Guo - HWK6.1

Andrew Guo

11/11/2021

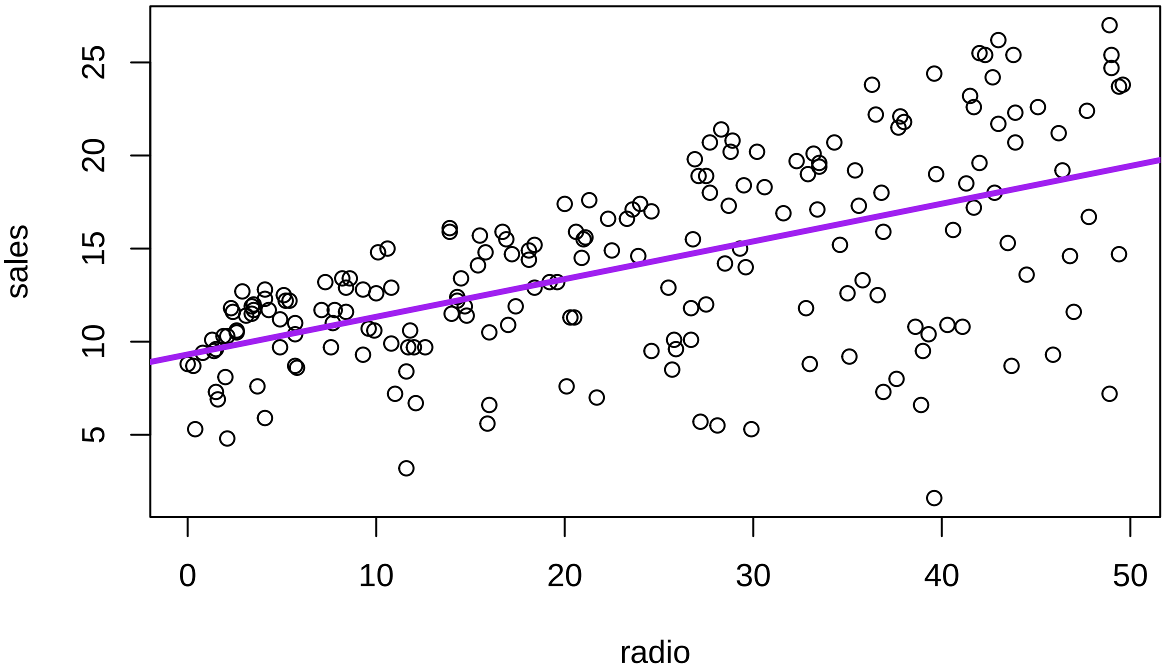
```
## Load libraries
library(lawstat)

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

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

```
##
## Call:
## lm(formula = sales ~ radio, data = Advertising)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.7305  -2.1324   0.7707   2.7775   8.1810
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.31164    0.56290  16.542   <2e-16 ***
## radio        0.20250    0.02041   9.921   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.275 on 198 degrees of freedom
## Multiple R-squared:  0.332,  Adjusted R-squared:  0.3287
## F-statistic: 98.42 on 1 and 198 DF,  p-value: < 2.2e-16
```

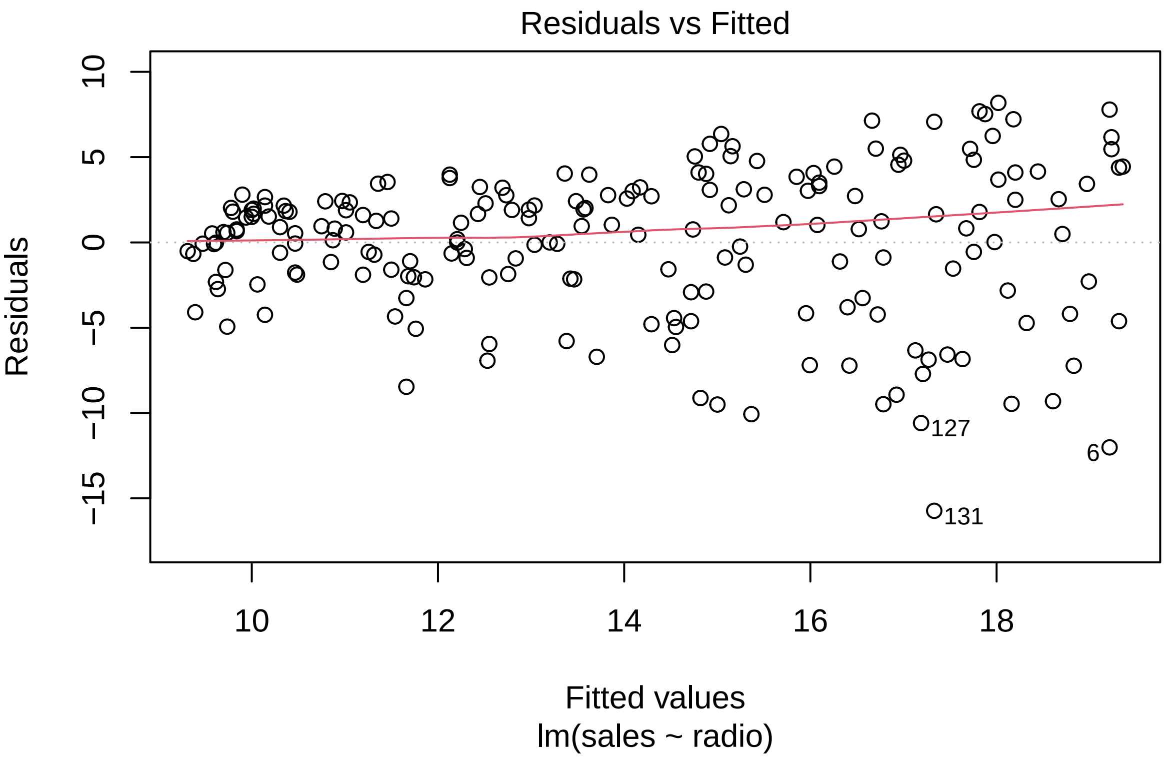
```
## Scatter plot along with linear fit
par(mfrow=c(1,1))
plot(radio,sales)
abline(lm.fit.Radio, col='purple', lwd=3)
```



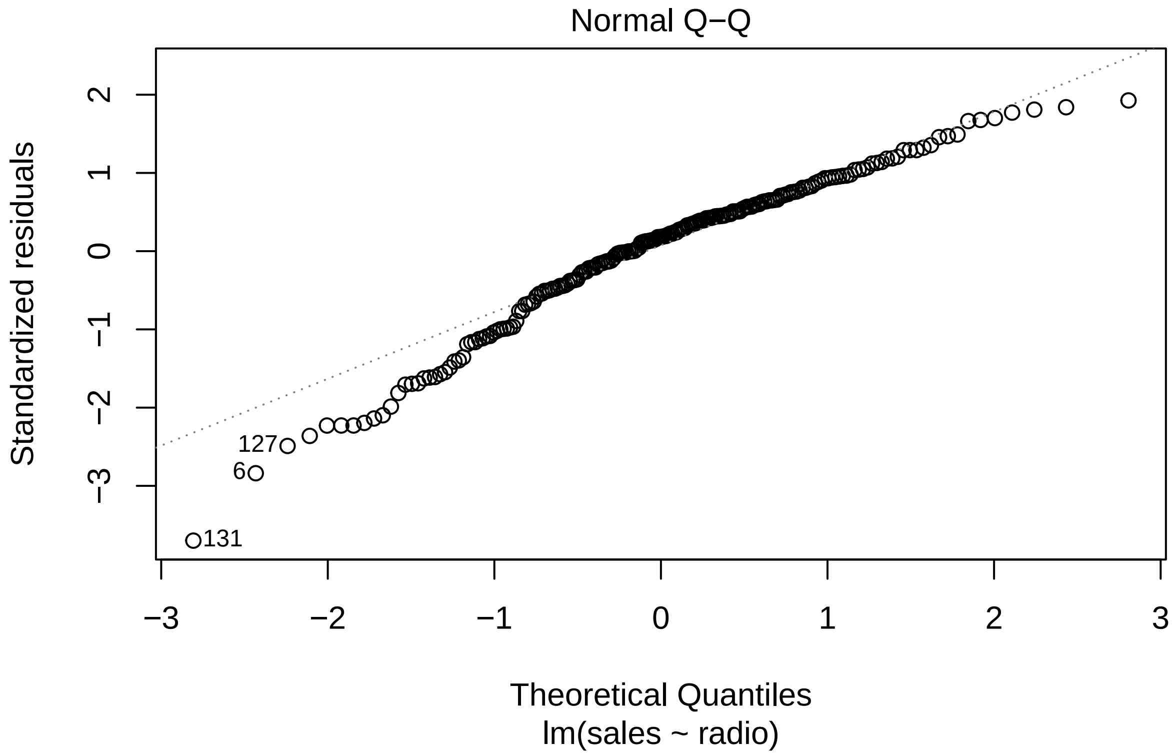
Based on the summary statistics, we observe the following: We have an LSRL of y = 0.20x + 9.31164. We observe a p-value of < 0.05 and an observable trend that as `radio` advertising expenditure increases, `sales` also increases. We can conclude that 'radio' and 'sales' have a linear relationship.

However, with an adjusted R-squared value of 0.3287, this means that only about 32.8% of the variation in sales is explained by the LSRL on sales. This places some skepticism as to how accurate a linear model for comparing these two variables can be. Thus, further diagnostics will be needed.

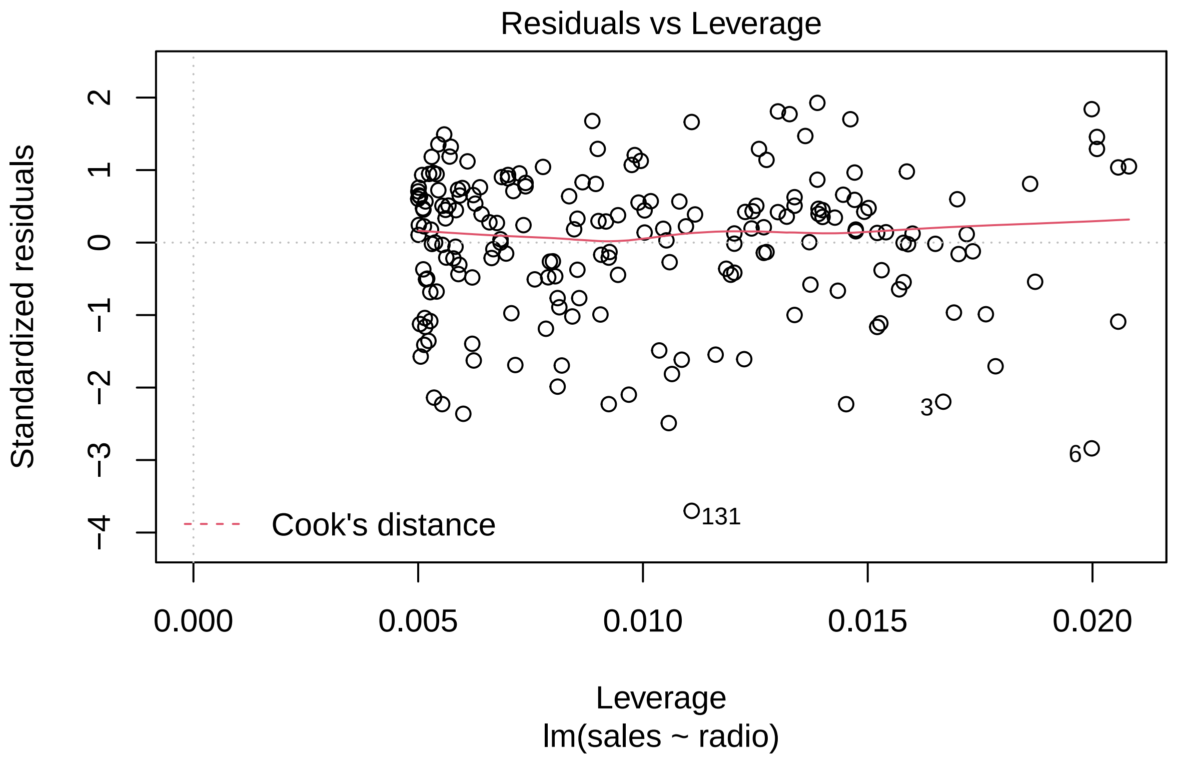
```
## Diagnostic plots
plot(lm.fit.Radio, which = c(1))
```



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

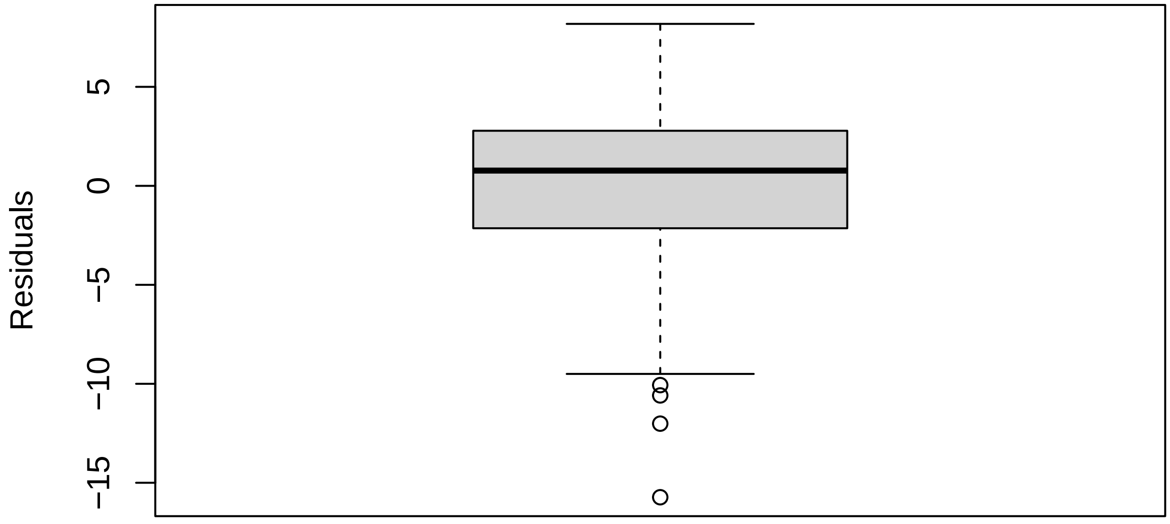


```
plot(lm.fit.Radio, which = c(5))
```

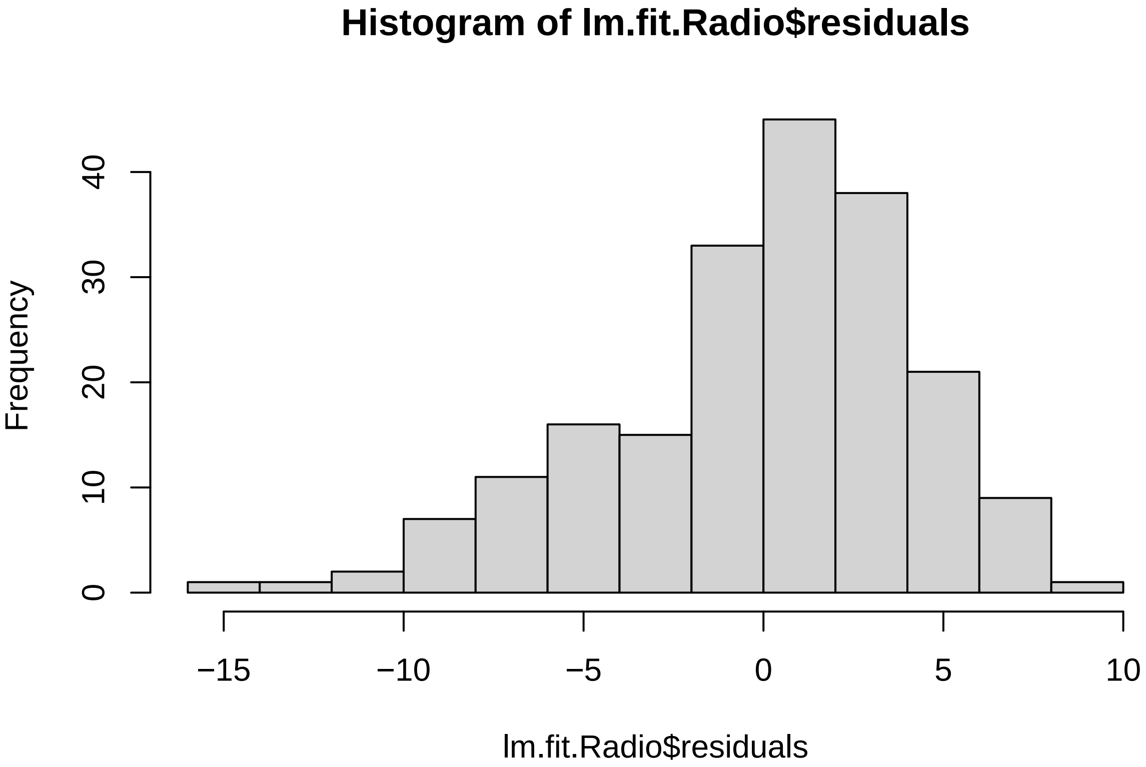


There doesn't appear to be amy curvature or a definite shape by the residuals, other than a noticeable fanning outwards to the right. This indicates that the data appears to not be homoscedastic. The Q-Q plot indicates a moderate to strong level of normality. Although, the quantiles at the ends of the plot have considerably more variation compared to the quantiles in the middle along the line of best fit. However, 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 or high leverage points. There are some values that could be outliers, such as observation 131, but while this value lies farther from the other points in this plot, its leverage statistic is still very small, so it has little impact on the overall shape of the graph.

```
boxplot(lm.fit.Radio$residuals, ylab="Residuals")
```



```
hist(lm.fit.Radio$residuals)
```



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

```
##
##  Shapiro-Wilk normality test
##
## data:  lm.fit.Radio$residuals
## W = 0.96072, p-value = 2.367e-05
```

```
Advertising$Group = rep("Group1",200)
indexRadio = (radio > median(radio))
Advertising$Group[indexRadio] = "Group2"
levene.test(lm.fit.Radio$residuals, Advertising$Group , location = c("median"))
```

```
##
##  Modified robust Brown-Forsythe Levene-type test based on the absolute
##  deviations from the median
##
## data:  lm.fit.Radio$residuals
## Test Statistic = 45.455, p-value = 1.671e-10
```

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 substantially 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. Additionally, by the Levene Test, we have a p-value that is less than p = 0.05, which backs up our initial suspicion that our data was heteroscedastic.

As such, our initial conclusions are that the normality and homoscedasticity assumptions are not met when regressing sales versus radio.

Thus, we need to apply transformations on `sales` to see if an SLR model is appropriate for our data.

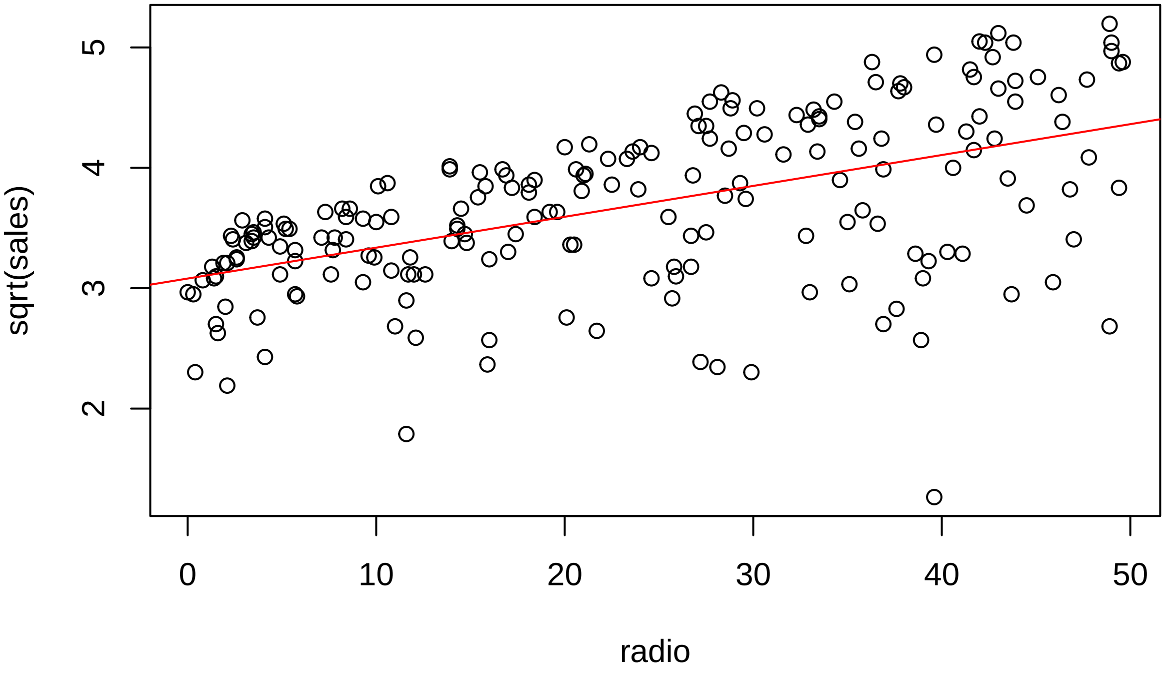
We first try the transformation:

$$Y' = \sqrt{Y}$$

```
lm.fit.Radio2 = lm(sqrt(sales)~radio)
summary(lm.fit.Radio2)
```

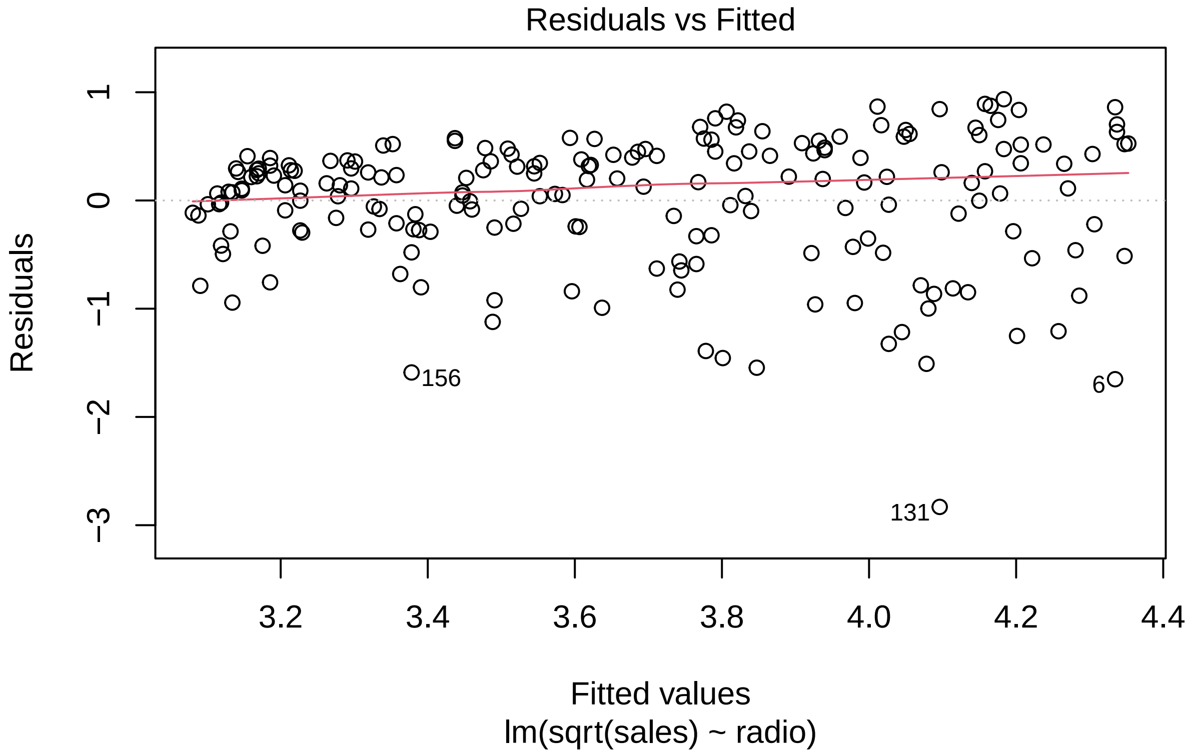
```
##
## Call:
## lm(formula = sqrt(sales) ~ radio)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8312 -0.2789  0.1403  0.4149  0.9353
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.080451   0.079078  38.955  < 2e-16 ***
## radio       0.025647   0.002867   8.944 2.68e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6006 on 198 degrees of freedom
## Multiple R-squared:  0.2878, Adjusted R-squared:  0.2842
## F-statistic:    80 on 1 and 198 DF,  p-value: 2.675e-16
```

```
par(mfrow=c(1,1))
plot(radio,sqrt(sales))
abline(lm.fit.Radio2, col='red')
```

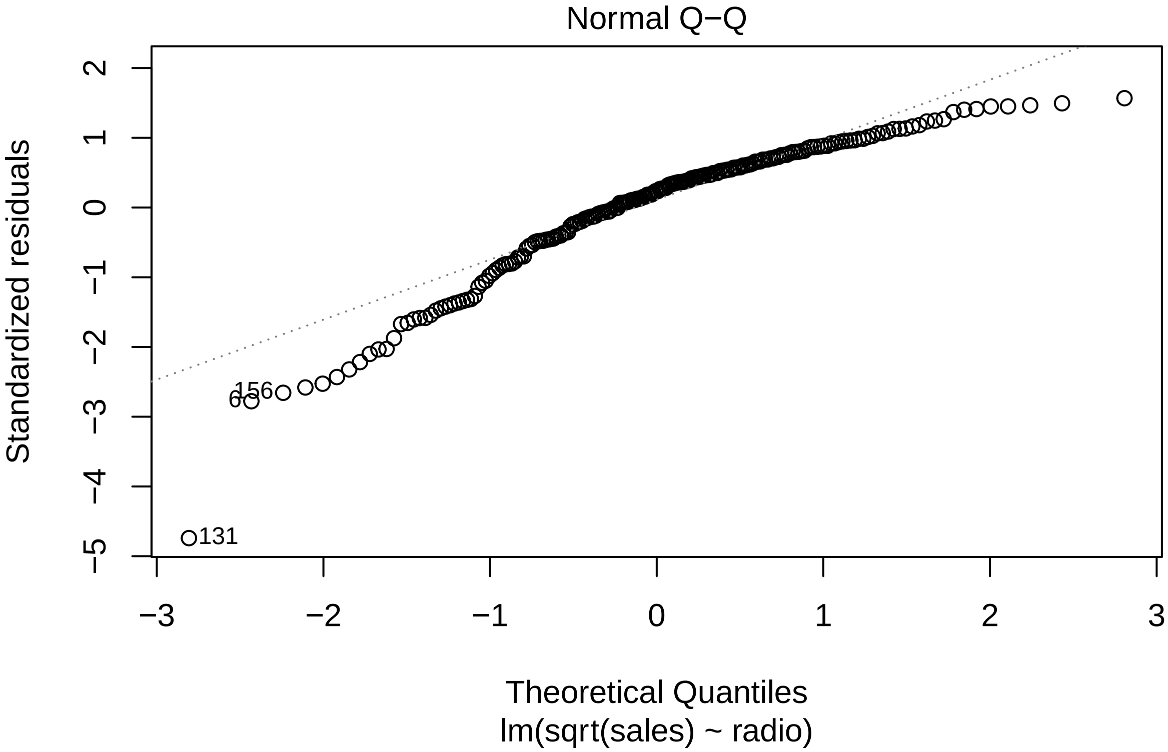


Since we obtained a p-value < .05, we can conclude that radio`` andsqrt(sales)have a linear association. Addionally, our graph shows an observable trend that asradioadvertising expenditure increases,sqrt(sales)‘ also increases, further supporting the notion that a linear association does exist.

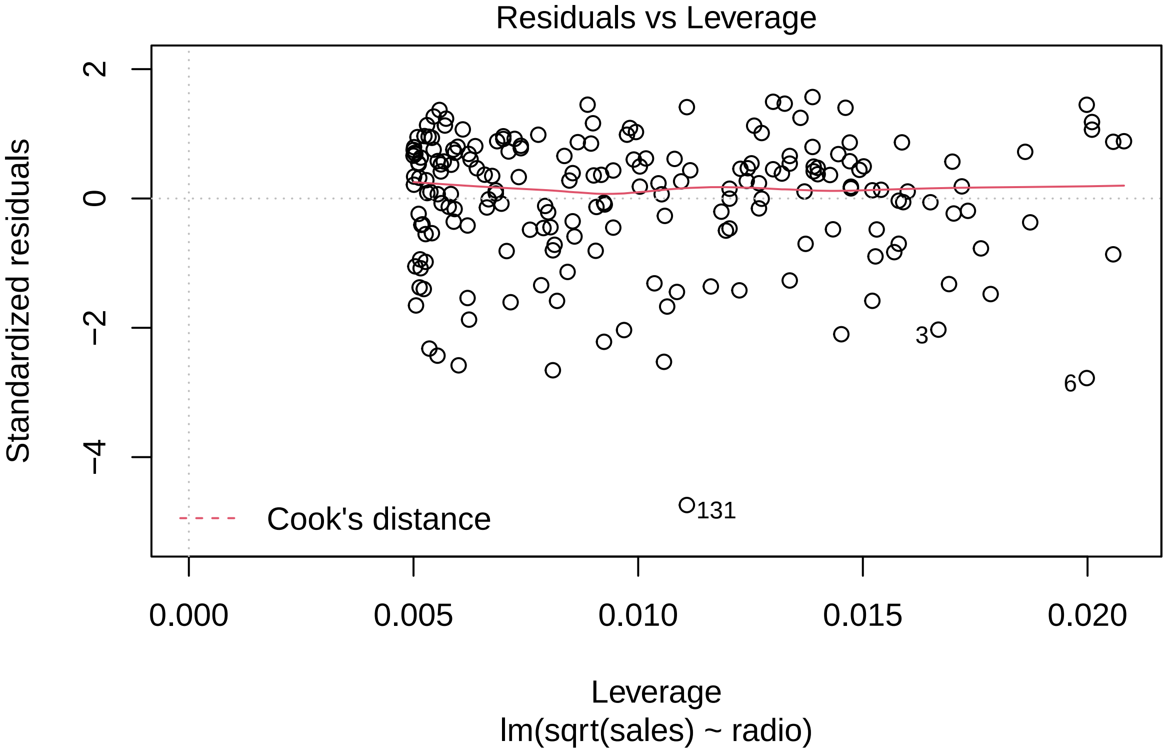
```
par(mfrow=c(1,1))
plot(lm.fit.Radio2, which = c(1))
```



```
plot(lm.fit.Radio2, which = c(2))
```



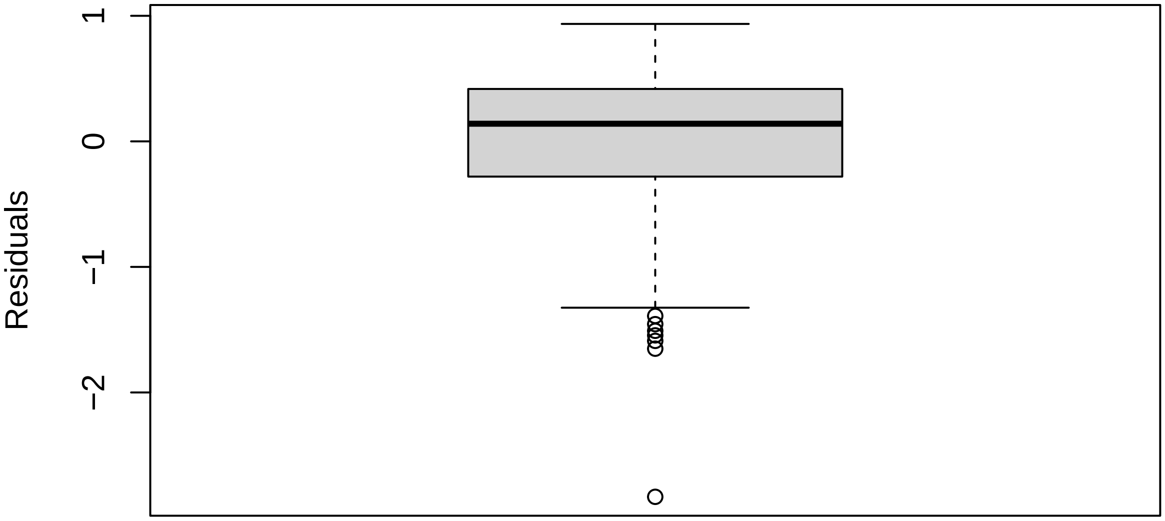
```
plot(lm.fit.Radio2, which = c(5))
```



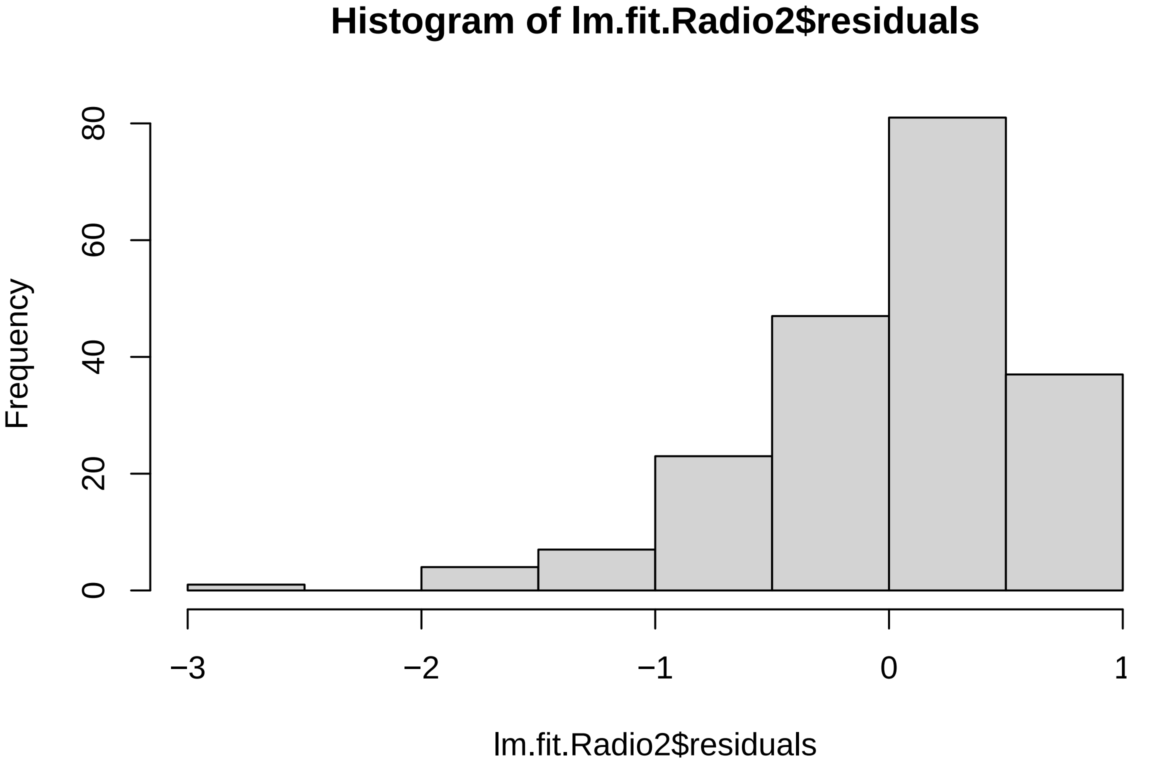
The quantiles on the Q-Q plot follow a moderate-strong linear relationship, indicating that our data appears to be approximately normal. However, the values at the tails of the plot show that there may be some abnormality in the distributions, but because the overall shape of the plot is linear, so we make the initial claim that our data is approximately normal. Additionally, the standardized residuals versus leverage plot do not indicate any outliers or high leverage points. However, our residual plot does indicate that there is a minor fanning out of the residuals as the fitted values increase, indicating that our data may not homoscedastic. Although, it is far less heteroscedastic than our residual plot without transformations, as the residuals here vary from around 1 to -2 residuals, whereas in our initial data, the data varied anywhere from -15 to 10 residuals.

We move onto quantitative calculations for normality and homoscedasticity:

```
boxplot(lm.fit.Radio2$residuals, ylab="Residuals")
```



```
hist(lm.fit.Radio2$residuals)
```



```
shapiro.test(lm.fit.Radio2$residuals)
```

```
##
```

```
##  Shapiro-Wilk normality test
##
## data:  lm.fit.Radio2$residuals
## W = 0.91685, p-value = 3.443e-09
```

```
levene.test(lm.fit.Radio2$residuals,Advertising$Group , location = c("median"))
```

```
##
##  Modified robust Brown-Forsythe Levene-type test based on the absolute
##  deviations from the median
##
## data:  lm.fit.Radio2$residuals
## Test Statistic = 22.188, p-value = 4.645e-06
```

Our boxplot indicates the data may be normally distributed, as the distribution does appear to be somewhat symmetrical. However, the histogram reveals significant asymmetry, as our data appears to be skewed to the left. The Shapiro-Wilk test no longer confirms normality, as the p-value = 3.443e-09 < 0.05. Thus, we reject the null hypothesis that the data are normally distributed. Additionally, the Brown-Forsythe-Levene test confirms the data are not homoscedastic, as the p-value = 4.645e-06 < 0.05. Thus, we reject the null hypothesis that the data are homoscedastic.

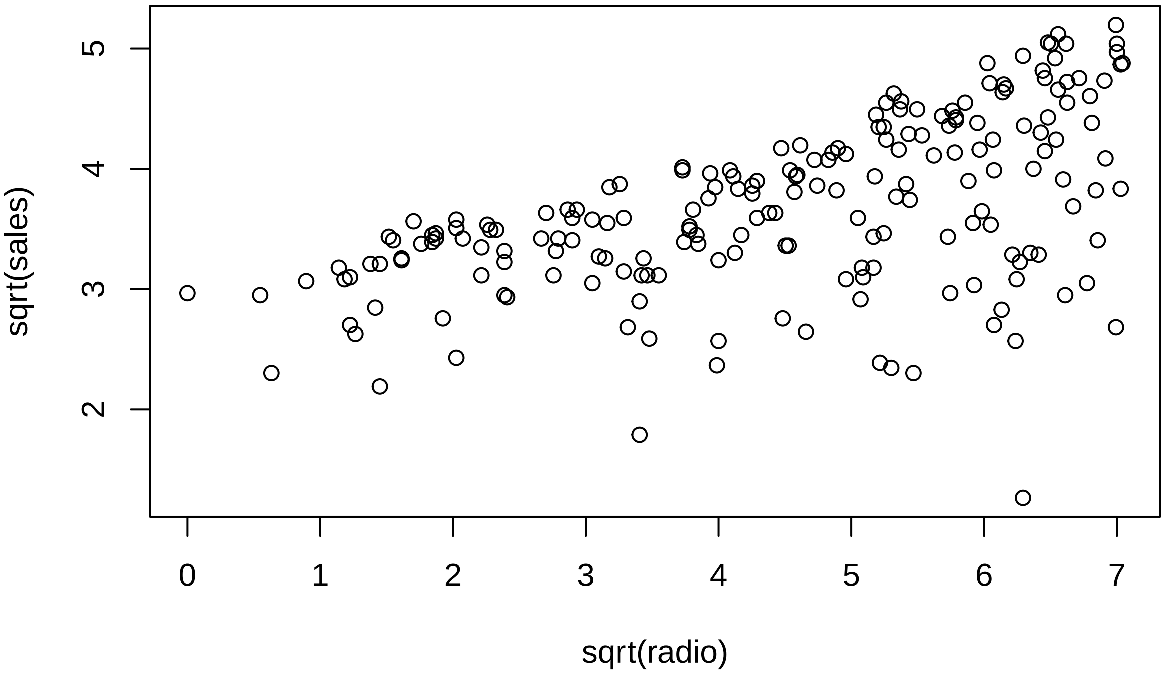
Because we were unable to obtain homoscedasticity, we need to continue to transform our y variable. Thus, we try:

$$Y' = 1/Y$$

```
lm.fit.Radio4 = lm(1/(sales)~radio)
summary(lm.fit.Radio4)
```

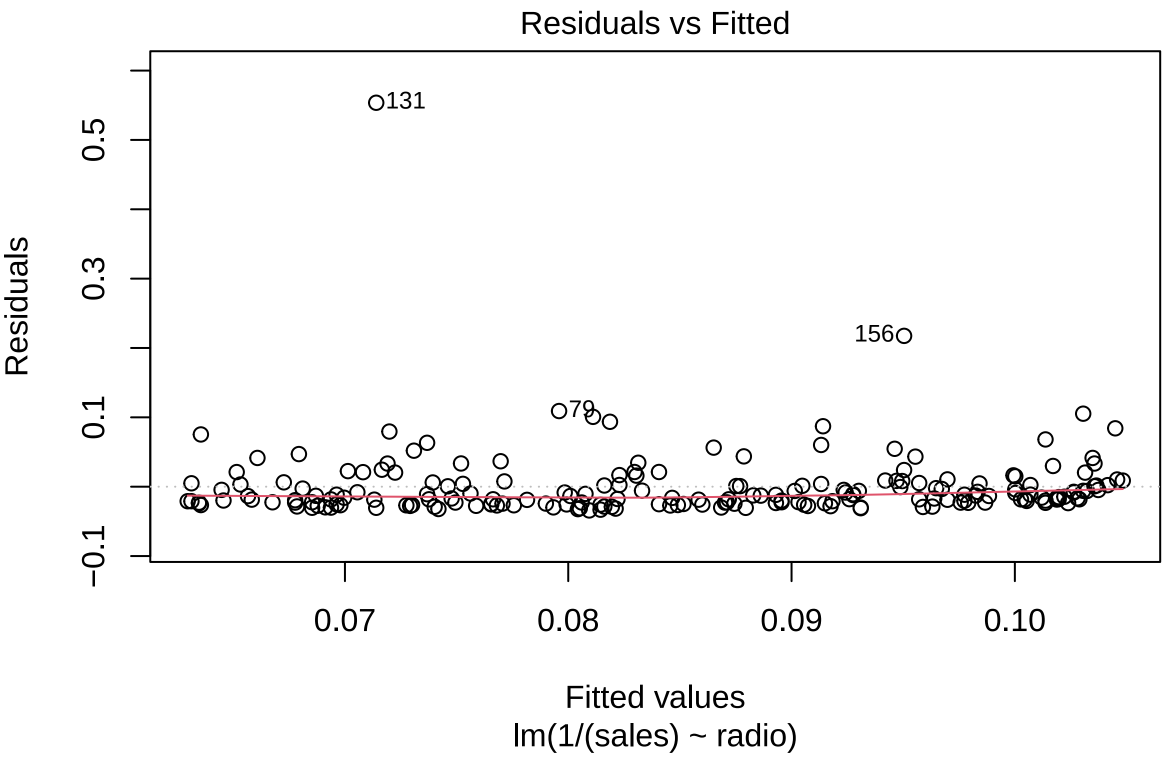
```
##
## Call:
## lm(formula = 1/(sales) ~ radio)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.03421 -0.02334 -0.01350  0.00500  0.55360
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.1048333  0.0067926  15.434  < 2e-16 ***
## radio       -0.0008443  0.0002463  -3.428  0.00074 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05159 on 198 degrees of freedom
## Multiple R-squared:  0.05602,    Adjusted R-squared:  0.05125
## F-statistic: 11.75 on 1 and 198 DF,  p-value: 0.0007403
```

```
par(mfrow=c(1,1))
plot(sqrt(radio),sqrt(sales))
abline(lm.fit.Radio4, col='purple')
```

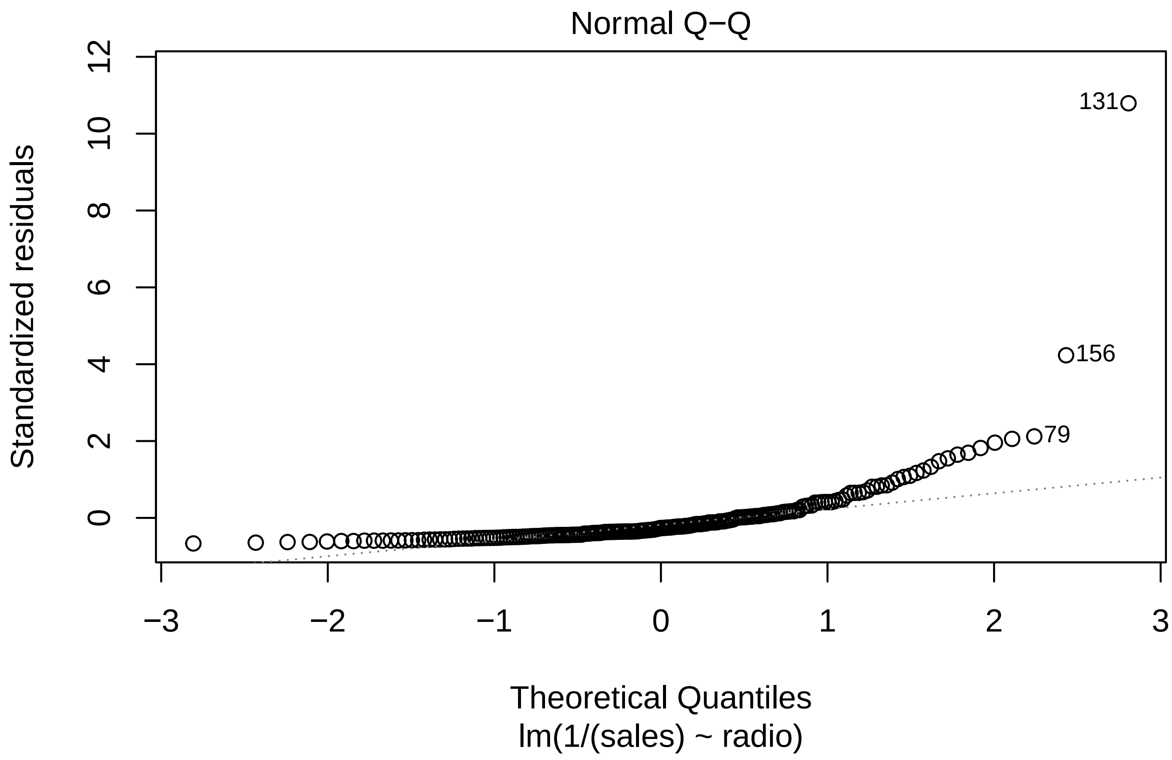


Since we obtained a p-value < .05, we can conclude that `log(radio)` and `sqrt(sales)` have a linear association. Addionally, our graph shows an observable trend that as `sqrt(radio)` advertising expenditure increases, `sqrt(sales)` also increases, further supporting the notion that a linear association does exist.

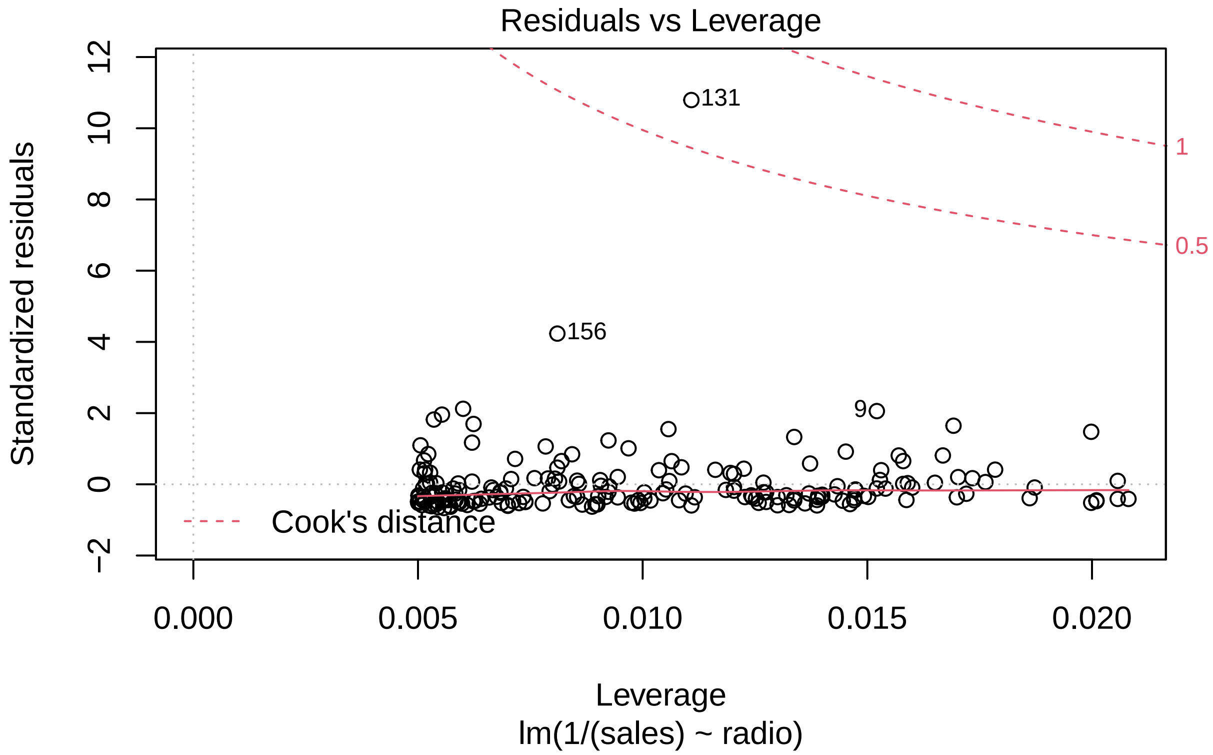
```
par(mfrow=c(1,1))
plot(lm.fit.Radio4, which = c(1))
```



```
plot(lm.fit.Radio4, which = c(2))
```

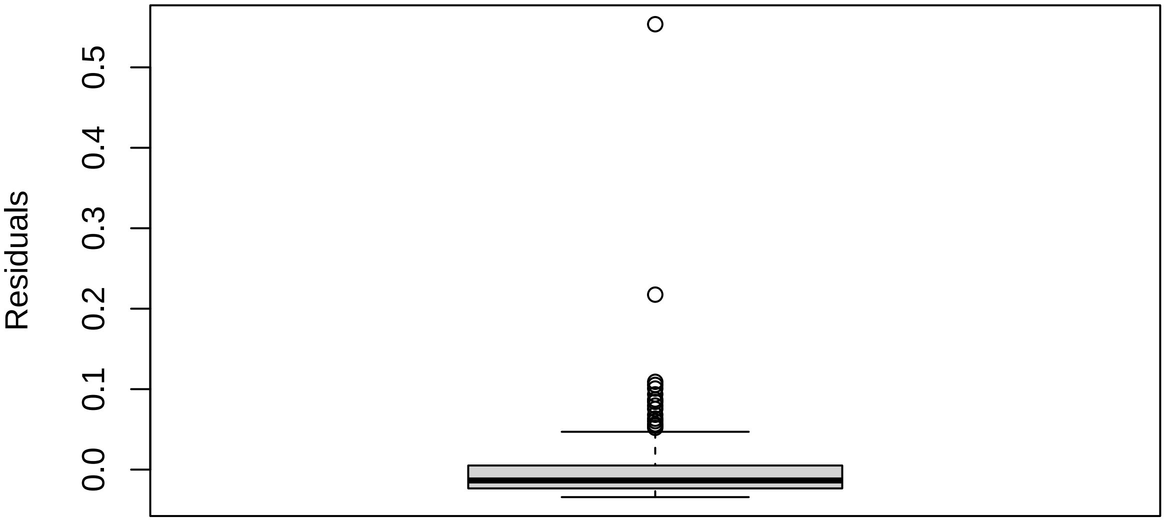


```
plot(lm.fit.Radio4, which = c(5))
```

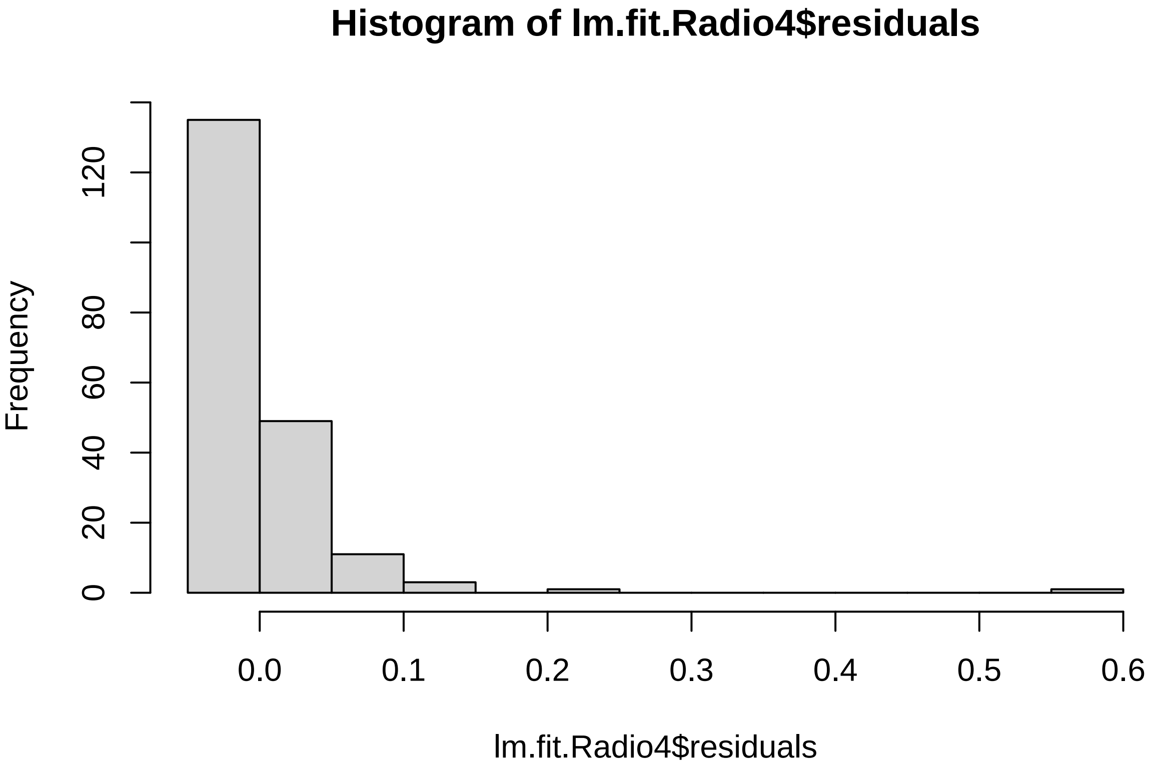


The quantiles on the Q-Q plot follow a moderate-strong linear relationship, indicating that our data appears to be approximately normal. However, the values at the tails of the plot show that there may be some abnormality in the distributions, but because the overall shape of the plot is linear, so we make the initial claim that our data is approximately normal. Additionally, the standardized residuals versus leverage plot indicates observation 131 as an outlier, for it falls outside of 3 standard deviations. However, this point does not yield high leverage, meaning that it should not affect the overall shape of the distribution drastically. Additionally, our residual plot does not indicate a strong presence of fanning out in the residuals, meaning that our data may not homoscedastic.

```
boxplot(lm.fit.Radio4$residuals, ylab="Residuals")
```



```
hist(lm.fit.Radio4$residuals)
```



```
shapiro.test(lm.fit.Radio4$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
```

```
## data:  lm.fit.Radio4$residuals
## W = 0.47953, p-value < 2.2e-16
```

```
levene.test(lm.fit.Radio4$residuals,Advertising$Group , location = c("median"))
```

```
##
##  Modified robust Brown-Forsythe Levene-type test based on the absolute
##  deviations from the median
##
## data:  lm.fit.Radio4$residuals
## Test Statistic = 0.91048, p-value = 0.3412
```

Our boxplot indicates the data may not be normally distributed, as the distribution is nowhere near symmetrical and there are a multitude of visible outliers. The histogram reveals significant asymmetry, as our data appears to be skewed to the right The Shapiro-Wilk test does not confirm normality, as the p-value = 7.05e-14 < 0.05. Thus, we reject the null hypothesis that the data are normally distributed. However, the Brown-Forsythe-Levene test confirms homoscedasticity, as the p-value = 0.3412 > 0.05. Thus, we have failed to reject the null hypothesis that the data is homoscedastic. We will therefore keep

$$Y' = 1/Y$$

as a transformation on the y-variable and now transform on the X-variable in an attempt to achieve a normal distribution. However, because there are values of zero in the data for `radio`, we are unable to use the transformations of

$$X' = 1/X$$

and

$$X' = log(X)$$

. Our tests are as follows:

Model	p-value (SW)	p-value (BFL)
1/(sales) ~ radio	2.2e-16	0.3412
1/(sales) ~ sqrt(radio)	2.2e-16	0.2836
1/(sales) ~ exp(radio)	2.2e-16	0.3802
1/sales ~ exp(-radio)	2.2e-16	0.2945
1/(sales) ~ radio^2	2.2e-16	0.3412

The transformation that yielded the least heteroscedastic data was $Y' = 1/Y$ and $X' = exp(X)$ however no subsequent transformation on `radio` yielded normally distributed residuals.

Thus, we cannot use a linear modeling approach with the `Advertising` data to compare `sales` and `radio`, as the assumptions for such a model are not met and cannot be fully met with an appropriate transformation.