205615894_stats101A_hw4

Takao

1/26/2022

Problem 1

```
ss_data <- read.csv("surface_streams.csv")
head(ss_data)
```

```
##      y    x
## 1  4.4 0.19
## 2  6.6 0.15
## 3  9.7 0.57
## 4 10.6 0.70
## 5 10.8 0.67
## 6 10.9 0.63
```

```
dim(ss_data)
```

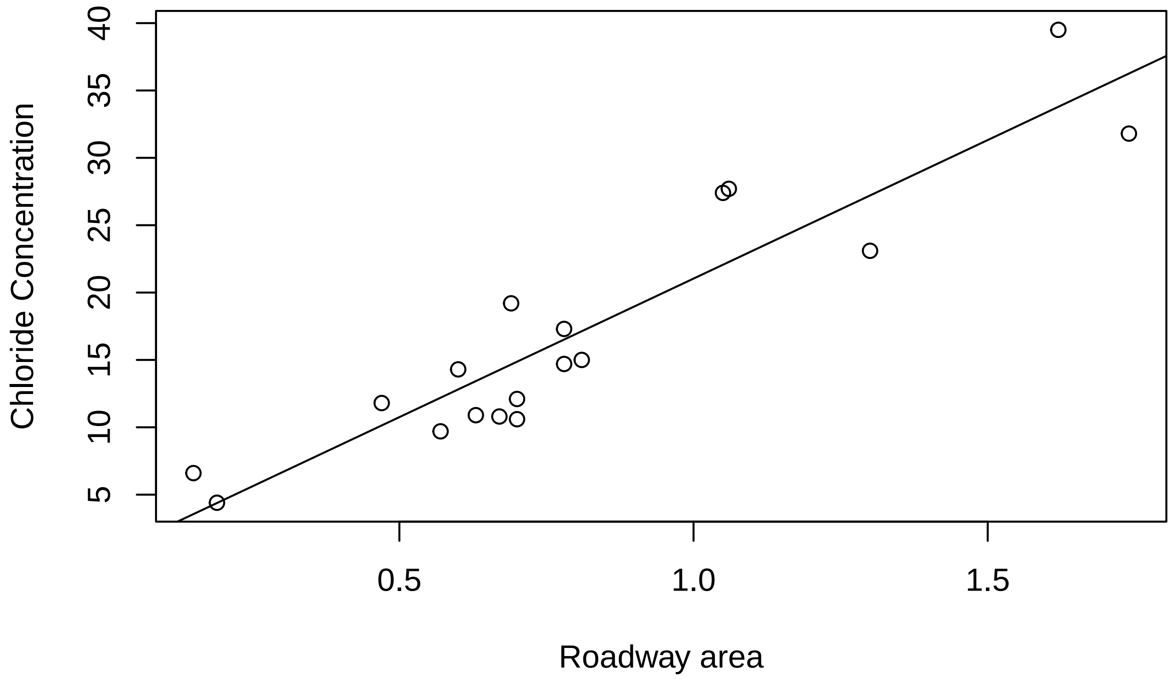
```
## [1] 18  2
```

(a)

```
slr.model1 <- lm(y ~ x, data = ss_data)
slr.model1
```

```
##
## Call:
## lm(formula = y ~ x, data = ss_data)
##
## Coefficients:
## (Intercept)            x
##      0.4705      20.5673
```

```
b.coeff1 <- coefficients(slr.model1)
plot(ss_data$x, ss_data$y, xlab = "Roadway area", ylab = "Chloride Concentration")
abline(a = b.coeff1[1], b = b.coeff1[2])
```



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

```
## Analysis of Variance Table
##
## Response: y
##           Df  Sum Sq Mean Sq F value   Pr(>F)
## x          1 1273.54 1273.54  92.224 4.81e-08 ***
## Residuals 16  220.95   13.81
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

predicted concentration = 20.5673*(roadway area) + 0.4705 (y hat) = 20.5673*x + 0.4705

The p-value for the observed test statistic is 0.0000000481. We have that, at a level of $\alpha = 0.05$, and the observed test statistics is less than the significance level. Thus, we can conclude that the regression model fits the data better than the model with no independent variables.

(b)

```
summary(slr.model1)
```

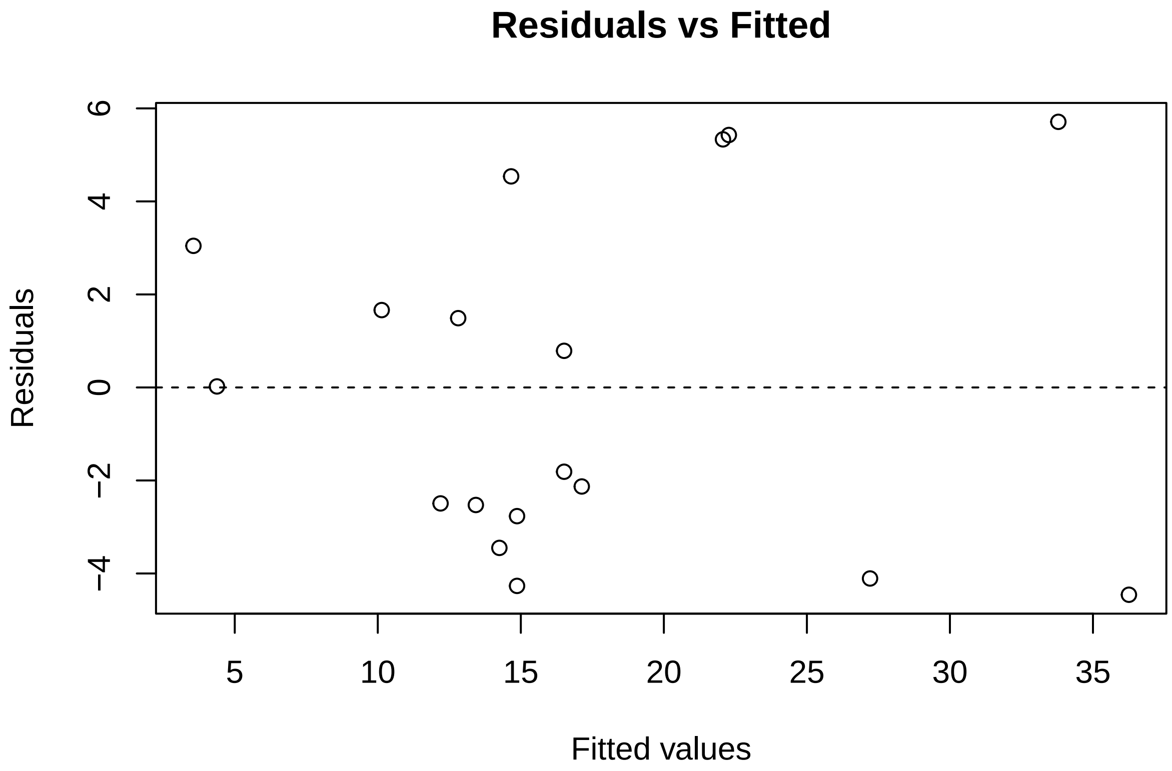
```
##
## Call:
```

```
## lm(formula = y ~ x, data = ss_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.4576 -2.7077 -0.8956  2.6991  5.7105
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.4705     1.9359   0.243    0.811
## x            20.5673     2.1417   9.603 4.81e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.716 on 16 degrees of freedom
## Multiple R-squared:  0.8522, Adjusted R-squared:  0.8429
## F-statistic: 92.22 on 1 and 16 DF,  p-value: 4.81e-08
```

The value of R^2 is shown at the output "Multiple R-squared: 0.8522" This means that the model $Y = \beta_0 + \beta_1 x + \epsilon$ explains 85.22% of the total variability in the chloride concentration.

(c)

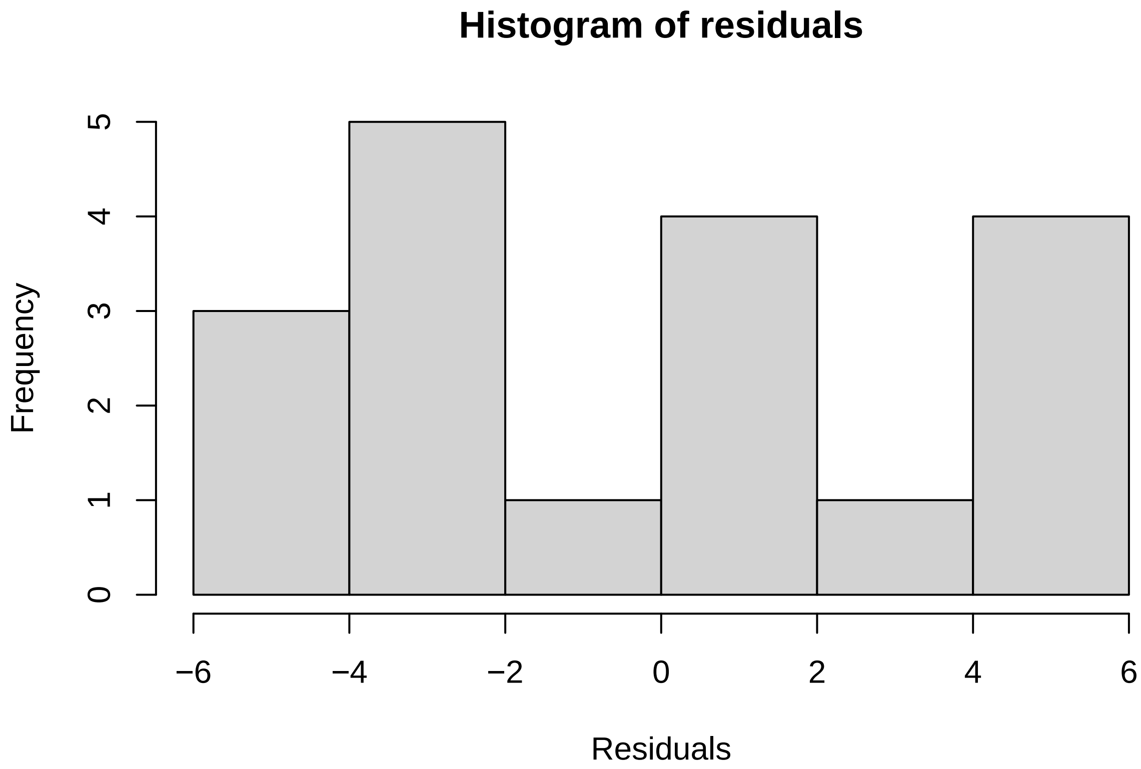
```
y.hat1 <- slr.model1$fitted.values
res1 <- slr.model1$residuals
plot(y.hat1, res1, xlab = 'Fitted values', ylab = 'Residuals',
     main = 'Residuals vs Fitted')
abline(h = 0, lty = 'dashed') # Add reference line.
```



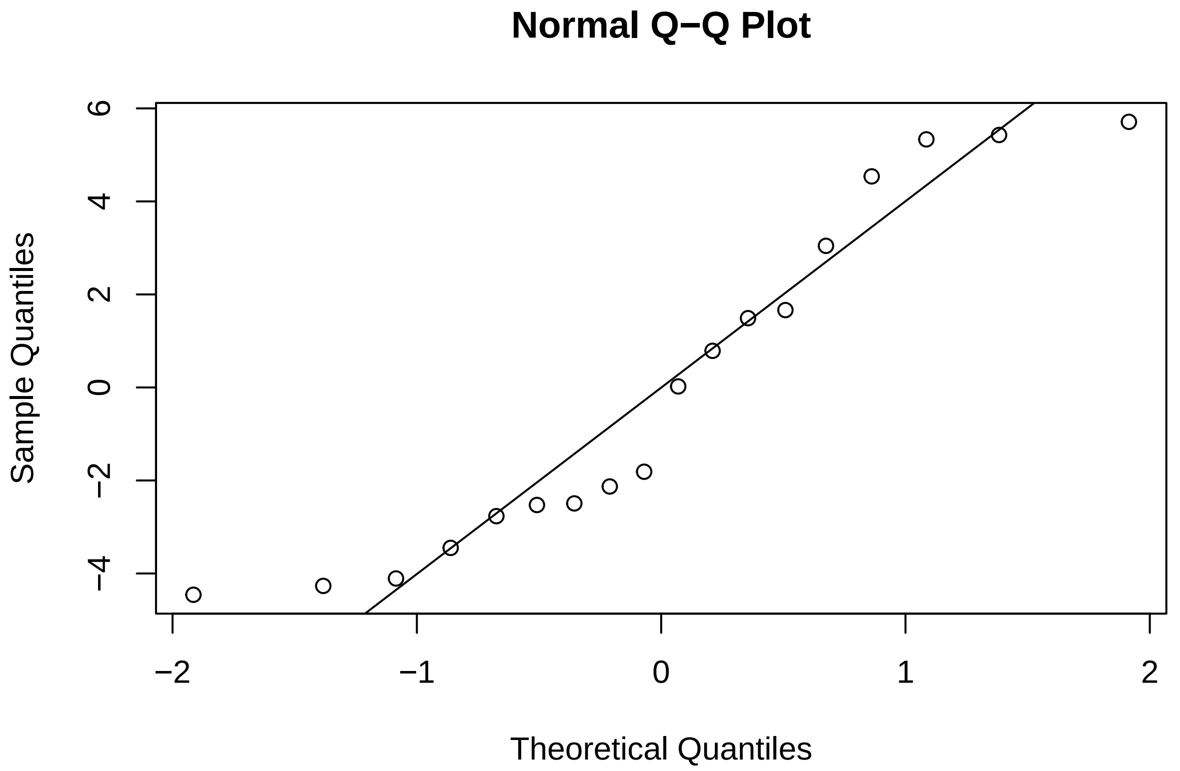
To conclude that the plot supports that the residuals have constant variance, there must not be any trends or patterns in the points. However, in this model, it seems as if the points are in a slight funnel shape indicating that there is non-constant variance. Thus, the assumption of constant variance does not seem to be satisfied.

(d)

```
hist(res1, main = "Histogram of residuals", xlab = 'Residuals')
```



```
qqnorm(res1)
qqline(res1)
```



The observations in the middle are close to the qqline, however, the end points branches out towards the extreme which does not support the normality assumption in the data, and the way the points are distributed along the line makes the data seem to be more uniformly distributed.

(e)

In part (a), we concluded that there is significance of the regression model. However, in part (c) we found out that the assumption of constant variance does not seem to be satisfied. Additionally, in part (d) we found that the normality assumption does not appear to be satisfied. Since these two assumptions are not satisfied, the prediction for the significance of the regression model will be inaccurate and the estimates of the intercept and slopes will tend to be biased.

Problem 2

```
library(readxl)
pine_data <- read_excel("pine.xlsx", sheet = "Sheet1")
```

```
## New names:
## * ‘Compressive Strength‘ -> ‘Compressive Strength...1‘
## * Density -> Density...2
## * ‘Compressive Strength‘ -> ‘Compressive Strength...3‘
## * Density -> Density...4
```

```
pine_data <- data.frame(pine_data)
head(pine_data)
```

```
##   Compressive.Strength...1 Density...2 Compressive.Strength...3 Density...4
## 1                     3040        29.2                     3840        30.7
## 2                     2470        24.7                     3800        32.7
## 3                     3610        32.3                     4600        32.6
## 4                     3480        31.3                     1900        22.1
## 5                     3810        31.5                     2530        25.3
## 6                     2330        24.5                     2920        30.8
```

```
typeof(pine_data)
```

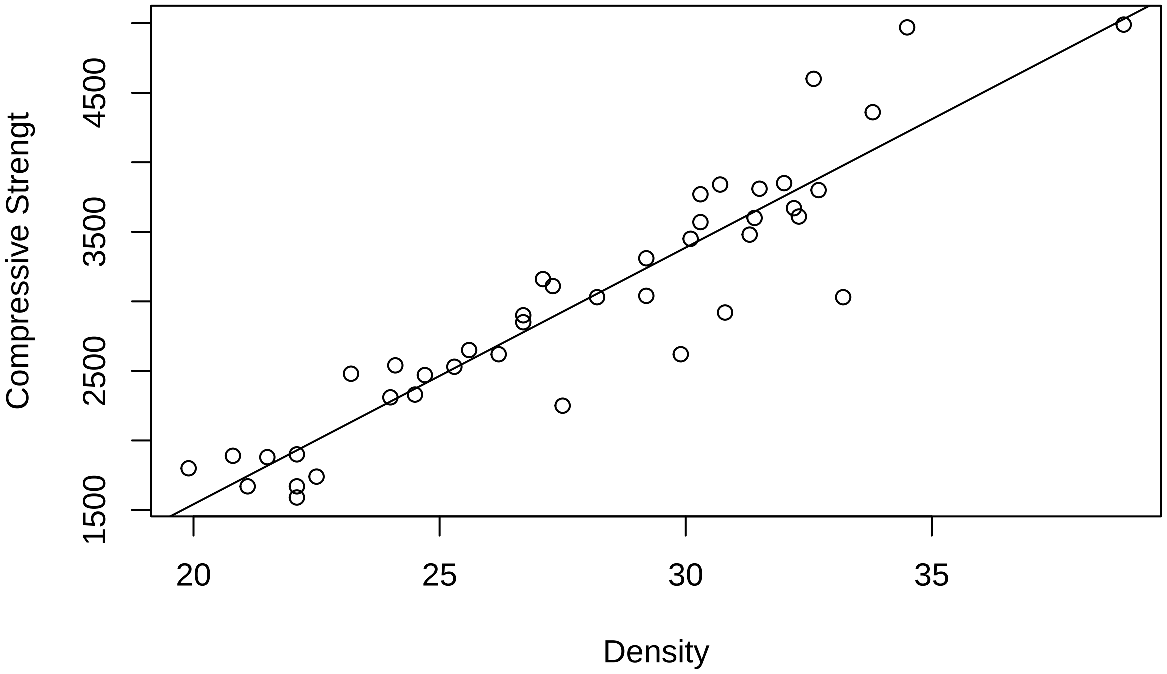
```
## [1] "list"
```

(a)

```
slr.model2 <- lm(c(pine_data[, 1], pine_data[, 3]) ~ c(pine_data[, 2], pine_data[, 4]))
slr.model2
```

```
## 
## Call:
## lm(formula = c(pine_data[, 1], pine_data[, 3]) ~ c(pine_data[, 
##     2], pine_data[, 4]))
## 
## Coefficients:
##                       (Intercept)  c(pine_data[, 2], pine_data[, 4])  
##                           -2149.6                              184.6
```

```
b.coeff2 <- coefficients(slr.model2)
plot(c(pine_data[, 1], pine_data[, 3]) ~ c(pine_data[, 2], pine_data[, 4]), xlab = "Density", ylab = "C
abline(a = b.coeff2[1], b = b.coeff2[2])
```



```
summary(slr.model2)
```

```
##
## Call:
## lm(formula = c(pine_data[, 1], pine_data[, 3]) ~ c(pine_data[,
##     2], pine_data[, 4]))
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -947.5 -113.5   37.5  187.3  752.6
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      -2149.65     332.52  -6.465 1.05e-07 ***
## c(pine_data[, 2], pine_data[, 4])   184.55      11.79  15.657  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 339.2 on 40 degrees of freedom
## Multiple R-squared:  0.8597, Adjusted R-squared:  0.8562
## F-statistic: 245.2 on 1 and 40 DF,  p-value: < 2.2e-16
```

predicted compressive strength = 184.6*(density) - 2149.6 (y hat) = 184.6*x - 2149.6

```
summary(slr.model2)
```

```
##
## Call:
## lm(formula = c(pine_data[, 1], pine_data[, 3]) ~ c(pine_data[,
##     2], pine_data[, 4]))
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -947.5 -113.5   37.5  187.3  752.6
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       -2149.65     332.52  -6.465 1.05e-07 ***
## c(pine_data[, 2], pine_data[, 4])   184.55      11.79  15.657  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 339.2 on 40 degrees of freedom
## Multiple R-squared:  0.8597, Adjusted R-squared:  0.8562
## F-statistic: 245.2 on 1 and 40 DF,  p-value: < 2.2e-16
```

t statistic = 15.657

To reject null hypothesis, the test statistic must be greater than critical value, thus to find critical value,

```
qt(1 - 0.025, 42 - 2)
```

```
## [1] 2.021075
```

critical value = 2.021075

Since the test statistic is greater than the critical value, we have significant evidence to reject the null hypothesis. Therefore, we conclude that there is an association between the density and the compressive strength.

(b)

```
summary(slr.model2)
```

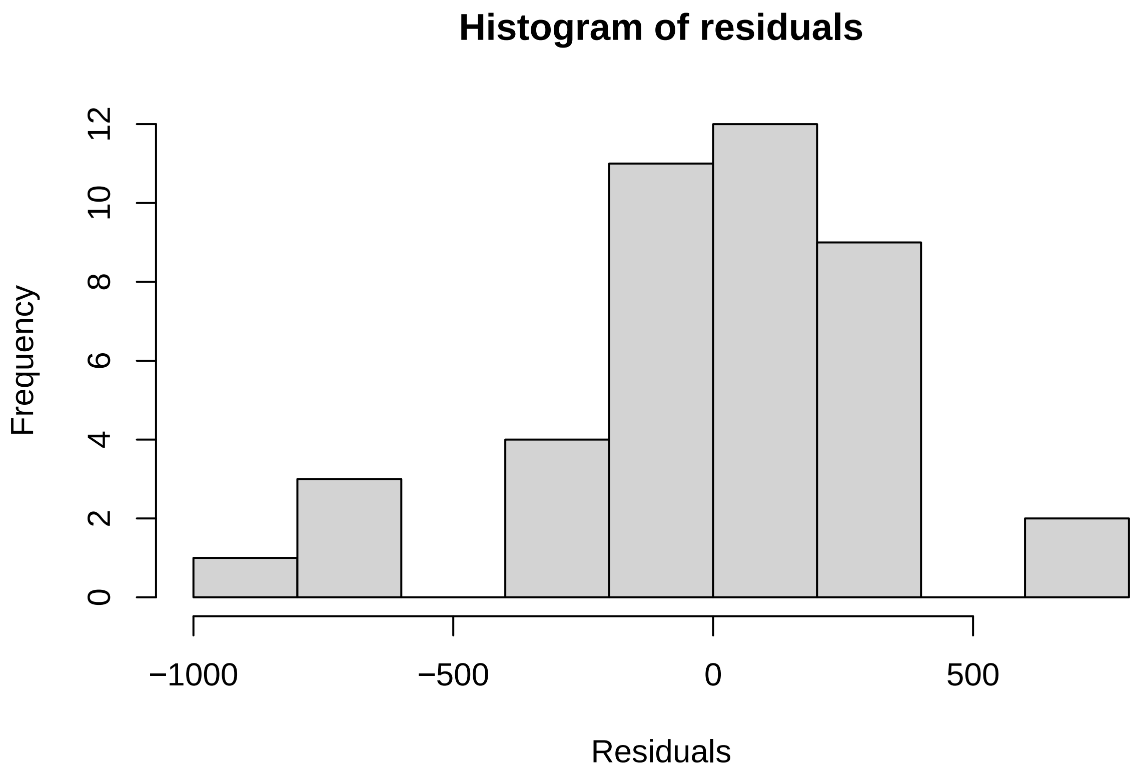
```
##
## Call:
## lm(formula = c(pine_data[, 1], pine_data[, 3]) ~ c(pine_data[,
##     2], pine_data[, 4]))
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -947.5 -113.5   37.5  187.3  752.6
##
## Coefficients:
```

```
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                        -2149.65     332.52  -6.465 1.05e-07 ***
## c(pine_data[, 2], pine_data[, 4])   184.55      11.79  15.657  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 339.2 on 40 degrees of freedom
## Multiple R-squared:  0.8597, Adjusted R-squared:  0.8562
## F-statistic: 245.2 on 1 and 40 DF,  p-value: < 2.2e-16
```

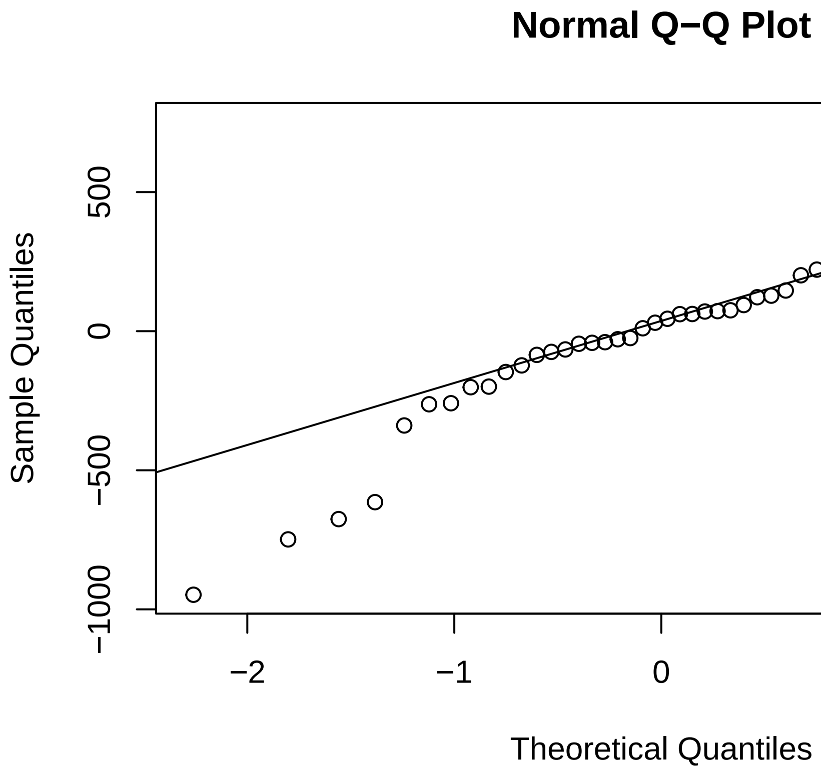
The value of R^2 is shown at the output "Multiple R-squared: 0.8597" This means that the model $Y = \beta_0 + \beta_1 x + \epsilon$ explains 85.97% of the total variability in the compressive strength.

(c)

```
res2 <- slr.model2$residuals
hist(res2, main = "Histogram of residuals", xlab = 'Residuals')
```



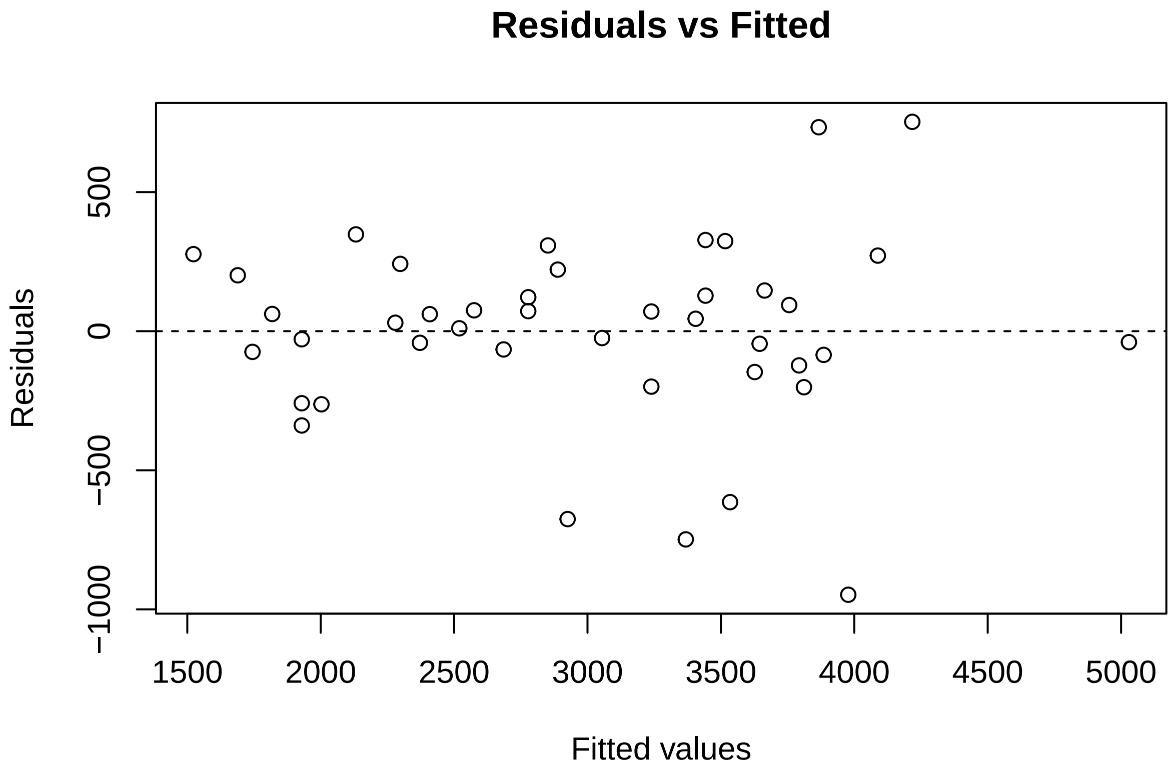
```
qqnorm(res2)
qqline(res2)
```



The distribution will not be normal if there is substantial departures from the straight line to the points. However, we can see that the majority of the plots fit closely with the qqline, and does not show any patterns that indicate that the data follows other types of distribution, which indicates that these data is approximately normal and thus the normality assumption is satisfied.

(d)

```r
y.hat2 <- slr.model2$fitted.values
plot(y.hat2, res2, xlab = 'Fitted values', ylab = 'Residuals',
     main = 'Residuals vs Fitted')
abline(h = 0, lty = 'dashed') # Add reference line.
```



To conclude that the plot supports that the residuals have constant variance, there must not be any trends or patterns in the points. In this plot, there does not seem to be any trends or patterns in the points. We conclude that the plot supports that the residuals have constant variance.

(e)

In part (a), we concluded that there is an association between the density and the compressive strength. We concluded from part (b) that the $Y = \beta_0 + \beta_1 x + \epsilon$ model explains a high percentage of variability in the compressive strength. In part (c), we concluded that the data is approximately normal. In part (d), we satisfied the assumption of constant variance. Thus, we satisfied all of the assumptions, and can state that the statement that we made in part (a) is reliable conclusion.