Linear Regression

1. Introduction

Linear regression is a statistical technique used to model the relationship between a dependent variable (response) and one or more independent variables (predictors). The key idea is to approximate this relationship by a linear function.

- **Dependent variable (Y):** also called outcome, response, or target.
- **Independent variable (X):** also called predictor, explanatory, or regressor.

Regression is used for:

- Prediction
- Estimation
- Hypothesis testing
- Modeling causal relationships

---

2. The Simple Linear Regression Model

For one predictor variable:

$$y=\beta 0+\beta 1x+\epsilon$$

where:

- Y is the dependent variable (the value we want to predict).
- X is the independent variable.
- β0 is the **y-intercept**, the value of Y when X=0.
- β1 is the **slope**, representing the change in Y for a one-unit change in X.
- ϵ is the error term, accounting for the variability in Y that is not explained by X.

Assumptions of Linear Regression

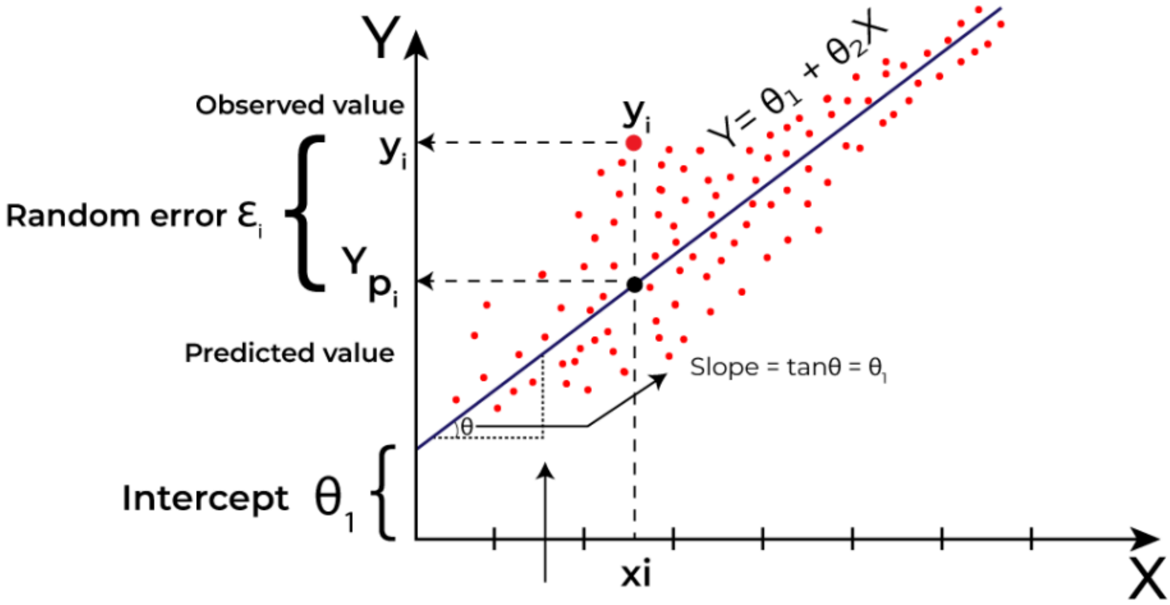
1. **Linearity: Relationship between predictors and target is linear.**

2. **Independence: Observations are independent of each other.**

3. **Homoscedasticity: Constant variance of errors across values of predictors.**

4. **Normality of Errors: Errors (residuals) follow a normal distribution.**
5. **No Multicollinearity: Independent variables should not be highly correlated.**

Example figure:



This is called a probabilistic model because for any fixed x, y is not exactly equal to β0 + β1x but varies around it.

Expected value:

E(Y | x) = β0 + β1 * x

Variance:

Var(Y | x) = σ²

---

3. Example

Suppose the regression line is:

$y = 7.5 + 0.5 * x$

- Interpretation of slope $\beta_1 = 0.5$: On average, for every 1 unit increase in x, y increases by 0.5.
- If $x = 20$, expected y is $7.5 + 0.5*20 = 17.5$.

---

4. Estimating Parameters (Least Squares)

We estimate β_0 and β_1 by minimizing the sum of squared errors:

$SSE = \Sigma (y_i - (\beta_0 + \beta_1 * x_i))^2$

The estimates are:

$\beta_{1\_hat} = \Sigma (x_i - x_{mean})(y_i - y_{mean}) / \Sigma (x_i - x_{mean})^2$

$\beta_{0\_hat} = y_{mean} - \beta_{1\_hat} * x_{mean}$

---

5. Predicted and Residual Values

For each observation x_i, the predicted value is:

$y_{hat\_i} = \beta_{0\_hat} + \beta_{1\_hat} * x_i$

The residual is:

$e_i = y_i - y_{hat\_i}$

Residuals are the difference between observed and predicted values.

---

6. Error Variance and R-squared

Error variance estimate:

$\sigma^2_{hat} = SSE / (n - 2)$

where n is the number of observations.

Coefficient of determination:

$R^2 = 1 - SSE / SST$

where $SST = \Sigma (y_i - y_{mean})^2$.

R^2 measures the proportion of variance in y explained by the regression model.

---

7. Multiple Linear Regression

Extension of simple regression to multiple predictors:

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon$

- Each β_i is a partial regression coefficient: the effect on y when x_i increases by 1, holding other predictors constant.