

Introduction to Linear Models

Introduction to Linear Models

Key concepts, assumptions, and typical applications

1. What Is a Linear Model?

A **linear model** expresses a response (dependent) variable (Y) as a linear combination of one or more predictor (independent) variables ($X_1, X_2, \dots, X_p$) plus an error term:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \epsilon$$

- β_0 – intercept (value of Y when all X 's are 0).
- β_j – slope/coefficients that quantify the effect of each predictor.
- ϵ – random error (captures variation not explained by the predictors).

The model is *linear* in the parameters (β), not necessarily in the raw predictors (e.g., you can include X^2 or $\log X$ as a predictor and still have a linear model).

2. Types of Linear Models

Model	Formulation	Typical Use
Simple Linear Regression	$Y = \beta_0 + \beta_1 X + \epsilon$	One predictor, continuous outcome.
Multiple Linear Regression	$Y = \beta_0 + \sum_{j=1}^p \beta_j X_j + \epsilon$	Several predictors, continuous outcome.
Analysis of Variance (ANOVA)	$Y = \mu + \tau_i + \epsilon$ (categorical predictors)	Compare means across groups.
Analysis of Covariance (ANCOVA)	Combination of ANOVA + regression terms	Adjust group comparisons for covariates.
General Linear Model (GLM)	$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ (matrix notation)	Unified framework for all above; facilitates estimation and inference.

Introduction to Linear Models

Model	Formulation	Typical Use
Generalized Linear Model (extension)	Links a non-normal response distribution (e.g., logistic, Poisson) to a linear predictor	Binary, count, or other non-continuous outcomes.

The focus here is on the classical (Gaussian) linear model.

3. Matrix Representation

Using vectors/matrices simplifies notation and computation.

- **Data matrix** ($\mathbf{X}$) (size $(n \times p)$): each row is an observation, each column a predictor; usually a column of 1's is added for the intercept.
- **Response vector** ($\mathbf{y}$) (size $(n \times 1)$).
- **Parameter vector** ($\boldsymbol{\beta}$) (size $(p \times 1)$).
- **Error vector** ($\boldsymbol{\varepsilon}$) (size $(n \times 1)$).

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

Ordinary Least Squares (OLS) estimate:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

provided $(\mathbf{X}^T \mathbf{X})$ is invertible (full column rank).

4. Core Assumptions

Assumption	Description	Diagnostic Tools
------------	-------------	------------------

Linearity		
-----------	--	--