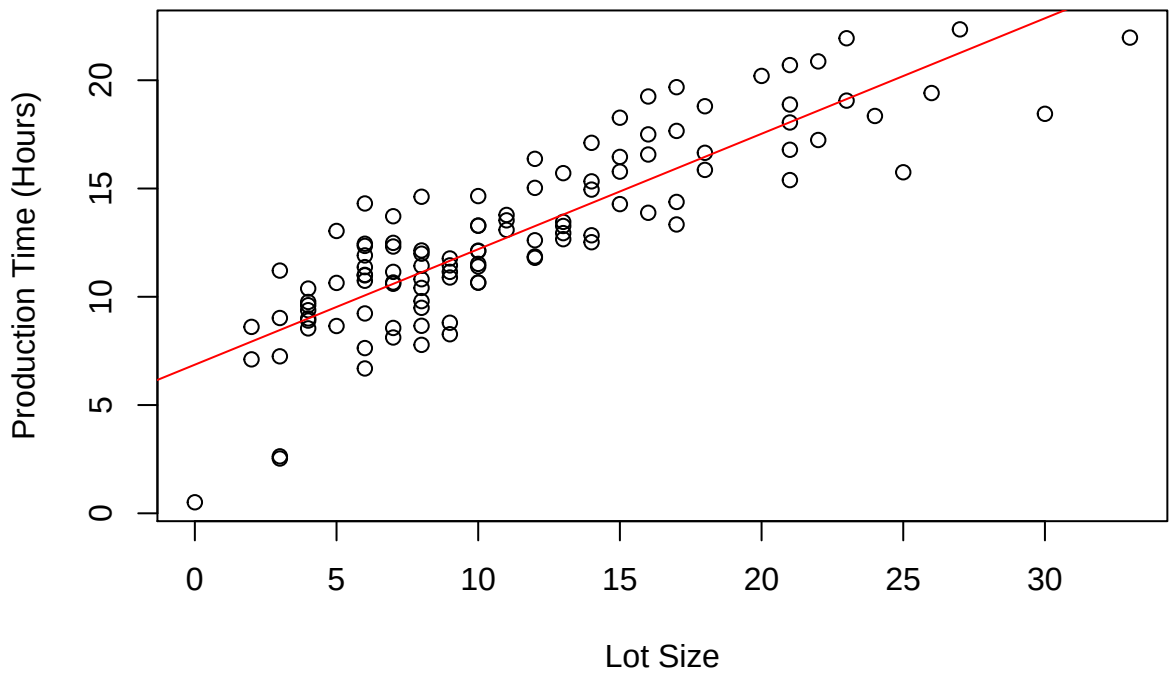
## 3. Do problem 3.18 on page 151.

**Problem 3.18: Production time. In a manufacturing study, the production times for 111 recent production runs were obtained. The table below lists for each run the production time in hours (Y) and the production lot size (X).**

**A. Prepare a scatter plot of the data. Does a linear relation appear adequate here? Would a transformation on X or Y be more appropriate here?**

```
df <- read.csv("data/CH. 3, PR 18.csv")
result <- lm(Hours ~ LotSize, data=df)
plot(x=df$LotSize, y=df$Hours, xlab="Lot Size", ylab="Production Time (Hours)")
abline(result, col="red")
```

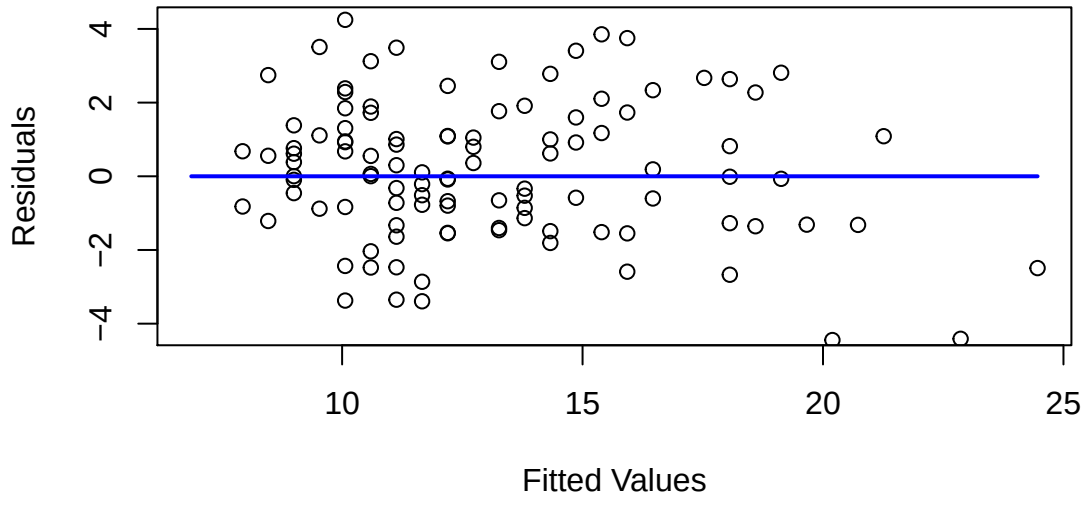



It is difficult to tell from the plot of Hours as a funtion of Lot size along. However, I would say it is reasonably linear.

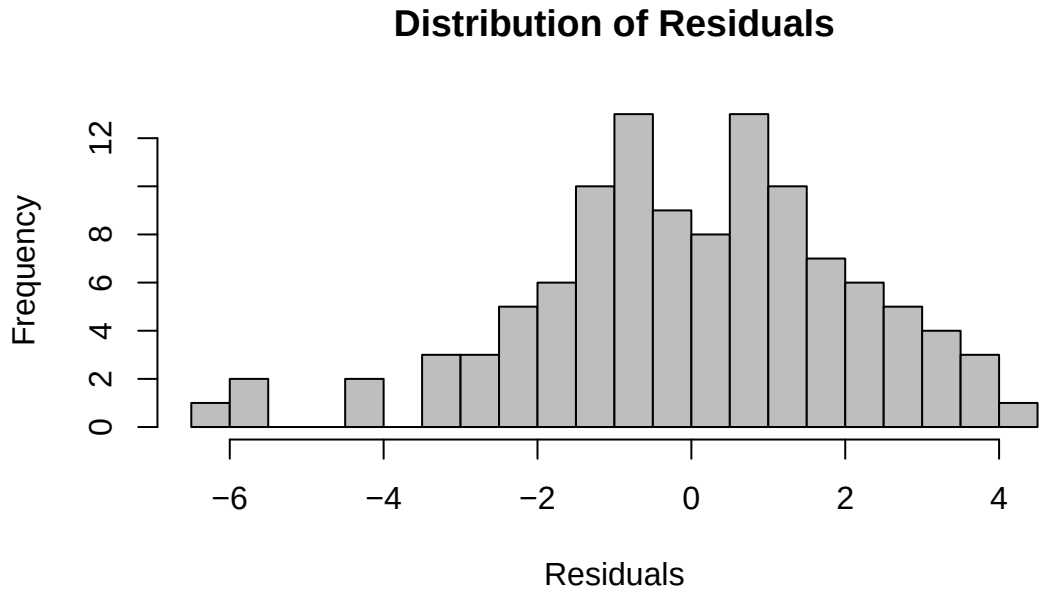
Below is a plot of the residuals vs fitted (predicted) values and a histogram of the residuals which does suggest some departures from linearity.

```
with(result, {
     plot(x=fitted.values, y=residuals,
          ylim=c(-max(residuals), max(residuals)),
          xlab="Fitted Values", ylab="Residuals")
```

```
    points(c(min(fitted.values), max(fitted.values)),
           c(0,0), type="l", lwd="2", col="blue")
    })
```




```
hist(result$residuals, main="Distribution of Residuals", xlab="Residuals", breaks=20, col="grey")
```




**B. Use the transformation $X' = \sqrt{X}$ and obtain the estimated linear regression function for the transformed data.**

```r
transformed_result <- lm(Hours ~ sqrt(LotSize), data=df)
summary(result)
```

```
##
## Call:
## lm(formula = Hours ~ LotSize, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.3535 -1.3154  0.0036  1.2405  4.2469
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.86349    0.39863   17.22   <2e-16 ***
## LotSize      0.53327    0.03028   17.61   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.118 on 109 degrees of freedom
## Multiple R-squared:   0.74,  Adjusted R-squared:  0.7376
## F-statistic: 310.2 on 1 and 109 DF,  p-value: < 2.2e-16
```

```r
summary(transformed_result)
```

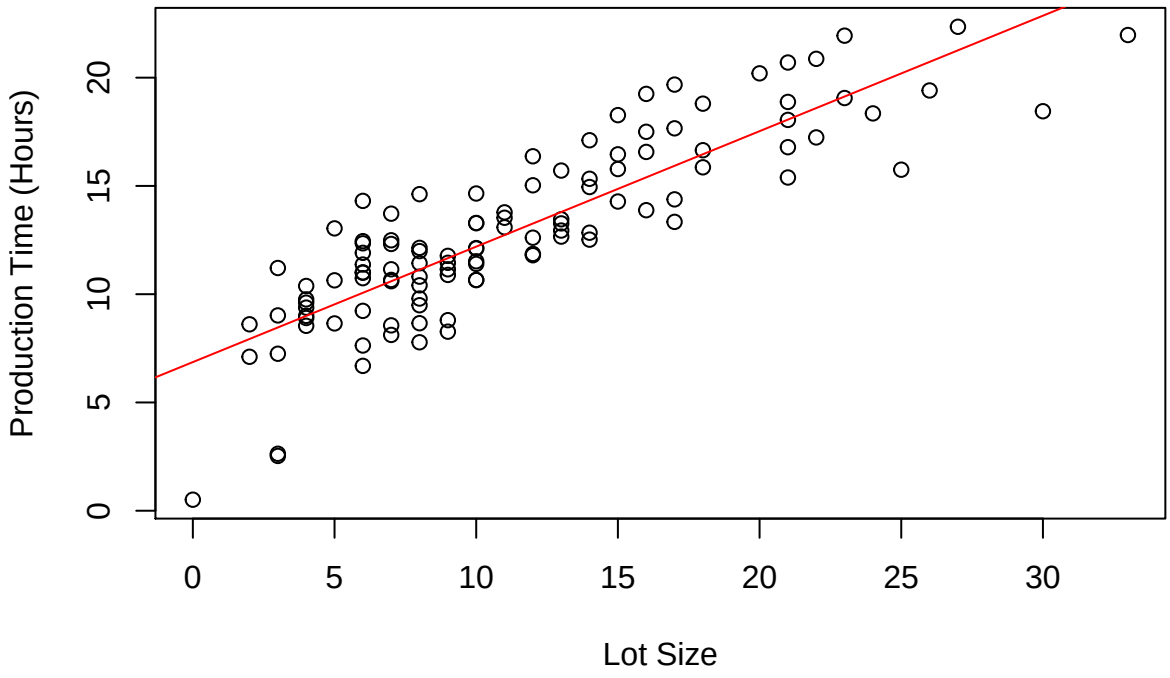
```
##
## Call:
## lm(formula = Hours ~ sqrt(LotSize), data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.0008 -1.2161  0.0383  1.3367  4.1795
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     1.2547     0.6389   1.964   0.0521 .
## sqrt(LotSize)   3.6235     0.1895  19.124   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.99 on 109 degrees of freedom
## Multiple R-squared:  0.7704, Adjusted R-squared:  0.7683
## F-statistic: 365.7 on 1 and 109 DF,  p-value: < 2.2e-16
```

```r
b0 <- coef(summary(transformed_result))[1,1]
b1 <- coef(summary(transformed_result))[2,1]
```

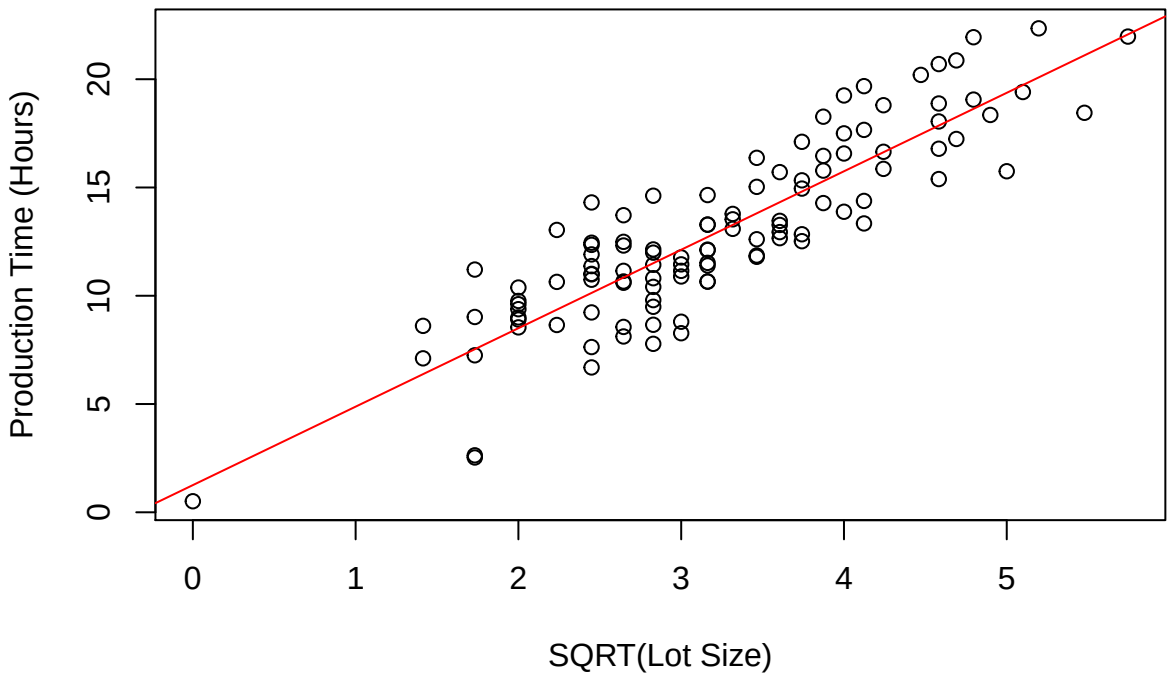
$\text{Hours} = 1.2546966 + 3.6235203 \text{ x } \sqrt{LotSize}$

**C. Plot the estimated regression line and the transformed data. Does the regression line appear to be a good fit to the transformed data?**

```r
plot(x=df$LotSize, y=df$Hours, xlab="Lot Size", ylab="Production Time (Hours)")
abline(result, col="red")
```



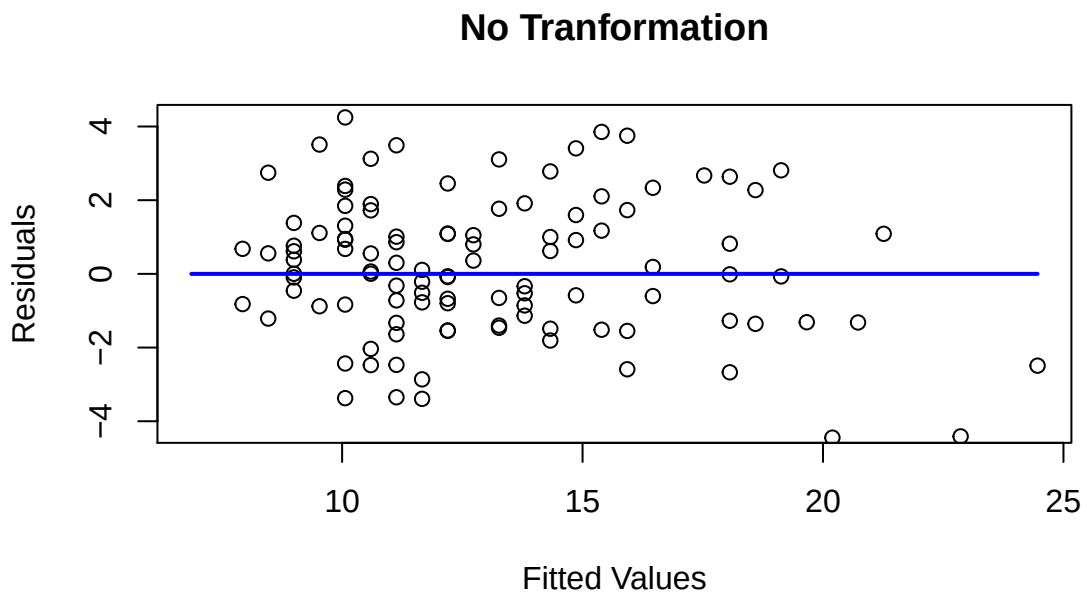

```r
plot(x=sqrt(df$LotSize), y=df$Hours, xlab="SQRT(Lot Size)", ylab="Production Time (Hours)")
abline(transformed_result, col="red")
```



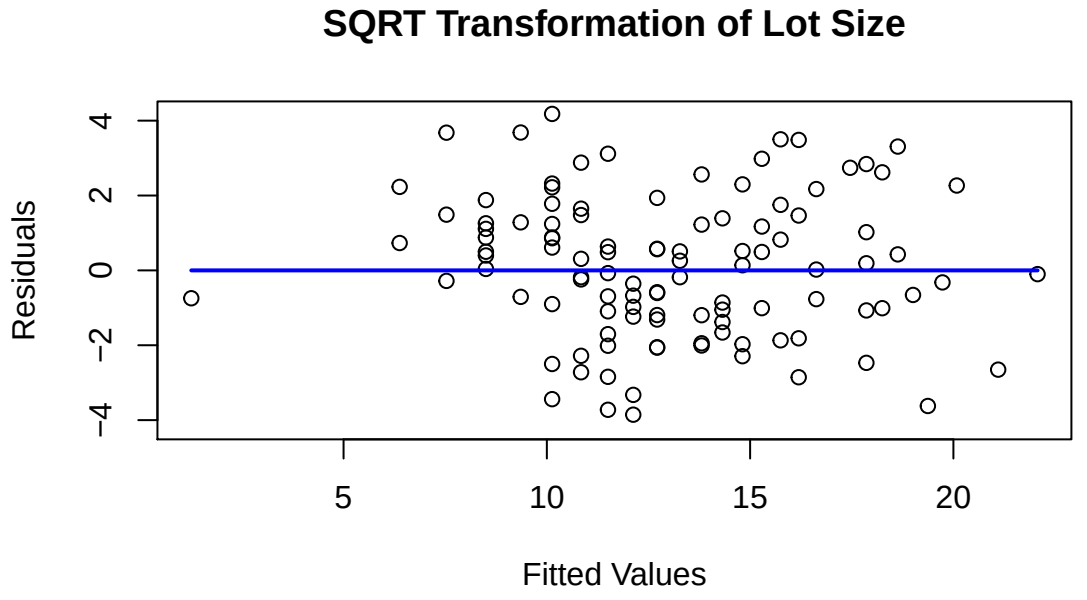

The tranformed explanatory variable does appear be a slightly better linear relation variable than the untransformed version but it is difficult to assess the level of difference.

**D. Obtain the residuals and plot them against the fitted values. Also prepare a normal probability plot. What do your plots show?**

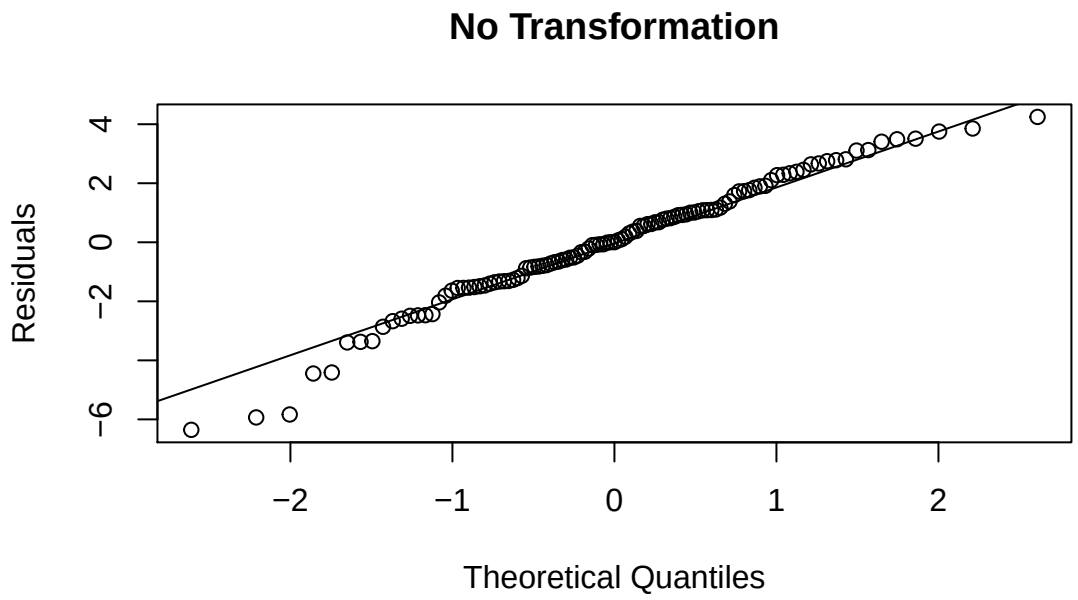
```
with(result, {
     plot(x=fitted.values, y=residuals,
          ylim=c(-max(residuals), max(residuals)),
          xlab="Fitted Values", ylab="Residuals", main="No Tranformation")

     points(c(min(fitted.values), max(fitted.values)),
            c(0,0), type="l", lwd="2", col="blue")
     })
```



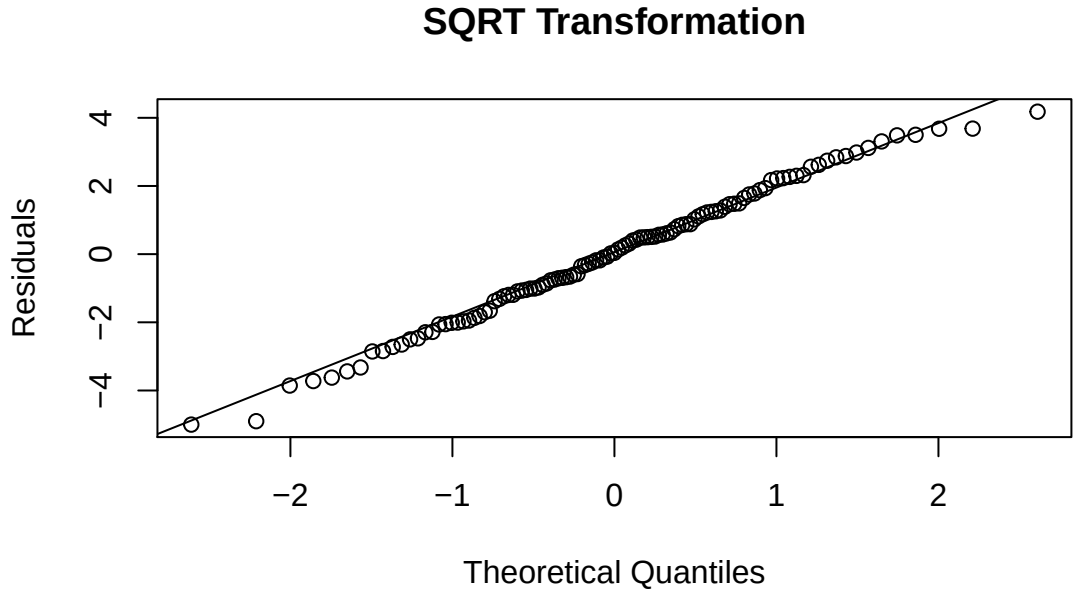

```
with(transformed_result, {
     plot(x=fitted.values, y=residuals,
          ylim=c(-max(residuals), max(residuals)),
          xlab="Fitted Values", ylab="Residuals", main="SQRT Transformation of Lot Size")

     points(c(min(fitted.values), max(fitted.values)),
            c(0,0), type="l", lwd="2", col="blue")
     })
```



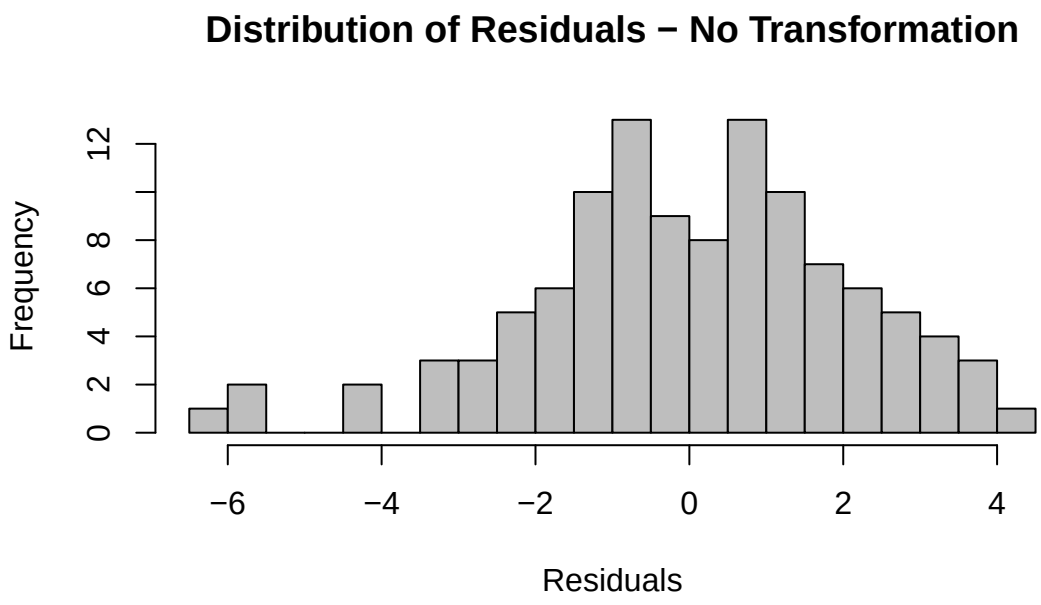
```
qqnorm(result$residuals, ylab="Residuals", main="No Transformation")
qqline(result$residuals)
```



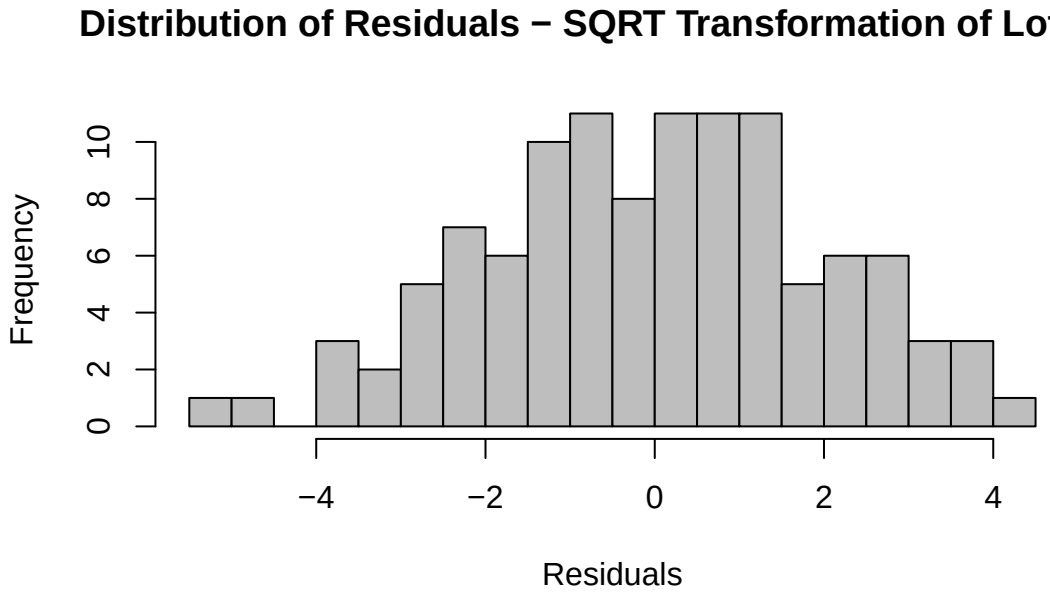

```
qqnorm(transformed_result$residuals, ylab="Residuals", main="SQRT Transformation")
qqline(transformed_result$residuals)
```



```r
hist(result$residuals, main="Distribution of Residuals - No Transformation", xlab="Residuals",
     breaks=20, col="grey")
```




```r
hist(transformed_result$residuals,
     main="Distribution of Residuals - SQRT Transformation of Lot Size", xlab="Residuals",
     breaks=20, col="grey")
```




The plots show that the model that uses tranformed explanatory variable Lot Size is a better choice as it has more normality (better constancy of variance) of the residuals.

**E. Express the estimated regression function in the original units.**

$Hours = 1.2547 + 3.6235\sqrt{(LotSize)}$