Homework 8

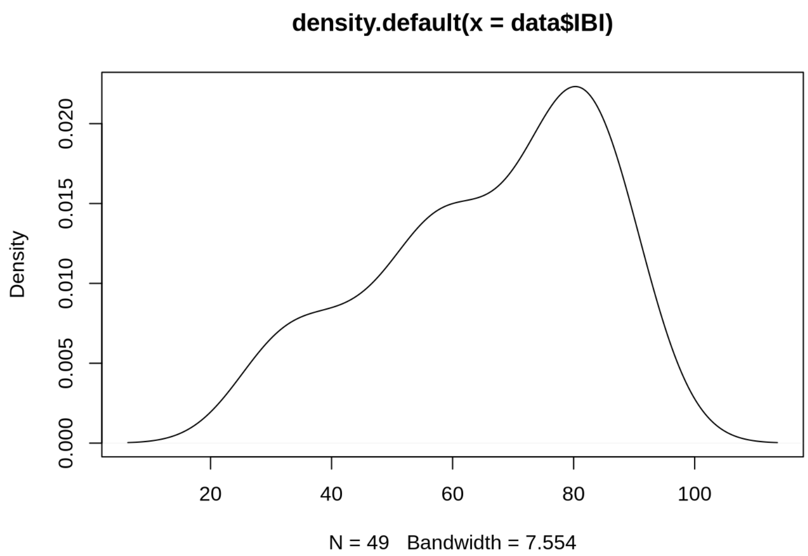
Kevin Ha

November 27, 2021

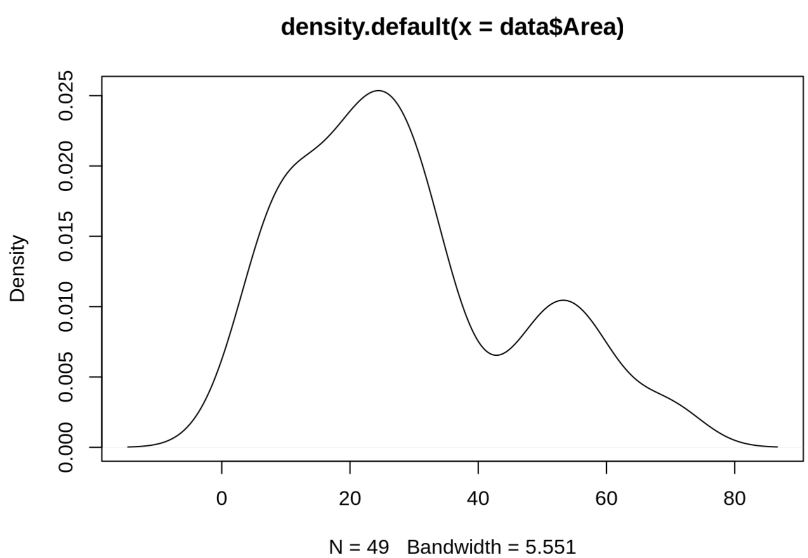
Problem 10.32

Problem 10.32a

```
plot(density(data$IBI))
```




```
plot(density(data$Area))
```



Graphical Summary: IBI is left skewed, Area is right skewed.

Numerical summary:

$\bar{x} = 28.2857143$

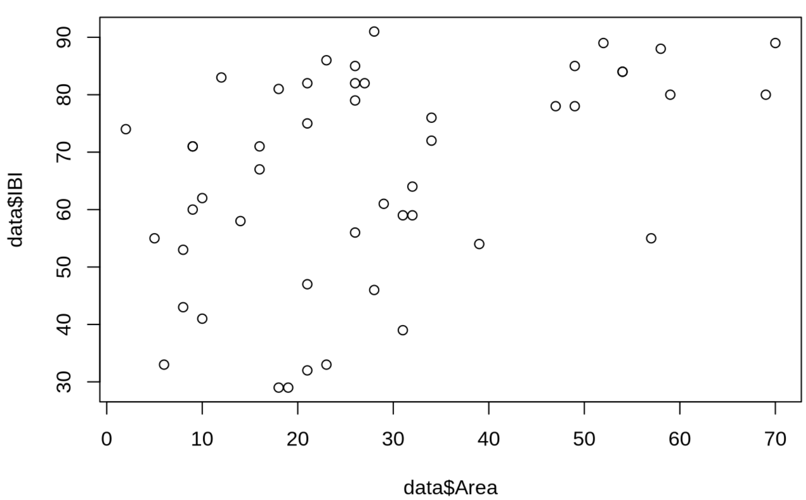
$s_x = 17.7141657$

$\bar{y} = 65.9387755$

$s_y = 18.2795516$

Problem 10.32b

```
plot(data$IBI ~ data$Area, data = data)
```



There is a weak and positive association between IBI and Area. No outliers or unusual patterns.

Problem 10.32c

$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ where $i = 1, 2, \ldots 49$

Problem 10.32d

$H_0 : B_1 = 0$

$H_a : B_1 \neq 0$

Problem 10.32e

```
fml = lm(IBI ~ Area, data = data)
summary(fml)
```

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

$\hat{IBI} = 52.92 + 0.46 \cdot Area$

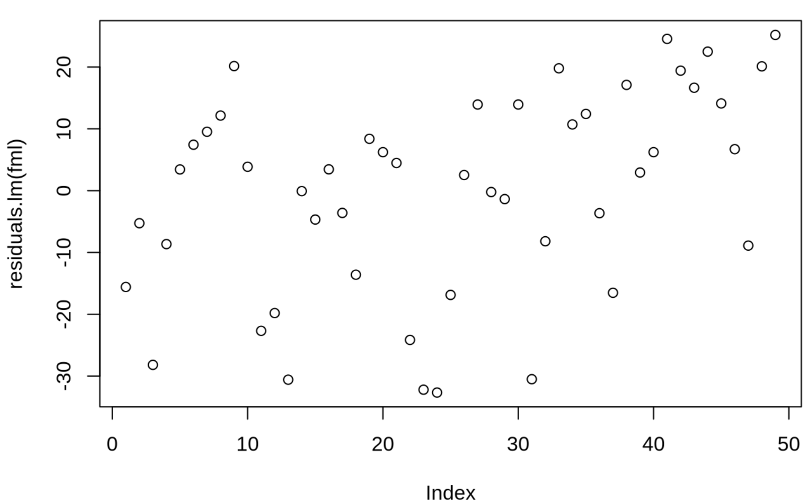
$s = 16.53$

$f = 11.67$

$p = 0.0013$

Problem 10.32f

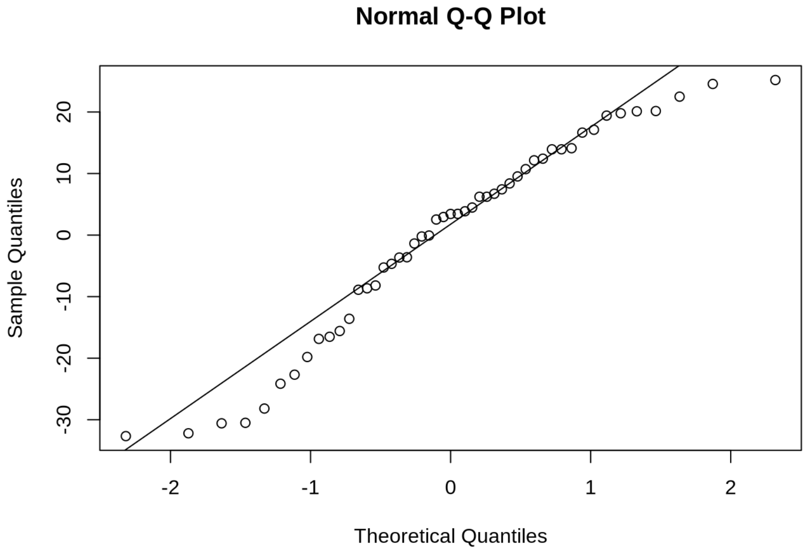
```
plot(residuals.lm(fml))
```



Nothing noticeably unusual.

Problem 10.32g

```
qqnorm(residuals(fml))
qqline(residuals(fml))
```



The residuals are not approximately normal since there appears to be a left skew among the data points on the scatterplot.

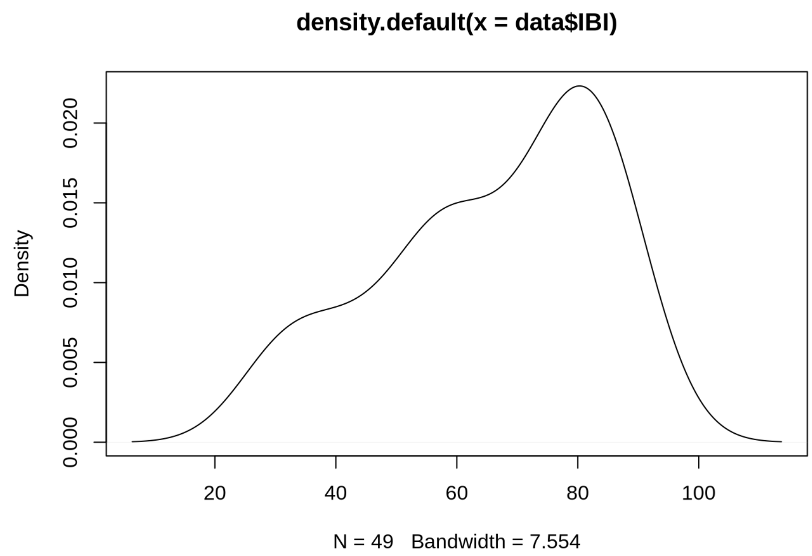
Problem 10.32h

The assumptions in part C do not appear to be reasonable since we assumed that the residuals (ϵ) is normally distributed, but the QQ plot above does not reflect this assumption.

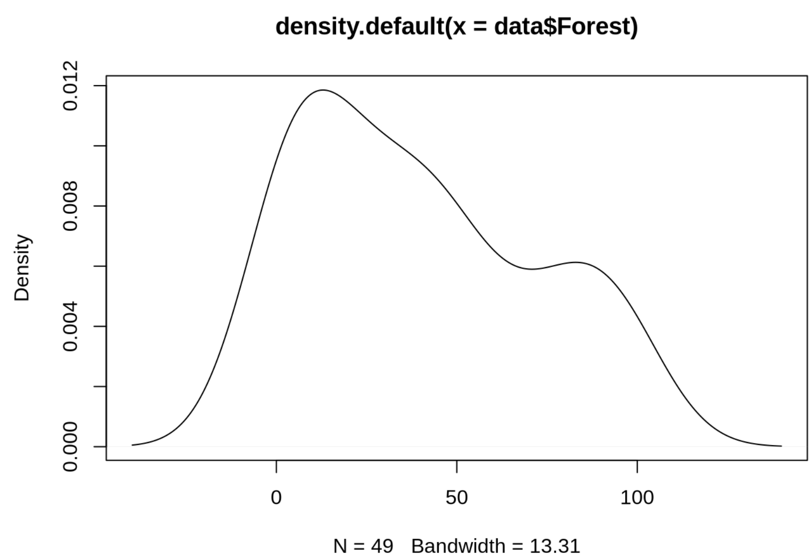
Problem 10.33

Problem 10.33a

```
plot(density(data$IBI))
```




```
plot(density(data$Forest))
```



Graphical Summary: IBI is left skewed, Area is right skewed.

Numerical summary:

$\bar{x} = 39.3877551$

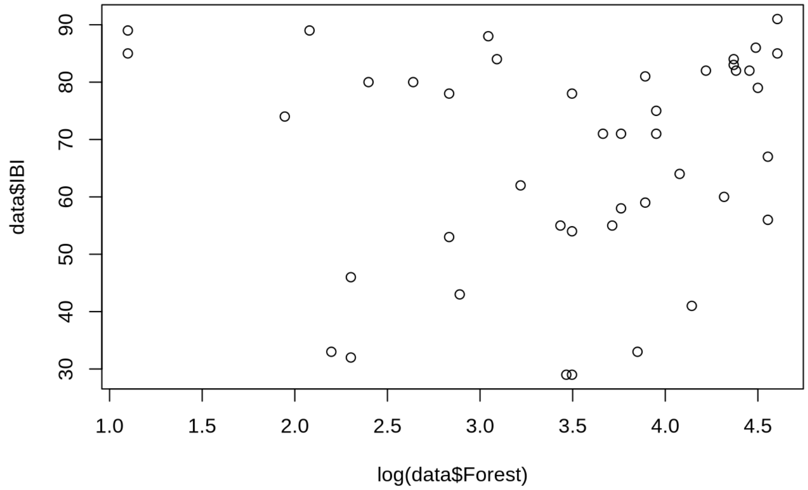
$s_x = 32.2043063$

$\bar{y} = 65.9387755$

$s_y = 18.2795516$

Problem 10.33b

```
plot(data$IBI ~ log(data$Forest), data = data)
```



There is a weak and positive association between IBI and Area. No outliers or unusual patterns.

Problem 10.33c

$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ where $i = 1, 2, \ldots 49$

Problem 10.33d

$H_0 : B_1 = 0$

$H_a : B_1 \neq 0$

Problem 10.33e

```
fml2 = lm(data$IBI ~ data$Forest)
summary(fml2)
```

```
##
## Call:
## lm(formula = data$IBI ~ data$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 ***
## data$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
```

$\hat{IBI} = 59.9 + 0.153 \cdot Forest$

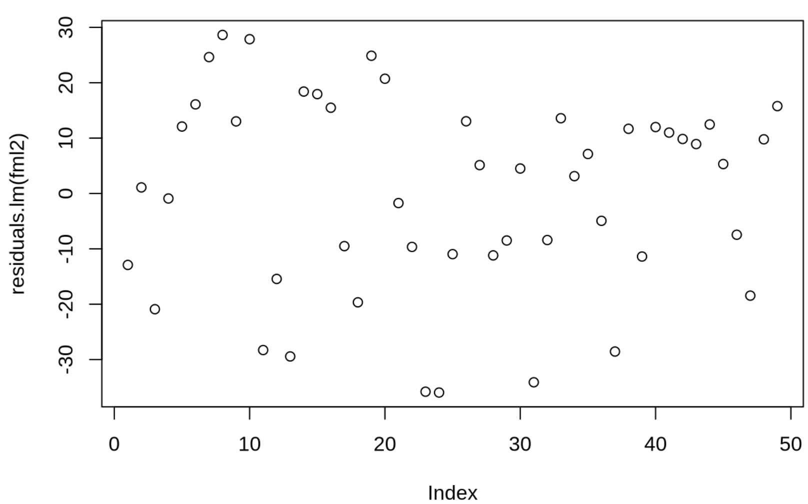
$s = 17.79$

$f = 3.689$

$p = 0.06$

Problem 10.33f

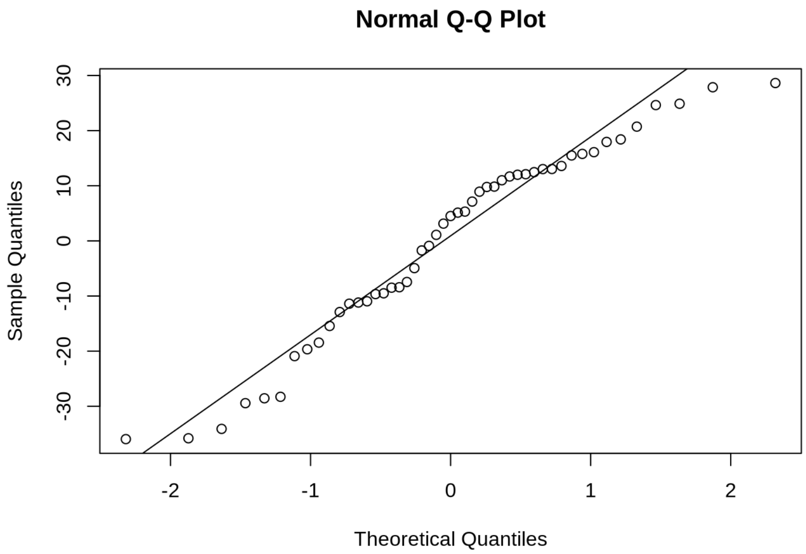
```
plot(residuals.lm(fml2))
```



Nothing noticeably unusual.

Problem 10.33g

```
qqnorm(residuals(fml2))
qqline(residuals(fml2))
```



The residuals are not approximately normal since there appears to be a left skew among the data points on the scatterplot.

Problem 10.33h

The assumptions in part C do not appear to be reasonable since we assumed that the residuals (ϵ) is normally distributed, but the QQ plot above does not reflect this assumption.

Problem 10.34

The area would be the preferred explanatory variable as it has a higher R-squared value compared to the Forest's R-squared value (0.1988 > 0.07) additionally, the Area has a lower p-value (0.001 < 0.06) which supports the claim as it indicates there is a statistically significant relationship between the two

Problem 10.35

The p-values are calculated below:

```
fml3 = lm(modified0$IBI ~ modified0$Forest)
fsum <- summary(fml3)$fstatistic
pf(fsum[1], fsum[2], fsum[3], lower.tail=FALSE)
```

```
##      value
## 0.03597982
```

```
fml4 = lm(modified100$IBI ~ modified100$Forest)
fsum2 <- summary(fml4)$fstatistic
pf(fsum2[1], fsum2[2], fsum2[3], lower.tail=FALSE)
```

```
##     value
## 0.6449941
```

When you decrease the IBI to 0 for an observation with 0% forest the p-value decreases thus a more significant relationship, the p-value decreases when the IBI is set to 0 for an observation with 100% forest resulting and a weaker association between the two.

Problem 10.36

Problem 10.36a

The 95% confidence interval is calculated using the following R function below:

```
new <- data.frame(Area = 40)
predict(fml, newdata = new, interval = 'confidence', level = 0.95)
```

```
##        fit      lwr      upr
## 1 71.32916 65.61416 77.04417
```

Confidence Interval: (65.61, 77.04)

Problem 10.36b

```
new <- data.frame(Area = 40)
predict(fml, newdata = new, interval = 'prediction', level = 0.95)
```

```
##        fit      lwr    upr
## 1 71.32916 37.57836 105.08
```

Prediction Interval: (37.58, 105.08)

Problem 10.36c

The **confidence interval for the mean response** gives an interval estimate on where the true population mean value of the response (dependent) variable IBI falls within when associated with an area of 40 km^2

The **prediction interval for the future response** gives an interval estimate on what the predicted IBI value should be when associated with an area of 40 km^2 is expected to fall in between for a future observation.

Problem 10.36d

These results cannot be applied exactly to other streams in Arkansas or other states since there a lot more different environments and other factors that influence IBI of other streams. However, the same process can be applied to determine the characteristics above given the proper data and influential factors.

Problem 10.37

$Area = 10$

Predicted IBI for Area of km^2

$\hat{y} = 52.92 + 0.46 * 10 = 57.52$

$Forest = 63$

Predicted IBI for Forest = 63

$\hat{y} = 59.9 + 0.153 * 63 = 69.55$

The reasoning for the differences of these estimates is due to the fact that prediction intervals have a lot of uncertainty due to their wide prediction intervals. In the previous problem above the standard error was about approximately 70 units which shows great uncertainty.