Julia Nelson

Ma-331-A

"I pledge my honor that I have
abided by the Steven's Honor System."

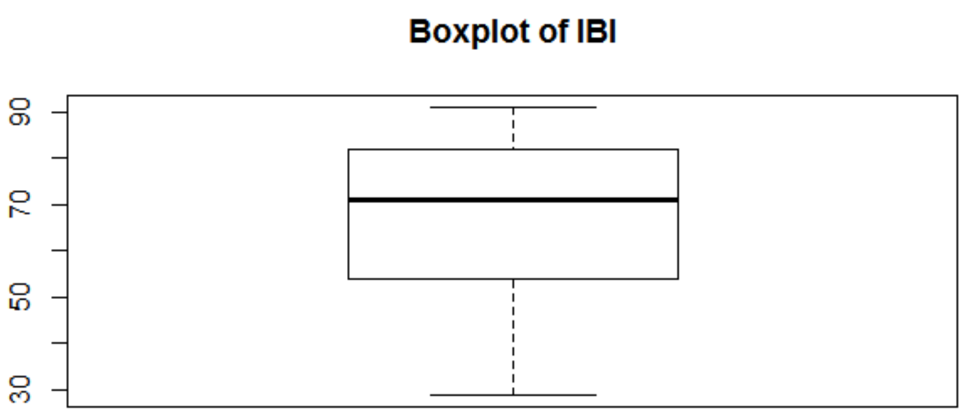
Homework 07

Exercise 10.32

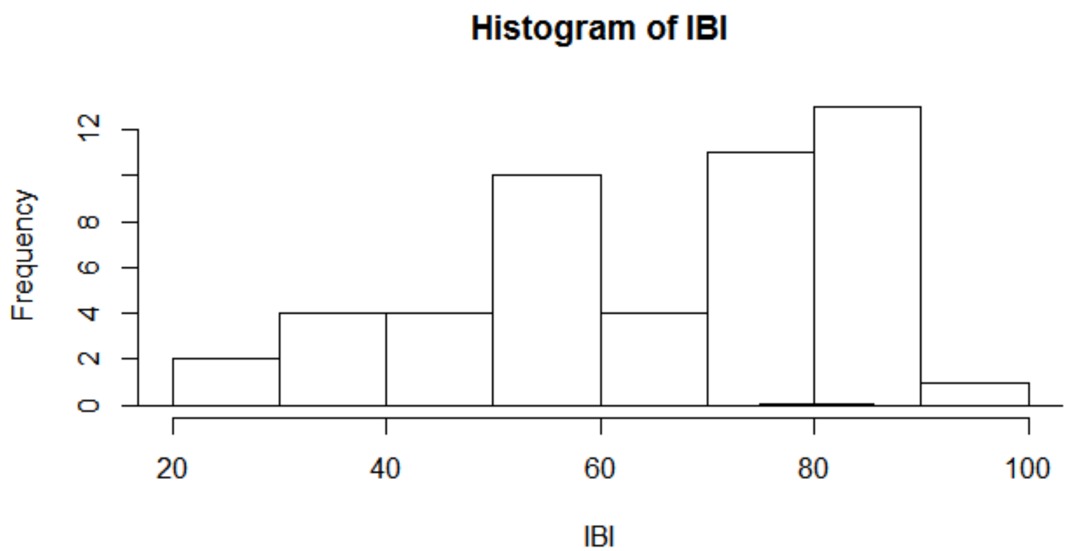
a) In this part of the task, the variable IBI is described using both numeric and graphical tools. After entering the data into R studio and performing some commands on it, the summary below was obtained of the variable IBI.

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  29.00   54.00   71.00   65.53   82.00   91.00
```

The above output provides the numerical description of the variable IBI. It can be seen that the mean of the variable is 65.53 and its median is 71. Given that the minimum and the maximum value for the variable is 29 and 91 respectively, the data can be presumed to be skewed. In order to be able to visualize the distribution of the data, the boxplot below was generated.

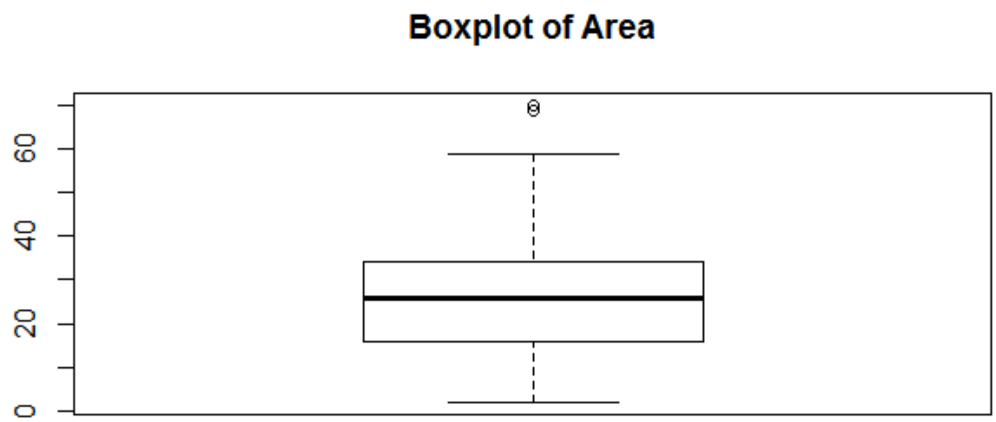


From the boxplot above, it is evident that there are no outliers for the variable IBI. It can also be seen clearly that the median value is closer to the upper quartile as compare to the lower quartile. This implies that there are more large values than they are small values for this variable. To get this point even clearer, a histogram of the variable was generated for the data and is as shown below. As seen in the histogram, there are more values to the right hence indicating negative skewness.

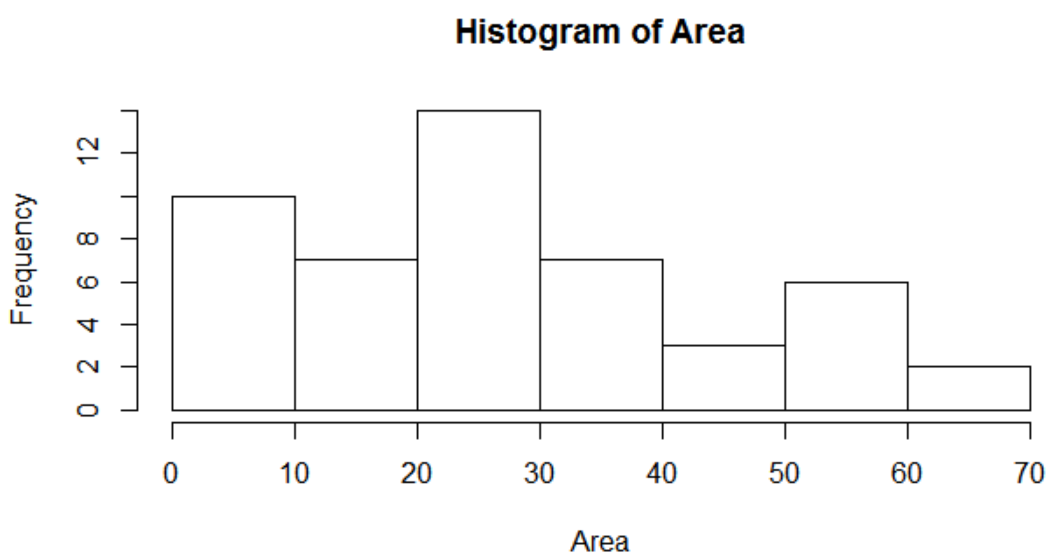


Another variable that was of interest to this study was the variable area, which represents the area of watershed in square kilometers for the streams. The numerical description of the data is as shown below. It can be seen that the data is nearly normal since the median and the mean are close to each other, but it indicates that more values are below the mean of the variable. Moreover, a wide gap is noticed between the upper quartile and the maximum value which is an indication of presence of outliers, which can be investigated using a boxplot.

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

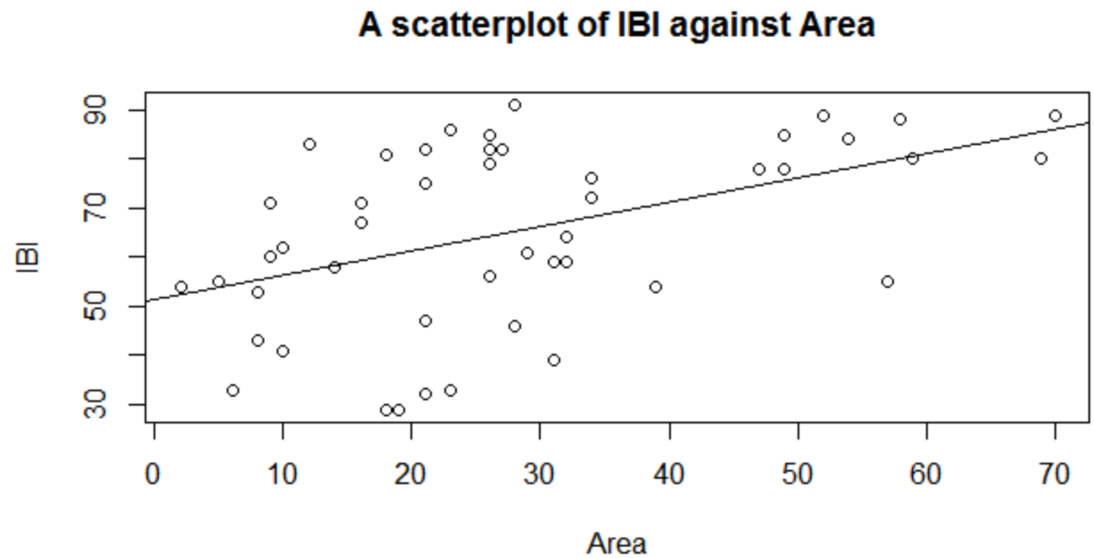


The boxplot above shows a nearly normally distributed data that consists of some outliers. Nonetheless, the normality of the data can be investigated using a histogram.



The above histogram shows that more values are concentrated to the left of the graph, this is an indication of positive skewness in the data.

b) The relationship between IBI and Area can be investigated using a scatter plot.



The scatter plot output above shows that the variable IBI and Area have a positive linear relationship. This means that as the value of Area increases, the value of IBI also increases.

c) The simple linear model for the relationship between IBI and Area can be statistically written as below:

$$IBI = \beta_0 + \beta_1 * Area$$

d) The null and alternative hypothesis for the model in part (c) can be stated statistically as below

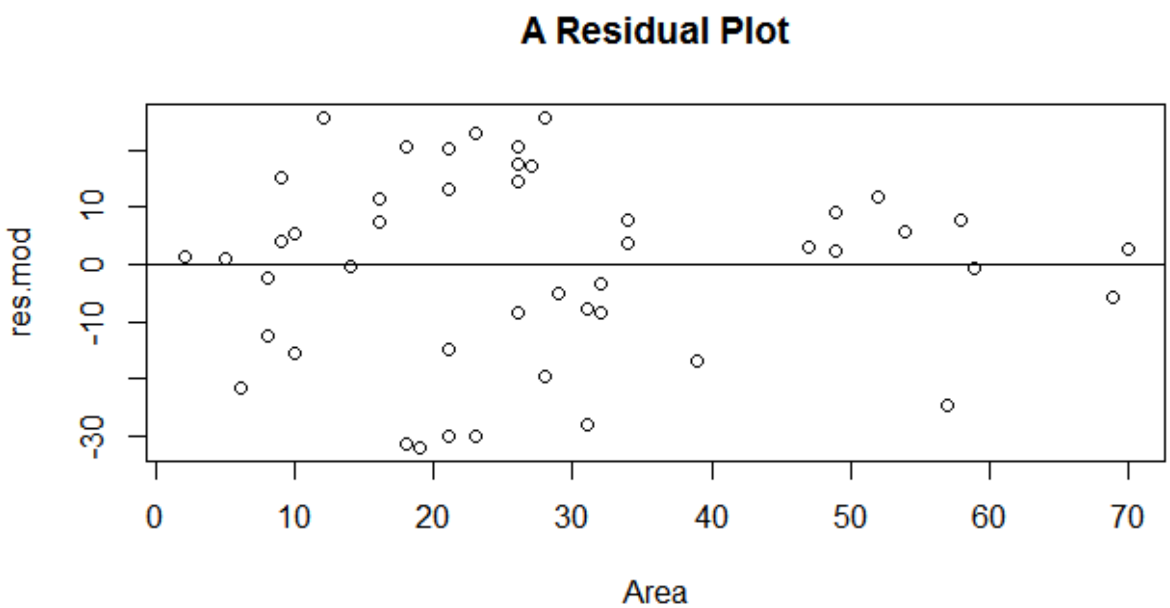
$$H_0: \beta_1 = 0 \; vs \;\; H_1: \beta_1 \neq 0$$

e) Regression results of the model in part (c) above as shown below

```
Call:
lm(formula = IBI ~ Area)

Residuals:
    Min      1Q  Median      3Q     Max 
-31.934  -8.399   2.818  11.729  25.611 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  51.5275     4.4074  11.691 1.63e-15 ***
Area          0.4951     0.1324   3.738 0.000502 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16.25 on 47 degrees of freedom
Multiple R-squared:  0.2292,  Adjusted R-squared:  0.2128 
F-statistic: 13.97 on 1 and 47 DF,  p-value: 0.0005024
```

The results above were obtain after running the model in c) in R. According to the results, the model is significant at 0.05 level of significance; F (1,47) = 13.97, p=0.000. Moreover, the results show that the predictor variable Area is a significant predictor of IBI at 0.05 level of significance; p=0.000. Therefore, IBI can be predicted using the Area of watershed in square kilometers for the streams. When the Area of the watershed increases by one square kilometer, the IBI increases by 0.495. Moreover, the model has a relatively weak coefficient of determination of 0.2292.

f) The plot of residuals against area is as shown below



The distribution of points in the residual plot above is generally random and distributed all over. This indicates the presence of unbiasness, homoscedasticity and linearly.

g) It can be presumed that the data is normally distributed bearing in mind that it the data point are randomly and evenly distributed on both sides of the 0 axis.
h) The Gauss Markov assumptions of a linear regression model are reasonable assumptions that have to be checked when performing a linear regression analysis. For instance, one of the assumptions is that the independent variable should be linearly related to the dependent variable. This is an important assumptions since the model is a linear model and the two

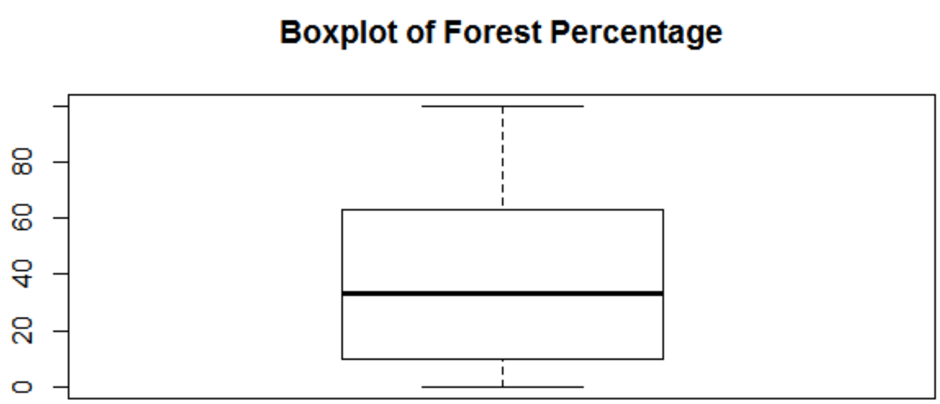
variables should have a linear correlation to bring sense to the model. Another assumption is the normality of the data. Linear regression works better with a data that is normally distributed. A linear regression is not useful for data that is not normal and thus attaining this assumption is crucial before using the data to run a regression analysis.

Exercise 10.33

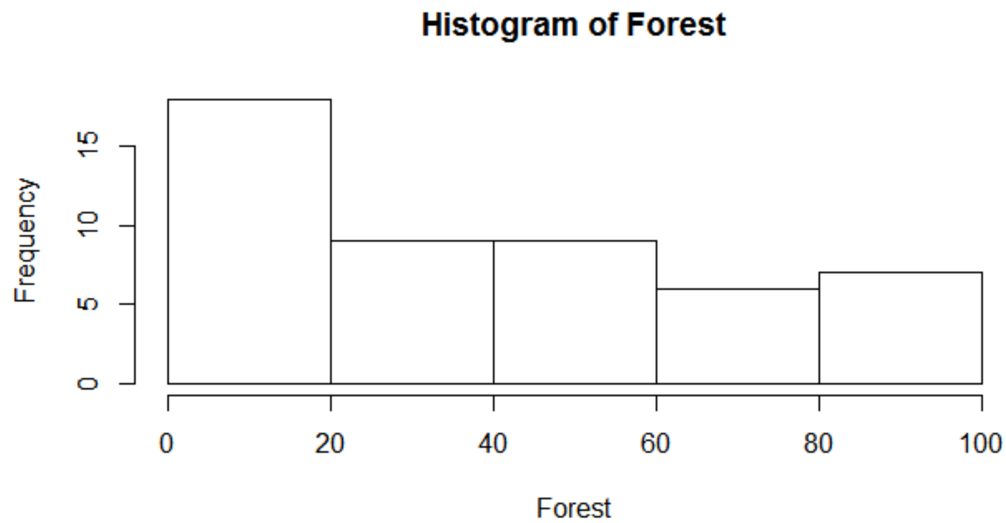
a) Since the summary of IBI was already obtained in the previous part exercise, the summary of the forest percentage will be done in this section.

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

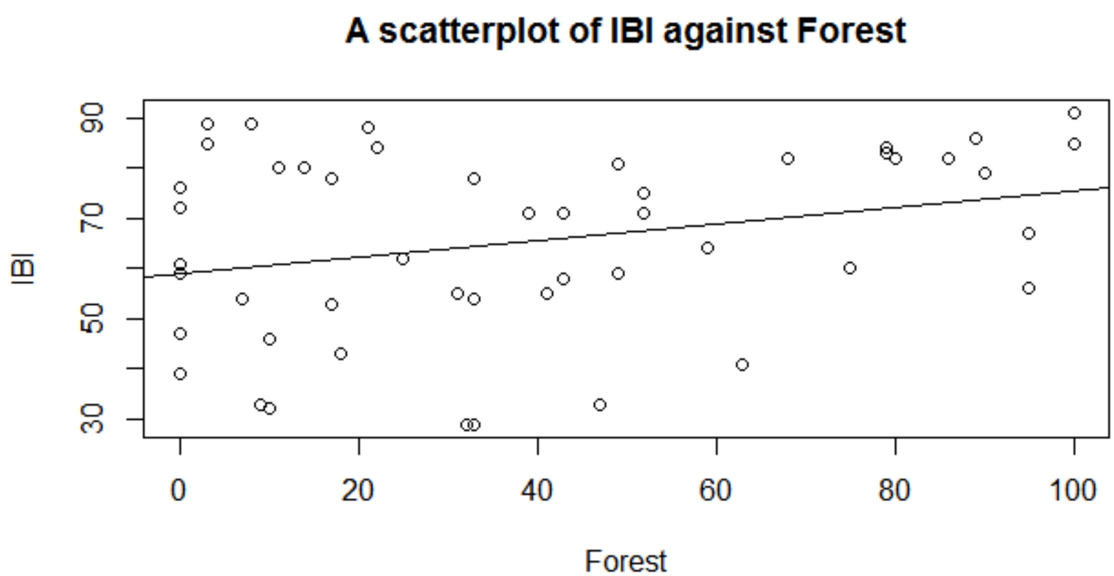
The summary above shows that the median and mean of the variable is 33 and 39.39 respectively. This shows that more data values are concentrated to the lower side of the mean. To put thing into more perspetive, a boxplot was generated.



Further, a histogram was generated for the same and it provided the evidence that more values were to the left side of the mean. Beside showing that the data is skewed to the left, the histogram also showed that the data is not normally distributed.



b) The relationship between IBI and Forest percentage can be investigated using a scatter plot.



The above scatter plot indicates that IBI and Forest have a positive linear relationship, but it can also be seen that not so many data points are close to the linear trend-line. This is an indication of a weak relationship between the two variables.

c) The simple linear model for the relationship between IBI and Area can be statistically written as below:

$$IBI = \beta_0 + \beta_1 * Forest$$

d) The null and alternative hypothesis for the model in part (c) can be stated statistically as below

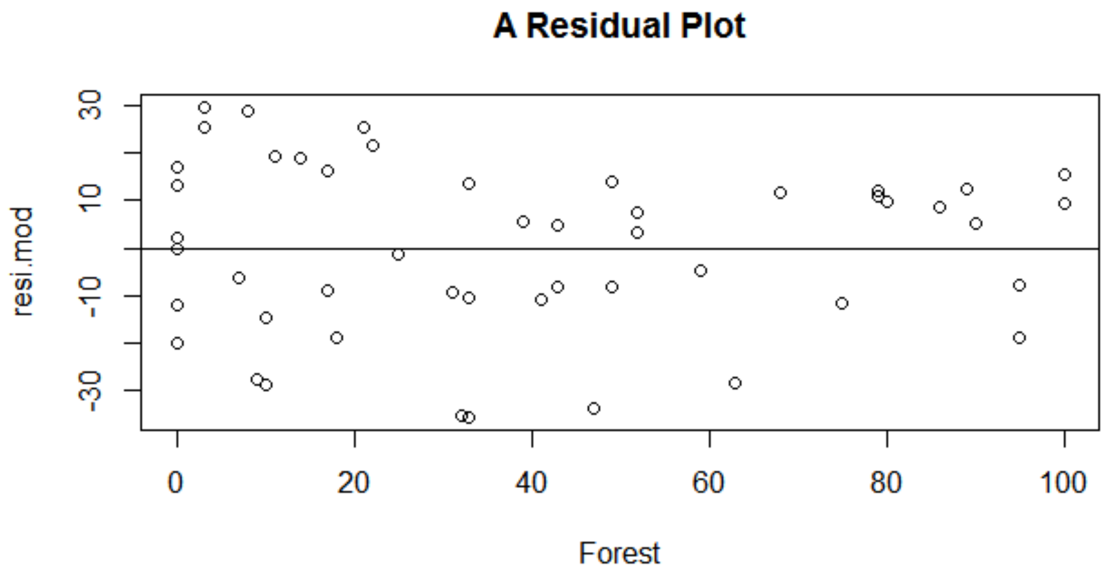
$$H_0: \beta_1 = 0 \; vs \;\; H_1: \beta_1 \neq 0$$

e) Regression results of the model in part (c) above as shown below

```
Call:
lm(formula = IBI ~ Forest)

Residuals:
    Min      1Q  Median      3Q     Max 
-35.469 -10.798   3.374  13.013  29.515 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 58.98657    4.02087  14.670   <2e-16 ***
Forest       0.16614    0.07936   2.094   0.0417 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 17.71 on 47 degrees of freedom
Multiple R-squared:  0.08531,	Adjusted R-squared:  0.06585 
F-statistic: 4.383 on 1 and 47 DF,  p-value: 0.04171
```

From the summary of the model above, it can be seen that the model is significant at 0.05 level of significance and hence the null hypothesis is rejected; F (1, 47) = 4.383, p=0.0417. This implies that the percentage of watershed area that was forest for each stream is a significant predictor of IBI. Also, the model has a very weak goodness of fit coefficient of 0.08531.

f) The residual plot is as shown below



This residual plot shows a slightly abnormal plot with more values to the left side of the graph compared to the right.

g) The residuals do not appear to be approximately normal because data points are not evenly distributed across the plot.

Exercise 10.34

The two models have turned out to be all significant at 0.05 level of significance. Even though the two models can predict IBI, it is evident from the results that one model is better than the other. When looking at the obtained p-values of the two model the first model of Area has a smaller p-value of 0.000, while the second model has a p-value of 0.0417. This indicates that the first model is more significant compared to the second one. The best measure of a powerful model is, however, not the p-value but the coefficient of determination. The model of Area has a coefficient of determination of 0.2292 which is larger than the second model which has a coefficient of 0.08531. Since the first model has a larger coefficient of determination, it has a better predictive power and hence better as compared to the second one.

Exercise 10.35

For this exercise, I change case 11 and 30 of the IBI data and run a regression on R and obtained the output below

```
Call:
lm(formula = ibi ~ forest)

Residuals:
    Min      1Q  Median      3Q     Max 
-69.274 -11.182   1.045  17.055  30.485 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 58.18233    5.05088  11.519 2.75e-15 ***
forest       0.11092    0.09968   1.113    0.272    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 22.24 on 47 degrees of freedom
Multiple R-squared:  0.02566,	Adjusted R-squared:  0.004934 
F-statistic: 1.238 on 1 and 47 DF,  p-value: 0.2715
```

The output above shows that the model has not become totally insignificant; p=0.2715. This shows that outliers can affect the significance of a model in a negative way. Having outliers affects the distribution of the data and affect its linearity. Presence of outliers makes data to be less linear and thus unable to produce a suitable linear regression model.

Exercise 10.36

a) Confidence interval at 95% level of confidence

```
        fit      lwr      upr
1  61.92376 56.86528 66.98224
2  68.35952 63.44646 73.27257
3  54.49788 46.94304 62.05271
4  74.79528 67.96292 81.62764
5  56.47811 49.72863 63.22760
6  75.78540 68.55498 83.01581
7  62.91387 58.03500 67.79275
8  67.36940 62.59452 72.14428
9  57.46823 51.09270 63.84376
10 59.44846 53.74454 65.15238
11 65.88423 61.20914 70.55931
12 75.78540 68.55498 83.01581
13 65.38917 60.71734 70.06100
14 55.48800 48.34436 62.63164
15 79.74586 70.78212 88.70961
```

16	55.98306	49.03881	62.92730
17	66.87434	62.14749	71.60120
18	56.47811	49.72863	63.22760
19	61.92376	56.86528	66.98224
20	64.39905	59.68831	69.10979
21	66.87434	62.14749	71.60120
22	77.27057	69.41307	85.12807
23	61.92376	56.86528	66.98224
24	55.48800	48.34436	62.63164
25	60.43858	55.02285	65.85432
26	54.00282	46.23684	61.76880
27	60.43858	55.02285	65.85432
28	64.39905	59.68831	69.10979
29	64.89411	60.21036	69.57786
30	64.39905	59.68831	69.10979
31	67.36940	62.59452	72.14428
32	52.51765	44.09939	60.93590
33	80.73598	71.31335	90.15861
34	80.24092	71.04873	89.43311
35	60.93364	55.64772	66.21956
36	58.45835	52.43281	64.48388
37	59.44846	53.74454	65.15238
38	55.98306	49.03881	62.92730
39	62.91387	58.03500	67.79275
40	65.38917	60.71734	70.06100
41	68.35952	63.44646	73.27257
42	86.18162	74.12586	98.23739
43	85.68657	73.87597	97.49716
44	78.26069	69.96866	86.55272
45	70.83481	65.36041	76.30921
46	55.98306	49.03881	62.92730
47	61.92376	56.86528	66.98224
48	78.26069	69.96866	86.55272
49	64.39905	59.68831	69.10979

>

After running the analysis in R the output above was produced indicating that the confidence interval of the IBI for an area of 40km squared is (65.39, 70.06). This implies that there is a 95% confidence that the true population mean response lies between 65.39 and 70.06.

b) Prediction interval at 95% level of confidence

	fit	lwr	upr
1	61.92376	28.83634	95.01118
2	68.35952	35.29402	101.42502
3	54.49788	20.93801	88.05775
4	74.79528	41.39063	108.19992
5	56.47811	23.09032	89.86591
6	75.78540	42.29707	109.27372
7	62.91387	29.85344	95.97431
8	67.36940	34.32415	100.41465
9	57.46823	24.15402	90.78244
10	59.44846	26.25624	92.64069
11	65.88423	32.85325	98.91520
12	75.78540	42.29707	109.27372
13	65.38917	32.35865	98.41968
14	55.48800	22.01830	88.95769

```
15 79.74586 45.84103 113.65070
16 55.98306 22.55535  89.41076
17 66.87434 33.83600  99.91269
18 56.47811 23.09032  89.86591
19 61.92376 28.83634  95.01118
20 64.39905 31.36301  97.43509
21 66.87434 33.83600  99.91269
22 77.27057 43.64128 110.89987
23 61.92376 28.83634  95.01118
24 55.48800 22.01830  88.95769
25 60.43858 27.29466  93.58250
26 54.00282 20.39479  87.61085
27 60.43858 27.29466  93.58250
28 64.39905 31.36301  97.43509
29 64.89411 31.86190  97.92631
30 64.39905 31.36301  97.43509
31 67.36940 34.32415 100.41465
32 52.51765 18.75293  86.28236
33 80.73598 46.70695 114.76501
34 80.24092 46.27498 114.20687
35 60.93364 27.81069  94.05659
36 58.45835 25.20934  91.70735
37 59.44846 26.25624  92.64069
38 55.98306 22.55535  89.41076
39 62.91387 29.85344  95.97431
40 65.38917 32.35865  98.41968
41 68.35952 35.29402 101.42502
42 86.18162 51.33151 121.03174
43 85.68657 50.92050 120.45263
44 78.26069 44.52722 111.99416
45 70.83481 37.68126 103.98837
46 55.98306 22.55535  89.41076
47 61.92376 28.83634  95.01118
48 78.26069 44.52722 111.99416
49 64.39905 31.36301  97.43509
```

After running the analysis in R, it emerged that the IBI when area is 40km squared has a prediction interval of (65.39, 98.42) at 95% level of confidence. This indicates that a predicted future value of IBI when area is 40km squared will lie within this interval with 95% confidence level.

c) After running the analysis in R the output above was produced indicating that the confidence interval of the IBI for an area of 40km squared is (65.39, 70.06). This implies that there is a 95% confidence that the true population mean response of Ozark highland streams lies between 65.39 and 70.06. Moreover, it emerged that the IBI when area is 40km squared has a prediction interval of (65.39, 98.42) at 95% level of confidence. This indicates that a predicted future value of IBI when area is 40km squared will lie within this interval with 95% confidence level.

d) These results cannot be applied to other streams in Arkansas because the sample was not to large enough to be used to make generalization and the sampling was not done randomly across all rivers in Arkansas.

Exercise 10.37

The estimates of the index of biotic integrity can be computed for 10km squared area and 63% forest as shown below

Area 10km squared

$$IBI = 51.53 + 0.5 * Area$$

$$IBI = 51.53 + 0.5 * 10 = 56.53$$

Forest 63%

$$IBI = 58.99 + 0.17 * forest$$

$$IBI = 58.99 + 0.17 * 63 = 69.7$$

The two estimates differ due to the rounding off issue. The percentages were rounded of and hence end up giving larger values as compared to the actual measurement.