STATISTICAL FOUNDATIONS FOR MACHINE LEARNING

Hypothesis Testing and Regression Analysis

Course: Foundations of Data Analysis (Group A)

Mediacrest Training College

Trainer: Victor Lumumba Wandera

Date: 9^{th} Feb 2026-11^{th} Feb 2026

Dataset: Hypertension Clinical Cohort (htn_dat.csv)

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.

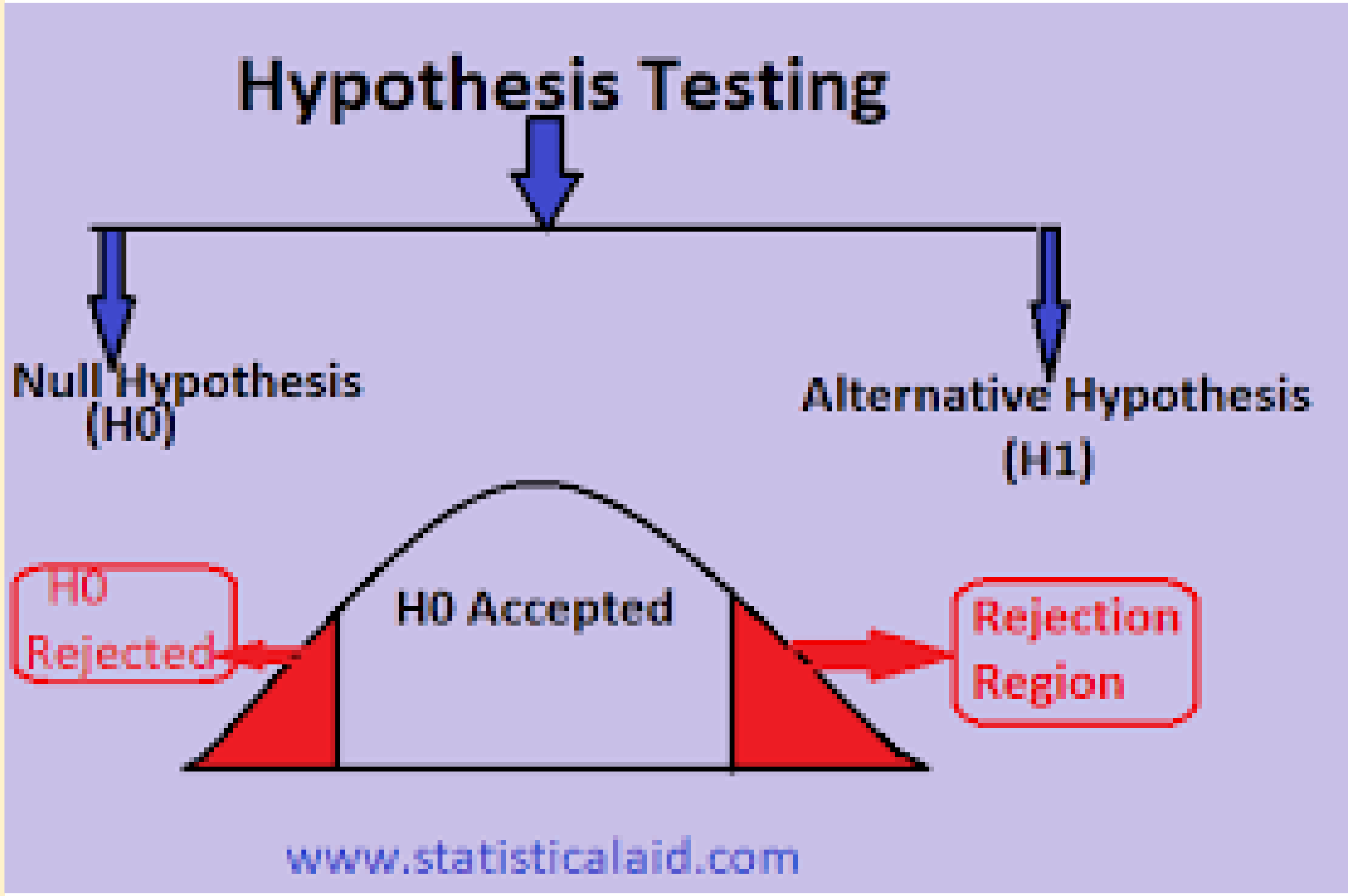
Null hypothesis (H_0):
❖No effect or no relationship

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



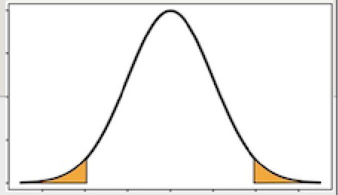
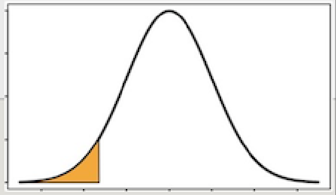
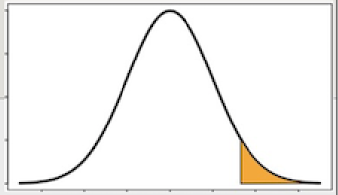
Decision Rule in Hypothesis Testing

Decision based on p-value:

- ❖ $p < 0.05 \rightarrow$ Reject H_0
- ❖ $p \geq 0.05 \rightarrow$ Fail to reject H_0

Important principle:

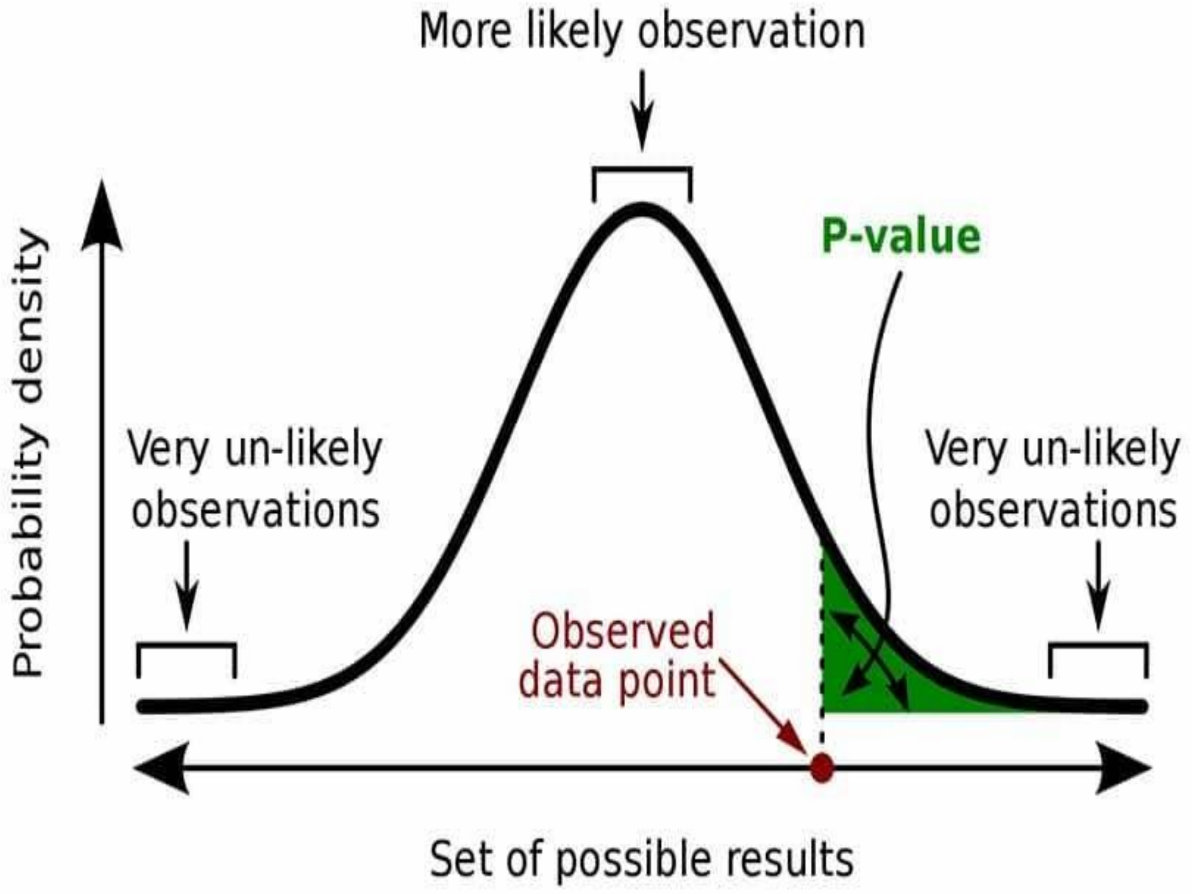
- ❖ We do not "accept" the null hypothesis.
- ❖ We only reject or fail to reject it.

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

- $p = 0.003$
- p-value is compared with alpha (5% (0.05), 10%(0.1), 1%(0.01)

Interpretation:

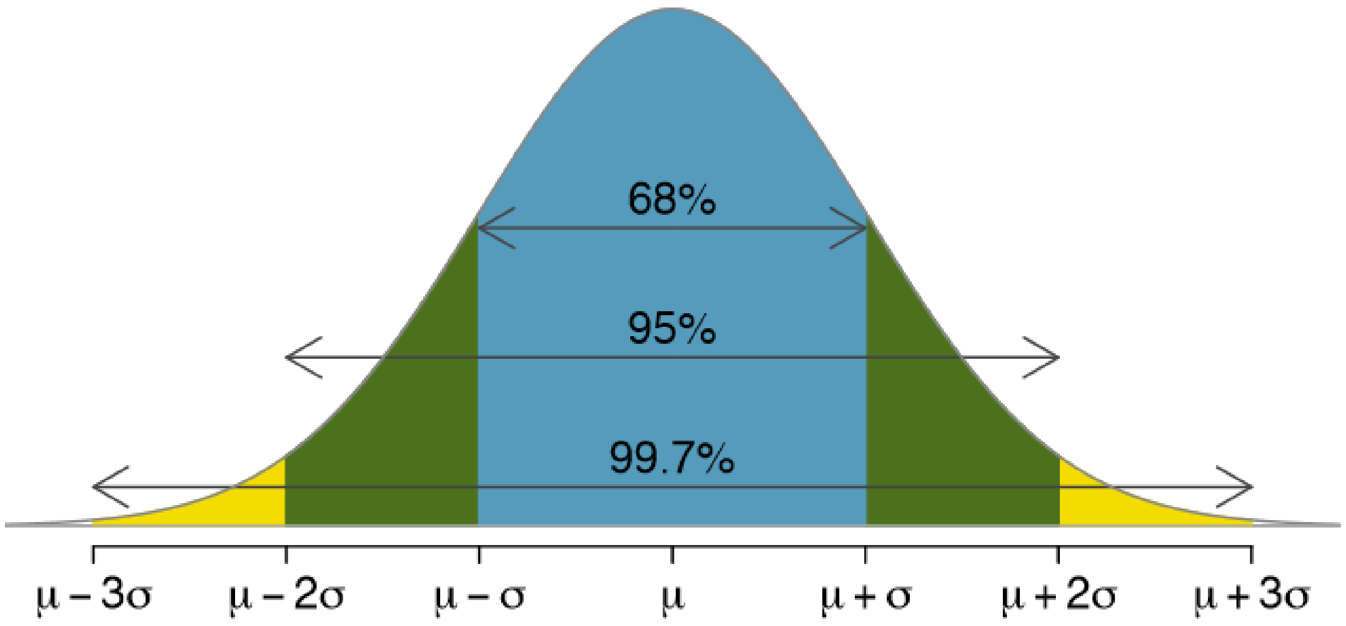
- There is strong evidence that clinic location is associated with systolic blood pressure.

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.

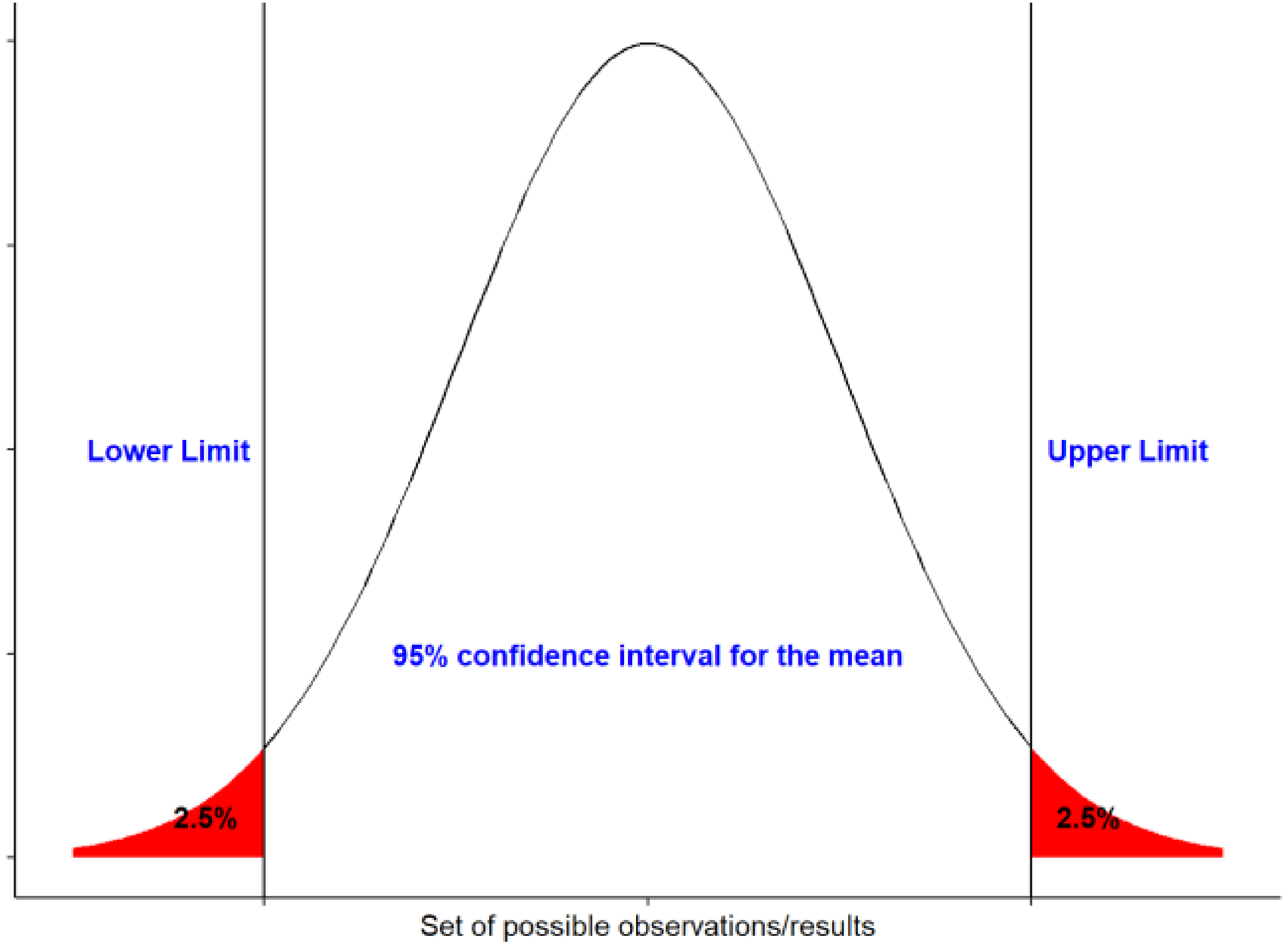
Confidence intervals show:
- Effect size
- Precision of estimates



Confidence Interval Example

Mean SBP difference (urban – rural) = 4.2 mmHg
• 95% CI: [2.1, 6.3] mmHg

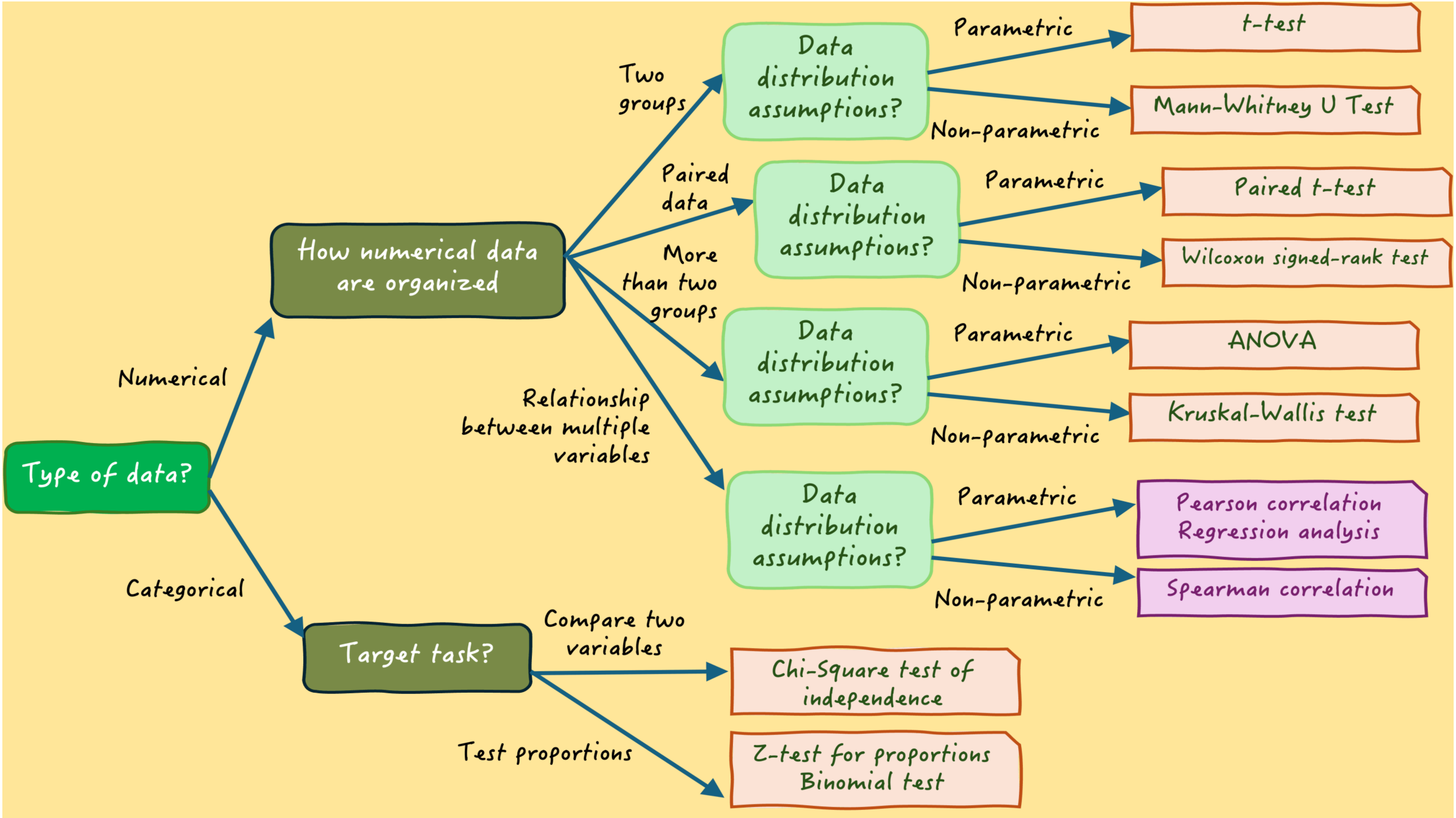
Interpretation:
• The true SBP difference lies within this range
• CI does not include zero → statistically significant
• Provides more information than p-value alone



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



Type of data?
Numerical
Categorical
How numerical data are organized
Two groups
Paired data
More than two groups
Relationship between multiple variables
Data distribution assumptions?
Parametric
Non-parametric
t-test
Mann-Whitney U Test
Data distribution assumptions?
Parametric
Non-parametric
Paired t-test
Wilcoxon signed-rank test
Data distribution assumptions?
Parametric
Non-parametric
ANOVA
Kruskal-Wallis test
Data distribution assumptions?
Parametric
Non-parametric
Pearson correlation
Regression analysis
Spearman correlation
Target task?
Compare two variables
Test proportions
Chi-Square test of independence
Z-test for proportions
Binomial test

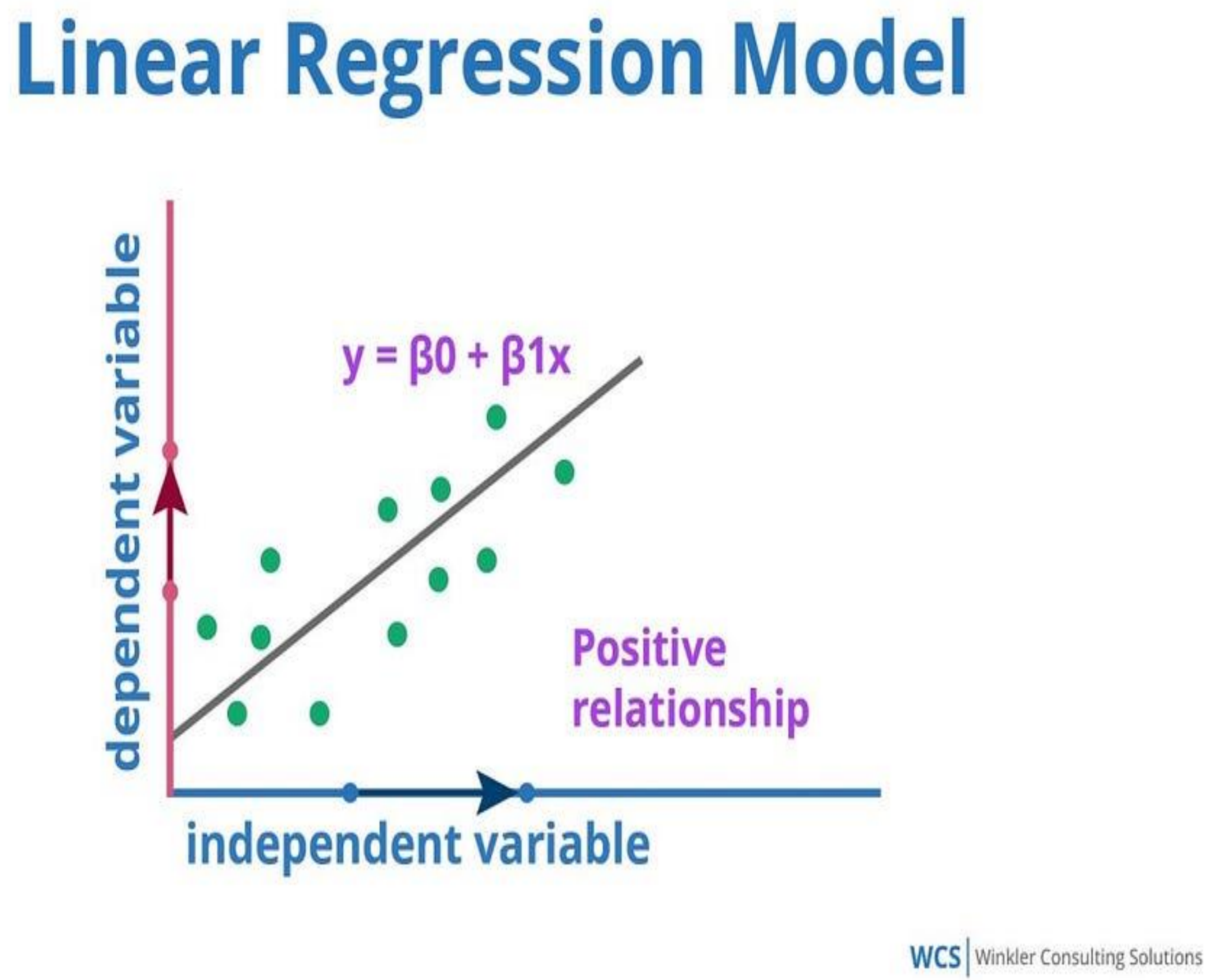
Introduction to Regression Analysis

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

General form:

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

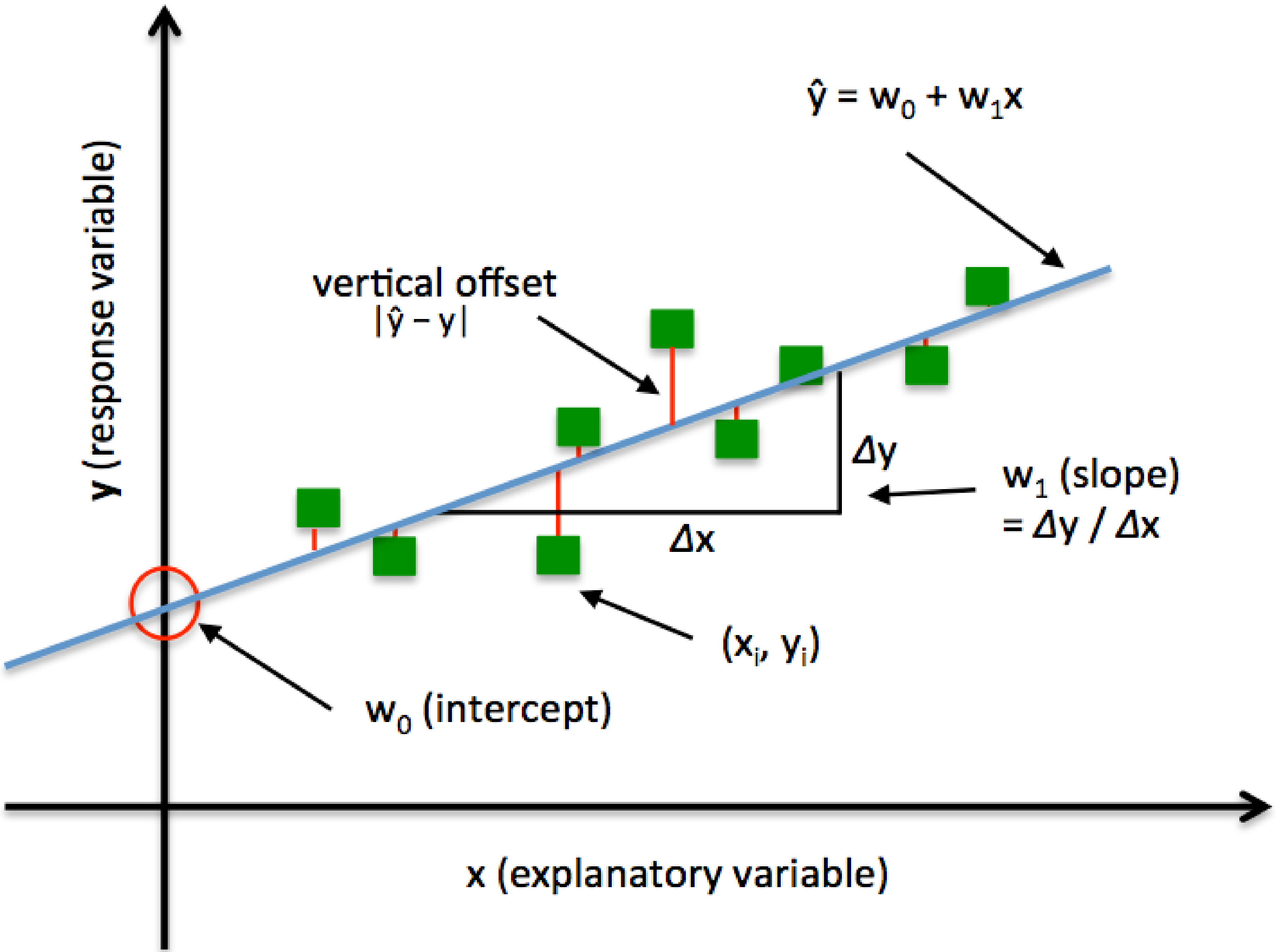
Regression is the foundation of many ML algorithms.



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:

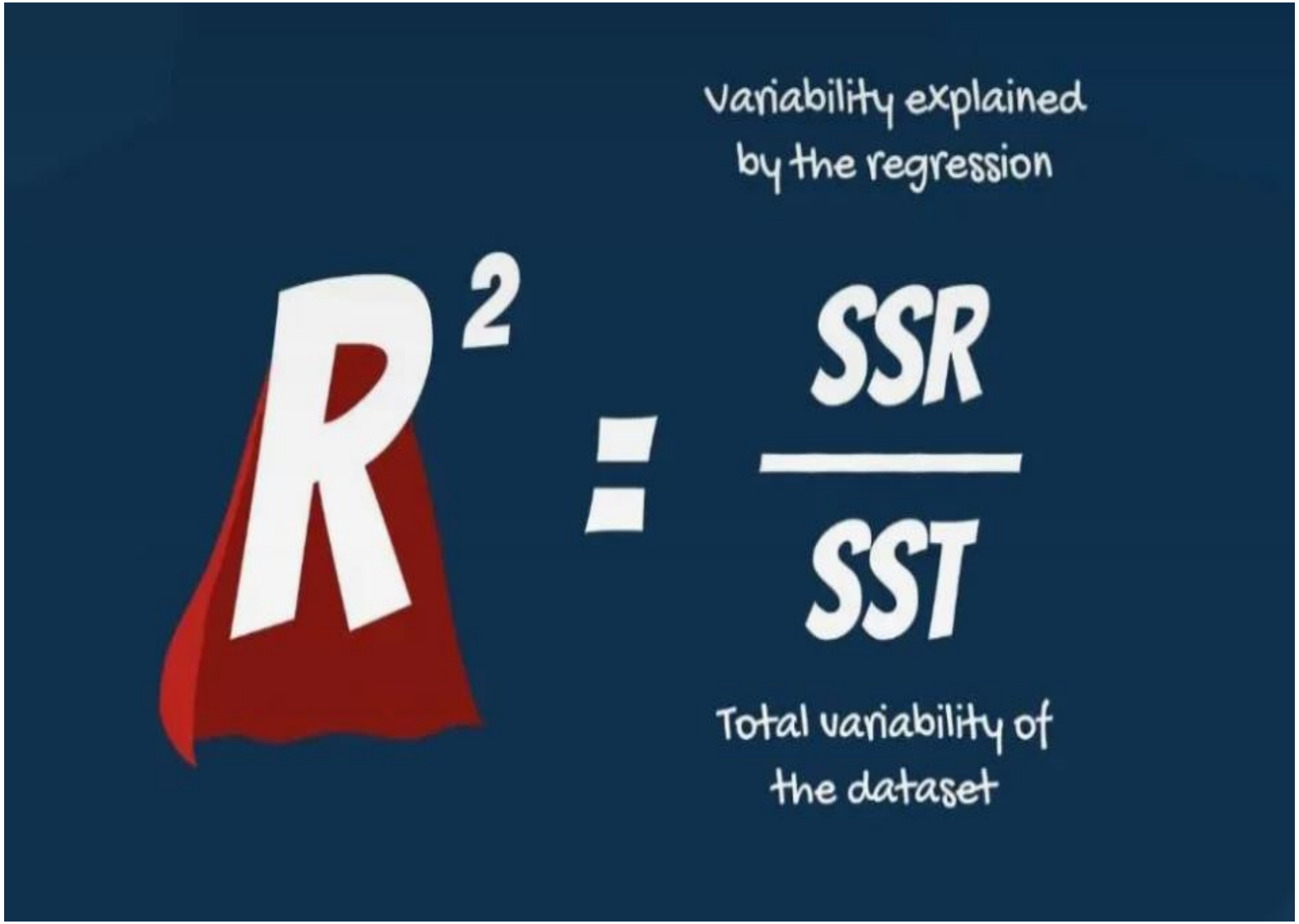
- ❖A BMI coefficient of 0.87 means SBP increases by 0.87 mmHg per unit BMI increase, holding other variables constant.

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.

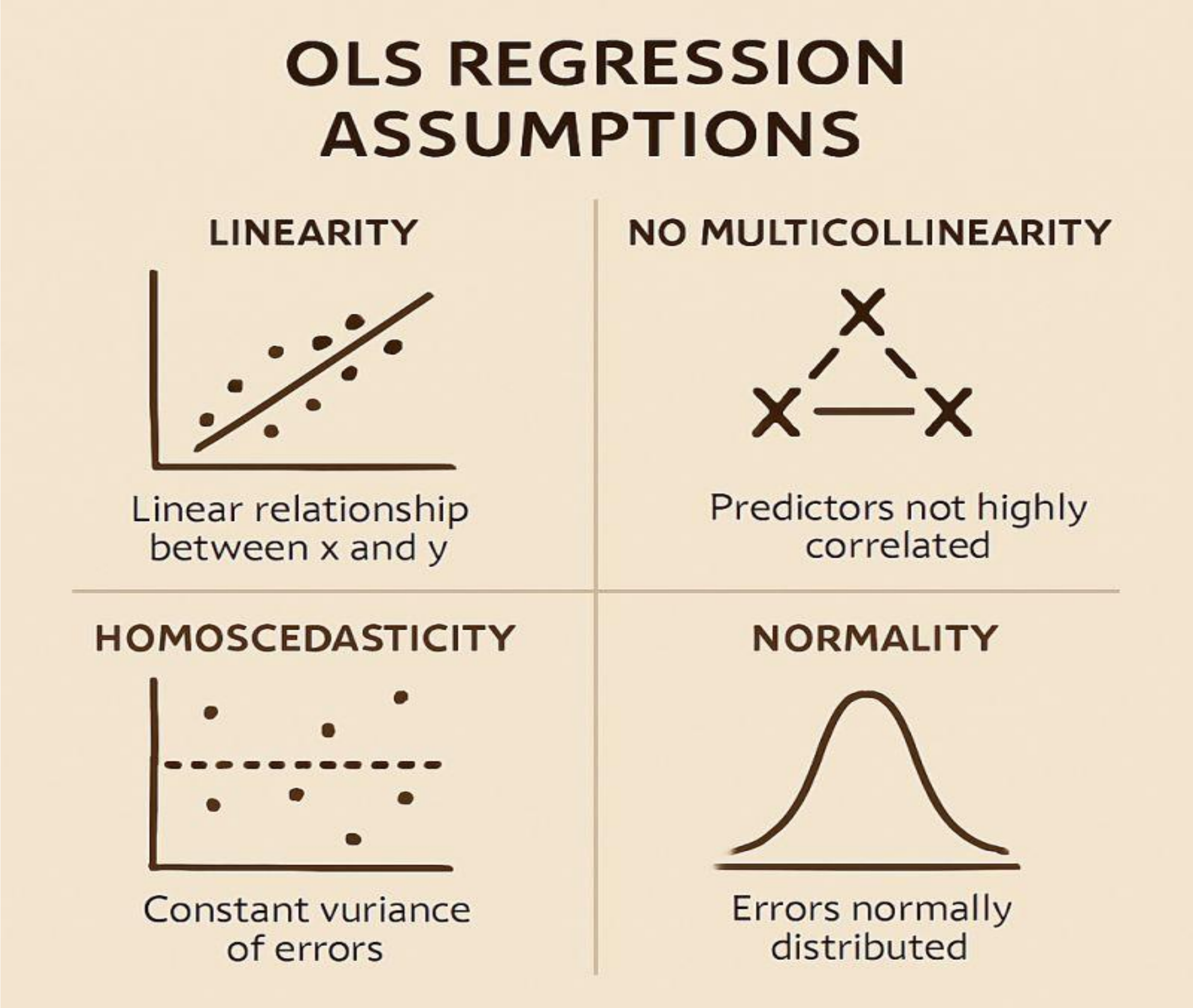


OLS Assumptions

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

Logistic Regression

Logistic regression is used when the outcome is binary.

- ❖ Medical application:
- ❖ Predict hypertension (yes/no)

Logistic regression models log-odds, not probabilities directly.

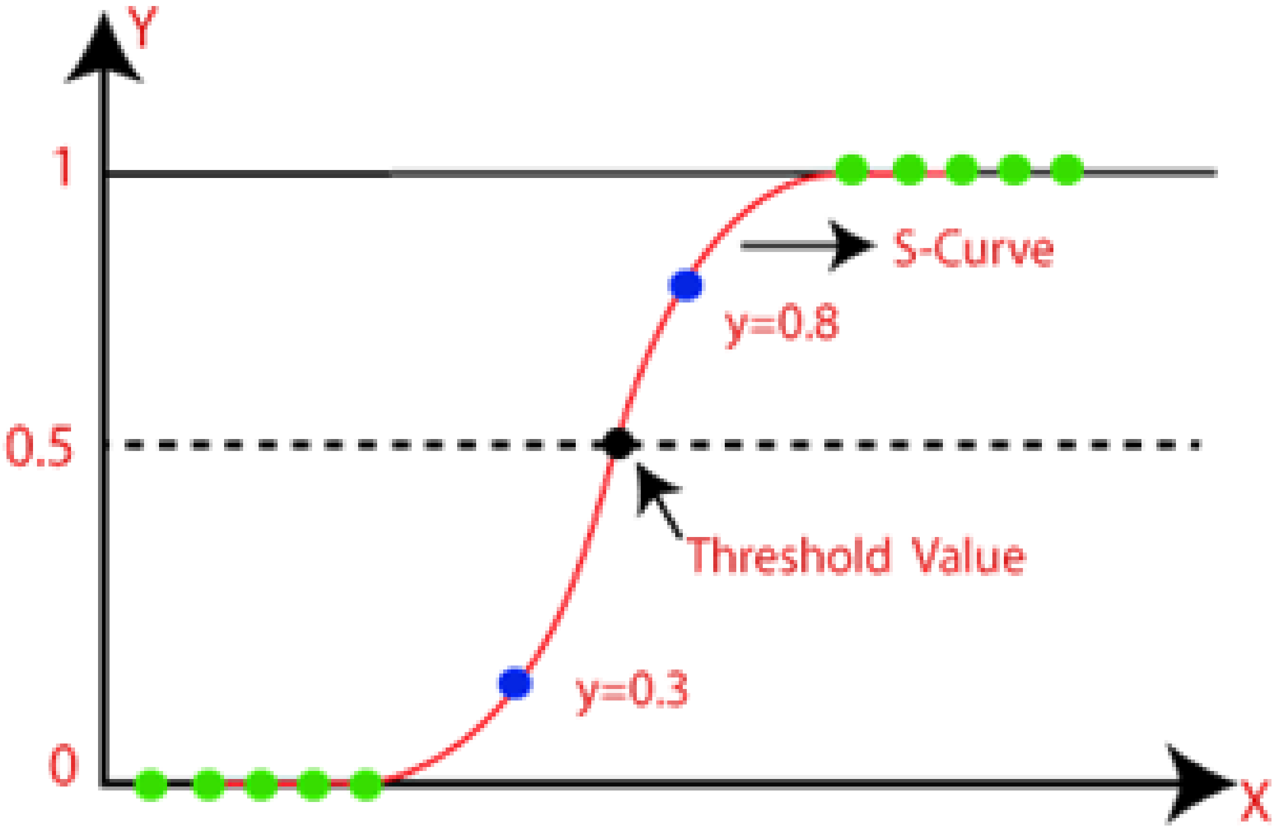


Fig 1:- Logistic Regression

Source: Javapoint

Logistic Regression Equation

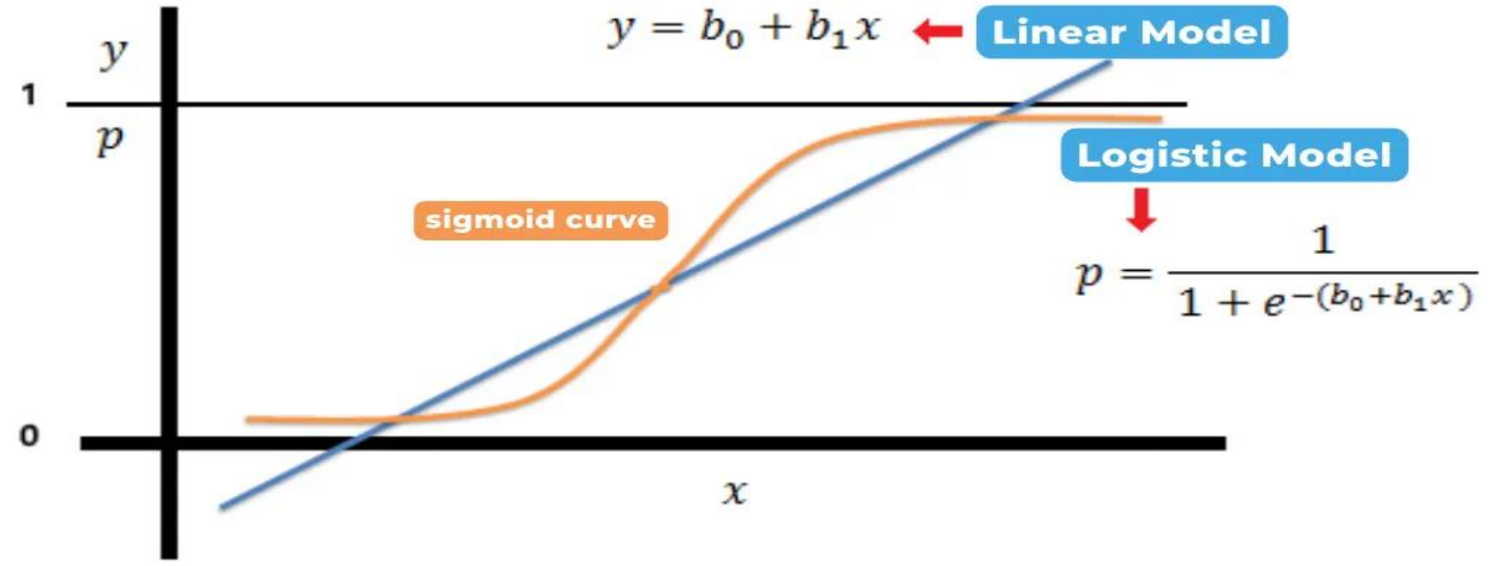
Logit model:

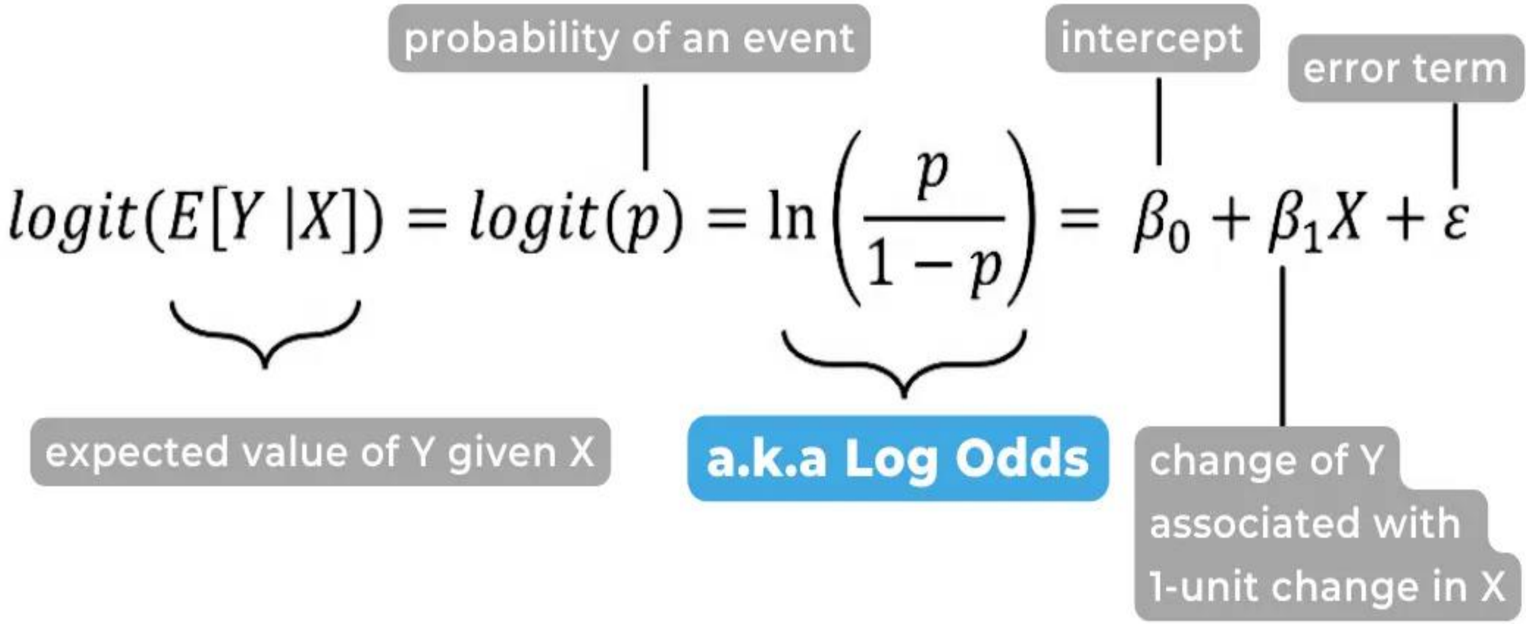
- $\log(p / (1 - p)) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots$

Where:

- p = probability of the event occurring

This ensures predictions lie between 0 and 1.





Interpreting Logistic Regression Results

Two interpretations:

- ❖Coefficients → Change in log-odds
- ❖Odds ratios (e^β) → Multiplicative change in odds

- ❑Odds ratio > 1 → Increased odds
- ❑Odds ratio < 1 → Decreased odds

Variable	Odds ratio	Interpretation
Age	$e^{-0.0292} = 0.9712$	For every one-year increase in age, the odds of having DR decreased by 2.88%
Diabetic Foot	$e^{-0.7343} = 0.480$	Patients without diabetic foot ulcer (Diabetic Foot=No) are less likely to have diabetic retinopathy compared to patients who suffered from diabetic foot ulcer (Diabetic Foot=Yes).
Duration of DM	$e^{0.1554} = 1.168$	For every one-year increase in duration of DM, the odds of having diabetic retinopathy increased by 16.8%.
HbA_{1C}	$e^{0.1853} = 1.204$	For everyone unit increase in HbA_{1C} level, the odds of having

Logistic Regression Example

Advanced HIV status:

Odds ratio = 2.51

Interpretation:

❖ Patients with advanced HIV have 2.5 times higher odds of hypertension compared to others.

Statistical Interpretation

The **Odds Ratio (OR)** is a measure of association between an exposure and an outcome. In this context:

- **Positive Association:** Since the $OR > 1$, advanced HIV is associated with higher odds of hypertension.
- **Magnitude:** The odds are **151% higher** in the advanced HIV group than in the control group. This is calculated as $(2.51 - 1) \times 100 = 151\%$.
- **Mathematical Expression:** The relationship is expressed as:

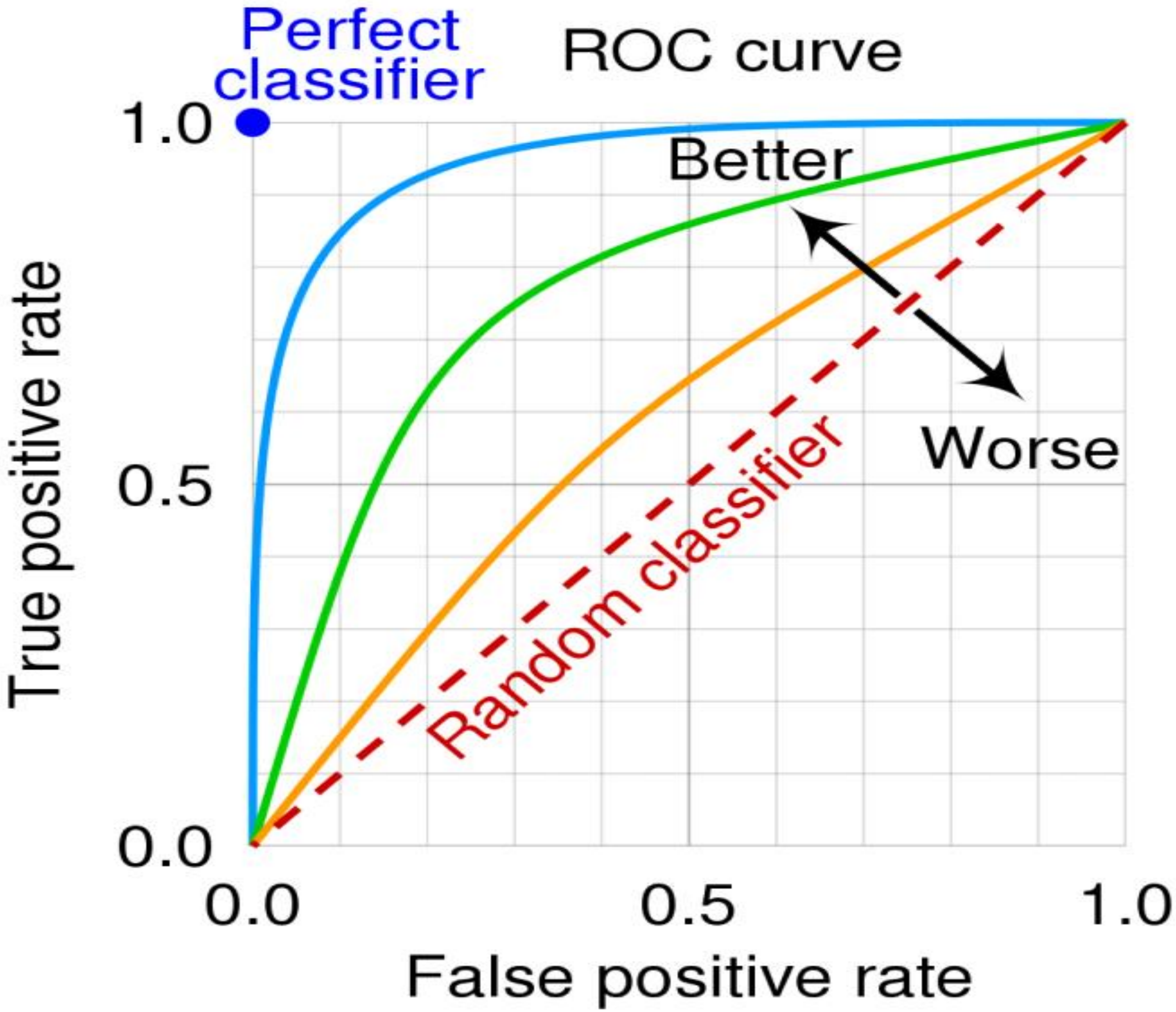
$$\text{Odds}_{\text{exposed}} = 2.51 \times \text{Odds}_{\text{unexposed}}$$

Model Evaluation for Classification

Key metrics for binary outcomes:

- ❖Accuracy
- ❖Sensitivity (Recall)
- ❖Specificity
- ❖AUC–ROC

Medical screening prioritize high sensitivity to reduce missed cases.



Confounding and Adjustment

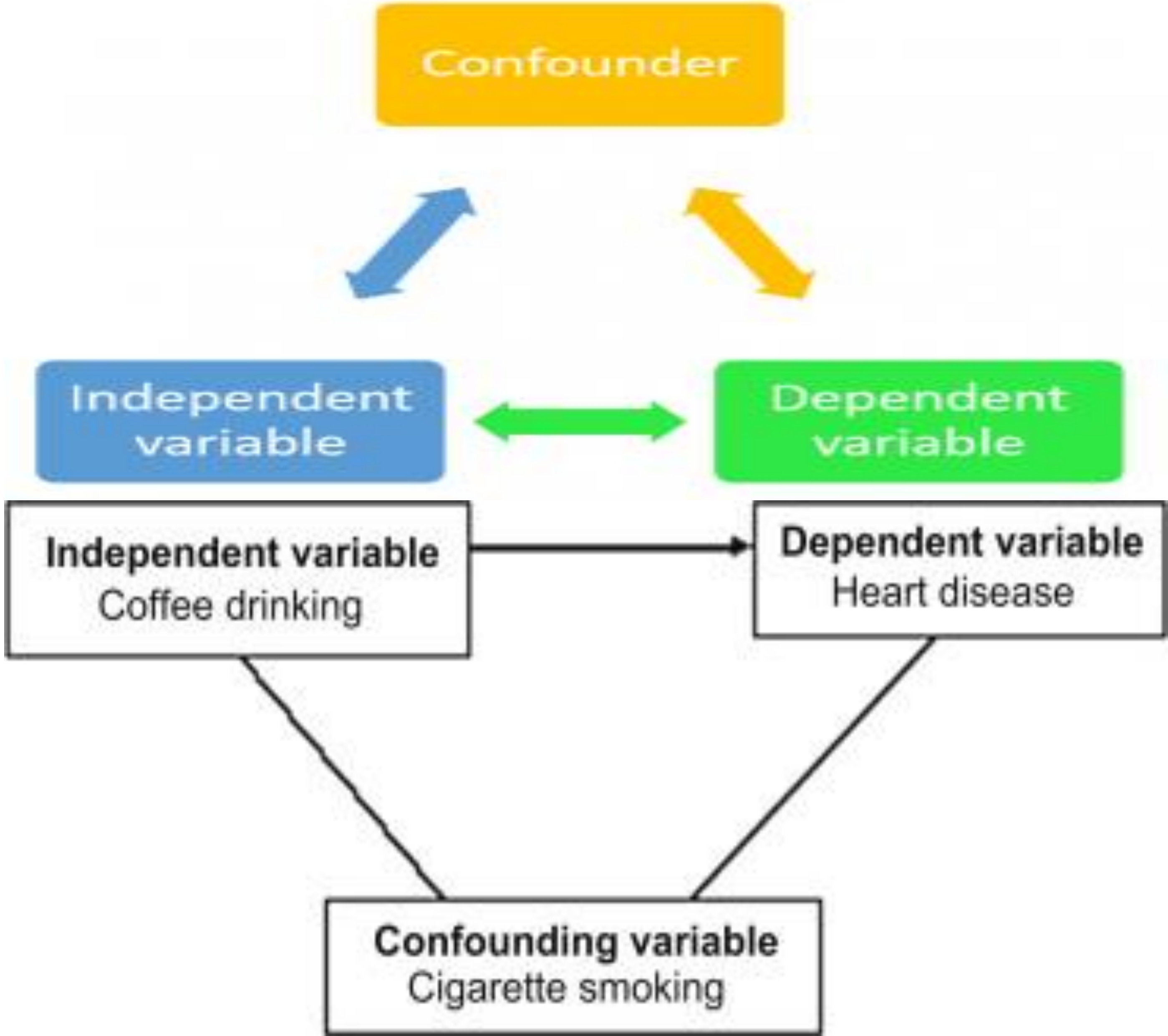
Confounding occurs when a third variable distorts the relationship between X and Y.

Example:

❖Urban clinics appear associated with higher SBP, but age explains the difference.

Solution:

❖Multivariable regression adjustment.

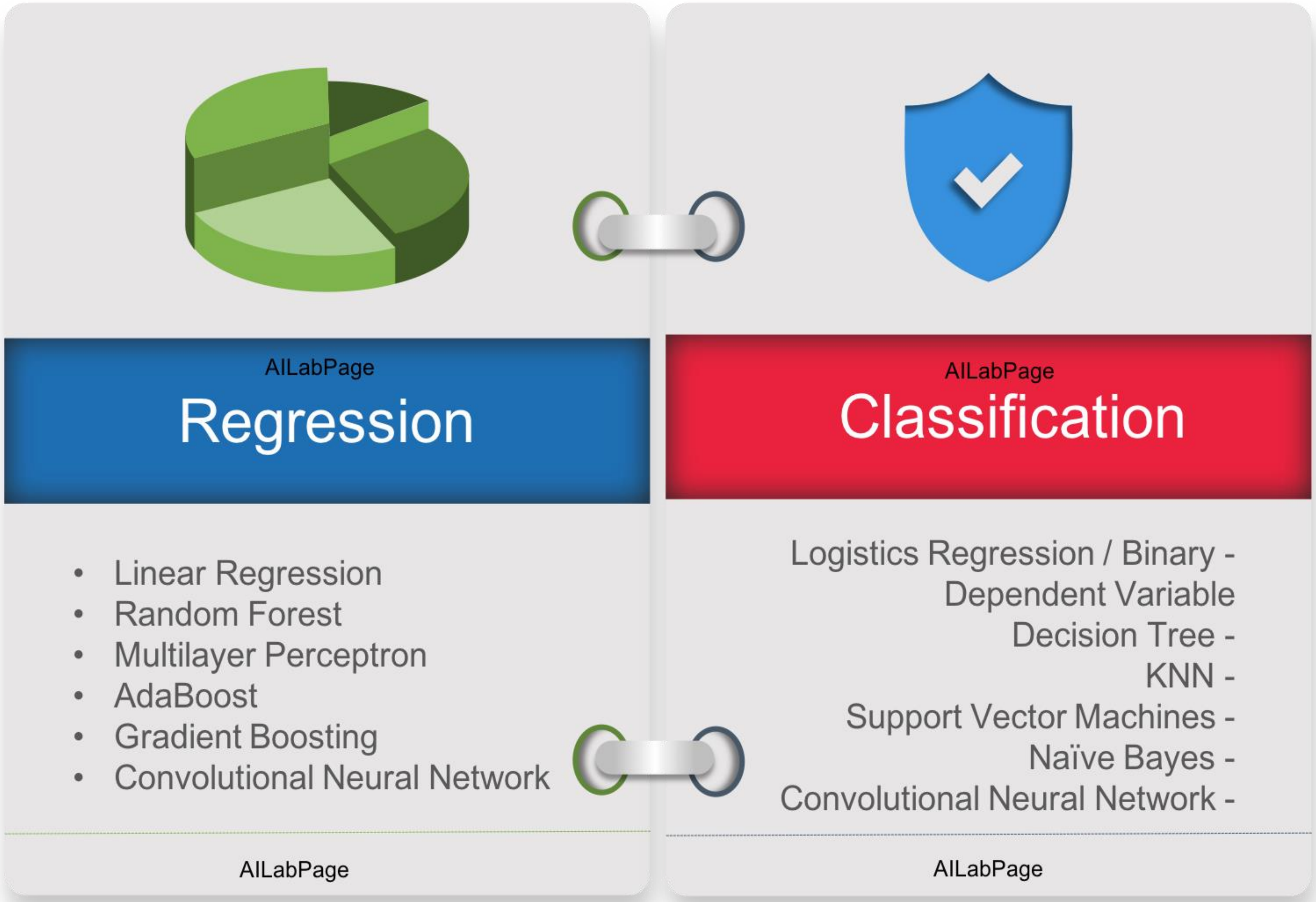


Regression as Machine Learning

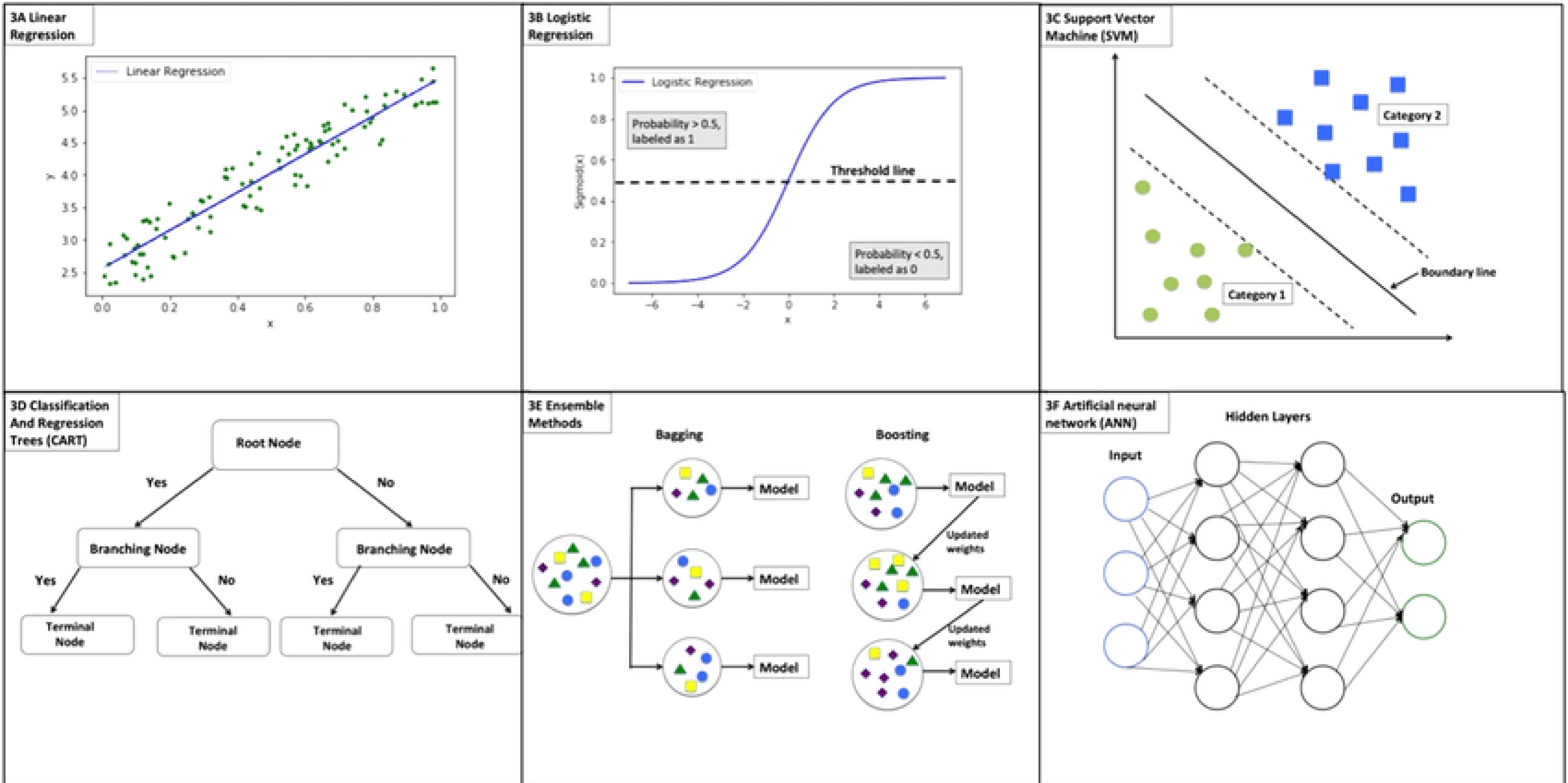
Regression models are machine learning algorithms.

Connections:
❖Linear regression → Linear ML model
❖Logistic regression → Classification ML model
❖Neural networks → Stacked regression layers

Understanding regression enables ML mastery.



Regression as Machine Learning



Common Statistical Pitfall

Avoid these errors:

- Data leakage
- P-hacking
- Ignoring confounding
- Overfitting models

Statistical rigor protects patient safety and model credibility.

Lab Assignment Overview

Tasks:

- Hypothesis testing
- OLS regression for SBP
- Logistic regression for hypertension
- Model evaluation and interpretation

Deliverable: Python notebook (.ipynb)

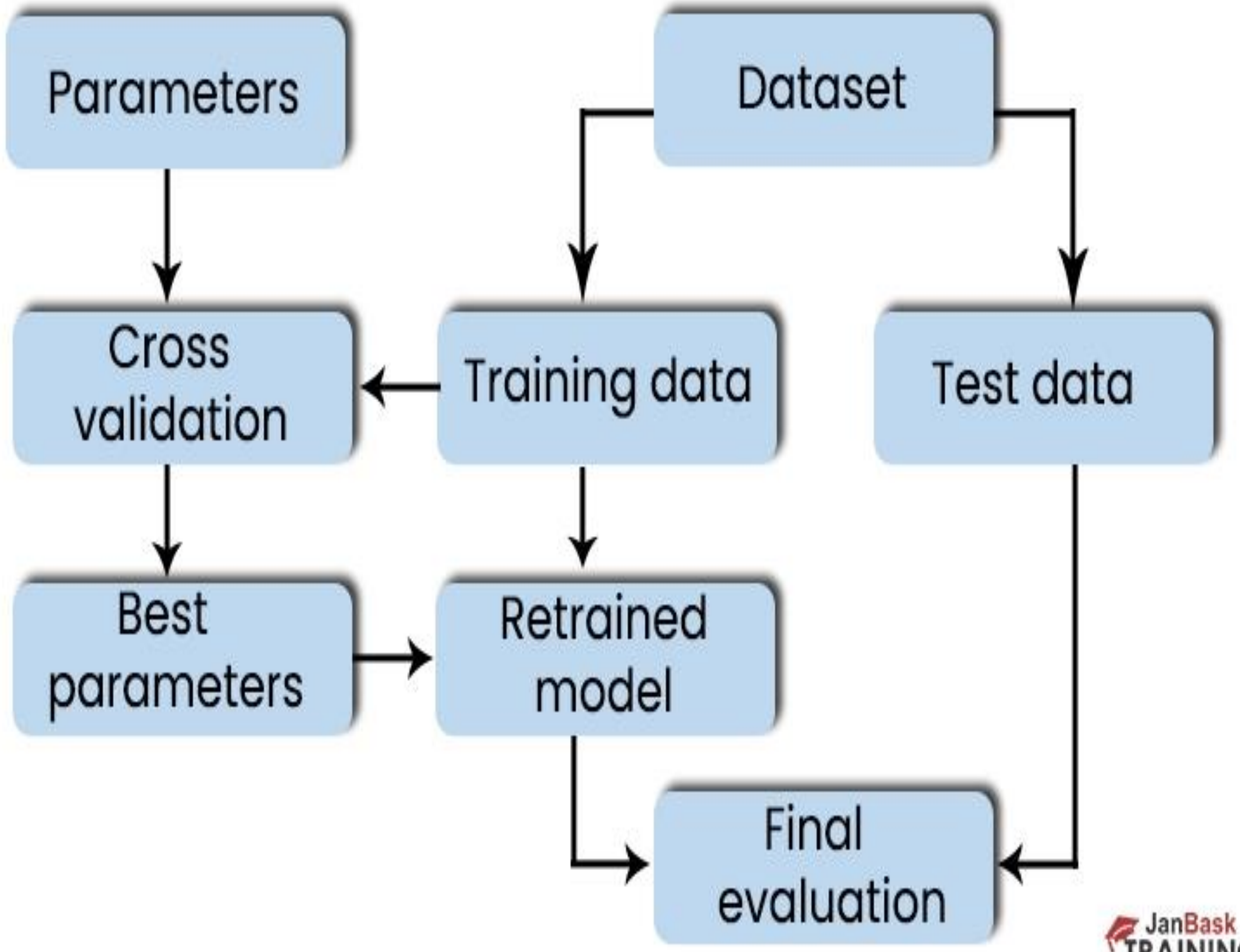
In [13]:
```python
# To perform feature selection based on p-value significance level in logistic regression, we can use the statsmodel

import statsmodels.api as sm

# Add a constant term to the features for the logistic regression model
X = sm.add_constant(X)

# Fit logistic regression model and calculate p-values
log_reg_model = sm.Logit(y, X)
result = log_reg_model.fit()

# Display summary with p-values
print(result.summary())
```

```
Optimization terminated successfully.
         Current function value: 0.464388
         Iterations 6
                           Logit Regression Results
==============================================================================
Dep. Variable:                Outcome   No. Observations:                  768
Model:                          Logit   Df Residuals:                      759
Method:                           MLE   Df Model:                            8
Date:                Fri, 15 Dec 2023   Pseudo R-squ.:                  0.2820
Time:                        18:28:30   Log-Likelihood:                -356.65
converged:                       True   LL-Null:                       -496.74
Covariance Type:            nonrobust   LLR p-value:                 6.750e-56
============================================================================================
                               coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------------------
const                       -9.0968      0.813    -11.195      0.000     -10.689      -7.504
Pregnancies                  0.1250      0.032      3.860      0.000       0.062       0.188
Glucose                      0.0374      0.004      9.630      0.000       0.030       0.045
BloodPressure               -0.0088      0.009     -1.028      0.304      -0.026       0.008
SkinThickness                0.0035      0.013      0.265      0.791      -0.022       0.029
Insulin                     -0.0008      0.001     -0.671      0.502      -0.003       0.002
BMI                          0.0931      0.018      5.219      0.000       0.058       0.128
DiabetesPedigreeFunction     0.8661      0.296      2.923      0.003       0.285       1.447
Age                          0.0131      0.010      1.382      0.167      -0.005       0.032
============================================================================================
```

Python Analysis Workflow

Steps:

1. Load and clean data
2. Conduct hypothesis tests
3. Fit OLS regression
4. Fit logistic regression
5. Evaluate model performance



Preparing for Week 3 Machine Learning

Before next session:

- ❖ Complete lab assignment
- ❖ Review regression concepts
- ❖ Understand train-test split
- ❖ Install required Python libraries

Week 3 focuses on scalable ML pipelines.

Conclusion

Key takeaways:

- ❖ Hypothesis testing guides evidence-based decisions
- ❖ Regression underpins machine learning algorithms
- ❖ Statistical literacy enables responsible AI
- ❖ Healthcare ML requires precision and ethics

Statistics is the foundation of trustworthy machine learning.