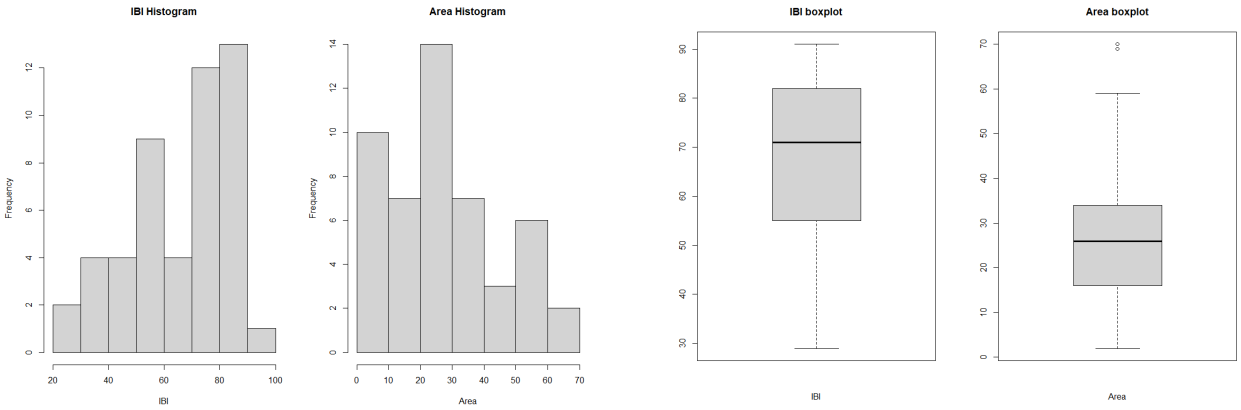
MA 331 ~ Hw #8

10.32

Part A:

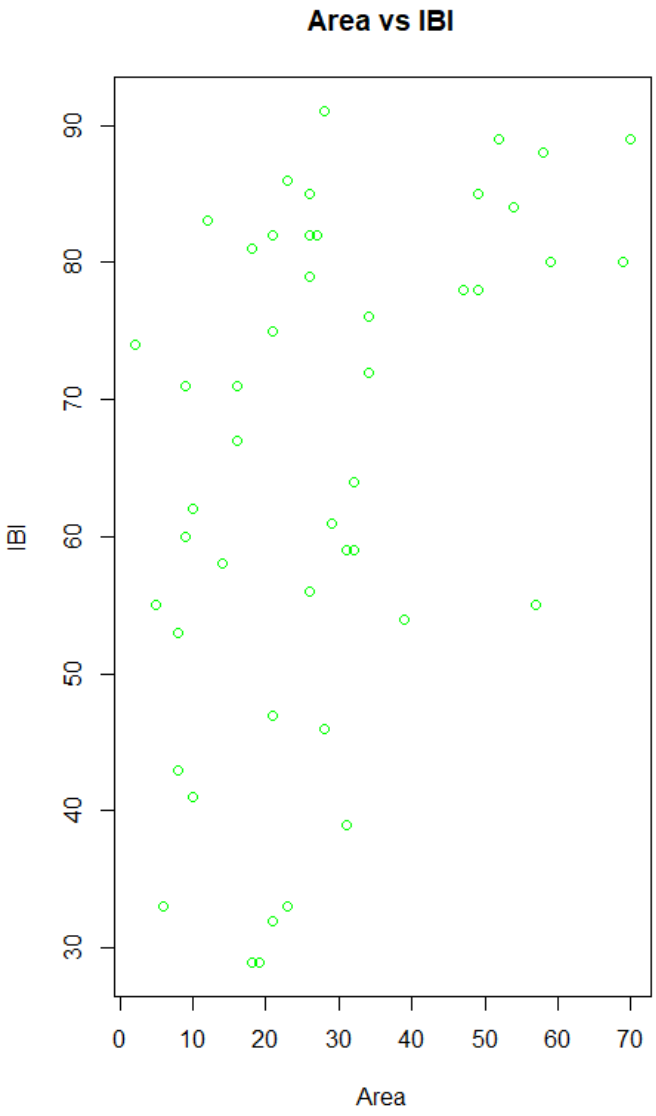
```
> cat("Mean of IBI: ", mean_IBI, sep = "\n")
Mean of IBI:
65.93878
> cat("SD of IBI: ", sd_IBI, sep = "\n")
SD of IBI:
18.27955
> cat("Variance of IBI: ", var_IBI, sep = "\n")
Variance of IBI:
334.142
>
> cat("Mean of Area: ", mean_Area, sep = "\n")
Mean of Area:
28.28571
> cat("SD of Area: ", sd_Area, sep = "\n")
SD of Area:
17.71417
> cat("Variance of Area: ", var_Area, sep = "\n")
Variance of Area:
313.7917
```

```
> # Based on plots:
> # IBI distribution appears to be left skewed
> # Area distribution appears to be right skewed
```



Part B:

Plot seems to suggest a positive moderately



strong linear association between Area & IBI.

Potential Outlier: (60, 54)

Part C:

$yi = \beta 0 + \beta 1 xi + \varepsilon i, i = 1, 2, ..., 49$

Where *xi* = Area in km^2 and *yi* = IBI, for the *ith data point*

Part D:

```
#Part D:
# H0: There is no relationship between IBI & Area --> B1 = 0
# Ha: There is a relationship between IBI & Area --> B1 != 0
```

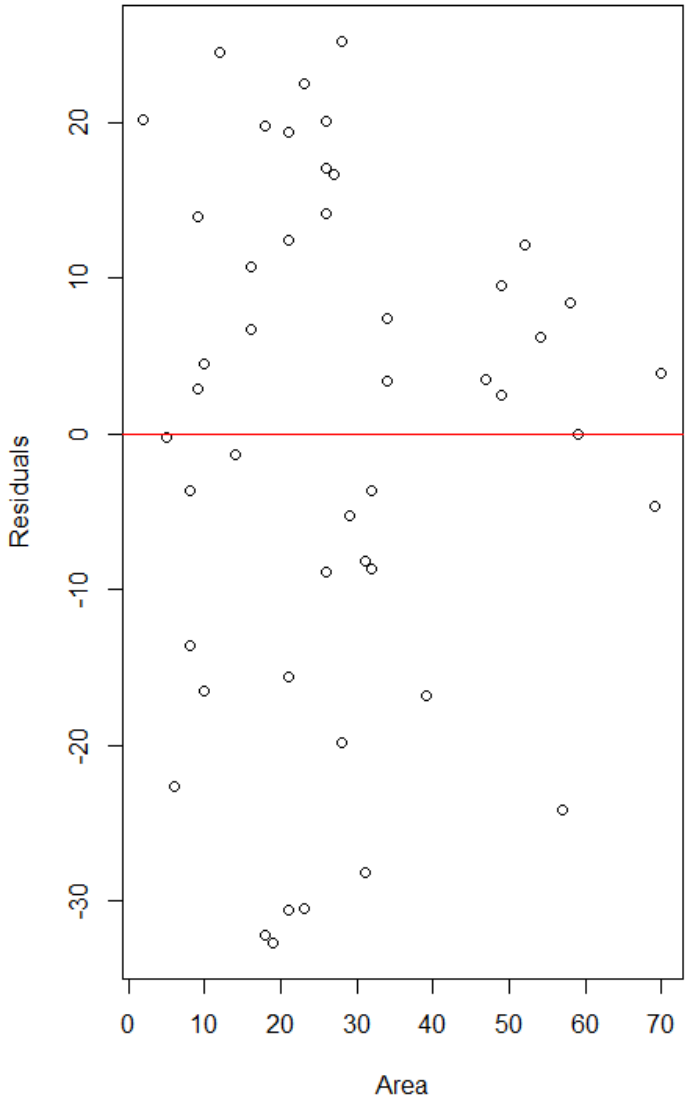
Part E:

```
lm(formula = IBI ~ Area, data = waterquality)

Residuals:
    Min      1Q  Median      3Q     Max
-32.666  -8.887   3.432  12.414  25.193

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  52.9230     4.4835  11.804 1.17e-15 ***
Area          0.4602     0.1347   3.415  0.00132 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16.53 on 47 degrees of freedom
Multiple R-squared:  0.1988,    Adjusted R-squared:  0.1818
F-statistic: 11.67 on 1 and 47 DF,  p-value: 0.001322

> cat(" y = 52.923 + 0.4602 * x", "B0 = 52.923 | B1 = 0.4602", sep = "\n " )
 y = 52.923 + 0.4602 * x
 B0 = 52.923 | B1 = 0.4602
>
> # Given that the p-value for the correlation = 0.001322 << a = 0.05 --> We reject the null and have significa
nt evidence to conclude there exists
> # a linear relationship between Area & IBI.
```

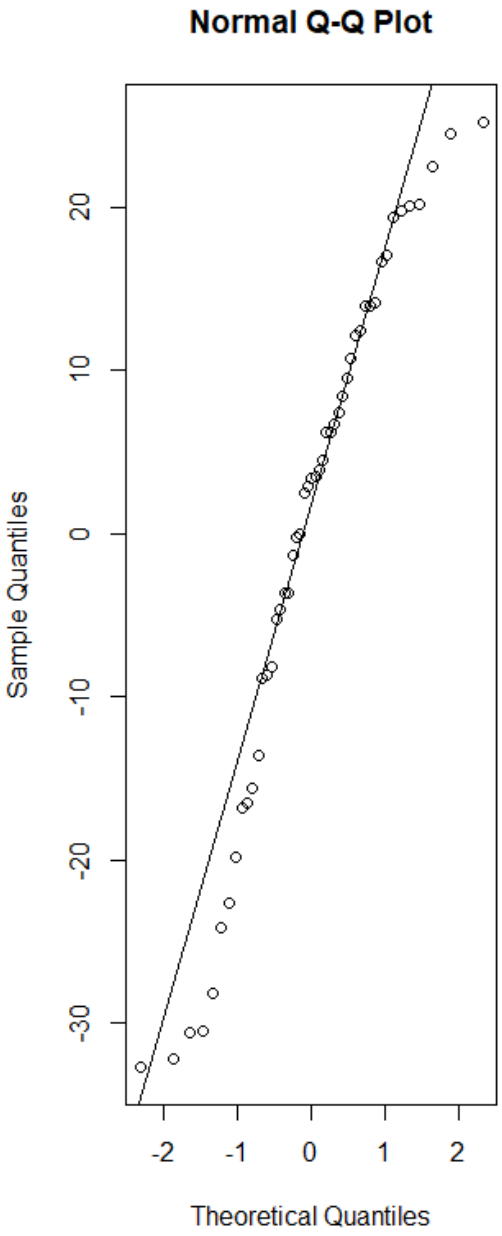
Part F:

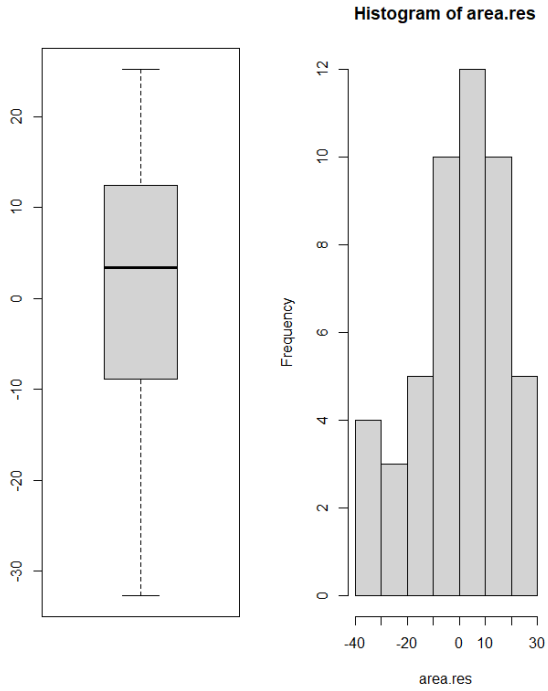
There is nothing unusual about the plot, they seem randomly scattered around 0, with no particular pattern. We can assume independence.

As there doesn't seem any clear pattern amongst the points we can assume linearity.



Part G:

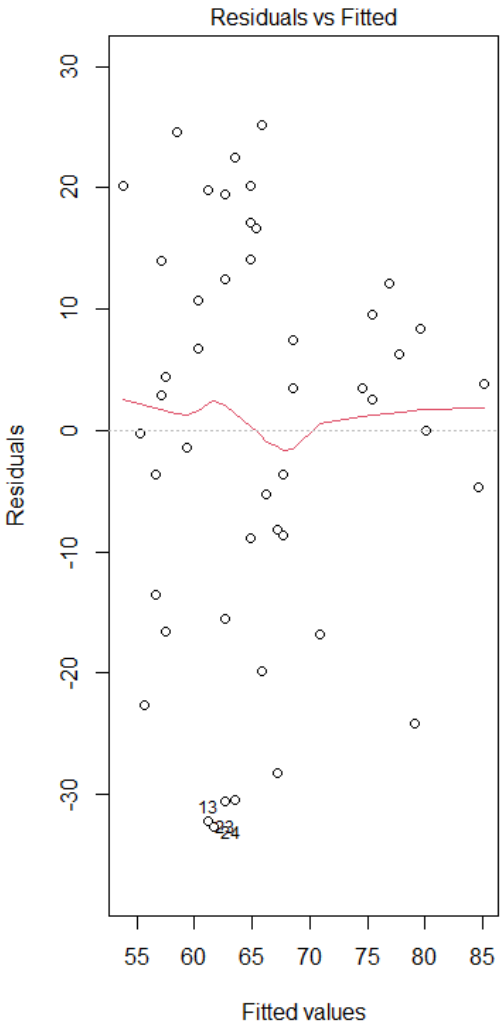


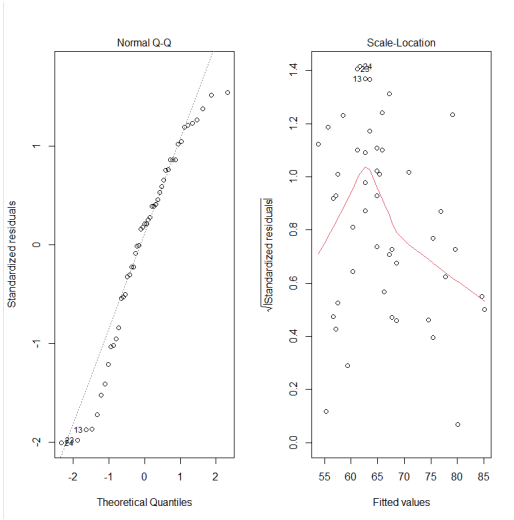


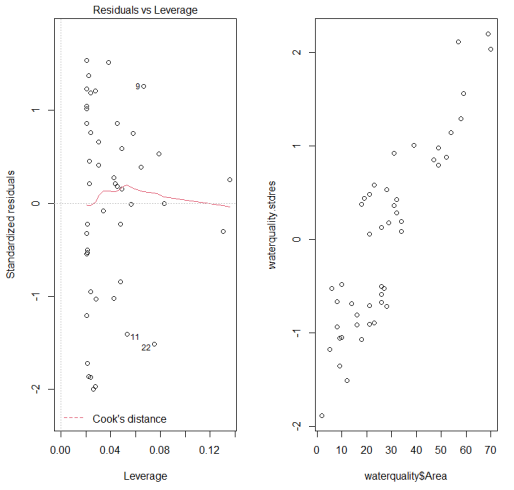
According to the qqnorm plot of waterquality's residuals, the residuals appear to closely fit the qqline, thus they are apx. normal

Part H:

```
#Checking Linearity (Residuals vs Fitted) && (Std Residual vs Area)
#Shown linearity in part F

#Checking Homoscedasticity (Residuals vs Fitted) **no pattern / equally spread around y = 0
#Since points of of Residuals vs Fitted seem to be randomly distributed we assume independence.

#checking Independence (Assumed by data set)
#Shown independence in part F

#Checking normality (Normal Q-Q)
#Shown normality in part G.

#All checks are good so all the assumptions made in part C are valid.
```



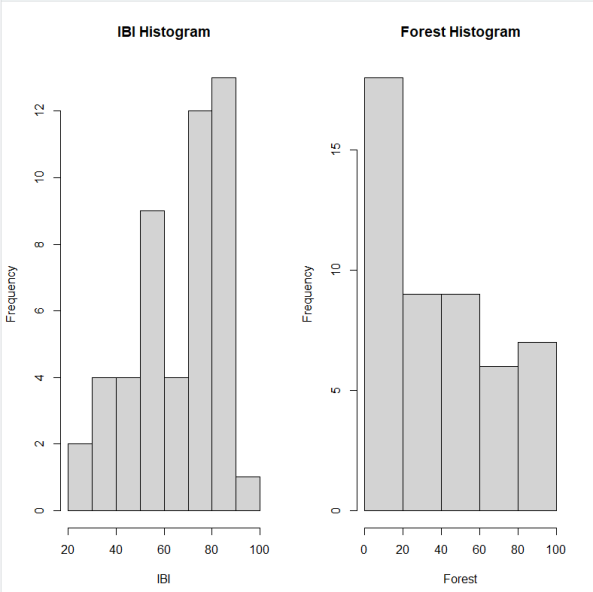




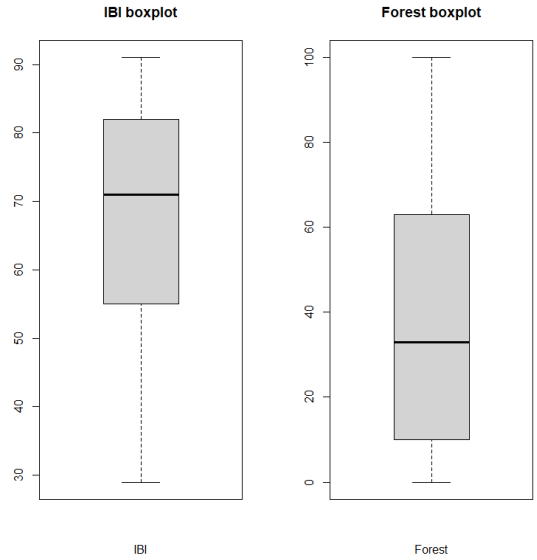
#10.33

Part A:

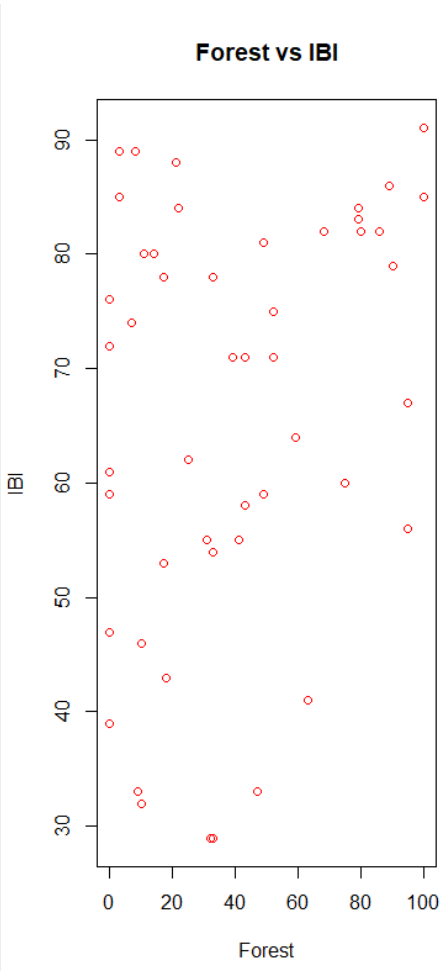
```
# Based on plots:
# IBI appears to be left skewed
# Forest % appears to be right skewed
```




```
> cat("Mean of IBI: ", mean_IBI, sep = "\n")
Mean of IBI:
65.93878
> cat("SD of IBI: ", sd_IBI, sep = "\n")
SD of IBI:
18.27955
> cat("Variance of IBI: ", var_IBI, sep = "\n")
Variance of IBI:
334.142
>
> cat("Mean of Area: ", mean_Forest, sep = "\n")
Mean of Area:
39.38776
> cat("SD of Area: ", sd_Forest, sep = "\n")
SD of Area:
32.20431
> cat("Variance of Area: ", var_Forest, sep = "\n")
Variance of Area:
1037.117
```



Part B:



Plot seems to suggest a weak linear relationship between Forest & IBI, with more outliers for small values of Forest %.

Part C:

$yi = \beta 0 + \beta 1xi + \varepsilon i, i = 1, 2, ..., 49$

Where *xi* = Area in km^2 and *yi* = IBI, for the *ith data point*

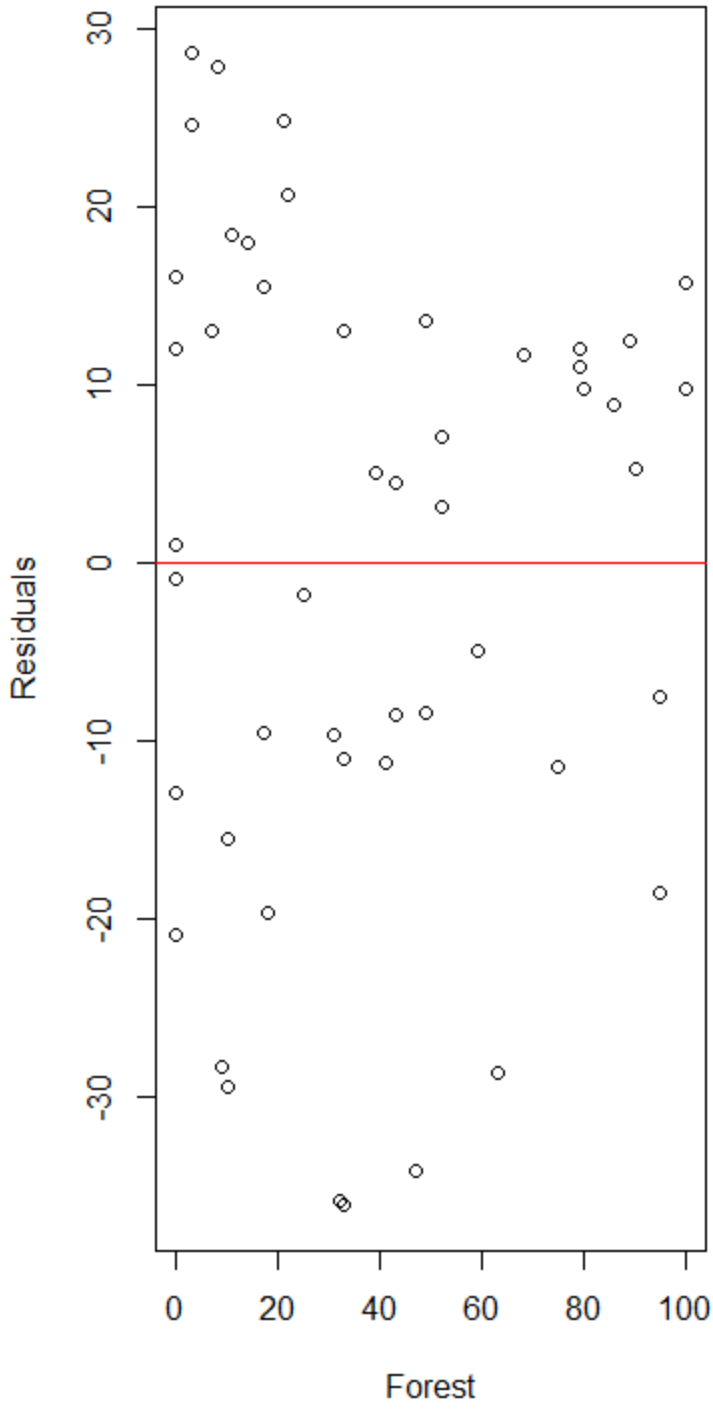
Part D:

```
#Part D:
# H0: There is no relationship between IBI & Forest --> B1 = 0
# Ha: There is a relationship between IBI & Forest --> B1 != 0
```

Part E:

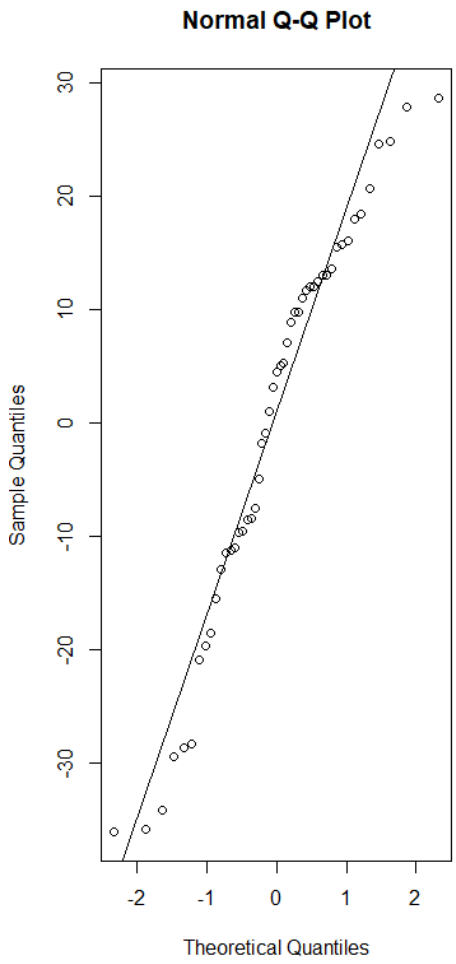
```
Residuals:
    Min      1Q  Median      3Q     Max
-35.961 -11.186   4.508  13.021  28.633

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 59.90725    4.03957  14.830   <2e-16 ***
Forest       0.15313    0.07972   1.921   0.0608 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 17.79 on 47 degrees of freedom
Multiple R-squared:  0.07278,   Adjusted R-squared:  0.05305
F-statistic: 3.689 on 1 and 47 DF,  p-value: 0.06084

>
> cat(" y = 59.90725 + 0.15313 * x", "B0 = 59.90725 | B1 = 0.15313", sep = "\n " )
 y = 59.90725 + 0.15313 * x
 B0 = 59.90725 | B1 = 0.15313
>
> # Given that the p-value for the correlation = 0.0608 > a = 0.05 --> We do not
> # have significant evidence to reject the null
>
```

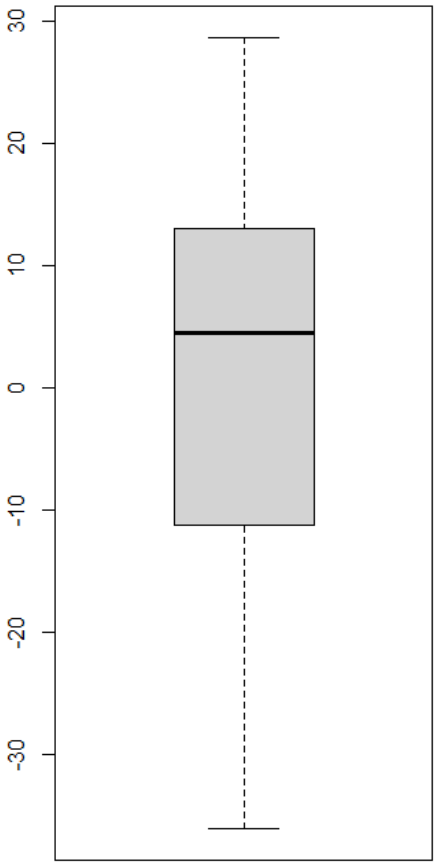
Part F:



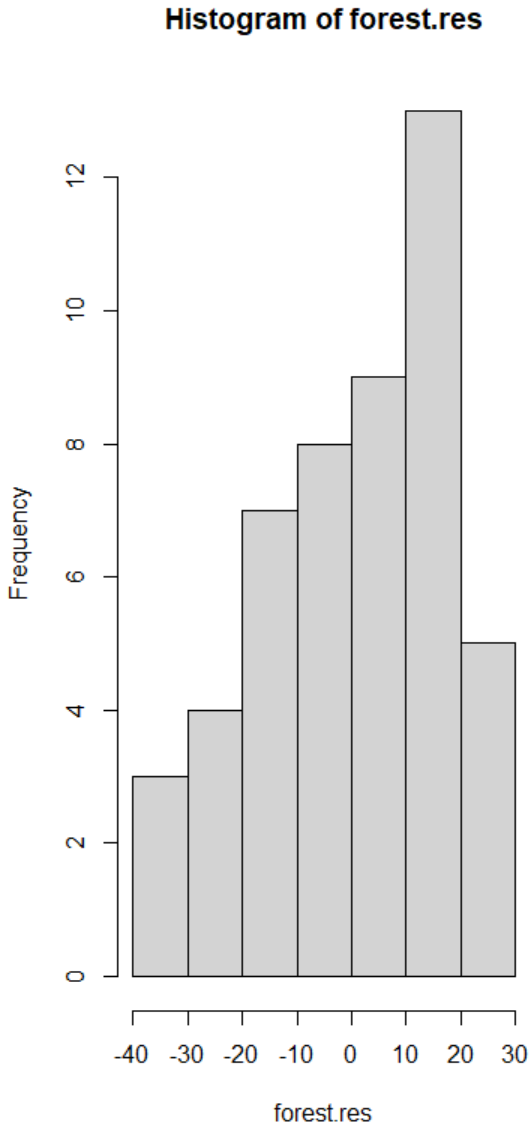

```
#Residual seems to show a slight convex curve.
```

Part G:

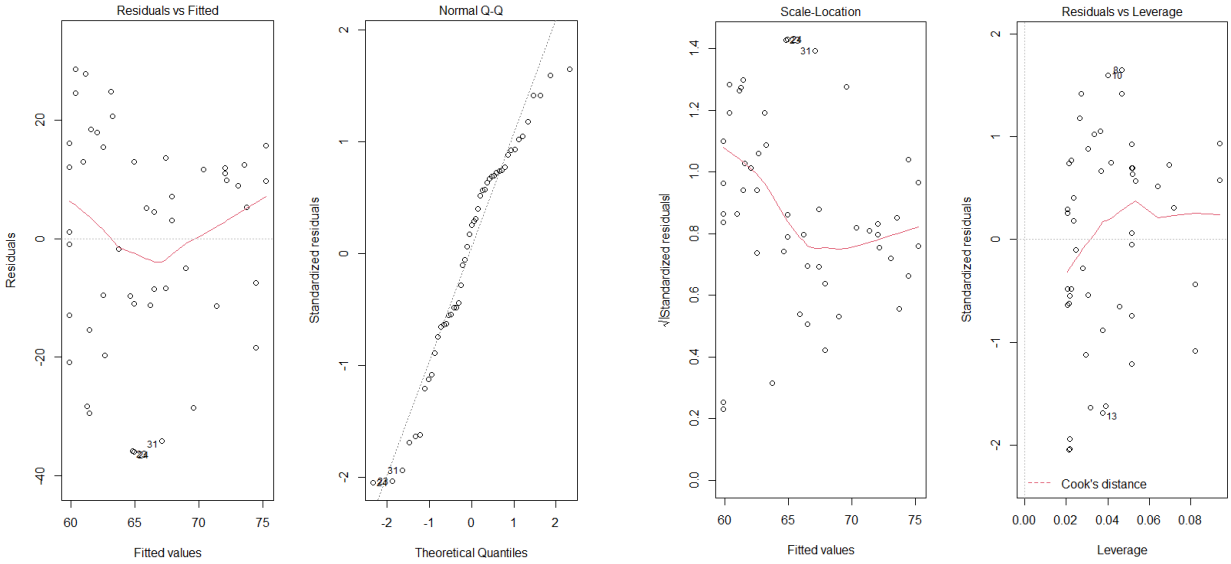




The qqnorm plot seems to suggest apx. normality as the qqline apx. fits the qqnorm plot. To confirm we take the boxplot and histogram of the residuals, and find the residuals are left skewed.


Histogram of forest.res
Frequency
0
2
4
6
8
10
12
-40
-30
-20
-10
0
10
20
30
forest.res

Part H:




```
#Checking Linearity (Residuals vs Fitted) && (Std Residual vs Area)
#Shown linearity in part F

#Checking Homoscedasticity (Residuals vs Fitted) **no pattern / equally spread around y = 0
#Since points of of Residuals vs Fitted seem to be randomly distributed we assume independence.

#checking Independence (Assumed by data set)
#Since there is a slight curve in the residuals vs fitted points graph, that is they show some pattern, we cannot assume independence.
#However since the association is weak we proceed with the assumption of independence and proceed with caution.

#Checking normality (Normal Q-Q)
#Shown normality in part G.
forest.stdres <- rstandard(forest.lm)
plot(waterquality$Forest, forest.stdres)

#Because independence is not guaranteed, we cautiously proceed with the assumptions made in part C,
#keeping in mind the results may not be particularly accurate.
```

10.34

```
# 10.34
# I'd prefer Area to IBI as the p-value of its occurance = way
# below alpha giving me high confidence that there is a definite
# relationship between Area & IBI. This does not hold true with
# the relationship between Forest% & IBI. Additionally, Forest % to IBI
# seems to lack independence, which prevents us from making conclusions
# on the data.
```

```
Call:
lm(formula = IBI ~ Area, data = waterquality)

Residuals:
    Min      1Q  Median      3Q     Max
-32.666  -8.887   3.432  12.414  25.193

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  52.9230     4.4835  11.804 1.17e-15 ***
Area          0.4602     0.1347   3.415  0.00132 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16.53 on 47 degrees of freedom
Multiple R-squared:  0.1988,    Adjusted R-squared:  0.1818
F-statistic: 11.67 on 1 and 47 DF,  p-value: 0.001322
```

```
Call:
lm(formula = IBI ~ Forest, data = waterquality)

Residuals:
    Min      1Q  Median      3Q     Max
-35.961 -11.186   4.508  13.021  28.633

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 59.90725    4.03957  14.830   <2e-16 ***
Forest       0.15313    0.07972   1.921   0.0608 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 17.79 on 47 degrees of freedom
Multiple R-squared:  0.07278,   Adjusted R-squared:  0.05305
F-statistic: 3.689 on 1 and 47 DF,  p-value: 0.06084
```

10.35

```
> # Original
> summary(forest.lm)

Call:
lm(formula = IBI ~ Forest, data = waterquality)

Residuals:
    Min      1Q  Median      3Q     Max
-35.961 -11.186   4.508  13.021  28.633

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 59.90725    4.03957  14.830   <2e-16 ***
Forest       0.15313    0.07972   1.921   0.0608 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 17.79 on 47 degrees of freedom
Multiple R-squared:  0.07278,   Adjusted R-squared:  0.05305
F-statistic: 3.689 on 1 and 47 DF,  p-value: 0.06084
```

```
> # Changing 0% to 0 IBI
> summary(adjust.lm1)

Call:
lm(formula = waterquality_temp$IBI ~ waterquality_temp$Forest)

Residuals:
    Min      1Q  Median      3Q     Max
-67.942  -8.614   6.875  16.658  28.771

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                 59.99030    5.02854  11.930 7.99e-16 ***
waterquality_temp$Forest  0.07952    0.09924   0.801    0.427
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 22.14 on 47 degrees of freedom
Multiple R-squared:  0.01348,   Adjusted R-squared:  -0.007514
F-statistic: 0.642 on 1 and 47 DF,  p-value: 0.427
```

```
> # Changing 100% to 0 IBI
> summary(adjust.lm2)

Call:
lm(formula = waterquality_temp$IBI ~ waterquality_temp$Forest)

Residuals:
    Min      1Q  Median      3Q     Max
-66.648 -10.134   6.384  16.511  26.459

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                 62.41420    4.62574  13.493   <2e-16 ***
waterquality_temp$Forest  0.04233    0.09129   0.464    0.645
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 20.37 on 47 degrees of freedom
Multiple R-squared:  0.004554,  Adjusted R-squared:  -0.01663
F-statistic: 0.215 on 1 and 47 DF,  p-value: 0.645
```

```
# Notice when we decrease the IBI to 0 of a point with 0% watershed area that
# was forest, we get a p-value = 0.03598, we also get a higher R-squared value
# That is by removing the outlier with 0% forest, we get a stronger linear
# association.
#
# In contrast, when we decrease the IBI to 0 of a point with 100% watershed area
# that was forest, we get a p-value = 0.645!!! with a much smaller R-Squared value
# That is we artificially created an outlier in our data, which greatly reduced
# the linear association between IBI & Forest %.

# Outliers have a drastic effect on fitting for a Linear Regression Model.
```

10.36

Part A:

Confidence interval:

```
(65.61416, 77.04417)
```

Part B:

Predicted interval:

```
(37.5783, 105.08)
```

Part C:

```
# The confidence interval tells us that if a large sample of Watersheds with
# Area (40km^2) were sampled, the mean IBI value of the data would fall
# between the interval (65.614, 77.044) with 95% confidence.

# In contrast, the predicted interval tells us according to our Linear Regression
# Model a randomly sampled Watershed with Area (40km^2) in the Ozark highlands
# will have an IBI within (37.578, 105.08) with 95% confidence.
```

Part D:

```
# We cannot extend these results to other streams in Arkansas or other states
# because our predictions are based on the observations made on Ozark streams,
# specifically. Streams in any other place could be drastically different to the
# environment & conditions we find in Ozark; which would introduce unknown
# confounding variables that would greatly disrupt our predictions.
```

10.37

```
       fit      lwr      upr
1 57.52451 23.55984 91.48918
```

```
       fit      lwr      upr
1 69.55457 33.20852 105.9006
```

```
# For the Area vs IBI model, we get a predicted value of IBI = 57.52 given Area = 10km^2
# For the Forest% vs IBI model, we get a predicted value of IBI = 69.5544 given forest% Area = 63
# Notice both predicted values have huge uncertainty:
# Area with around 68 units & Forest% with 72.7 units allowing both predicted values to
# differ drastically
```