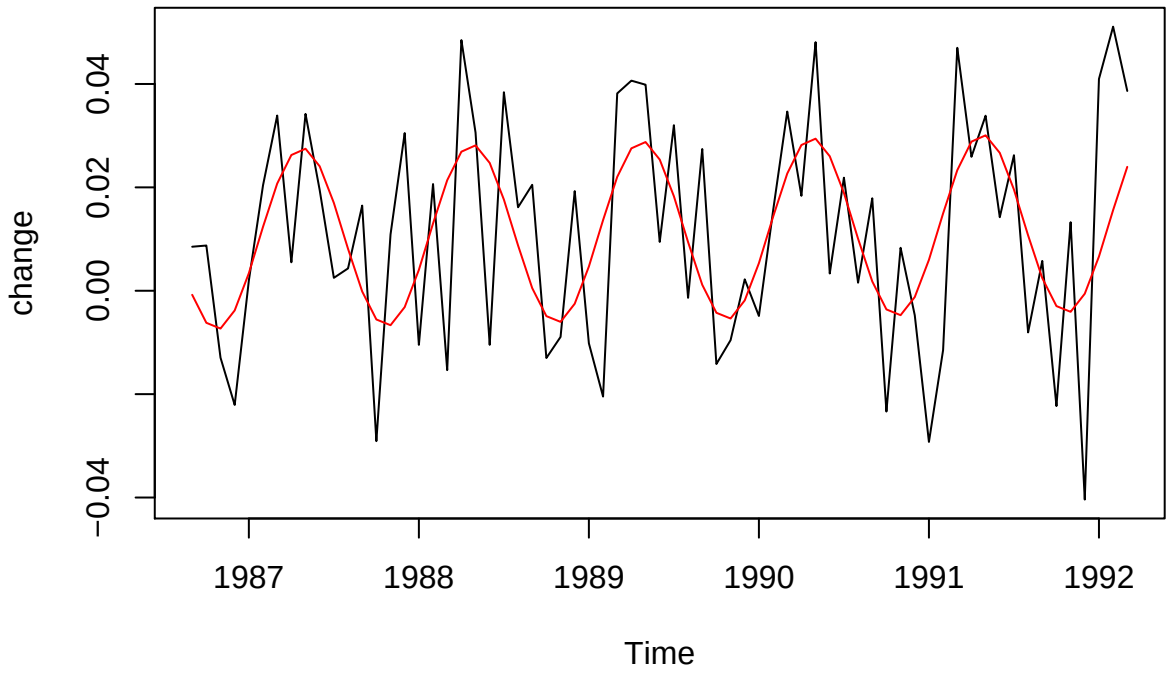
This plot clearly depicts seasonality, as months such as March, April, and May consistently appear at the peaks, and the months October and December are often at the troughs.

(c) Use least squares to fit a cosine trend with fundamental frequency 1/12 to the percentage change series. Interpret the regression output. Save the standardized residuals.

```
nlm <- lm(change ~ time(change) + harmonic(change))
summary(nlm)
```

```
##
## Call:
## lm(formula = change ~ time(change) + harmonic(change))
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.039822 -0.013468  0.002466  0.014213  0.035480
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  -1.2759708  2.8282982  -0.451   0.6534
## time(change)                  0.0006472  0.0014217   0.455   0.6505
## harmonic(change)cos(2*pi*t) -0.0066376  0.0032580  -2.037   0.0458 *
## harmonic(change)sin(2*pi*t)  0.0160672  0.0032302   4.974 5.35e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01874 on 63 degrees of freedom
## Multiple R-squared:  0.3149, Adjusted R-squared:  0.2823
## F-statistic: 9.652 on 3 and 63 DF,  p-value: 2.489e-05
```

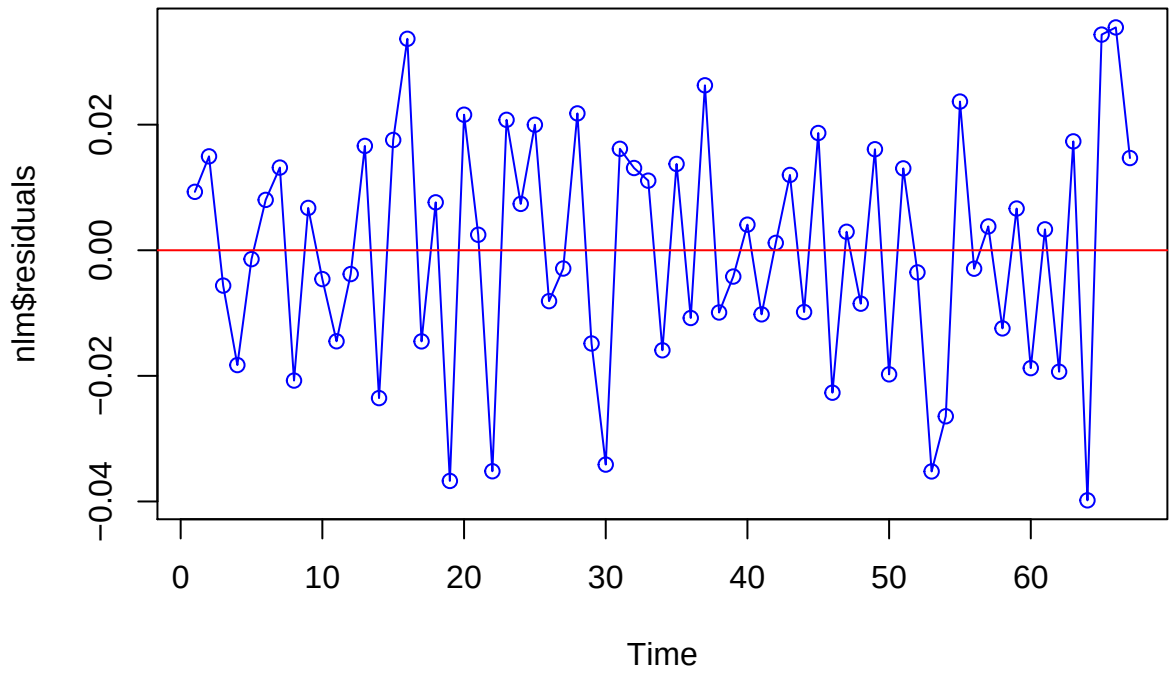
```
plot(change)
lines(y=predict(nlm), x=time(change), type="l", col="red")
```



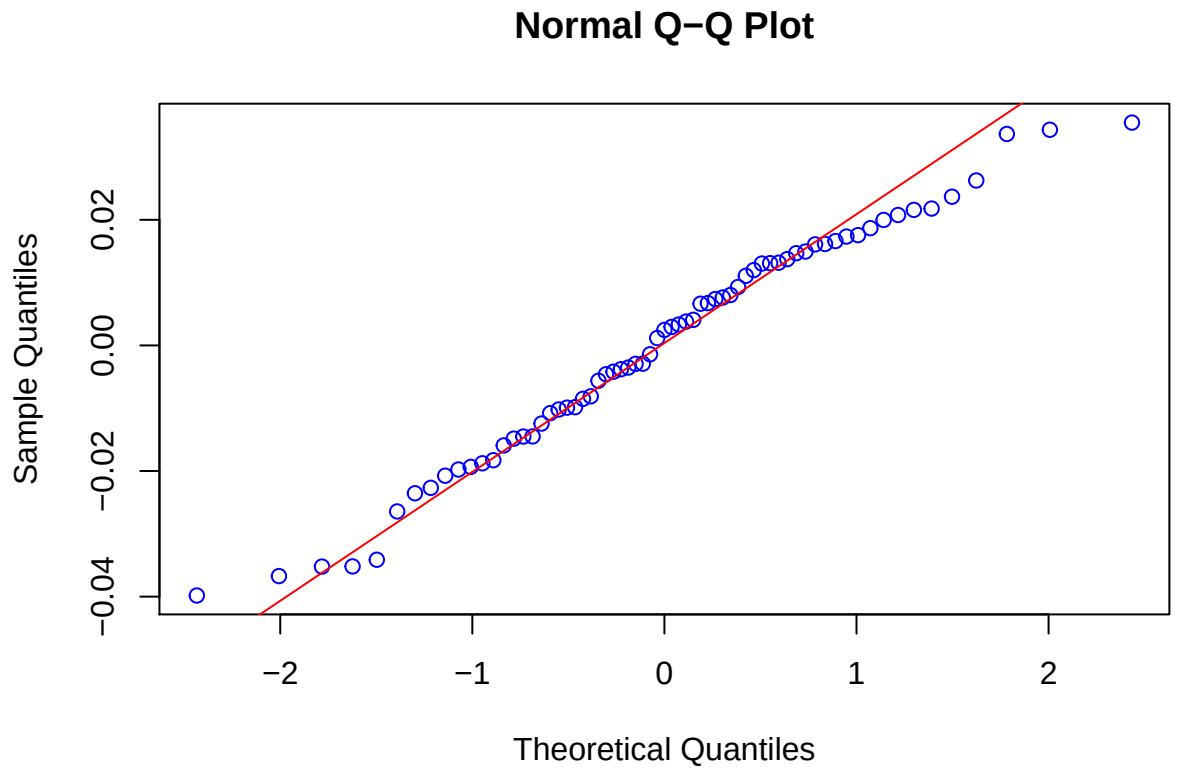
Only the sine term is very statistically significant, while p-value for the cosine term is a moderate .0458. The intercept and slope terms are not statistically significant, and we therefore cannot reject the null hypothesis that they are zero. However, the F Test suggests that the terms are different from zero when all others are held constant. Still, the R-squared value is a meager .3149, meaning a significant amount of information is lost in this model, making it a very poor choice.

(d) Plot the sequence of standardized residuals to investigate the adequacy of the cosine trend model. Interpret the plot.

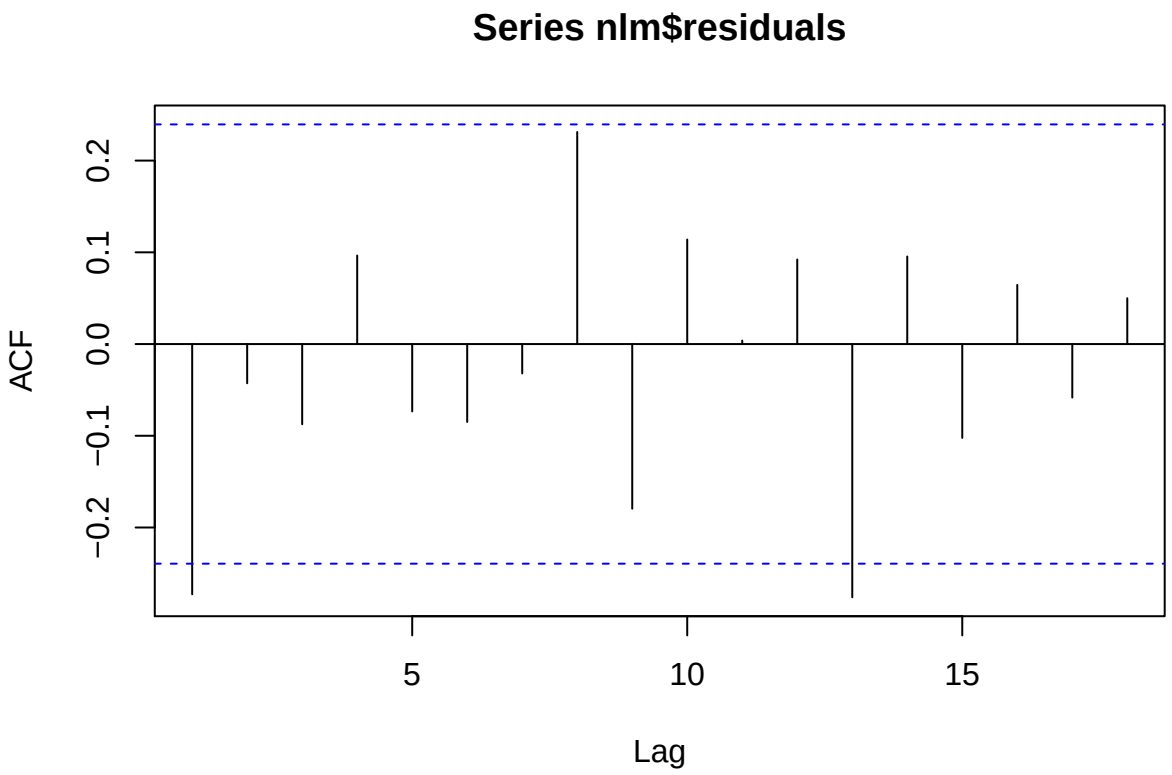
```
plot.ts(nlm$residuals, col="blue")
points(nlm$residuals, col="blue")
abline(h=0, col="red")
```



```
qqnorm(nlm$residuals, col="blue")
qqline(nlm$residuals, col="red")
```




```
acf(nlm$residuals)
```



The QQ plot has concavity on the tails, however, some of these points appear to be outliers. The residual plot is well distributed about 0 and the acf suggests that there is no autocorrelation. These findings suggest the error term is likely normally distributed, but the terrible R-squared value means this normality is likely due to the model consistently failing to capture any of the volitility in these data, and merely sticking close to the mean.