Fundamentals of Genomic Prediction and Data-Driven Crop Breeding (August 4-8, 2025)

IRRI International Rice Research Institute

Understanding Regression and Ordinary Least Squares

Module 2
August, 2025

Waseem Hussain and Mahender Anumalla
Rice Breeding Innovations Platform
IRRI

Ordinary Least squares (OLS)

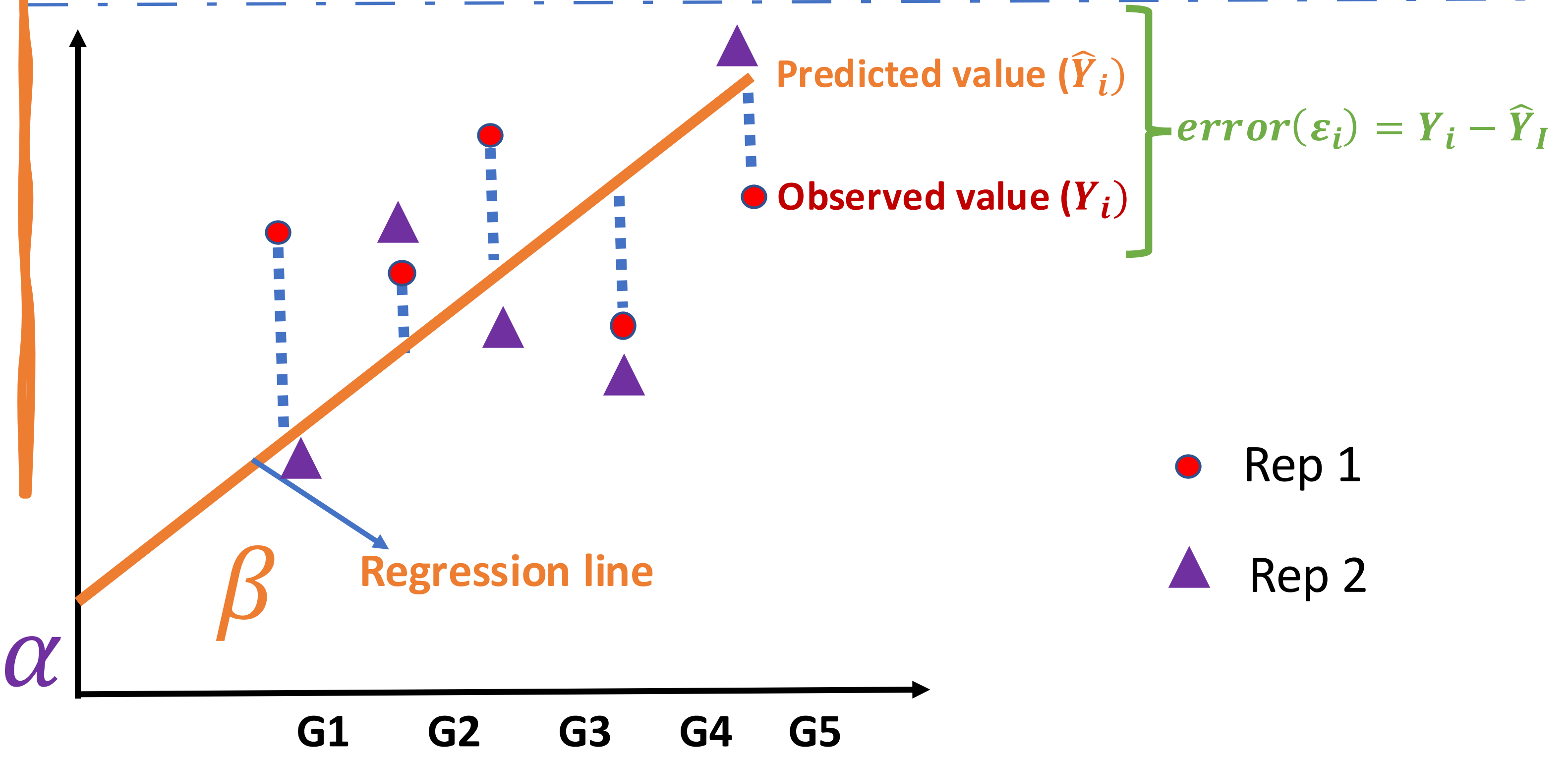
The aim is to estimate α and β (fixed) parameters bminimizing the squared errors

Simple Linear Regression:

$$\hat{Y}_i = \alpha + \beta x_i + \varepsilon_i$$

dependent variable | intercept | coefficient | Independent variable | error

Parameters (un-observed)



OLS Extended to Markers

Simple marker regression model

$$Y_n = \mu_m + X_m\beta_n + \varepsilon$$

y_1
y_2
y_3
y_4
.
.
y_n

Y_n are the phenotypic values of *n*th plants/individuals

μ_m is mean of *n*th marker

	SNP_1	SNP_m
y_{11}	0	0
y_{21}	2	2
y_{31}	2	2
y_{41}	0	0
y_{51}	0	0
.	.	.
.	.	.
.	.	.
y_{m1}	2	2

X_m is a vector or design matrix containing alleles (0,2 or 1) and connecting it to phenotypic values.

β_n is the additive effect of *n*th marker

error for marker assumed $\varepsilon \sim N(0, \sigma_e^2)$, with mean 0 and marker variance σ_e^2

Aim is to minimize residual squares

$$argmin(\varepsilon`\varepsilon) = \text{argmin}(y - X\beta)`(y - x\beta)$$

...

$$\beta = (\acute{X}X)^{-1}\,\acute{X}Y$$

Determines β such that residual squares are minimal called as Least Squares

$$V_\beta = (\acute{X}X)^{-1}\,\sigma_e^2$$

$$\text{where, } \sigma_e^2 = \frac{1}{n-1}\sum(y_i - \beta_i)^2$$

variance-covariance estimate for the sample estimates

Numeric Conversion is Key for Regression

Nucleotide Format

	SNP1	SNP2	SNP3	SNP4	SNP5	.....	SNPm
Allele	**A/G**	**C/T**	**G/A**	**T/C**	**A/G**	**C/T**	**A/G**
Genotype 1	AA	CC	GA	TT	GG	CC	AA
Genotype 2	AA	TT	AA	TT	GG	TT	AA
Genotype 3	AG	TT	GG	TT	GG	TT	AG
Genotype 4	GG	CC	AA	TC	AA	CC	NA
Genotype n	AG	TT	AA	TC	AG	TT	GG

Numeric Format

	SNP1	SNP2	SNP3	SNP4	SNP5	.....	SNPm
Reference Allele	**A/G**	**C/T**	**G/A**	**T/C**	**A/G**	**C/T**	**A/G**
Genotype 1	2	2	1	2	0	2	2
Genotype 2	2	0	0	2	0	0	2
Genotype 3	1	0	2	2	0	0	2
Genotype 4	0	2	0	1	2	2	NA
Genotype n	1	0	0	1	1	0	0

General Linear Model

$$y = X\beta + \varepsilon$$

where,

y = vector of dependent values (observed)

X= Design matrix for observations

β=unknow parameter to estimate

ε=residuals (deviations) and are equal to $y - X\beta$

Ordinary Least Square (OLS)

$$\varepsilon \sim MNV(0, \sigma_\varepsilon^2 I)$$

$$\beta = (\acute{X}X)^{-1}\,\acute{X}Y$$

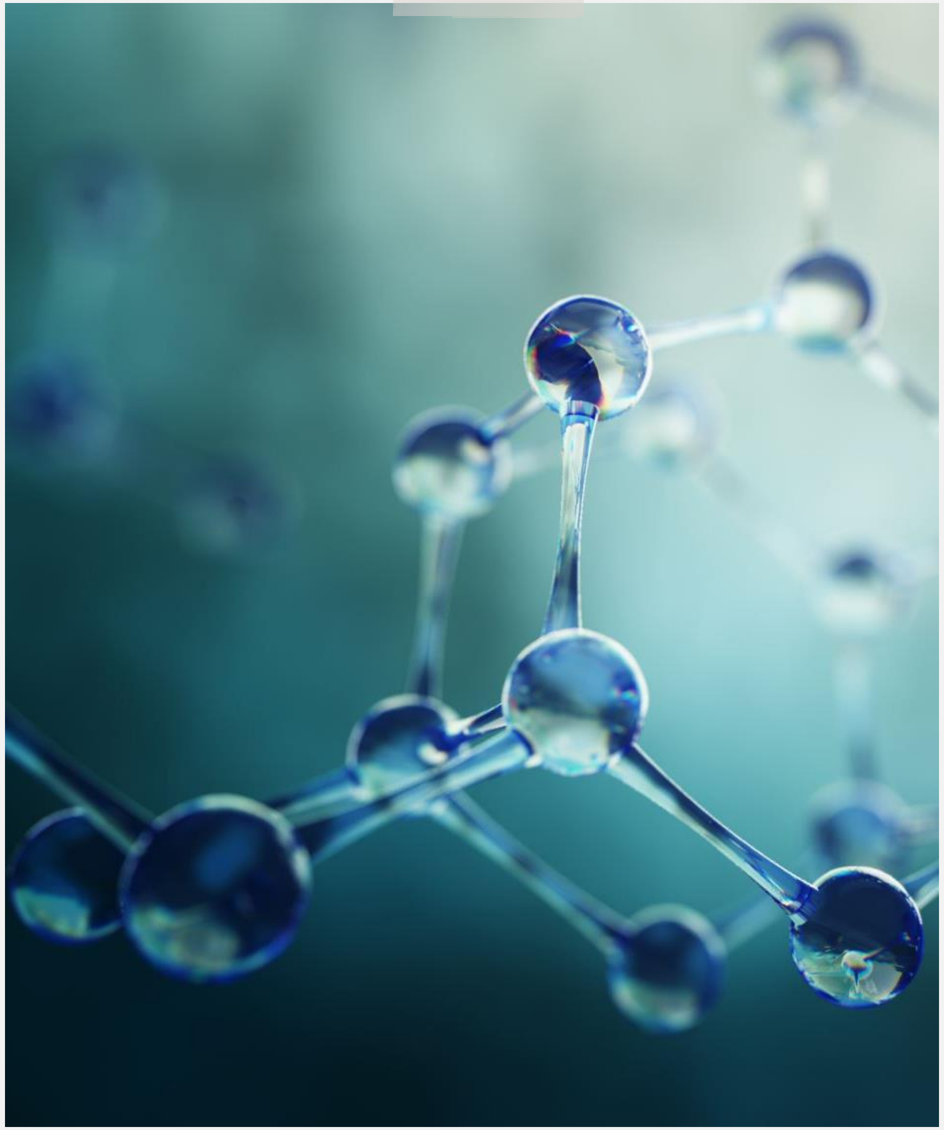
Residuals are homoscedastic and uncorrelated

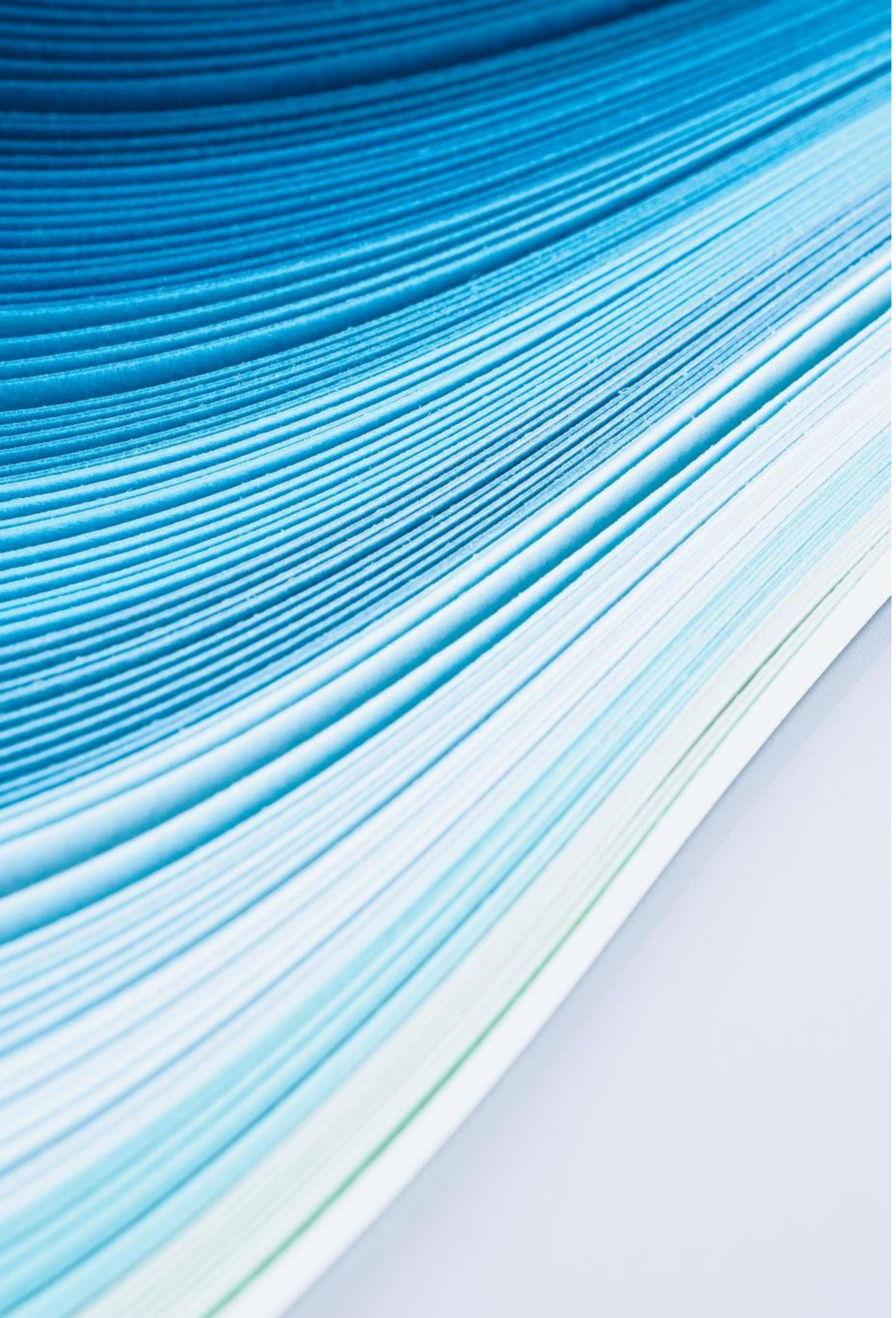
Generalized Least Square (GLS)

$$\varepsilon \sim MNV(0, V)$$

$$\beta = (\acute{X}V^{-1}X)^{-1}\,X\acute{V}^{-1}Y$$

Residuals are heteroscedastic and/or dependent,





OLS is BLUE?

Expected value:

$$E(\widehat{\beta}) = (\acute{X}X)^{-1}\, X\, \acute{E}(Y)$$

$$= (\acute{X}X)^{-1}\, X\acute{}\,(\beta)$$

$$\boldsymbol{E(\widehat{\beta}) = \beta}$$

estimation is true β, and when this condition is met, it is called *unbiased*

When Gauss Markov Theorem is met

1. $E(\varepsilon) = 0\ (expectation\ of\ error\ is\ 0)$
2. $variance\ = I\sigma^2\ (errors\ are\ uncorrleated)$
3. Homoscedasticity of errors

Then, **OLS** → **BLUE**

Baseline Model of GS

$$\mathbf{y} = \boldsymbol{\mu} + \sum_k x_k \,\boldsymbol{\beta}_k + \boldsymbol{\varepsilon}$$

How the marker effects (β) are distributed

Solution

- Markers are fitted as random
- We constrain these markers (penalty).
- What distribution are these markers sampled from (Optimization of the constraints)