Fixed Linear Effects Models

Peter von Rohr

2024-03-18

Extension of Dataset on Body Weight

Animal	BC	Body Weight	Breed
1	176	471	Angus
2	177	463	Angus
3	178	481	Simmental
4	179	470	Angus
5	179	496	Simmental
6	180	491	Simmental
7	181	518	Limousin
8	182	511	Limousin
9	183	510	Limousin
10	184	541	Limousin

Include Breed into Model

- Breed has an influence on body weight
- Predictor variables must be numeric
- Breed must be converted to numeric code
- Assignment of codes to breeds is rather arbitrary

Breed Codes

Code	Breed
1	Angus
2	Limousin
3	Simmental

In R: Encoding based on alpha-numeric order of factor names

```
levels(as.factor(tbl_bw_bc_breed$Breed))
```

```
[1] "Angus"     "Limousin"  "Simmental"
```

```
as.integer(as.factor(tbl_bw_bc_breed$Breed))
```

```
 [1] 1 1 3 1 3 3 2 2 2 2
```

Dataset with Breed Codes

Animal	Body Weight	Breed	Breed Code
1	471	Angus	1
2	463	Angus	1
3	481	Simmental	3
4	470	Angus	1
5	496	Simmental	3
6	491	Simmental	3
7	518	Limousin	2
8	511	Limousin	2
9	510	Limousin	2
10	541	Limousin	2

Modelling Effect of Breed

- Simplification: “breed” is the only predictor, ignore BC
- Take breed code of animal i as the predictor value x_i
- Expected body weight (y_i) for animal i

$$E(y_i) = b_0 + b_1 x_i$$

Problems

- Nothing wrong with previous model
- But the following relations might give a hint to some problems

$$\left.\begin{array}{l}\text{Animal i of breed Angus}\\ \text{Animal j of breed Limousin}\\ \text{Animal k of breed Simmental}\end{array}\right\}\rightarrow\left\{\begin{array}{l}E(y_i) = b_0 + b_1 * 1\\ E(y_j) = b_0 + b_1 * 2\\ E(y_k) = b_0 + b_1 * 3\end{array}\right.$$

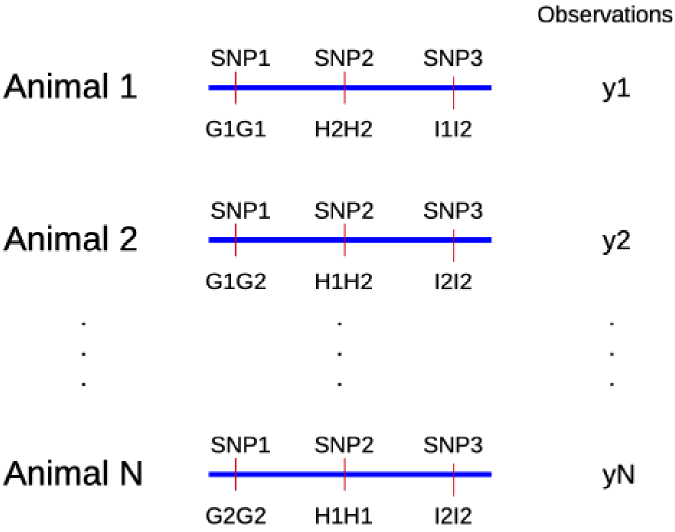
This means, for expected differences between body weights of animals of different breeds

$$E(y_j) - E(y_i) = E(y_k) - E(y_j) = b_1$$
$$E(y_k) - E(y_i) = 2 * b_1$$

Consequences

- Allocation of numerical codes imposes relations between expected values
- Relations might be unreasonable
- Regression analysis only yields estimates for b_0 and b_1, effects of other breeds are determined
- Conclusion: regression on numerical codes of discrete variables are in most cases unreasonable
- Exception: Estimation of marker effects

Linear Regression Analysis for Genomic Data



Observations

Animal 1
SNP1 SNP2 SNP3
G1G1 H2H2 I1I2
y1

Animal 2
SNP1 SNP2 SNP3
G1G2 H1H2 I2I2
y2

.
.
.

Animal N
SNP1 SNP2 SNP3
G2G2 H1H1 I2I2
yN

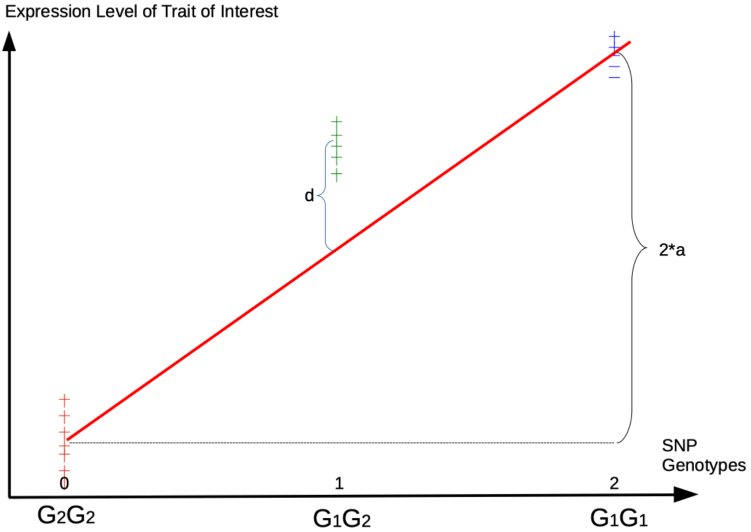
Marker Effect Estimation

- Assume: marker and QTL are very close, such they can no longer be distinguished
- Fit regression of observations (y) on marker genotypes of locus G
- Assume G_1 is the allele with a positive effect on observed trait
- Use the following encoding of marker genotypes to numeric values

Genotype	Code
G_1G_1	2
G_1G_2	1
G_2G_2	0

→ Biological meaning of genotype code: count number of G_1 alleles

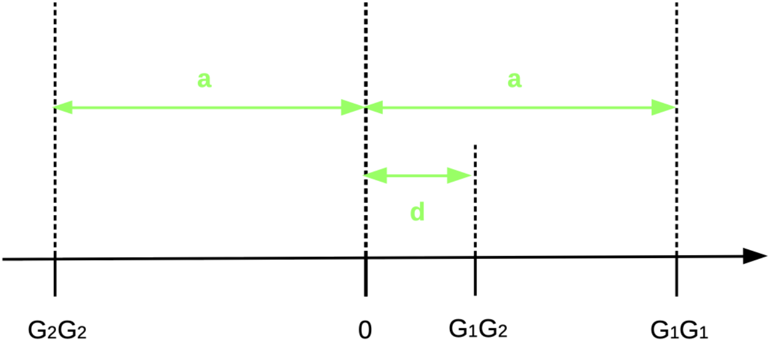
Genomic Regression



Model Fit

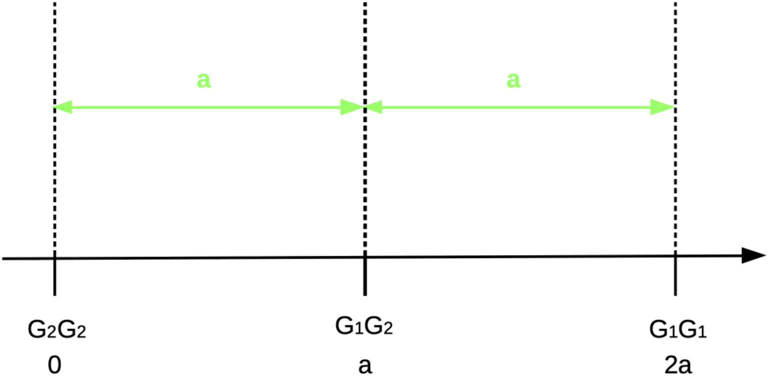
- Take only homozygous genotypes G_1G_1 and G_2G_2 from dataset
- Why only homozygotes? Look at extreme values for d with over- and under-dominance
- Fit regression line and compute marker effect a
- Shift x- axis, such that homozygotes have values $-a$ and a
- Compute d as deviation of heterozygotes from 0

Single Locus Model



- Assuming $d = 0 \rightarrow$ genotypic value of G_1G_2 between homozygotes
- Shifting origin to genotypic value of G_2G_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 5, Problem 1

Animal	SNP G	SNP H	Observation
1	G_1G_1	H_1H_2	510
2	G_1G_2	H_1H_1	528
3	G_1G_2	H_1H_1	505
4	G_1G_1	H_2H_2	539
5	G_1G_1	H_1H_1	530
6	G_1G_2	H_1H_2	489
7	G_1G_2	H_2H_2	486
8	G_2G_2	H_1H_1	485
9	G_1G_2	H_2H_2	478
10	G_2G_2	H_1H_2	479

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

- Instead of $E(y_i) = b_0 + b_1 * x_i$
- Observation y_{ij} stands for birth weight for animal j in breed i

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

- Sort animals according to breeds

Sorted Data

Animal	Body Weight	Breed
1	471	Angus
2	463	Angus
4	470	Angus
7	518	Limousin
8	511	Limousin
9	510	Limousin
10	541	Limousin
3	481	Simmental
5	496	Simmental
6	491	Simmental

Matrix - Vector Notation

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

with

- vectors y and e defined as in linear regression
- vector b contains intercept b_0 and separate effects for each breed

$$b = \left[\begin{array}{c} b_0 \\ b_{Angus} \\ b_{Limousin} \\ b_{Simmental} \end{array}\right]$$

Matrix X

- Matrix X is an incidence matrix linking factor levels to observations

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

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

$$Ab = r$$

- Solution

$$b = Gr$$

if G is a generalized inverse of A.

$$b = Gr$$

$$Ab = AGr$$

$$Ab = AGAb$$

Non Uniqueness

- Solution $b = Gr$ is not unique

$$\tilde{\mathbf{b}} = \mathbf{G}\mathbf{r} + (\mathbf{G}\mathbf{A} - \mathbf{I})\mathbf{z}$$

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

$$\mathbf{A}\tilde{\mathbf{b}} = \mathbf{A}\mathbf{G}\mathbf{r} + (\mathbf{A}\mathbf{G}\mathbf{A} - \mathbf{A})\mathbf{z}$$

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

$\rightarrow$ Require different concept: estimable functions