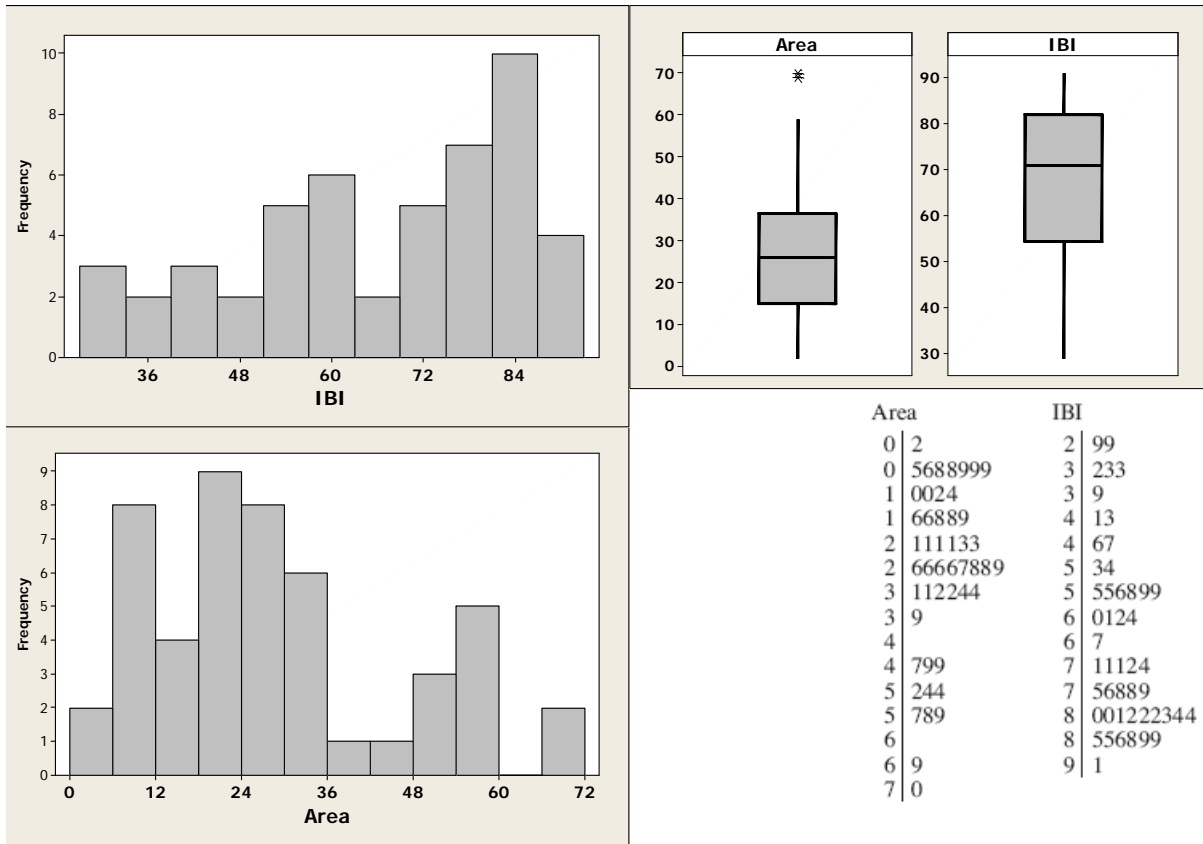


Homework # 13 — Solutions

(1)

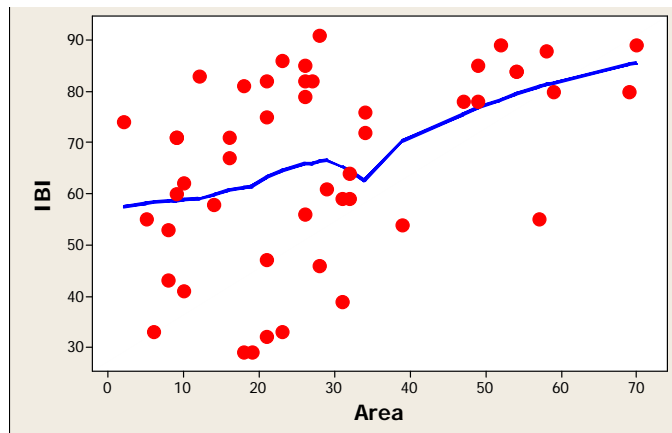


Variable	N	Mean	StDev	Minimum	Q1	Median	Q3	Maximum
IBI	49	65.94	18.28	29.00	54.50	71.00	82.00	91.00
Area	49	28.29	17.71	2.00	15.00	26.00	36.50	70.00

IBI is left-skewed (long left tail), best described by the five-number summary.

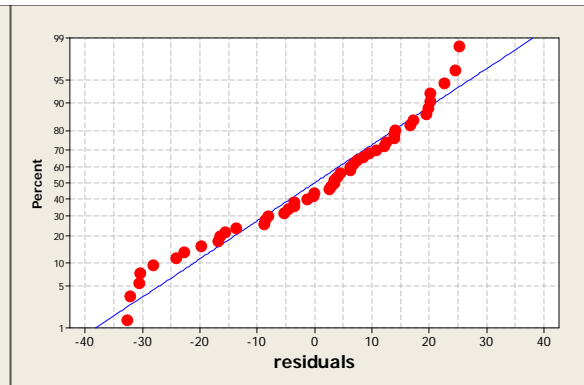
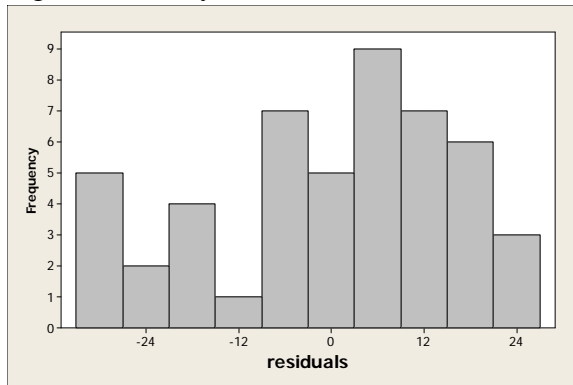
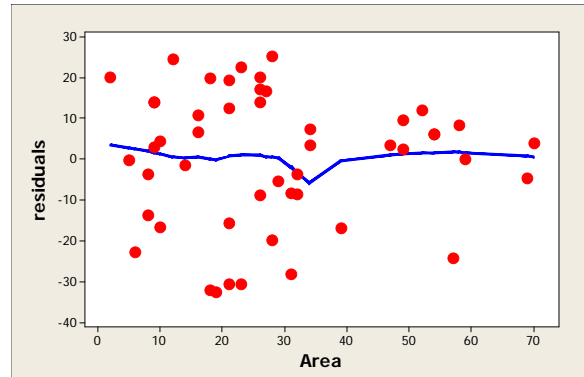
Area is right skewed (long right tail), with possibly a second peak at 50–60 km² and/or two high outliers; the five-number summary again is the better description.

(2) The scatterplot shows a weak positive association. There is more scatter in *IBI* for small values of *Area*. There are two clusters of points, corresponding to the two peaks in the distribution of *Area*, but the relationship between the variables does not seem to differ between these clusters. There are no distinct outliers or unusual observations.



- (3) (a) $y_i = e_0 + e_1x_i + h_i$, $i = 1, 2, \dots, 49$; h_i are independent $N(0, \sigma^2)$ variables.
- (b) The hypotheses are $H_0: e_1 = 0$ vs. $H_a: e_1 \neq 0$.
- (c) The regression equation is $IBI = 52.923 + 0.4602 \text{ Area}$. The estimated standard deviation is $s = 16.53$, and r^2 is 0.199. For testing the hypotheses in (b), $t = 3.42$ and $P = 0.001$; the null hypothesis would be rejected and we would conclude that there is a positive linear relationship.

- (4) (a) The residual plot shows again that there is more variation for small x . There is no indication of nonlinearity (curvature).
- (b) As the histogram and Normal probability plot below show, the residuals are somewhat left-skewed but with a somewhat truncated left tail. The distribution thus is not Normal, but considering the sample size ($n = 49$), the non-Normality is not severe enough to invalidate the least-squares regression analysis.



- (5) Opinions may vary. The two apparent deviations from the model are (i) a possible change in standard deviation as x changes and (ii) possible nonnormality of error terms.
- (6) With the regression equation $IBI = 52.92 + 0.4602 \text{ Area}$, the predicted mean response when $\text{Area} = 30 \text{ km}^2$ is $\hat{p}_y = 66.73$. While it is possible to find $SE[\hat{p}]$ and $SE[\hat{y}]$ using the formulas from Section 10.2, it is easier (and less prone to mistakes) to use software, as in the Minitab output shown below. ($SE[\hat{p}]$ is reported by Minitab as “Stdev.fit.” Note that $SE[\hat{y}] = \sigma \sqrt{s^2 + SE^2[\hat{p}]}$, where $s = 16.53$.)

Minitab output

Fit	Stdev.Fit	95.0% C.I.	95.0% P.I.
66.73	2.37	(61.95, 71.50)	(33.12, 100.34)

- (a) The 95% confidence interval for μ_y is 61.95 to 71.50.
- (b) The 95% prediction interval for a future response is 33.12 to 100.34.

- (c) Among many streams with watershed area 30 km^2 , we estimate the mean IBI to be between about 61.95 and 71.50. For an individual stream with watershed area 30 km^2 , we expect its IBI to be between about 33.12 and 100.34.
- (d) We probably cannot reliably apply these results elsewhere; it is likely that the particular characteristics of the Ozark Highland region play some role in determining the regression coefficients.

- (7) (a) The table on the right shows the correlations and the corresponding test statistics.

	r	t	P
IBI/area	0.4459	3.42	0.0013
IBI/forest	0.2698	1.92	0.0608
area/forest	-0.2571	-1.82	0.0745

There is a fairly strong (and statistically significant) positive correlation between *IBI* and *Area*, a weaker

positive correlation between *IBI* and the percent forested, and a negative correlation (similar in strength to that between *IBI* and % forested) between *Area* and % forested; the latter two correlations would not be significant at the conventional $\alpha = 0.05$, but the P -values are fairly small.

- (b) The results for the correlation between *IBI* and *Area* (first row of the table) agree with the results of (3)(c): r is the square root of r^2 given there, and t and P are identical to the regression results.