CHAPTER 4: Scatterplots and Correlation

Basic Practice of Statistics

7th Edition

Lecture PowerPoint Slides

In Chapter 4, we cover …

- Explanatory and response variables
- Displaying relationships: Scatterplots
- Interpreting scatterplots
- Adding categorical variables to scatterplots
- Measuring linear association: Correlation
- Facts about correlation

Response Variables and Explanatory Variables

- Interested in studying the relationship between two variables by measuring both variables on the same individuals.
  - A **response variable** measures an outcome of a study.
  - An **explanatory variable** may explain or influence changes in a response variable.
  - sometimes there is no distinction

Question

In a study to determine whether surgery or chemotherapy results in higher survival rates for a certain type of cancer, whether or not the patient survived is one variable, and whether they received surgery or chemotherapy is the other. Which is the explanatory variable and which is the response variable?

Explanatory and Response

A study examines whether state political repression increases the chances for popular revolution. What is the response variable in this study?

a) the state
b) popular revolution
c) state political repression
d) repression

Displaying Relationships

If a data set consists of two variables measured on each of 20 individuals, how many dots are in the scatterplot?

a) 10
b) 20
c) 30
d) 40

Univariate and Bivariate Data

Table 1: Univariate and Bivariate Data

	Univariate Data	Bivariate Data
The number of variables	one	two
Examples	•Mid term scores of 1070 class •heights of 1070 class	•Mid term score and amount of time spent studying for midterm of 1070 class •heights and weights of 1070 class
Useful graph	histogram	scatterplot

Scatterplot (1 of 3)

- The most useful graph for displaying the relationship between two quantitative variables is a **scatterplot**.

- A scatterplot shows the relationship between two quantitative variables that are 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. As a reminder, 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.

Scatterplot (2 of 3)

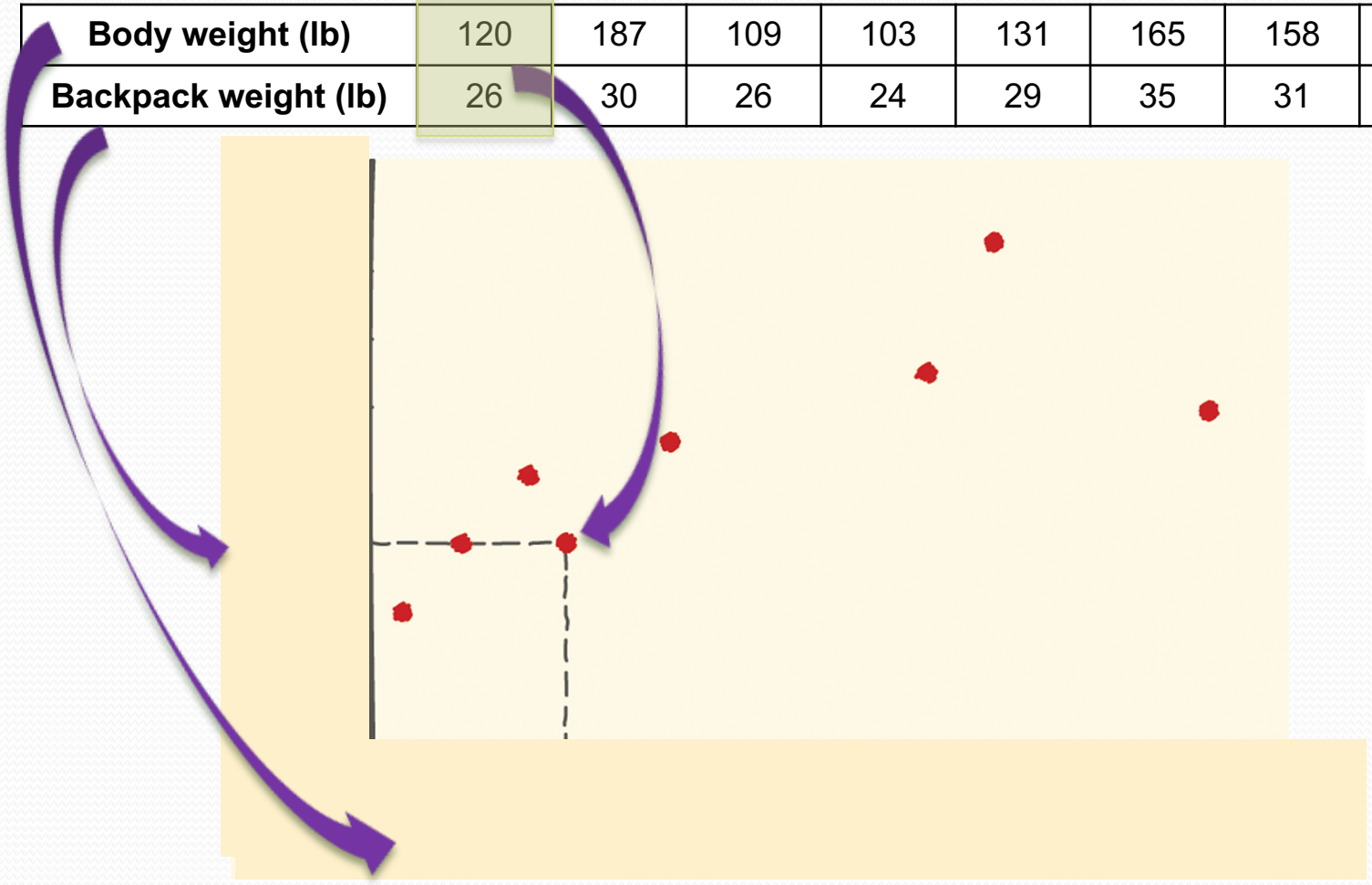
Making the scatterplot in the context of our four-step process:

- STATE: The research question of interest is stated as a statement (or a query about a statement) of the association between two variables in your data.
- PLAN: The solution of your problem is planned by plotting the variables according to the guidelines on the previous slide.
- SOLVE: Examine the scatterplot, taking note of any relationship present.
- CONCLUDE: We will explore this step later …

Scatterplot (3 of 3)

Example: Make a scatterplot of the relationship between body weight and backpack weight for a group of hikers.

Body weight (lb)	120	187	109	103	131	165	158	116
Backpack weight (lb)	26	30	26	24	29	35	31	28



Interpreting Scatterplots

To interpret a scatterplot, follow the basic strategy of data analysis from Chapters 1 and 2. Look for patterns and important departures from those patterns.

EXAMINING A SCATTERPLOT

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 departure is an **outlier—**an individual value that falls outside the overall pattern of the relationship.

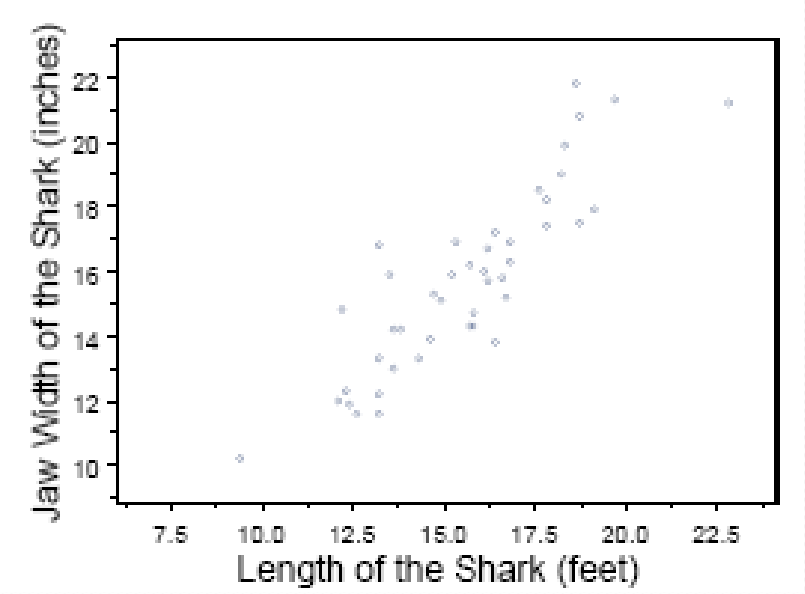
Direction of Association

POSITIVE ASSOCIATION, NEGATIVE ASSOCIATION

- 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 tend to accompany below-average values of the other, and vice versa.

Interpreting Scatterplots

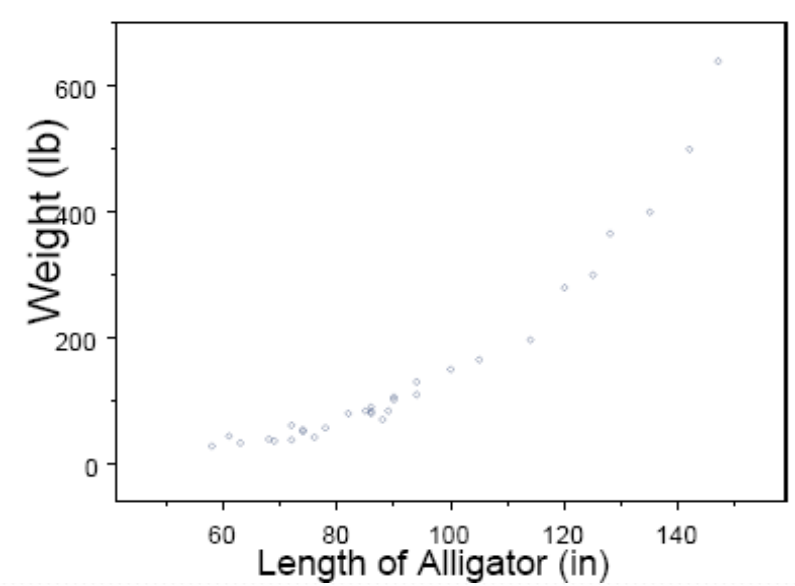
Look at the following scatterplot. Which description best fits the plot?



a) *direction*: positive; *form*: linear; *strength*: strong
b) *direction*: negative; *form*: linear; *strength*: strong
c) *direction*: positive; *form*: non-linear; *strength*: weak
d) *direction*: negative; *form*: non-linear; *strength*: weak
e) no relationship

Interpreting Scatterplots

Look at the following scatterplot. Which description best fits the plot?

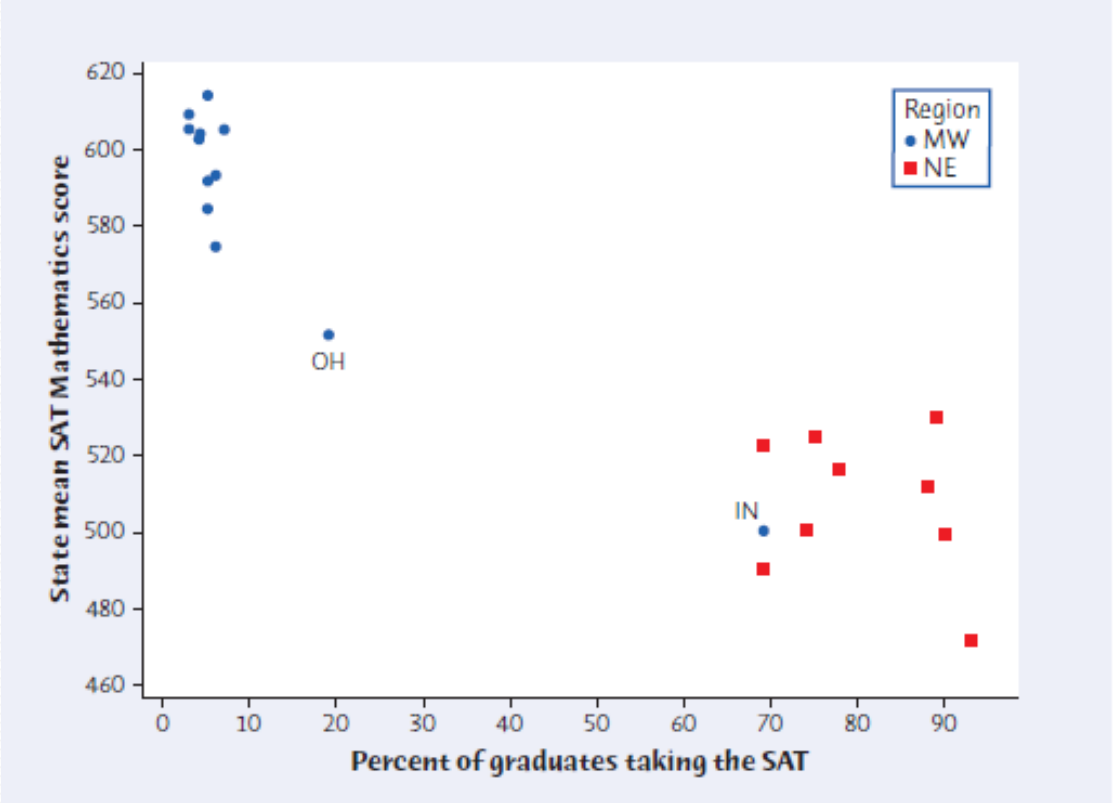


a) *direction*: positive; *form*: non-linear; *strength*: strong
b) *direction*: negative; *form*: linear; *strength*: strong
c) *direction*: positive; *form*: linear; *strength*: weak
d) *direction*: positive; *form*: non-linear; *strength*: weak
e) no relationship

Adding Categorical Variables

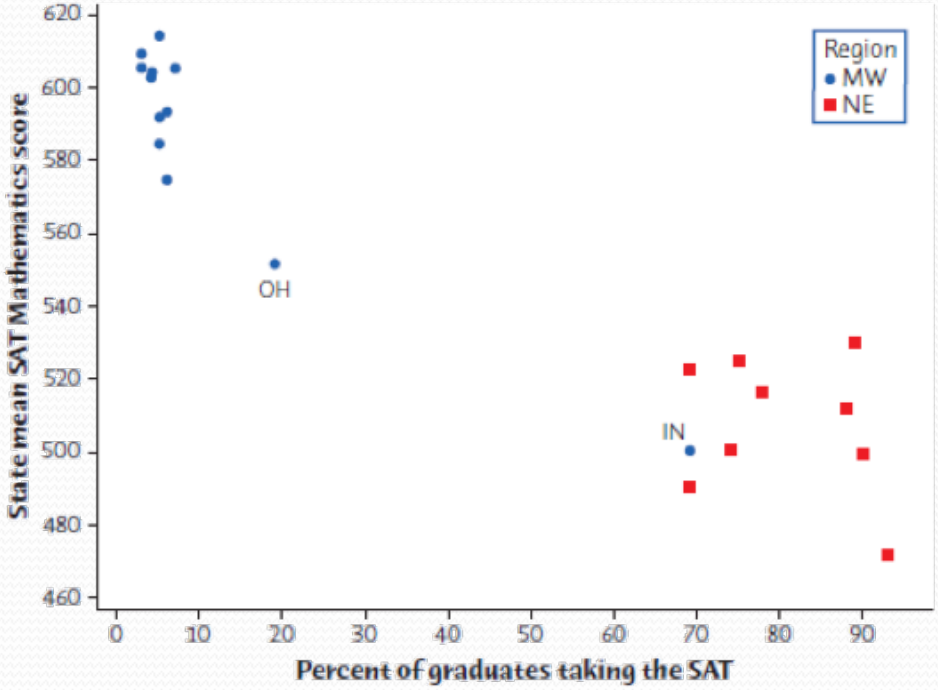
Consider the relationship between mean SAT verbal score and percent of high-school grads taking SAT for each state.

Mean SAT Mathematics score and percent of high school graduates who take the test for only the Midwest (blue) and Northeast (red) states.



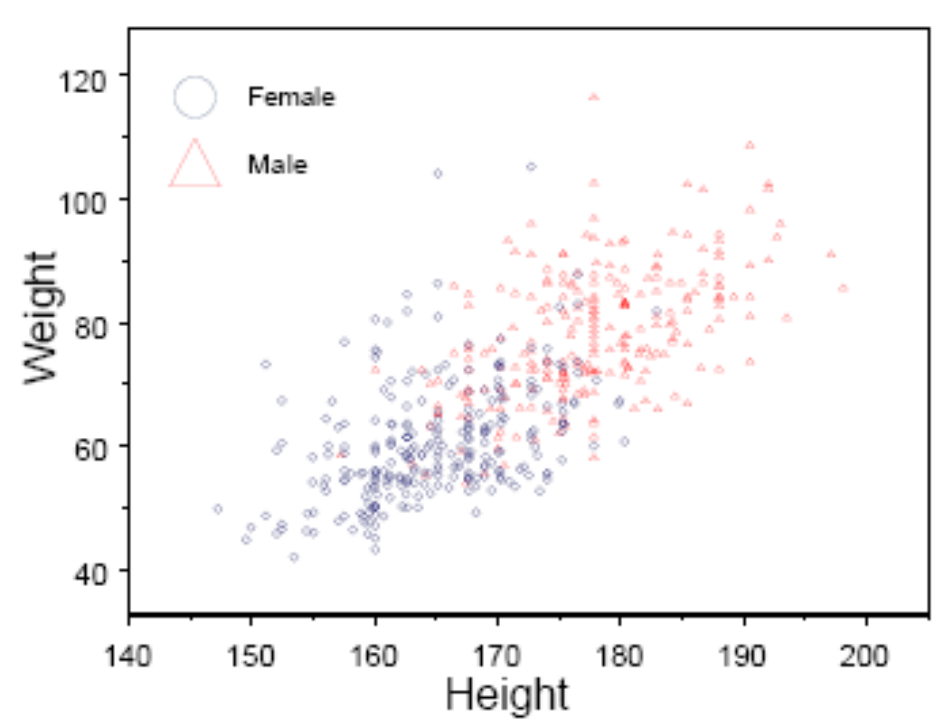
Categorical Variables in Scatterplots

To add a categorical variable, use a different plot color or symbol for each category.



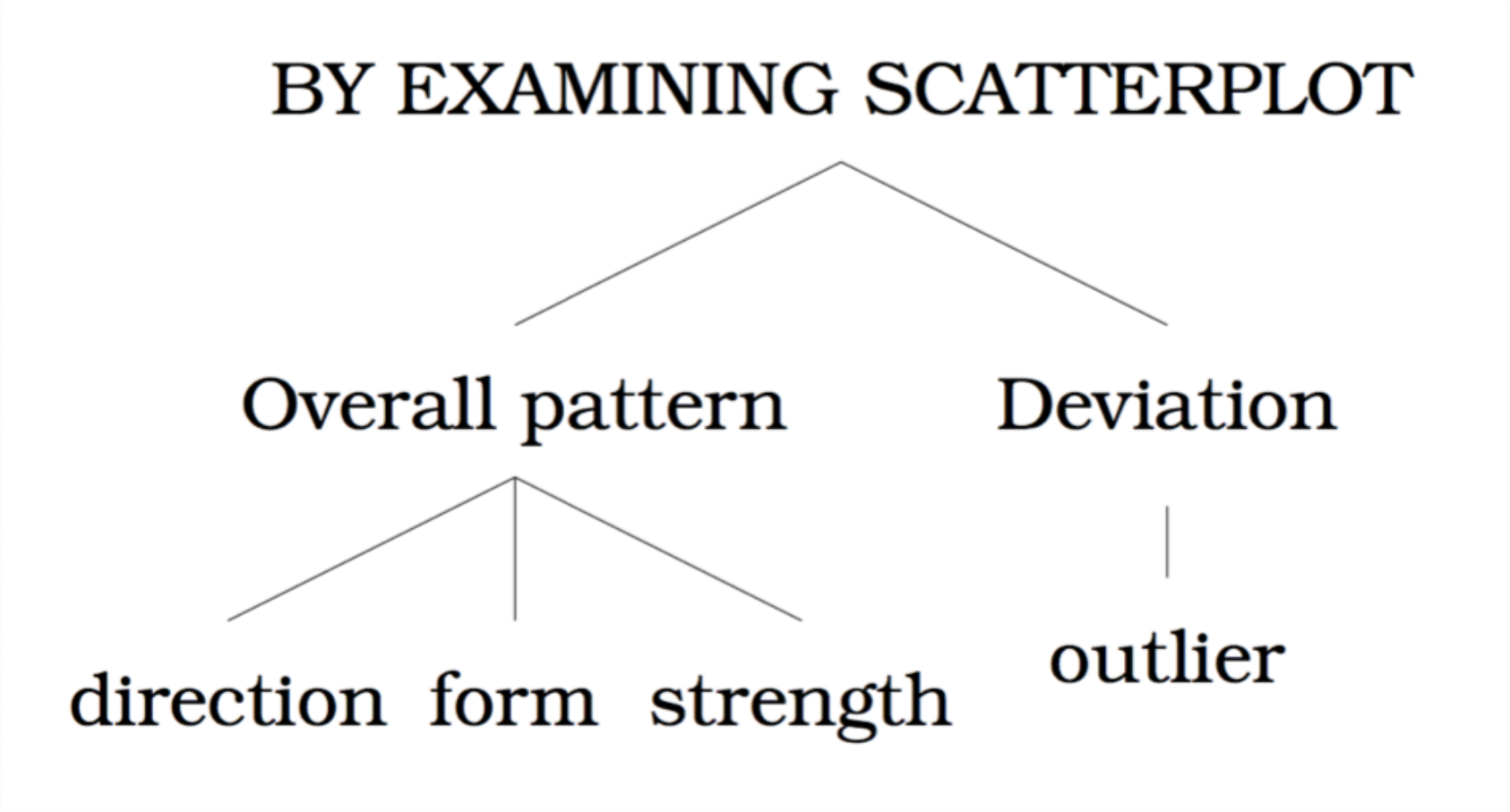
Adding Categorical Variables

Look at the following scatterplot. Which variable is categorical?



a) height
b) weight
c) gender

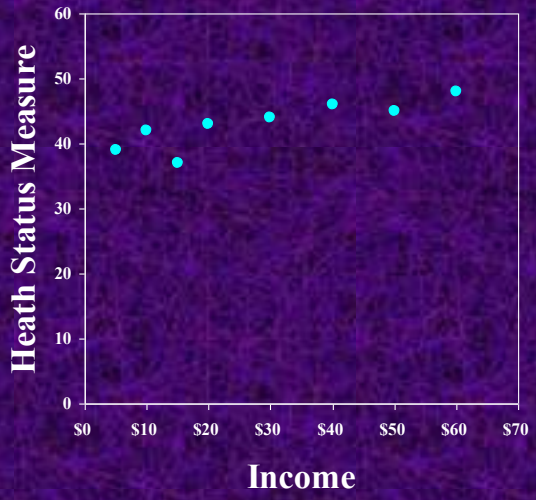
Interpreting scatterplots

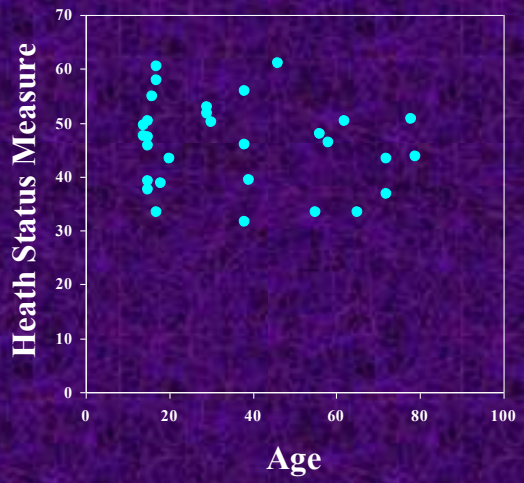


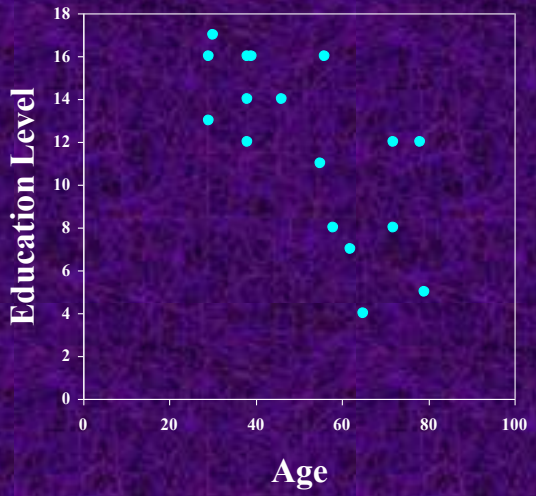
Definitions

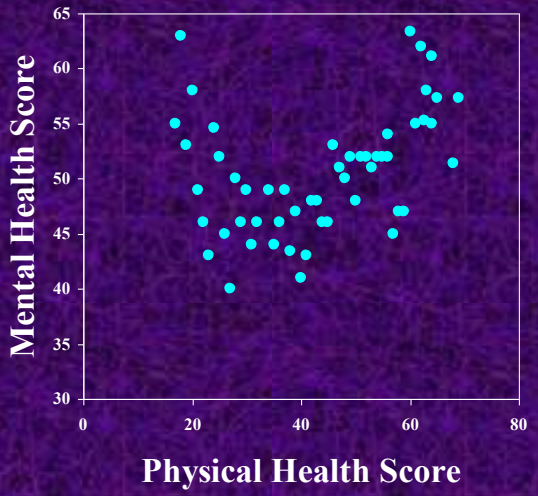
- Direction: If the relationship has a clear direction, we speak of either **positive association**(high values of the two variables tend to occur together) or **negative association**(high values of one variable tend to occur with low values of the other variable).
- Form : **Linear relationships**, where the points show a straight-line pattern, are an important form of relationship between two variables. **Curved relationships** and **clusters** are other forms to watch for.
- Strength: The strength of a relationship is determined by how close the points in the scatterplot lie to a simple form such as a line.

Examples of Relationships









Measuring Linear Association

A scatterplot *displays* the strength, direction, and form of the relationship between two quantitative variables.

- The **correlation**, r, *measures* the ***strength*** of the linear relationship between two quantitative variables. (the stronger the relationship, the larger the magnitude of 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 $\bar{x}$ and s_x for the x-values, and $\bar{y}$ and s_y for the y-values. The correlation r between x and y is

$$r = \frac{1}{n-1}\left[\left(\frac{x_1-\bar{x}}{s_y}\right)\left(\frac{y_1-\bar{y}}{s_y}\right)+\left(\frac{x_2-\bar{x}}{s_x}\right)\left(\frac{y_2-\bar{y}}{s_y}\right)+\cdots+\left(\frac{x_n-\bar{x}}{s_x}\right)\left(\frac{y_n-\bar{y}}{s_y}\right)\right]$$

shorter form:

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

Facts about Correlation (1 of 3)

- Correlation makes no distinction between explanatory variables and response variables.
- r has no units and does not change when we change the units of measurement of x, y, or both.
- Positive r indicates positive association between the variables, and negative r indicates negative association.
- The correlation r is always a number between –1 and 1.

Facts about Correlation (2 of 3)

Cautions:

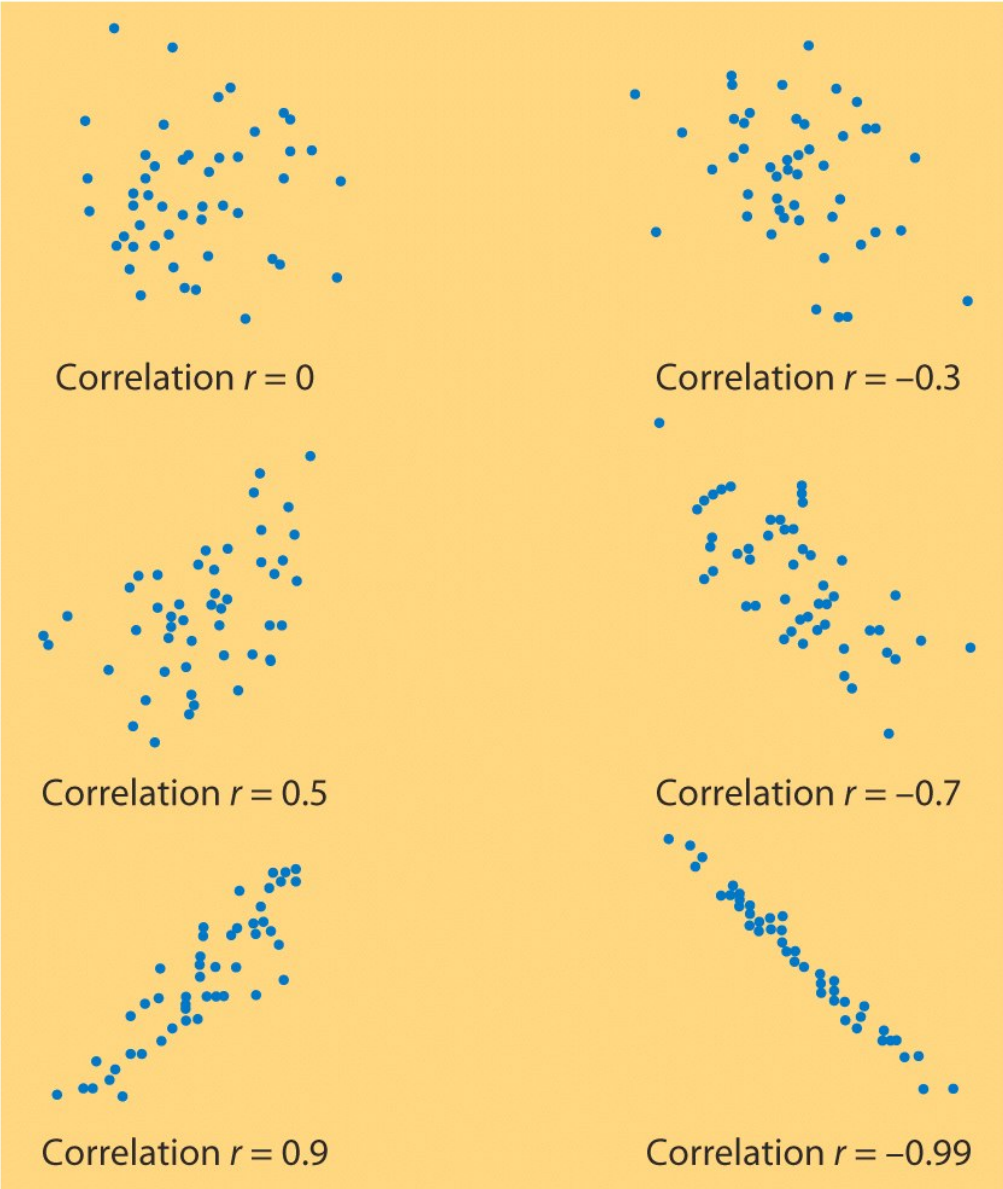
- Correlation requires that both variables be quantitative, so it makes sense to do the arithmetic indicated by the formula for r.
- Correlation does not describe curved relationships between variables, no matter how strong the relationship is between them.
- Correlation is not resistant; *r* is strongly affected by a few outlying observations.
- Correlation is ***not*** a complete summary of two-variable data.

Facts about Correlation (3 of 3)

Special values for r :

- a perfect positive linear relationship would have r = +1
- a perfect negative linear relationship would have r = -1
- if there is no *linear* relationship, or if the scatterplot points are best fit by a horizontal line, then r = 0
- *Note: r must be between -1 and +1, inclusive*

Examples of Correlations

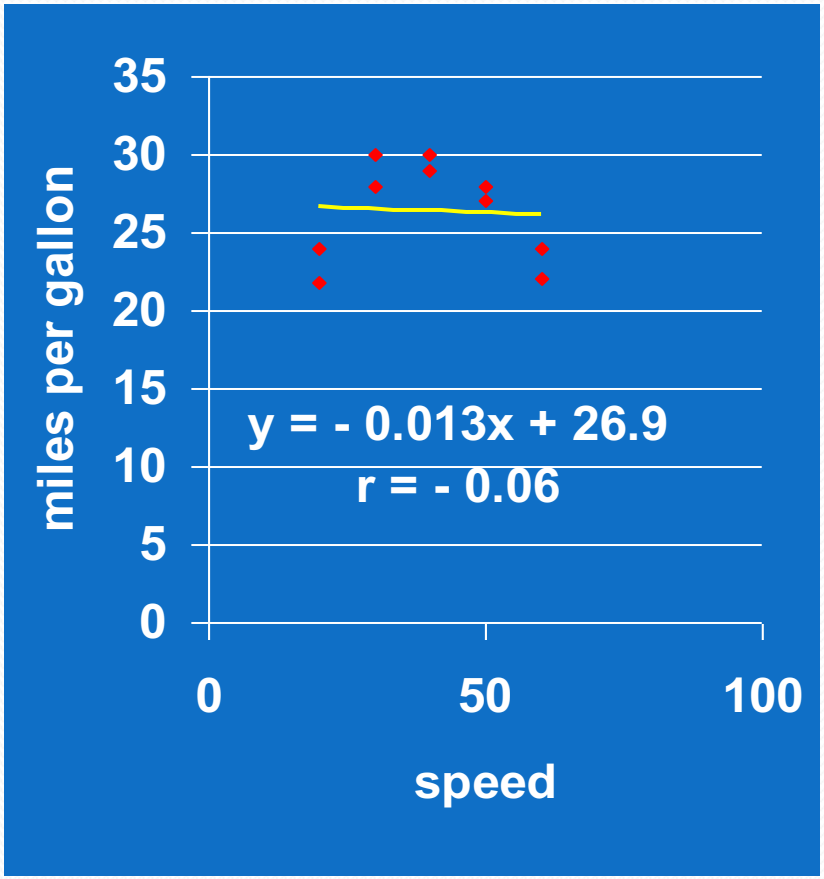


Examples of Correlations

- Husband's versus Wife's ages
  - r = .94
- Husband's versus Wife's heights
  - r = .36
- Professional Golfer's Putting Success: Distance of putt in feet versus percent success
  - r = -.94

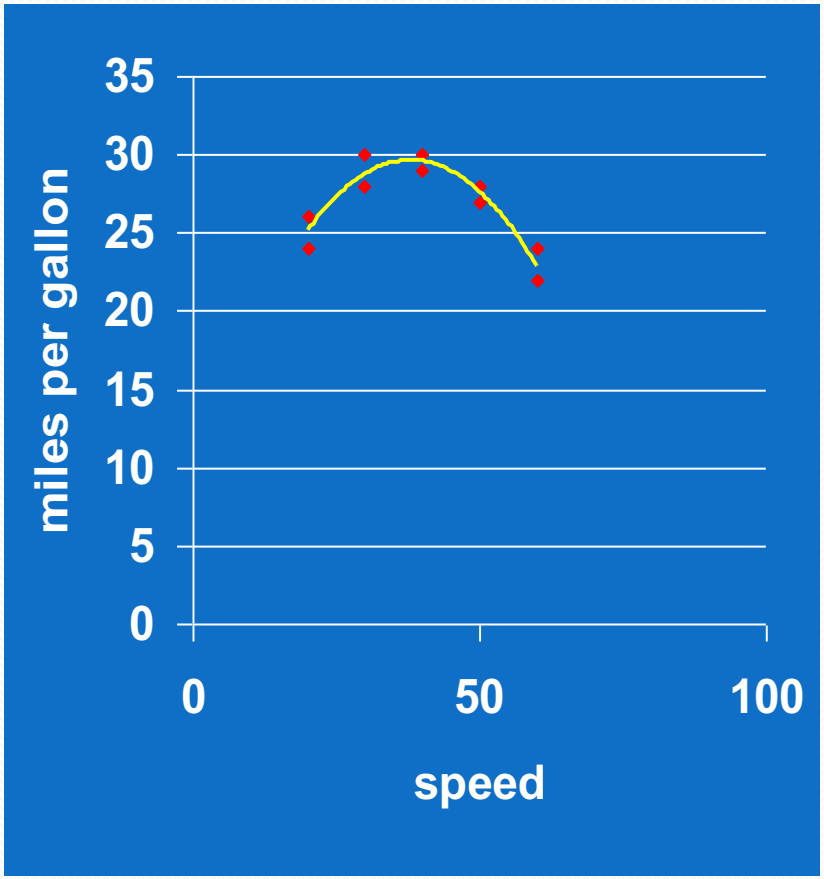
Not all Relationships are Linear
Miles per Gallon versus Speed

- Linear relationship?
- Correlation is close to zero.



Not all Relationships are Linear
Miles per Gallon versus Speed

- Curved relationship.
- Correlation is misleading.

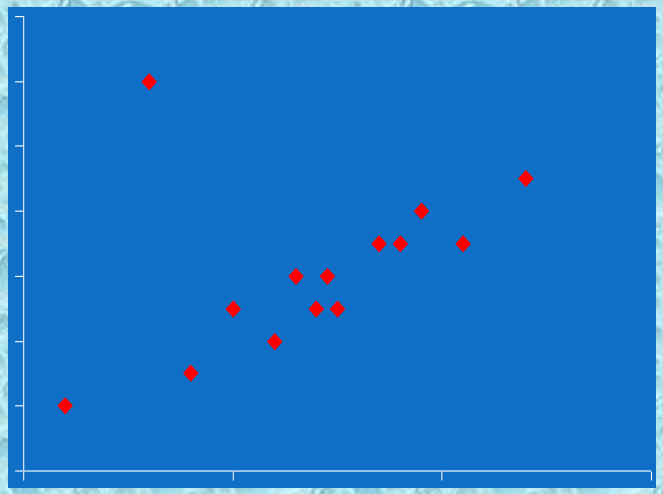


Problems with Correlations

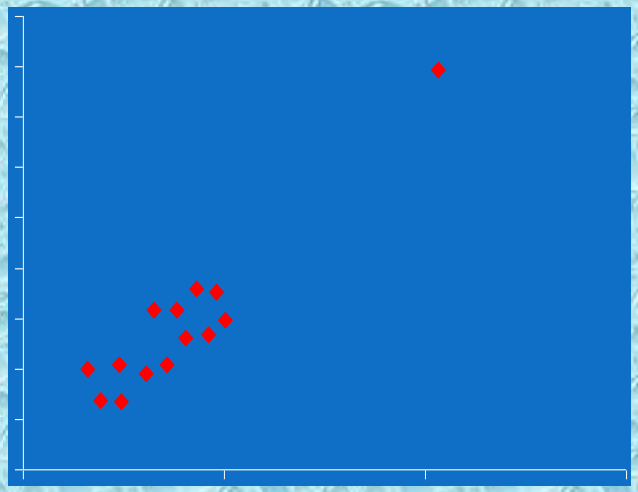
- Outliers can inflate or deflate correlations (see next slide)
- Groups combined inappropriately may mask relationships (a third variable)
  - groups may have different relationships when separated

Outliers and Correlation

A



B



For each scatterplot above, how does the outlier affect the correlation?

A: outlier decreases the correlation

B: outlier increases the correlation

Correlation Calculation

- Suppose we have data on variables X and Y for n individuals:

  $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_n$

- Each variable has a mean and std dev:

  $(\bar{x}, s_x)$ and $(\bar{y}, s_y)$ (see ch. 2 for s)

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

Example

student's name	Choo	Kang	Lee	Park	Ryu
Time studied(hours)	2	10	6	8	5
Score	60	95	78	88	72

- Draw scatterplot

- Calculate r

Correlation (1 of 6)

For which of the following situations would it be appropriate to calculate *r*, the correlation coefficient?

a) the time spent studying for a statistics exam and the score on the exam
b) income for county employees and their respective counties
c) eye color and hair color of selected participants
d) the party affiliation of senators and their votes on presidential impeachment

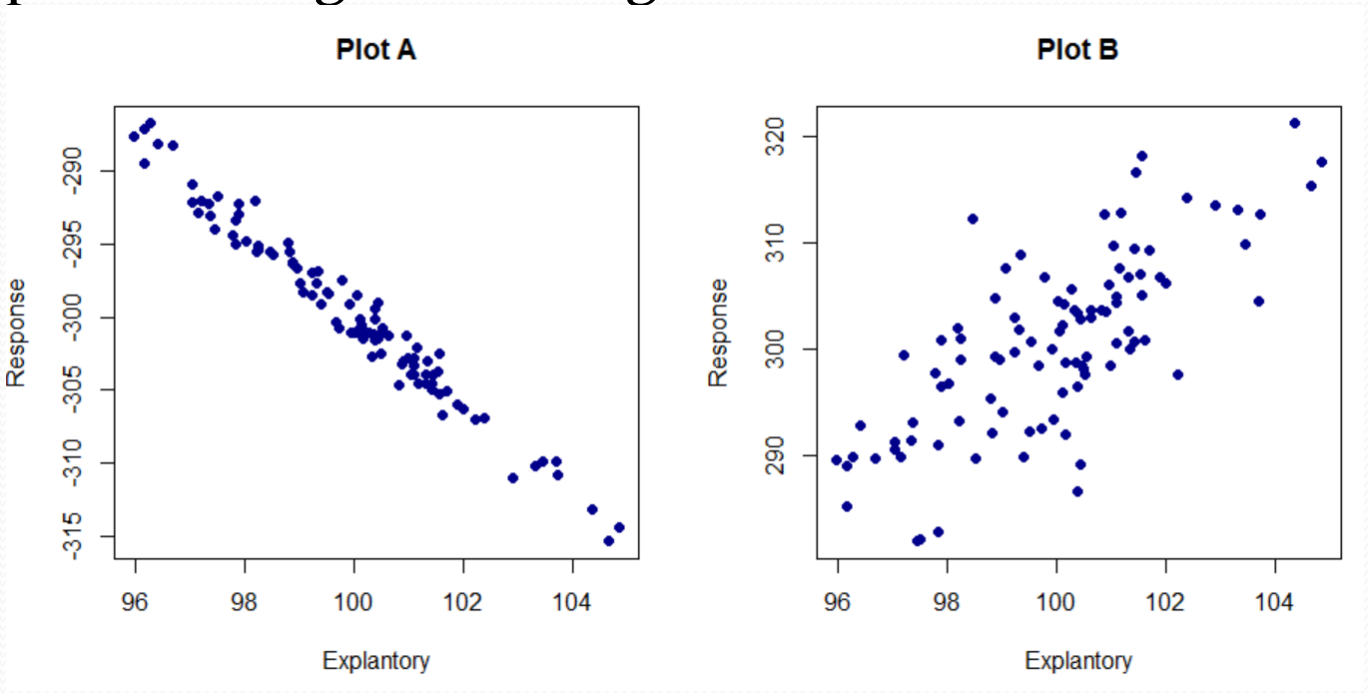
Correlation (2 of 6)

Which is a false statement about *r*, the correlation coefficient?

a) It can range in value from –1 to 1.
b) It measures the strength and direction of the linear relationship between X and Y.
c) It is measured in units of the X variable.

Correlation (3 of 6)

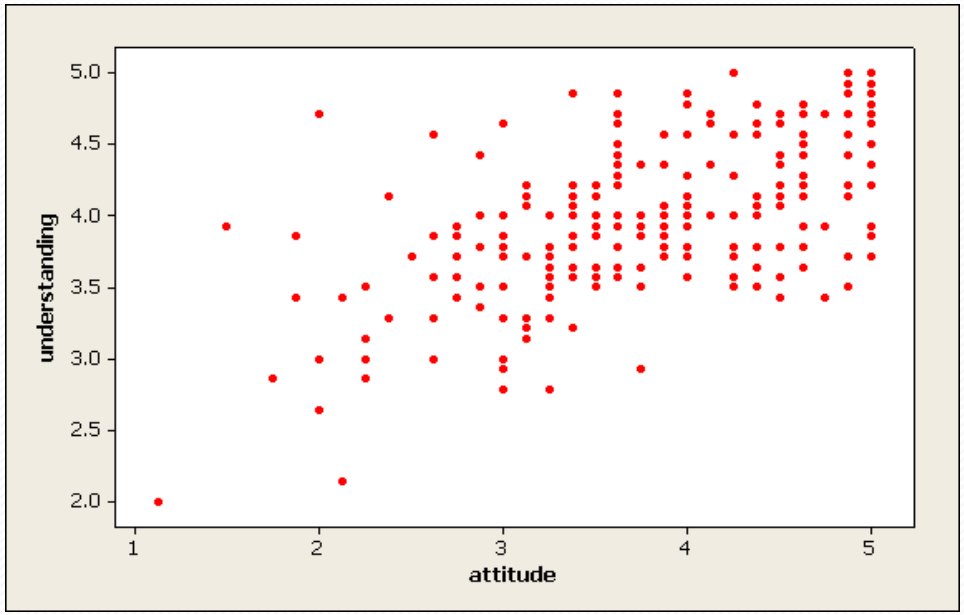
Which scatterplot would give a stronger value for *r*?



a) Plot A
b) Plot B
c) It would be the same for both plots.

Correlation (4 of 6)

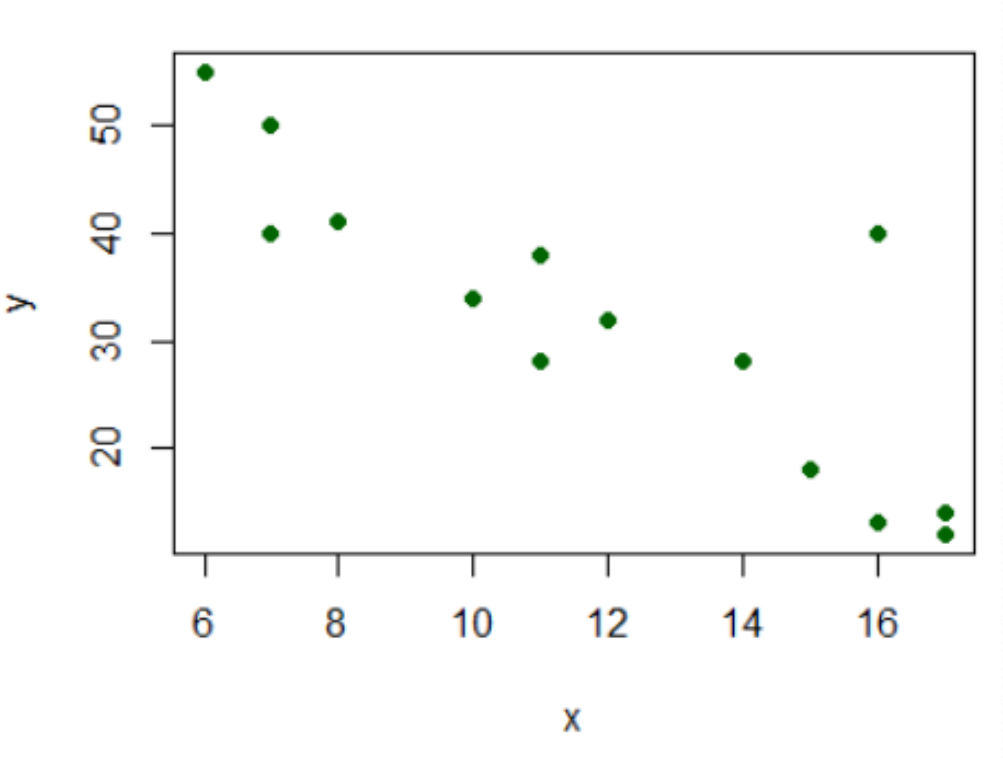
Which of the following is the correlation of understanding (of biology) with attitude for the following scatterplot?



a) −0.75
b) −0.04
c) 0.63
d) 0.98

Correlation (5 of 6)

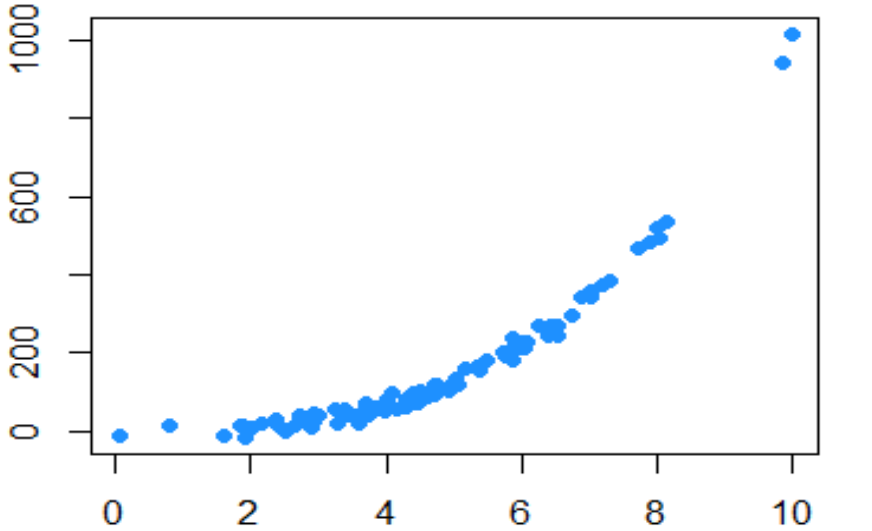
Look at the following scatterplot. Choose the number that best estimates the correlation.



a) −0.87
b) −0.15
c) 0.63
d) 0.98

Correlation (6 of 6)

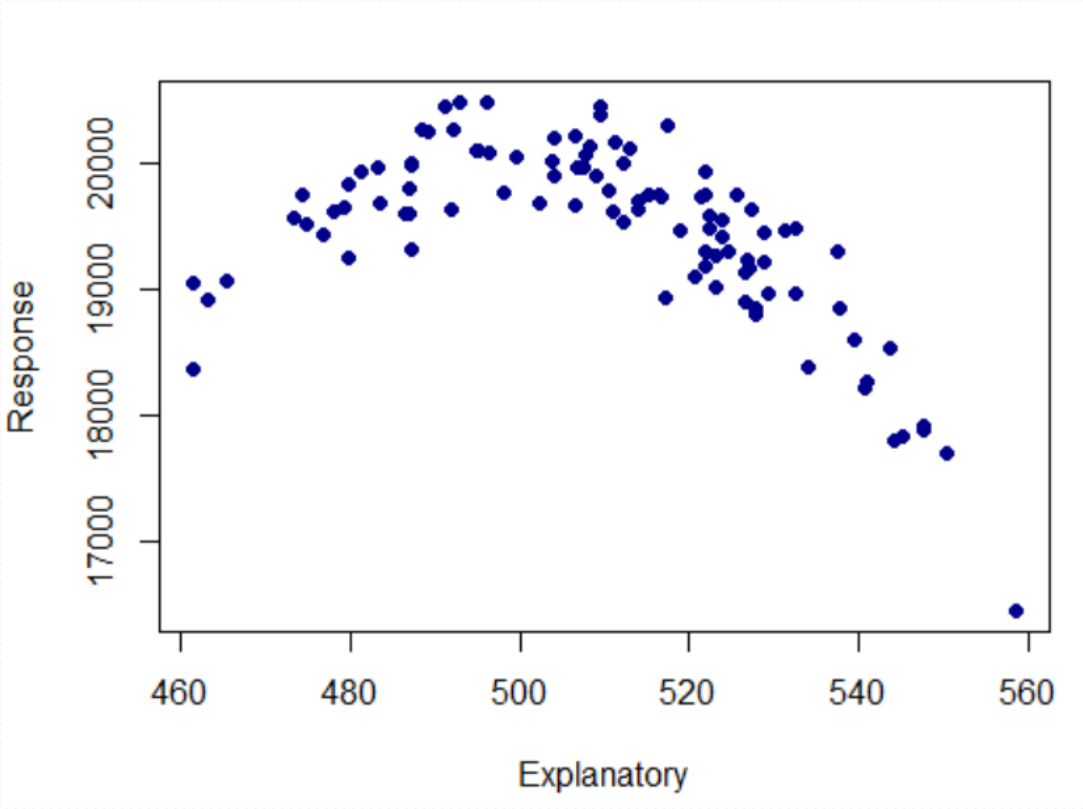
Look at the following scatterplot. Choose the number that best estimates the correlation.



a) 0.99
b) 0.75
c) −0.75
d) None is appropriate given the form of the relationship.

Facts about Correlation (1 of 4)

Computing *r* as a measure of the strength of the relationship between *X* and *Y* is appropriate for the data in the following scatterplot.



a) true
b) false

Facts about Correlation (2 of 4)

The correlation coefficient will not be affected by an outlier in the data.

a) true
b) false

Facts about Correlation (3 of 4)

If we change the height measure (X) from inches into centimeters and the weight measure (Y) from pounds into kilograms, what will happen to the correlation coefficient?

a) Correlation coefficient will increase.
b) Correlation coefficient will decrease.
c) Correlation coefficient will remain the same.

Facts about Correlation (4 of 4)

Given that the correlation between MPG (miles per gallon) and a car's weight is –0.85, the correlation between a car's weight and MPG should become +0.85 when reversed.

a) true
b) false