Homework07

Ahilan Subbaian

4/30/2020

"I pledge my honor that I have abided by the Stevens Honor System"

10.32) a.

```
library(readxl)
data <-read.csv("Water-Quality.csv")
area<- data$Area
IBI<-data$IBI
forest<-data$Forest


chrt <- matrix(c(mean(IBI),sd(IBI),median(IBI),max(IBI),min(IBI),
                 mean(area),sd(area),median(area),max(area),min(area)),ncol=5,byrow=TRUE)
colnames(chrt) <- c("mean","standard deviation","median","max","min")
rownames(chrt) <- c("IBI","area")
tabl <- as.table(chrt)
chrt
```

```
##          mean standard deviation median max min
## IBI  65.93878           18.27955     71  91  29
## area 28.28571           17.71417     26  70   2
```

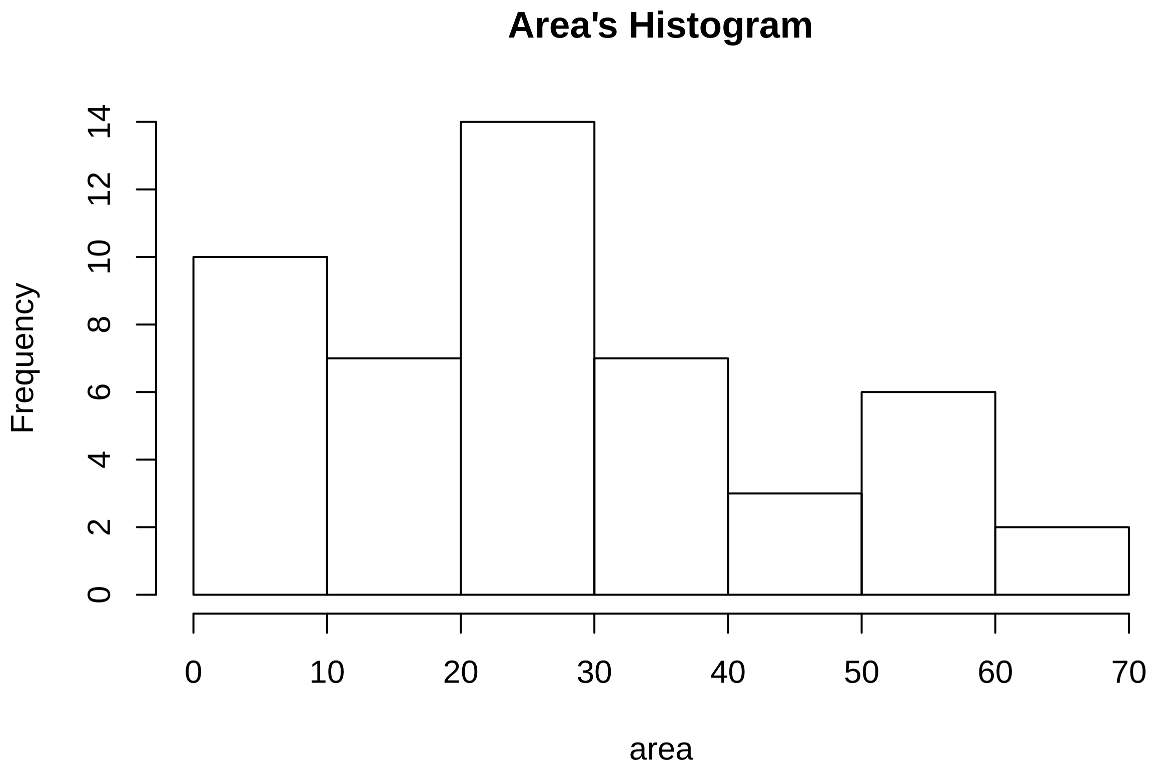
```
summary(area)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.00   16.00   26.00   28.29   34.00   70.00
```

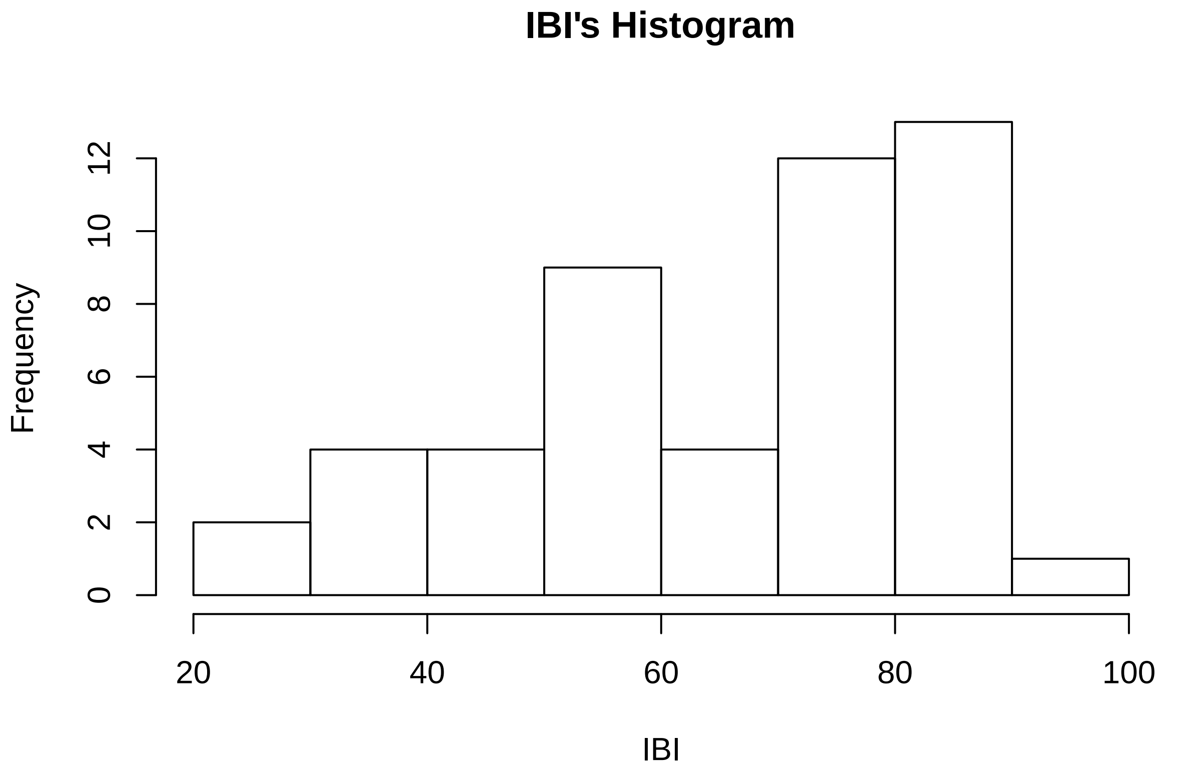
```
summary(IBI)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   29.00   55.00   71.00   65.94   82.00   91.00
```

```
hist(area, main = "Area's Histogram")
```

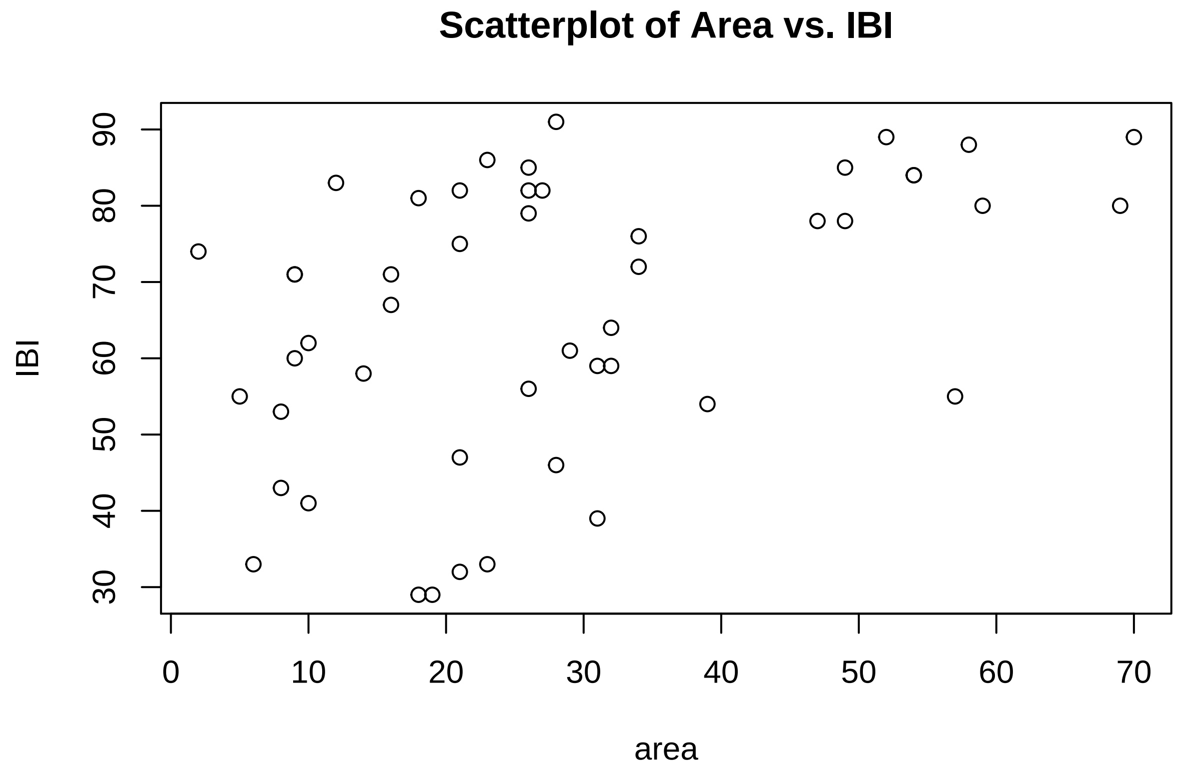


```r
hist(IBI, main = "IBI's Histogram")
```



Because the sample size is greater than 30 we can assume that both distributions are normal.Area is right skewed and IBI is left skewed. Area has multiple outliers while IBI has no outliers. b.

```
plot(area, IBI, main = "Scatterplot of Area vs. IBI")
```



There appears to be no outliers or unusual patters. There is a weak positive relationship between area and IBI and more variance in smaller values of area.

c. for i in 1:49 let yi = b0 + b1 * xi + ei

d. h0: b1 = 0 ha: b1 != 0

e.

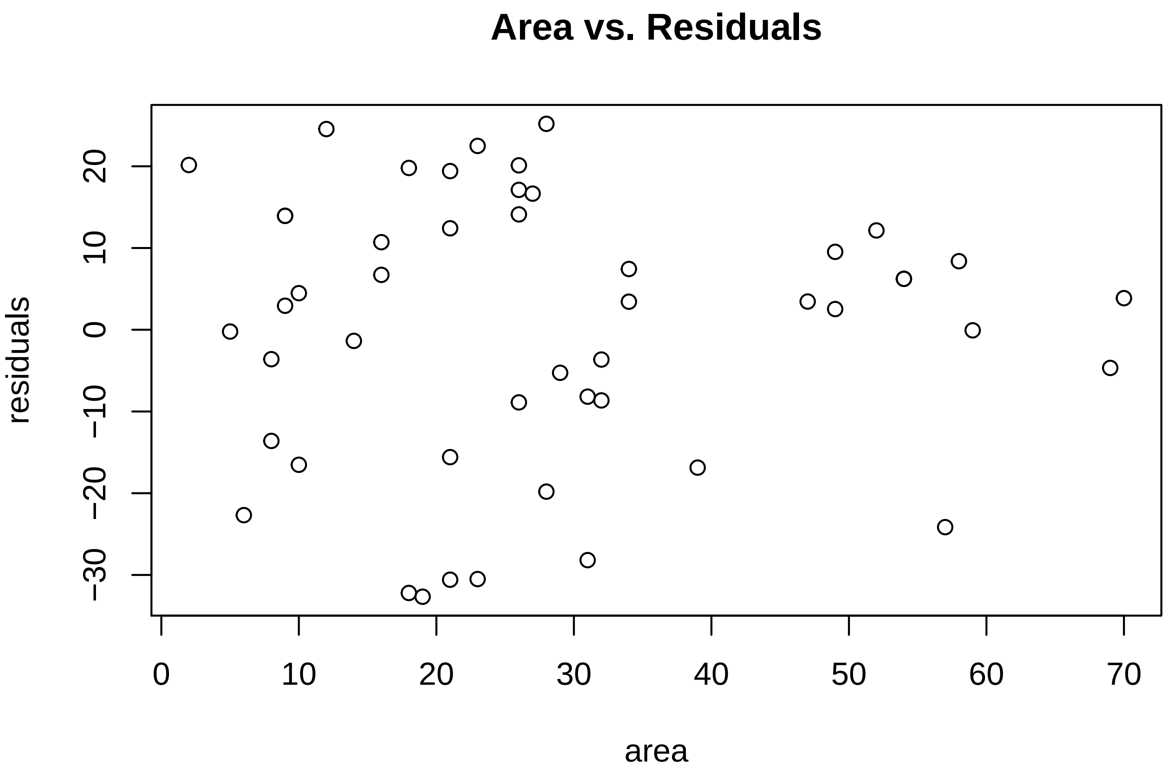
```
summary(lm(IBI~area))
```

```
##
## Call:
## lm(formula = IBI ~ area)
##
## 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
```

The coefficient of multiple determination is .1988, so 19.88% of the changes in IBI is caused by Area. the coefficient of area is .4602, so for every increase of 1 in area, IBI increases by .4602.

f.

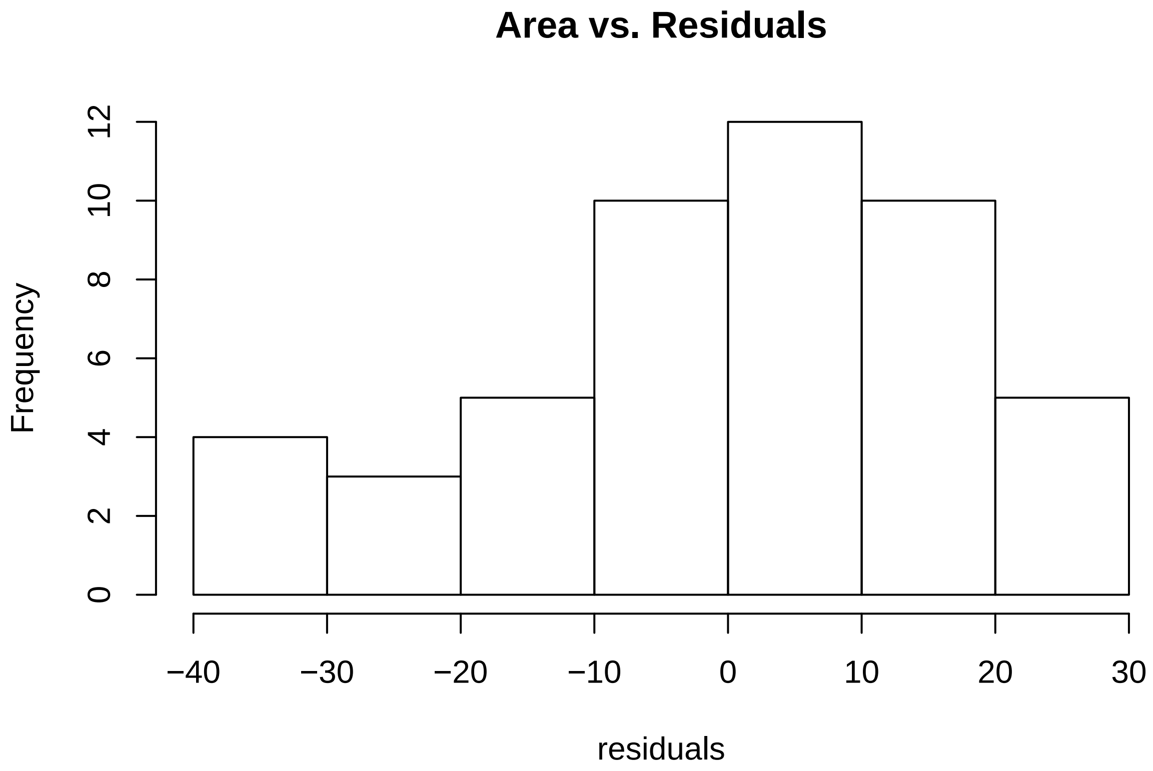
```
residuals<- resid(lm(IBI~area))
plot(area, residuals, main = "Area vs. Residuals")
```



Nothing unusual, there appears to be no pattern

g.

```
hist(residuals, main = "Area vs. Residuals")
```



The residuals appear approximately normal because the scatterplot did not seem to have any relationship and the histogram is approximately normal.

h. My assumptions from c seem reasonable as the residuals appear to be independent of area.

10.33) a.

```
chrt1 <- matrix(c(mean(IBI),sd(IBI),median(IBI),max(IBI),min(IBI),
                 mean(forest),sd(forest),median(forest),max(forest),min(forest)),ncol=5,byrow=TRUE)
colnames(chrt1) <- c("mean","standard deviation","median","max","min")
rownames(chrt1) <- c("IBI","Forest")
tabl1 <- as.table(chrt1)
chrt1
```

```
##             mean standard deviation median max min
## IBI    65.93878           18.27955     71  91  29
## Forest 39.38776           32.20431     33 100   0
```

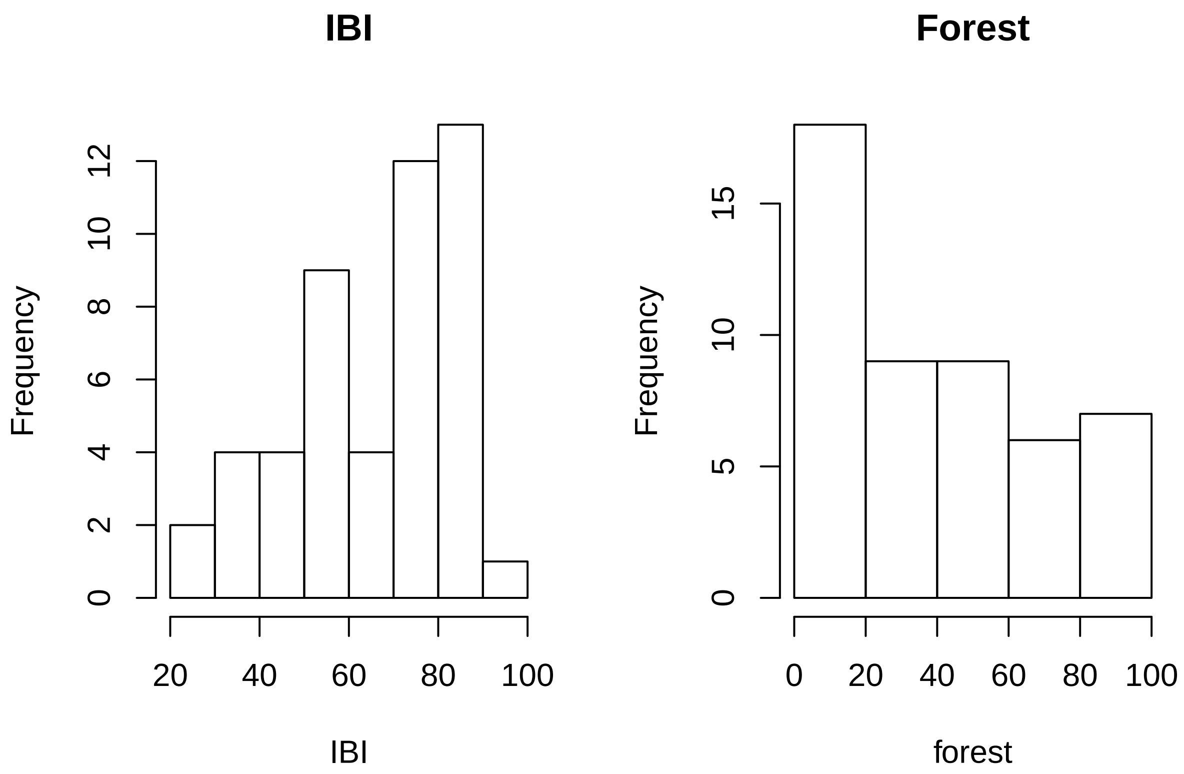
```
summary(IBI)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   29.00   55.00   71.00   65.94   82.00   91.00
```

```r
summary(forest)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    0.00   10.00   33.00   39.39   63.00  100.00
```

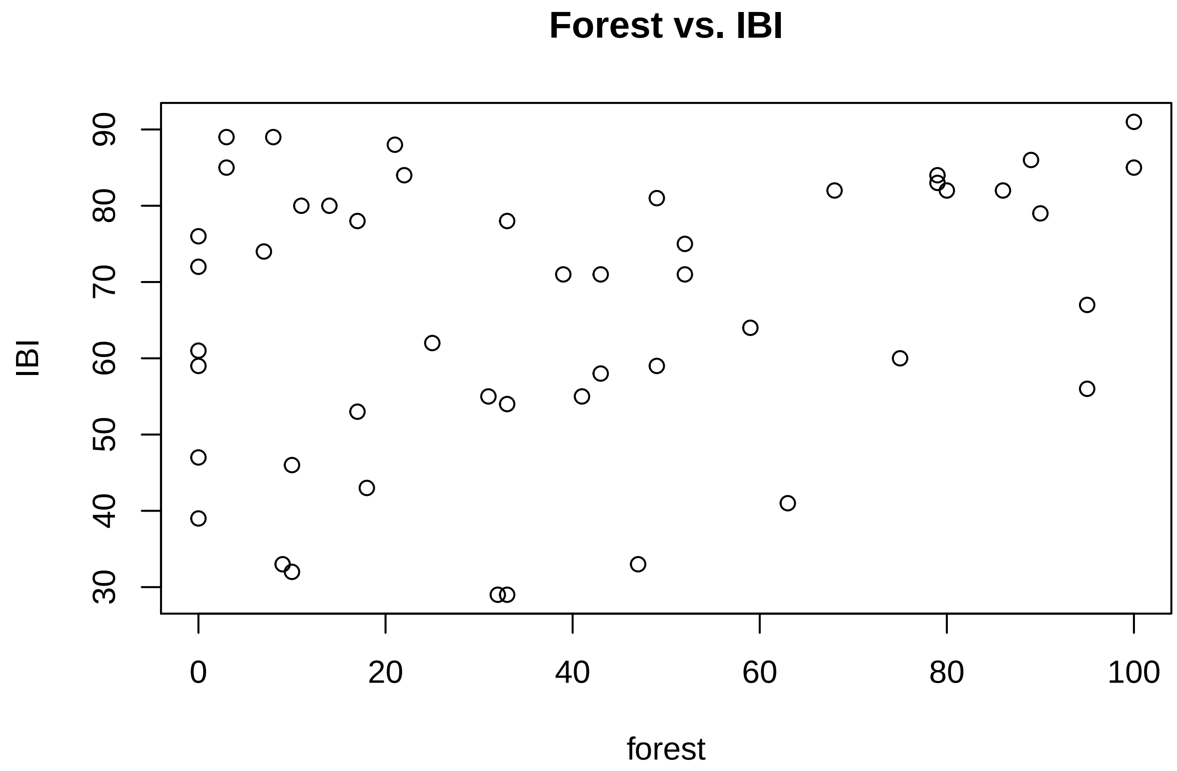
```r
par(mfrow=c(1,2))
hist(IBI, main = "IBI")
hist(forest, main = "Forest")
```



Because the sample size is larger than 30 we can assume that the distributions are normal. Forest is right skewed and IBI is left skewed. Niether seem to have an outlier.

b.

```r
plot(forest, IBI, main = "Forest vs. IBI")
```



There seems to be a very weak positive relationship between forest and IBI, and no weird patters.

c. for i in 1:49 let yi = a0 + a1 * xi + ei ei is independant and normally distributed at mean 0

d. h0 b1 = 0 ha b1 != 0

e.

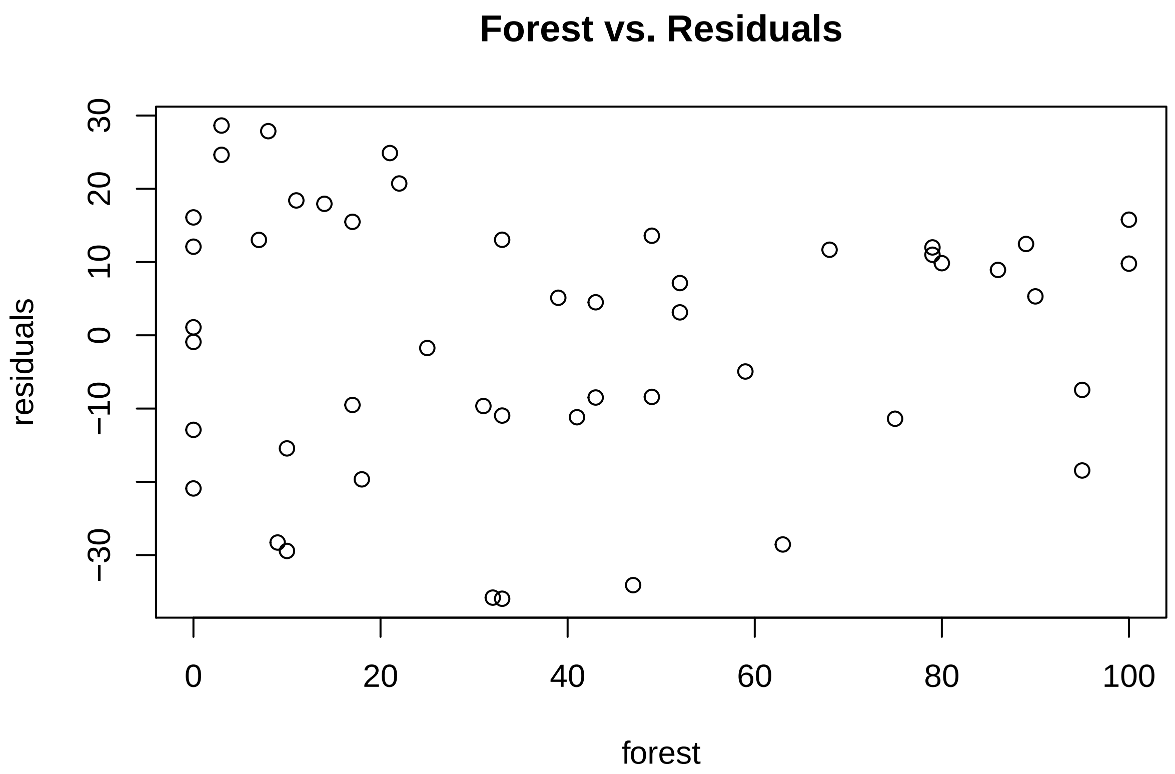
```r
summary(lm(IBI ~ forest))
```

```
##
## Call:
## lm(formula = IBI ~ forest)
##
## 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
```

the coefficient of multiple determination is .07278 so 7.27% of the change in IBI is explained by forest. and the coefficent of forest is .15313 so for every 1 change in forest there is a .15313 change in IBI

f.

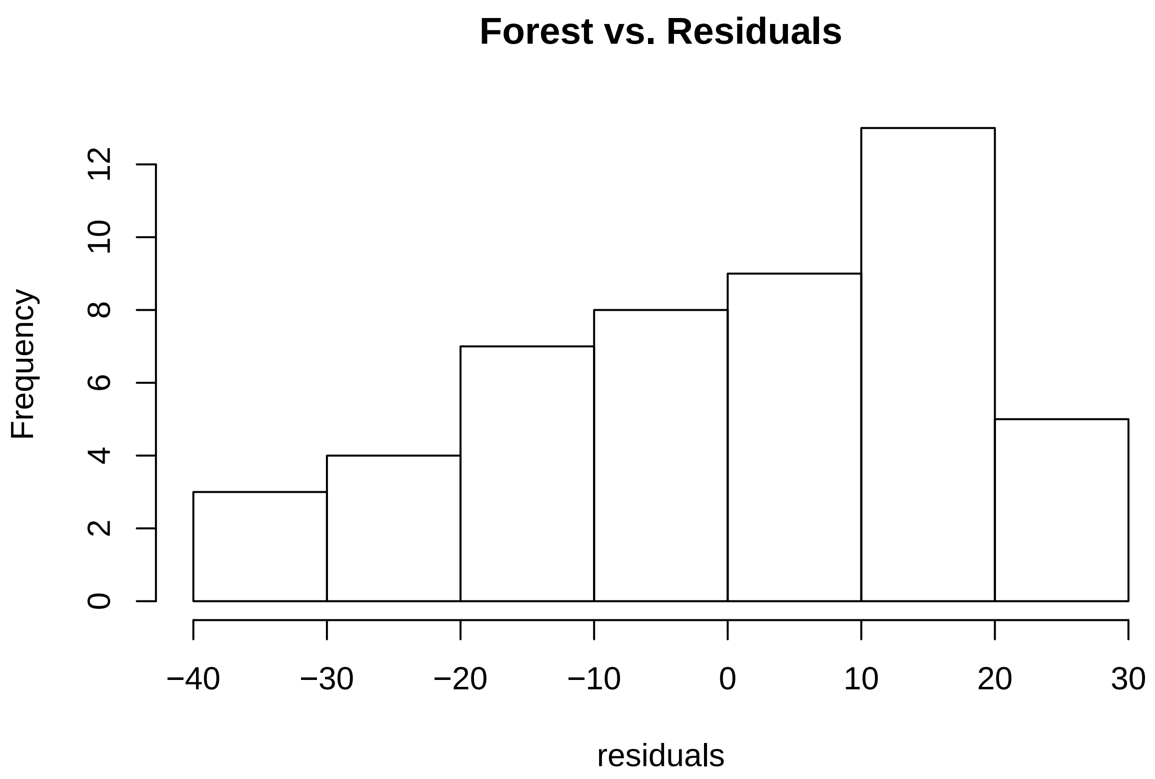
```
residuals<- resid(lm(IBI~forest))
plot(forest, residuals, main = "Forest vs. Residuals")
```



there does not appear to be a pattern in the scatterplot

g.

```r
hist(residuals, main = "Forest vs. Residuals")
```



the histogram appears to be left skewed

h. the assumption is not reasonable because the residuals appear to be left skewed showing a pattern in the data.

10.34) I pick area because it has a higher r squared so it explains more change and has a lower p value.

10.35) a. Because the relationship between the two variables is almost significant at the alpha level of 0.05, an observation with 0% forest would decrease the p value The regression equation is IBI = 51.3 + 0.483 Area

b. Because the relationship between the two variables is almost significant at the alpha level of 0.05, an observation with 100% forest would increase the p value The regression equation is IBI = 51.0 + 0.462 Area

10.36) a.

```r
71.4735 - (-2.011) * sqrt( 314.34 * ((1/49) + (40-25*4)^(2) / (-4272069)))
```

```
## [1] 76.4607
```

```
71.4735 + (-2.011) * sqrt( 314.34 * ((1/49) + (40-25*4)^2/ (-4272069)))
```

```
## [1] 66.4863
```

(66.4863, 76.4607)

b.

```
71.4735 + (-2.011) * sqrt( 314.34 * (1 + (1/49) + (40-25*4)^2/ (-4272069)))
```

```
## [1] 35.47209
```

```
71.4735 - (-2.011) * sqrt( 314.34 * (1+ (1/49) + (40-25*4)^2/ (-4272069)))
```

```
## [1] 107.4749
```

(35.47209, 107.4749)

c. the confidence interval tell us that we are 90% confident that the true population mean is between 66.48 and 76.46. and the prediction interval tell us that we are 95% confident that the true prediction response lies between 35.47 and 107.47

d. We cannot use this data for other streams in Arkansas or other states because we dont know if other streams will have the same conditions.

10.37) At an area of 10, the prediction index of biotic integrity was 57.52. At 63% forest produced a predicted value of 69.55. This has a very large range of change in values, because they both use different base data.