Fixed Linear Effects Models

general form

regression

- regression on dummy variables
- models of non-full rank

rank of X

Peter von Rohr

2022-03-07

Extension of Dataset on Body Weight

response y

additional predictor

Animal	BC	Body Weight	BCS	HEI	Breed	add 1 column in X
1	176	471	5.0	161	Angus	→ 1
2	177	463	4.2	121	Angus	1
3	178	481	4.9	157	Simmental	3
4	179	470	3.0	165	Angus	1
5	179	496	6.8	136	Simmental	3
6	180	491	4.9	123	Simmental	3
7	181	518	4.4	163	Limousin	2
8	182	511	4.4	149	Limousin	1
9	183	510	3.5	143	Limousin	
10	184	541	4.7	130	Limousin	2

Include Breed into Model

predictors are columns in X ; components of X are numeric

$$X = \begin{bmatrix} 1 & 176 & \cdot & \cdot & \cancel{\text{Angus}} \\ 1 & 178 & & & \\ 1 & \vdots & & & \\ \vdots & & & & \\ 1 & & & & \end{bmatrix}$$

- 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

$$X = \begin{array}{c} \begin{array}{cccc} & & & \text{Sock} \end{array} \\ \begin{bmatrix} 1 & - & \cdots & 1 \\ 1 & & & 1 \\ & & & 3 \\ & & & \end{bmatrix} \end{array}$$

Code	Breed
1 ↔	Angus → less BW
2 ↔	Limousin → ?
3 ↔	Simmental

Assignment arbitrary

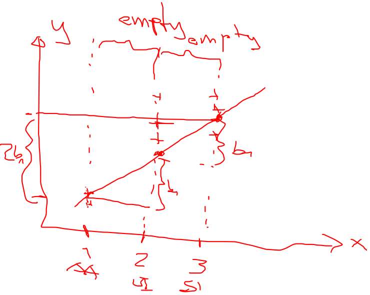
Modelling Effect of Breed

$$X = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ \vdots & 3 \\ & \vdots \\ 1 & 2 \end{bmatrix}$$

numeric code

- Simplification: “breed” is the only predictor
- Expected body weight (y_i) for animal i

$$E(y_i) = \underbrace{b_0}_{\text{intercept}} + \overset{\text{reg. coef}}{b_1}\underbrace{x_i}_{\text{breed code}}$$



Problems

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

$$\left.\begin{aligned}
\rightarrow\; \overbrace{E(\text{BW Angus})}^{\text{Expected value for Body Weight}} &= \underline{b_0 + 1\cdot b_1} \\
\rightarrow\; E(\text{BW Limousin}) &= b_0 + 2b_1 \\
E(\text{BW Simmental}) &= b_0 + 3b_1
\end{aligned}\right\}\; E[y_i] = b_0 + b_1 \underset{\underset{\text{breed code}}{\downarrow}}{x_i}$$

This means, for example, that

$$\underbrace{E(\text{BW Limousin})}_{} \overset{b_0 + 2b_1 \;-\; [b_0 + b_1] \;=\; b_0 - b_0 + 2b_1 - b_1 = b_1}{} - E(\text{BW Angus}) =$$

$$\overset{b_0 + 3b_1 \;-\; [b_0 + 2b_1] \;=\; b_1}{E(\text{BW Simmental}) - E(\text{BW Limousin})}$$

$$E(\text{BW Simmental}) - E(\text{BW Angus}) =$$

$$2\,[E(\text{BW Limousin}) - E(\text{BW Angus})]$$

Consequences

Regression → y X

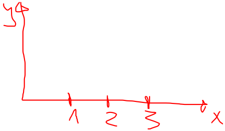
y → BW

continuous

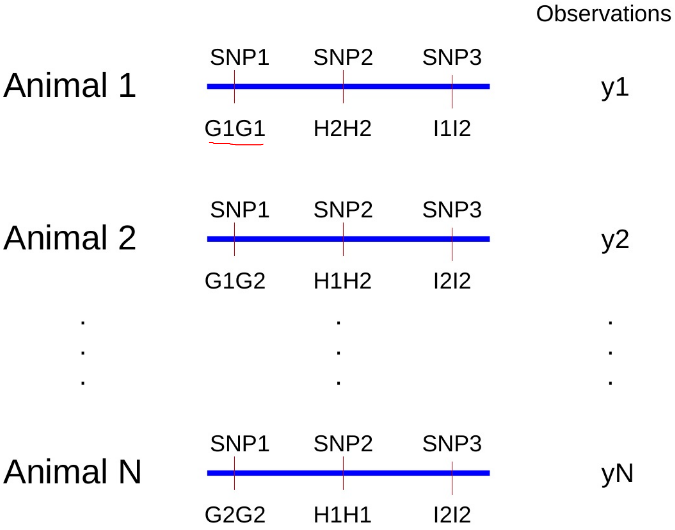
X → BC, BC, HEI

- 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

Breed #2,75



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

.
.
.

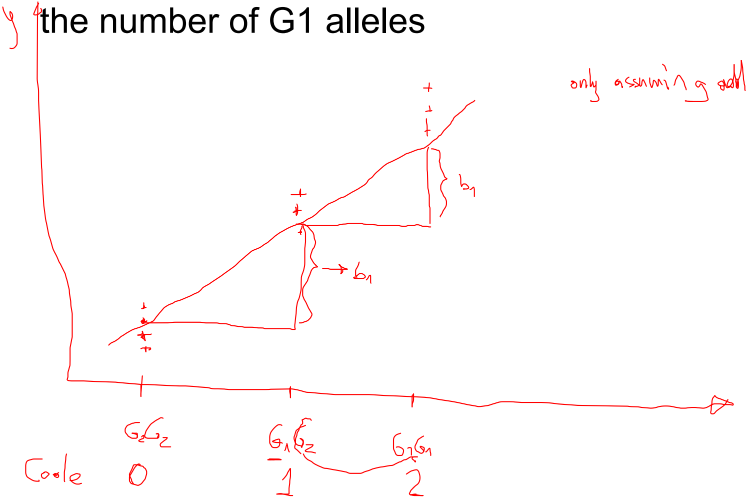
Animal N

SNP1 SNP2 SNP3

G2G2 H1H1 I2I2

yN

For a single SNP: Assuming G1 to be the favorable allele, the numeric code for a given genotype will be the number of G1 alleles



only assuming add. Model