Chapter 4. Scatterplots and Correlation

Explanatory and Response Variables

Definition. A **response variable** measures an outcome of a study. An **explanatory variable** may explain or influence changes in a response variable.

Example. Exercise 4.2 page 92.

Displaying Relationships: Scatterplots

Definition. A **scatterplot** shows the relationship between two quantitative variables measured on the same individuals. The values of one variable appear on the horizontal axis, and the values of the other variable appear on the vertical axis. Each individual in the data appears as the point in the plot fixed by the values of both variables for that individual. Always plot the explanatory variable, if there is one, on the horizontal axis (the $x-$axis) of a scatterplot. We usually call the explanatory variable x and the response variable y. If there is no explanatory-response distinction, either variable can go on the horizontal axis.

Interpreting Scatterplots

Note. In any graph of data, look for the overall pattern and for striking deviations from that pattern You can describe the overall pattern of a scatterplot by the direction, form, and strength of the relationship. An important kind of deviation is an outlier, an individual value that falls outside the overall pattern of the relationship.

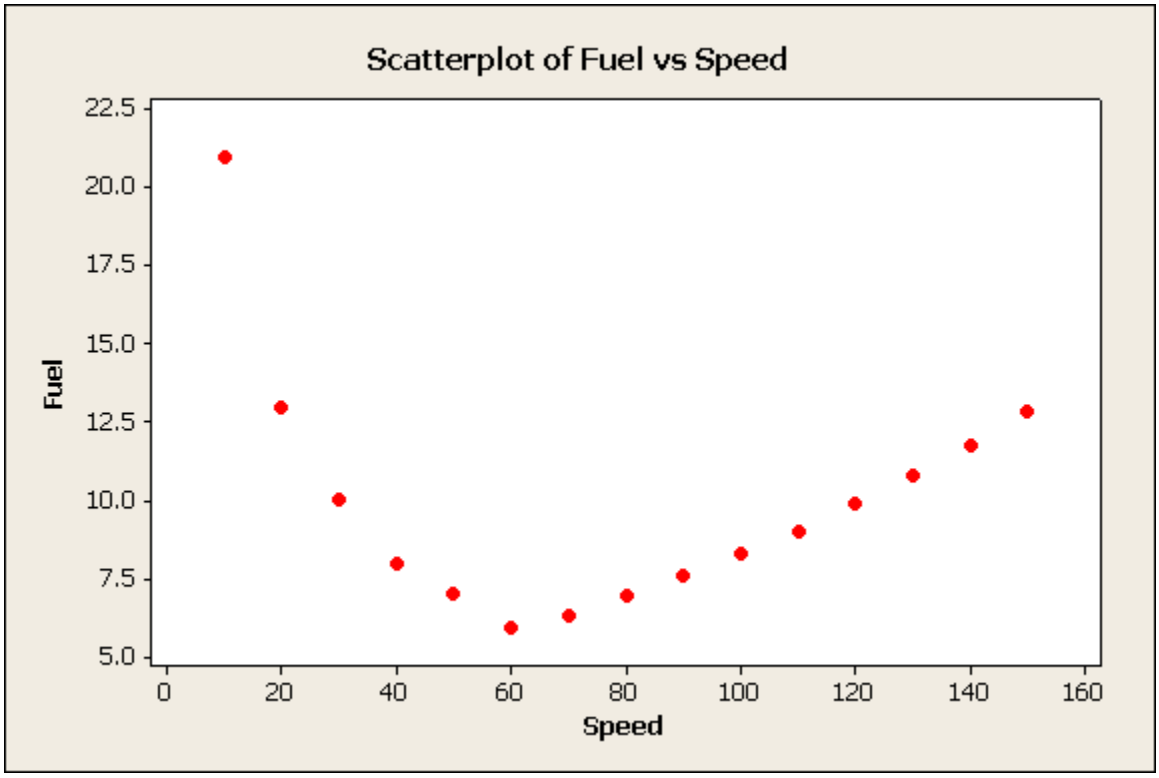
Definition. Two variables are **positively associated** when above-average values of one tend to accompany above average values of the other and below-average values also tend to occur together. Two variables are **negatively associated** when above-average values of one accompany below-average values of the other, and vice versa.

Example. Exercise 4.39 page 113.

Definition. If the points of a scatterplot lie roughly along a straight line, the relationship is said to be **linear**.

Example. Exercise 4.6 page 96.

Partial Solution. We can create a scatterplot using Minitab (enter the data into a worksheet, click on `Graph` and `Scatter plot`). We then get:



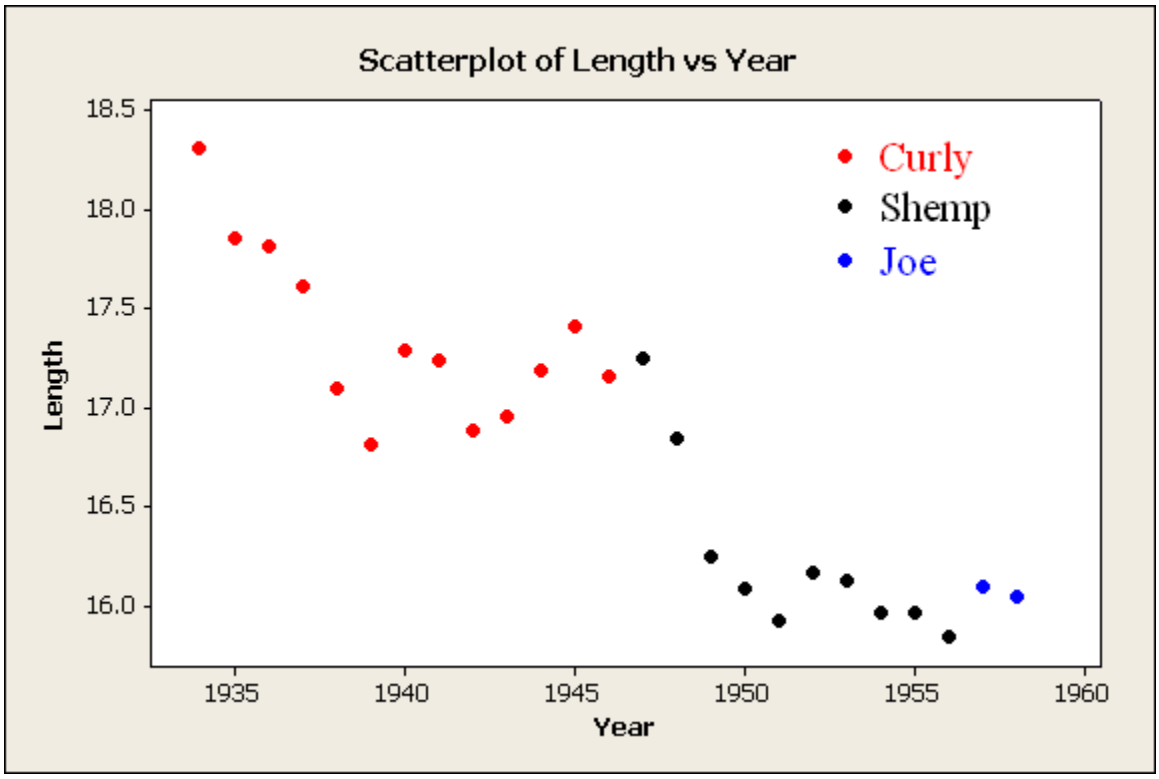
Adding Categorical Variables to Scatterplots

Note. We can subdivide the data in a scatterplot by adding a categorical variable which we represent by a different symbol or color than that used in plotting the remainder of the data.

Example S.4.1. Scattered Stooges.
Based on the length of individual Stooge films as reported in *The Complete Three Stooges* by Jon Solomon, one can arrive at the following average film lengths:

Year	Average Length	Year	Average Length
1934	18.31	1947	17.25
1935	17.85	1948	16.84
1936	17.81	1949	16.25
1937	17.61	1950	16.09
1938	17.10	1951	15.92
1939	16.81	1952	16.17
1940	17.29	1953	16.13
1941	17.24	1954	15.96
1942	16.88	1955	15.96
1943	16.95	1956	15.84
1944	17.19	1957	16.10
1945	17.41	1958	16.05
1946	17.16		

Here is a scatterplot of this information from Minitab:



Here, colors have been used to indicate the three categories of Stooges films: Curly films, Shemp films, and Joe films. Do you think there is a linear relationship between these two variables? Are the variables positively associated, negatively associated, or neither?

Measuring Linear Association: Correlation

Definition. The **correlation** measures the strength and direction of the linear relationship between two quantitative variables. Correlation is usually written as r. Suppose that we have data on variables x and y for n individuals. The values for the first individual are x_1 and y_1, the values for the second individual are x_2 and y_2, and so on. The means and standard deviations of the two variables are $\overline{x}$ and s_x for the $x-$values, and $\overline{y}$ and s_y for the $y-$values. The correlation r between x and y is

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{s_x} \right) \left(\frac{y_i - \overline{y}}{s_y} \right).$$

Facts about Correlation

Note. Here is what you need to know in order to interpret correlation r.

1. Correlation makes no distinction between explanatory and response variables. The formula for r is symmetric with respect to x and y.

2. Because r uses the standardized values of the observations, r does not change when we change the units of measurement of x, y, or both.

3. Positive r indicates positive association between the variable, and negative r indicates negative association.

4. The correlation r is always a number between -1 and 1. Values of r near 0 indicate a very weak linear relationship. Values of r close to -1 and 1 indicate that the points in a scatterplot lie close to a straight line.

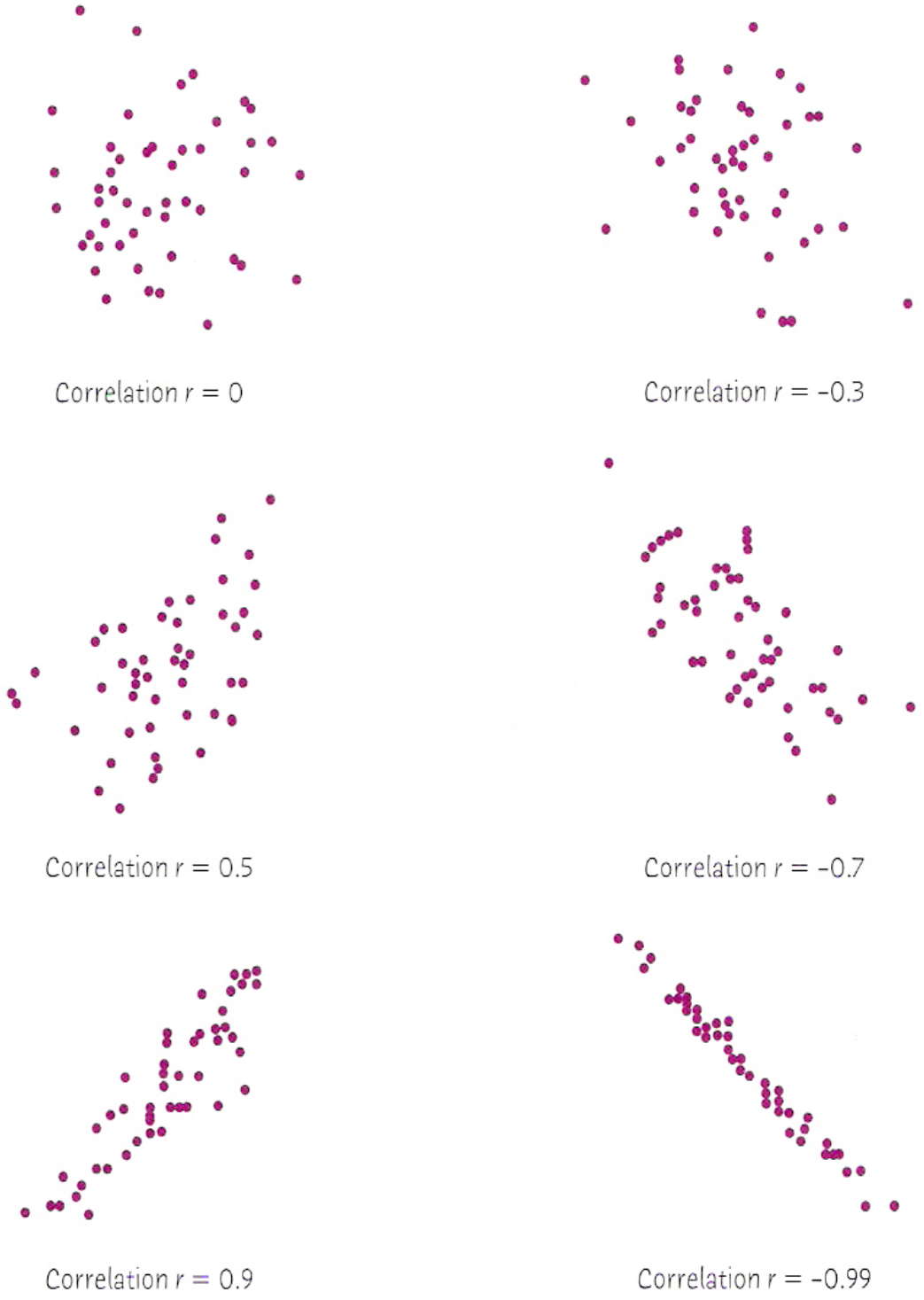


Figure 4.5 page 102.

Note. Correlation measures the strength of the linear relationship between two variables, not other ("curved") relationships. The correlation is not resistant: r is strongly affected by a few outlying observations.

Example. Exercise 4.21 page 106. The data is:

Knee height x	57.7	47.4	43.5	44.8	55.2
Height y	192.1	153.3	146.4	162.7	169.1

Solution. From the formulas of Chapter 2 we have:

$$\overline{x} = 49.72 \quad s_x = 6.36$$
$$\overline{y} = 164.72 \quad s_y = 17.60$$

We then have:

x_i	$(x_i - \overline{x})/s_x$	y_i	$(y_i - \overline{y})/s_y$	$(x_i - \overline{x})/s_x \times (y_i - \overline{y})/s_y$
57.7	1.2547	192.1	1.5557	1.9519
47.4	-0.3648	153.3	-0.6489	0.2367
43.5	-0.9780	146.4	-1.0409	1.0180
44.8	-0.7736	162.7	-0.1148	0.0888
55.2	0.8616	169.1	0.2489	0.2145
				sum = 3.5099

Therefore

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{s_x}\right)\left(\frac{y_i - \overline{y}}{s_y}\right) = \frac{1}{4}(3.5099) = 0.8775.$$

Minitab reports $r = 0.877$. This indicates a fairly strong linear correlation and a positive association.

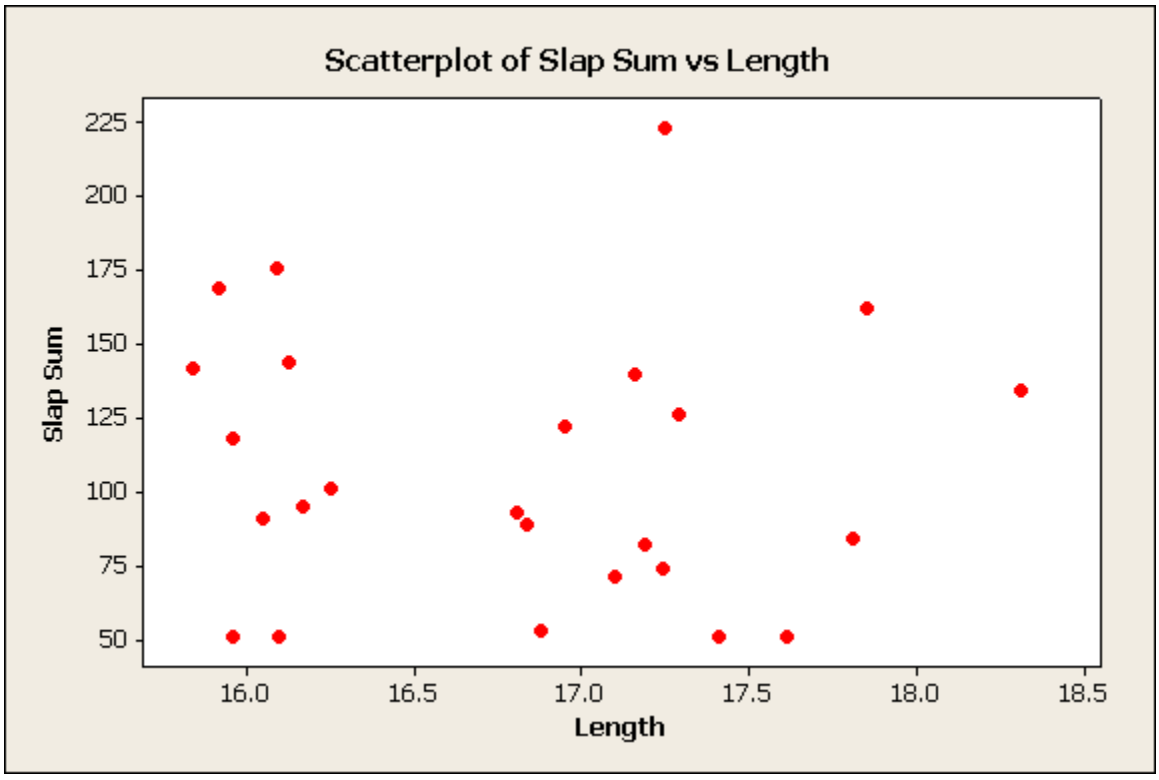
Example S.4.2. Stooge Correlation.
Use Minitab to calculate the correlation r for the length of film data in Example S.4.1.

Solution. First, load the year and length data into two separate columns of a worksheet. Under the `Stat` menu click on `Basic Statistics` and `Correlation`. Select the two columns containing the data from the menu and click `OK`. Minitab will display `Pearson correlation of Year and Length =` -0.904. This means that the correlation is strong (near 1 in absolute value) and that there is a negative association. That is, as one variable increases ("year" say), the other ("length") decreases.

Note. Often times it is hard to find variables representing "real world data" that have a strong linear relationship. We might expect the number of face slaps per film in the Stooges' work would be larger in the longer films, since there is more time for action in the longer films. The following example explores this idea.

Example S.4.3. Stooges Uncorrelated.
Consider the average number of face slaps per film as given in Example S.1.2 and the length of films as given in Example S.4.1. Use Minitab to find the correlation between these two variables and to create a scatterplot.

Solution. Minitab gives $r = -0.040$ indicating that there really is no linear correlation between the variables. The scatterplot is:



rbg-12-30-2008