# Chapter 6: Scatterplots, Association, and Correlation

We often want to examine the relationship or association between two quantitative variables. ( bivariate data )

We will choose one variable to be the **response variable** and one variable to be the **explanatory variable**:

- **Response Variable**: Variable of interest that we want to predict or explain.
- **Explanatory Variable:** Variable that accounts for or explains the outcome of the response variable.

**Example:**

- Maximum daily temperature vs cooling cost.

  explanatory response

- Number of loan applications vs interest rate.

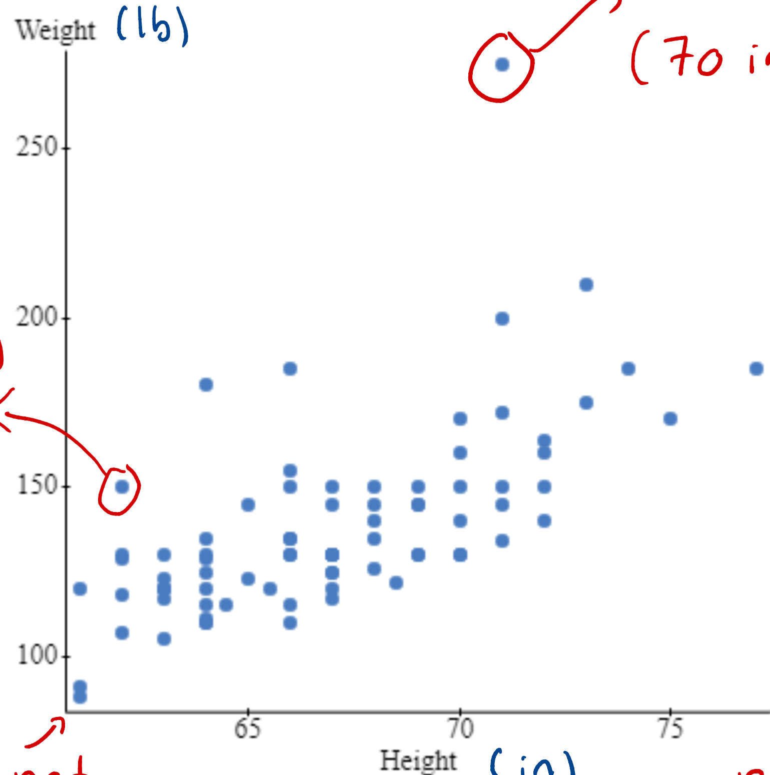
  response explanatory

A **scatterplot** is a display for two quantitative variables:

- The explanatory variable is placed on the horizontal axis ($x$-axis).
- The response variable is placed on the vertical axis ($y$-axis).
- The values for the two variables for a subject are represented by a point.
- If there are $n$ subjects, then the scatterplot will have $n$ points.

**Example:** Heights vs weights of 80 people:

outlier
(70 in, 275 lb)




(61 in, 150 lb)

moderate
positive
linear
association

need not
intersect at (0,0)

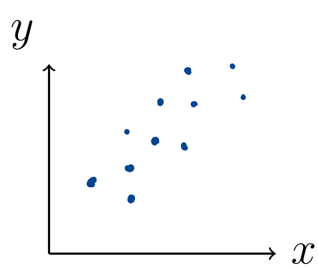
r = 0.644

## What to Look for in a Scatterplot

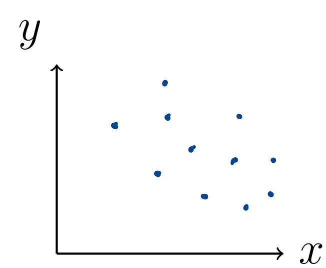
When we look at a scatterplot, we want to watch for trends or overall patterns in the scatterplot:

a) **Direction:**

- If the points tend to rise to the right, then there is a **positive association**. (As $x$ increases, $y$ increases.)
- If the points fall to the right, then there is a **negative association**. (As $x$ increases, $y$ decreases.)



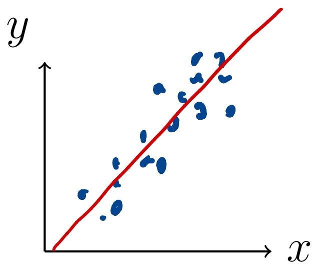

positive association



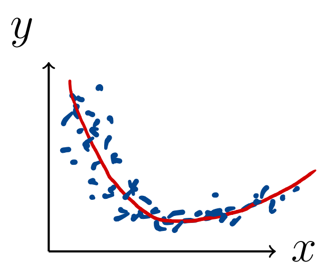

negative association

b) **Form of the Relationship:**

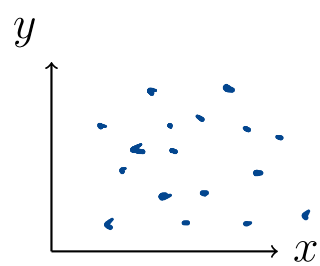
- **Linear Relationship:** points roughly follow a straight line.
- curved
- no pattern




linear



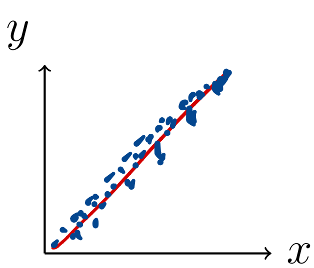

Curved



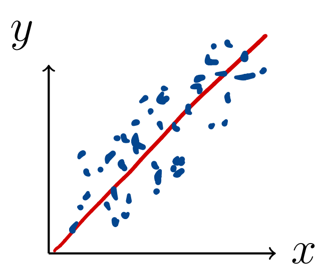

no pattern

c) **Strength of the Relationship:** How much scatter?

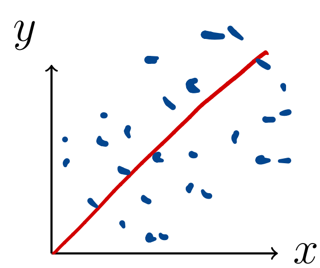
- Defined by how close the points lie to the form.
- The more tightly clustered the points are around the form, the stronger the relationship.




Strong




moderate




weak

d) **Unusual Features or Outliers**

# Correlation

If a scatterplot appears to show a linear relationship, then we can measure the direction and strength of the linear relationship between the two variables by computing the **correlation coefficient**, denoted $r$.

Suppose that we have data for variables $x$ and $y$ for $n$ individuals. Let

$$(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots, (x_n, y_n)$$

be the $n$ pairs of observations. Let $\bar{x}$ and $s_x$ be the mean and standard deviation for the $x$-values and let $\bar{y}$ and $s_y$ be the mean and standard deviation for the $y$-values.

To calculate $r$, we can use the formula:

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

standardized

We often let $z_{x_i} = \dfrac{x_i - \bar{x}}{s_x}$ and $z_{y_i} = \dfrac{y_i - \bar{y}}{s_y}$, so that this formula becomes

$$r = \frac{1}{n-1}\left(\sum_{i=1}^{n} z_{x_i} z_{y_i}\right)$$

**Example:** Class Absences vs Final Grades

| Student | Number of Absences ($x$) | Final Grade ($y$) | $z_x$ | $z_y$ | $z_x z_y$ |
|---|---|---|---|---|---|
| 1 | 6 | 82 | -0.490 | 0.536 | -0.263 |
| 2 | 2 | 86 | -1.404 | 0.775 | -1.088 |
| 3 | 15 | 43 | 1.567 | -1.788 | -2.802 |
| 4 | 9 | 74 | 0.196 | 0.060 | 0.012 |
| 5 | 12 | 58 | 0.882 | -0.894 | -0.789 |
| 6 | 5 | 90 | -0.718 | 1.013 | -0.727 |
| 7 | 8 | 78 | -0.033 | 0.298 | -0.01 |

For (5, 90):

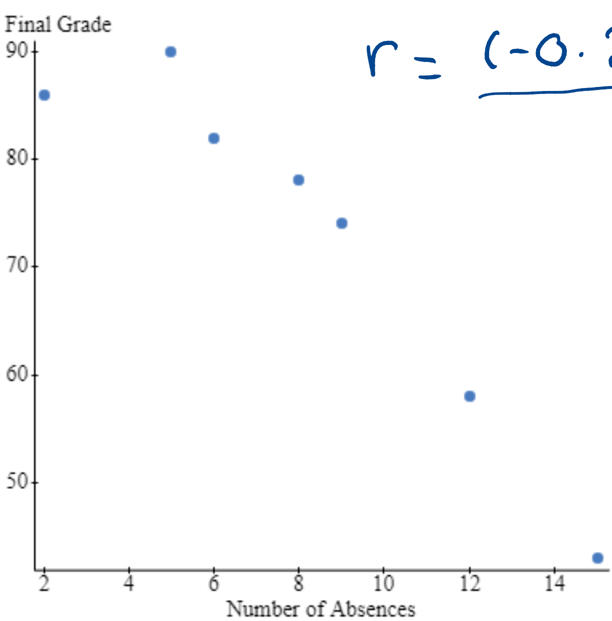
$$z_x = \frac{5 - 8.143}{4.375} = -0.718$$

$$z_y = \frac{90 - 73}{16.783} = 1.013$$

$$z_x z_y = (-0.718)(1.013) = -0.727$$

$n = 7$ exp. res.

**Note:** $\bar{x} = 8.143$, $s_x = 4.375$, $\bar{y} = 73$, $s_y = 16.783$




$$r = \frac{(-0.263) + (-1.088) + \dots + (-0.01)}{6} = -0.94$$

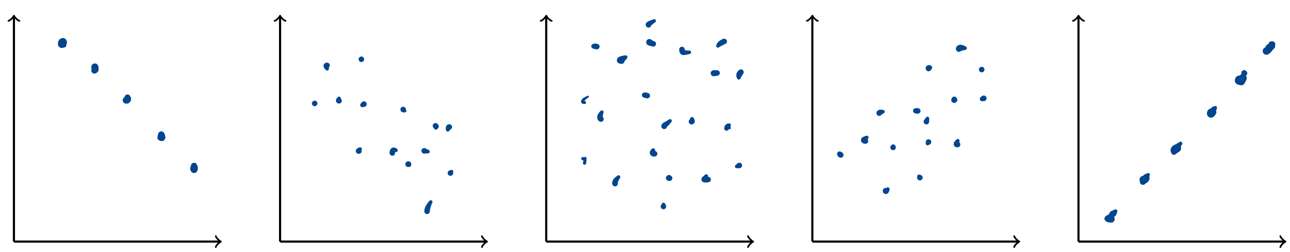
strong, negative, linear association

# Properties of the Correlation Coefficient

- The correlation coefficient is a value between $-1$ and 1. $-1 \leq r \leq 1$

- The sign (positive or negative) of the correlation coefficient gives us the direction of the association:
  - A positive value ($r > 0$) means there is a positive association.
  - A negative value ($r < 0$) means there is a negative association.

- The further $r$ falls from 0, the stronger the linear association between the two variables, that is, the closer the points fall to a straight line.



$r = -1$ $\quad$ $r = -0.5$ $\quad$ $r = 0$ $\quad$ $r = 0.5$ $\quad$ $r = 1$

strong $|r| > 0.8$

- Correlation is symmetric: the correlation of $x$ with $y$ is the same as the correlation of $y$ with $x$. $\rightarrow$ get same $r$ value

if interchange roles of variables

- Correlation is sensitive to outliers. Outliers can drastically affect the value of $r$.

- Correlation is not affected by changes in the centre or scale of either variable.

ex. same $r$ value if convert data units : kg $\rightarrow$ lb, cm $\rightarrow$ in

- Correlation has no units.

Correlation measures the strength of the linear association between two quantitative variables.

Before we use a correlation, we should check three conditions:

- **Quantitative Variable Condition:** Both variables must be quantitative variables. Does not apply to categorical variables.
- **Straight Enough Condition:** Correlation does not apply to non-linear relationships. Look at the scatterplot to see if it is reasonably straight.
- **No Outliers Condition:** Outliers can drastically distort a correlation. It can change the sign of a correlation. It can make a weak association seem strong or a strong association seem weak. Scan the scatterplot for outliers.

properly randomized experiment

**Warning:** Correlation does not imply causation! There may be a lurking variable (a variable which is hidden, but may be influencing our understanding of the relationship between the two variables).

**Example:** There is a strong positive correlation between the number of firefighters at the scene of a house fire and the amount of damage (measured in dollars) sustained by the house.

lurking variable : Size of fire