Regression On Dummy Variables

Peter von Rohr

2023-03-27

Why

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

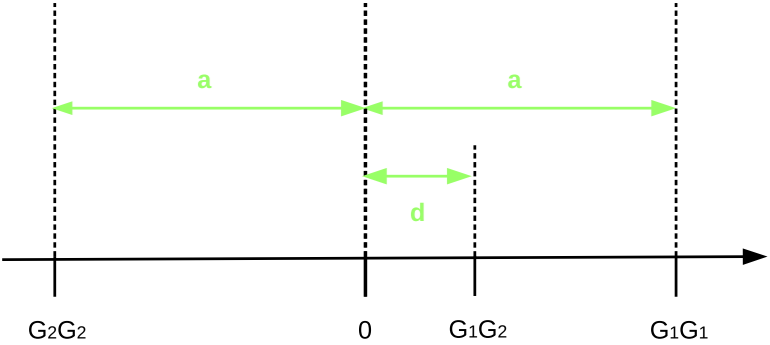
$$
\begin{aligned}
E(\text{BW Angus}) &= b_0 + b_1 \\
E(\text{BW Limousin}) &= b_0 + 2b_1 \\
E(\text{BW Simmental}) &= b_0 + 3b_1
\end{aligned}
$$

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

Linear Regression in Genomic Analysis

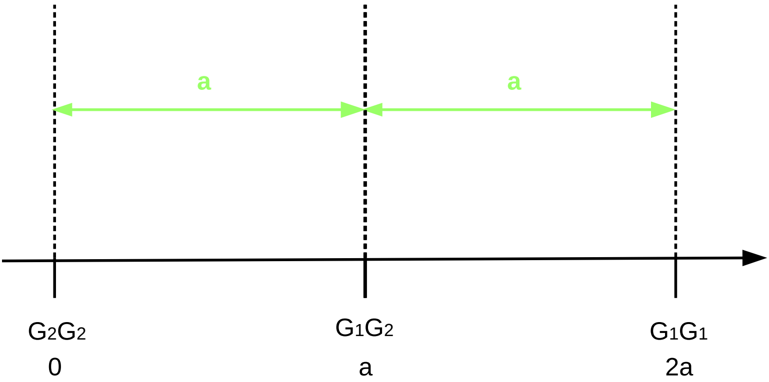
- Regression on the number of positive alleles
- Estimate for slope b_1 corresponds to estimate of marker effect
- Review single-locus model from Quantitative Genetics

Single Locus Model



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

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

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

- Differences

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

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

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

Models Not Of Full Rank

- Model

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

- Least squares normal equations

$$\mathbf{X}^T\mathbf{X}\mathbf{b}^{(0)} = \mathbf{X}^T\mathbf{y}$$

Solutions

- matrix $\mathbf{X}$ not of full rank, use Matrix::rankMatrix() to check
- $\mathbf{X}^T\mathbf{X}$ cannot be inverted
- solution

$$\mathbf{b}^{(0)} = (\mathbf{X}^T\mathbf{X})^{-}\mathbf{X}^T\mathbf{y}$$

where $(\mathbf{X}^T\mathbf{X})^{-}$ stands for a **generalized inverse**

Generalized Inverse

- matrix $\mathbf{G}$ is a generalized inverse of matrix $\mathbf{A}$, if

$$\mathbf{AGA} = \mathbf{A}$$

$$(\mathbf{AGA})^T = \mathbf{A}^T$$

- Use `MASS::ginv()` in R

Systems of Equations

- For a consistent system of equations

$$Ax = y$$

- Solution

$$x = Gy$$

if G is a generalized inverse of A.

$$x = Gy$$
$$Ax = AGy$$
$$Ax = AGAx$$

Non Uniqueness

- Solution $x = Gy$ is not unique

$$\tilde{\mathbf{x}} = \mathbf{G}\mathbf{y} + (\mathbf{G}\mathbf{A} - \mathbf{I})\mathbf{z}$$

yields a different solution for an arbitrary vector $\mathbf{z}$

$$\mathbf{A}\tilde{\mathbf{x}} = \mathbf{A}\mathbf{G}\mathbf{y} + (\mathbf{A}\mathbf{G}\mathbf{A} - \mathbf{A})\mathbf{z}$$

Least Squares Normal Equations

- Instead of $Ax = y$, we have

$$\mathbf{X}^T\mathbf{X}\mathbf{b}^{(0)} = \mathbf{X}^T\mathbf{y}$$

- With generalized inverse $\mathbf{G}$ of $\mathbf{X}^T\mathbf{X}$

$$\mathbf{b}^{(0)} = \mathbf{G}\mathbf{X}^T\mathbf{y}$$

is a solution to the least squares normal equations

Parameter Estimator

But $\mathbf{b}^{(0)}$ is not an estimator for the parameter $\mathbf{b}$, because

- it is not unique
- Expectation $E(\mathbf{b}^{(0)}) = E(\mathbf{G}\mathbf{X}^T\mathbf{y}) = \mathbf{G}\mathbf{X}^T\mathbf{X}\mathbf{b} \neq \mathbf{b}$

Estimable Functions

Animal	Breed	Observation
1	Angus	16
2	Angus	10
3	Angus	19
4	Limousin	27
5	Simmental	11
6	Simmental	13

Model

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

$$\mathbf{y} = \begin{bmatrix} 16 \\ 10 \\ 19 \\ 27 \\ 11 \\ 13 \end{bmatrix}, \ \mathbf{X} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix} \text{ and } \mathbf{b} = \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix}$$

Normal Equations

$$X^T X b^{(0)} = X^T y$$

$$\begin{bmatrix} 6 & 3 & 1 & 2 \\ 3 & 3 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 2 & 0 & 0 & 2 \end{bmatrix} \begin{bmatrix} \mu^0 \\ \alpha_1^0 \\ \alpha_2^0 \\ \alpha_3^0 \end{bmatrix} = \begin{bmatrix} 96 \\ 45 \\ 27 \\ 24 \end{bmatrix}$$

A solution

$$b^{(0)} = \begin{bmatrix} 13.5 \\ 1.5 \\ 13.5 \\ -1.5 \end{bmatrix}$$

Solutions to Normal Equations

Elements of Solution	b_1^0	b_2^0	b_3^0	b_4^0
μ^0	16	14	27	-2982
α_1^0	-1	1	-12	2997
α_2^0	-4	-2	-15	2994
α_3^0	11	13	0	3009

Functions of Solutions

Linear Function	b_1^0	b_2^0	b_3^0	b_4^0
$\alpha_1^0 - \alpha_2^0$	3.0	3.0	3.0	3.0
$\mu^0 + \alpha_1^0$	15.0	15.0	15.0	15.0
$\mu^0 + 1/2(\alpha_2^0 + \alpha_3^0)$	19.5	19.5	19.5	19.5

- $\alpha_1^0 - \alpha_2^0$: estimate of the difference between breed effects for Angus and Simmental
- $\mu^0 + \alpha_1^0$: estimate of the general mean plus the breed effect of Angus
- $\mu^0 + 1/2(\alpha_2^0 + \alpha_3^0)$: estimate of the general mean plus mean effect of breeds Simmental and Limousin

Definition of Estimable Functions

$$\mathbf{q}^T\mathbf{b} = \mathbf{t}^T E(\mathbf{y})$$

- Why is $\mathbf{q}^T\mathbf{b}$ estimable?
- Based on the defintion of $\mathbf{b}$ and $E(\mathbf{y})$

$$\mathbf{q}^T\mathbf{b} = \mathbf{t}^T\mathbf{X}\mathbf{G}\mathbf{X}^T\mathbf{y}$$

where $\mathbf{X}\mathbf{G}\mathbf{X}^T$ is the same for all choices of $\mathbf{G}$

Examples

$$E(y_{1j}) = \mu + \alpha_1$$

with $\mathbf{t}^T = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \end{bmatrix}$ and $\mathbf{q}^T = \begin{bmatrix} 1 & 1 & 0 & 0 \end{bmatrix}$

$$E(y_{2j}) = \mu + \alpha_2$$

$$E(y_{3j}) = \mu + \alpha_3$$

Property

Based on the definition, the following property can be derived

$$\mathbf{q}^t = \mathbf{t}^T\mathbf{X}$$

with the definition of an estimable function $\mathbf{q}^T\mathbf{b}$, we get

$$\mathbf{q}^T\mathbf{b} = \mathbf{t}^T E(\mathbf{y})$$
$$\mathbf{q}^T\mathbf{G}\mathbf{X}^T\mathbf{y} = \mathbf{t}^T\mathbf{X}\mathbf{G}\mathbf{X}^T\mathbf{y}$$

hence for any $\mathbf{G}$, $\mathbf{q}^t = \mathbf{t}^T\mathbf{X}$ which is helpful to find $\mathbf{q}$ for a given $\mathbf{t}$

Test

When we want to test whether a certain vector $\mathbf{q}$ can establish an estimable function, we can test wheter

$$\mathbf{q}^T\mathbf{H} = \mathbf{q}^T$$

with $\mathbf{H} = \mathbf{G}\mathbf{X}^T\mathbf{X}$

Setting $\mathbf{q}^T = \mathbf{t}^T\mathbf{X}$, we get

$$\mathbf{q}^T\mathbf{H} = \mathbf{t}^T\mathbf{X}\mathbf{H} = \mathbf{t}^T\mathbf{X} = \mathbf{q}^T$$