

Chapter 2 Looking at Data --- Relationships

Let X and Y are two variables measured on the same individuals. Usually, there are two types of relationships between X and Y based on different purposes:

- To explore simple nature of the relationship. The status of two variables are equivalent.
- To show that one of the variables can explain variation in the other.
 - Changes of X cause changes of Y.
 - X --- explanatory variable (independent variable): explains or causes changes in the response variables.
 - Y --- response variable(dependent variable): measures an outcome of a study.

2.1 Scatterplots

1. Scatterplot for quantitative variables X and Y

- A scatterplot shows the relationship between two quantitative variables measured in the same individual. The values of one variable appear on the horizontal axis, and the values of the other appear on the vertical axis.
- Always plot the explanatory variable, if there is one, on the horizontal axis of a scatterplot.
- If there is no explanatory-response distinctions, either variable can go on the horizontal axis.

Example 1

Over a million American students take the SAT college entrance examination each year. This scatterplot examines the relationship between the mean SAT mathematics score in each state and the percent of that state's high school seniors who take the SAT.

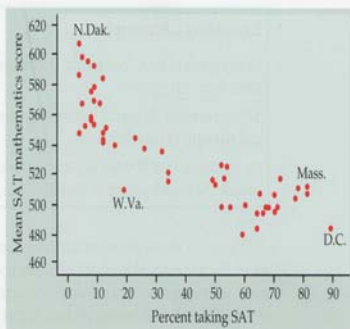


FIGURE 21 State mean SAT mathematics scores plotted against the percent of high school seniors in each state who take the SAT exams.

Example 2 (adding categorical var. to scatterplots)

The Census Bureau groups the states into four broad regions, named Midwest (MW), Northeast (NE), South (S) and West (W). This plot repeats part of the 1st one, but only restricted to the Northeast and Midwest groups of states, with symbol 'e' for the northeastern states, and 'm' for midwestern states.

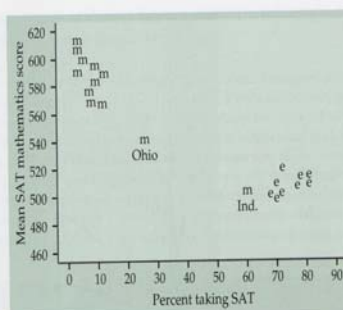


FIGURE 22 State mean SAT mathematics scores and percent taking the SAT for the northeastern states (plot symbol "e") and the midwestern states (plot symbol "m").

2. Interpreting scatterplots

To examine a scatterplot, we look at it from two aspects:

•Overall pattern

•Form

- ✓ Any distinct clusters? How many?
- ✓ (roughly) linear relationship?

•Direction

- ✓ Positive association: X increasing causes Y increasing
- ✓ Negative association: X increasing causes Y decreasing

•Strength

- ✓ How closely the points follow a clear form (straight line)? Strong, moderate, or weak

•Striking deviation

- Outliers: points fall outside the overall pattern of the relationship.

3. Categorical explanatory variables

- Better to use side-by-side boxplots instead of scatterplots especially for large data.

Lecture 4

3

Example 3

Archaeopteryx is an extinct beast having feathers like a bird but teeth and a long bony tail like a reptile. Only six fossil specimens are known. This plot displays data on the lengths in centimeters of the femur (a leg bone) and the humerus (a bone in the upper arm) for the five specimens that preserve both bones. (no explanatory-response distinction)

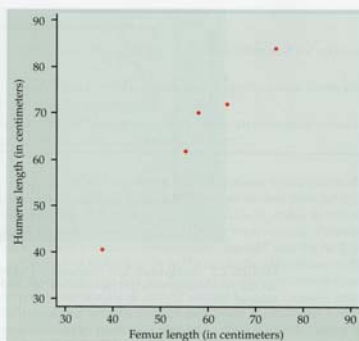


FIGURE 2.3 The lengths of two bones in the five surviving fossil specimens of the extinct beast *Archaeopteryx*.

Example 4

How much corn per acre should a farmer plant to obtain the highest yield? Table 2.1 shows the results of several years of field experiments in Nebraska. Each entry is the mean yield of four small plots planted at the same rate per acre. All plots were irrigated, fertilized, and cultivated identically. The experiment lasted several years.

TABLE 2.1 Corn yields (bushels per acre) in an agricultural experiment

Plants per acre	1956	1958	1959	1960	Mean
12,000	150.1	113.0	118.4	142.6	131.0
16,000	166.9	120.7	135.2	149.8	143.2
20,000	165.3	130.1	139.6	149.9	146.2
24,000		134.7	138.4	156.1	143.1
28,000			119.0	150.5	134.8
Mean	160.8	124.6	130.1	149.8	

Lecture 4

4

Example 4 (continued)

The below plot displays the results of this experiment. The planting rate (explanatory variable) is plotted as the X variable, and the yield is the response variable. The vertical spread of points over each planting rate shows the year-to-year variation in yield. We show the overall pattern by plotting the mean yield for each planting rate (averaged over all years).

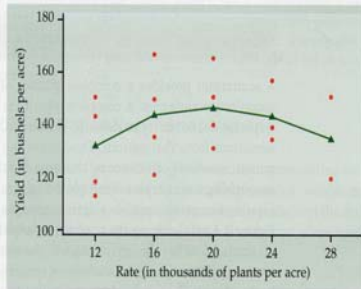


FIGURE 2.4 Yield of corn plotted against planting rate, from an agricultural experiment.

Example 5 (categorical explanatory variables)

We suspect that there is a positive association between how much education a person has and his or her income. This plot summarizes the income information of 55,899 people (between the age of 25 and 65 years old) in March 2000 according to six education categories: 1 = less than high school; 2 = some high school; ...; 6 = postgraduate degree. We arrange the boxplots in order of increasing education.

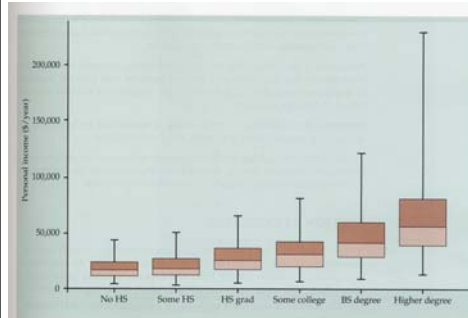


FIGURE 2.6 Boxplots for the incomes of people aged 25 to 65, by highest education level attained. The lines reach from the 5th percentile of incomes to the 95th percentile.

Lecture 4

5

2.2 Correlation

1. Definition

- Correlation r numerically measures the direction and strength of the linear association between two quantitative variables X and Y .

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

Recall:

$$s_x = \sqrt{\frac{1}{n-1} \sum (x_i - \bar{x})^2} = \sqrt{\frac{1}{n-1} (\sum x_i^2 - n\bar{x}^2)}$$

2. Properties

- r is symmetric about X and Y .
- r measures the strength of only the linear relationship between two quantitative variables X & Y .
- If $r > 0$ then positive association between X and Y
- $r < 0$ then negative association between X and Y
- $-1 \leq r \leq 1$ or $0 \leq |r| \leq 1$; when $|r|$ goes from 0 to 1, the strength of linear association goes from weak to strong; $|r| = 1$ means the points in a scatterplot lie exactly along a straight line.

Lecture 4

6

A plot to show the limit of graph method by changing the plotting scales

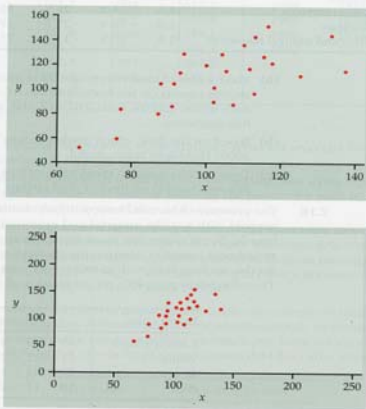


FIGURE 2.9 Two scatterplots of the same data; the linear pattern in the lower plot appears stronger because of the surrounding white space.

A plot to show the relationship of r and the direction&strength of linear association

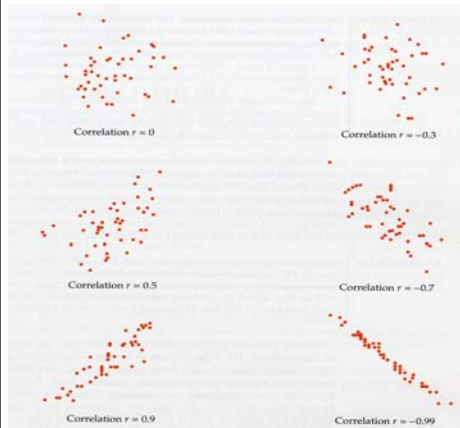


FIGURE 2.10 How the correlation r measures the direction and strength of linear association.

Note:

- correlation r is not a complete description of two variable data.
- r^2 is called coefficient of determination. $0 \leq r^2 \leq 1$.