Guo - HWK6.2

Andrew Guo

11/11/2021

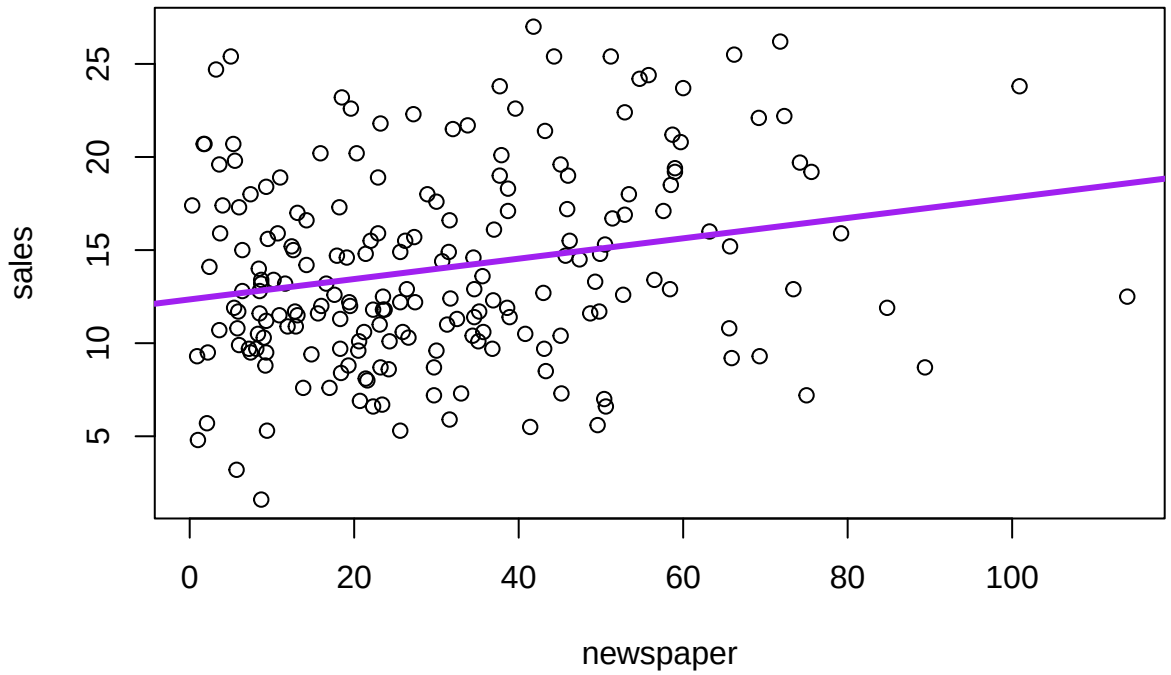
```
## Load libraries
library(lawstat)

## Read in the data
Advertising <- read.csv("Advertising.csv")
attach(Advertising)

#Part 1: Advertising and Newspaper
lm.fit.Newspaper = lm(sales~newspaper, data = Advertising)
summary(lm.fit.Newspaper)
```

```
##
## Call:
## lm(formula = sales ~ newspaper, data = Advertising)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.2272  -3.3873  -0.8392   3.5059  12.7751
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.35141    0.62142   19.88  < 2e-16 ***
## newspaper    0.05469    0.01658    3.30  0.00115 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.092 on 198 degrees of freedom
## Multiple R-squared:  0.05212,    Adjusted R-squared:  0.04733
## F-statistic: 10.89 on 1 and 198 DF,  p-value: 0.001148
```

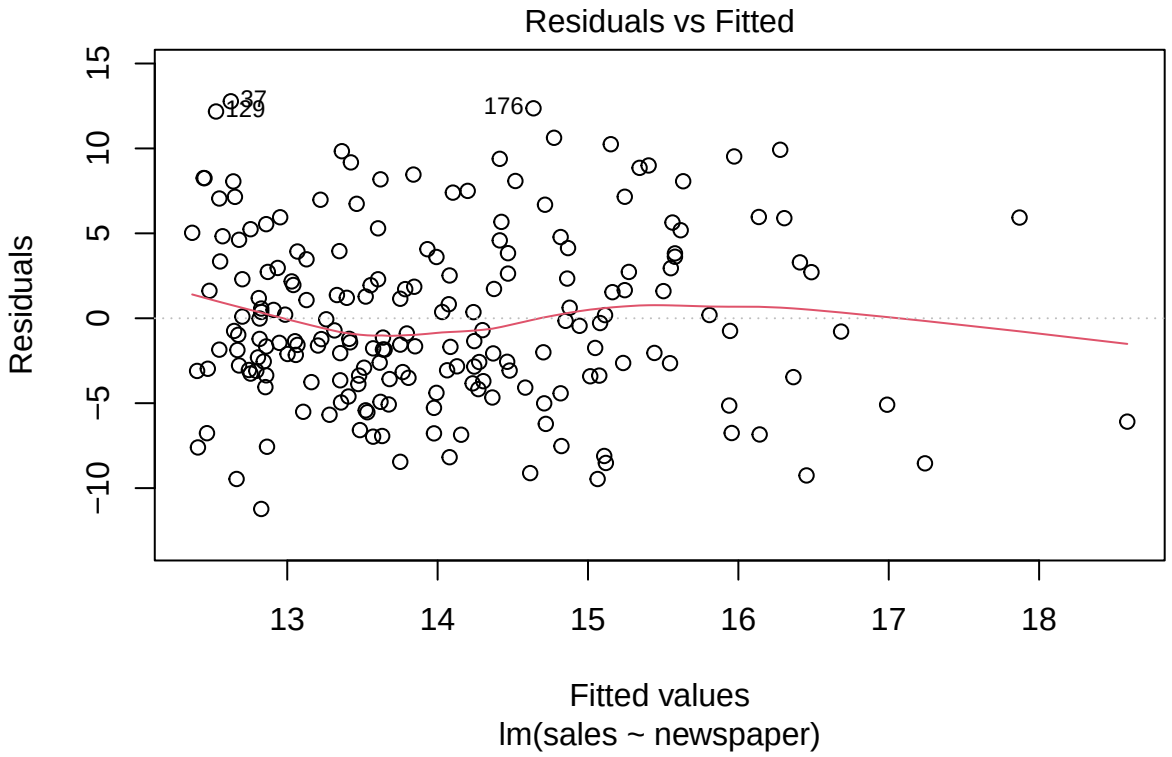
```
## Scatter plot along with linear fit
par(mfrow=c(1,1))
plot(newspaper,sales)
abline(lm.fit.Newspaper, col='purple', lwd=3)
```



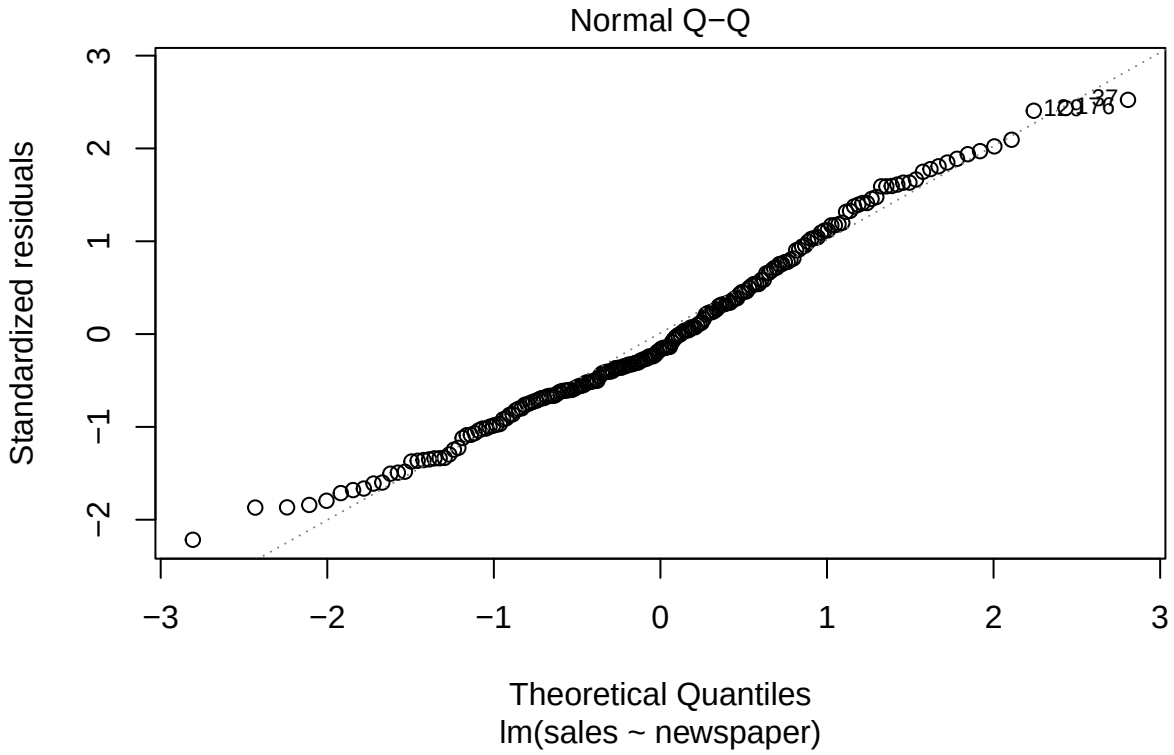
Based on the summary statistics, we observe the following: We have an LSRL of y = 0.05469x + 12.35141. We observe a p-value of < 0.05 and an observable trend that as `newspaper` advertising expenditure increases, `sales` also increases, we can conclude that 'newspaper' and 'sales' have a linear relationship.

However, with an adjusted R-squared value of 0.04733, this means that only about 4.7% of the variation in sales is explained by the LSRL on newspaper This places great skepticism as to how accurate a linear model for comparing these two variables can be. Thus, further diagnostics will be needed.

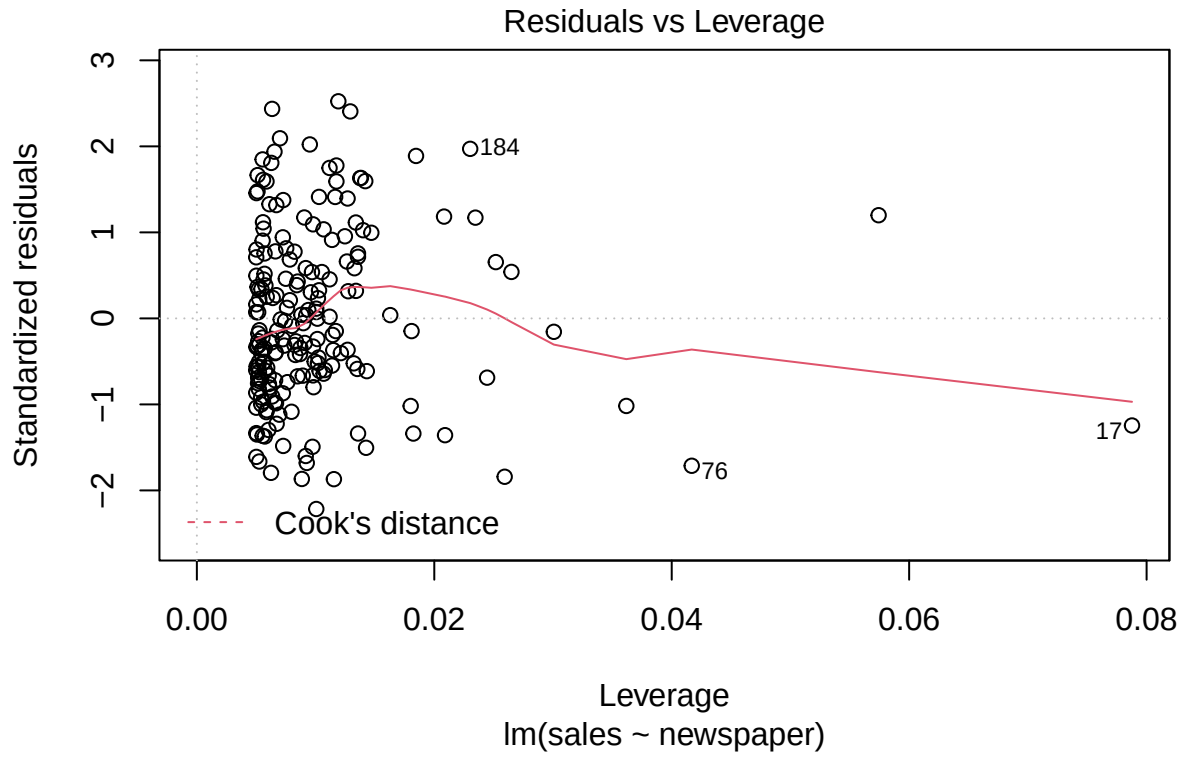
```
## Diagnostic plots
plot(lm.fit.Newspaper, which = c(1))
```




```
plot(lm.fit.Newspaper, which = c(2))
```

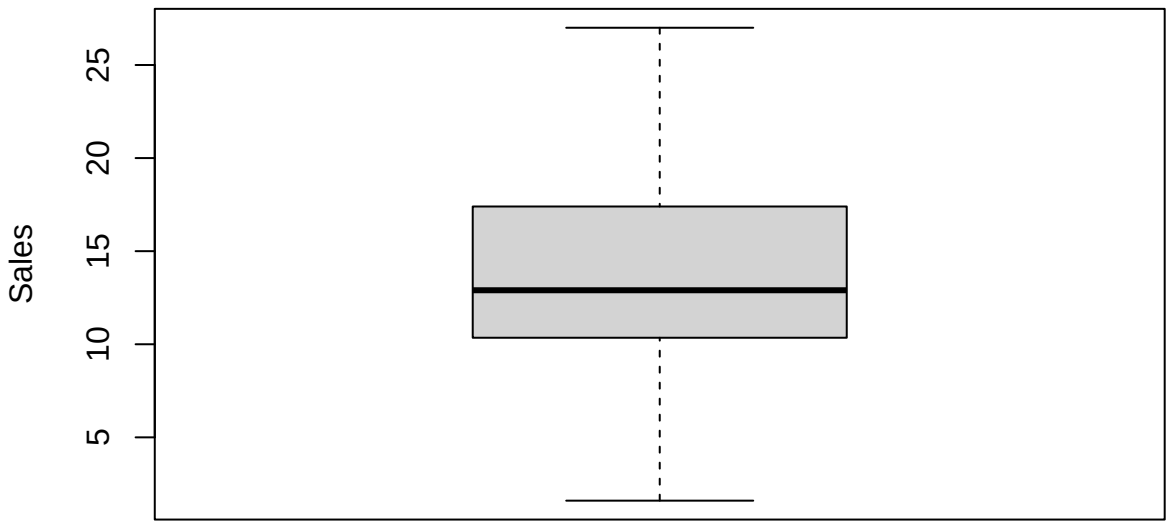



```
plot(lm.fit.Newspaper, which = c(5))
```

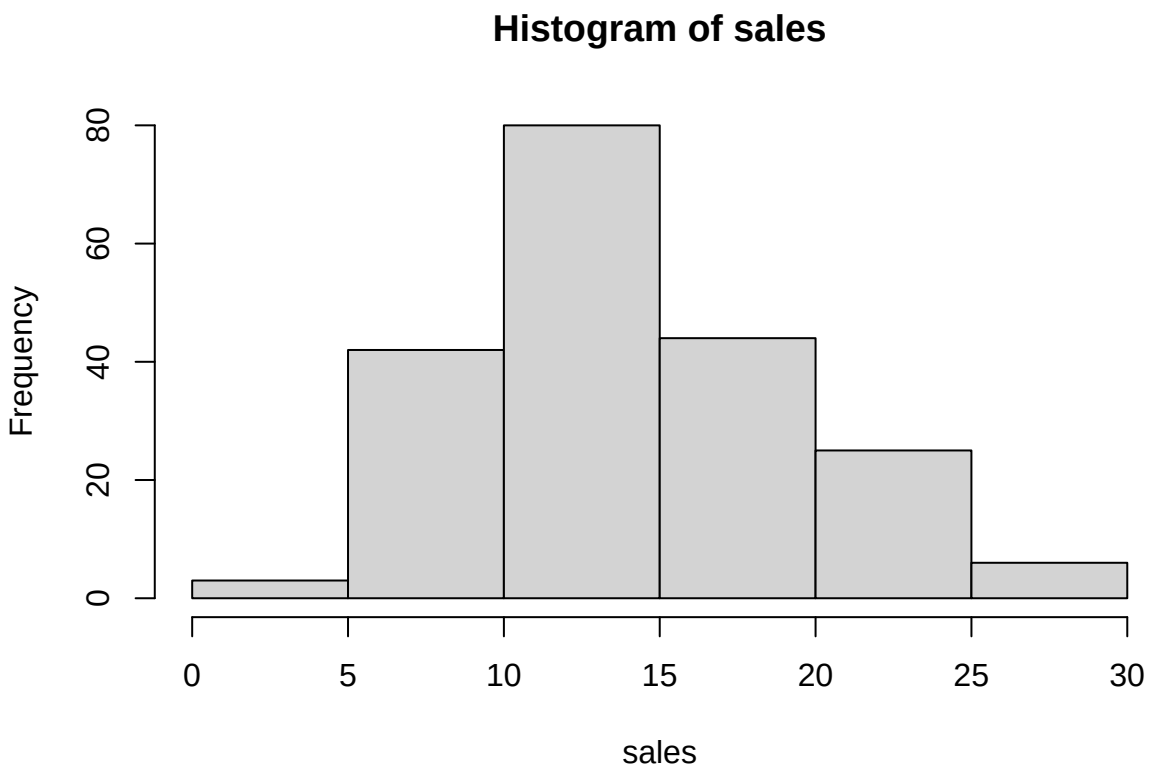


There doesn't appear to be amy curvature or a definite shape by the residuals, nor does there appear to be any fanning, which indicates a level of homoscedasticity. The values in the Q-Q plot also follows a strong linear pattern, indicating that our data may be normally distributed. However, the values at the tails of the plot show that there may be some abnormality in the distributions, but because the overall shape of the plot is linear, we can say that our data is approximately normal. Additionally, the standardized residuals versus leverage plot do not indicate any outliers that fall outside of +-3 standard deviations. However, observations 76, 17, as well as the observation located in between those two points are noticeably far from the rest of the data. However, the leverage statistic values of these points are not too high, as they are very far from 1, which indicates that these points are not those of high leverage.

```
boxplot(sales,data=Advertising, ylab="Sales")
```




```
hist(sales)
```




```
shapiro.test(lm.fit.Newspaper$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  lm.fit.Newspaper$residuals
## W = 0.98197, p-value = 0.0114
```

```
Advertising$Group = rep("Group1", 200)
indexNewspaper = (newspaper > median(newspaper))
Advertising$Group[indexNewspaper] = "Group2"
levene.test(lm.fit.Newspaper$residuals, Advertising$Group, location = c('median'))
```

```
##
##  Modified robust Brown-Forsythe Levene-type test based on the absolute
##  deviations from the median
##
## data:  lm.fit.Newspaper$residuals
## Test Statistic = 2.7715, p-value = 0.09754
```

The boxplot is fairly symmetrical with the median located roughly in the center, indicating that the data may be normally distributed. The histogram has no observable outliers, no gaps, is fairly unimodal, slightly skewed to the right, and the median of the histogram lies roughly in the center.

The Shapiro test yields a p-value less than p = 0.05, which indicates that the data is not normally distributed, and we therefore reject the null hypothesis that the data is normally distributed.

The Levene Test yields a p-value of 0.09754, whic his greater than p= 0.05, meaning that we have failed to reject the null hypothesis that our data is homoscedastic.

As such, our initial conclusions are that while homoscedasticity assumptions are met, normality is not met when regressing `sales` versus `newspaper`.

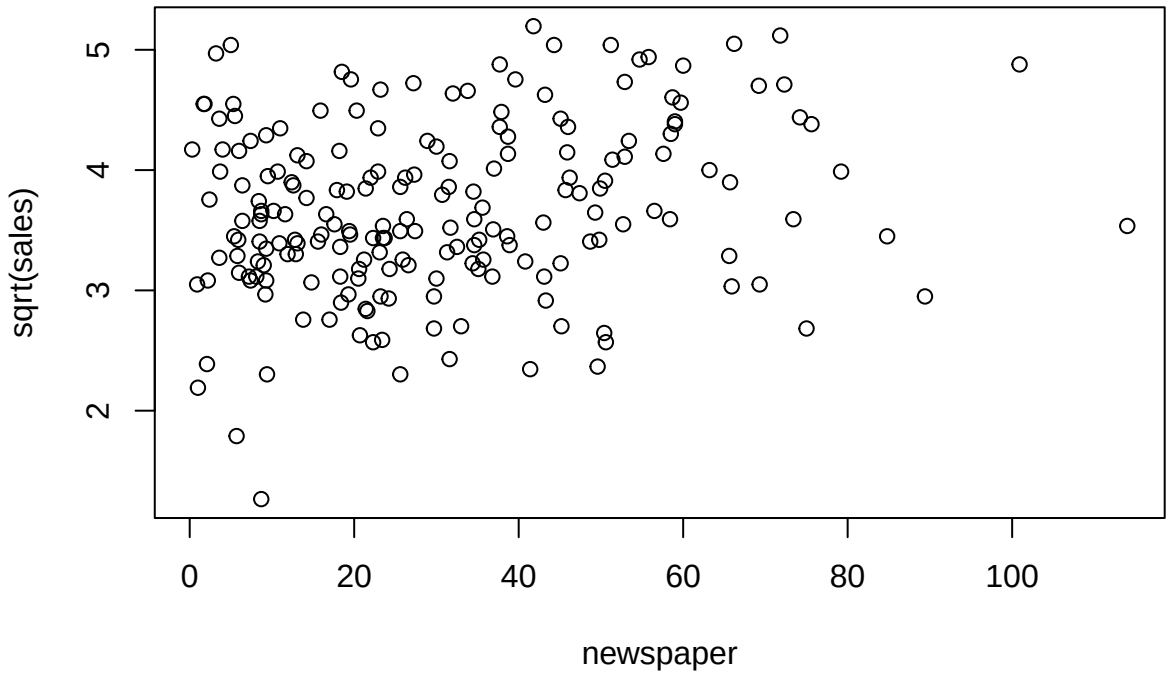
Thus, we need to apply transformations on `newspaper` to see if an SLR model is appropriate for our data. We first try the transformation:

$$X' = \sqrt{X}$$

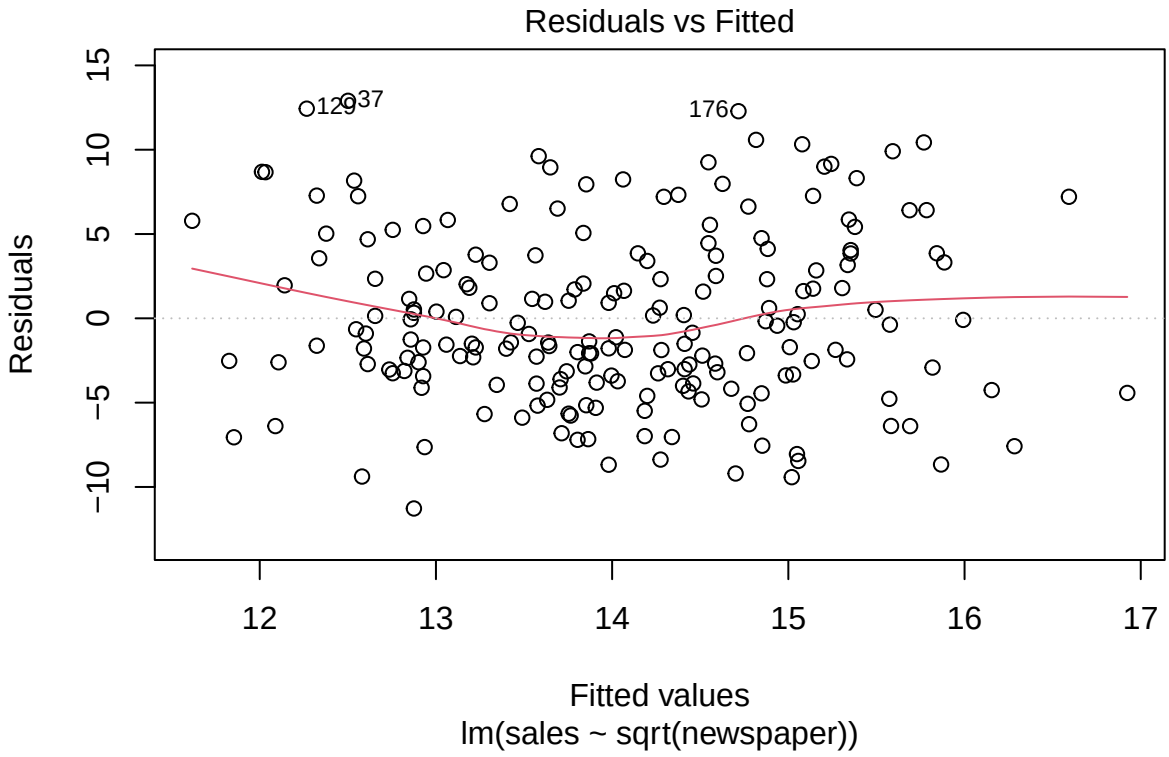
```
lm.fit.Newspaper3 = lm(sales~sqrt(newspaper))
summary(lm.fit.Newspaper3)
```

```
##
## Call:
## lm(formula = sales ~ sqrt(newspaper))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.2750  -3.4038  -0.9149   3.5997  12.8989
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      11.3296     0.9840  11.514  < 2e-16 ***
## sqrt(newspaper)   0.5239     0.1780   2.943  0.00364 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.12 on 198 degrees of freedom
## Multiple R-squared:  0.04191,    Adjusted R-squared:  0.03707
## F-statistic: 8.662 on 1 and 198 DF,  p-value: 0.003638
```

```
par(mfrow=c(1,1))
plot(newspaper,sqrt(sales))
abline(lm.fit.Newspaper3, col='red')
```

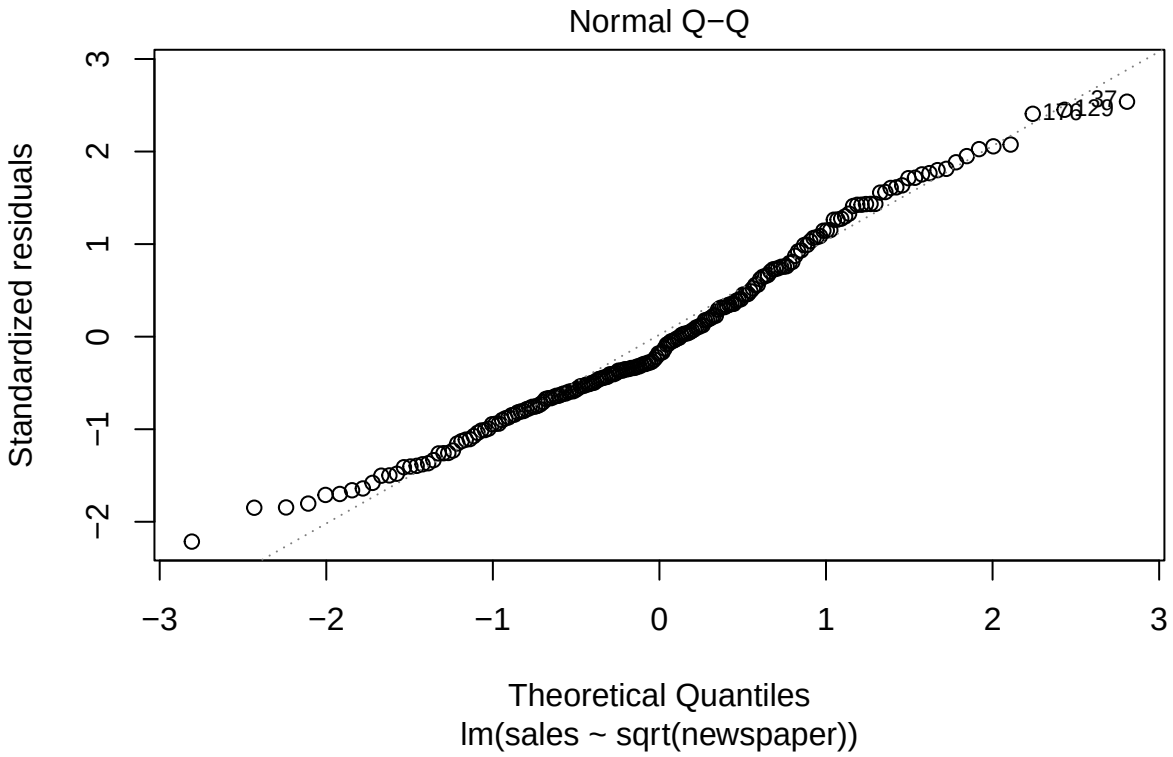


```
plot(lm.fit.Newspaper3, which = c(1))
```

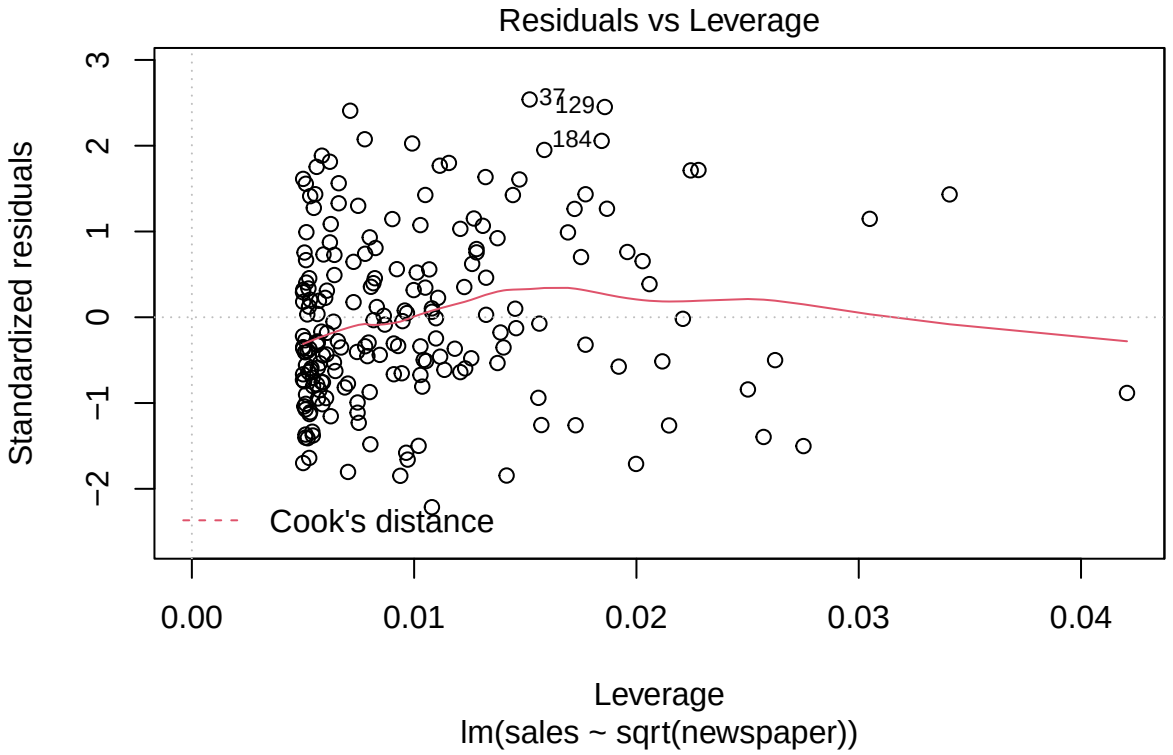


The residual plot here is spread out and there is no visible curvature or fanning. Thus, our data visually appears to be homoscedastic. However, the low adjusted r-squared value of 0.03707 is troubling, as this means that only about 3.7% of the variation in sales can be explained by our LSRL on newspaper advertising expenditures. This puts into question as to whether a linear model is the best model to use for finding a relationship between the two variables.

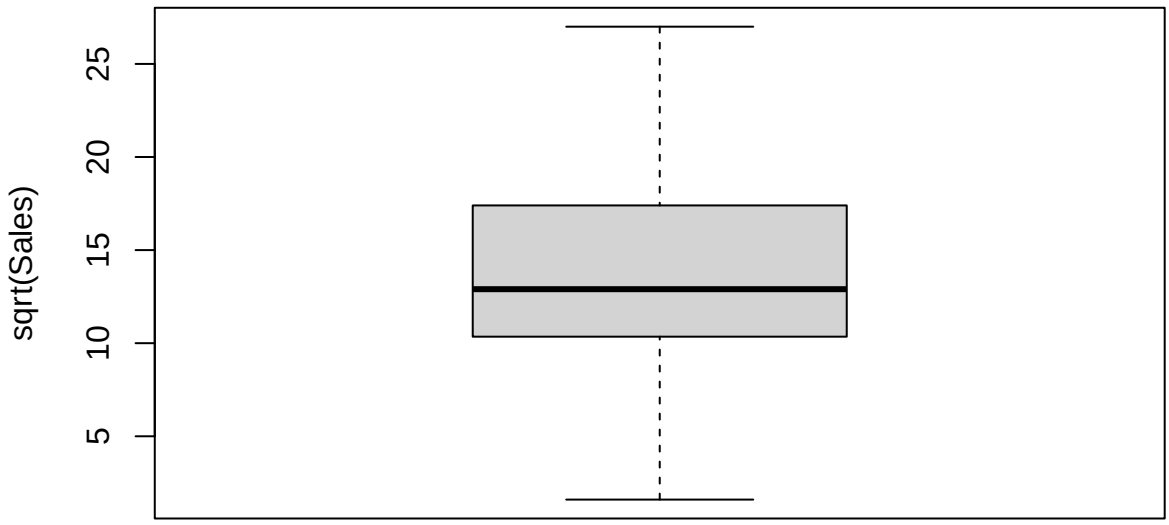
```
par(mfrow=c(1,1))
plot(lm.fit.Newspaper3, which = c(2))
```



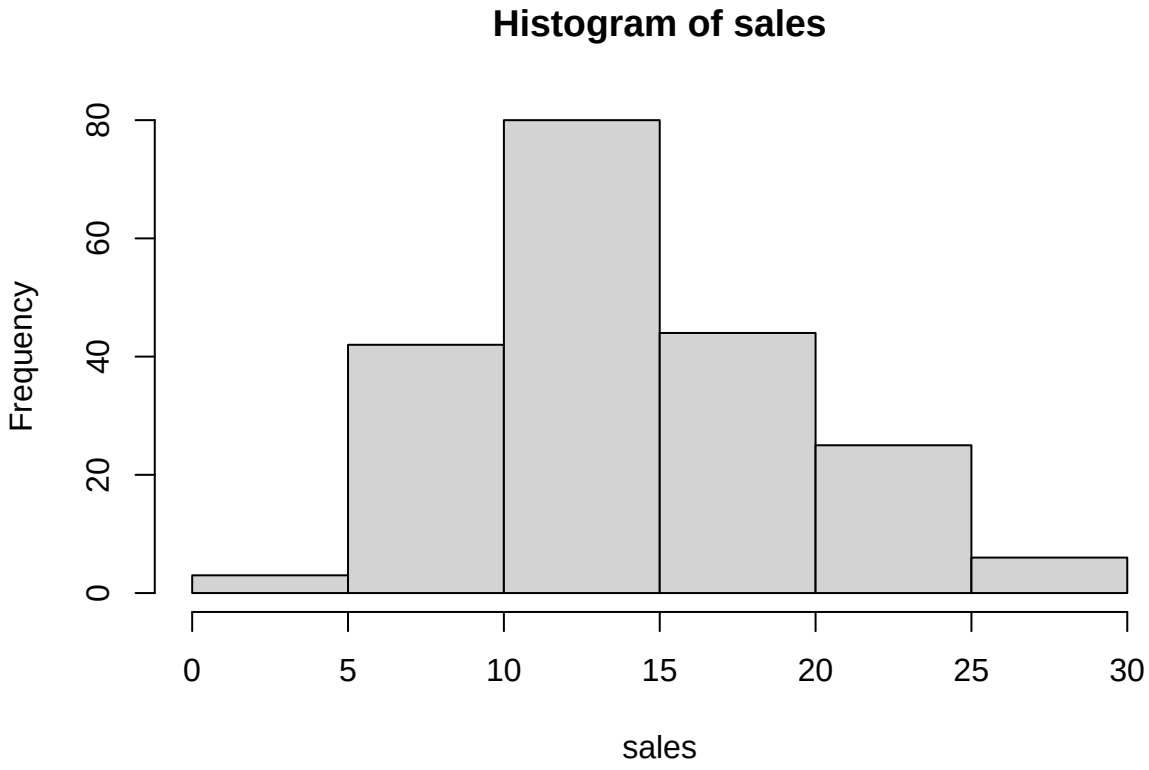

```
plot(lm.fit.Newspaper3, which = c(5))
```




```
boxplot(sales,data=Advertising, ylab="sqrt(Sales)")
```




```
hist(sales)
```




```
shapiro.test(lm.fit.Newspaper3$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  lm.fit.Newspaper3$residuals
## W = 0.97949, p-value = 0.005035
```

```
Advertising$Group1 = rep("Group1", 200)
indexNewspaper = (newspaper > median(newspaper))
Advertising$Group1[indexNewspaper] = 'Group2'
levene.test(lm.fit.Newspaper3$residuals, Advertising$Group1, location = c('median'))
```

```
##
##  Modified robust Brown-Forsythe Levene-type test based on the absolute
##  deviations from the median
##
## data:  lm.fit.Newspaper3$residuals
## Test Statistic = 2.5636, p-value = 0.1109
```

The values in the Q-Q plot also follows a strong linear pattern, indicating that our data may be normally distributed. However, the values at the tails of the plot show that there may be some abnormality in the distributions, but because the overall shape of the plot is linear, we can say that our data is approximately normal. Additionally, the standardized residuals versus leverage plot do not indicate any outliers that fall outside of +-3 standard deviations. While some observations may be leaning towards the higher ends of variance, the leverage statistic values of these points are not too high, as they are very far from 1, which indicates that these points are not those of high leverage. The boxplot is fairly symmetrical with the median located roughly in the center, indicating that the data may be normally distributed. The histogram has no observable outliers, no gaps, is fairly unimodal, and the median of the histogram lies roughly in the center. The histogram is roughly symmetric, though it is slightly skewed to the right. Thus, our data visually appears to be approximately normal.

The Shapiro test yields a p-value less than p = 0.05 (p = 0.005035), which indicates that we reject the null that our data is normally distributed. Although, we do retain homoscedasticity, as our p-value is greater than p = 0.05 (p = 0.1109).

Our following tests are as follows:

Model	p-value (SW)	p-value (BFL)
sales ~ newspaper	0.0114	0.09754
sales ~ sqrt(newspaper)	0.005035	0.1109
sales ~ log10(newspaper)	0.001358	0.09634
sales ~ exp(newspaper)	0.001993	0.04944
sales ~ newspaperˆ2	0.0114	0.09754
sales ~ 1/(newspaper)	0.001683	0.04504
sqrt(sales) ~ (newspaper)	0.308	0.3703
sqrt(sales) ~ sqrt(newspaper)	0.2702	0.4016

The transformation that yielded the greatest normality was $Y' = \sqrt{Y}$ and $X' = X$, while the transformation that yielded the greatest homoscedasticity was $Y' = \sqrt{Y}$ and $X' = \sqrt{X}$. Regardless, both of these transformations yielded p-values in the Shapiro (p = 0.308 and p = 0.2702) and Levene tests (p = 0.3703 and p = 0.4016) that were greater than p = 0.05 , and for both of these transformations, we have failed to reject the null hypotheses that the data are normally distributed and homoscedastic.

Conclusions

- There is a linear association between: 1.) `sqrt(sales)` and `newspaper` expenditure, and 2.) `sqrt(sales)` and `sqrt(newspaper)` expenditure.
- As newspaper advertising expenditure increases by $1,000, sqrt(sales) increases by 0.007207. As sqrt(newspaper) advertising expenditure increases by $1,000, sqrt(sales) increases by 0.07019.
- A linear model is appropriate between 1.) `newspaper` and `sqrt(sales)`, and 2.) `sqrt(newspaper)` and `sqrt(sales)`
- The data when `sales` is transformed into `sqrt(sales)` are normally distributed and homoscedastic.
- The data when `newspaper` is transformed into `sqrt(newspaper)` are normally distributed and homoscedastic.
- There are no clear outliers or high leverage points in the data set.