Regression On Dummy Variables

Peter von Rohr

2022-03-14

Why

- Discrete valued predictor variables like `Breed`
- Assignment of numeric codes to different breeds creates dependencies between expected values of different breeds

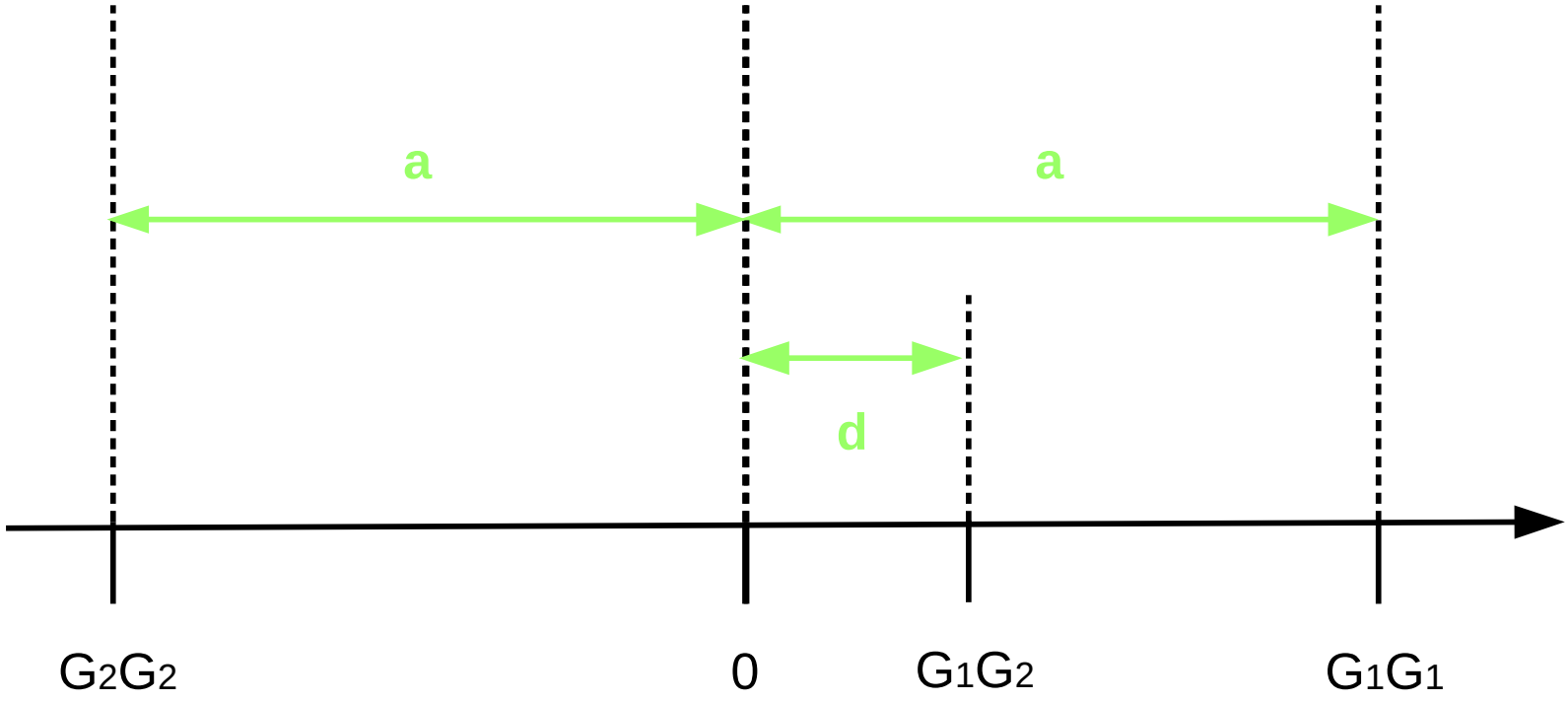
$$
\begin{aligned}
E(\text{BW Angus}) &= b_0 + b_1 \\
E(\text{BW Limousin}) &= b_0 + 2b_1 \\
E(\text{BW Simmental}) &= b_0 + 3b_1
\end{aligned}
$$

- Only estimates are b_0 and b_1
- Usually unreasonable, with one exception

Linear Regression in Genomic Analysis

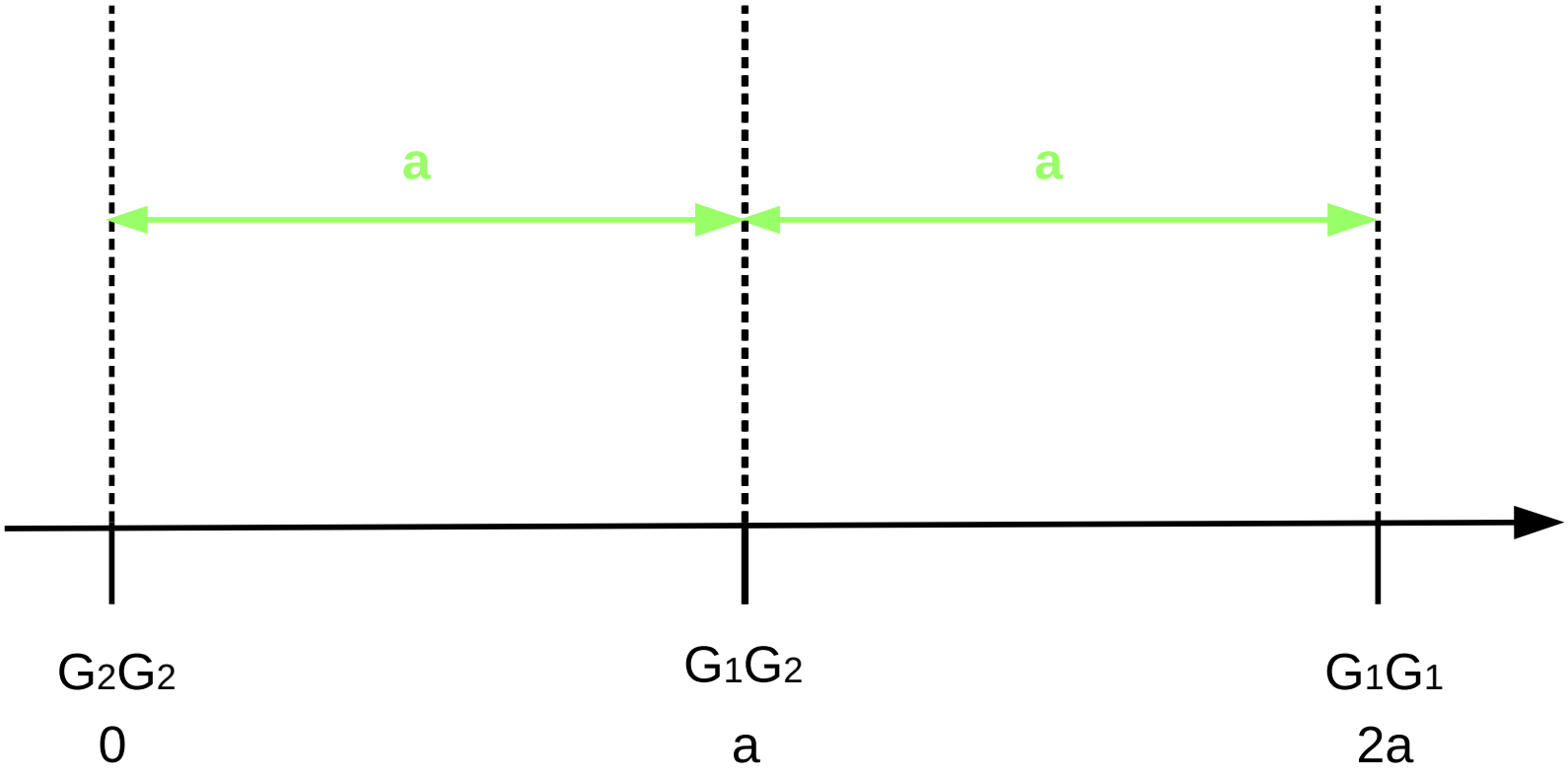
- Regression on the number of positive alleles
- Estimate for slope b_1 corresponds to estimate of marker effect
- Review single-locus model from Quantitative Genetics

Single Locus Model



- Assuming $d = 0 \rightarrow$ genotypic value of $G_1 G_2$ between homozygotes

Modified Single Locus Model



- Transformation of regression on genotypes to regression on number of "positive" alleles (G_1)
- Relationships imposed by regression are meaningful

Relationships

- Expected value for observation for a given genotype

$$E(G_2G_2) = b_0 + 0 * b_1$$
$$E(G_1G_2) = b_0 + 1 * b_1$$
$$E(G_1G_1) = b_0 + 2 * b_1$$

- Differences

$$E(G_1G_2) - E(G_2G_2) = E(G_1G_1) - E(G_1G_2) = b_1$$
$$E(G_1G_1) - E(G_2G_2) = 2b_1$$

Example Dataset

- Exercise 3, Problem 1

Regression On Dummy Variables

- Cases that are not like genomic data
- Example with breeds
- Discrete independent variables are called **Factors** (e.g. Breed)
- Different values that a factor can take are called **Levels**
- Levels for our example factor `Breed` are: `Angus`, `Limousin` and `Simmental`

Levels To Independent Variables

Use "separate" x-variable for each level, hence each of the breeds

Breed	Independent Variable
Angus	x_1
Limousin	x_2
Simmental	x_3

Model

- Observation y_{ij} stands for birth weight for animal j in breed i

$$
\begin{aligned}
y_{11} &= b_0 + b_1 * 1 + b_2 * 0 + b_3 * 0 + e_{11} \\
y_{12} &= b_0 + b_1 * 1 + b_2 * 0 + b_3 * 0 + e_{12} \\
\cdots &= \cdots \\
y_{33} &= b_0 + b_1 * 0 + b_2 * 0 + b_3 * 1 + e_{33}
\end{aligned}
$$

- Sort animals according to breeds

Matrix - Vector Notation

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$$