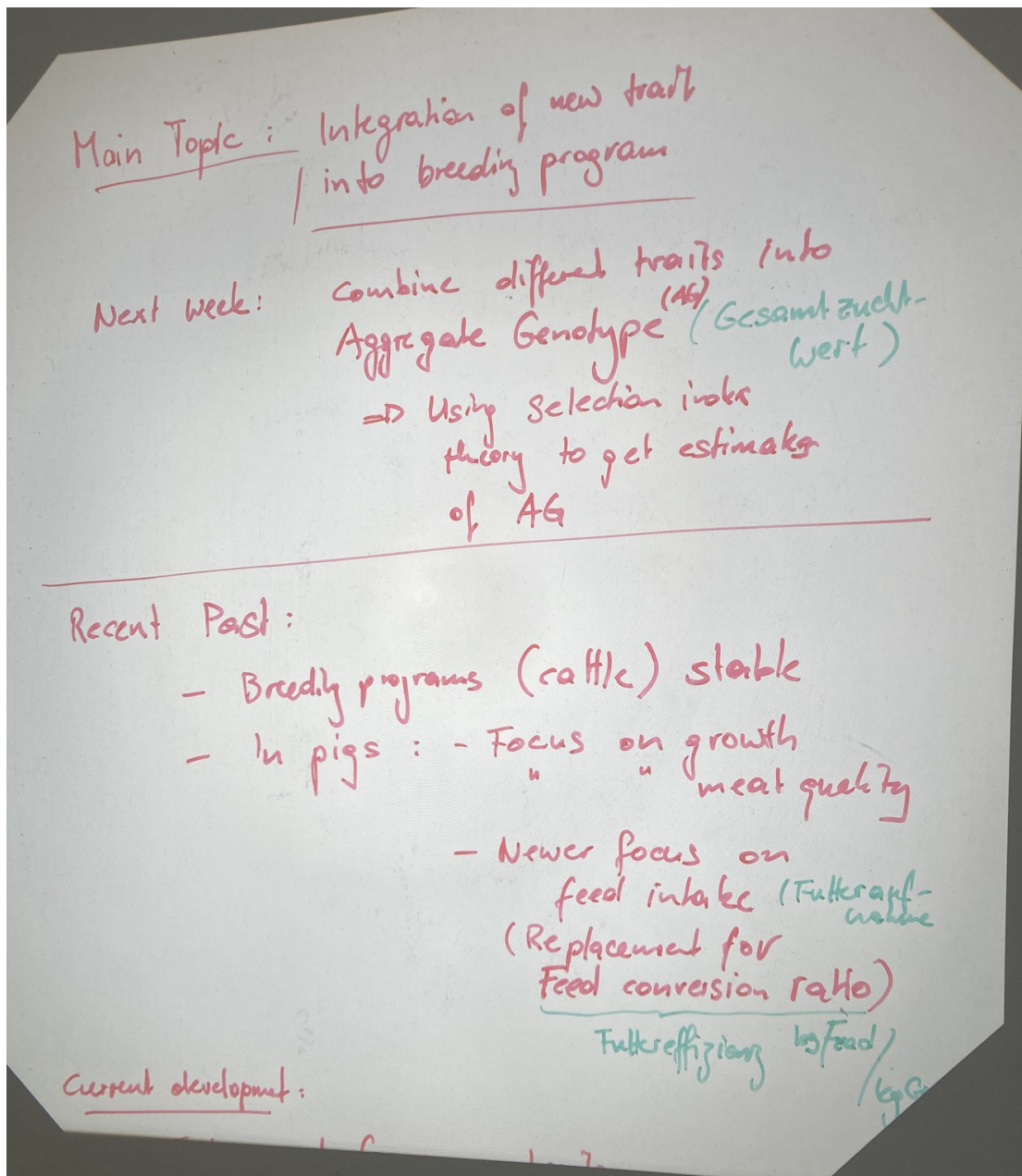
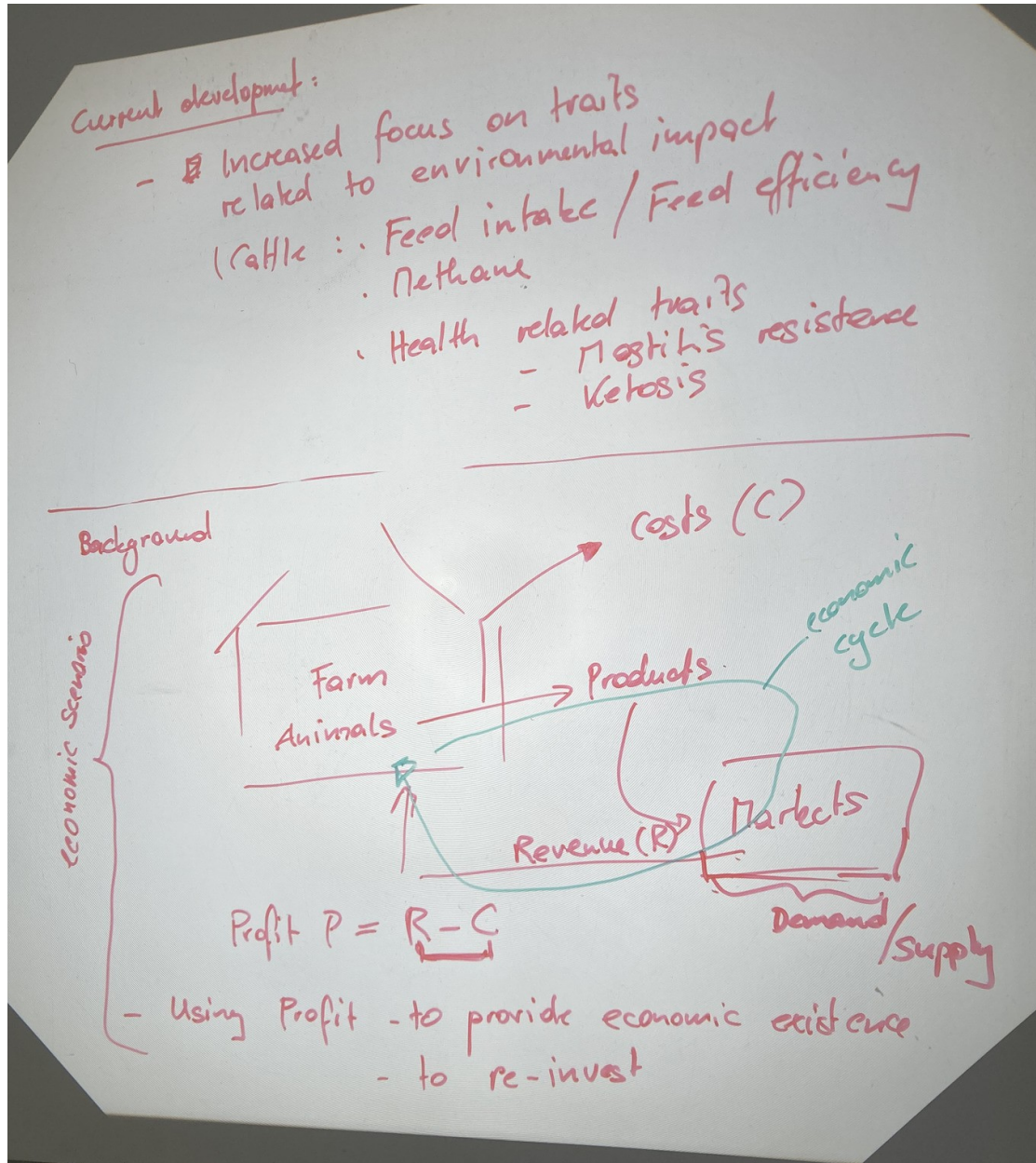


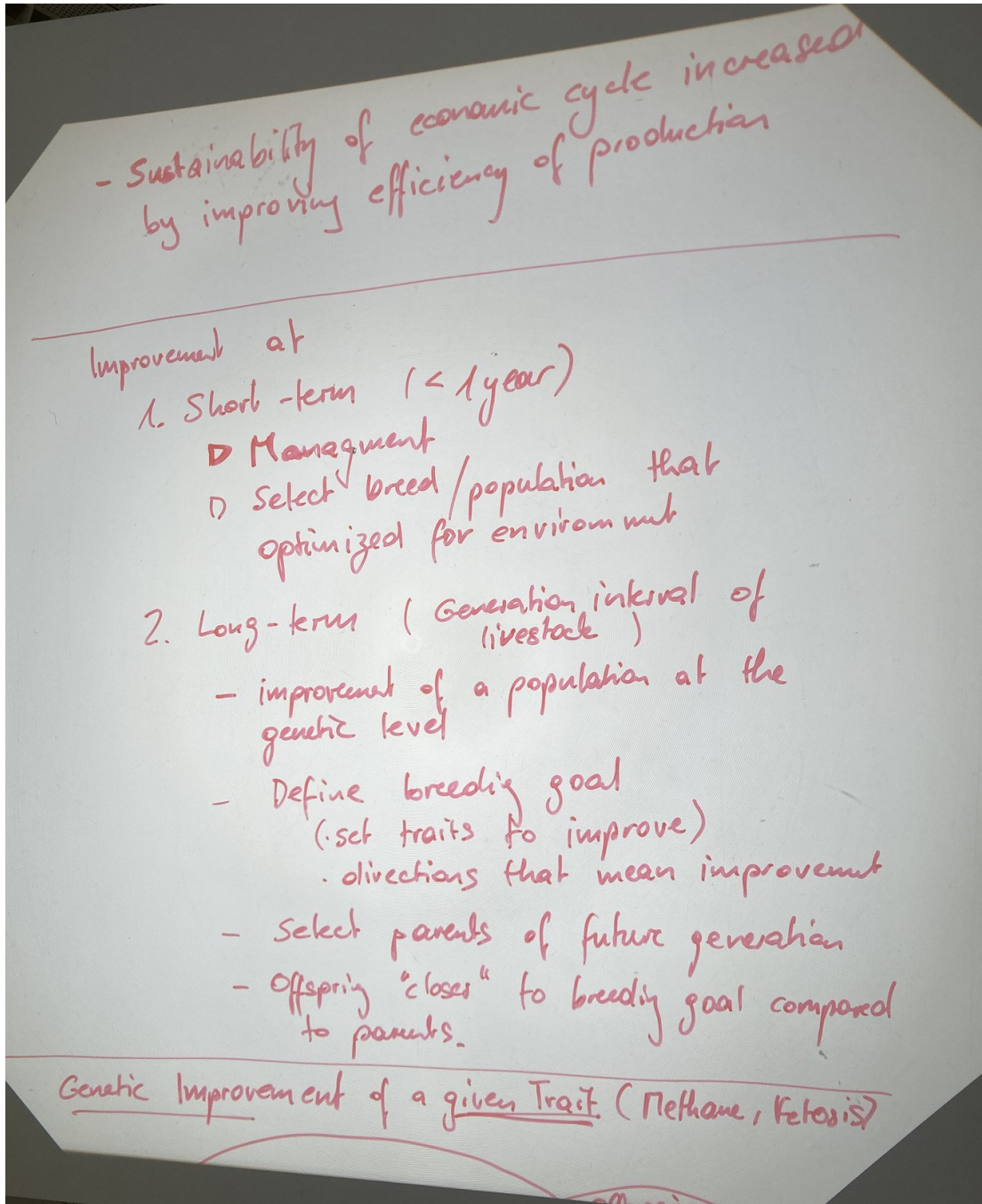
OHP Picture 1



OHP Picture 2



OHP Picture 3



OHP Picture 4

Genetic Improvement of a given Trait (Methane, Fetal)

The diagram illustrates the process of genetic improvement. It starts with a group of 'Parents' with various phenotypes: 100, 150, 250, and 85. A selection process is shown where 50% of the parents are chosen to form an 'Offspring' group. The offspring group consists of individuals with phenotypes 85, 170, and 250. An arrow points from the offspring group towards the right, indicating further genetic improvement.

- How do we find parents with alleles that are favorable with respect to trait of interest
- Most (quantitative) traits: Location of \pm genes that are important for the expression of the trait are unknown
- For every trait, there is a certain genetic component which has an influence on the expression of the phenotypic level of the trait.

Why? \rightarrow Central Dogma of Molecular Biology

Why? → Central Dogma of Molecular Biology

DNA

↓

RNA

↓

Proteins

⇒ Most traits of interests are related to Proteins.

→ Most traits are a composition of different Proteins. The regulation and architecture of all proteins is encoded in the DNA.

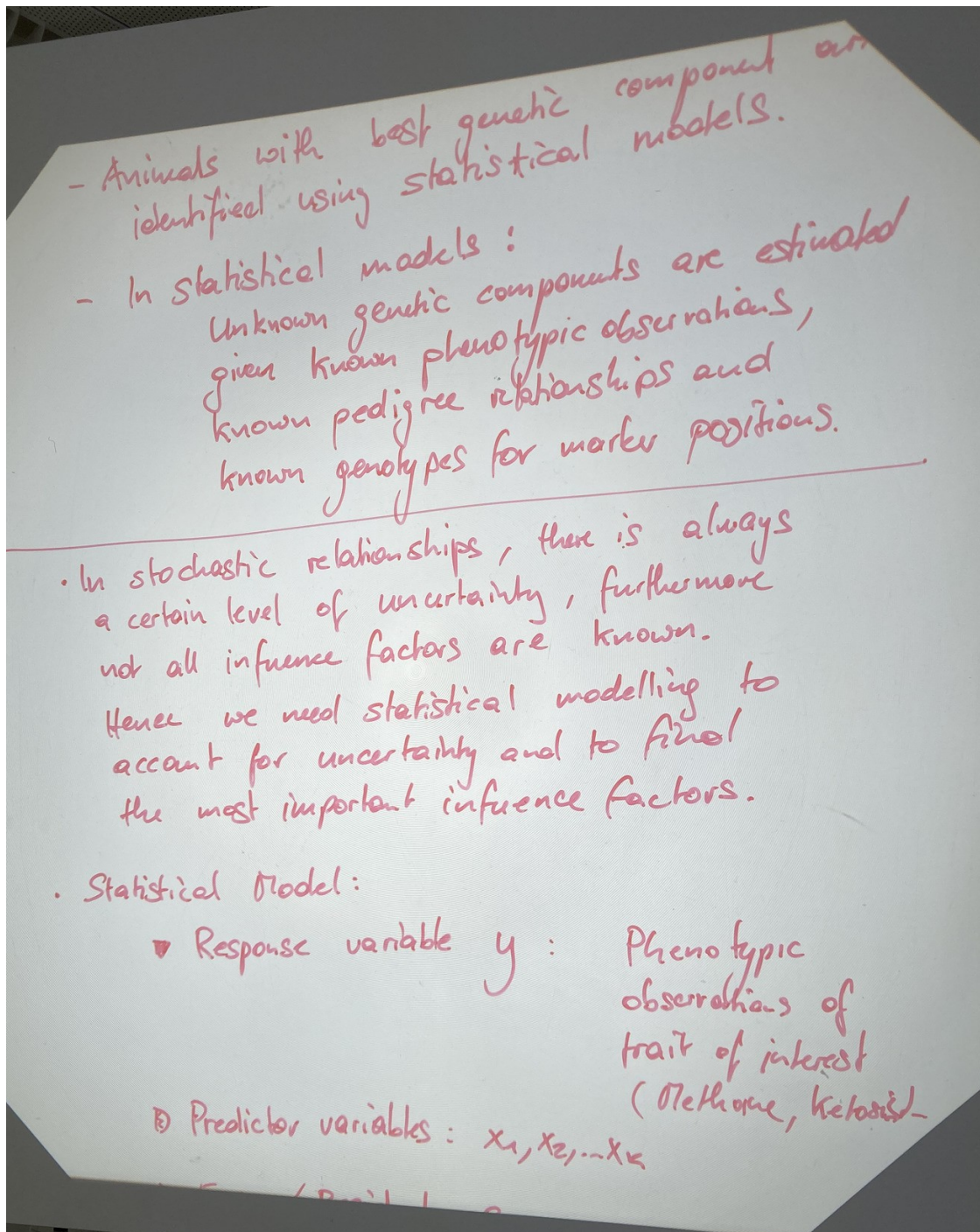
▶ The positions in the genome where protein encoding information is stored are the locations which are relevant for our trait of interest. These locations are called "Quantitative Trait Loci" (QTL)

This can be expressed as the following genetic model

$$P = G + E \rightarrow \text{environment}$$

↓ phenotype ↓ genetic component

OHP Picture 6



• Statistical Model:

- ▷ Response variable y : Phenotypic observations of trait of interest (Methone, ketone)
- ▷ Predictor variables: x_1, x_2, \dots, x_k
- ▷ Error / Residual: e
- ▷ function $m(x_1, x_2, \dots, x_k)$ with $\underline{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \end{bmatrix}$
 $\Leftrightarrow m(\underline{x})$

• Function $m(\underline{x})$ is used to relate the predictor (\underline{x}) to the response y

□ The simplest class of relationships between \underline{x} and y is a linear relationship, for animal i :

$$y_i = m(\underline{x}_i) + e_i$$

□ Example Dataset

Animal	Body Weight	Breast Circumference
1	471	176

OHP Picture 8

□ Example Data set

Animal	Body Weight	Breast Circumference
1	471	176
2	⋮	⋮
⋮	⋮	⋮
10	541	184

Response Predictor

□ Specify $m(x)$ for the example dataset:
For animal i :

$$y_i = b \cdot x_i + e_i$$

Known Unknown Known
 Estimate from data Unknown

□ Data used as input:

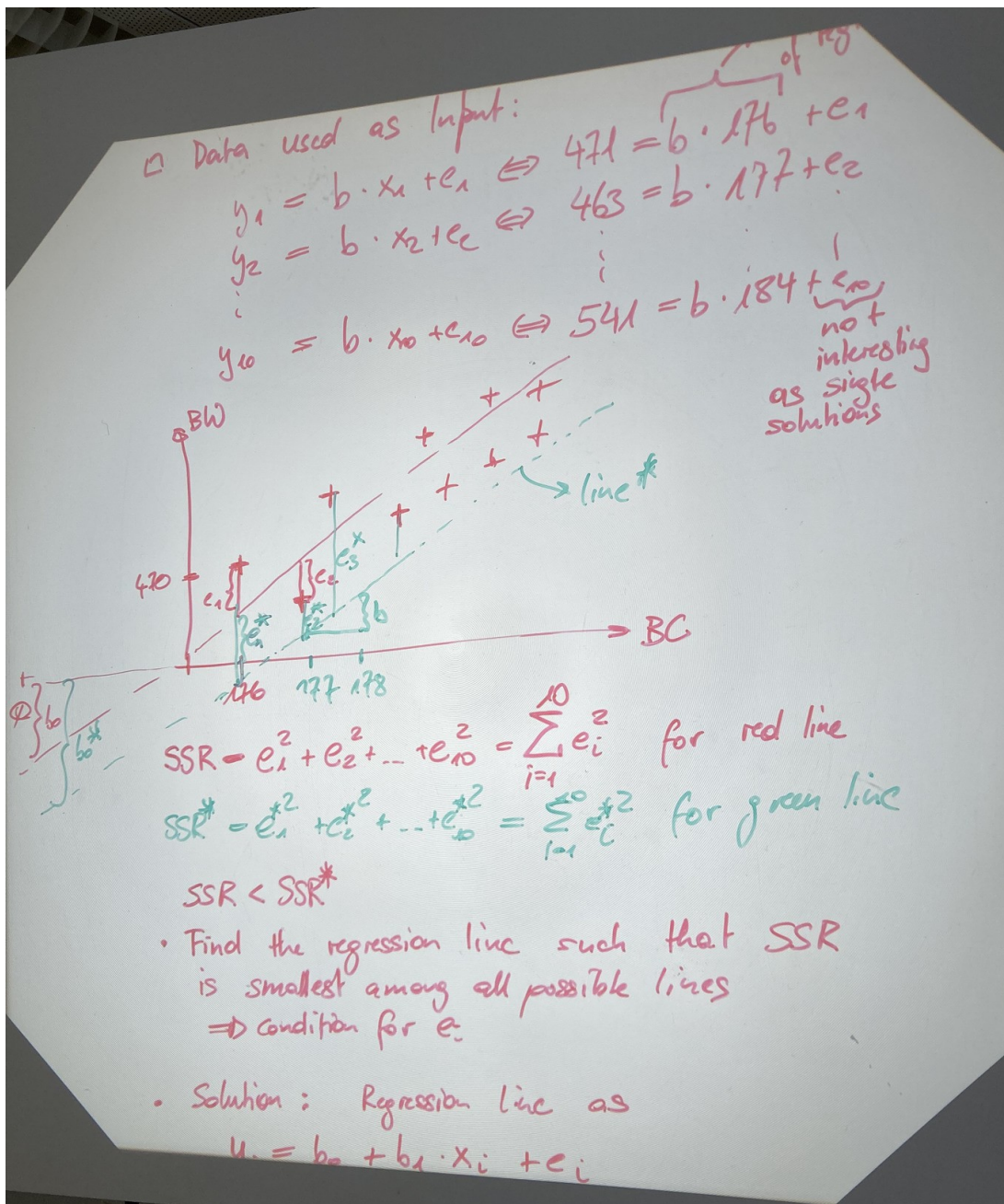
$$y_1 = b \cdot x_1 + e_1 \Leftrightarrow 471 = b \cdot 176 + e_1$$
$$y_2 = b \cdot x_2 + e_2 \Leftrightarrow 463 = b \cdot 177 + e_2$$
$$\vdots$$
$$y_{10} = b \cdot x_{10} + e_{10} \Leftrightarrow 541 = b \cdot 184 + e_{10}$$

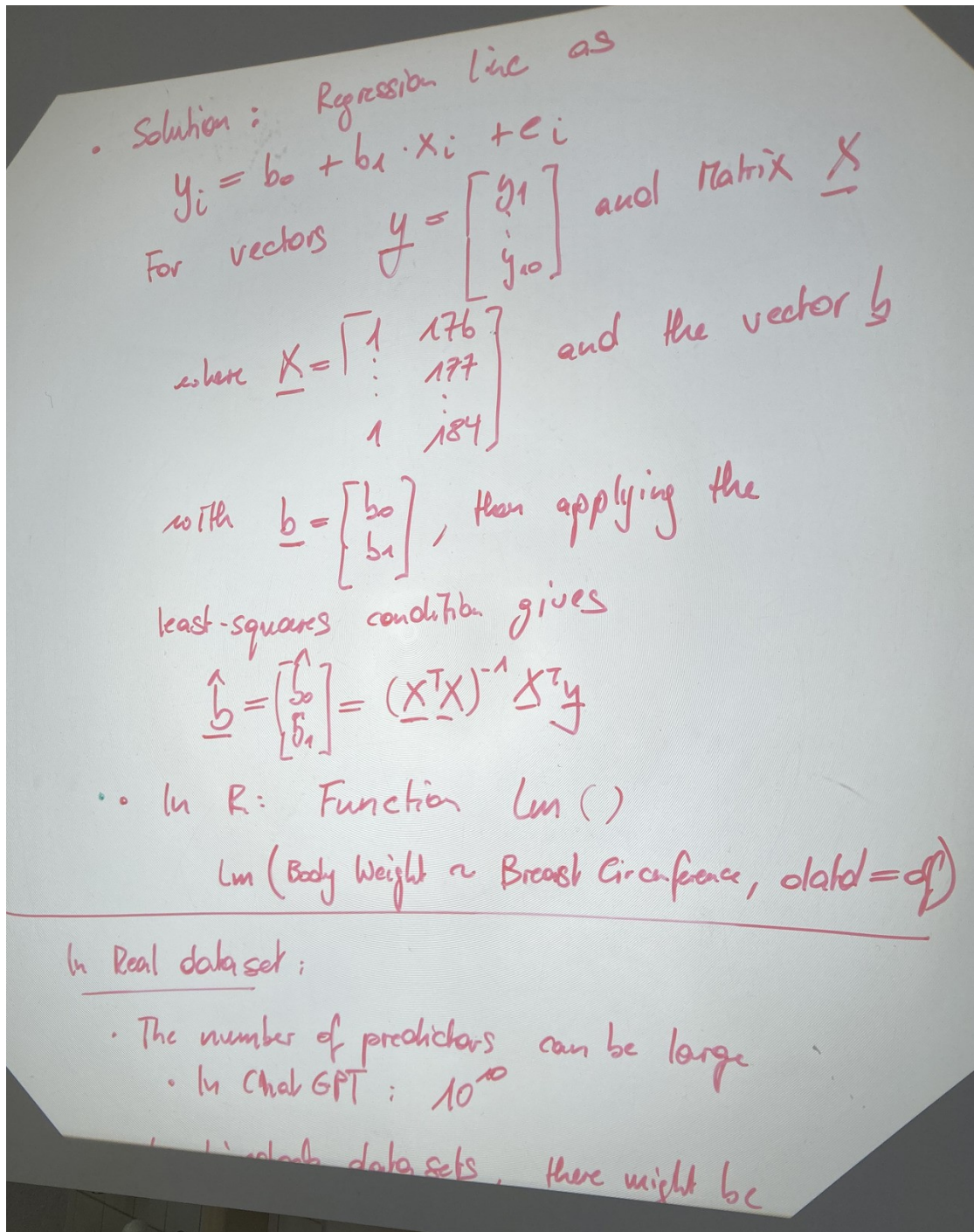
defines the slope of regression line

no + interesting as sign

qBW

OHP Picture 9





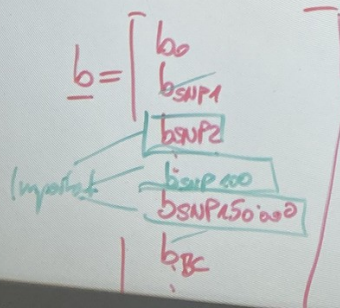
In Real dataset:

- The number of predictors can be large
- In Chab GPT: 10^{10}
- In Livestock datasets, there might be 20 - 100 predictors, for environmental factors
- Genomic datasets: Genotypes at 150K positions.

Animal	Body Weight	SNP ₁	SNP ₂	...	SNP ₁₅₀₀₀₀	...
1						
⋮						
N						

Linear Model:

$$y_i = b_0 + b_{SNP_1} \cdot x_{i,1} + b_{SNP_2} \cdot x_{i,2} + \dots + b_{SNP_{150000}} \cdot x_{i,150000} + \epsilon_i$$



Least Squares Solution can no longer be used because $(X^T X)^{-1}$ cannot be computed.

Linear Model: $y_i = b_0 + b_{SNP1} \cdot X_{i,1} + b_{SNP2} \cdot X_{i,2} + \dots + b_{SNP_{50000}} \cdot X_{i,50000} + \epsilon_i$

$\underline{b} = \begin{bmatrix} b_0 \\ b_{SNP1} \\ b_{SNP2} \\ b_{SNP_{50000}} \\ b_{\epsilon} \end{bmatrix}$

Least Squares Solution can no longer be used because $(X^T X)^{-1}$ cannot be computed.

In R: `lm(---)` \Rightarrow Error

Solution:

1. Separate meaningful predictors from predictors which are not important \Rightarrow Model selection.