Multivariate Analysis

Like father, like son.

1. THE SCATTER DIAGRAM

The methods discussed in part II are good for dealing with one variable at a time. Other methods are needed for studying the relationship between two vari- ables.[1] Sir Francis Galton (England, 1822–1911) made some progress on this front while he was thinking about the degree to which children resemble their parents. Statisticians in Victorian England were fascinated by the idea of quanti- fying hereditary influences and gathered huge amounts of data in pursuit of this goal. We are going to look at the results of a study carried out by Galton's disciple Karl Pearson (England, 1857–1936).[2]

As part of the study, Pearson measured the heights of 1,078 fathers, and their sons at maturity. A list of 1,078 pairs of heights would be hard to grasp. But the relationship between the two variables—father's height and son's height—can be brought out in a *scatter diagram* (figure 1 on the next page). Each dot on the diagram represents one father-son pair. The x -coordinate of the dot, measured along the horizontal axis, gives the height of the father. The y-coordinate of the dot, along the vertical axis, gives the height of the son.

Figure 1. Scatter diagram for heights of 1,078 fathers and sons. Shows positive association between son's height and father's height. Families where the height of the son equals the height of the father are plotted along the 45-degree line $y \quad x$. Families where the father is 72 inches tall (to the nearest inch) are plotted in the vertical strip.

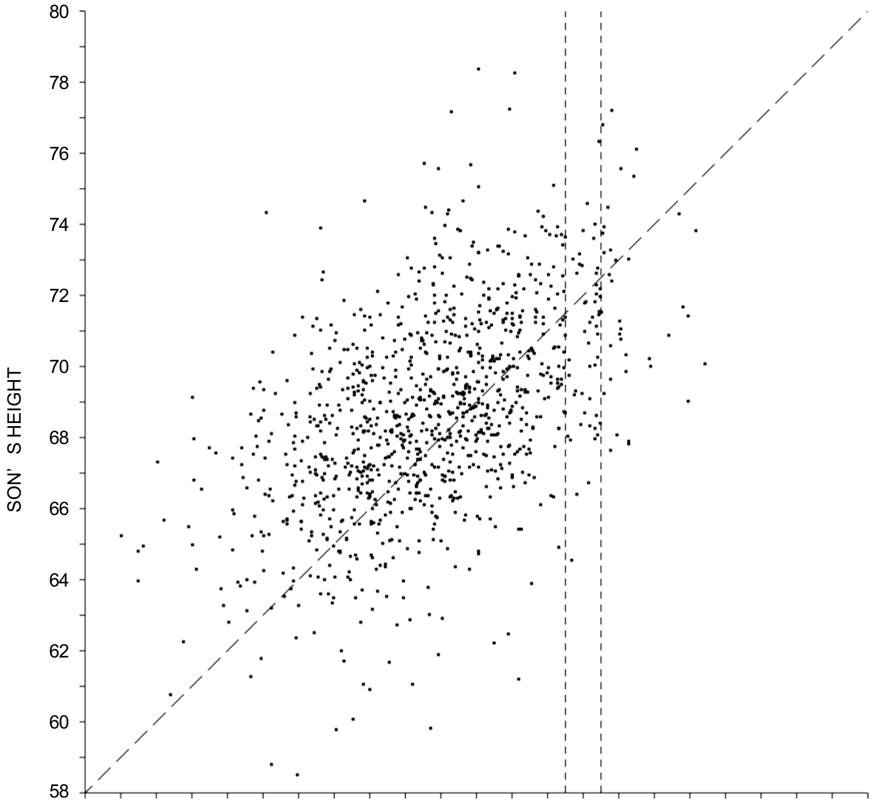


Figure 2a illustrates the mechanics of plotting scatter diagrams. (Chapter 7 has details.) The scatter diagram in figure 1 is a cloud shaped something like a football, with points straggling off the edges. When making a rough sketch of such a scatter diagram, it is only necessary to show the main oval portion—figure 2b.

The swarm of points in figure 1 slopes upward to the right, the *y*-coordinates of the points tending to increase with their *x* -coordinates. A statistician might say there is a *positive association* between the heights of fathers and sons. As a rule, the taller fathers have taller sons. This confirms the obvious. Now look at the 45- degree line in figure 1. This line corresponds to the families where son's height equals father's height. Along the line, for example, if the father is 72 inches tall then the son is 72 inches tall; if the father is 64 inches tall, the son is too; and so forth. Similarly, if a son's height is close to his father's height, then their point on the scatter diagram will be close to the line, like the points in figure 3.

Figure 2a. A point on a scatter diagram.

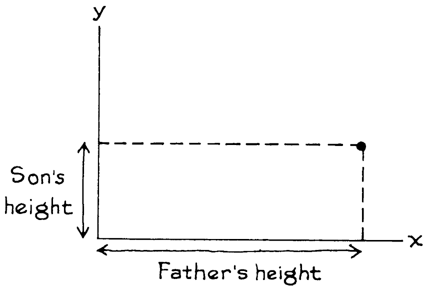
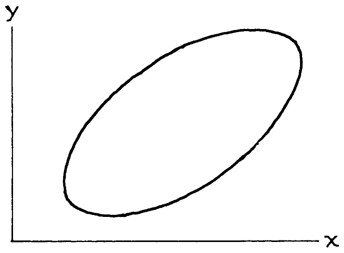


Figure 2b. Rough sketch.



There is a lot more spread around the 45-degree line in the actual scatter diagram than in figure 3. This spread shows the weakness of the relationship be- tween father's height and son's height. For instance, suppose you have to guess the height of a son. How much help does the father's height give you? In figure 1, the dots in the chimney represent all the father-son pairs where the father is 72 inches tall to the nearest inch (father's height between 71.5 inches and 72.5 inches, where the dashed vertical lines cross the *x* -axis). There is still a lot of variability in the heights of the sons, as indicated by the vertical scatter in the chimney. Even if you know the father's height, there is still a lot of room for error in trying to guess the height of his son.

> If there is a strong association between two variables, then knowing one helps a lot in predicting the other. But when there is a weak association, information about one variable does not help much in guessing the other.

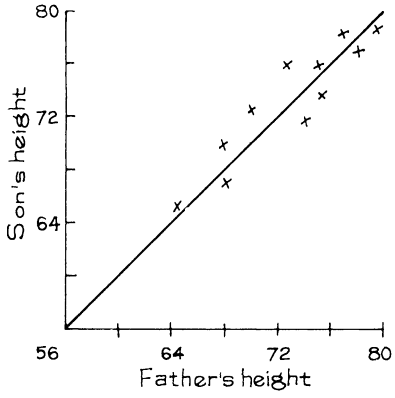


Figure 3. Son's height close to father's height.

Sir Francis Galton (England, 1822–1911)

Source: *Biometrika* (November, 1903).

In social science studies of the relationship between two variables, it is usual to label one as *independent* and the other as *dependent*. Ordinarily, the independent variable is thought to influence the dependent variable, rather than the other way around. In figure 1, father's height is taken as the independent variable and plotted along the *x* -axis: father's height influences son's height. However, there is nothing to stop an investigator from using son's height as the independent vari- able. This choice might be appropriate, for example, if the problem is to guess a father's height from his son's height.

Exercise Set A

1. Use figure 1 (p. 120) to answer the following questions:
   (a) What is the height of the shortest father? of his son?
   (b) What is the height of tallest father? of his son?
   (c) Take the families where the father was 72 inches tall, to the nearest inch. How tall was the tallest son? the shortest son?
   (d) How many families are there where the sons are more than 78 inches tall? How tall are the fathers?
   (e) Was the average height of the fathers around 64, 68, or 72 inches?
   (f) Was the SD of the fathers' heights around 3, 6, or 9 inches?

Covariance as an Extension of Variance

Variance is a statistical measure that quantifies how much the values of a single variable deviate from their mean. Mathematically, it is defined as the average of the squared deviations from the mean:

$$\sigma^2 = \frac{\sum (xi - \bar{x})^2}{N}$$

It captures the **spread** or **dispersion** of one variable but does not say anything about how two variables move together.

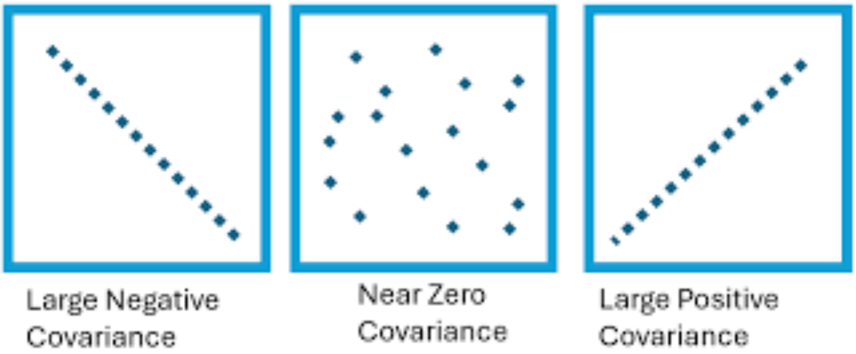
Covariance, on the other hand, generalizes the concept of variance to **two variables**. Instead of squaring the deviation of one variable, we take the product of deviations of two variables:

$$\boxed{COV(X,Y) = \frac{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{n-1}}$$

his definition shows that covariance is essentially **variance extended to two dimensions**:

- When X=Y = X=Y, covariance reduces to variance:

Cov(X,X)=Var(X)



- A **positive covariance** indicates that as XXX increases, YYY tends to increase.
- A **negative covariance** indicates that as XXX increases, YYY tends to decrease.
- A **zero covariance** suggests no linear relationship between the two variables.

Thus, while variance measures how a single variable spreads around its mean, covariance measures how **two variables co-vary**, i.e., how their fluctuations are related.

This idea forms the foundation for more advanced concepts like the **covariance matrix** (a generalization to multiple variables). The range of covariance is –∞ to +∞

Key Note

Covariance is **not standardized** — its value depends on the units of the variables.

- Example: If X is measured in meters and Y in kilograms, covariance will be in “meter–kilogram” units.
- This makes comparisons across different datasets difficult.

👉 To overcome this, we use **correlation**, which is a standardized version of covariance, always ranging between **–1 and +1**.

The co-variance values are standardised by the standard deviation of variables under test to get correlation.

$$\boxed{\rho_{XY} = \frac{Covariance(\boldsymbol{X},\boldsymbol{Y})}{\sigma_X \sigma_Y}}$$

Both **covariance** and **correlation** are essential tools in Machine Learning (ML), but they are used in slightly different contexts:

Covariance in Machine Learning

- **Feature Relationships**: Covariance is used to measure how two features vary together. For example, if “study time” and “exam scores” have a positive covariance, it indicates they increase together.
- **Covariance Matrix**: In ML, we often use the **covariance matrix** to understand relationships among multiple features.
    - This is crucial in **Principal Component Analysis (PCA)**, where eigenvalues and eigenvectors of the covariance matrix help reduce dimensionality while preserving maximum variance.
- **Portfolio Optimization & Finance ML Models**: In models predicting financial returns, covariance between assets is used to balance risk.

Correlation in Machine Learning

- **Feature Selection**: Correlation is often used to check for **multicollinearity** (when two features are highly correlated, e.g., height in cm vs. height in inches). Highly correlated features provide redundant information, which can hurt model performance.
- **Data Preprocessing**: Correlation helps identify which features are strongly related to the target variable, guiding feature engineering.

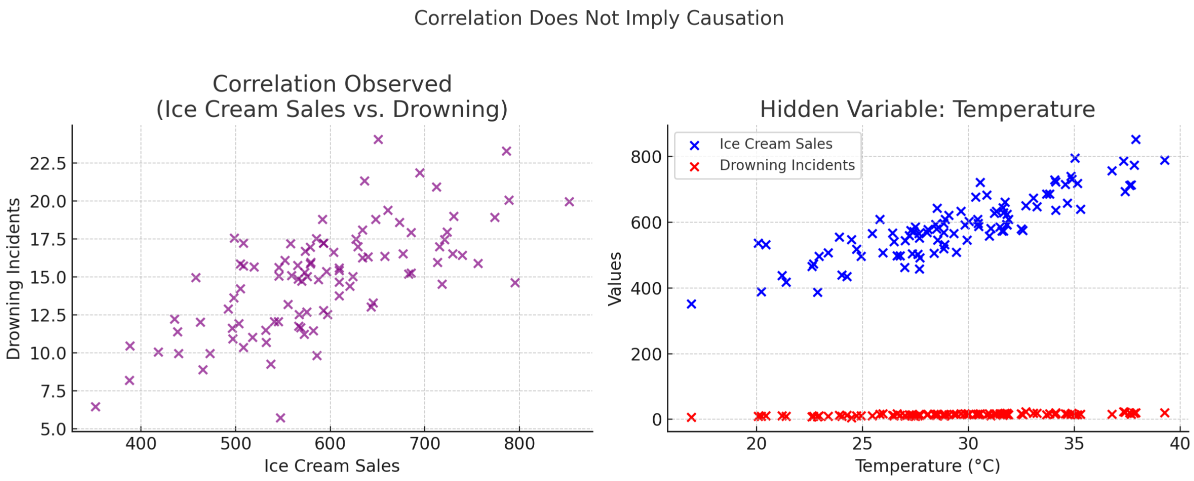
- **Interpretability**: Since correlation is standardized between –1 and +1, it provides an easy-to-interpret measure of the strength and direction of relationships.

Correlation and Causation..

Correlation measures how strongly two variables move together, but it does not explain **why** they move together. A common mistake in data analysis is to assume that if two variables are correlated, then one must cause the other. This is not necessarily true.

There are three main reasons why correlation may not mean causation:

1. **Hidden Variable (Confounder):** A third factor influences both variables.
2. **Reverse Causation:** The relationship may exist in the opposite direction than assumed.
3. **Coincidental Correlation:** Some correlations arise by chance with no meaningful link.



Here's a **real-world example image** showing why **correlation does not imply causation**:

- **Left plot (purple)**: Ice cream sales and drowning incidents appear positively correlated. One might wrongly conclude that ice cream causes drowning.
- **Right plot (blue & red)**: The hidden variable **temperature** explains both trends — hotter days increase ice cream sales **and** swimming activity, which can lead to more drownings.

This demonstrates that correlation may arise due to a **third factor** rather than a direct causal link.