Correlation and Regression

Adejumo Ridwan Suleiman

Correlation

What is Correlation?

- Measures the strength and direction of relationship between two variables.
- Does not imply causation.
- Change in a variable influences another variable

Types of Correlation

- **Pearson Correlation** ()
- **Spearman Rank Correlation** ()
- Kendall's Tau Correlation ρ
- Point-Biserial Correlation

Pearson and Spearman

Feature	Pearson Correlation	Spearman Correlation
Type of relationship	Linear	Monotonic (increasing or decreasing)
Data Type	Interval or ratio, normally distributed	Ordinal, interval, or ratio; non-normal distribution is fine
Outlier Sensitivity	Sensitive to outliers	Less sensitive to outliers

Measurement of Correlation

- Ranges from **-1** to **+1**
  - **+1** indicates a perfect positive correlation
  - **-1** indicates a perfect negative correlation
  - **0** indicates no correlation

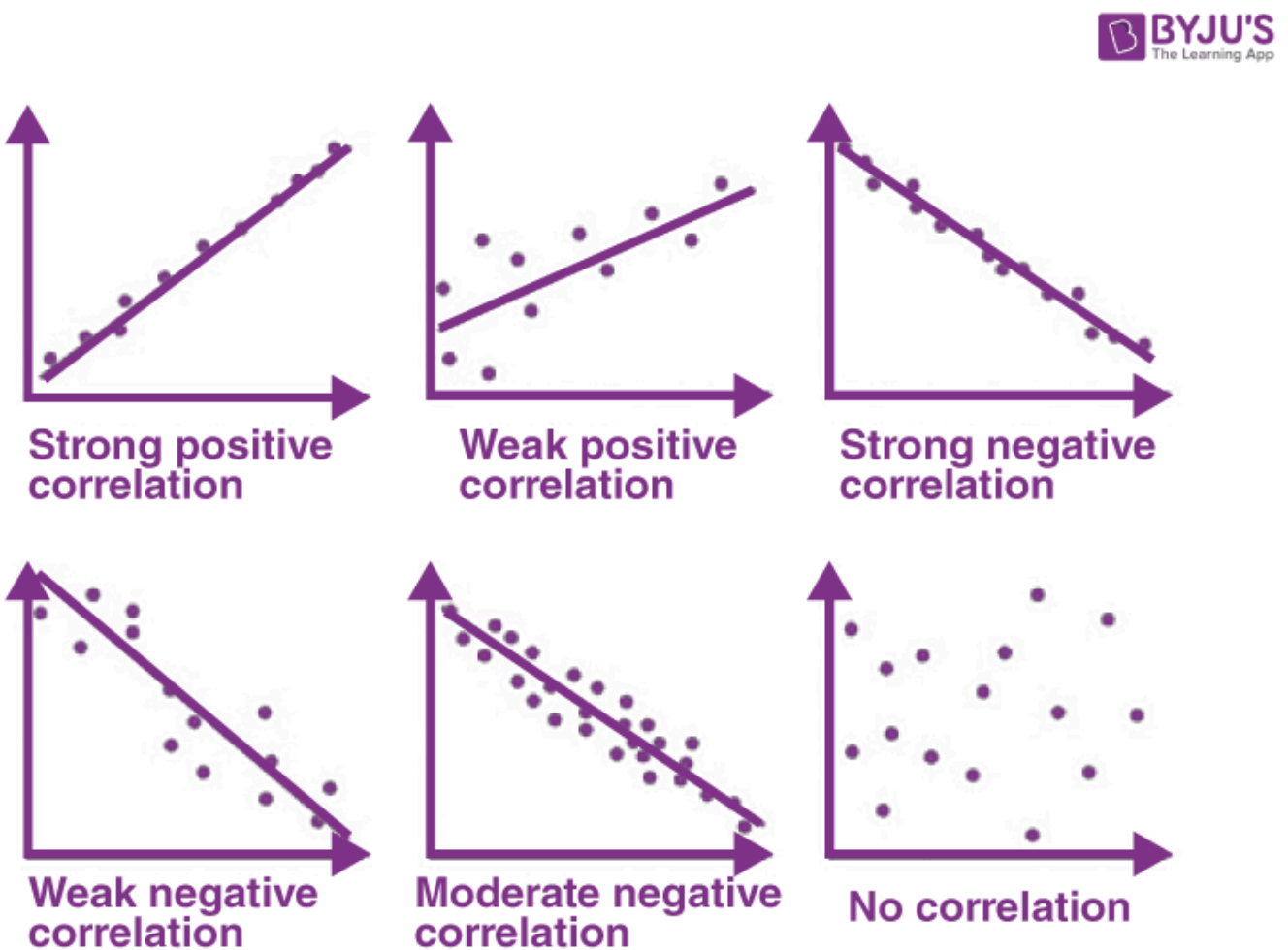
Measurement of correlation

Correlation Coefficient	Interpretation
-1	Very strong negative correlation
Between -1 and <= -0.6	Strong negative correlation
Between > -0.6 and <= -0.4	Moderate negative correlation
Between > -0.4 and < 0	Weak negative correlation

Measurement of correlation

Correlation Coefficient	Interpretation
0	No correlation
Between 0 and < 0.4	Weak positive correlation
Between >= 0.4 and < 0.6	Moderate positive correlation
Between >= 0.6 and < 1	Strong positive correlation
1	Very strong positive correlation

Visualizing Correlation



Source. Byju

Examples

- Correlation between hypertension and heart disease is **0.7.**
- Correlation between obesity and lung cancer is **0.2.**

Regression

What is regression?

- Understand relationships between variables and make predictions.
- Modelling the relationship between one or more independent variable and an outcome variable.
- Estimate how changes in predictors impact the dependent variable.

Simple Linear Regression

- Only one independent variable and one dependent variable.
- Simple Linear Regression equation is given as:

$$y = \beta_0 + \beta_1 x_1 + \epsilon$$

Where

- - dependent variable
- y - independent variable
- x - the intercept, representing the expected value of when
- β_0 - the slope, representing the change in for a unit change in
- β_1 - the error term capturing the difference between predicted
6 and actual values of

y $x = 0$

y

x.

y.

Multiple Linear Regression

- Extends SLR by allowing more than one independent variable.
- The multiple linear regression equation is given as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n$$

Where

- y - the dependent variable
- $x_1, x_2, \ldots, x_n$ - the independent variables
- β_0 - the intercept
- $\beta_1, \beta_2, \ldots, \beta_n$ - The coefficients representing the effect of each independent variable on y.
- ϵ: The error term

Assumptions of Multiple Linear Regression

- Linearity of the relationship
- Independence of Errors
- Homoscedasticity of variance (Constant Variance of Errors)
- Normality of Errors
- No perfect of multicollinearity

Interpretation of MLR

Coefficients

- Represent the change in the dependent variable for a one-unit change in the corresponding predictor variable.
- The sign indicates direction of the relationship.

β_0

- Represent expected value of Y when all predictor variables are zero. X_1, X

Coefficient of determination

- Measures variation in the dependent variable accounted by independent variables.
- Ranges between to
  - : perfect fit
  - : explains none of the variability
  - : explains that **75%** of the variation in the dependent variable is explained by the model.

R^2

$0 \quad 1$

$R^2 = 1$

$R^2 = 0$

$R^2 = 0.75$

Logistic Regression

What is logistic regression?

- Unlike regression that predicts continuous outcome
- Logistic regression is designed for categorical outcomes
- Example
  - What are the predictors of **heart disease(yes/no)**

Assumptions of Logistic Regression

- Binary outcome
- Linear relationship
- Independence of observations
- Absence of Multicollinearity

Coefficients

- Each coefficient represents the effect of a predictor on the log odds of the outcome.
- Log odds is the

Log-Odds and Odds

- **Logistic regression** doesn't predict probabilities directly, it first predicts **log odds.**
- **Log odds** is transformation of probabilities that makes the relationship between features and the outcome roughly linear.
- **Odds** is the ratio of probability of an event happening versus not happening.
- **Coefficients** are converted to **log-odds**s by taking their exponential.

Interpreting Effect of a Predictor

- A feature effect is interpreted through the log-odds.
- If **odds ratio > 1:** outcome increases by 1
- If **odds ratio < 1:** outcome decreases by 1
- If **odds ratio = 1:** Feature has no effect on the odds of the outcome.