Statistical Science 1024

Chapter 4

Scatterplots and Correlation

RESPONSE VARIABLE, EXPLANATORY VARIABLE

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

SCATTERPLOT

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. 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.

Purpose of Scatterplots

- to see how two variables are related to one another
- to see if one variable can explain the behaviour of the other variable

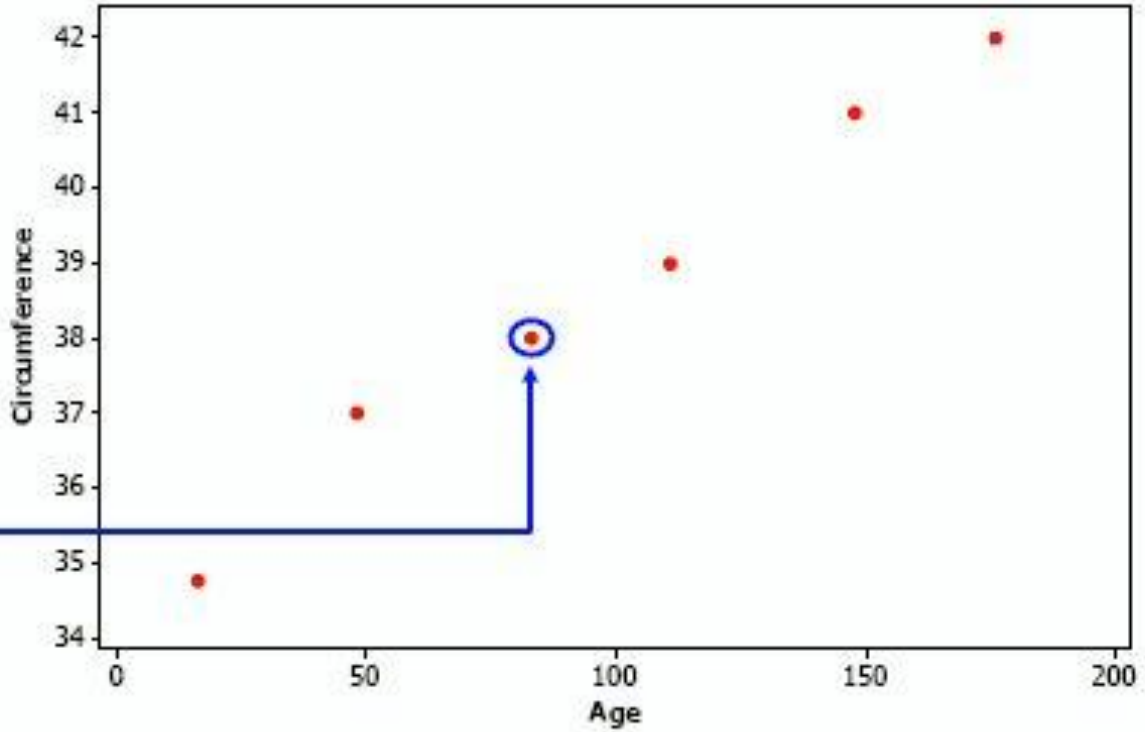
Erika's Head Measurements

When she was a baby, each time my eldest daughter visited our doctor, he took a number of measurements such as height (or length), weight and head circumference. The latter measurement is made to see if the head, or more particularly cranial capacity, is growing quickly enough.

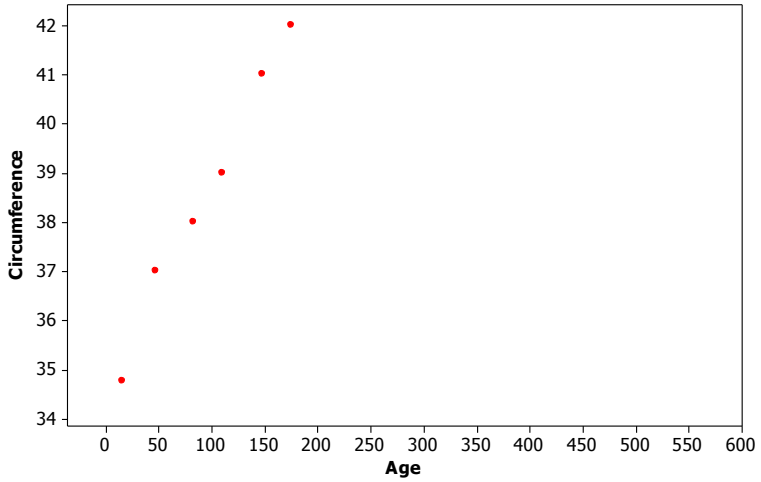
Age in days	Circumference in cm
16	34.75
48	37
83	38
111	39
148	41
176	42

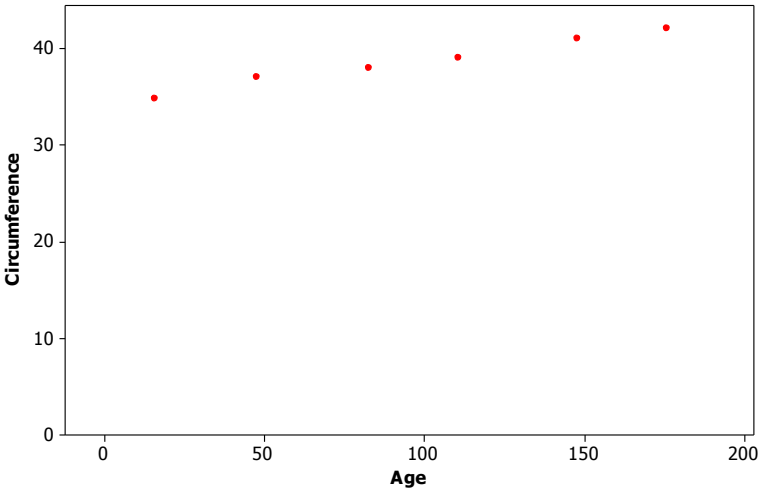
Scatterplot Construction

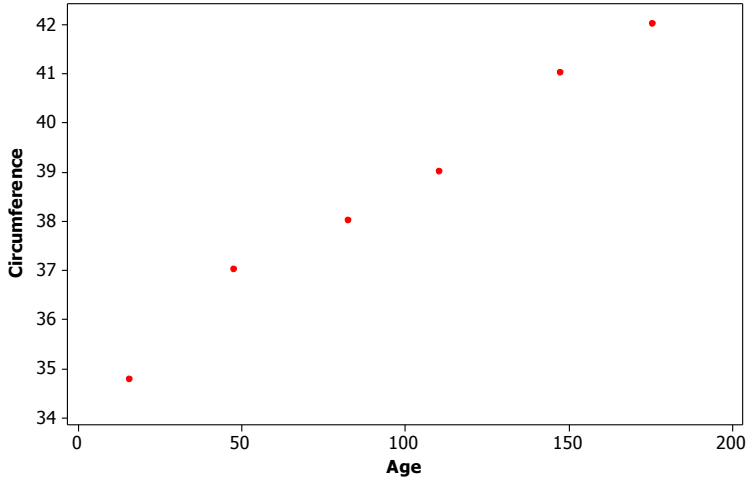
Age in Days	Circumference in cm
16	34¾
48	37
83	38
111	39
148	41
176	42

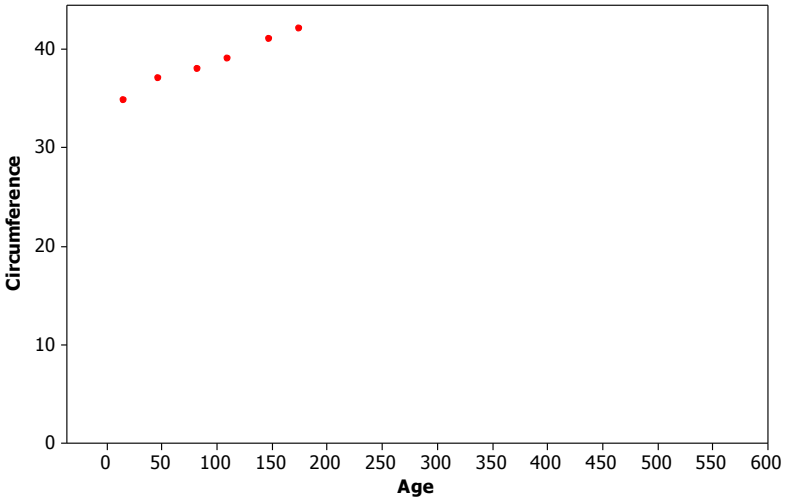


Scatterplot Scale





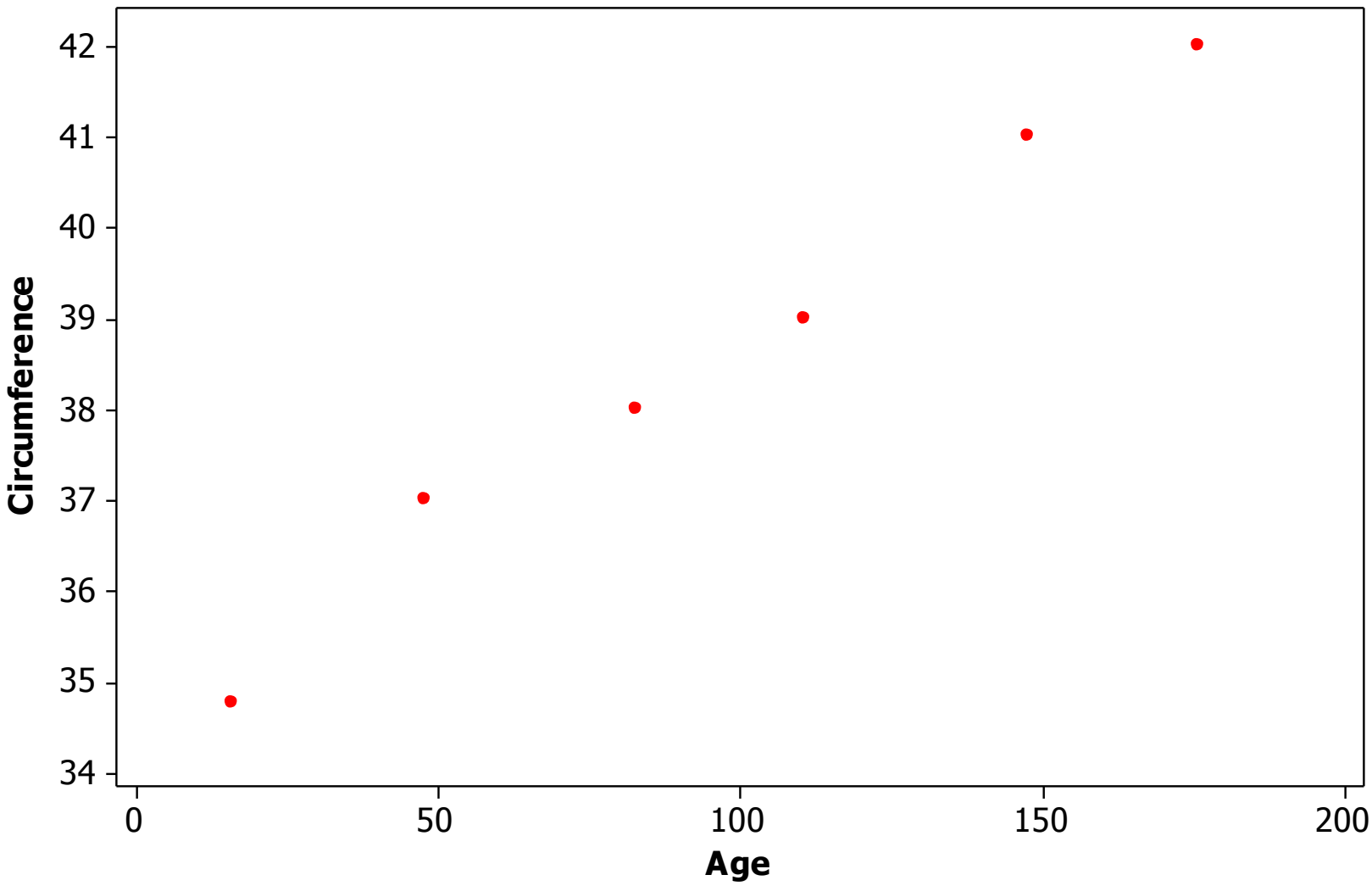




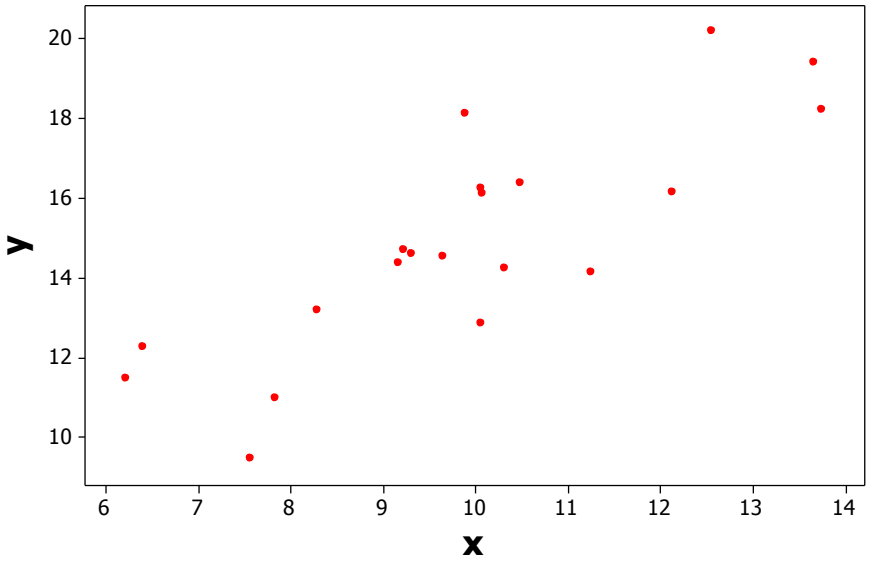
Both variables should be given a similar amount of space:

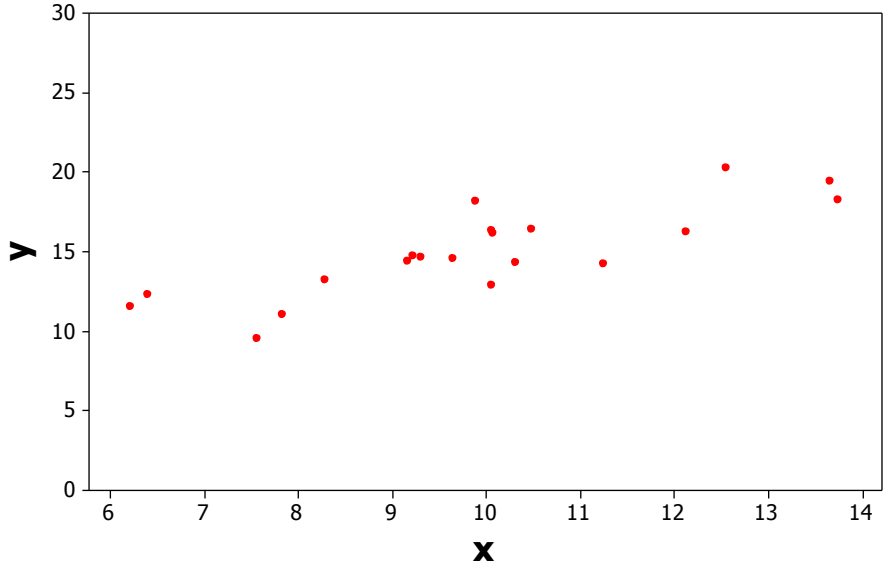
- plot roughly a square shape (the above are approximately golden rectangles)
- points should occupy all the plot space (no blank space)

Scatterplot of Head Measurements



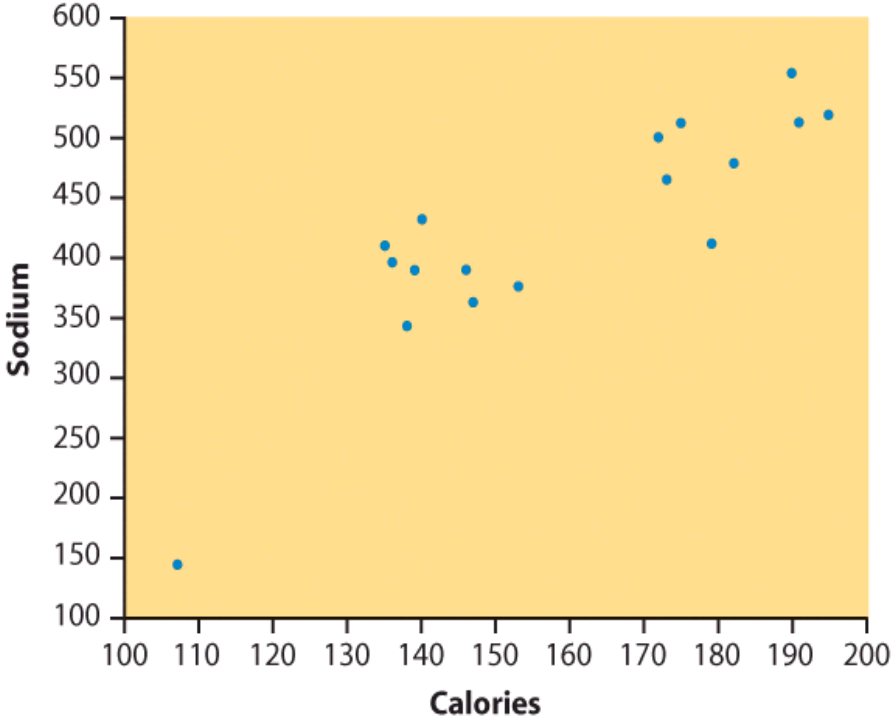
Other Effects of Scale
Example – simulated data



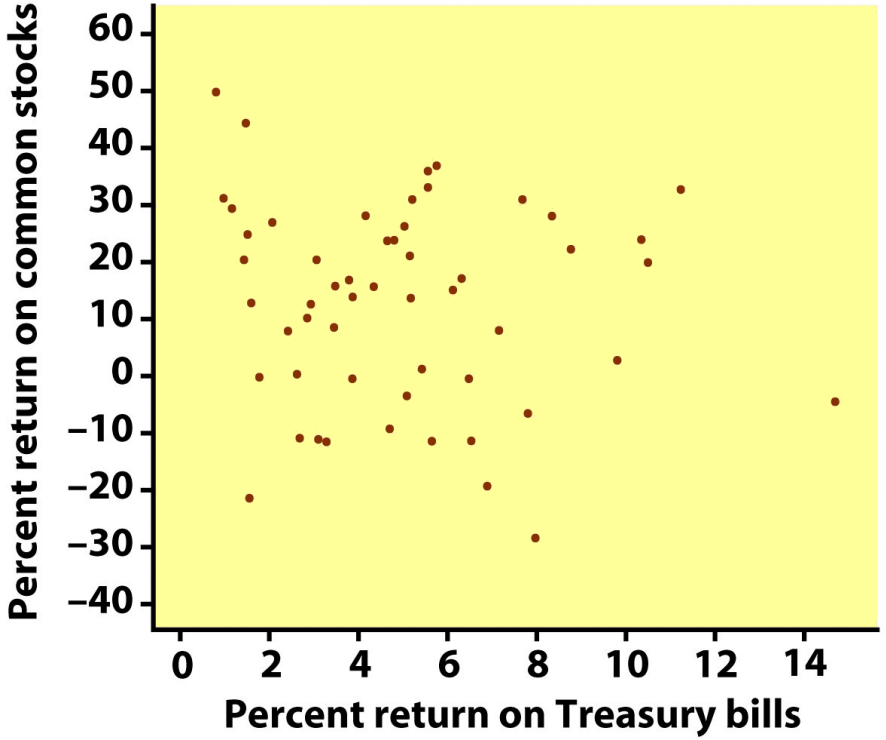


Same data but the relationship appears stronger in the graph on the right

Some plots don't have clear explanatory and response variables.



Do calories explain sodium amounts?



Does percent return on Treasury bills explain percent return on common stocks?

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 deviation is an **outlier,** an individual value that falls outside the overall pattern of the relationship.

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.

Types of Association

Positive association

- larger values of one variable are associated with larger values of the other, likewise for smaller values tend to occur together
- as one variable increases in value the other has the same tendency

Negative association

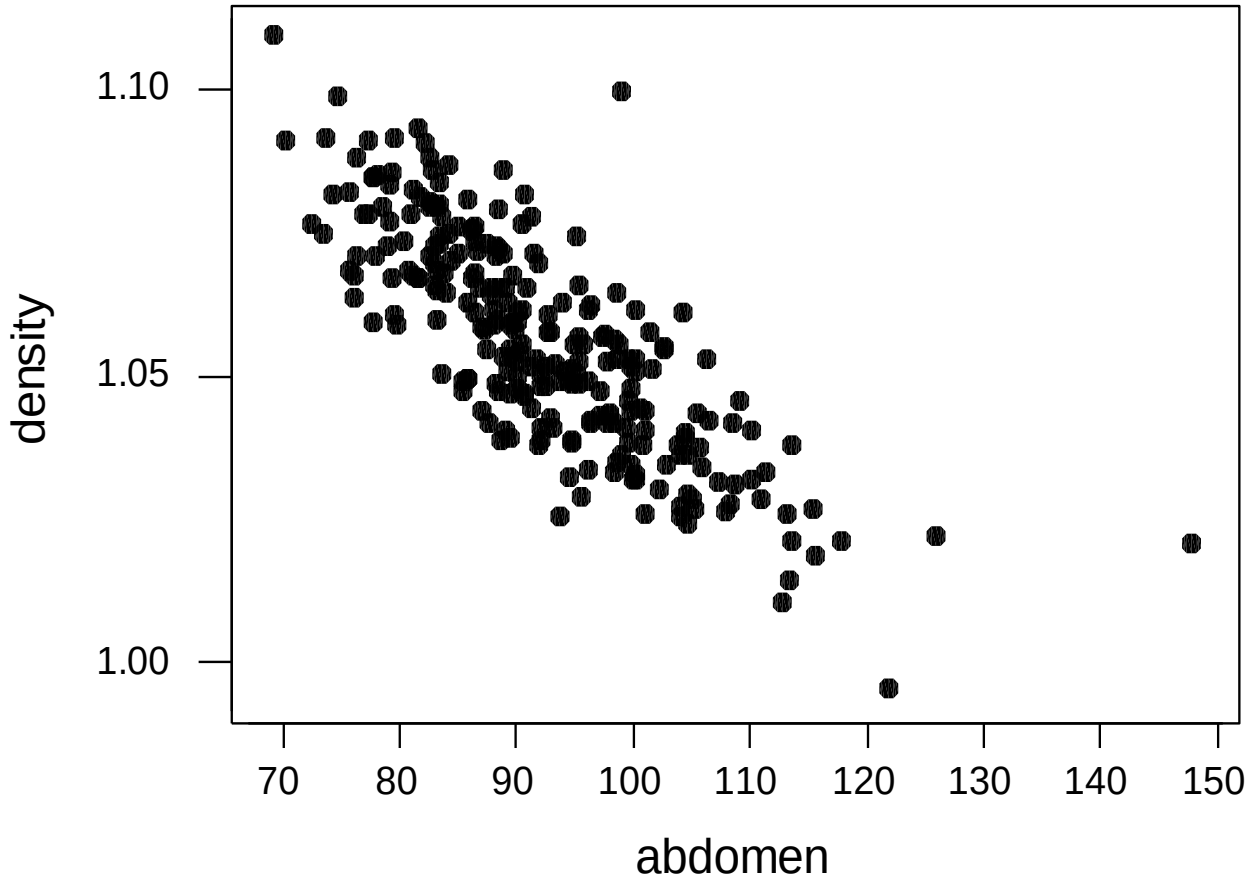
- larger values of one variable are associated with smaller values of the other
- as one variable increases in value the other tends to decrease in value

Example – Body Density and Body Mass

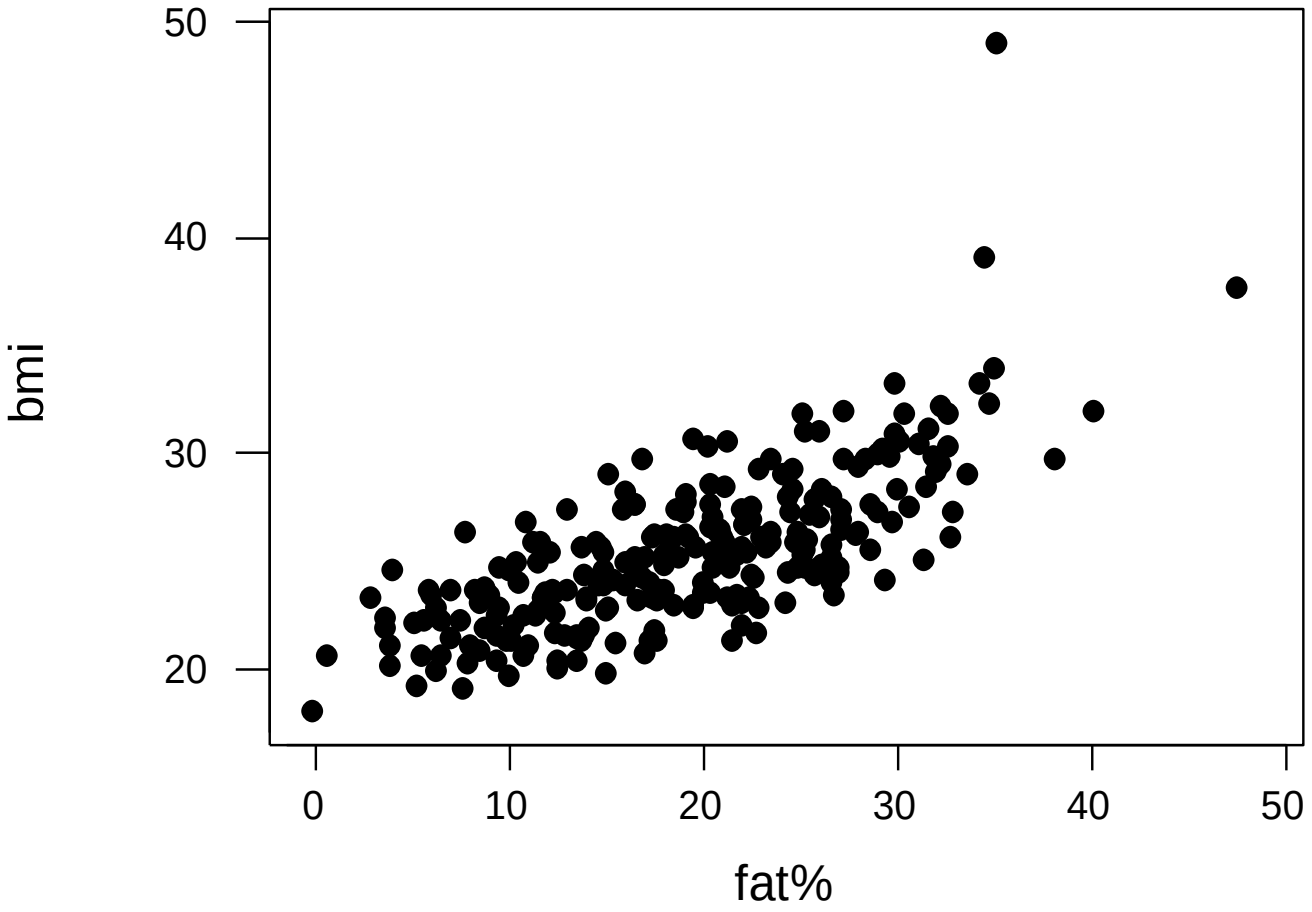
252 men had a variety of measurements taken on them relative to their body sizes and densities

- abdomen circumference (cm)
- chest circumference (cm)
- wrist circumference (cm)
- body mass index (weight in kg/height in meters squared)
- proportion of fat tissue in the body
- density (grams per cubic cm)

Negative Association



Positive Association



Strength of Association

Strong

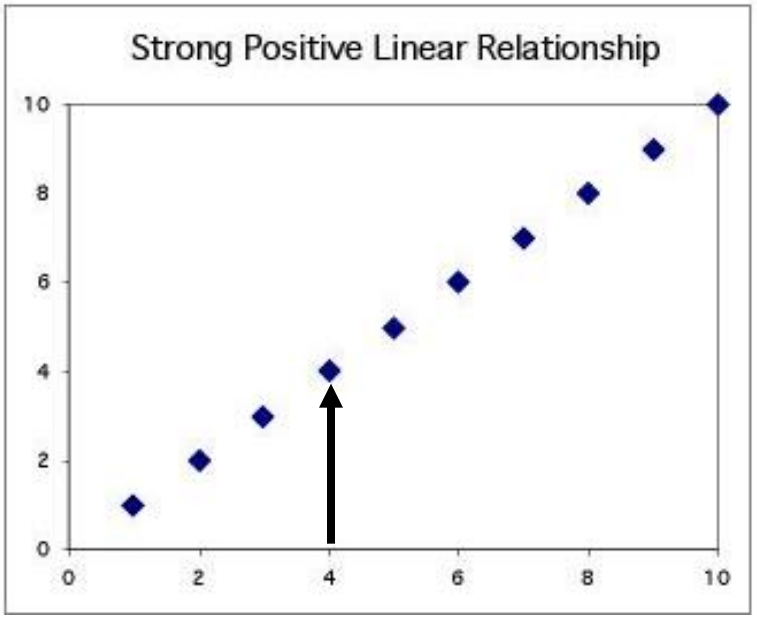
- the scatterplot points will tend to fall along a straight line

Weak

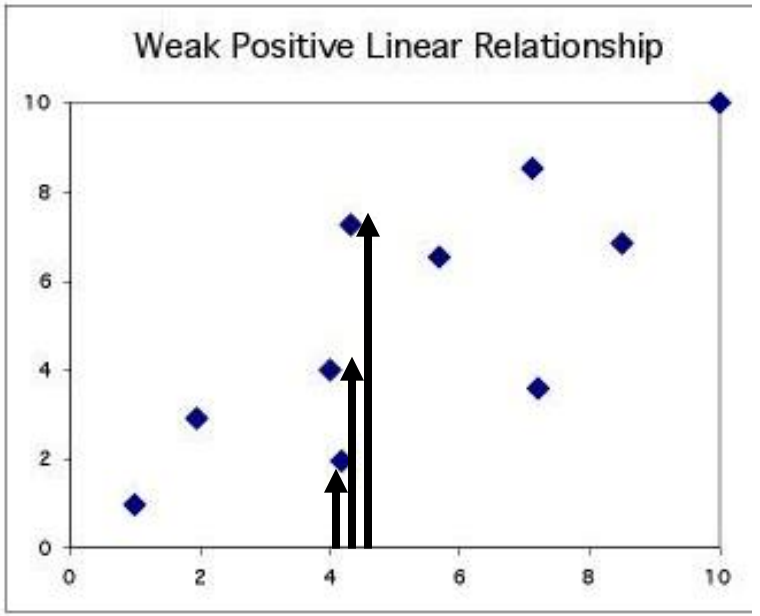
- the scatterplot points will be highly variable about the possible trend line
- no association if the trend line appears to be horizontal

Strength of the association

The **strength** of the relationship between the two variables can be seen by how much variation, or **scatter,** there is around the main form.

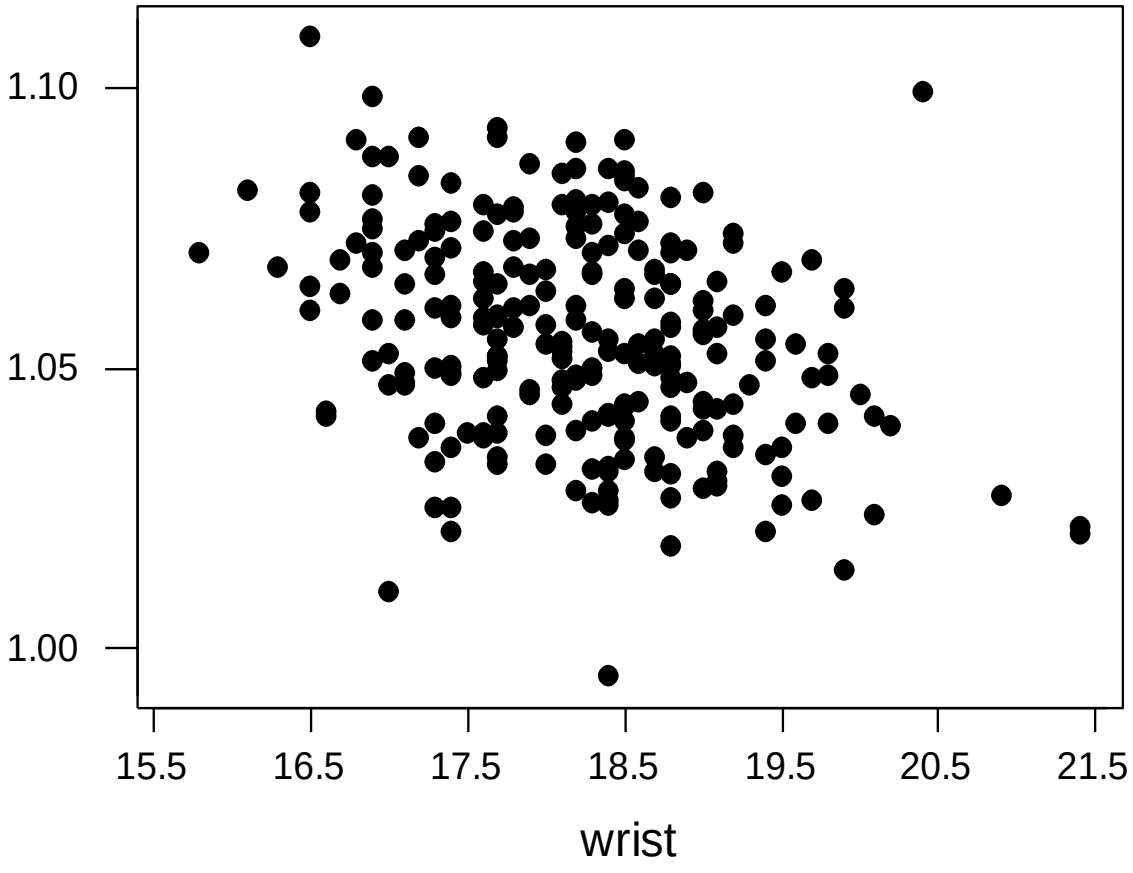


With a strong relationship, you can get a pretty good estimate of *y* if you know *x*.

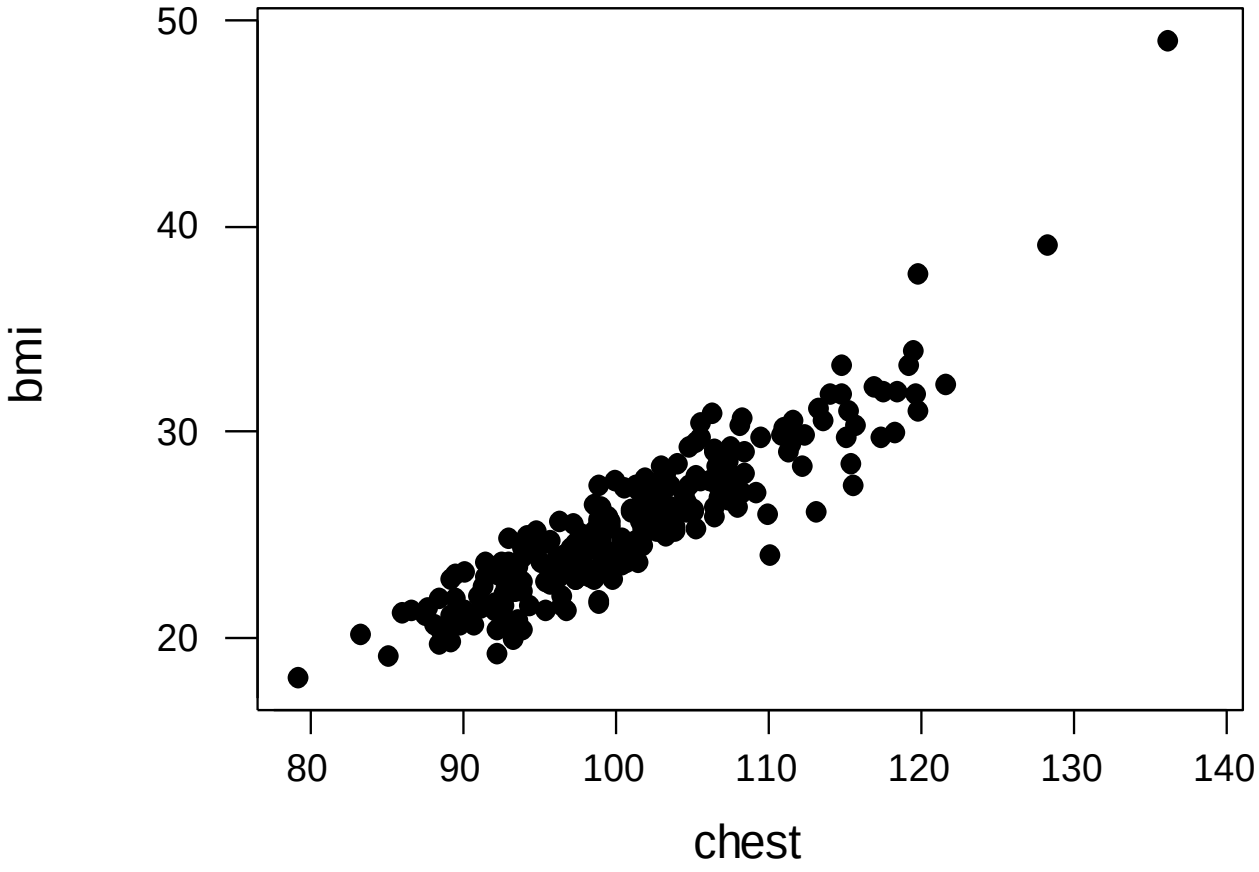


With a weak relationship, for any *x* you might get a wide range of *y* values.

Weak (Negative) Association



Strong (Positive) Association



CATEGORICAL VARIABLES IN SCATTERPLOTS

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

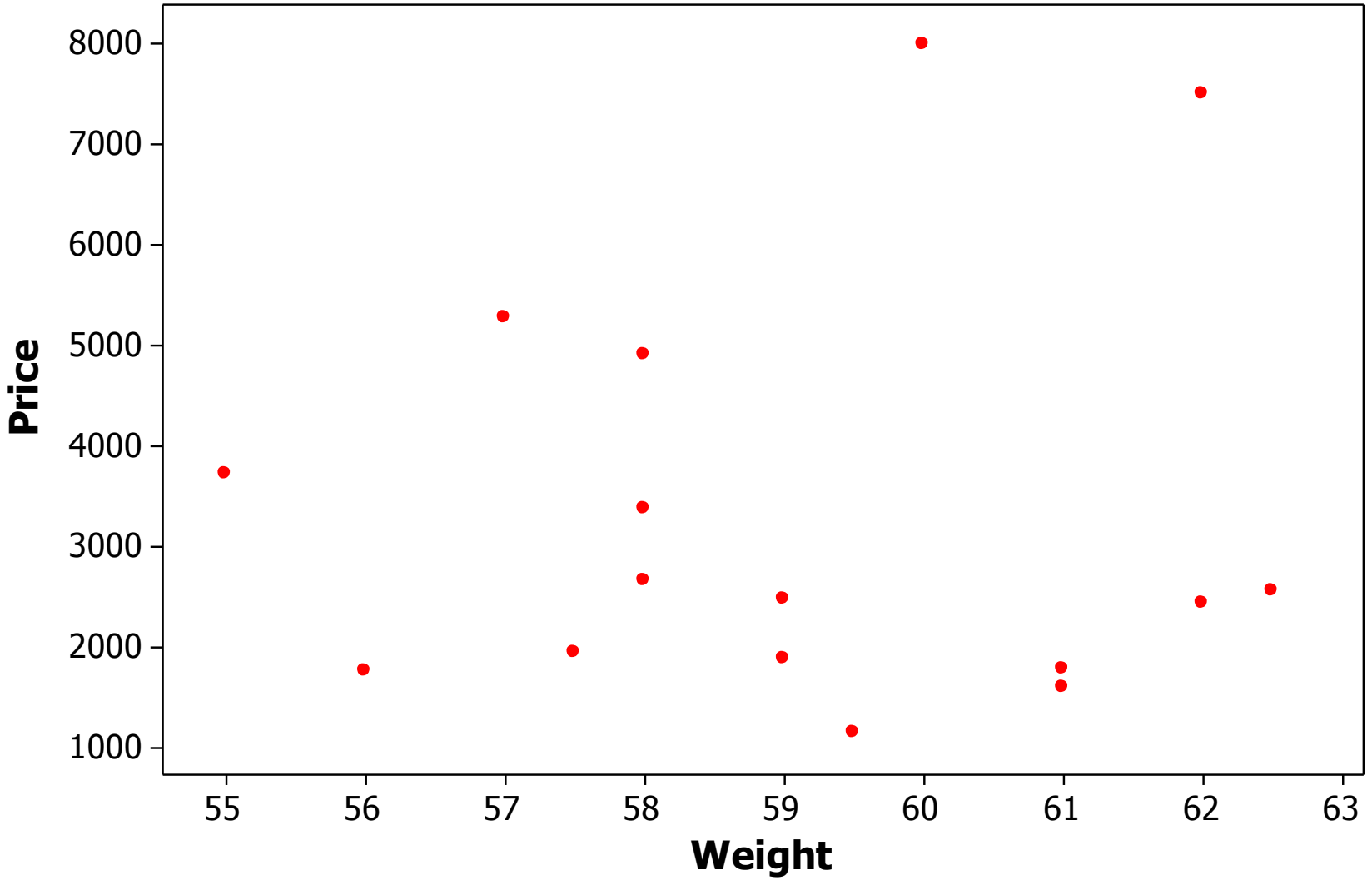
Scatterplots and Categorical Variables

- relationship between the price of a violin bow and its attributes such as age, weight, shape and ornamentation (gold, tortoise shell or pearl) on the bow

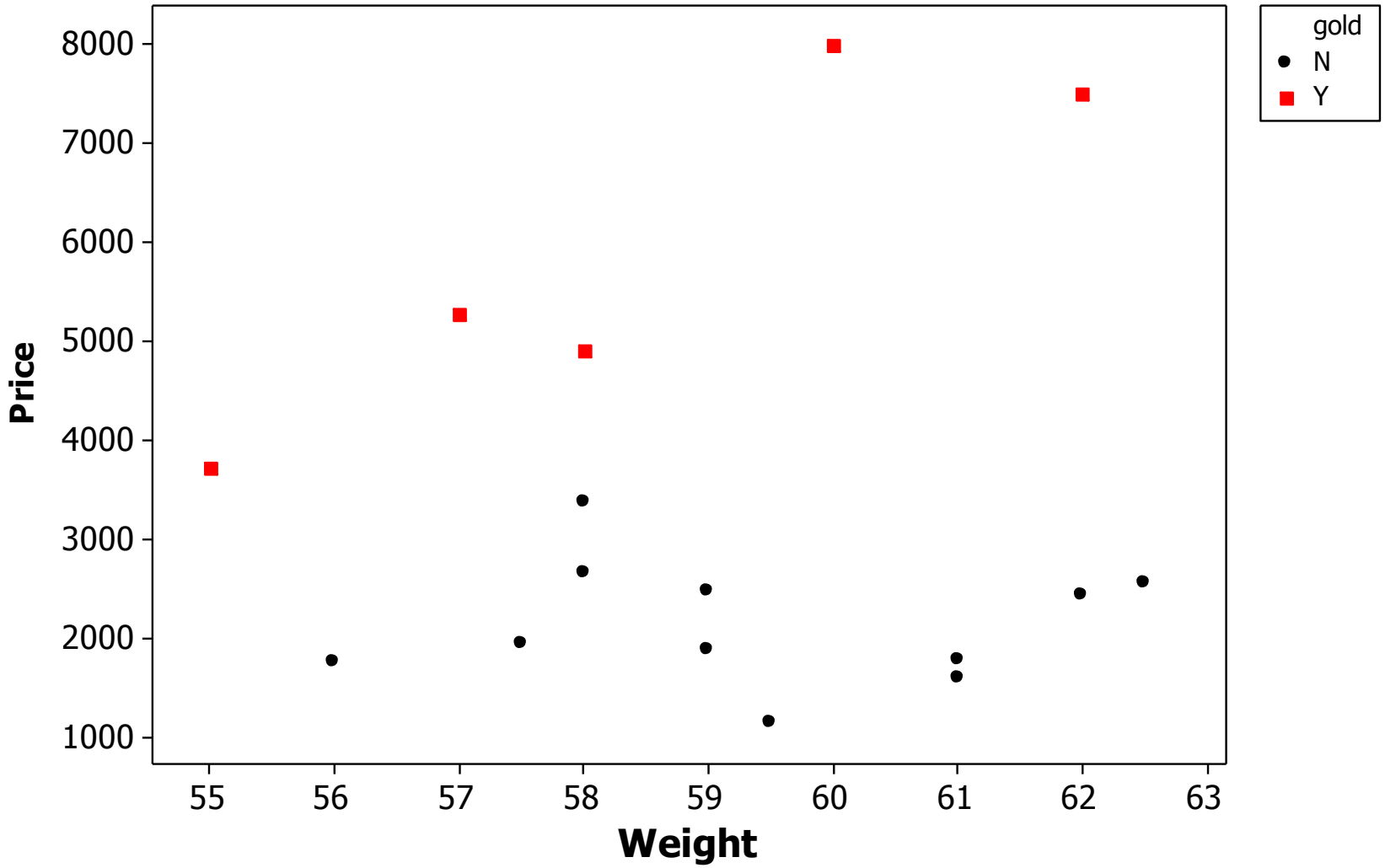
Source of the Data

- bowmaker – W.H. Hill and Sons, London
- purchase prices of bows sold at auction by Sotheby's, 1994 - 1997

Price in U.S. dollars versus weight in grams



Price by weight with or without gold ornamentation



CORRELATION

The **correlation** measures the direction and strength 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

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

or, more compactly,

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

Explanatory and Response Variables?

- correlation makes no distinction between explanatory and response variables
- can call either variable *y* and the other one *x*

Method of Calculation

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

- standardize each of the *x* observations
- standardize each of the *y* observations
- find the products of the standardized values for each observation
- sum these products over all the observations and divide by $n - 1$.

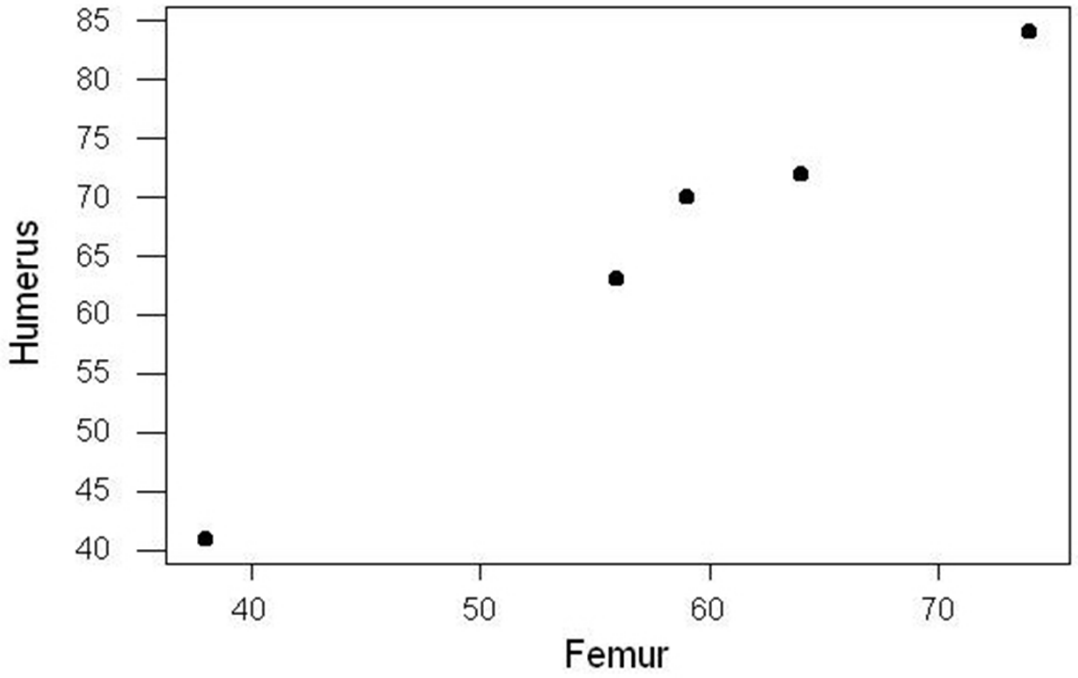
Example: Classifying Fossils

Archaeopteryx is an extinct beast having feathers like a bird but teeth and a long tail like a reptile. Only six fossil specimens are known. Because these fossils differ greatly in size, some scientists think they are different species rather than individuals from the same species.

The scientists examined data on the lengths in centimeters of the femur (a leg bone) and the humerus (a bone in the upper arm) for the five fossils that preserve both bones.

Data and Scatterplot

Femur length in cm (x)	38	56	59	64	74
Humerus length in cm (y)	41	63	70	72	84



Correlation Calculation

x	38	56	59	64	74
y	41	63	70	72	84

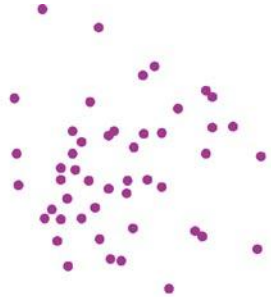
$$\bar{x} = 58.2, \bar{y} = 66.0, s_x = 13.20, s_y = 15.89$$

$$r = \frac{1}{4}\left(\frac{(38-58.2)(41-66.0)}{(13.20)(15.89)} + \frac{(56-58.2)(63-66.0)}{(13.20)(15.89)} + \frac{(59-58.2)(70-66.0)}{(13.20)(15.89)} + \frac{(64-58.2)(72-66.0)}{(13.20)(15.89)} + \frac{(74-58.2)(84-66.0)}{(13.20)(15.89)}\right)$$

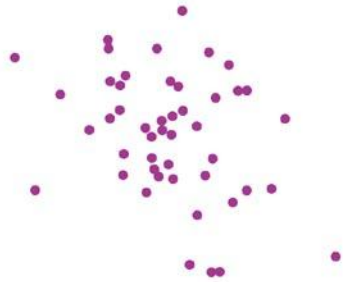
$$= 0.994$$

What is Correlation?

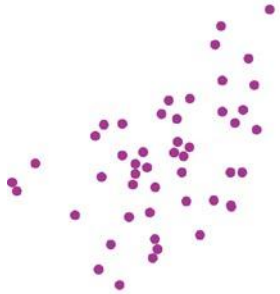
- correlation measures the direction and strength of the linear relationship between two quantitative variables.
- value of the correlation is denoted by *r*
- *r* takes on values between –1 and +1: –1 is a perfect linear relationship between the variables (with a negative slope on the line) and +1 is a perfect linear relationship with a positive slope.



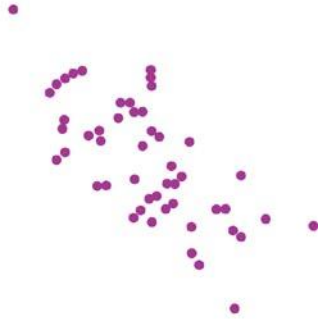
Correlation $r = 0$



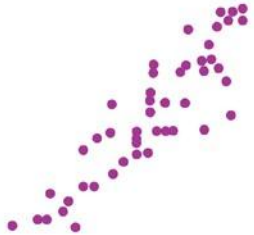
Correlation $r = -0.3$



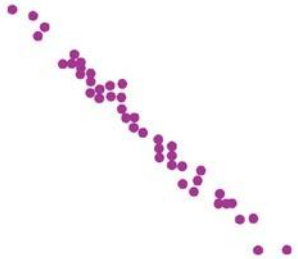
Correlation $r = 0.5$



Correlation $r = -0.7$

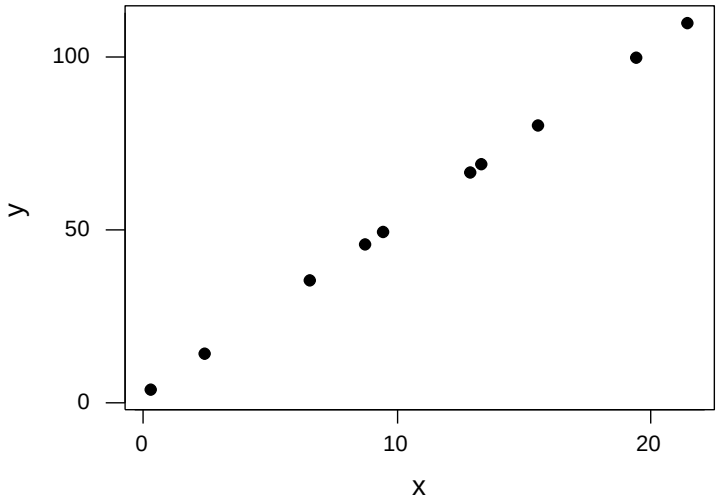


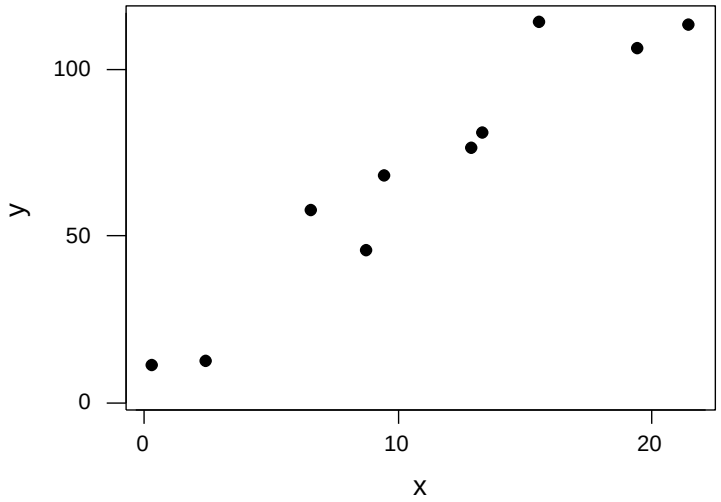
Correlation $r = 0.9$



Correlation $r = -0.99$

Positive Correlation

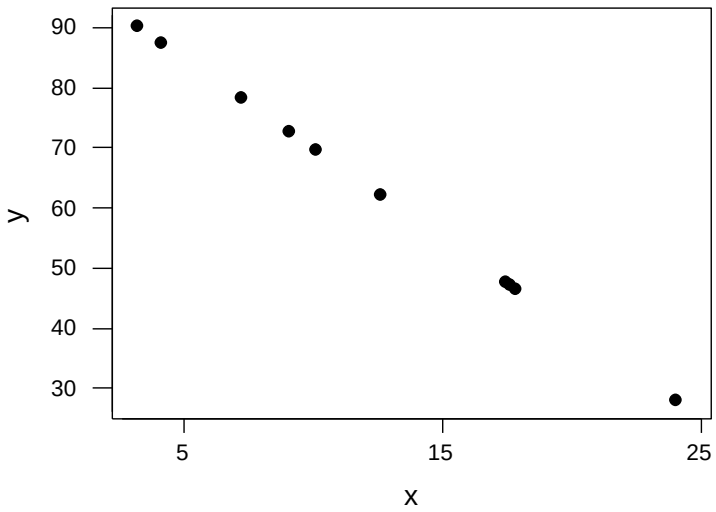


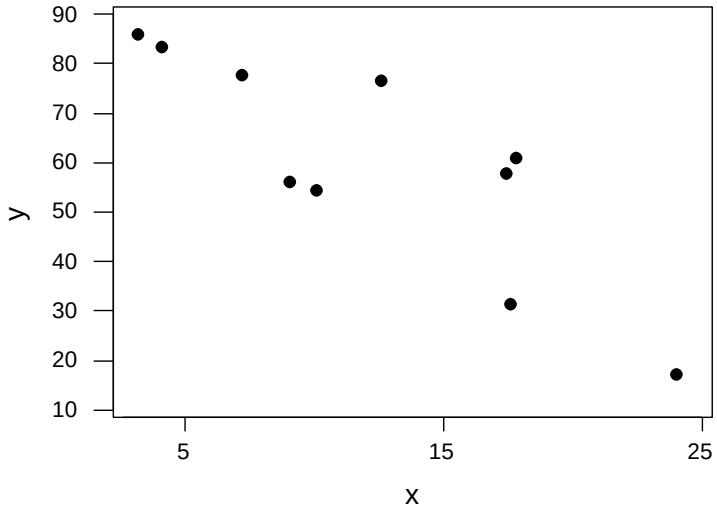


$r = 1$ $0 < r < 1$

positive association between y and x

Negative Correlation





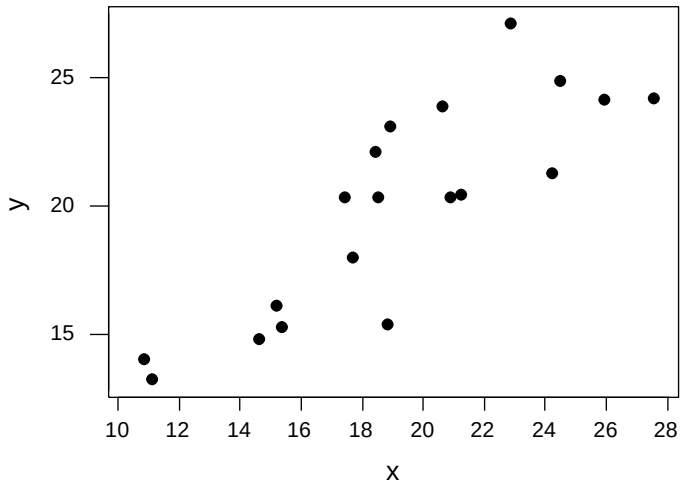
$r = -1$ $-1 < r < 0$

negative association between y and x

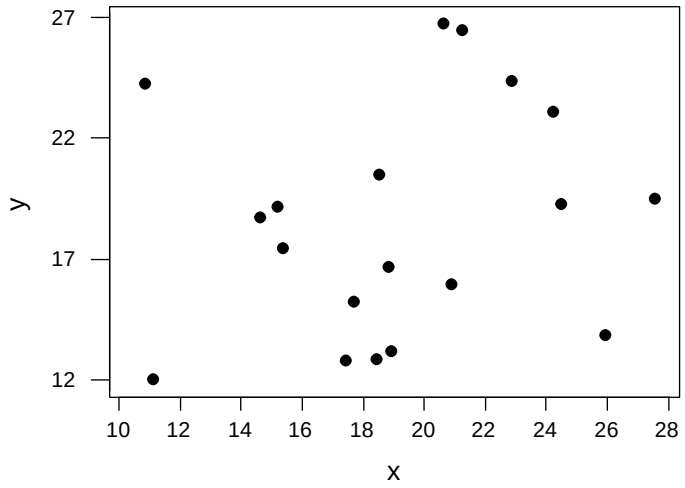
Strength of Association

- values of *r* near 0 indicate a weak association
- the strength of the association increases as you move away from 0 toward either –1 or +1
- values close to –1 or +1 indicate the points in the scatterplot are close to a straight line

Positive Association

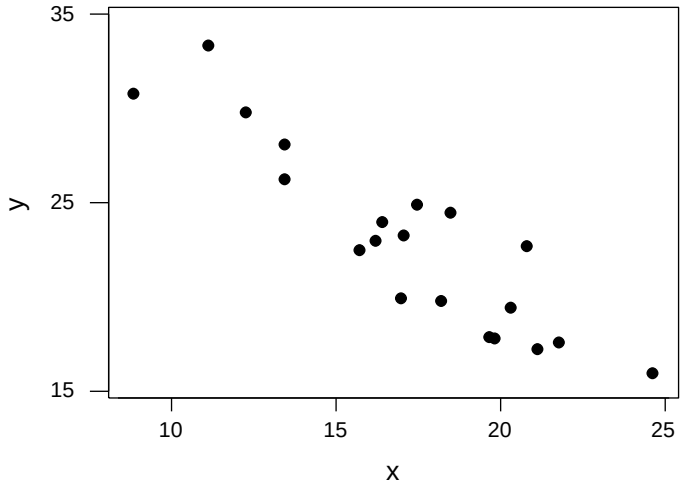


strong $r = 0.843$

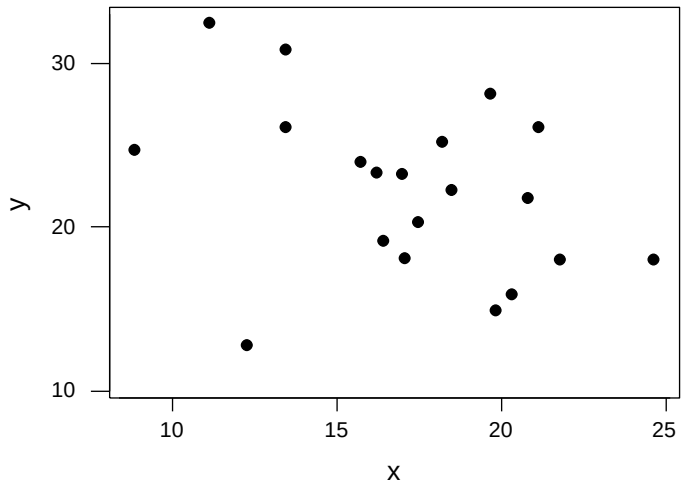


weak $r = 0.189$

Negative Association

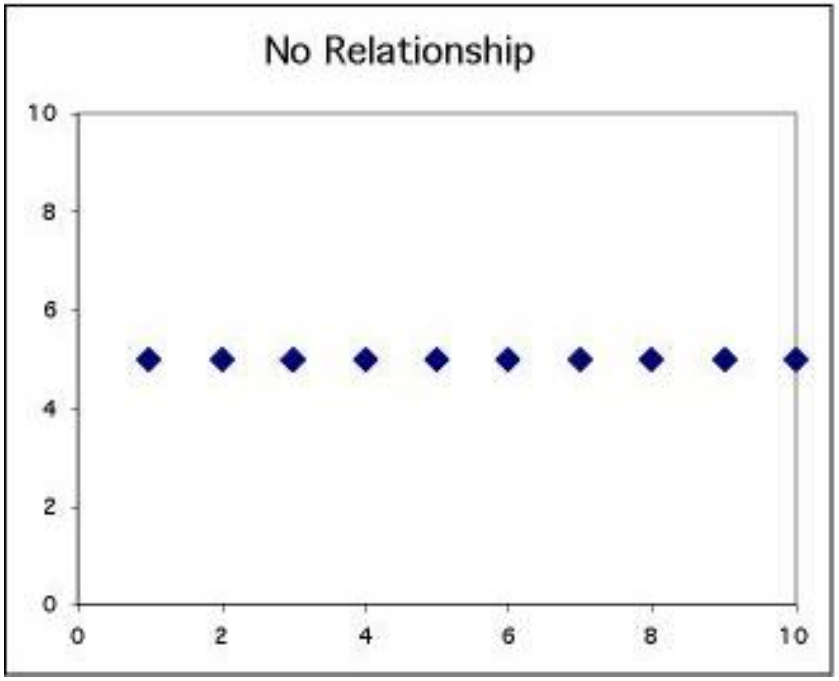


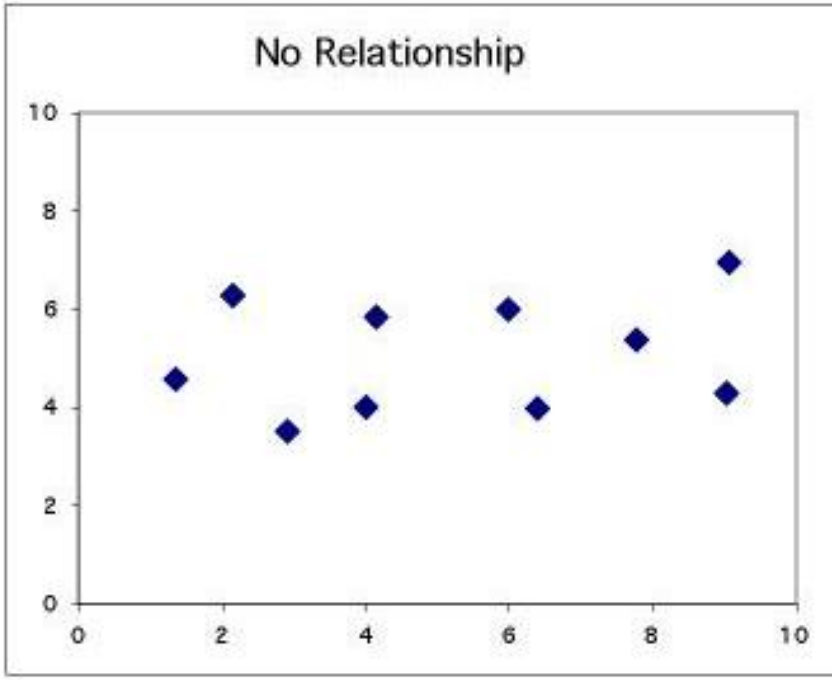
strong $r = -0.906$



weak $r = -0.368$

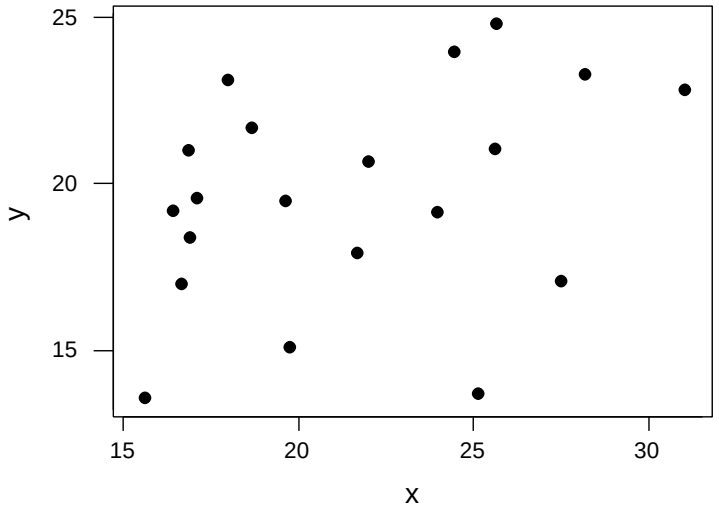
No relationship:
x and *y* vary independently. Knowing *x* tells you nothing about *y*.

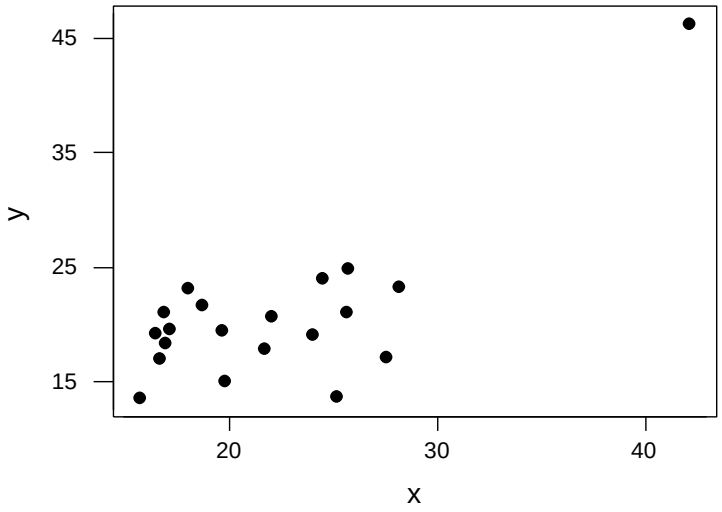




One way to remember this:
The equation for this line is y = 5.
x is not involved.

Effect of Outliers on Correlation



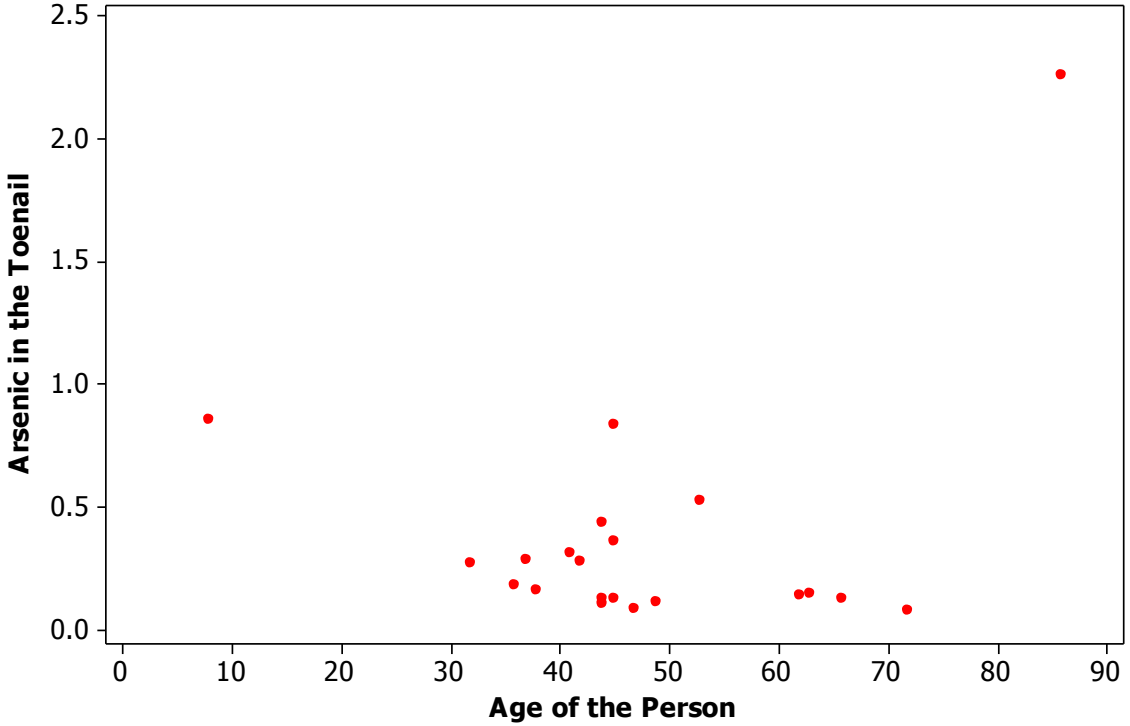


$r = 0.342$ $r = 0.753$

- only one number has been changed between the graphs, the largest *x*
- correlation is not a resistant measure

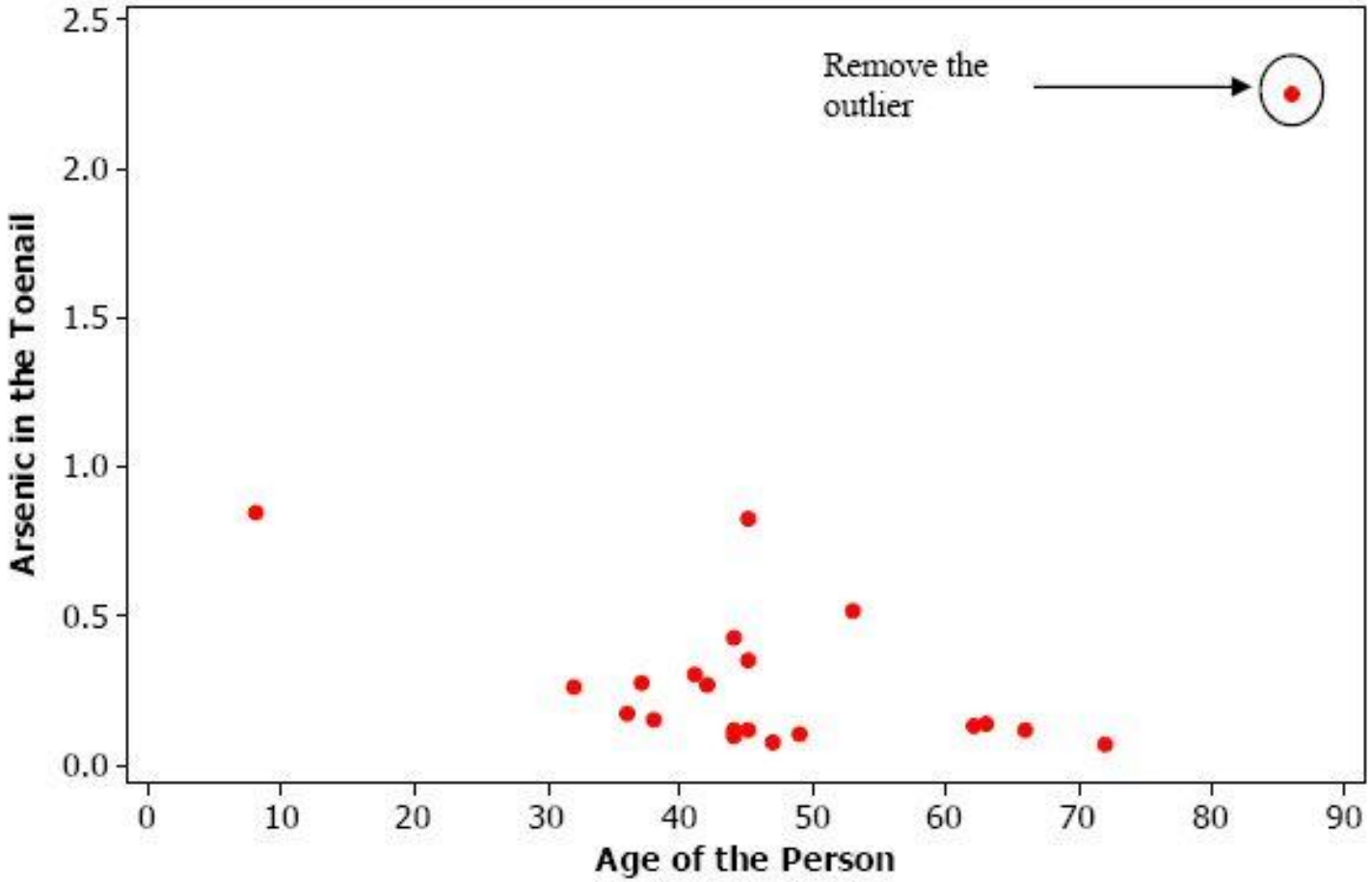
More on Outliers

- study to find the relationship between arsenic in well water in a community and arsenic in the body (measured in toe nails)
- graph shows arsenic content versus age

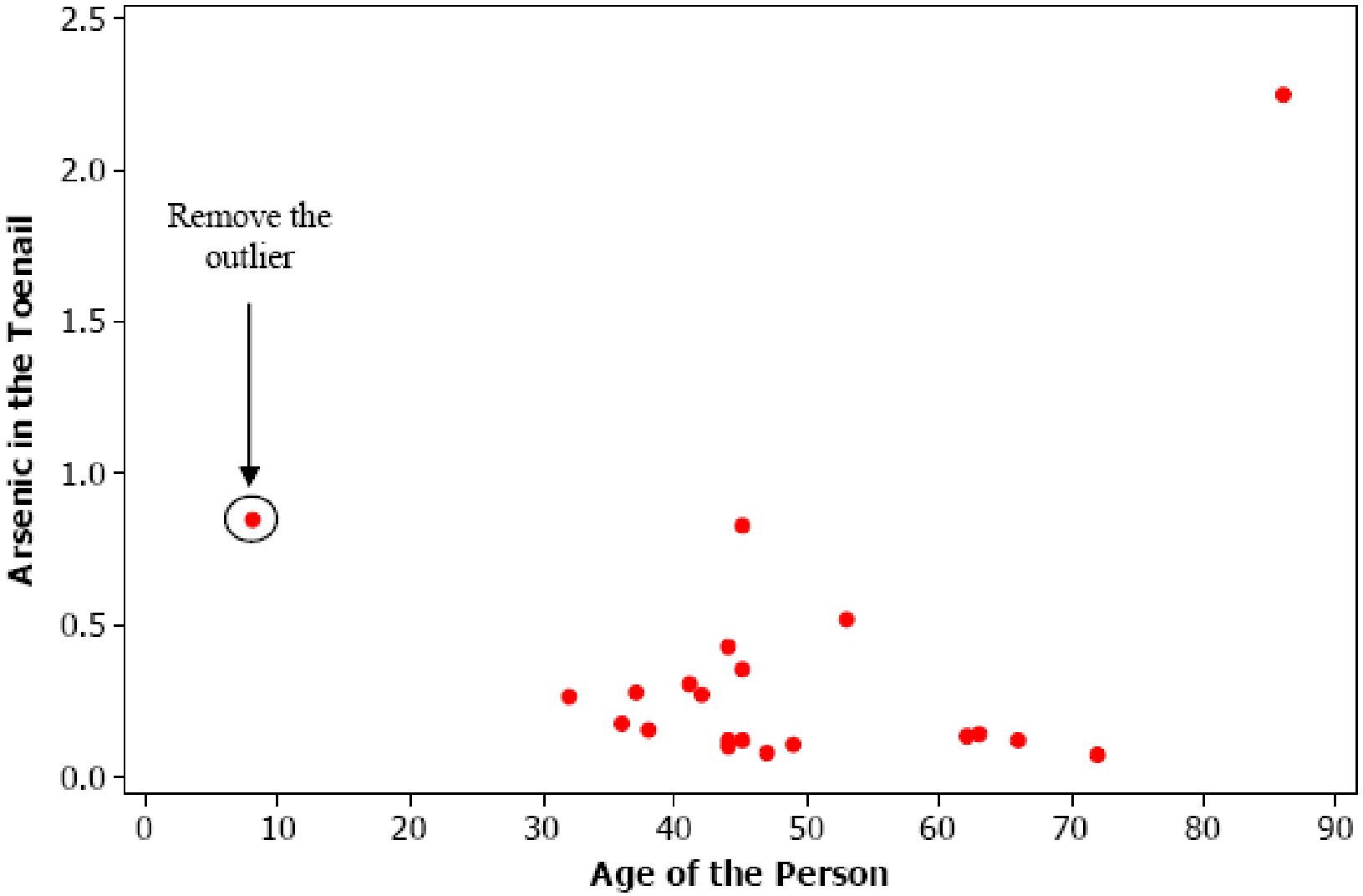


$r = 0.281$

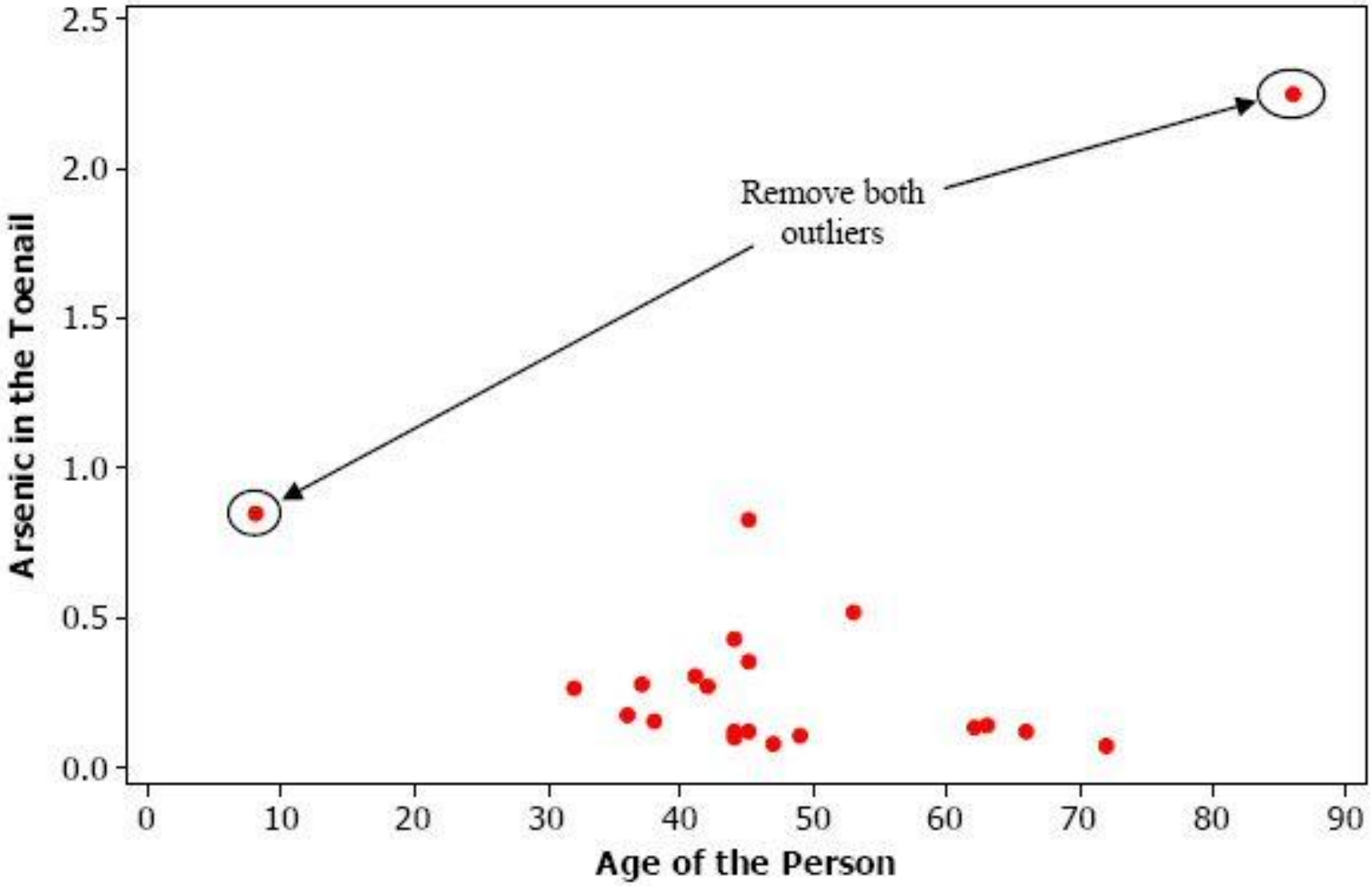
Effect of Outliers on Correlation



$r = -0.532$


Remove the outlier
Arsenic in the Toenail
Age of the Person
2.5
2.0
1.5
1.0
0.5
0.0
0
10
20
30
40
50
60
70
80
90

$r = 0.022$


Remove both
outliers
Arsenic in the Toenail
Age of the Person
2.5
2.0
1.5
1.0
0.5
0.0
0
10
20
30
40
50
60
70
80
90

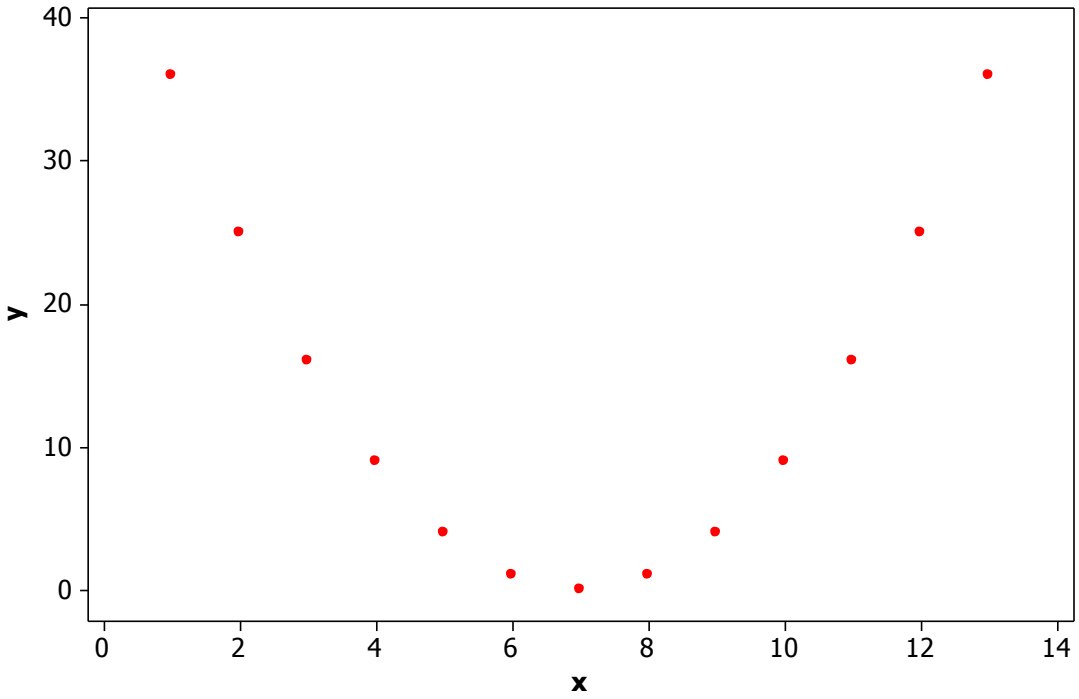
$r = -0.247$

Units of Measurement and Correlation

***r* has no unit of measurement; it is a number between –1 and +1**

- *r* does not change when the unit of measurement is changed – a correlation between weights and heights will be the same if weights are measured in kilograms or pounds and height are measured in centimeters or inches

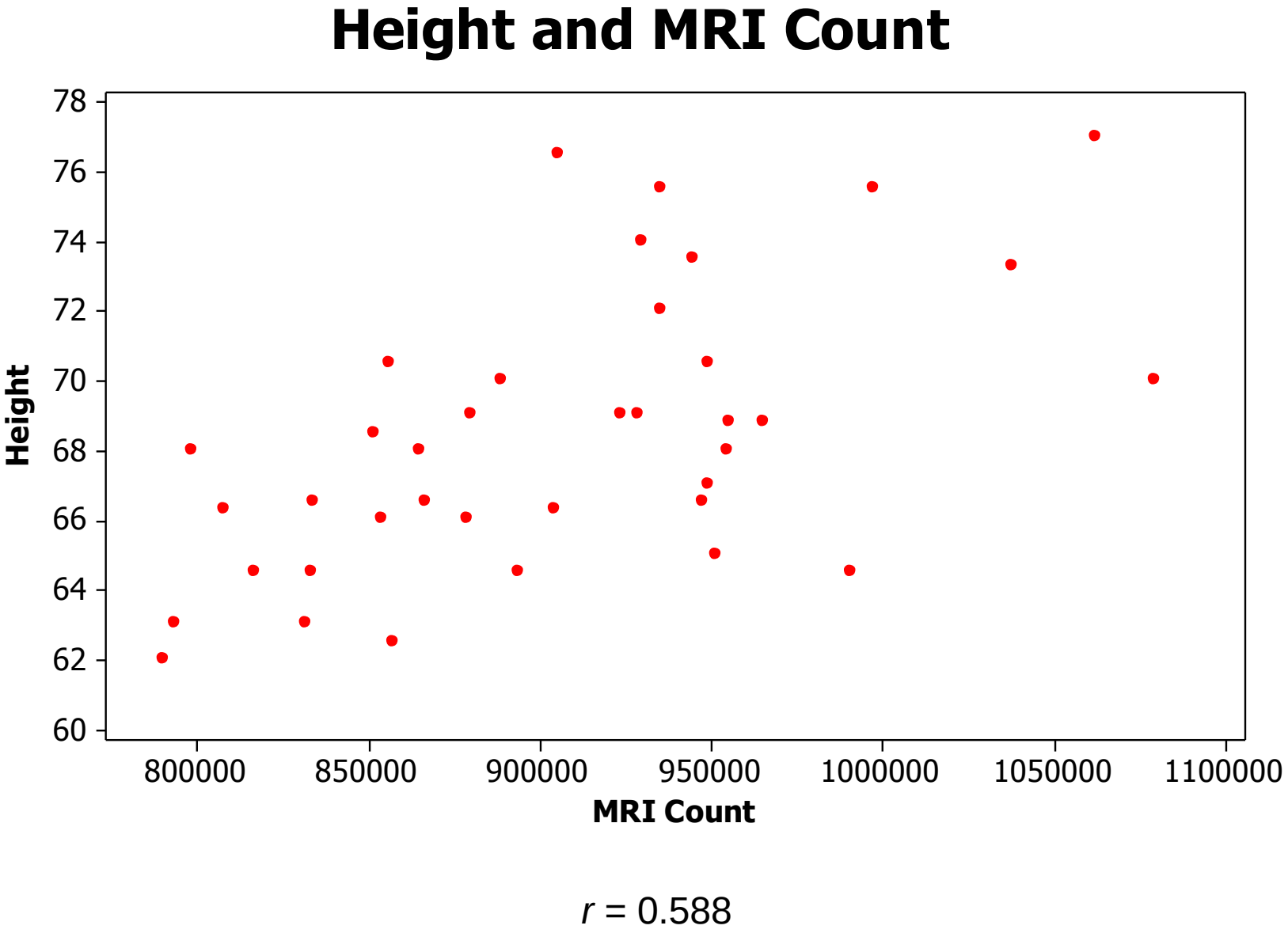
Correlation Measures Linear Association Only



- the graph shows a perfect quadratic relationship between y and x.
- correlation $r = 0$
- correlation does **not** measure the strength of curved relationships

Final Example – Brain Size

- Lee Willerman and his colleagues (*Intelligence*, 1991) conducted a study at a large southwestern university. They selected a sample of 40 right-handed Anglo introductory psychology students who had indicated no history of alcoholism, unconsciousness, brain damage, epilepsy, or heart disease. These subjects were drawn from a larger pool of introductory psychology students with total Scholastic Aptitude Test Scores higher than 1350 or lower than 940 who had agreed to satisfy a course requirement by allowing the administration of four subtests (Vocabulary, Similarities, Block Design, and Picture Completion) of the Wechsler Adult Intelligence Scale-Revised.

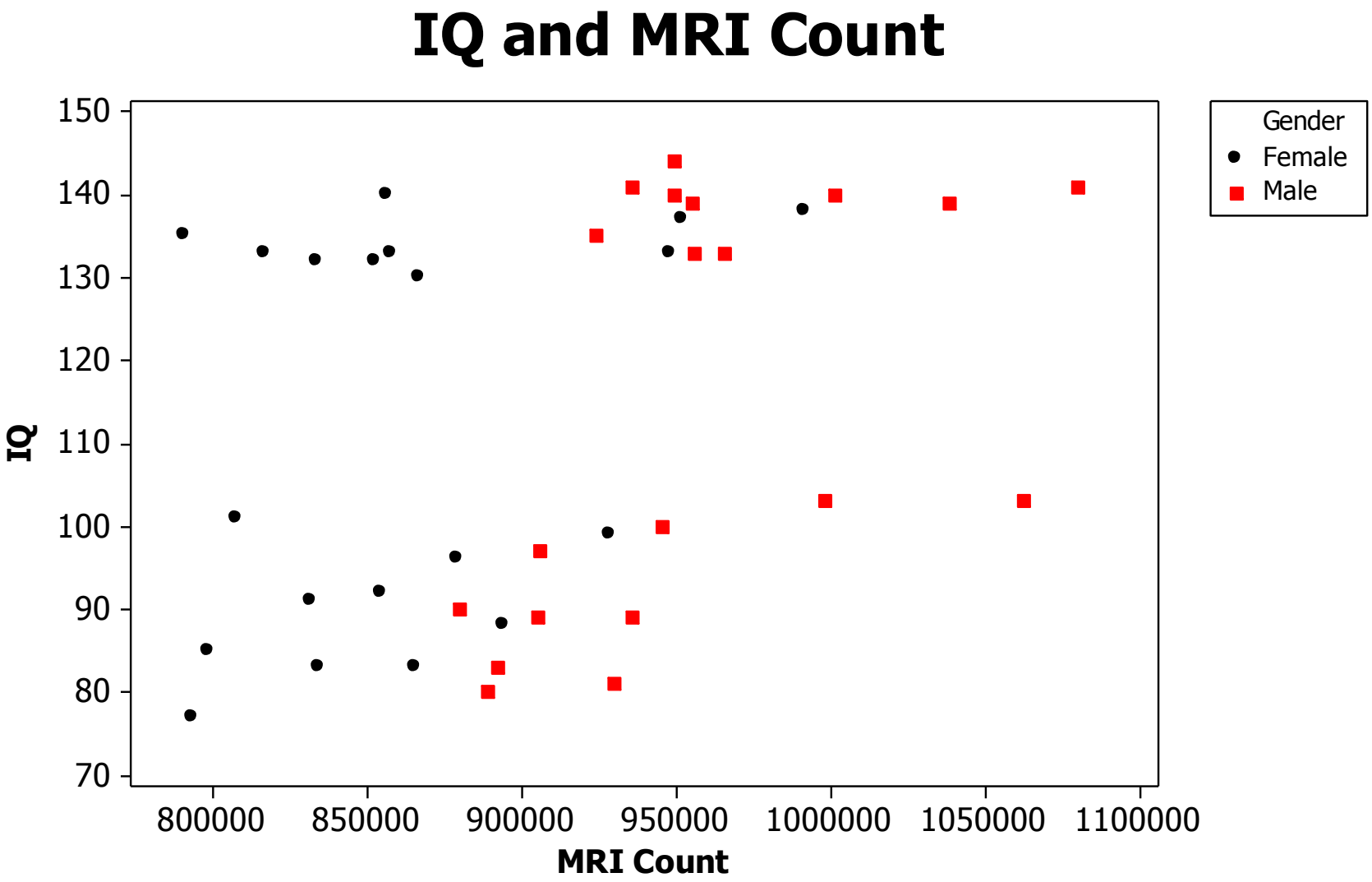

Height and MRI Count
78
76
74
72
70
68
66
64
62
60
Height
800000
850000
900000
950000
1000000
1050000
1100000
MRI Count
r = 0.588

Brain Size and IQ

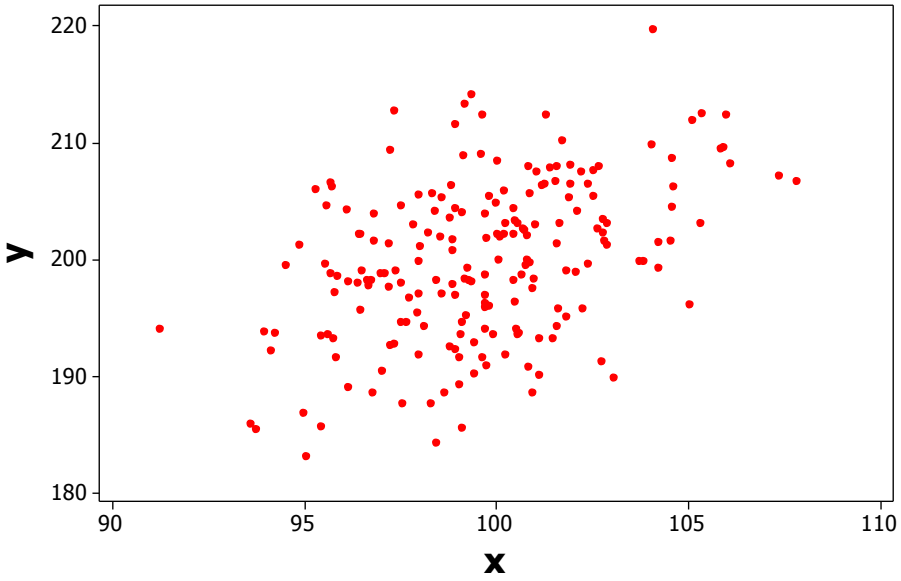
$$r = \frac{1}{39}\sum_{i=1}^{40}\left(\frac{x_i - 113.45}{24.08}\right)\left(\frac{y_i - 908755}{72282}\right)$$

$$= 0.358$$

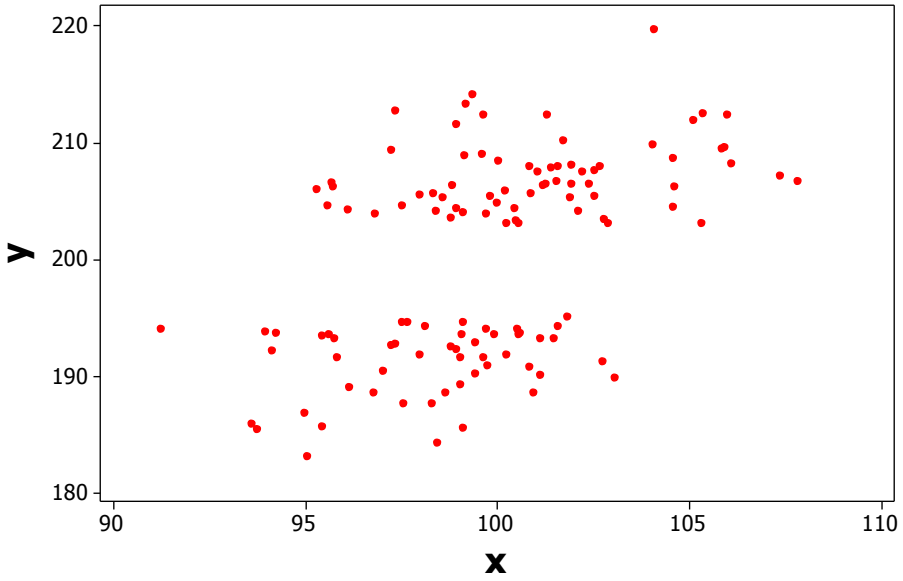
But remember the study design



Effect of the Study Design on Correlation Example – simulated data



$r = 0.426$



$r = 0.496$

Removing the middle part of the data tends to increase the value of the correlation coefficient