Notes on Analysis of Molecular Variance

Xingyuan Su

January 28, 2026

1 General Linear Model

General linear model is expressed as:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

The matrix form of the equation:

$$
\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_n \end{bmatrix}_{n \times 1}
=
\begin{bmatrix}
x_{11} & x_{12} & x_{13} & \cdots & x_{1p} \\
x_{21} & x_{22} & x_{23} & \cdots & x_{2p} \\
x_{31} & x_{32} & x_{33} & \cdots & x_{3p} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
x_{n1} & x_{n2} & x_{n3} & \cdots & x_{np}
\end{bmatrix}_{n \times p}
\begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \vdots \\ \beta_p \end{bmatrix}_{p \times 1}
+
\begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \vdots \\ \epsilon_n \end{bmatrix}_{n \times 1}
$$

The goal of general linear model is to find $\boldsymbol{\beta}$ that minimizes the distance between $\boldsymbol{Y}$ and $\boldsymbol{X}\boldsymbol{\beta}$.

2 Analysis of Molecular Variance

The model was developed using human mitochondrial DNA restriction site data.

A restriction haploptype is the combination of presence (1) or absence (0) of restriction sites with length S.

It can be present as a vector (p is either 1 or 0):

$$\begin{bmatrix} p_1 \\ p_2 \\ p_3 \\ \vdots \\ p_S \end{bmatrix}$$

A linear model is set up as:

$$p_{jig} = p + a_g + b_{ig} + c_{jig}$$

where j represents the haplotype, i represents the population, g represents the group.

Example: There are two populations with two individuals in each population.

A system of equations is written as:

$$p_{11} = p + b_1 + \epsilon$$
$$p_{12} = p + b_1 + \epsilon$$
$$p_{21} = p + b_2 + \epsilon$$
$$p_{22} = p + b_2 + \epsilon$$

Written as matrix form (each value is a column vector of length S):

$$\underbrace{\begin{bmatrix} p_{11} \\ p_{12} \\ p_{21} \\ p_{22} \end{bmatrix}}_{\text{observed values of each individual}} = \underbrace{\begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix}}_{\text{a matrix with linear dependent columns}} \begin{bmatrix} p \\ b_1 \\ b_2 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \end{bmatrix}$$

The matrix with 1 and 0 is used to form the X matrix in general linear model by removing redundant columns.

Two examples of X matrix:

Remove the third column, X matrix is

$$X = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ \underbrace{1}_{\text{intercept}} & 1 \end{bmatrix}$$

The column with all 1 is called the intercept.

Remove the first column, X matrix is

$$X = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}$$

The two X matrices are row equivalent.

New X matrices can also be obtained using elementary row operations.

For example,

$$\begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \xrightarrow[]{R1 \to R1 - R3} \begin{bmatrix} 1 & -1 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \xrightarrow[]{R2 \to R2 - R3} \begin{bmatrix} 1 & -1 \\ 1 & -1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \xrightarrow[R4 \to R4 \times 2]{R3 \to R3 \times 2} \begin{bmatrix} 1 & -1 \\ 1 & -1 \\ 0 & 2 \\ 0 & 2 \end{bmatrix} \xrightarrow[R4 \to R4 + R2]{R3 \to R3 + R1} \begin{bmatrix} 1 & -1 \\ 1 & -1 \\ 1 & 1 \\ 1 & 1 \end{bmatrix}$$

The row equivalent X matrix is $\begin{bmatrix} 1 & -1 \\ 1 & -1 \\ 1 & 1 \\ 1 & 1 \end{bmatrix}$.

Different X matrices give different equations:

$$\underbrace{\begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \end{bmatrix}}_{X} \underbrace{\begin{bmatrix} p_1 \\ p_2 - p_1 \end{bmatrix}}_{\beta} = \underbrace{\begin{bmatrix} p_1 \\ p_1 \\ p_2 \\ p_2 \end{bmatrix}}_{\text{predicted values}} .$$

$$\begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \end{bmatrix} = \begin{bmatrix} p_1 \\ p_1 \\ p_2 \\ p_2 \end{bmatrix}$$

p_1 and p_2 are means of populations 1 and 2.

Note that they have different X matrices and β vector. But they give the same predicted values.

Choosing the columns of X matrix is called **contrasts**.