University of Petra

307304 Business Intelligence and Data Mining

Simple Linear Regression

Topics and Learning Outcomes

What You'll Learn:

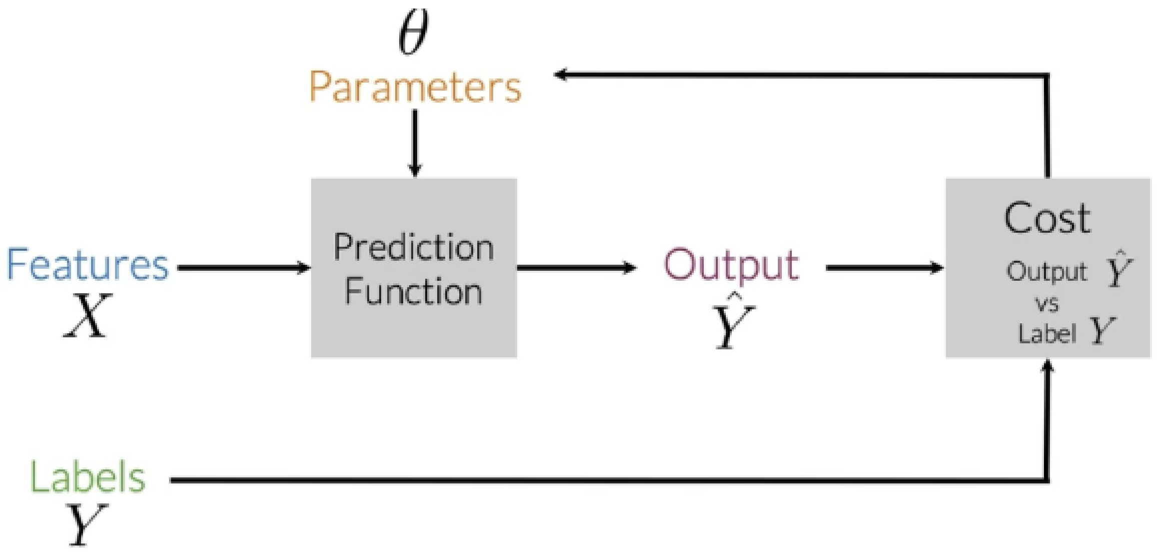
- Introduce the concept of supervised learning
- Understanding regression analysis fundamentals
- Types of linear regression models
- Best-fit line, hypothesis function, and OLS method
- Closed-form solution implementation
- Gradient descent algorithm
- Linear regression assumptions
- Evaluation metrics and performance assessment
- Practical implementation with scikit-learn

Supervised Machine Learning

Supervised machine learning is a type of learning where a model learns from labeled data.

The process follows these key steps:

- **Input Features (X):** The model takes input data (features) which represents the information used to make predictions.
- **Prediction Function:** The model, with some parameters (θ), processes the input data to generate predictions (Ŷ).
- **Output (Ŷ):** This is the model’s prediction based on the input features.

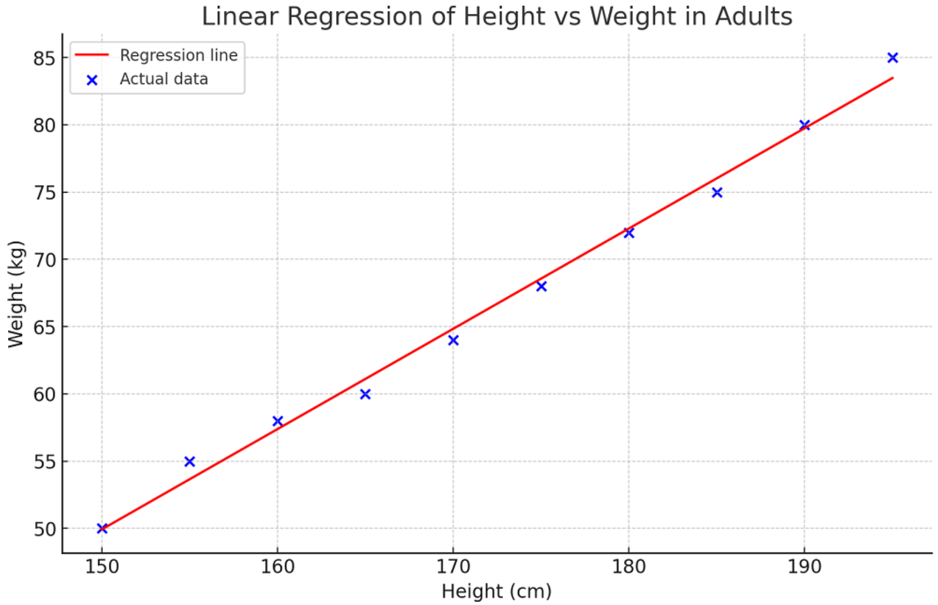


- **Cost Function:** The predicted output (Ŷ) is compared to the actual labels (Y) to calculate an error or "cost."
- **Optimization:**
  - The model adjusts its parameters (θ) of the cost function to minimize the error, improving its predictions over time.
  - This process repeats iteratively until the model achieves good accuracy.
  - It's commonly used for tasks like classification (e.g., spam detection) and regression (e.g., predicting house prices).

What is Regression Analysis?

Definition and Purpose

- **Linear Regression** is a supervised machine learning algorithm that models the linear relationship between:
  - **Dependent variable** (target): What we want to predict
  - **Independent variable(s)** (features): What we use to make predictions
- **Key Concept:** Fits a linear equation to observed data to make predictions on new, unseen data.
- **Real-World Applications:**
  - Predicting house prices based on size
  - Estimating sales revenue from advertising spend
  - Forecasting stock prices using economic indicators



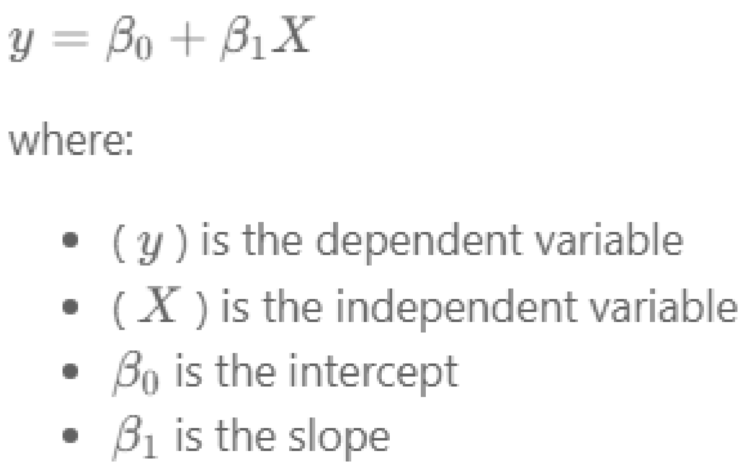
Types of Linear Regression

There are two main types of linear regression:

1- Simple Linear Regression

This is the simplest form of linear regression, and it involves only one independent variable and one dependent variable.

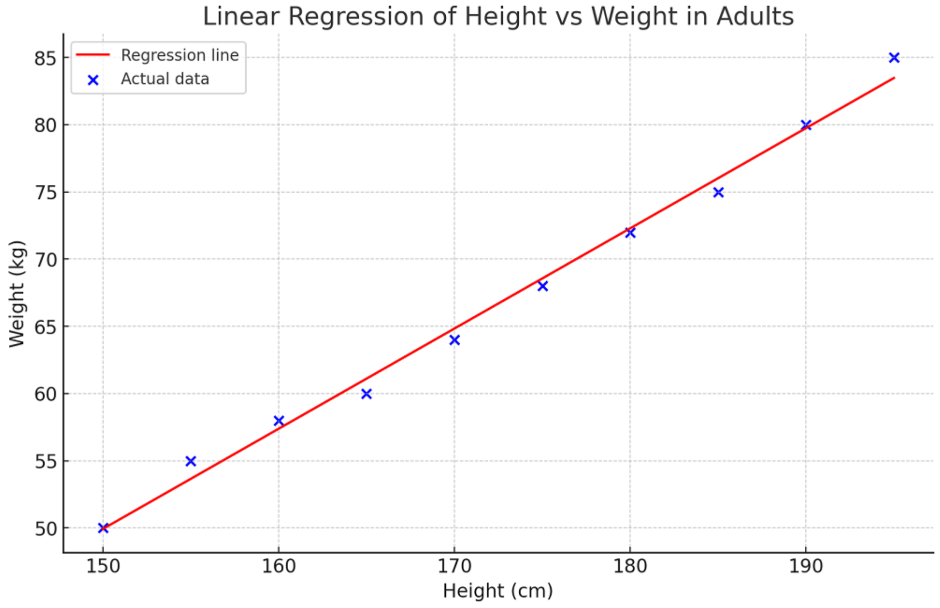
The equation for simple linear regression is:



$$y = \beta_0 + \beta_1 X$$

where:

- (y) is the dependent variable
- (X) is the independent variable
- β_0 is the intercept
- β_1 is the slope

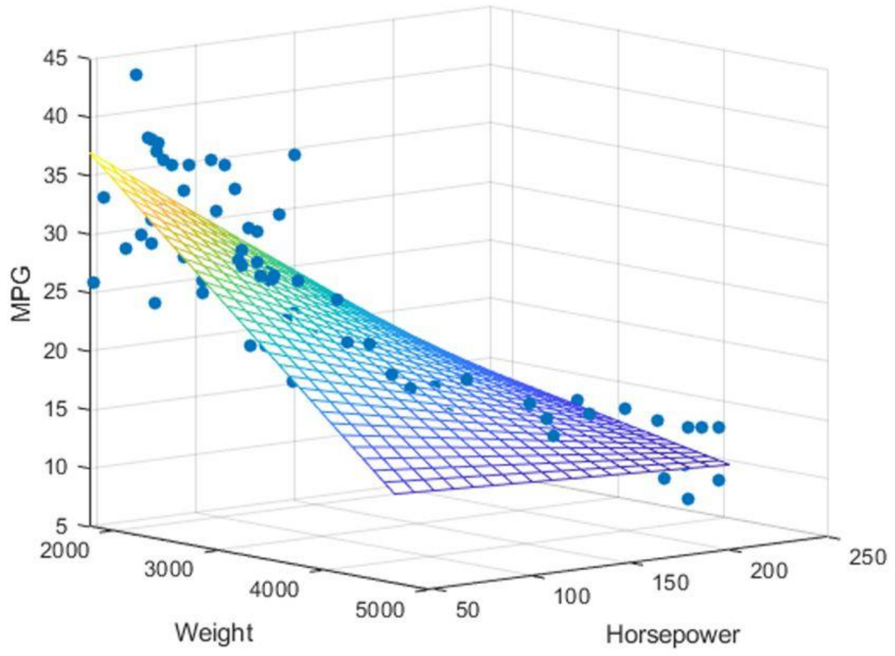


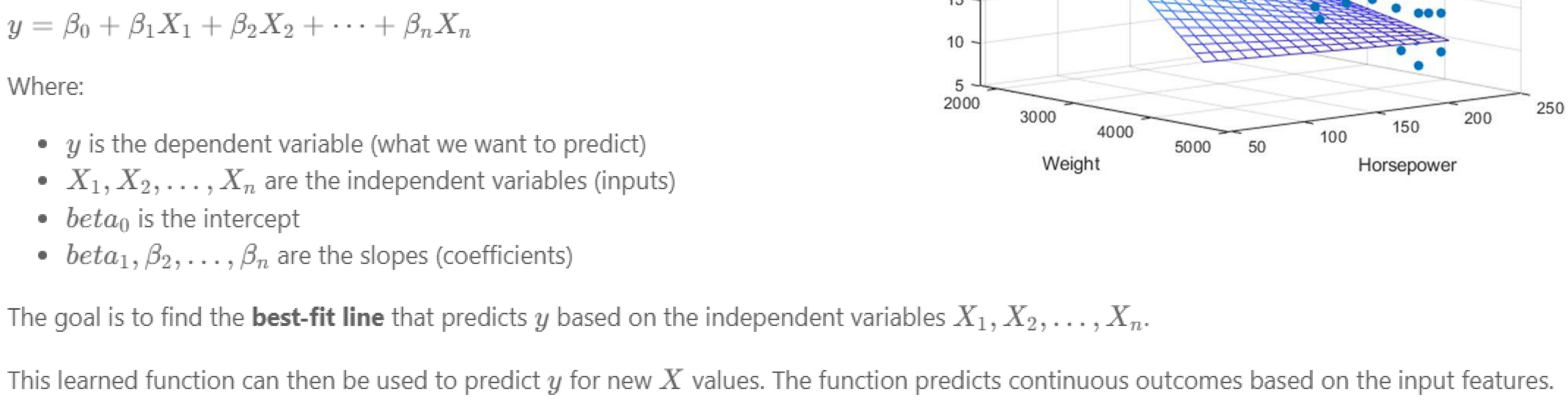
Types of Linear Regression

2- Multiple Linear Regression (Simplified)

In multiple linear regression, there is one dependent variable and more than one independent variable.

The equation is:





$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$$

Where:

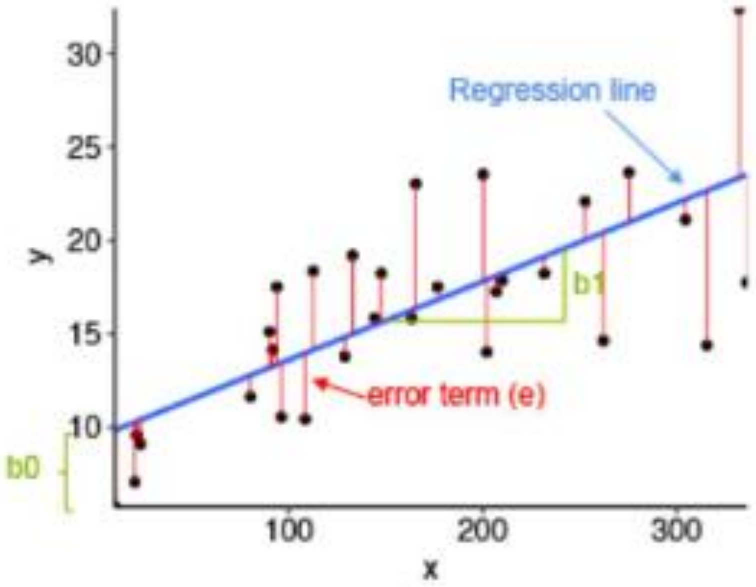
- y is the dependent variable (what we want to predict)
- $X_1, X_2, \ldots, X_n$ are the independent variables (inputs)
- $beta_0$ is the intercept
- $beta_1, \beta_2, \ldots, \beta_n$ are the slopes (coefficients)

The goal is to find the **best-fit line** that predicts y based on the independent variables $X_1, X_2, \ldots, X_n$.

This learned function can then be used to predict y for new X values. The function predicts continuous outcomes based on the input features.

What is the Best Fit Line?

- In linear regression, the best-fit line is the line that minimizes the error between the predicted and actual values.
- It represents the relationship between the dependent and independent variables.
- The goal is to change the slope and the intercept of the line to make the most accurate predictions.



- Linear regression predicts Y based on X, using a straight line to model the relationship. For example, X could be work experience, and Y could be salary.
- The goal is to find the best-fit line that minimizes the error between predicted and actual values.
- To find this line, we use a cost function, which helps determine the best values for the line's coefficients.

Why Linear Regression Matters

Four Key Benefits:

- **Interpretability**
  - Clear, understandable coefficients
  - Easy to explain relationships between variables
- **Simplicity**
  - Transparent and easy to implement
  - Foundation for more complex algorithms
- **Basis for Advanced Models**
  - Building block for regularization (Ridge, Lasso)
  - Foundation for support vector machines
- **Assumption Testing**
  - Validates key data assumptions
  - Critical for statistical analysis

The Best-Fit Line Concept

What Makes a Line "Best-Fit"?

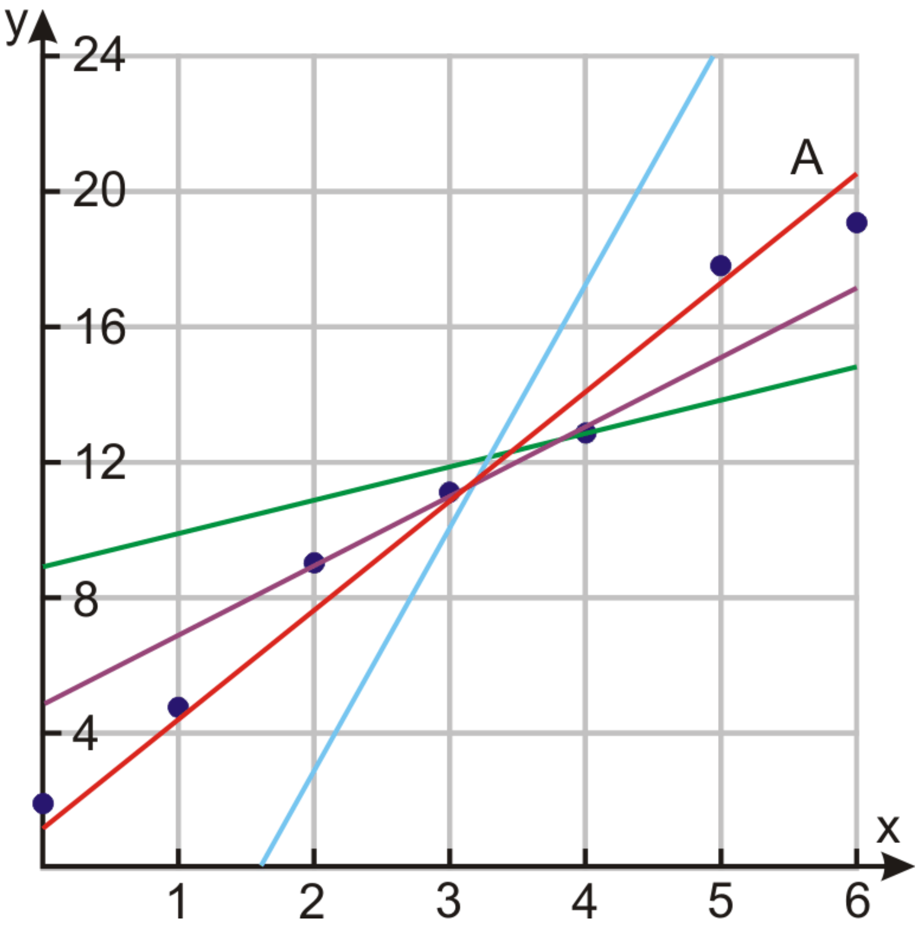
The best-fit line minimizes the error between predicted and actual values.

Key Components:

- **Slope (β_1):** How much Y changes for each unit change in X
- **Intercept (β_0):** Value of Y when X = 0

- Goal: Find the line that makes the most accurate predictions
- Mathematical Representation: $\hat{Y} = \beta_0 + \beta_1 X$

Where $\hat{Y}$ (pronounced Y Hat) is the predicted value.



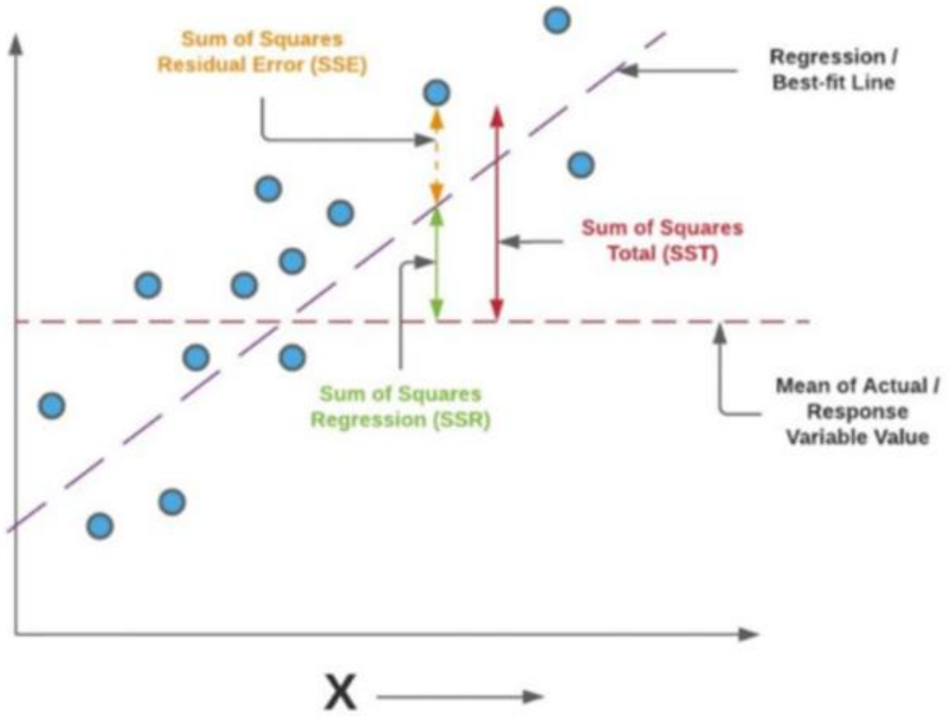
Ordinary Least Squares (OLS) Method

The Foundation of Linear Regression

OLS Principle: Find the line that minimizes the sum of squared differences between observed and predicted values.

Why "Least Squares"?

- Squares prevent positive and negative errors from canceling out
- Penalizes larger errors more heavily
- Provides unique, mathematically optimal solution



Formula:

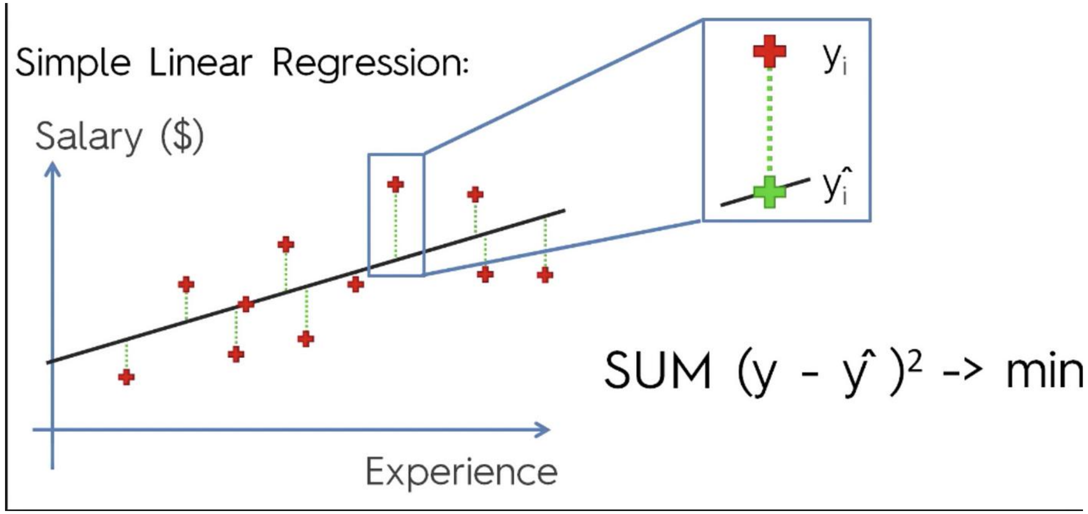
$$\overset{\text{Dependent Variable}}{Y_i} = \underbrace{\overset{\text{Population Y intercept}}{\beta_0} + \overset{\text{Population Slope Coefficient}}{\beta_1}\overset{\text{Independent Variable}}{X_i}}_{\text{Linear component}} + \underbrace{\overset{\text{Random Error term}}{\varepsilon_i}}_{\text{Random Error component}}$$

Cost Function - Mean Squared Error

Measuring Prediction Quality

- The **Mean Squared Error (MSE)** quantifies how well our model performs.
- **Formula:**

$$J = \frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2$$



Key Points:

- Lower MSE = Better model performance
- Units are squared (e.g., dollars² for price prediction)
- Always positive (due to squaring)
- Heavily penalizes large errors

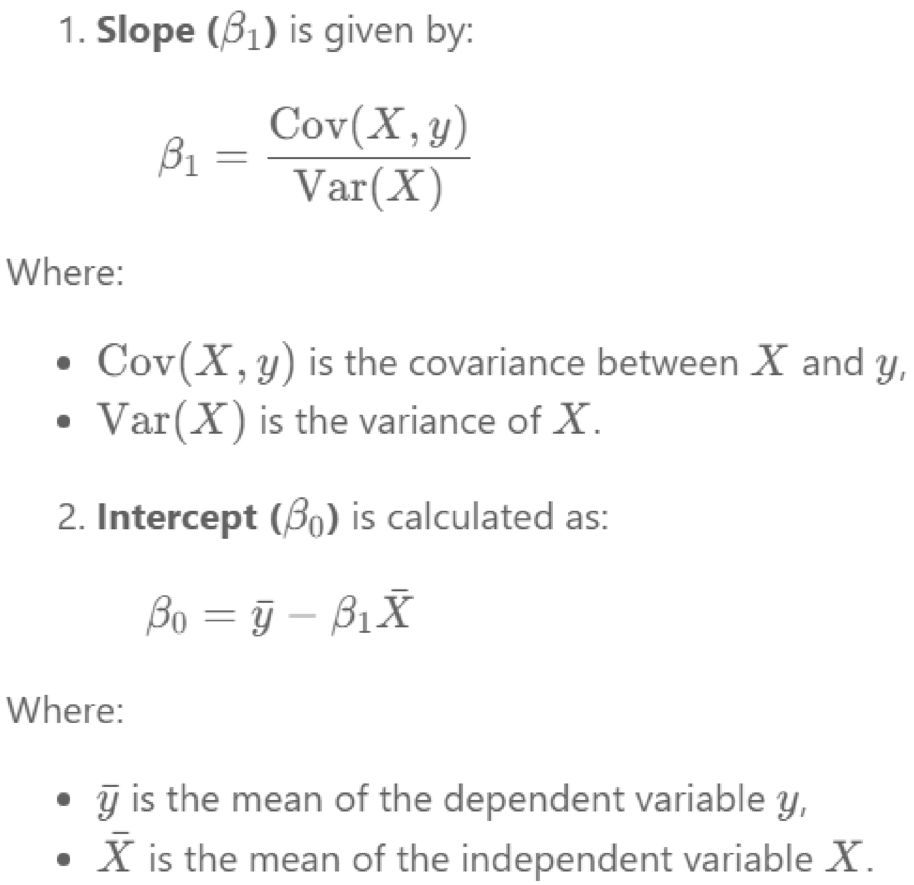
Solving Linear Regression with the Closed-Form Solution

- The closed-form solution is efficient for small to medium-sized datasets, especially in **simple linear regression**, where slope and intercept can be computed directly without matrix operations.
- However, for **multiple linear regression**, the closed-form solution involves computing the inverse of X^T.X, which becomes computationally expensive as dataset size and dimensionality grow.

It works well when:

- The dataset is small (typically up to a few thousand rows).
- The number of features is low.
- There is no multicollinearity (i.e., features are not highly correlated).

For larger or more complex problems, **iterative methods** like **Gradient Descent** are preferred due to their scalability and efficiency.



1. **Slope (β_1)** is given by:

$$\beta_1 = \frac{\mathrm{Cov}(X, y)}{\mathrm{Var}(X)}$$

Where:

- $\mathrm{Cov}(X, y)$ is the covariance between X and y,
- $\mathrm{Var}(X)$ is the variance of X.

2. **Intercept (β_0)** is calculated as:

$$\beta_0 = \bar{y} - \beta_1 \bar{X}$$

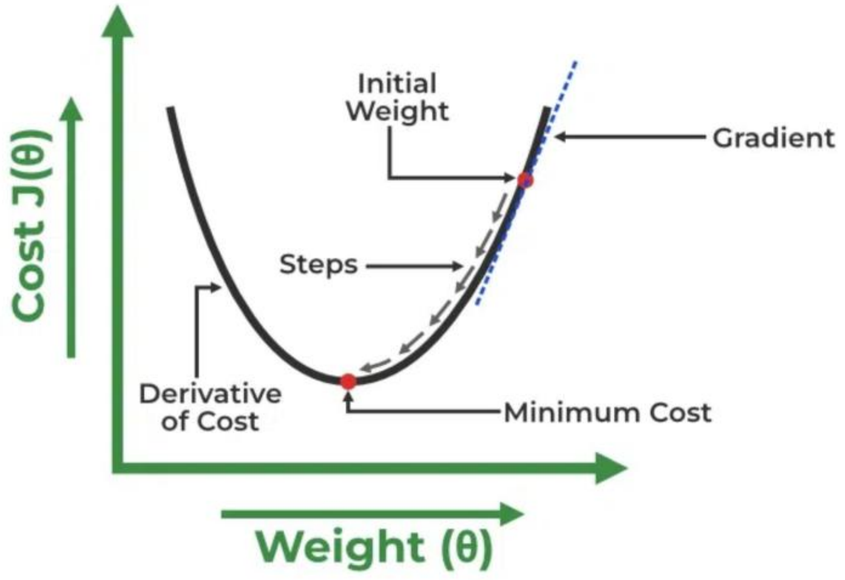
Where:

- $\bar{y}$ is the mean of the dependent variable y,
- $\bar{X}$ is the mean of the independent variable X.

python™

Solving Linear Regression Using The Gradient Descent Method

- A linear regression model can be trained using ***Gradient Descent***, which adjusts the model's parameters to minimize the mean squared error (MSE).
- To update (*Beta 0)* and (*Beta 1)* and reduce the cost function (minimizing the RMSE), gradient descent starts with random values for (*Beta 0)* and (*Beta 1)* and ***iteratively*** improves them to find the best-fit line.
- A gradient is simply the derivative, showing how small changes in inputs affect the output.
- By moving in the direction of the Mean Squared Error negative gradient with respect to the coefficients, the coefficients can be changed.



Solving Linear Regression Using The Gradient Descent Method

Computing β_0 Rate of Change:

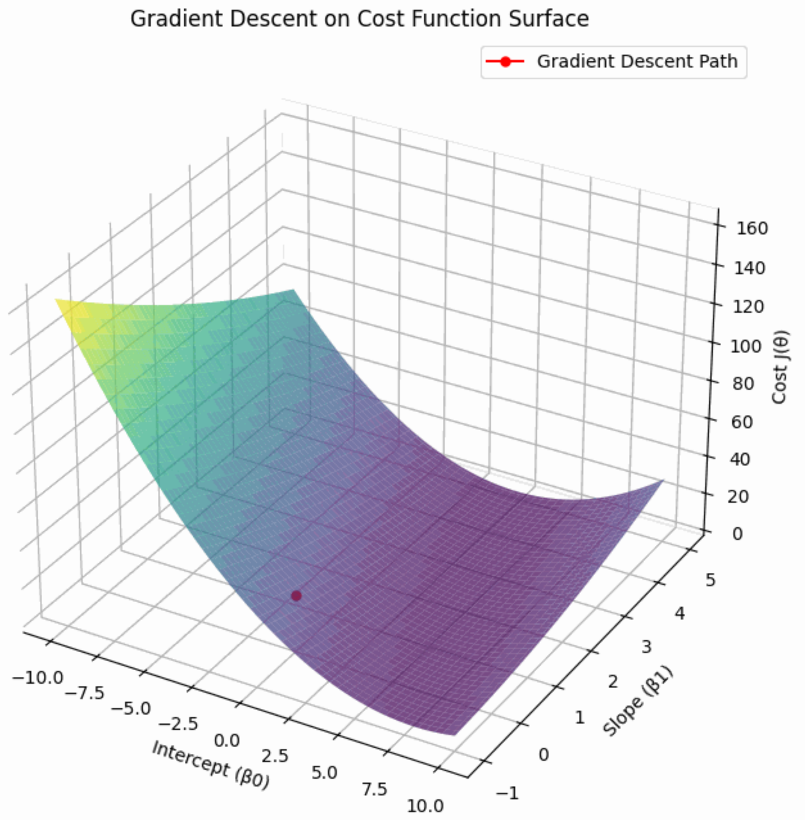
$$\beta_0 = \beta_0 - \alpha\left(J'_{\beta_0}\right)$$

$$= \beta_0 - \alpha\left(\frac{2}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)\right)$$

Computing β_1 Rate of Change:

$$\beta_1 = \beta_1 - \alpha\left(J'_{\beta_1}\right)$$

$$= \beta_1 - \alpha\left(\frac{2}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)\cdot x_i\right)$$

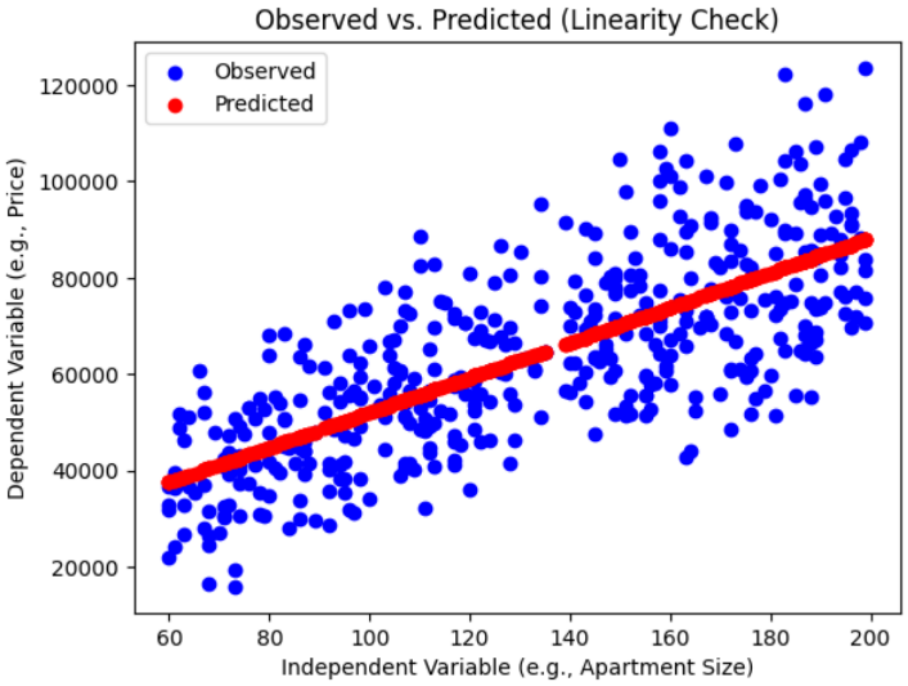


python™

Assumptions of Simple Linear Regression

Linear regression is a powerful tool for understanding and predicting the behavior of a variable, however, it needs to meet a few conditions in order to be accurate and dependable solutions.

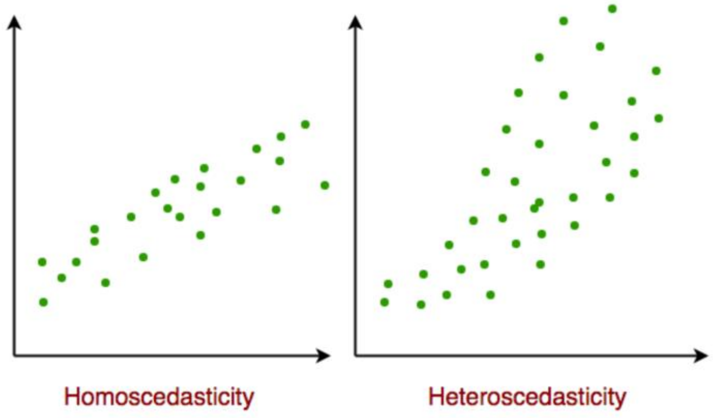
1. Linearity: The independent and dependent variables have a linear relationship with one another. This implies that changes in the dependent variable follow those in the independent variable(s) in a linear fashion. This means that there should be a straight line that can be drawn through the data points. If the relationship is not linear, then linear regression will not be an accurate model.

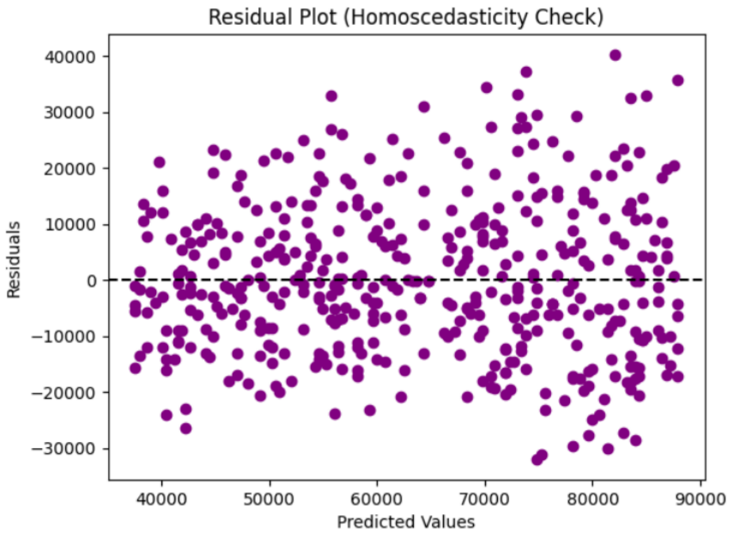


Assumptions of Simple Linear Regression

2. Independence: The observations in the dataset must be independent, meaning the value of the dependent variable for one observation should not be influenced by the value of another. If this condition is violated, linear regression may produce inaccurate results.

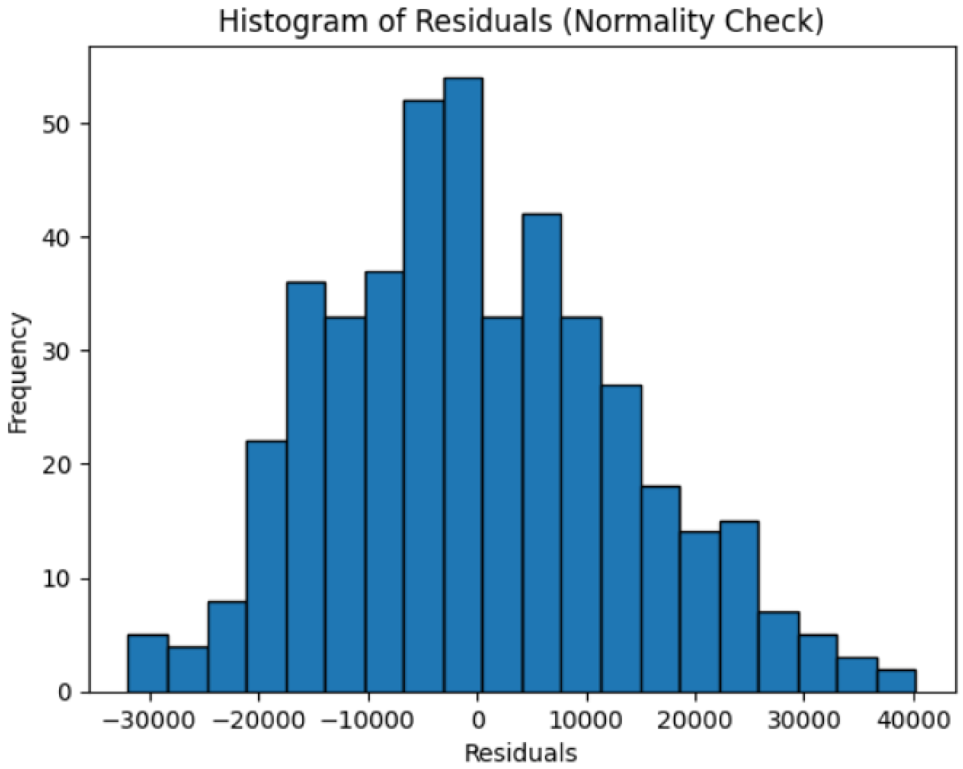
3. Homoscedasticity: The variance of the errors should be constant across all levels of the independent variable(s). This means that changes in the independent variable(s) should not affect the spread of the errors. If the error variance is not constant (heteroscedasticity), the linear regression model may not be reliable.

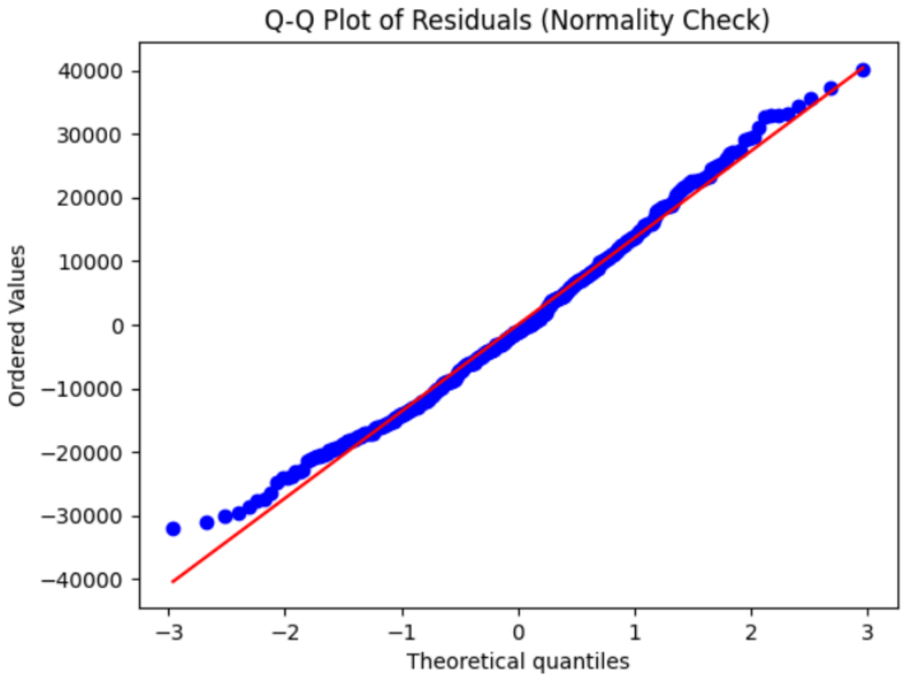




Assumptions of Simple Linear Regression

4. Normality: The residuals should be normally distributed. This means that the residuals should follow a bell-shaped curve. If the residuals are not normally distributed, then linear regression will not be an accurate model.





Assumptions of Simple Linear Regression

5. No Multicollinearity

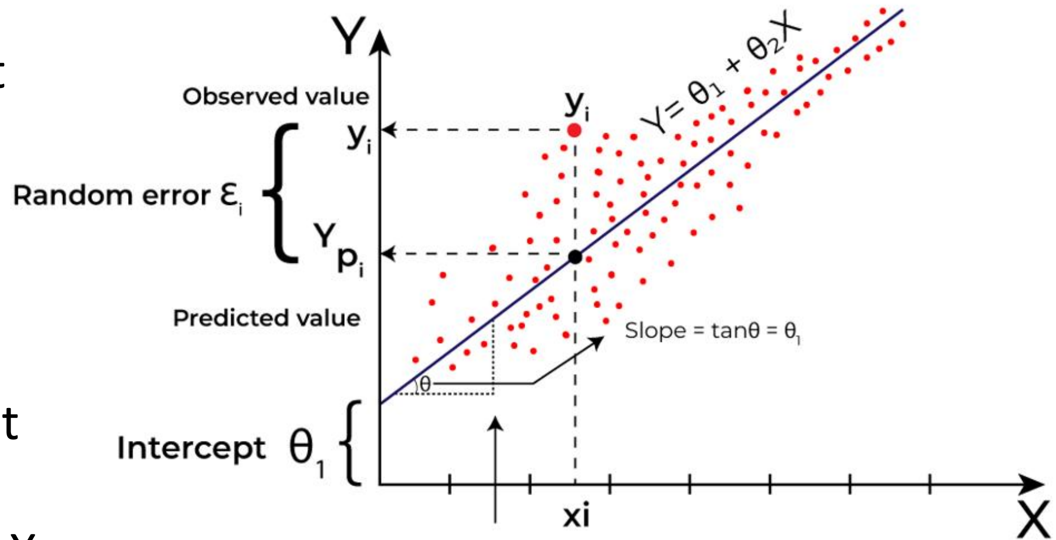
- The independent variables should not be highly correlated with each other.
- Multicollinearity occurs when two or more independent variables are strongly related, making it difficult to isolate the effect of each variable on the dependent variable.
- If multicollinearity exists, it can lead to inaccurate results in multiple linear regression.
- For multiple linear regression, multicollinearity among independent variables can be assessed using the Variance Inflation Factor (VIF). A VIF value greater than 10 indicates high multicollinearity.

What is Regressions Analysis?

- Linear regression is a type of supervised machine learning algorithm that computes the linear relationship between a dependent variable and one or more independent variables by fitting a linear equation that models the relationship between these variables.
- In LR, we use a linear formula in the form of:

  $$Y = B_0 + B_1X$$

  to represent the relationship between an independent variable (X) and a dependent variable (Y).
- In LR, both X and Y are numerical values, for example, X could be the employee's years of experience and Y could be his predicted salary, and we use LR to model that relation.


Y
$Y = \theta_1 + \theta_2 X$
Observed value
y_i
y_i
Random error ε_i
Y_{p_i}
Predicted value
Slope = tanθ = θ_1
θ
Intercept θ_1
xi
X