Chapter 2 - Part 3

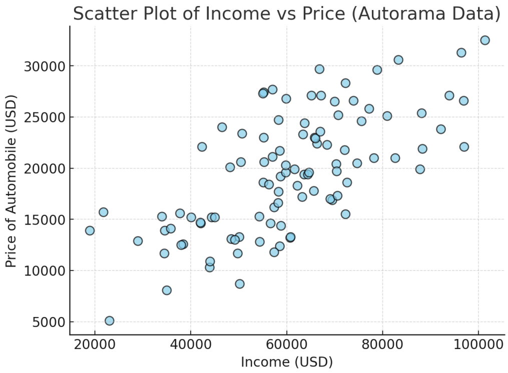
# Covariance and Correlation Coefficient

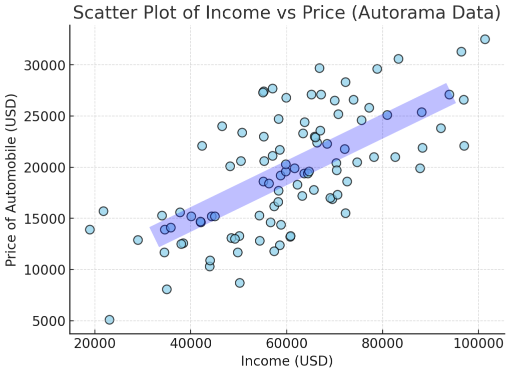
# Quantifying Relationship Between Two Numerical Variables

- Our next aim is to quantifying the relationship between two numerical variables.
- For this, we will define covariance and correlation coefficient.

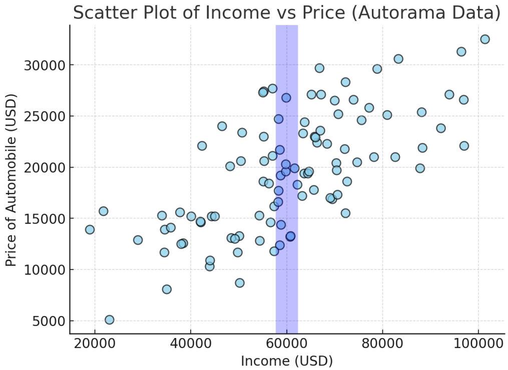
Consider data set of 100 credit applications to a bank, with the intention of buying a car. Our variables:

- Income: Annual income of the applicant, in USD.
- Price: The price of the car applicant plans to buy, in USD, (applying for credit to buy a car).



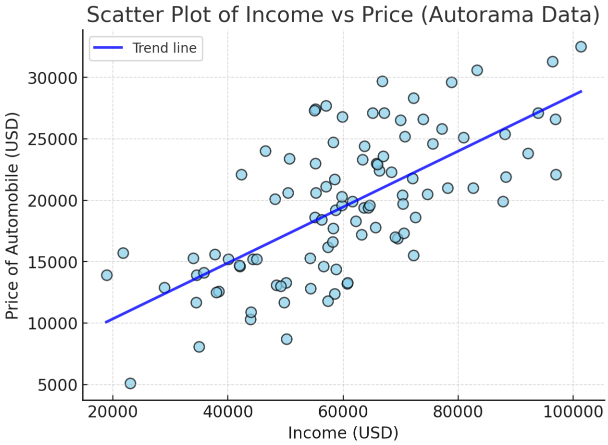



We see an increasing trend. Higher values of income are more likely to be observed with higher values of price, lower values of income are more likely to be observed with lower price. There is a linear increasing relationship.




Even though some customers have very similar income levels, the prices of cars they plan to buy may be very different. This means that the relationship between income and price is not very strong.

# Regression Line (Trend Line)



# Covariance

If data is $\{(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots, (x_n, y_n)\}$, then covariance is

$$s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

where

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}, \qquad \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

# Example

- **Example.** Calculate the covariance of $(1,7)$, $(2,13)$, $(4,9)$, $(5,10)$, $(8,12)$

$$\bar{x} = \frac{1+2+4+5+8}{5} = 4, \qquad \bar{y} = \frac{7+13+9+10+12}{5} = 10.2$$

| $x_i$ | $y_i$ | $x_i - \bar{x}$ | $y_i - \bar{y}$ | $(x_i - \bar{x})(y_i - \bar{y})$ |
|---|---|---|---|---|
| 1 | 7 | $-3$ | $-3.2$ | 9.6 |
| 2 | 13 | $-2$ | 2.8 | $-5.6$ |
| 4 | 9 | 0 | $-1.2$ | 0 |
| 5 | 10 | 1 | $-0.2$ | $-0.2$ |
| 8 | 12 | 4 | 1.8 | 7.2 |
| | | | | 11 |

$$s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{11}{4} = 2.75$$

# Variance as Covariance

Variance can be seen as a special case of covariance, where a variable is measured against itself.

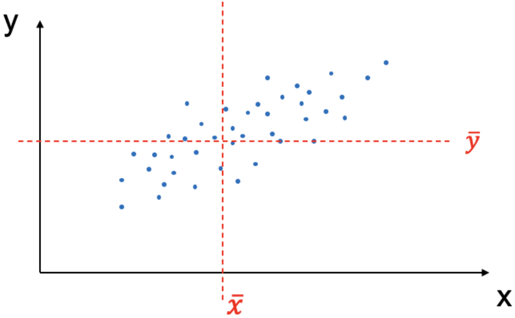
$$s_{xx} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})}{n-1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = s_x^2$$

# Interpretation

- $s_{xy} > 0 \implies$ a positive linear relationship (when one increases, the other tends to increase)
- $s_{xy} < 0 \implies$ a negative linear relationship (when one increases, the other tends to decrease)

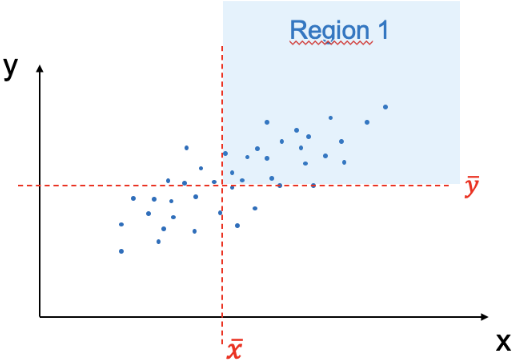
# Understanding Covariance Formula: A Linear Increasing Case

Divide the coordinate plane into four regions:



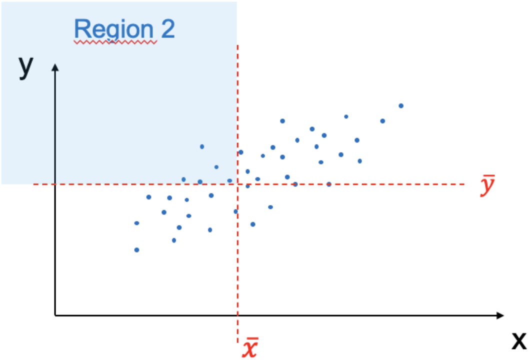

$$s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

$$s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$



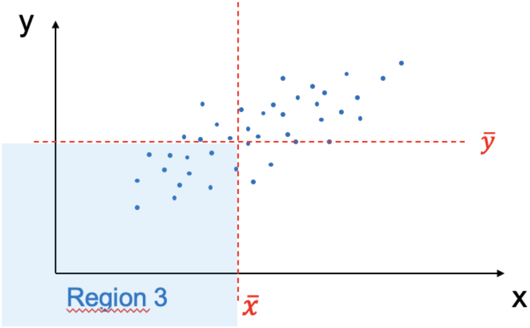

$$(x_i - \bar{x}) > 0 \quad \text{and} \quad (y_i - \bar{y}) > 0 \quad \Longrightarrow \quad (x_i - \bar{x})(y_i - \bar{y}) > 0$$

$$s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$



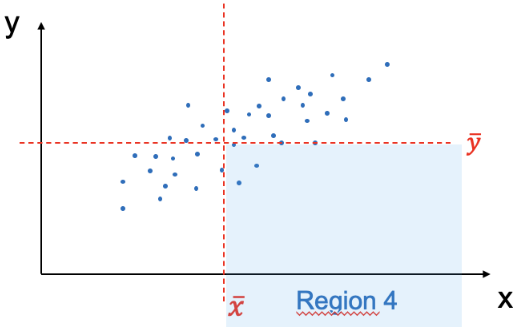

$$(x_i - \bar{x}) < 0 \quad \text{and} \quad (y_i - \bar{y}) > 0 \quad \Longrightarrow \quad (x_i - \bar{x})(y_i - \bar{y}) < 0$$

$$s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$




$$(x_i - \bar{x}) < 0 \quad \text{and} \quad (y_i - \bar{y}) < 0 \quad \Longrightarrow \quad (x_i - \bar{x})(y_i - \bar{y}) > 0$$

$$s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$



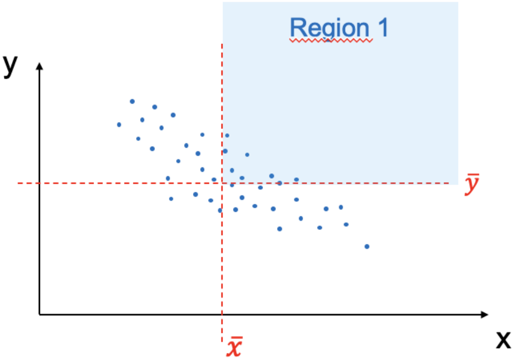

$$(x_i - \bar{x}) > 0 \quad \text{and} \quad (y_i - \bar{y}) < 0 \quad \Longrightarrow \quad (x_i - \bar{x})(y_i - \bar{y}) < 0$$

$$s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

- Summing all $(x_i - \bar{x})(y_i - \bar{y})$:
    - Regions 1 and 3: many points $\rightarrow$ large positive contributions
    - Regions 2 and 4: few points $\rightarrow$ small negative contributions
- Hence $s_{xy} > 0$.

# Understanding Covariance Formula: A Linear Decreasing Case

Use the same method.



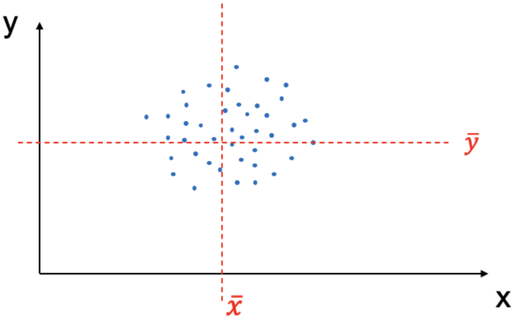

$$s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

$$s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

- Summing all $(x_i - \bar{x})(y_i - \bar{y})$:
  - Regions 1 and 3: few points $\rightarrow$ small positive contributions
  - Regions 2 and 4: many points $\rightarrow$ large negative contributions
- Hence $s_{xy} < 0$.

# Understanding Covariance Formula: A Case with No Relationship

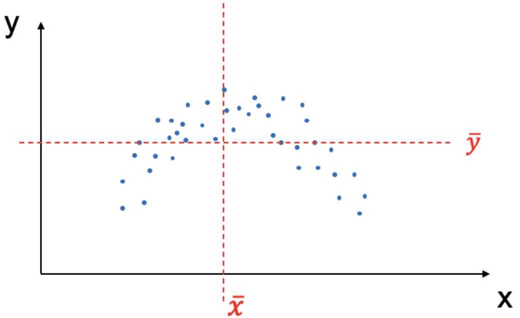
$$s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$




$$s_{xy} \approx 0$$

# Understanding Covariance Formula: A Non-Linear Case

$$s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$




$$s_{xy} \approx 0$$

Note: Although there is a nonlinear relationship, covariance cannot capture it. Covariance only measures the direction of linear relationships.

# Covariance

- $s_{xy} > 0$: a positive linear relationship
- $s_{xy} < 0$: a negative linear relationship
- $s_{xy} \approx 0$: little or no linear relationship
- Covariance is not designed to capture nonlinear relationships —other measures are needed for that (beyond the scope of MATH 201).

# Covariance

- Covariance $s_{xy}$ measures the direction of linear relationships.
- It does not measure strength.
- **Example.** The covariance of $(1, 7)$, $(2, 13)$, $(4, 9)$, $(5, 10)$, $(8, 12)$ is 2.75.
- **Example.** The covariance of $(1, 70)$, $(2, 130)$, $(4, 90)$, $(5, 100)$, $(8, 120)$ is 27.5.
- **Example.** The covariance of $(10, 70)$, $(20, 130)$, $(40, 90)$, $(50, 100)$, $(80, 120)$ is 275.
- Size of covariance depends on the units of $X$ and $Y$.

# Correlation Coefficient

To normalize covariance, we define the correlation coefficient

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

where

- $s_{xy}$ is the covariance between $X$ and $Y$
- $s_x$ is the standard deviation of $X$
- $s_y$ is the standard deviation of $Y$

# Example

- **Example.** Calculate the correlation coefficient of $(1,7)$, $(2,13)$, $(4,9)$, $(5,10)$, $(8,12)$

$$\bar{x} = 4, \qquad \bar{y} = 10.2$$

| $x_i$ | $y_i$ | $x_i - \bar{x}$ | $y_i - \bar{y}$ | $(x_i - \bar{x})(y_i - \bar{y})$ | $(x_i - \bar{x})^2$ | $(y_i - \bar{y})^2$ |
|---|---|---|---|---|---|---|
| 1 | 7 | $-3$ | $-3.2$ | 9.6 | 9 | 10.24 |
| 2 | 13 | $-2$ | 2.8 | $-5.6$ | 4 | 7.84 |
| 4 | 9 | 0 | $-1.2$ | 0 | 0 | 1.44 |
| 5 | 10 | 1 | $-0.2$ | $-0.2$ | 1 | 0.04 |
| 8 | 12 | 4 | 1.8 | 7.2 | 16 | 3.24 |
| | | | | 11 | 30 | 22.8 |

$$s_{xy} = \frac{11}{4} = 2.75, \quad s_x = \sqrt{\frac{30}{4}} = \sqrt{7.5} \approx 2.74, \quad s_y = \sqrt{\frac{22.8}{4}} = \sqrt{5.7} \approx 2.39$$

$$r_{xy} = \frac{s_{xy}}{s_x s_y} \approx 0.4206$$

# Correlation Coefficient

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

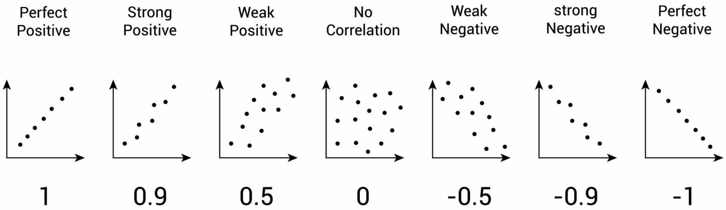
- Correlation coefficient measures the strength and direction of a linear relationship.
- Unit-free measure of linear association.
- Not designed to capture nonlinear relationships.

# Correlation Coefficient

$$-1 \leq r_{xy} \leq 1$$

- $r_{xy} \approx 1$: a strong positive linear relationship
- $r_{xy} \approx -1$: a strong negative linear relationship
- $r_{xy} \approx 0$: little or no linear relationship
- All the other values of $r_{xy}$ in between give us information about the direction and strength of the linear relationship between the two variables.

# Correlation Coefficient



## Correlation Coefficient

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{\frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}}{\sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}\sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}}$$

$$= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

# The Case of Perfect Linear Relation

Assume that $y_i = mx_i + c$ for some constants $m \neq 0$ and $c$. Then,

$$
\begin{aligned}
r_{xy} &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \\
&= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(mx_i + c - (m\bar{x} + c))}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(mx_i + c - (m\bar{x} + c))^2}} \\
&= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(mx_i - m\bar{x})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(mx_i - m\bar{x})^2}} \\
&= \frac{m\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})}{\sqrt{m^2}\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(x_i - \bar{x})^2}} \\
&= \frac{m}{\sqrt{m^2}} = \frac{m}{|m|} \\
&= \begin{cases} 1, & \text{if } m > 0 \\ -1 & \text{if } m < 0 \end{cases}
\end{aligned}
$$

# Correlation Does Not Imply Causation

- A high correlation does not prove that one variable causes the other.
- **Example.** Ice cream sales and shark attacks are positively correlated. (Both increase in the summer, but one does not cause the other.)

# Correlation does not imply causation!

cum hoc ergo propter hoc
('with this, therefore because of this')