Fixed Linear Effects Models

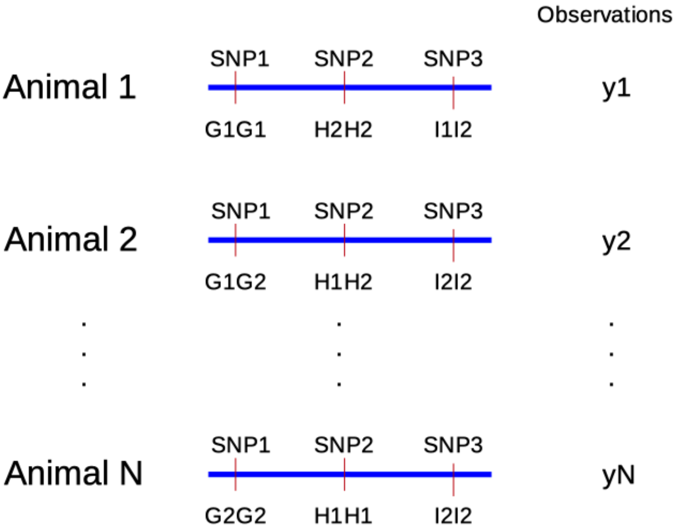
Peter von Rohr

24.02.2020

Background

- Given a population of N animals
- Each animal has information on genotypes at loci G, H and I
- Each animal has an observation for one quantitative trait of interest y
- **Goal**: Predict genomic breeding values

Data



	SNP1	SNP2	SNP3	Observations
Animal 1	G1G1	H2H2	I1I2	y1
Animal 2	G1G2	H1H2	I2I2	y2
⋮		⋮		⋮
Animal N	G2G2	H1H1	I2I2	yN

Two Types Of Models

1. **Genetic** Model: How can we decompose the phenotype into genetic part and non-genetic environmental part
2. **Statistical** Model: How to estimate unknown parameters from a dataset

Goals:

1. Use genetic model to show how observations and genetic information can be used to predict breeding values.
2. Use statistical techniques to do the prediction

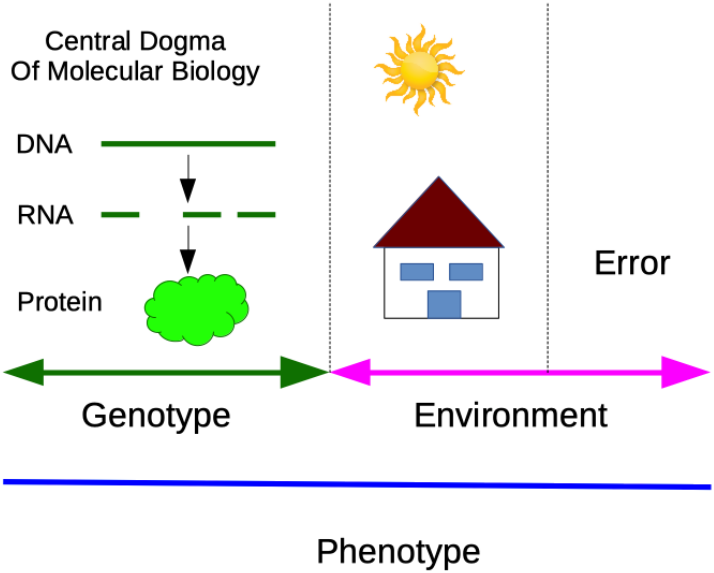
Genetic Model

- simple model from quantitative genetics to split phenotypic observation into
  - genetic part g and
  - environmental part e

$$y = g + e$$

- environment: split into
  - known environmental factors: `herd`, `year`, ... (β)
  - unknown random error (ϵ)
- polygenic model: use a finite number of loci to model genetic part of phenotypic observation

Genetic Model (II)



Polygenic Model

- Component g can be decomposed into contributions g_j of single loci

$$g = \sum_{j=1}^{k} g_j$$

- Assume that loci are additive, hence genotypic values g_j depends on a_j with $d_j = 0$
- Genotypic values at locus j can either be $-a_j$, 0 or a_j
- Breeding values based on locus j depends on a_j.

Genotypic Value

- Genotypic value g_i for animal i over all loci

$$g_i = M_i \cdot a$$

where M_i is a row vector with elements -1, 0 and 1 and a is the vector of all genotypic values of the positive homozygous genotypes of all loci.

Phenotypic Value

- Collecting all components for an observation y_i for animal i

$$y_i = W_i \cdot \beta + M_i \cdot a + \epsilon_i$$

- all animals in the population

$$y = W \cdot \beta + M \cdot a + \epsilon$$

- combining $b^T = \left[\begin{array}{cc} \beta & a \end{array}\right]$ and $X = \left[\begin{array}{cc} W & M \end{array}\right]$

$$y = X \cdot b + \epsilon$$

Statistical Model

- genetic model from statistics point of view
- phenotypic observation as response y
- vector b (known environment and genotypic values) as unknown parameter
- fixed predictor variales in matrix X
- vector ϵ as random error terms

$\rightarrow$ Fixed Linear Effects Model

Parameter Estimation

- use regression model
- regression means both response and predictors are continuous
- example dataset: body weight on breast circumference

Regression Dataset

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

Regression Model

- response y: body weight
- predictor x: breast circumference
- model for observation y_i

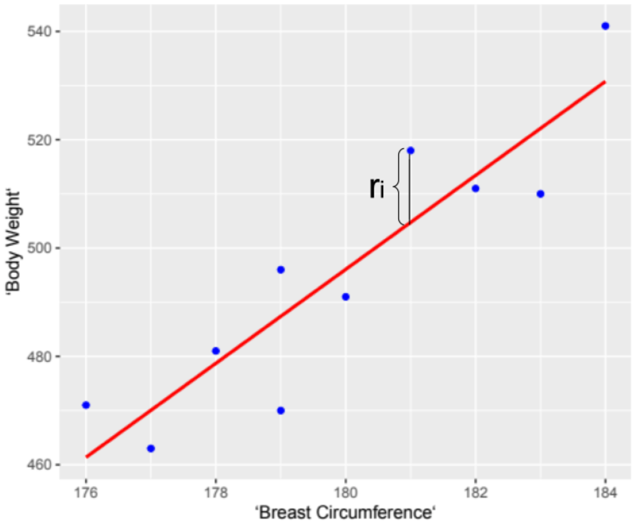
$$y_i = x_i * b + \epsilon_i$$

- meaning of b: change x_i by one unit $\rightarrow$ y_i changes on average by b units.
- use case: measure x_{N+1} for animal $N+1$ with unknown weight and use b to predict y_{N+1}

Least Squares

- How to find b such that y is best approximated by x
- Residuals $r_i = y_i - x_i * \hat{b}$
- Minimization of sum of squared residuals (LS)
- Use $\hat{b}$ at minimal LS as estimate

LSQ Diagram


540
520
500
480
460
'Body Weight'
r_i
176
178
180
182
184
'Breast Circumference'

Sum of squared residuals

$$LS = \sum_{i=1}^{n} r_i^2$$

- In matrix-vector notation with r denoting the vector of all residuals

$$LS = ||r||^2 = r^T r$$

where $||.||$ stands for the norm ("length in 2D") of a vector

- Replacing r with $r = y - X\hat{b}$

$$LS = (y - X\hat{b})^T (y - X\hat{b}) = y^T y - y^T X\hat{b} - \hat{b}^T X^T y + \hat{b}^T X^T X\hat{b}$$

Minimization

- Set partial derivative of LS with respect to $\hat{b}$ to 0

$$\frac{\partial LS}{\partial \hat{b}} = -y^T X - y^T X + 2\hat{b}^T X^T X = 0$$

- Take the $\hat{b}$ that satisfies the above equation as the least squares estimate $\hat{b}_{LS}$

$$X^T X \hat{b}_{LS} = X^T y$$

- Solution

$$\hat{b}_{LS} = (X^T X)^{-1} X^T y$$

Variance of Error Terms

- Least Squares Procedure does not yield an estimate for σ^2
- The estimator based on the residuals

$$\hat{\sigma^2} = \frac{1}{n-p} \sum_{i=1}^{n} r_i^2$$

Different Types of Regressions

- Regression through the origin

$$y_i = x_i * b + e_i$$

- Regression with intercept

$$y_i = b_0 + x_i * b + e_i$$

Predictions

- One of the use-cases for regression is **prediction**
- Prediction means that given a regression model with estimated regression coefficients based on a data set, values of responses are to be predicted for new predictor values (x_{new})

$$\hat{y} = x_{new} * \hat{b}$$

- No predictions outside of the range of x used to estimate $\hat{b}$

Multiple Linear Regression

- Use more than one predictor variable
- Example: Conformation traits `BCS` and `HEI` besides `BC`
- New model:

$$y_i = b_0 + BC_i * b_1 + BCS_i * b_2 + HEI_i * b_3 + e_i$$

- In matrix vector notation:

$$y = Xb + e$$

with $b^T = \left[\begin{array}{cccc} b_0 & b_1 & b_2 & b_3 \end{array} \right]$

New data set

Table 2: Dataset for Multiple Linear Regression

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

Goal

- Find solution for $\hat{b}_{LS}$
- Same principle of least squares as with simple linear regression
- Different dimensions for X and b

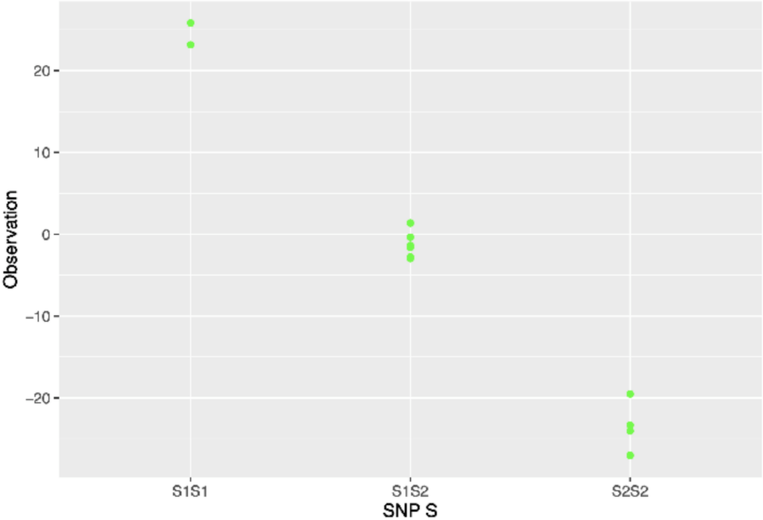
$\rightarrow$ Problem 1 in Exercise 2

Regression on Dummy Variables

- What happens when predictor variables X are no longer continuous
- Examples: SNP-Genotypes
- X can only take a few discrete values, e.g., $0, 1$ or $-1, 0, 1, \ldots$

$\rightarrow$ regression on dummy variables or just general fixed linear model.

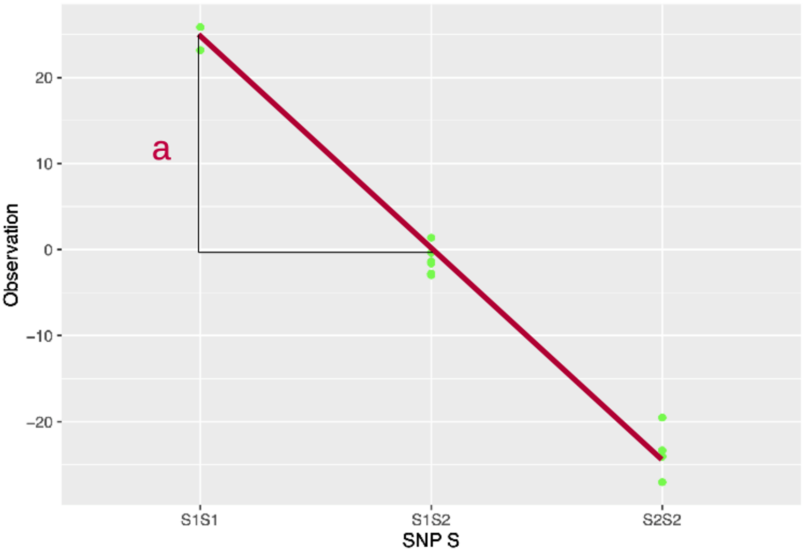
Example: SNP-Data



Goal

- Same as in linear regression: fit line through points such that least squares criterion holds
- Interpretation: Difference between effect levels
- For SNP-data: differences correspond to marker effects

Dummy Regression Line



Problem

- In many datasets X does not have full column-rank
- That means some columns of X show linear dependence
- As a consequence of that $(X^T X)$ cannot be inverted

Solution

- Use a generalised inverse $(X^T X)^-$ to get a solution $\hat{b}_{LS}$ for least squares normal equations
- Use estimable functions of $\hat{b}_{LS}$ which are independent of the choice of $(X^T X)^-$
- One example for estimable functions are differences between effect levels
- For example of SNP-data these correspond to marker effects.

Generalised Inverse

- Reminder: the (ordinary) inverse A^{-1} of A is given by $A^{-1}A = I$, but A^{-1} exists only, if A is of full rank.
- A generalised inverse G of matrix A satisfies: $AGA = A$
- For the system of equations $Ax = y$, the vector $x = Gy$ is a solution, if $AGA = A$
- For a generalised inverse G of A, the system of equation $Ax = y$ has solutions

$$\tilde{x} = Gy + (GA - I)z$$

for an arbitrary vector z.

Estimable Functions

- linear function of the parameter (b) that is identical to linear function of expected values of observations y, i.e.,

$$q^T b = t^T E(y)$$

- estimable functions are invariant (do not change) with different generalised inverses.