

STATISTICAL FOUNDATIONS FOR MACHINE LEARNING

Hypothesis Testing and Regression Analysis

Course: Foundations of Data Analysis (Group A)

Trainer: Victor Lumumba Wandera

Date: 5th February 2026

Dataset: Hypertension Clinical Cohort (htn_dat.csv)

Foundations of Data Analysis – Group A
Bridge Module to Week 3 Machine Learning

Why Statistics Before Machine Learning?

Machine Learning is applied statistics at scale.

- ❖ Key statistical foundations:
- ❖ Hypothesis testing for decision-making
- ❖ Regression for prediction and inference
- ❖ p-values and confidence intervals for uncertainty

Healthcare relevance in Kenya:

- ❖ Logistic regression for hypertension risk prediction
- ❖ Survival analysis for HIV treatment outcomes
- ❖ Statistics ensures transparent and responsible AI

Key message:

- ❖ Machine learning without statistics becomes a black box.

Learning Objectives

By the end of this session, you will be able to:

- Formulate and test statistical hypotheses
- Interpret regression coefficients correctly
- Apply OLS regression for continuous outcomes
- Apply logistic regression for binary outcomes
- Evaluate models using R^2 , accuracy, and AUC
- Connect regression concepts to ML algorithms

Dataset Overview

Dataset: htn_dat.csv

Records: 4,900 patients from Kenyan health facilities

Outcome variables:

- ❑ SBP (continuous systolic blood pressure)
- ❑ SBP_ge120 (binary hypertension indicator)

Predictor variables:

- ❖ Age, BMI, DBP
- ❖ Gender, marital status
- ❖ Urban clinic indicator
- ❖ HIV and ART status

Research question:

- ❖ Which factors significantly predict hypertension risk?

What is Hypothesis Testing?

Hypothesis testing evaluates claims about population parameters.

Null hypothesis (H_0):

❖ No effect or no relationship

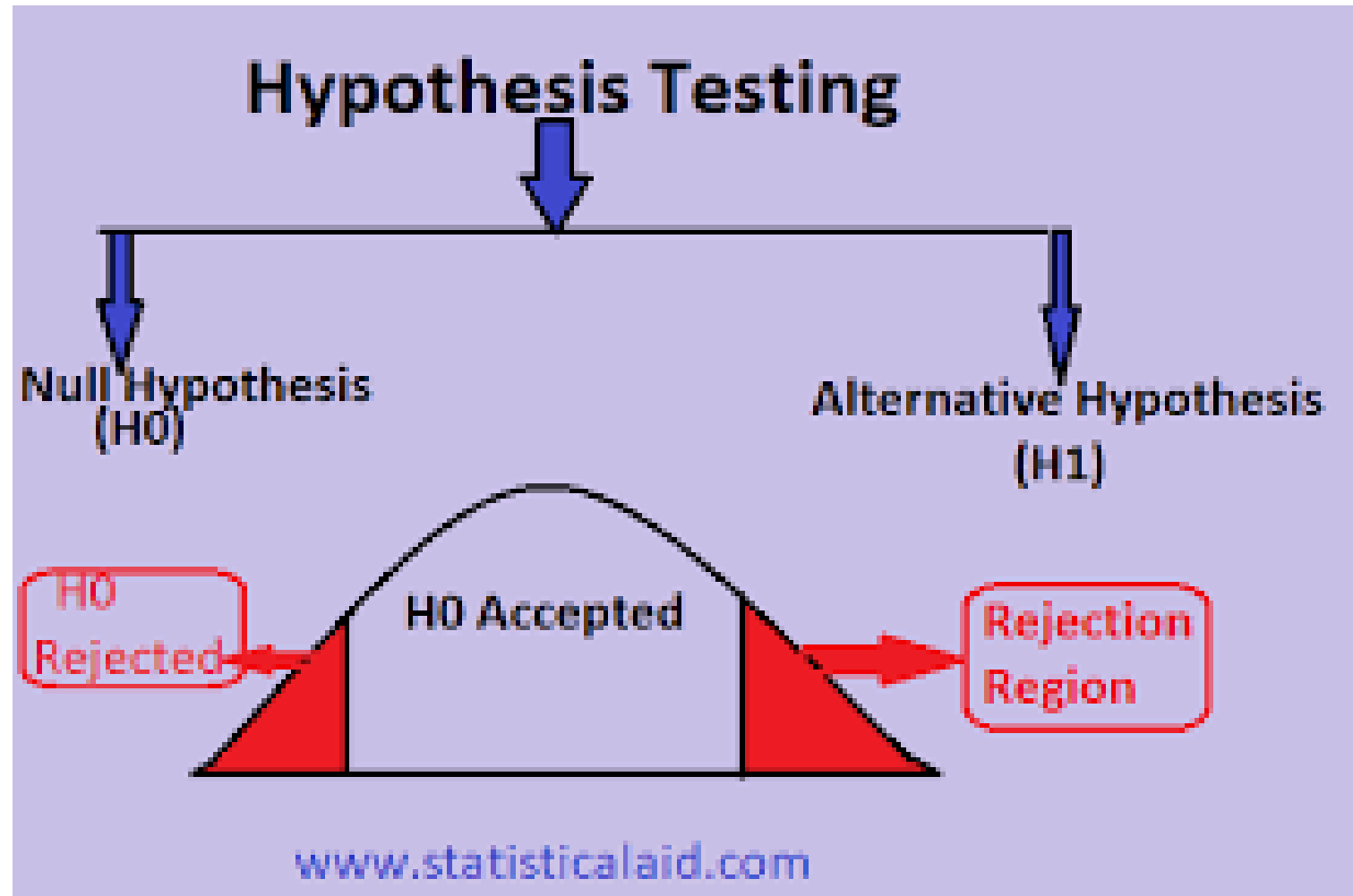
Alternative hypothesis (H_1):

❖ An effect or relationship exists

Example:

➤ H_0 : BMI is not associated with hypertension

➤ H_1 : BMI is associated with hypertension



Decision Rule in Hypothesis Testing

Decision based on p-value:

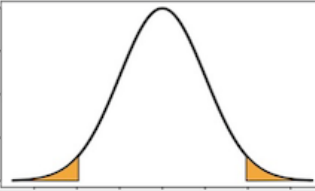
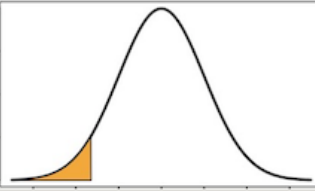
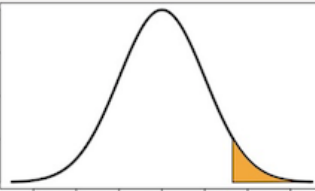
❖ $p < 0.05 \rightarrow$ Reject H_0

❖ $p \geq 0.05 \rightarrow$ Fail to reject H_0

Important principle:

❖ We do not "accept" the null hypothesis.

❖ We only reject or fail to reject it.

Hypothesis	Decision Rule	
$H_0 : \mu = \mu_0$	if $ t^* \leq t_{1-\alpha/2, r-1}$ Fail to reject H_0	
$H_a : \mu \neq \mu_0$	if $ t^* > t_{1-\alpha/2, r-1}$ Reject H_0 & accept H_a	
$H_0 : \mu \geq \mu_0$	if $t^* \geq t_{\alpha, r-1}$ Fail to reject H_0	
$H_a : \mu < \mu_0$	if $t^* < t_{\alpha, r-1}$ Reject H_0 & accept H_a	
$H_0 : \mu \leq \mu_0$	if $t^* \leq t_{\alpha, r-1}$ Fail to reject H_0	
$H_a : \mu > \mu_0$	if $t^* > t_{\alpha, r-1}$ Reject H_0 & accept H_a	

Understanding p-values

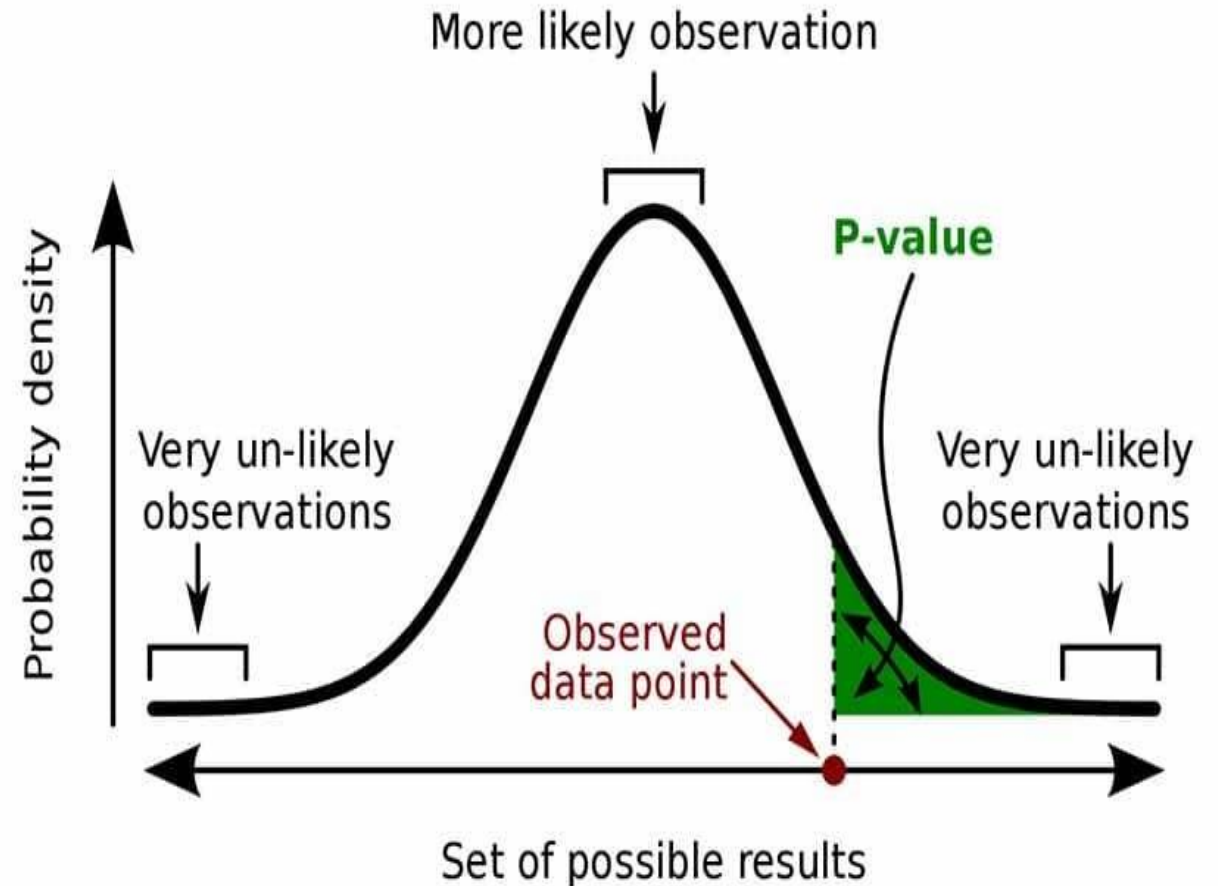
A p-value measures evidence against the null hypothesis.

Definition:

- ❖ Probability of observing the data (or more extreme)
- ❖ assuming the null hypothesis is true.

Interpretation:

- ❖ Small p-value → Strong evidence against H_0
- ❖ Large p-value → Weak evidence against H_0



A **p-value** (shaded green area) is the probability of an observed (or more extreme) result assuming that the null hypothesis is true.

Medical Example of a p-value

Hypothesis:

❖ H_0 : Mean SBP is equal in urban and rural clinics

❖ H_1 : Mean SBP differs between clinics

Result:

❖ $p = 0.003$

Interpretation:

❖ There is strong evidence that clinic location is associated with systolic blood pressure.

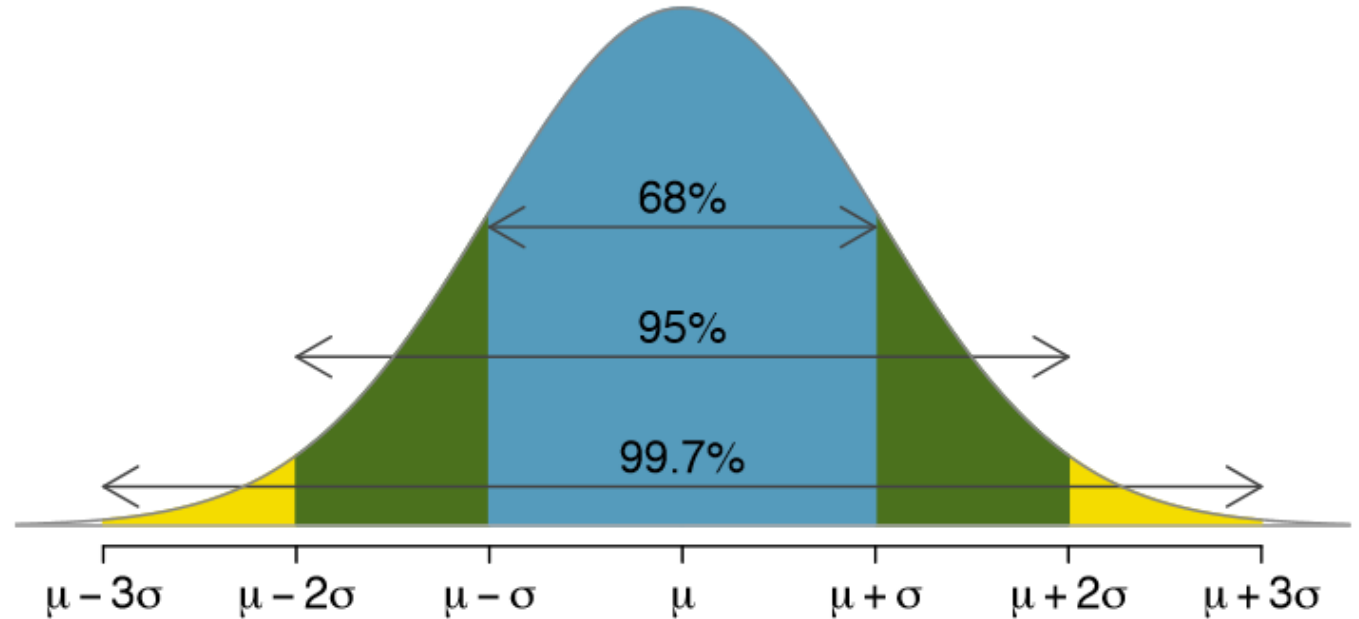
Confidence Interval

A confidence interval provides a range of plausible values for a population parameter.

95% confidence interval:
If the study were repeated many times, 95% of such intervals would contain the true value.

Confidence intervals show:

- Effect size
- Precision of estimates



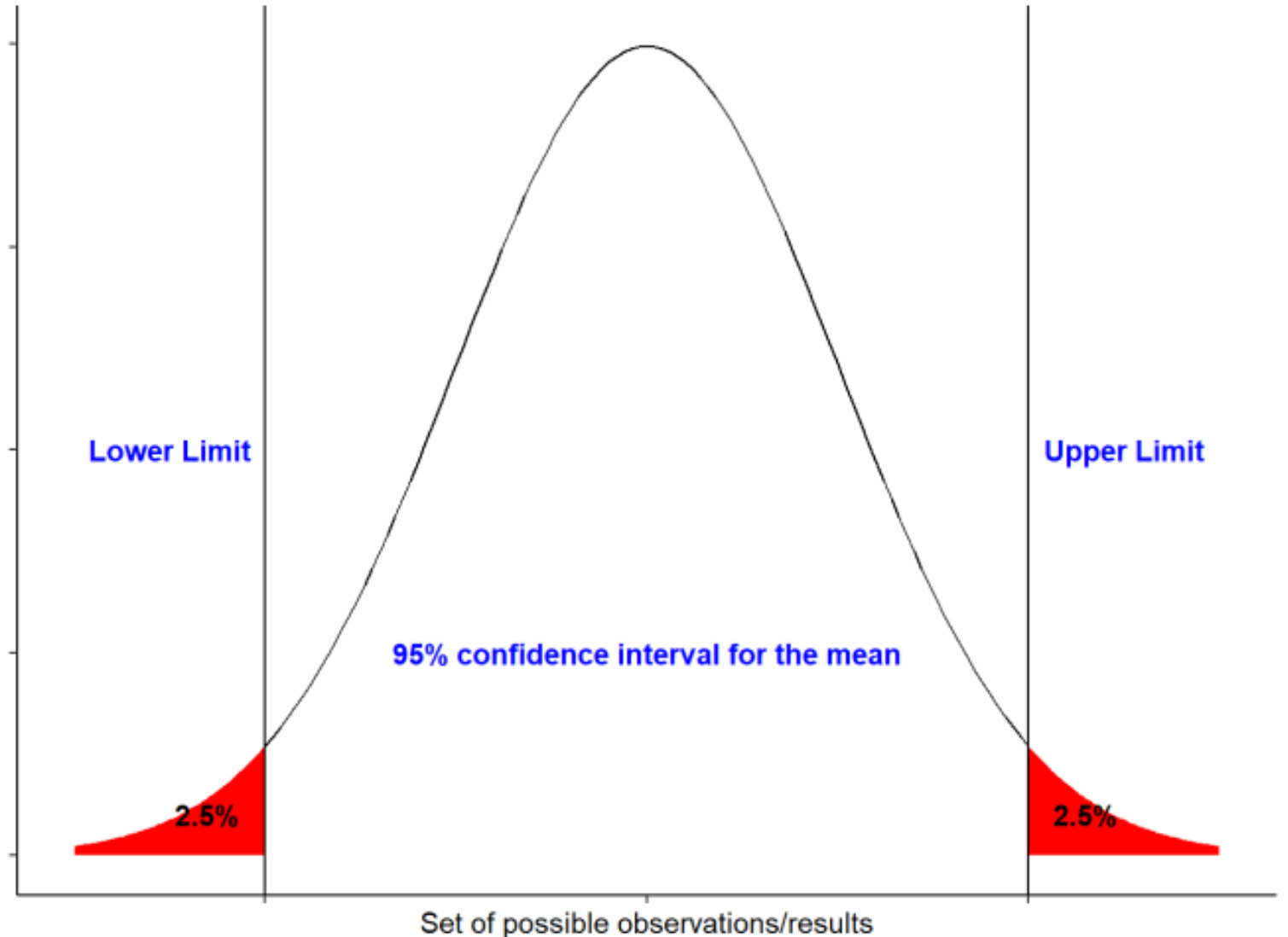
Confidence Interval Example

Mean SBP difference
(urban – rural) = 4.2
mmHg

- 95% CI: [2.1, 6.3] mmHg

Interpretation:

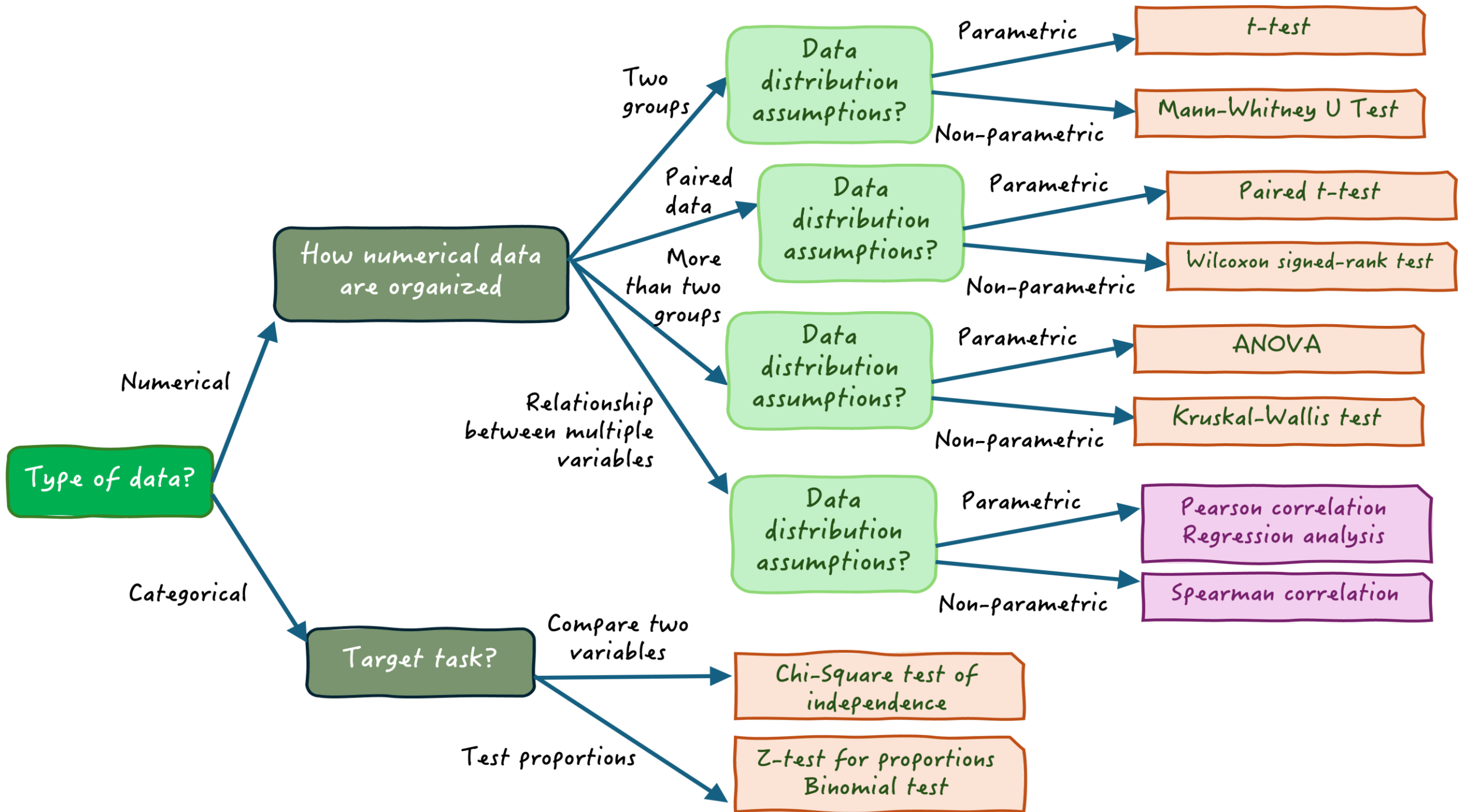
- The true SBP difference lies within this range
- CI does not include zero → statistically significant
- Provides more information than p-value alone



Choosing the Right Statistical Test

Test selection depends on variable types:

- Two means → Independent t-test
- More than two means → ANOVA
- Two categorical variables → Chi-square test
- Two continuous variables → Correlation
- Continuous outcome → OLS regression
- Binary outcome → Logistic regression



Introduction to Regression Analysis

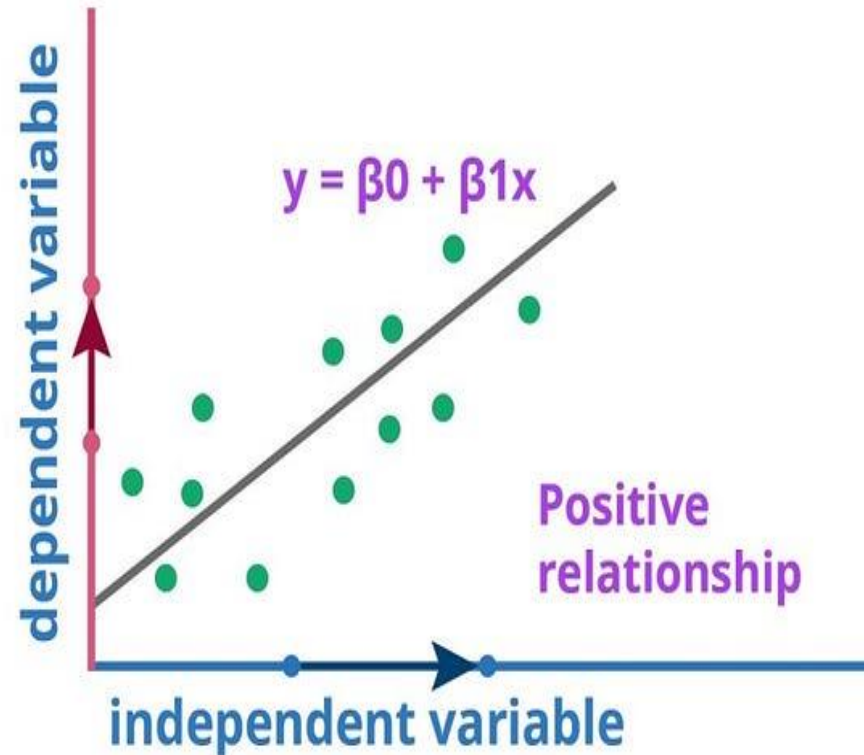
Regression models the relationship between predictors (X) and an outcome variable (Y).

General form:

- $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \varepsilon$

Regression is the foundation of many ML algorithms.

Linear Regression Model



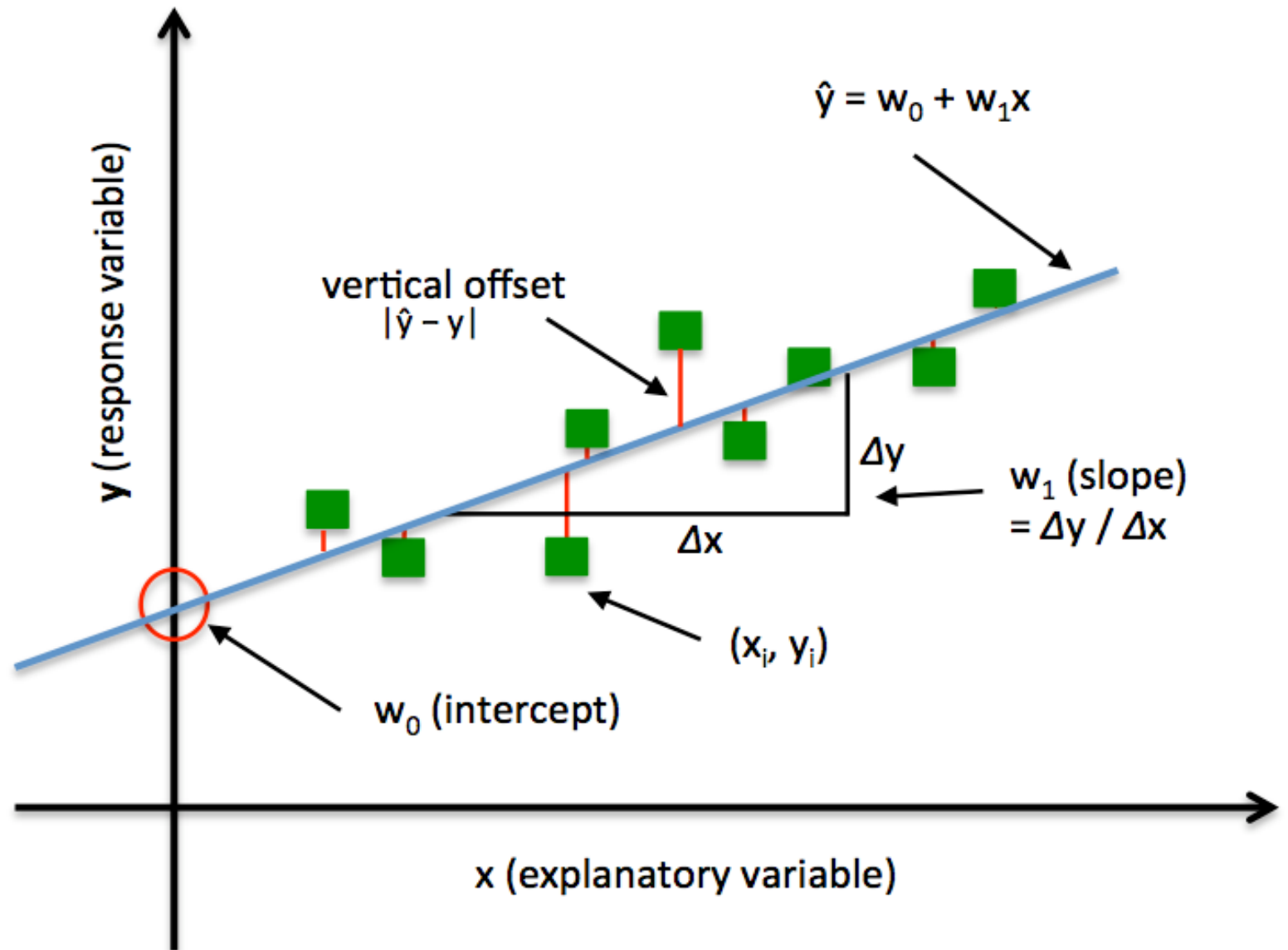
Ordinary Least Squares (OLS) Regression

OLS regression predicts continuous outcomes.

Medical application:
Predict systolic blood pressure using:

- BMI
- Age
- Gender

OLS minimizes the sum of squared prediction errors.



Interpreting OLS Coefficients

Coefficient interpretation:

- ❖ Sign (+/−): Direction of relationship
- ❖ Magnitude: Size of effect
- ❖ p-value: Statistical significance

Example:

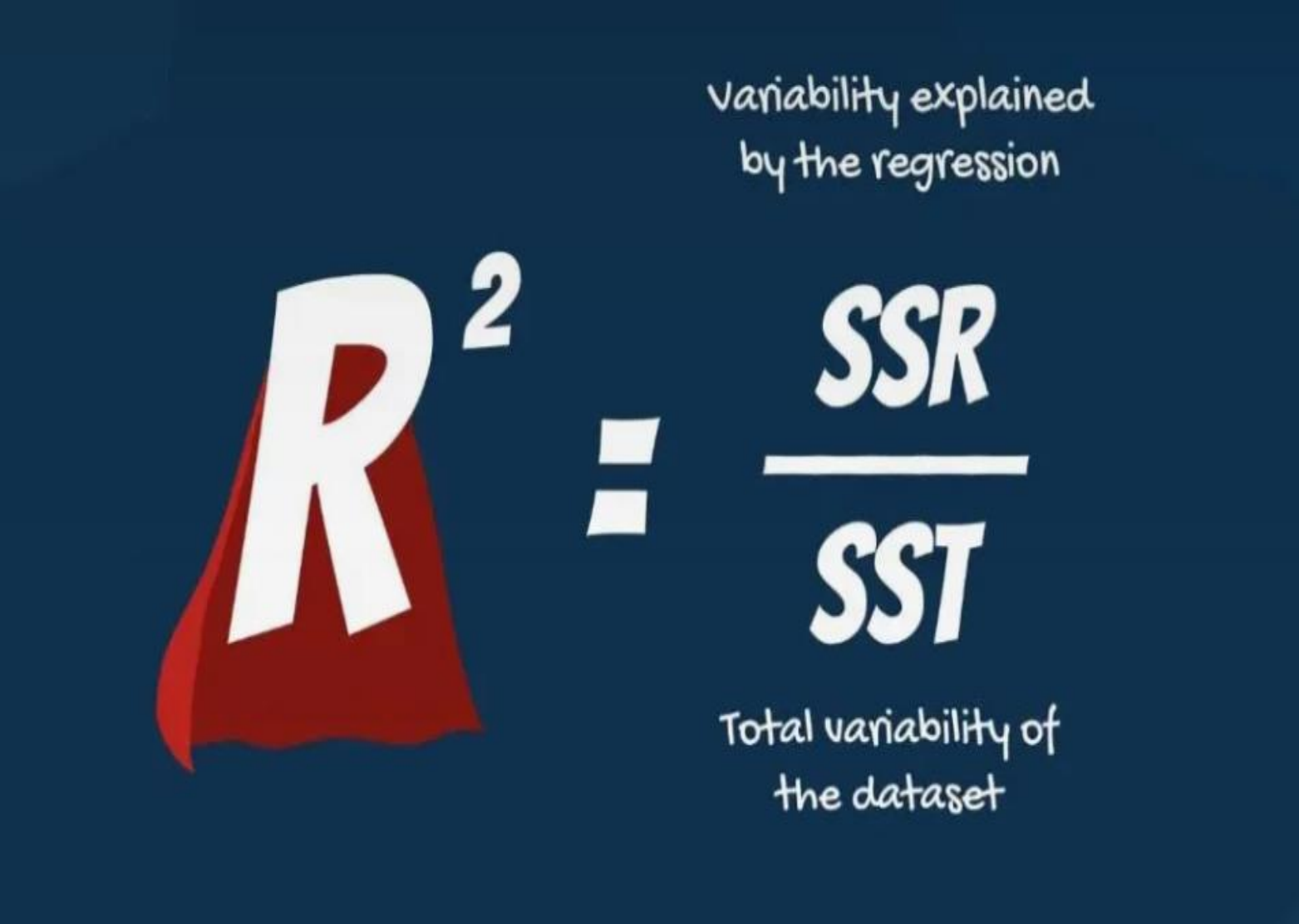
- ❖ A BMI coefficient of 0.87 means SBP increases by 0.87 mmHg per unit BMI increase, holding other variables constant.

Model Fit in OLS Regression

Key evaluation metrics:

- ❖ R^2 : Proportion of variance explained
- ❖ RMSE: Prediction error magnitude
- ❖ MAE: Average absolute error

R^2 does not imply causation, only explanatory power.


$$R^2 = \frac{SSR}{SST}$$

variability explained by the regression

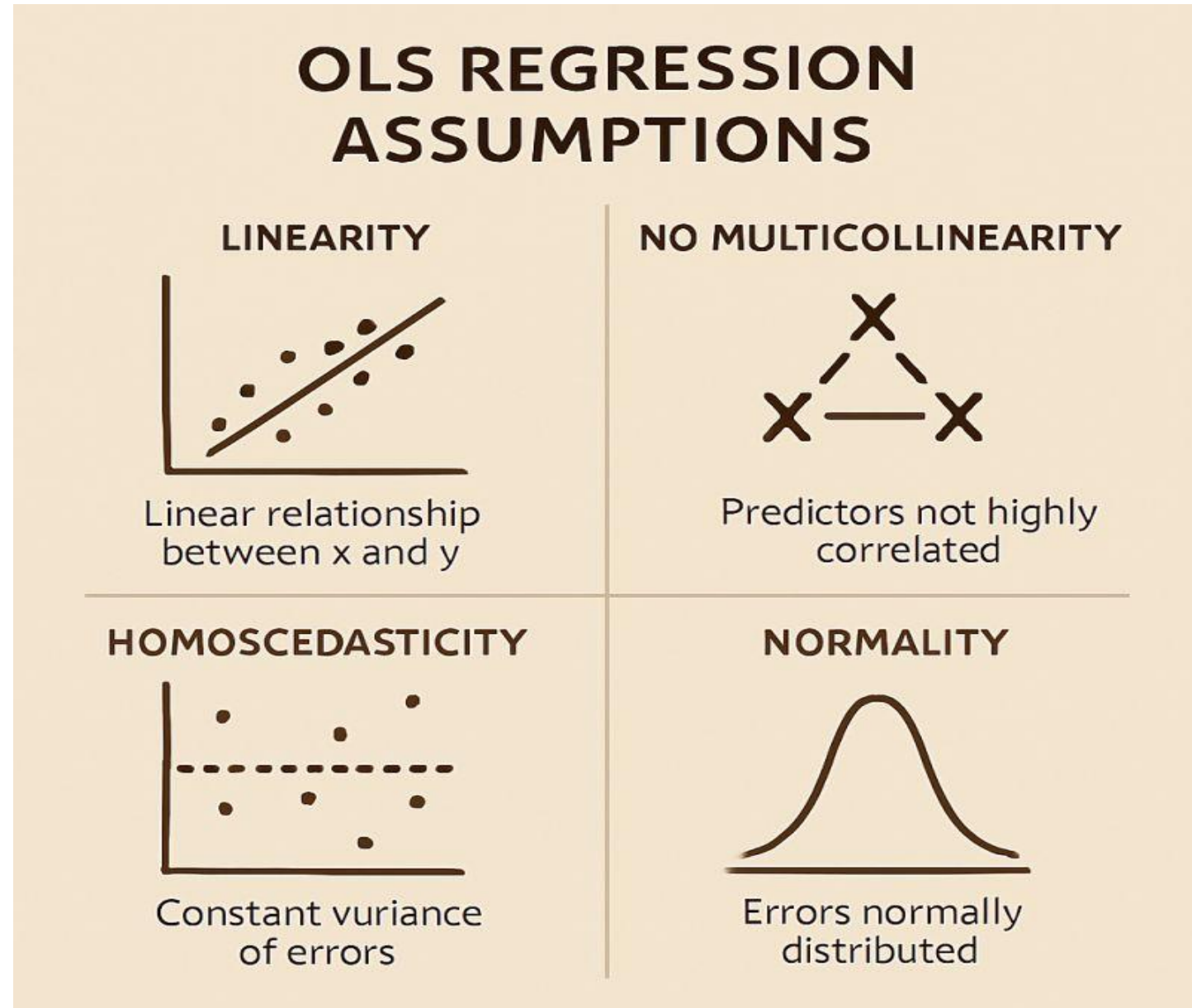
Total variability of the dataset

OLS Assumptions

OLS regression assumptions:

- 1. Linearity
- 2. Independence of observations
- 3. Homoscedasticity
- 4. Normality of residuals
- 5. No multicollinearity

Violations require transformations or alternative modeling approaches.



Logistic Regression Analysis