Correlation

Correlation: the process of establishing a relationship or connection between two or more things.

Correlation in Statistics: This section shows how to calculate and interpret correlation coefficients for ordinal and interval level scales. Methods of correlation summarize the relationship between two variables in a single number called the correlation coefficient. The correlation coefficient is usually represented using the symbol r, and it ranges from -1 to +1.

A correlation coefficient quite close to 0, but either positive or negative, implies little or no relationship between the two variables. A correlation coefficient close to plus 1 means a positive relationship between the two variables, with increases in one of the variables being associated with increases in the other variable.

A correlation coefficient close to -1 indicates a negative relationship between two variables, with an increase in one of the variables being associated with a decrease in the other variable. A correlation coefficient can be produced for ordinal, interval or ratio level variables, but has little meaning for variables which are measured on a scale which is no more than nominal

For ordinal scales, the correlation coefficient can be calculated by using Spearman's rho. For interval or ratio level scales, the most commonly used correlation coefficient is Pearson's r, ordinarily referred to as simply the correlation coefficient.

Statistical Measures of Correlation:

- Pearson correlation coefficient.
- Spearman's rank correlation.
- Kendall's tau correlation.
- Point-biserial correlation.

What Does Correlation Measure?

In statistics, Correlation studies and measures the direction and extent of relationship among variables, so the correlation measures co-variation, not causation. Therefore, we should never interpret correlation as implying cause and effect relation. For example, there exists a correlation between two variables X and Y, which means the value of one variable is found to change in one direction, the value of the other variable is found to change either in the same direction (i.e. positive change) or in the opposite direction (i.e. negative change). Furthermore, if the correlation exists, it is linear, i.e. we can represent the relative movement of the two variables by drawing a straight line on graph paper.

Co-variation (Correlation): Co-variation, or correlation, refers to a statistical measure that indicates the extent to which two variables change together. It does not imply any causal relationship between the variables.

Example: There might be a positive correlation between ice cream sales and temperatures. As temperatures increase, ice cream sales also increase. However, this does not mean that higher temperatures cause people to buy ice cream.

Causation: Causation indicates a relationship where one variable directly affects another. This means that changes in one variable are responsible for changes in another.

Example: Smoking causes lung cancer. Here, smoking is a direct cause of lung cancer, implying a clear directional influence.

Nature of Relationship:

- **Co-variation**: Observes association or co-occurrence without implying any causal link.
- **Causation**: Establishes a direct cause-and-effect relationship.

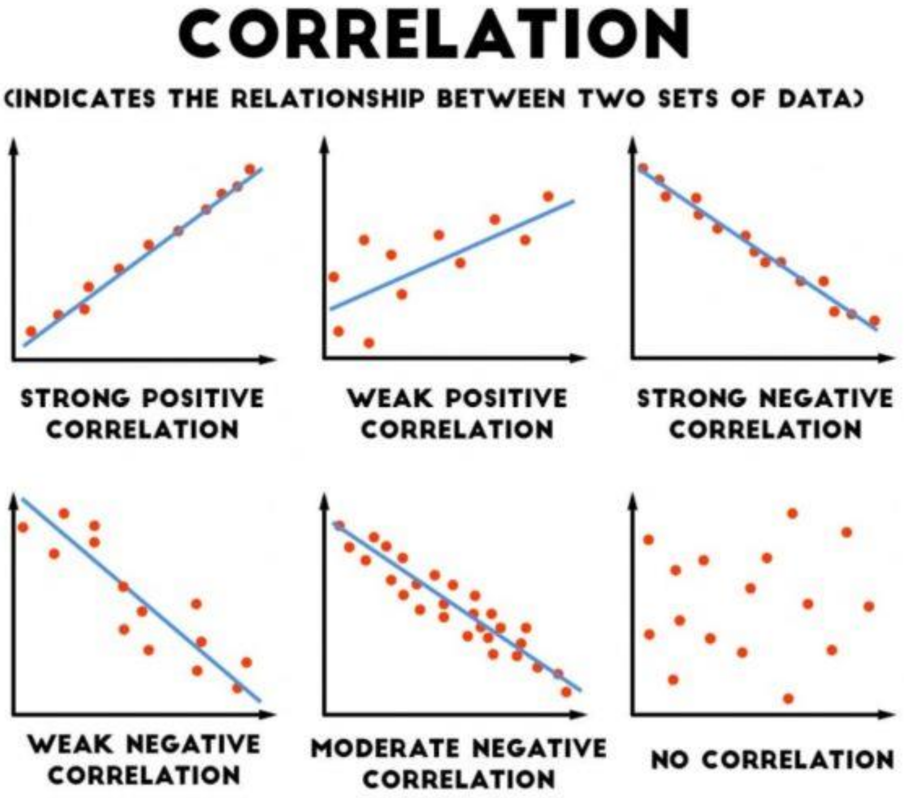
Determination:

- **Co-variation**: Can be identified using statistical methods like correlation coefficients.
- **Causation**: Requires more rigorous methods like controlled experiments, longitudinal studies, or statistical techniques like regression analysis that account for confounding variables.

Correlation Coefficient

The correlation coefficient, r, is a summary measure that describes the extent of the statistical relationship between two interval or ratio level variables. The correlation coefficient is scaled so that it is always between -1 and +1. When r is close to 0 this means that there is little relationship between the variables and the farther away from 0 r is, in either the positive or negative direction, the greater the relationship between the two variables.

The two variables are often given the symbols X and Y. In order to illustrate how the two variables are related, the values of X and Y are pictured by drawing the scatter diagram, graphing combinations of the two variables. The scatter diagram is given first, and then the method of determining Pearson's r is presented. From the following examples, relatively small sample sizes are given. Later, data from larger samples are given.



Types of Correlation

The scatter plot explains the correlation between the two attributes or variables. It represents how closely the two variables are connected. There can be three such situations to see the relation between the two variables –

- Positive Correlation – when the values of the two variables move in the same direction so that an increase/decrease in the value of one variable is followed by an increase/decrease in the value of the other variable.
- Negative Correlation – when the values of the two variables move in the opposite direction so that an increase/decrease in the value of one variable is followed by decrease/increase in the value of the other variable.
- No Correlation – when there is no linear dependence or no relation between the two variables.

Applications of Correlation in Machine Learning:

1) Feature Selection

Feature selection is a critical step in building machine learning models. Correlation analysis helps in identifying and selecting the most relevant features.

Details:

- **Highly Correlated Features**: Features that are highly correlated with the target variable are usually more important for prediction.
- **Redundant Features**: Features that are highly correlated with each other might be redundant. Removing one can reduce multicollinearity and simplify the model.

Example: In a dataset predicting house prices, features like the number of bedrooms and square footage may be highly correlated with the price. By analyzing correlations, redundant features can be identified and removed.

2) Multicollinearity in Regression Models

Multicollinearity occurs when independent variables in a regression model are highly correlated, which can lead to unreliable estimates of coefficients.

Details:

- **Detection**: Correlation matrices and Variance Inflation Factor (VIF) are used to detect multicollinearity.

- **Solution**: Dropping or combining correlated features to improve model stability.

Example: In a linear regression model predicting car prices, features like engine size and horsepower might be highly correlated. Identifying and addressing this can improve model performance.

Evaluating Model Performance

Correlation is used to evaluate the performance of machine learning models, especially in regression tasks.

Details:

- **Correlation Coefficient**: The correlation coefficient between predicted and actual values indicates the strength and direction of the relationship.
- **Residual Analysis**: Analyzing the correlation between residuals and predicted values helps in diagnosing model issues.

Example: In predicting stock prices, a high correlation between predicted and actual prices suggests a good model fit.

In machine learning, most of the relationships identified between variables are typically correlations rather than causations …. Why?

- **Observational Data**: Most machine learning models are trained on observational data, which inherently reflects correlations. Observational data lacks the controlled conditions required to infer causation.

- **Complexity of Causal Inference**: Determining causation is complex and requires techniques such as randomized controlled trials (RCTs), natural experiments, or advanced statistical methods like instrumental variables and causal graphs (e.g., using directed acyclic graphs, or DAGs).

Predictive Focus: Machine learning models are often used for prediction rather than understanding the underlying causal mechanisms. For prediction, knowing that two variables are correlated can be sufficient.

In summary, machine learning models predominantly deal with correlations because they are typically based on observational data, and establishing causation requires more rigorous methods. While correlations can be powerful for prediction, distinguishing between correlation and causation is crucial for accurate interpretation and decision-making. Advanced techniques in causal inference can help bridge this gap when causal relationships are of interest.