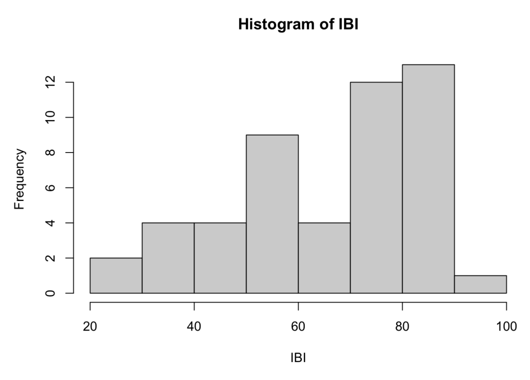
HW8

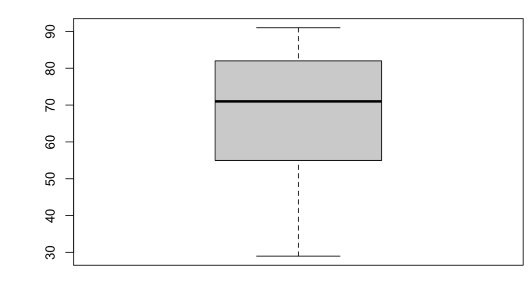
Aparajita Rana

12/4/2020

I pledge my honor that I have abided by the Stevens Honors System. **10.32 Predicting water quality. a.**

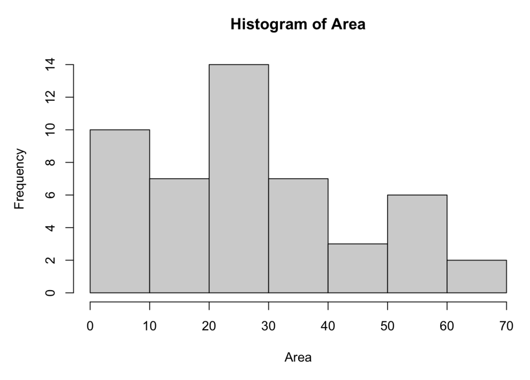
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   29.00   55.00   71.00   65.94   82.00   91.00 
```

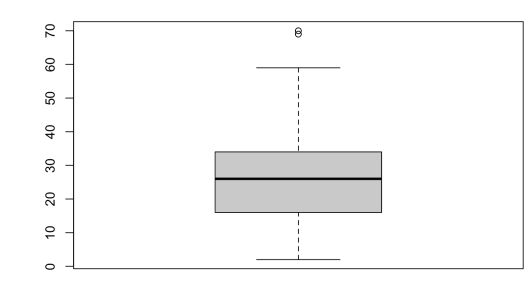




IBI data has a strong skew to the left. No outliers via the box plot.

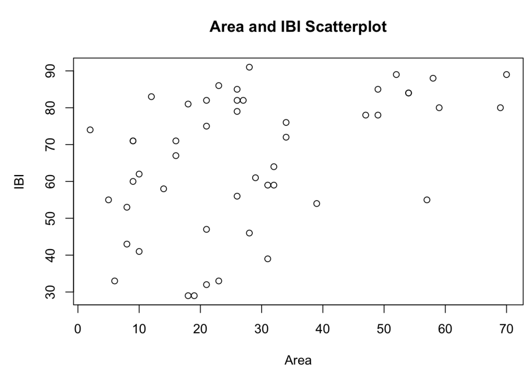
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    2.00   16.00   26.00   28.29   34.00   70.00 
```





Area data has a slight skew to the right. Upper outlier(s) via the box plot.

b.



IBI is more varied when the area is lower, this has no outliers visible

c.

$IBI \ = \ \beta_0 + \beta_1 \left(Area\right) \ + \ \xi_i, i = 1,2,\ldots,49$

$IBI = 52.92 + 0.46 * Area + \epsilon$

d.

Null Hypothesis: $\beta_1 = 0$

Alternative Hypothesis: $\beta_1 \neq \ 0$

e.

Significance Level= 0.05

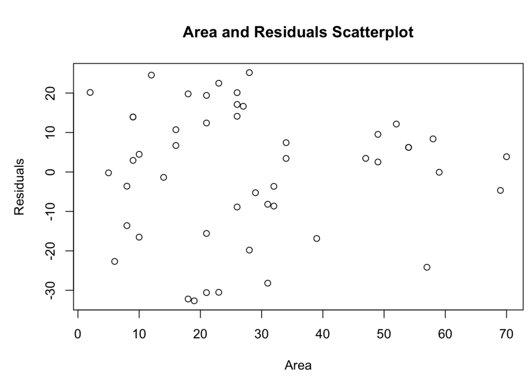
$$T = \frac{\hat{\beta_1}}{SE_{\hat{\beta_1}}} = 3.41$$

$$SE_{\hat{\beta}_1} = \sqrt{\frac{S^2}{\sum_{i=1}^{n}\left(x_i = \hat{x}\right)^2}} = 0.13$$

$$S^2 = \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{n-2} = 273.4$$

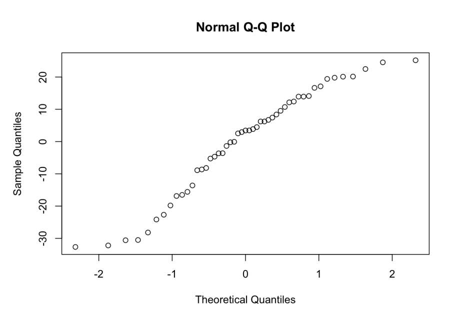
Degrees of Freedom = n-2 = 49 - 2 = 47 P-value = 2*P(T>3.41) = 0.0013 P-value is 0.0013 < α. Therefore, with this evidence, we reject H_0.

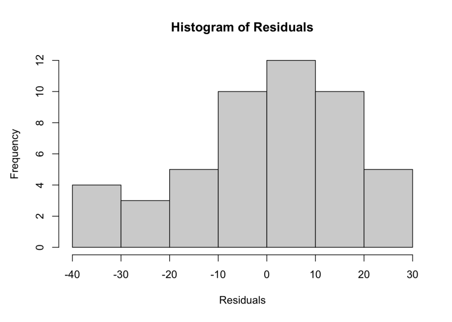
f.



As area increases, the residual approaches 0 and decreases.

g.





The residuals are skewed left and have a normal distribution.

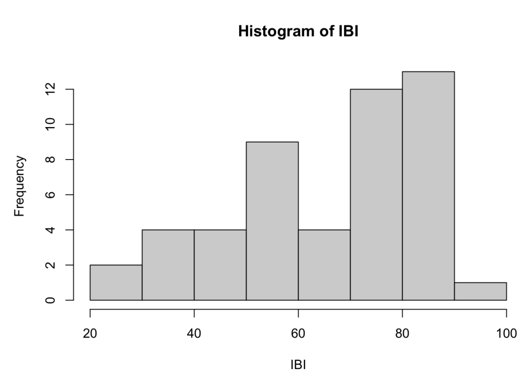
h.

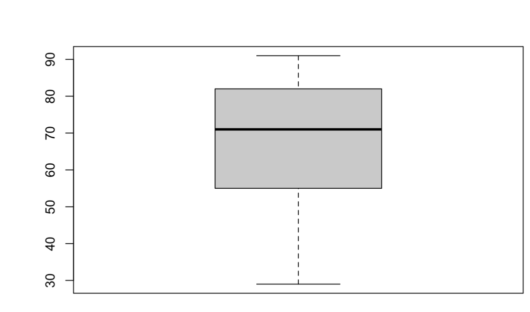
Looking at the graphs and residuals, the assumptions are reasonable.

10.33 More on predicting water quality.

a.

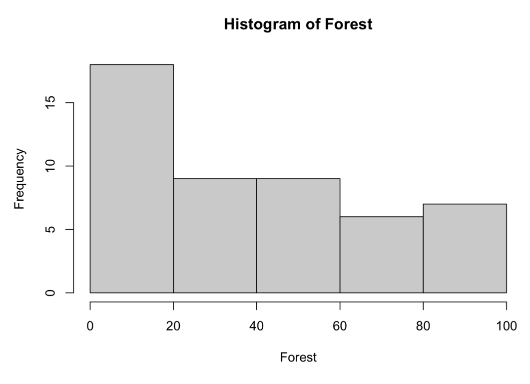
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   29.00   55.00   71.00   65.94   82.00   91.00 
```

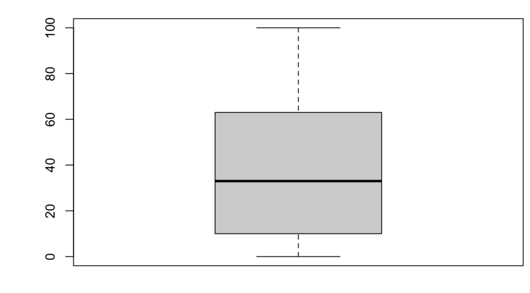




IBI data has a strong skew to the left. No outliers via the box plot.

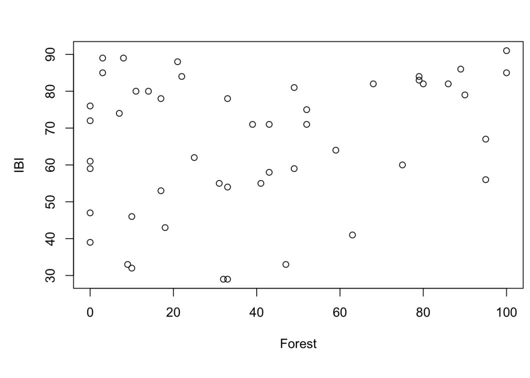
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    0.00   10.00   33.00   39.39   63.00  100.00 
```





Forest data has a strong skew to the right. No outliers via the box plot.

b.



IBI varies more when Forest is lowered. There is no strong association between Forest and IBI and there are no outliers.

c.

$IBI_i = \beta_0 + \beta_1 \cdot \ Forest_i + \epsilon_i$ for $i = 1,2,3,…,49

$IBI = 59.91 + 0.153 * Forest + \epsilon$

d.

Null Hypothesis: $\beta_1 = 0$

Alternative Hypothesis: $\beta \neq 0$

e.

$\alpha = 0.05$

$$T = \frac{\hat{\beta_1}}{SE_{\hat{\beta_1}}} = \frac{0.153}{0.08} = 1.9$$

$$SE_{\hat{\beta_1}} \ = \sqrt{\frac{S^2}{\sum_{i=1}^{n}\left(x_i - x^{-}\right)^2}} = \sqrt{\frac{316.4}{49781.6}} = 0.08$$

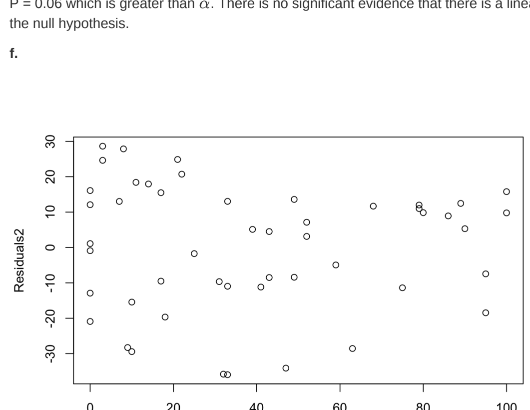
$$S^2 = \frac{\sum_{i=1}^{n} \ \left(y_i - \hat{y}_i\right)^2}{n-2} = 316.4$$

Degrees of Freedom: n-2 = 47

P-value: P = 2*P(T>1.91) = 0.06

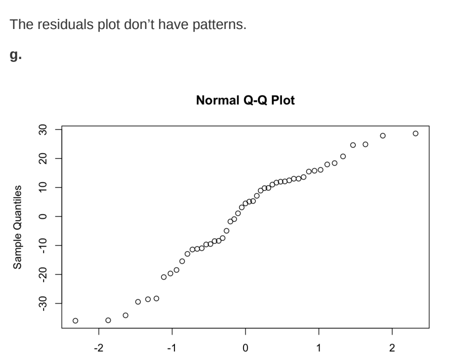
P = 0.06 which is greater than α. There is no significant evidence that there is a linear relationship between IBI and Forest. Hence, we fail to reject the null hypothesis.

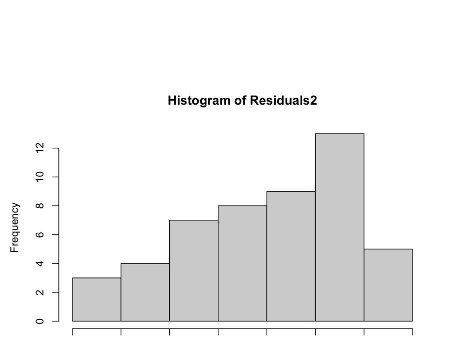
f.



The residuals plot don't have patterns.

g.





The residuals are strongly skewed to the left and they are not normal.

h.

The assumptions are not reasonable based on the graphs. The residuals are not normally distributed.

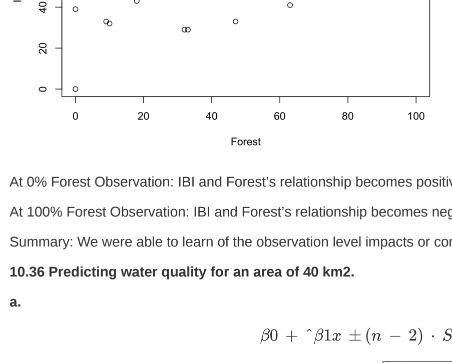
10.34 Comparing the analyses.

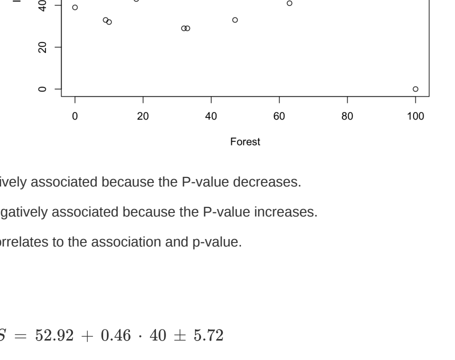
The first analysis of Area and IBI shows significant evidence of a linear relationship. The residuals are normally distributed.

The second analysis of Forest and IBI shows no significant evidence of a linear relationship. The residuals are strongly skewed to the left and they are not normal.

Given the choice, I believe that the first analysis is a better choice because regression seems to have worked better. The second analysis did not have normal distribution and lacked a linear relationship so is not a good choice.

10.35 How an outlier can affect statistical significance.





At 0% Forest Observation: IBI and Forest's relationship becomes positively associated because the P-value decreases.

At 100% Forest Observation: IBI and Forest's relationship becomes negatively associated because the P-value increases.

Summary: We were able to learn of the observation level impacts or correlates to the association and p-value.

10.36 Predicting water quality for an area of 40 km2.

a.

$$\beta 0 \ + \ \hat{}\beta 1x \ \pm \left(n \ - \ 2\right) \ \cdot \ S \ = \ 52.92 \ + \ 0.46 \ \cdot \ 40 \ \pm \ 5.72$$

$$SE_{\mu\hat{Y}} = \sqrt{\left(\frac{1}{n} + \frac{\left(x^{\cdot} - \hat{x}\right)^2}{\sum_{i=1}^{n}\left(x_{i - \hat{x}}\right)^2}\right) S^2} = 2.8$$

$$S^2 = \sum_{i=1}^{n}\frac{\left(y_i - \hat{y}_1\right)^2}{n-2} = 273.4$$

95% Confidence Interval: $71.33 \ \pm \ 5.72 \ = \ \left(65.61, \ 77.04\right)$

b.

$$\beta 0 \ + \ \hat{}\beta 1x \ \pm \ \left(n \ - \ 2\right) \ \cdot \ SE_Y \ = \ 52.92 \ + \ 0.46 \ \cdot \ 40 \ \pm \ 33.75 * \ t1 - \alpha 2EY\hat{}*$$

$$SE_{\hat{Y}} = \sqrt{\left(1 + \frac{1}{n} + \frac{\left(x^{\cdot} - \hat{x}\right)^2}{\sum_{i=1}^{n}\left(x_{i - \hat{x}}\right)^2}\right) S^2} = 16.8$$

$$S^2 = \sum_{i=1}^{n}\frac{\left(y_i - \hat{y}_1\right)^2}{n-2} = 273.4$$

95% Prediction Interval: $71.33 \ \pm \ 33.75 \ = \ \left(37.58, \ 105.08\right)$

c.

Using the confidence interval, we can say that we can be 95% certain that the mean of the area of $40km^2$ is between 65.61 and 77.04.

Using the prediction interval, we can say that we can be 95% certain that the mean of the next new observation is between 37.58 and 105.08.

d.

I believe this can be applied to other streams in Arkansas because the area's setting overall is probably very similar. However, other states and locations are less likely to be similar because of how their settings might be different.

10.37 Compare the predictions.

Area: IBI = 52.92 + (0.46 * Area) = 57.5 Forest: IBI = 59.91 + (0.153 * Forest) = 69.5

The prediction interval is broad which has resulted in the forest estimate being greater than the area estimate. This can cause overall uncertainty with the results.