Problem set 3. Relationship between variables

Exercise 1. Average temperatures. Here are the average temperatures in degrees for Lafayette, Indiana, during the months of February through May:

Month	February	March	April	May
Temperature (degrees F)	30	41	51	62

(a) Explain why month should be the explanatory variable for examining this relationship.
(b) Make a scatterplot and describe the relationship.

Exercise 2. Explanatory and response variables. In each of the following situations, is it more reasonable to simply explore the relationship between the two variables or to view one of the variables as an explanatory variable and the other as a response variable? In the latter case, which is the explanatory variable and which is the response variable?
(a) The weight of a child and the age of the child from birth to 10 years.
(b) High school English grades and high school math grades.
(c) The rental price of apartments and the number of bedrooms in the apartment.
(d) The amount of sugar added to a cup of coffee and how sweet the coffee tastes.
(e) The student evaluation scores for an instructor and the student evaluation scores for the course.

Exercise 3. Reading ability and IQ. A study of reading ability in schoolchildren chose 60 fifth-grade children at random from a school. The researchers had the children's scores on an IQ test and on a test of reading ability.

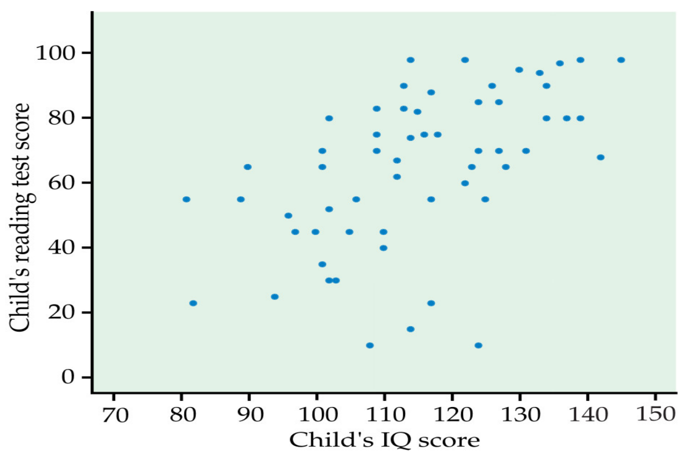

FIGURE 2.6 IQ and reading test scores for 60 fifth-grade children, for Exercise 2.11.

(a) Explain why we should expect a positive association between IQ and reading score for children in the same grade. Does the scatterplot show a positive association?
(b) A group of four points appear to be outliers. In what way do these children's IQ and reading scores deviate from the overall pattern?
(c) Ignoring the outliers, is the association between IQ and reading scores roughly linear? Is it very strong? Explain your answers.

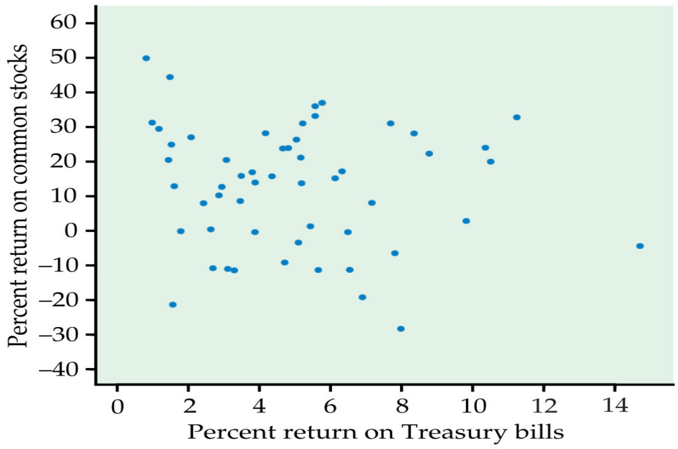

FIGURE 2.7 Percent return on Treasury bills and common stocks for the years 1950 to 2003, for Exercise 2.12.

plots the annual returns on stocks for the years 1950 to 2003 against the returns on Treasury bills for the same years.

(a) The best year for stocks during this period was 1954. The worst year was 1974. About what were the returns on stocks in those two years?
(b) Treasury bills are a measure of the general level of interest rates. The years around 1980 saw very high interest rates. Treasury bill returns peaked in 1981. About what was the percent return that year?
(c) Some people say that high Treasury bill returns tend to go with low returns on stocks. Does such a pattern appear clearly in figure 2.7? Does the plot have any clear pattern?

Exercise 5. Can children estimate their reading ability? The main purpose of the study cited in Exercise 3 was to ask whether schoolchildren can estimate their own reading ability. The researchers had the children's scores on a test of reading ability. They asked each child to estimate his or her reading level, on a scale from 1 (low) to 5 (high). figure 2.8

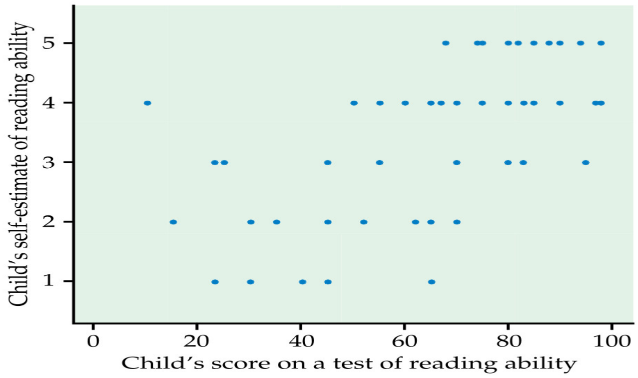

FIGURE 2.8 Reading test scores for 60 fifth-grade children and the children's estimates of their own reading levels, for Exercise 2.13.

is a scatterplot of the children's estimates (response) against their reading scores (explanatory).

(a) What explains the "stair-step" pattern in the plot?

(b) Is there an overall positive association between reading score and self-estimate?

(c) There is one clear outlier. What is this child's self-estimated reading level? Does this appear to over- or underestimate the level as measured by the test?

Exercise 6. Literacy of men and women. Table 1.2

Country	Female (%)	Male (%)	Country	Female (%)	Male (%)
Algeria	60	78	Morocco	38	68
Bangladesh	31	50	Saudi Arabia	70	84
Egypt	46	68	Syria	63	89
Iran	71	85	Tajikistan	99	100
Jordan	86	96	Tunisia	63	83
Kazakhstan	99	100	Turkey	78	94
Lebanon	82	95	Uzbekistan	99	100
Libya	71	92	Yemen	29	70
Malaysia	85	92			

shows the percent of men and women at least 15 years old who were literate in 2002 in the major Islamic nations for which data were available. Make a scatterplot of these data, taking male literacy as the explanatory variable. Describe the direction, form, and strength of the relationship. Are there any identical observations that plot as the same point? Are there any clear outliers?

Exercise 7. World records for the 10K. Table 2.3 shows the progress of world record times (in seconds) for the 10,000-meter run up to mid-2004. Concentrate on the women's world record times. Make a scatterplot with year as the explanatory variable. Describe the pattern of improvement over time that your plot displays.

World record times for the 10,000-meter run					
Men				Women	
Record year	Time (seconds)	Record year	Time (seconds)	Record year	Time (seconds)
1912	1880.8	1962	1698.2	1967	2286.4
1921	1840.2	1963	1695.6	1970	2130.5
1924	1835.4	1965	1659.3	1975	2100.4
1924	1823.2	1972	1658.4	1975	2041.4
1924	1806.2	1973	1650.8	1977	1995.1
1937	1805.6	1977	1650.5	1979	1972.5
1938	1802.0	1978	1642.4	1981	1950.8
1939	1792.6	1984	1633.8	1981	1937.2
1944	1775.4	1989	1628.2	1982	1895.3
1949	1768.2	1993	1627.9	1983	1895.0
1949	1767.2	1993	1618.4	1983	1887.6
1949	1761.2	1994	1612.2	1984	1873.8
1950	1742.6	1995	1603.5	1985	1859.4
1953	1741.6	1996	1598.1	1986	1813.7
1954	1734.2	1997	1591.3	1993	1771.8
1956	1722.8	1997	1587.8		
1956	1710.4	1998	1582.7		
1960	1698.8	2004	1580.3		

Table 2.3

Exercise 8. How do icicles grow? How fast do icicles grow? Japanese researchers measured the growth of icicles in a cold chamber under various conditions of temperature, wind, and water flow. Table 2.4 contains data produced under two sets of conditions. In both cases, there was no wind and the temperature was set at -11° C. Water flowed over the icicle at a higher rate (29.6 milligrams per second) in Run 8905 and at a slower rate (11.9 mg/s) in Run 8903.

Growth of icicles over time

Run 8903				Run 8905			
Time (min)	Length (cm)	Time (min)	Length (cm)	Time (min)	Length (cm)	Time (min)	Length (cm)
10	0.6	130	18.1	10	0.3	130	10.4
20	1.8	140	19.9	20	0.6	140	11.0
30	2.9	150	21.0	30	1.0	150	11.9
40	4.0	160	23.4	40	1.3	160	12.7
50	5.0	170	24.7	50	3.2	170	13.9
60	6.1	180	27.8	60	4.0	180	14.6
70	7.9			70	5.3	190	15.8
80	10.1			80	6.0	200	16.2
90	10.9			90	6.9	210	17.9
100	12.7			100	7.8	220	18.8
110	14.4			110	8.3	230	19.9
120	16.6			120	9.6	240	21.1

Table 2.4

(a) Make a scatterplot of the length of the icicle in centimeters versus time in minutes, using separate symbols for the two runs.
(b) Write a careful explanation of what your plot shows about the growth of icicles.

Exercise 9. Records for men and women in the 10K. Table 2.3 above shows the progress of world record times (in seconds) for the 10,000-meter run for both men and women.
(a) Make a scatterplot of world record time against year, using separate symbols for men and women. Describe the pattern for each sex. Then compare the progress of men and women.
(b) Women began running this long distance later than men, so we might expect their improvement to be more rapid. Moreover, it is often said that men have little advantage over women in distance running as opposed to sprints, where muscular strength plays a greater role. Do the data appear to support these claims?

Exercise 10. Coffee prices and deforestation. Coffee is a leading export from several developing countries. When coffee prices are high, farmers often clear forest to plant more coffee trees. Here are data for five years on prices paid to coffee growers in Indonesia and the rate of deforestation in a national park that lies in a coffee-producing region:

Price (cents per pound)	Deforestation (percent)
29	0.49
40	1.59
54	1.69
55	1.82
72	3.10

(a) Make a scatterplot. Which is the explanatory variable? What kind of pattern does your plot show?
(b) find the correlation r step-by-step. That is, find the mean and standard deviation of the two variables. Then find the five standardized values for each variable and use the formula for r. Explain how your value for r matches your graph in (a).
(c) Now enter these data into your calculator or software and use the correlation function to find r. Check that you get the same result as in (b).

Exercise 11. First test and final exam. We return to the relationship between the score on the first test and the score on the final exam in an elementary statistics course. The data for eight students from such a course are presented in the following table.

first-test score	153	144	162	149	127	118	158	153
final-exam score	145	140	145	170	145	175	170	160

(a) find the correlation between these two variables.
(b) Is the observed relationship weak? Does your calculation of the correlation support this statement? Explain your answer.

Exercise 12. Second test and final exam. Refer to the previous exercise. Here are the data for the second test and the final exam for the same students:

Second-test score	158	162	144	162	136	158	175	153
final-exam score	145	140	145	170	145	175	170	160

(a) find the correlation between these two variables.
(b) Is the observed relationship between these two variables stronger than the relationship between the two variables in the previous exercise. How do the values of the correlations that you calculated support this statement? Explain your answer.

Exercise 13. IQ and reading scores. Figure 2.6 (above) displays the positive association between the IQ scores of fifth-grade students and their reading scores. Do you think the correlation between these variables is closest to $r = 0.1$, $r = 0.6$, or $r = 0.9$? Explain the reason for your guess.

Exercise 14. An interesting set of data. Make a scatterplot of the following data.

x	1	2	3	4	10	10
y	1	3	3	5	1	11

Use your calculator to show that the correlation is about 0.5. What feature of the data is responsible for reducing the correlation to this value despite a strong straight-line association between x and y in most of the observations?

Exercise 15. City and highway gas mileage. Table 1.10

Two-Seater Cars			Minicompact Cars		
Model	City	Highway	Model	City	Highway
Acura NSX	17	24	Aston Martin Vanquish	12	19
Audi TT Roadster	20	28	Audi TT Coupe	21	29
BMW Z4 Roadster	20	28	BMW 325CI	19	27
Cadillac XLR	17	25	BMW 330CI	19	28
Chevrolet Corvette	18	25	BMW M3	16	23
Dodge Viper	12	20	Jaguar XK8	18	26
Ferrari 360 Modena	11	16	Jaguar XKR	16	23
Ferrari Maranello	10	16	Lexus SC 430	18	23
Ford Thunderbird	17	23	Mini Cooper	25	32
Honda Insight	60	66	Mitsubishi Eclipse	23	31
Lamborghini Gallardo	9	15	Mitsubishi Spyder	20	29
Lamborghini Murcielago	9	13	Porsche Cabriolet	18	26
Lotus Esprit	15	22	Porsche Turbo 911	14	22
Maserati Spyder	12	17			
Mazda Miata	22	28			
Mercedes-Benz SL500	16	23			
Mercedes-Benz SL600	13	19			
Nissan 350Z	20	26			
Porsche Boxster	20	29			
Porsche Carrera 911	15	23			
Toyota MR2	26	32			

gives the city and highway gas mileages for 21 two-seater cars, including the Honda Insight gas-electric hybrid car.

(a) Make a scatterplot of highway mileage y against city mileage x for all 21

cars. There is a strong positive linear association. The Insight lies far from the other points. Does the Insight extend the linear pattern of the other cars, or is it far from the line they form?
(b) find the correlation between city and highway mileages both without and with the Insight. Based on your answer to (a), explain why r changes in this direction when you add the Insight.

Exercise 16. A property of the least-squares regression line. Use the equation for the least-squares regression line to show that this line always passes through the point $(\bar{x}, \bar{y})$.

Exercise 17. Icicle growth. The data for Run 8903 in Table 2.4 (above) describe how the length y in centimeters of an icicle increases over time x. Time is measured in minutes.
(a) What are the numerical values and units of measurement for each of x, s_x, y, s_y, and the correlation r between x and y?
(b) There are 2.54 centimeters in an inch. If we measure length y in inches rather than in centimeters, what are the new values of y, s_y , and the correlation r?
(c) If we measure length y in inches rather than in centimeters, what is the new value of the slope b_1 of the least-squares line for predicting length from time?

Exercise 18. Full-time and part-time college students. The Census Bureau provides estimates of numbers of people in the United States classified in various ways. Let's look at college students. The following table gives us data to examine the relation between age and full-time or part-time status. The numbers in the table are expressed as thousands of U.S. college students.

U.S. college students by age and status

	Status	
Age	Full-time	Part-time
15–19	3388	389
20–24	5238	1164
25–34	1703	1699
35 and over	762	2045

(a) What is the U.S. Census Bureau estimate of the number of full-time college students aged 15 to 19?
(b) Give the joint distribution of age and status for this table.
(c) What is the marginal distribution of age? Display the results graphically.
(d) What is the marginal distribution of status? Display the results graphi-

cally.

Exercise 19. Predicting text pages. The editor of a statistics text would like to plan for the next edition. A key variable is the number of pages that will be in the final version. Text files are prepared by the authors using a word processor called LATEX, and separate files contain figures and tables. For the previous edition of the text, the number of pages in the LATEX files can easily be determined, as well as the number of pages in the final version of the text. The table presents the data.

Chapter	1	2	3	4	5	6	7	8	9	10	11	12	13
LATEX pages	77	73	59	80	45	66	81	45	47	43	31	46	26
Text pages	99	89	61	82	47	68	87	45	53	50	36	52	19

(a) Plot the data and describe the overall pattern.
(b) find the equation of the least-squares regression line and add the line to your plot.
(c) find the predicted number of pages for the next edition if the number of LATEX pages is 62.
table
(d) Write a short report for the editor explaining to her how you constructed the regression equation and how she could use it to estimate the number of pages in the next edition of the text.

Exercise 20. Endangered animals and habitat. Endangered animal species often live in isolated patches of habitat. If the population size in a patch varies a lot (due to weather, for example), the species is more likely to disappear from that patch in a bad year. Here is a general question: Is there less variation in population size when a patch of habitat has more diverse vegetation? If so, maintaining habitat diversity can help protect endangered species. A researcher measured the variation over time in the population of a cricket species in 45 habitat patches. He also measured the diversity of each patch. He reported his results by giving the least-squares equation

population variation$= 84.4 - 0.13\times$ diversity

along with the fact that $r^2 = 0.34$. Do these results support the idea that more diversity goes with less variation in population size? Is the relationship very strong or only moderately strong?