205615894_stats101a_hw5

Takao

2/9/2022

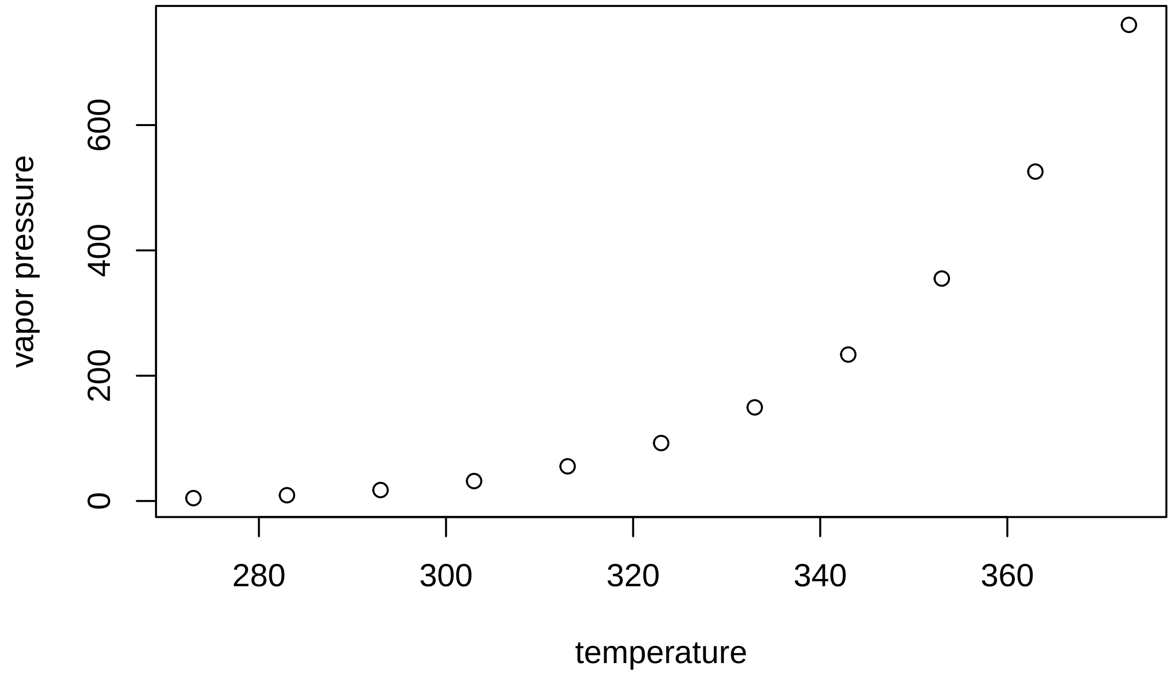
(1)

```
library(readxl)
water_data <- read_excel("water.xlsx", sheet = "Sheet1")
head(water_data)
```

```
## # A tibble: 6 x 3
##   ‘Observation number‘ ‘Temperature (K)‘ ‘Vapor pressure (mm Hg)‘
##                  <dbl>             <dbl>                    <dbl>
## 1                    1               273                      4.6
## 2                    2               283                      9.2
## 3                    3               293                     17.5
## 4                    4               303                     31.8
## 5                    5               313                     55.3
## 6                    6               323                     92.5
```

a

```
plot(water_data$`Vapor pressure (mm Hg)` ~ water_data$`Temperature (K)`, xlab = "temperature", ylab = "
```



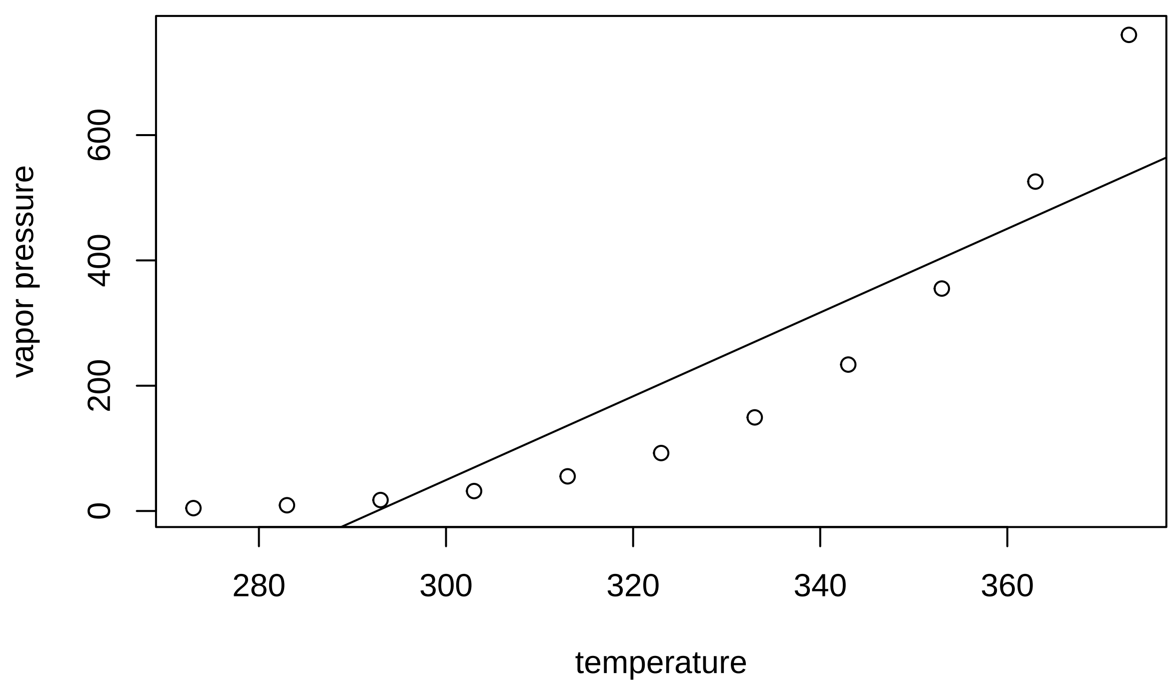
Looking at the scatter diagram, a non-linear relationship seems appropriate in relating pressure (y) to temperature (x).

b

```
slr.model1 <- lm(water_data$`Vapor pressure (mm Hg)` ~ water_data$`Temperature (K)`)

b.coeff1 <- coefficients(slr.model1)

plot(water_data$`Vapor pressure (mm Hg)` ~ water_data$`Temperature (K)`, xlab = "temperature", ylab = "
abline(a = b.coeff1[1], b = b.coeff1[2])
```



c

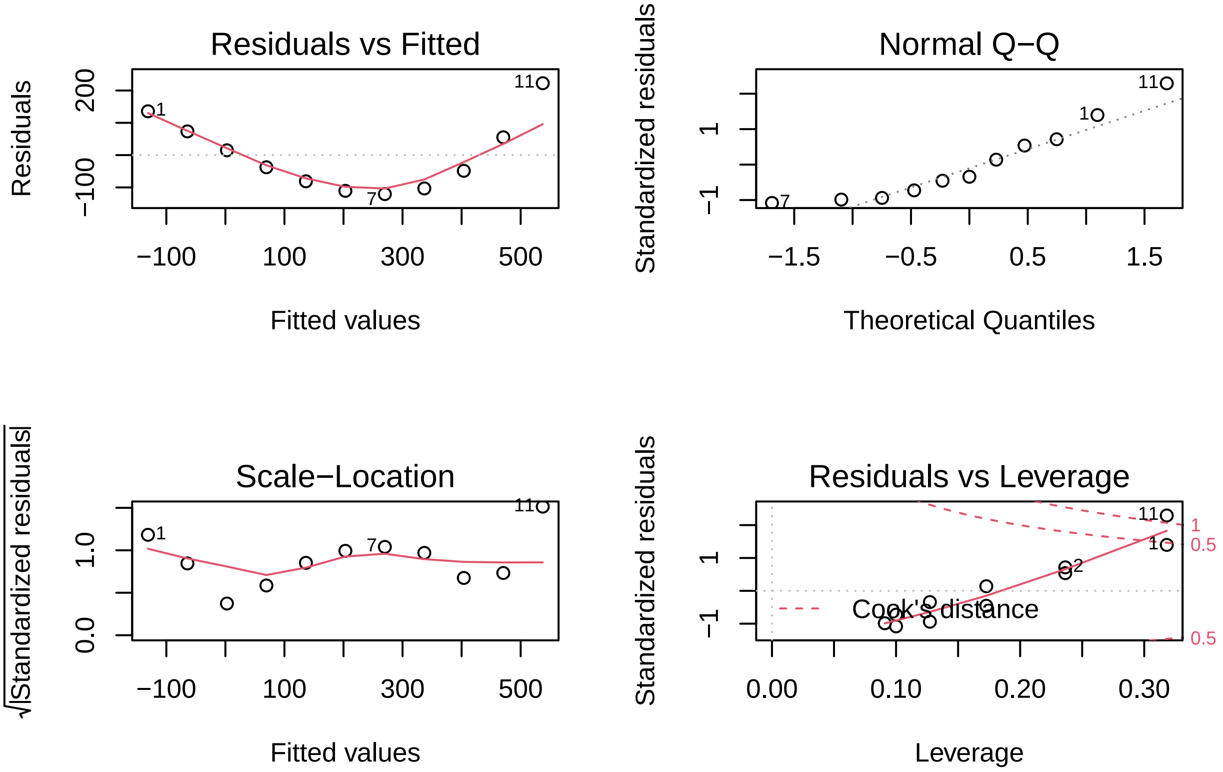
```
anova(slr.model1)
```

```
## Analysis of Variance Table
##
## Response: water_data$‘Vapor pressure (mm Hg)‘
##                              Df Sum Sq Mean Sq F value    Pr(>F)
## water_data$‘Temperature (K)‘  1 491662  491662  35.569 0.0002117 ***
## Residuals                     9 124403   13823
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

The ANOVA table shows a p-value of 0.0002117 Based on a significance level of $\alpha = 0.05$, we have significance evidence to reject the null hypothesis. since $0.0002117 < 0.05$. There is sufficient evidence that the linear regression model is appropriate.

d

```
par(mfrow=c(2,2)); plot(slr.model1)
```



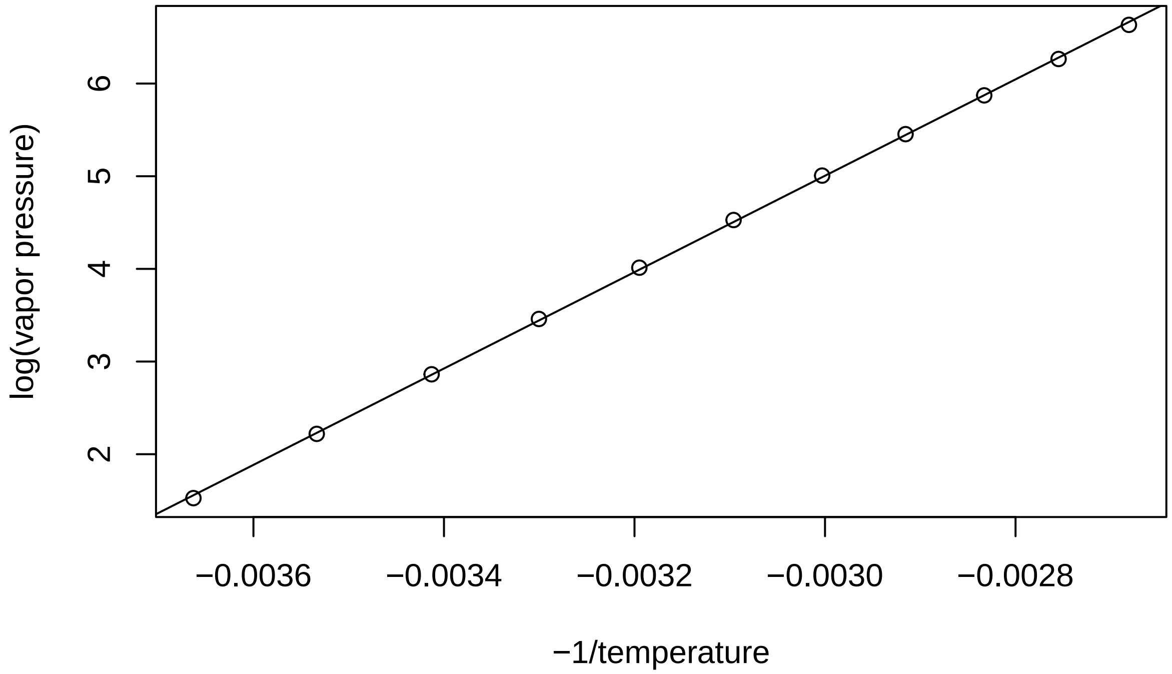
Looking at the residuals vs fitted plot for this data set, we can see that there is a clear trend. The red line does not seem to be straight on residuals = 0 and is curvy with the residuals spreading approximately from -100 to more than a 100. Thus, we can conclude that the linear regression model is not adequate.

e

```
# a
Pv <- log(water_data$`Vapor pressure (mm Hg)`)
ttemp <- (-1/water_data$`Temperature (K)`)

plot(Pv ~ ttemp, xlab = "-1/temperature", ylab = "log(vapor pressure)")

# b
slr.model3 <- lm(Pv ~ ttemp)
b.coeff3 <- coefficients(slr.model3)
plot(Pv ~ ttemp, xlab = "-1/temperature", ylab = "log(vapor pressure)")
abline(a = b.coeff3[1], b = b.coeff3[2])
```

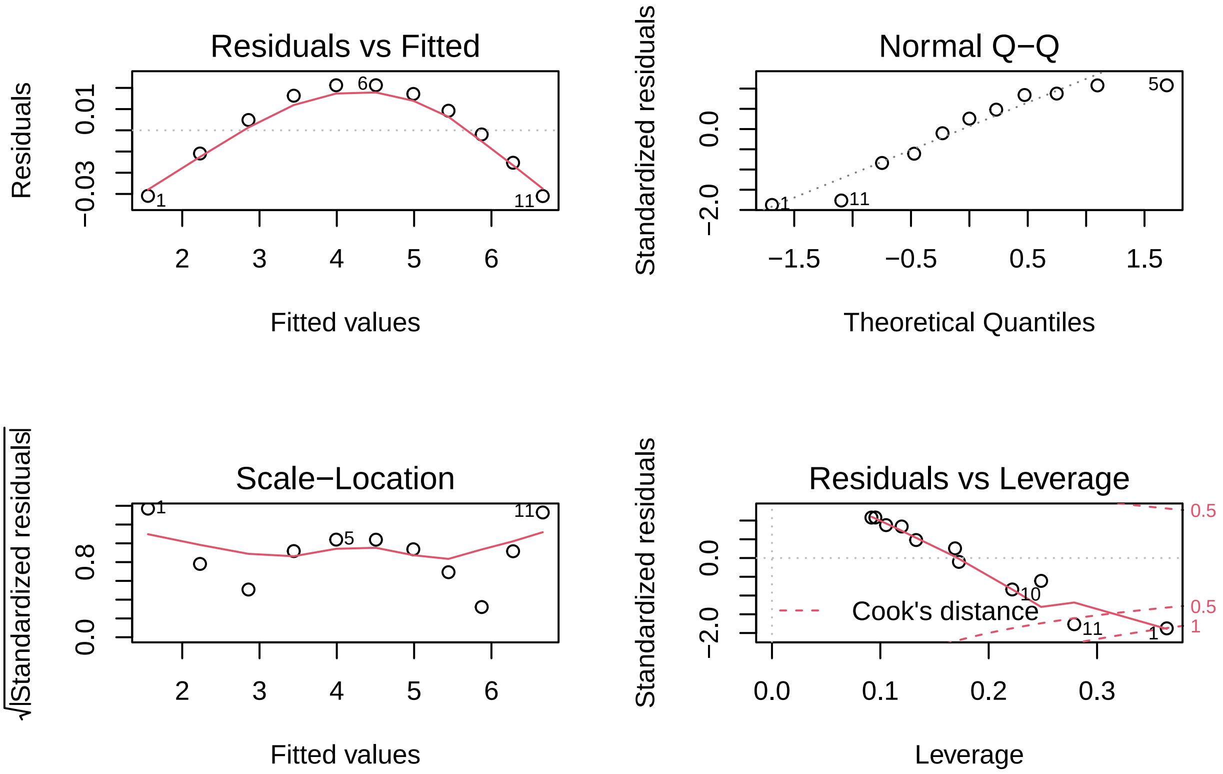


```
# c
anova(slr.model3)
```

```
## Analysis of Variance Table
##
## Response: Pv
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## ttemp      1 28.5110 28.5110   66715 < 2.2e-16 ***
## Residuals  9  0.0038  0.0004
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The ANOVA table shows a p-value of 0.00000000000000022 Based on a significance level of $\alpha = 0.05$, we have significance evidence to reject the null hypothesis. since $0.00000000000000022 < 0.05$. There is sufficient evidence that the new linear regression model is appropriate.

```
# d
par(mfrow=c(2,2)); plot(slr.model3)
```



Looking at the residuals vs fitted plot for the new model, we can see how there is a clear trend. The red line in the plot seems to be in a upside down u shaped. However, if we look at the values of the residuals, we can see that they are very small since the range for the residuals values are vetween approximately -0.03 and 0.03. Thus, although the shape of the graphs indicate otherwise, we can say that this new relationship fits the linear regression model better than the original.

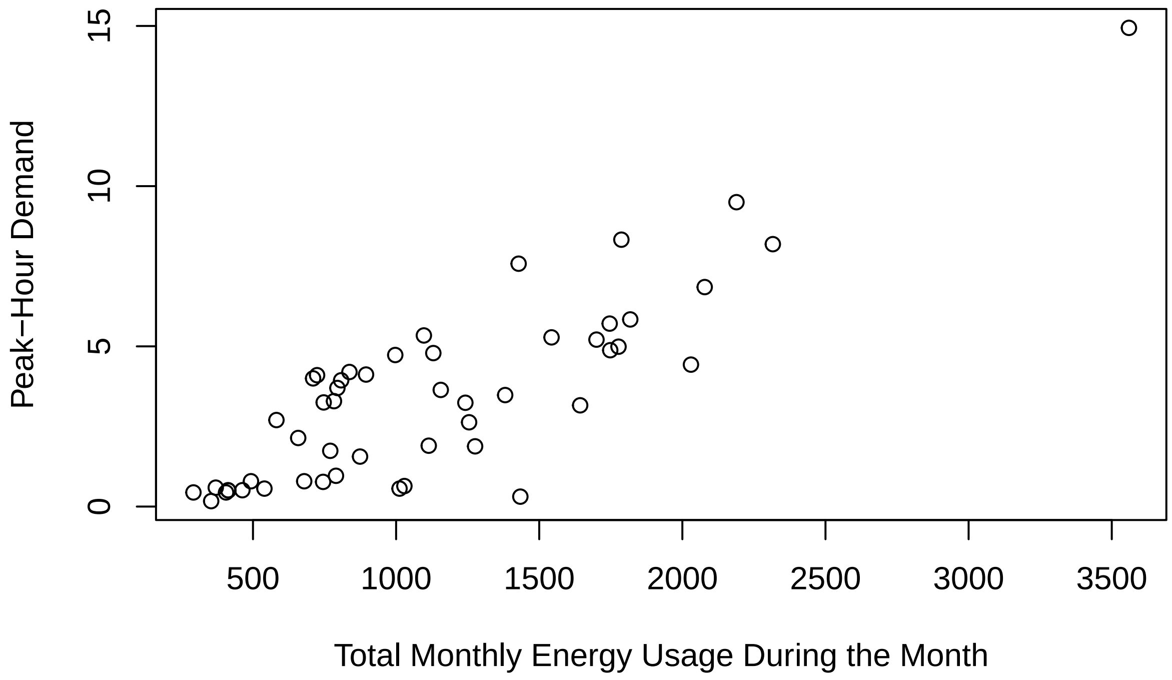
(2)

```
electric_utility_data <- read_excel("electric_utility.xlsx", sheet = "Sheet1")
head(electric_utility_data)
```

```
## # A tibble: 6 x 3
##   Customer     x     y
##      <dbl> <dbl> <dbl>
## 1        1   679  0.79
## 2        2   292  0.44
## 3        3  1012  0.56
## 4        4   493  0.79
## 5        5   582  2.7
## 6        6  1156  3.64
```

a

```
plot(electric_utility_data$y ~ electric_utility_data$x, xlab = "Total Monthly Energy Usage During the M
```

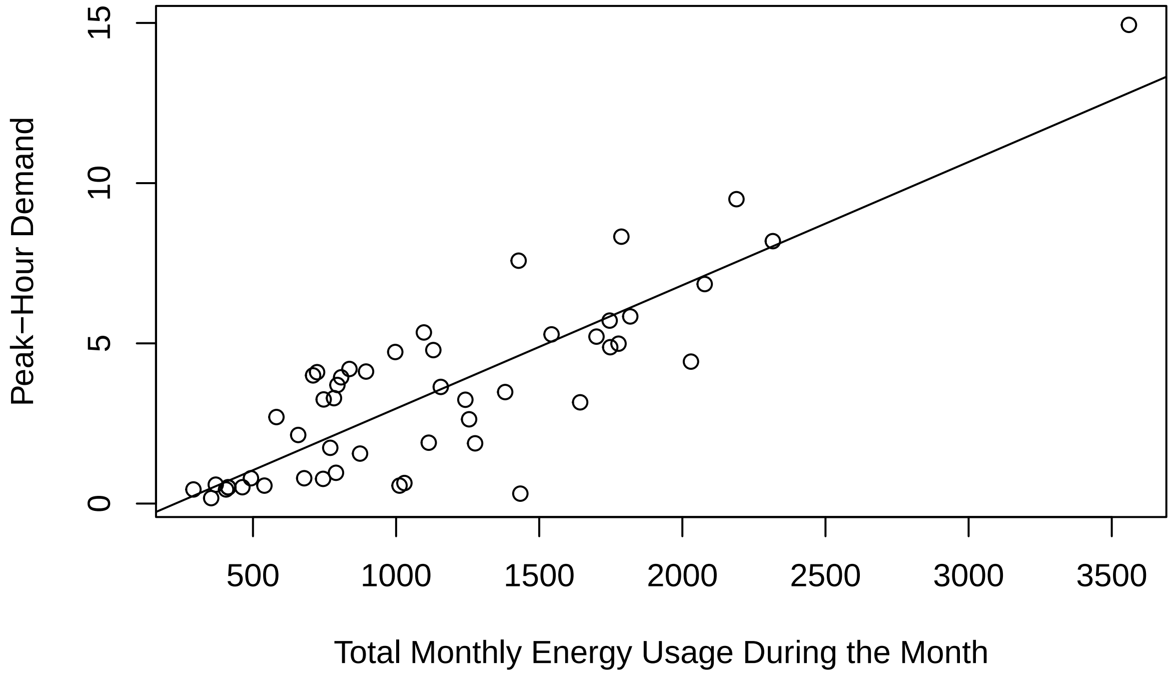


b

```
slr.model2 <- lm(electric_utility_data$y ~ electric_utility_data$x)

b.coeff2 <- coefficients(slr.model2)

plot(electric_utility_data$y ~ electric_utility_data$x, xlab = "Total Monthly Energy Usage During the M
abline(a = b.coeff2[1], b = b.coeff2[2])
```



c

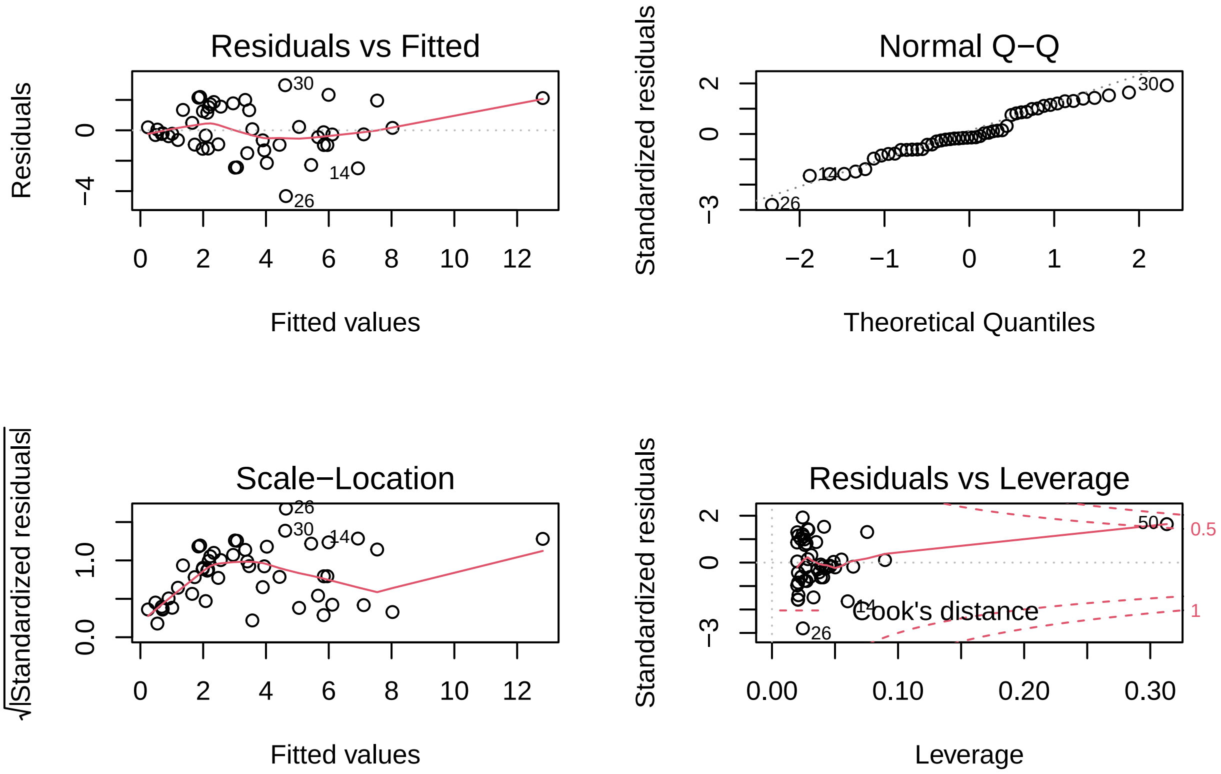
```
anova(slr.model2)
```

```
## Analysis of Variance Table
##
## Response: electric_utility_data$y
##                          Df Sum Sq Mean Sq F value    Pr(>F)
## electric_utility_data$x  1 297.67 297.674  122.03 8.808e-15 ***
## Residuals                48 117.09   2.439
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The ANOVA table shows a p-value of 0.000000000000008808 Based on a significance level of $\alpha = 0.05$, we have significance evidence to reject the null hypothesis. since $0.000000000000008808 < 0.05$. There is sufficient evidence that the linear regression model is appropriate.

d

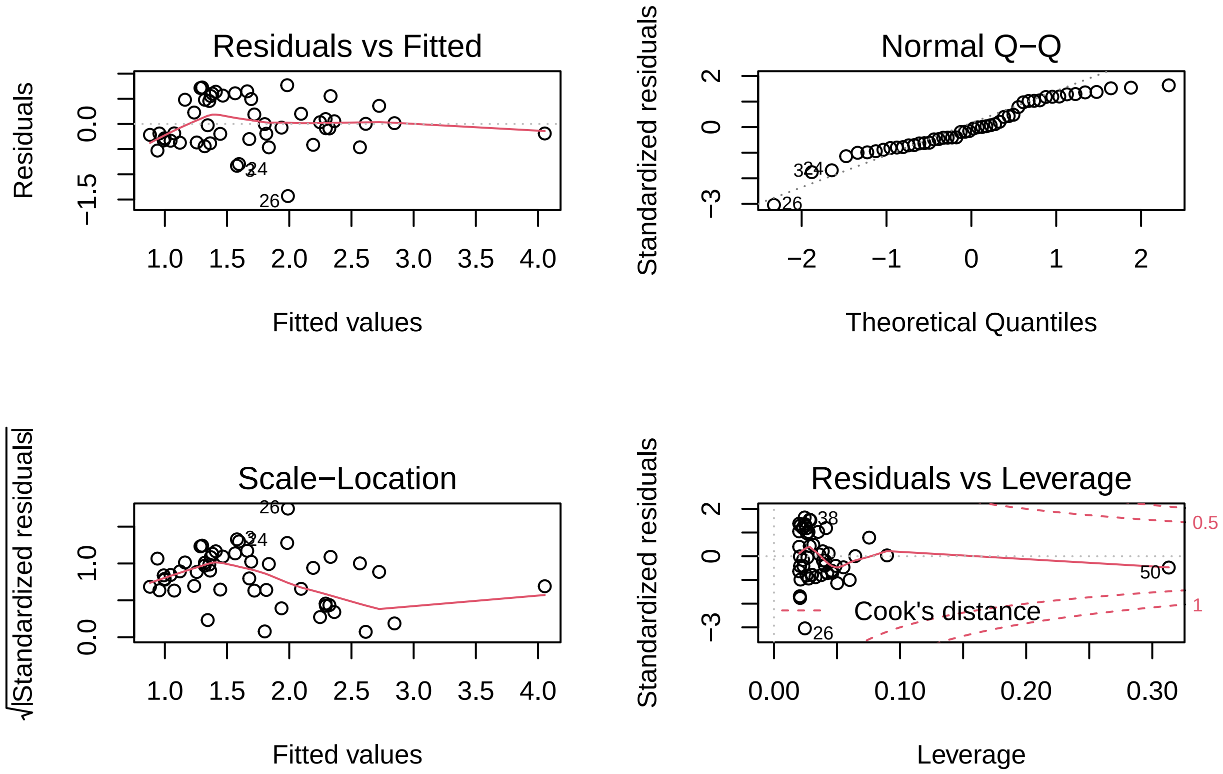
```
par(mfrow=c(2,2)); plot(slr.model2)
```



Looking at the residuals vs fitted plot, there does not seem to be a funnel shape, however, there seems to be a slight trend in the red line in this plot. Because of the trend, there seems to be more residuals that are positive. Thus, we can say that the equality of variance assumption is not completely satisfied. Since the equality of variance assumption is not satisfied, we cannot say that the linear model assumption is adequate for this data set.

e

```
bc.slr1 <- lm((electric_utility_data$y^{0.5})~electric_utility_data$x)
par(mfrow=c(2,2)); plot(bc.slr1)
```



Looking at the residuals vs fitted plot for this newly transformed data set, we can see that there seems to be much less of a trend in the residuals. There also doesn't seem to be a funnel shape indicating that the residuals are not increasing as the fitted values increase. Thus, we can say that the equality of variance assumption is satisfied and the transformation on y using the square-root of y as the response stabilize the inequality of variance problem noted in part d.