

Max Shi

MA 331

Professor Li

November 29, 2019

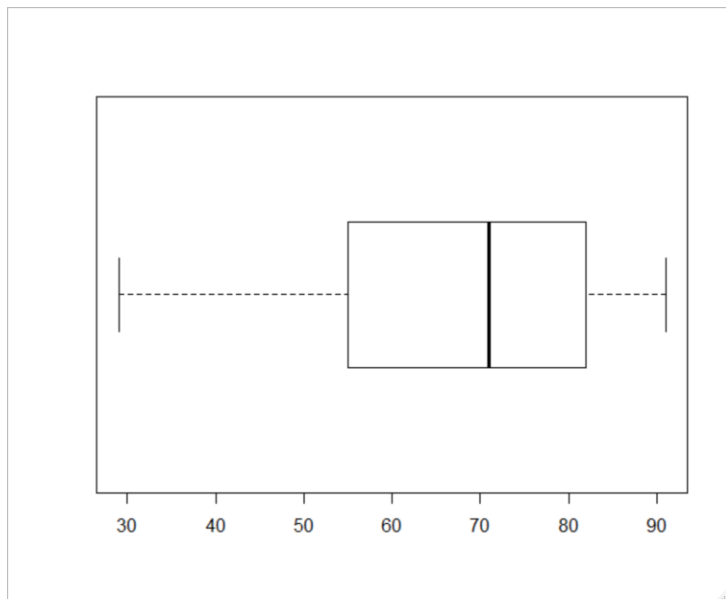
I pledge my honor that I have abided by the Stevens Honor System.

Homework 7

10.32.

a.

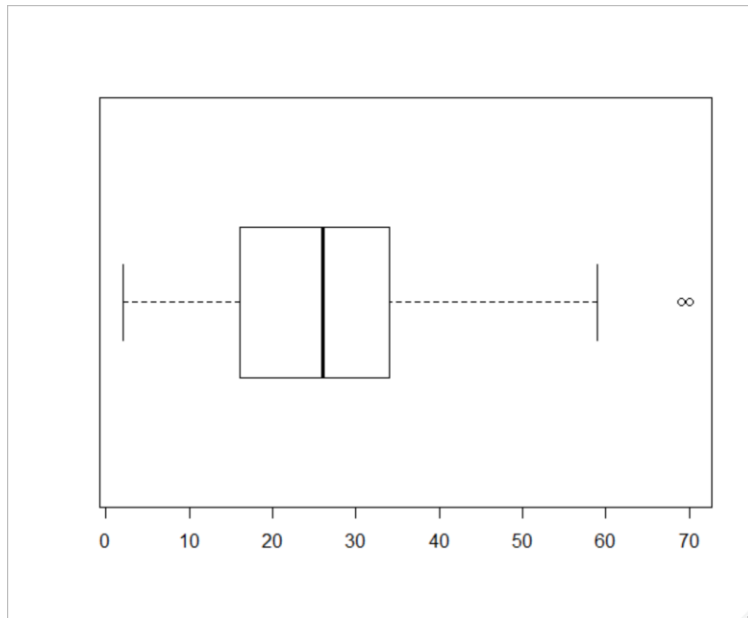
(ibi)



boxplot of ibi shows a left skew.

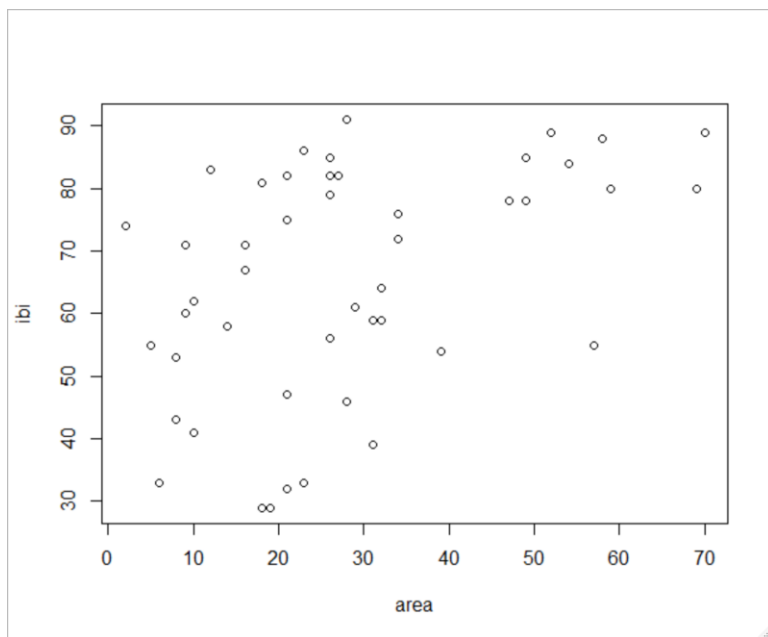
Mean of ibi = 65.94, standard deviation = 18.28.

(area)



The boxplot shows a slight right skew, as well as a few outliers in the data. The mean is 28.29 and the standard deviation is 17.71.

b.



There seems to be a positive correlation, enforced by a calculated correlation of 0.445. There seems to be an outlier near the (58,55) point.

c.

In our plot, where IBI = y and area = x , the linear regression model is $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ for all i in our dataset, where x and y are independent, normally distributed variables.

d. $H_0: \beta_1 = 0$

$H_a: \beta_1 \neq 0$.

e.

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

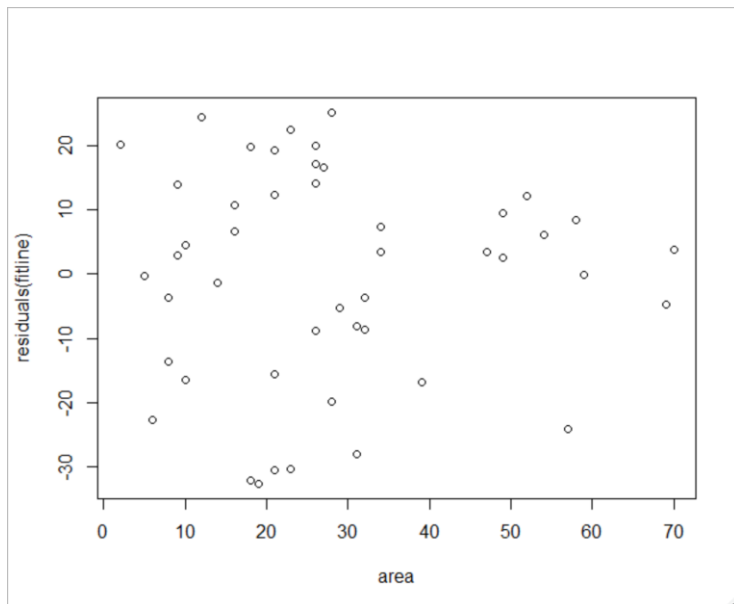
Residual standard error: 16.53 on 47 degrees of freedom

Multiple R-squared: 0.1988, Adjusted R-squared: 0.1818

F-statistic: 11.67 on 1 and 47 DF, p-value: 0.001322

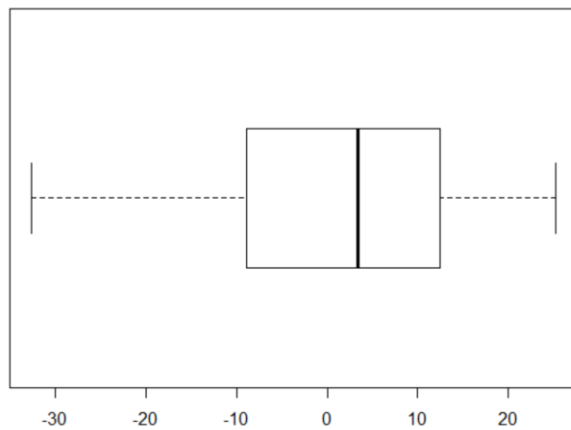
The linear regression gives $\beta_0 = 52.92$ and $\beta_1 = 0.460$. The standard error here is 16.53, and for the hypothesis tests, the t value is 3.415 and the p value is 0.00132.

f.



There does not seem to be anything too unusual in the plot.

g.



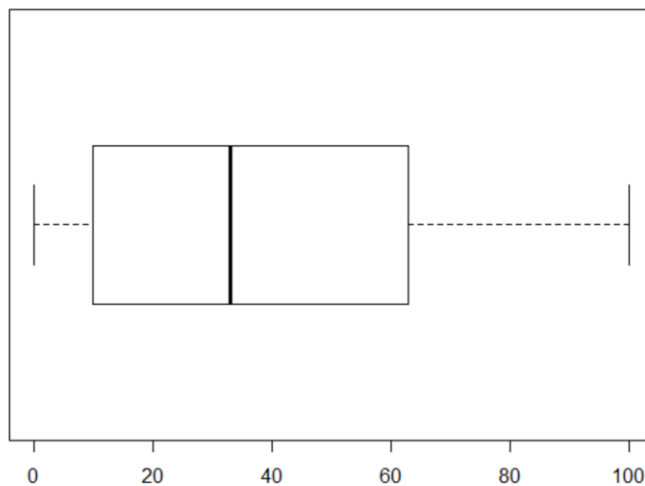
The residuals are left skewed.

g.

The assumptions here seem to be reasonable. There is nothing unusual about the plot of the area vs the residuals, and although the boxplots of the data and the residuals are skewed slightly, the large number of observations dismiss this concern.

10.33.

a. (percent forested)

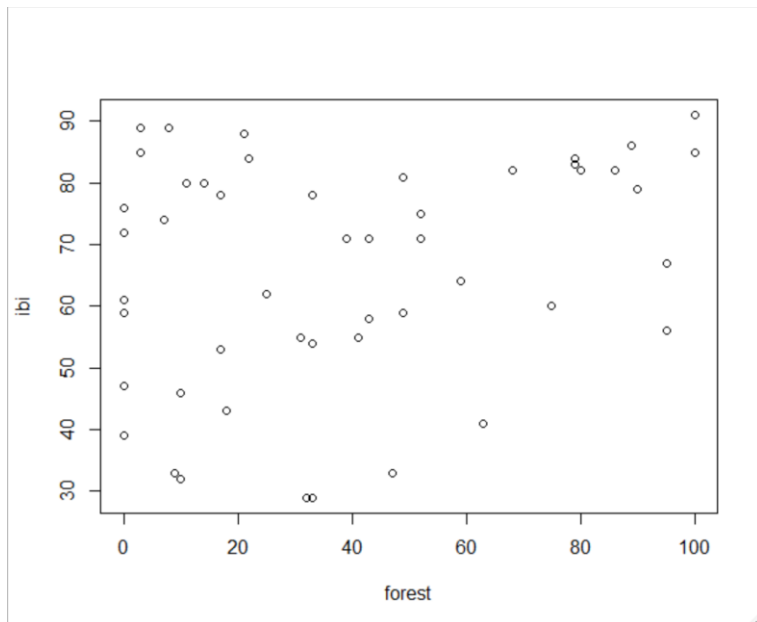


Boxplot shows a right skew.

The average is 39.39%, with a standard deviation of 32.02%.

(The observations of IBI can be taken from the previous exercise)

b.



From the plot of forest vs ibi, there seems to be a sort of positive correlation between the two variables. The calculated correlation in R is 0.269. However, among small values of x, the values of y are scattered.

c.

In our plot, where IBI = y and forest = x, the linear regression model is $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ for all i in our dataset, where x and y are independent, normally distributed variables.

d. $H_0 : \beta_1 = 0$

$H_a : \beta_1 \neq 0$.

e.

Linear Regression:

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

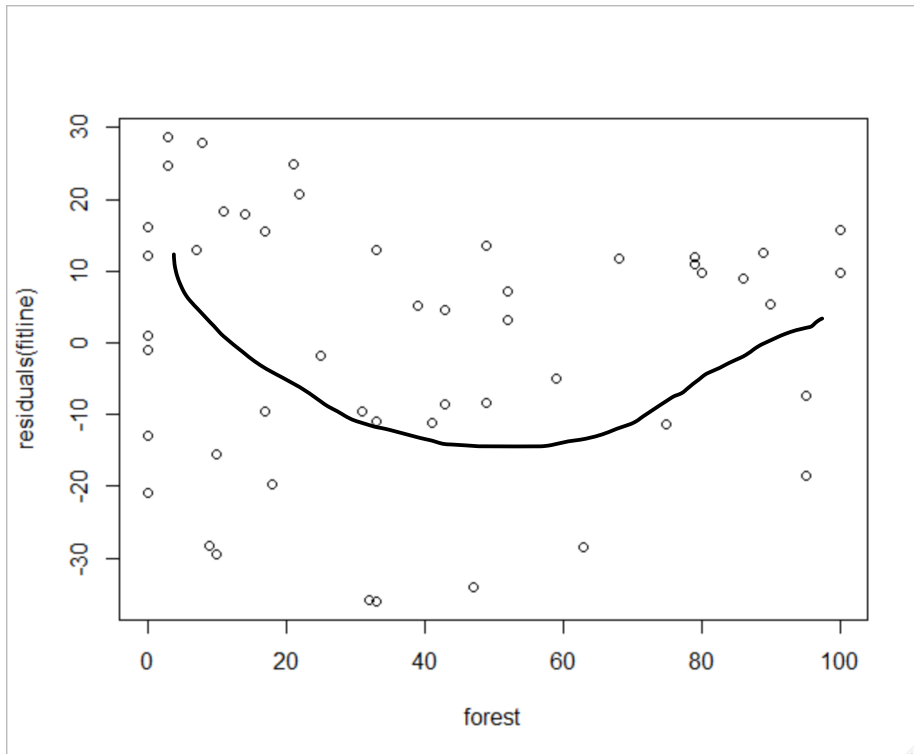
Residual standard error: 17.79 on 47 degrees of freedom

Multiple R-squared: 0.07278, Adjusted R-squared: 0.05305

F-statistic: 3.689 on 1 and 47 DF, p-value: 0.06084

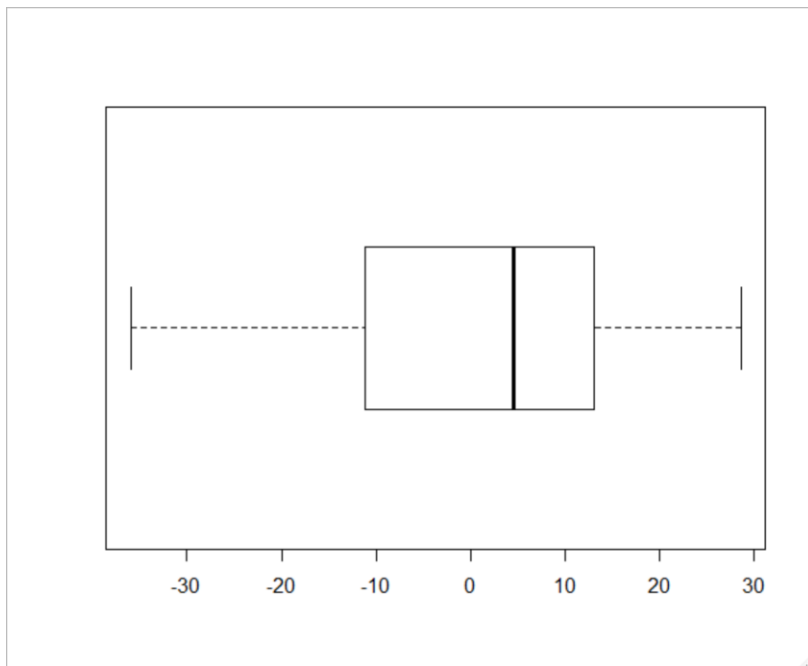
Here, $\beta_0 = 59.91$ and $\beta_1 = 0.153$ with error being 17.79. Hypothesis tests from d result in the t value being 1.921 and the P value being 0.0608.

f.



Residual plot shows a small curve, hopefully illustrated by the drawn line.

g.



The boxplot of the residuals shows a left skew, possibly suggesting a non-normal distribution.

h. They appear to be not very reasonable. The slight curve seen in the area vs residuals plot may suggest a use of a non-linear plot, and the residuals themselves do not seem to be of normal distribution either. Furthermore, although the data themselves are heavily left and right skewed, also suggesting use of a non-linear model.

10.34.

The two analyses see that using area instead of percent forest yields a much lower p value, as well as the lack of a curve in the x vs residual plot. It also shows a much stronger positive correlation. Thus, I would prefer to use area to predict IBI.

10.35.

With a (forest = 0, ibi = 0) observance added in, the regression becomes:

Residuals:

	Min	1Q	Median	3Q	Max
	-56.969	-10.064	4.111	14.292	31.436

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	56.96922	4.32222	13.18	<2e-16 ***
forest	0.19821	0.08617	2.30	0.0258 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19.52 on 48 degrees of freedom

Multiple R-squared: 0.09928, Adjusted R-squared: 0.08052

F-statistic: 5.291 on 1 and 48 DF, p-value: 0.02583

With a (forest = 100, ibi = 0) observance added in, the regression becomes:

Residuals:

	Min	1Q	Median	3Q	Max
	-68.744	-10.065	5.901	15.655	26.991

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	61.80108	4.60916	13.408	<2e-16 ***
forest	0.06943	0.08844	0.785	0.436

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 20.43 on 48 degrees of freedom

Multiple R-squared: 0.01268, Adjusted R-squared: -0.007892

F-statistic: 0.6163 on 1 and 48 DF, p-value: 0.4363

The statistical significance of the positive correlation goes up when the (0,0) value is added because it reinforces the idea of a positive correlation. The opposite happens when the (100,0) is added, because it does not reinforce the positive correlation that the regression suggests is happening.

10.36.

a. From R, calculating the 95% confidence interval yields

	fit	lwr	upr
1	71.32916	65.61416	77.04417

Thus, the confidence interval for a mean response corresponding to an area of 40 sq km = (65.614, 77.044).

b. From R, calculating the 95% prediction interval yields

	fit	lwr	upr
1	71.32916	37.57836	105.08

Thus, the prediction interval for a future response of 40 sq km = (37.578, 105.08)

c. The confidence interval predicts the average ibi of a piece of land with 40 sqkm of watershed. This is in comparison to the prediction interval, which gives the confidence interval of ibi with only one observation. More specifically, the confidence interval for a mean response gives the confidence interval for an average over multiple responses, while the prediction interval for a future response give the confidence interval for only one response. Thus, the intervals are centered on the same number, but the prediction interval is wider to account for error of prediction of only one observation.

d. I think these results could be applied to other streams in Arkansas, with minimal changes in other features such as climate, rainfall, etc. However, I would not apply this data to other streams in locations with many different changes in factors compared to this stream in Arkansas, for example, a stream in Texas with much less rainfall. It is likely there that percent forest or area of watershed is not the only factor playing into the IBI measurement.

10.37.

Using area = 10, the prediction is:

	fit	lwr	upr
1	57.52451	23.55984	91.48918

Using forest percentage = 63, the prediction is:

	fit	lwr	upr
1	69.55457	33.20852	105.9006

These two prediction intervals differ because of the criteria they use to make the prediction. Because both are linear regressions with a positive correlation, the lower the value, the smaller the prediction will be. This entry has a small area value, but a high forest value, meaning that the predicted ibi using the area as the factor will be lower than the prediction using forest percentage as the factor. However, the prediction intervals overlap, meaning that there is a good chance that an even better prediction could be combining these two factors into a multiple linear regression to examine the relationship there. This is also a good suggestion, as the prediction intervals are very wide, each both about 70 units, so another model may help to strength the confidence of the prediction.