Quantitative Genomics and Genetics

BioCB 4830/6830; PBSB.5201.03

Lecture 14: Intro to Genetic Model (Regression) Inference

Jason Mezey

March 12, 2024 (T) 8:40-9:55

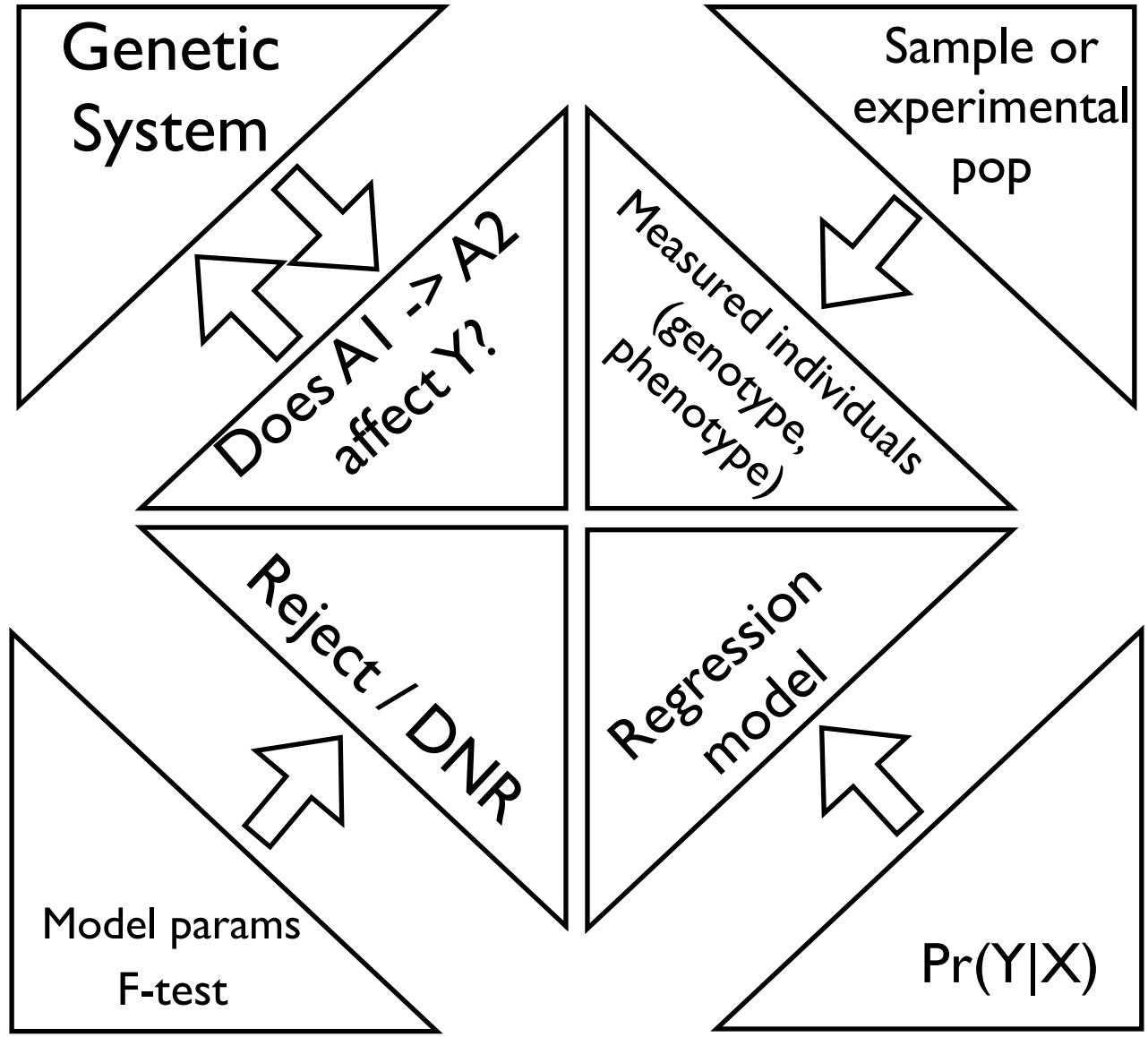
Announcements

- Another typo in homework #3 now corrected (!!) and posted as V3 on canvas - for problem 2a had n=5 but the paragraph before it had n=10 (now both switched to n=5) but PLEASE NOTE:
  - We will give full credit for n=5 OR n=10 in 2a (just as we will give full credit if you used the older / incorrect critical values for 2h and 2j)
  - That is: if you have handed in your homework already no need to change it (!!)
- I have updated the syllabus (!!) where please note
  - We will have one more homework (#4) that will be available next week
  - Your "midterm" will be AFTER Cornell, Ithaca Spring break (available April 9)
- On Thurs (March 14) I will be in lecturing from Ithaca and the following two weeks I will be lecturing by zoom (although lecture rooms will be available) - I will announce more details on Thurs

Summary of lecture 14: Genetic Probability Models

- Last lecture, we began our introduction to Genetic Models (Regressions!)
- Today we will complete our introduction to Regression models (=families of probability models!)
- …and we will begin discussing how to do inference for these models (specifically MLE!)

Conceptual Overview



Review: Genetic system

- **causal mutation** - a position in the genome where an experimental manipulation of the DNA would produce an effect on the phenotype under specifiable conditions
- Formally, we may represent this as follows:

$$A_1 \rightarrow A_2 \Rightarrow \Delta Y | Z$$

- Note: that this definition considers "under specifiable" conditions" so the change in genome need not cause a difference under every manipulation (just under broadly specifiable conditions)
- Also note the symmetry of the relationship
- Identifying these is the core of quantitative genetics/genomics (why do we want to do this!?)
- What is the perfect experiment?
- Our experiment will be a statistical experiment (sample and inference!)

The statistical model I

- As with any statistical experiment, we need to begin by defining our sample space
- In the most general sense, our sample space is:

$$\Omega = \{ \text{ Possible Individuals } \}$$

- More specifically, each individual in our sample space can be quantified as a pair of sample outcomes so our sample space can be written as:

$$\Omega = \{\Omega_g \cap \Omega_P\}$$

- Where Ω_g is the genotype sample space at a locus and Ω_P is the phenotype sample space
- Note that genotype $g_i = A_j A_k$ is the set of possible genotypes, where for a diploid, with two alleles:

$$\Omega_g = \{A_1A_1, A_1A_2, A_2A_2\}$$

- For the phenotype, this can be any type of measurement (e.g. sick or healthy, height, etc.)

The statistical model II

- Next, we need to define the probability model on the sigma algebra of the sample space ($\mathcal{F}_{\{g,P\}}$):

$$Pr(\mathcal{F}_{\{g,P\}})$$

- Which defines the probability of each possible genotype and phenotype pair:

$$Pr\{g, P\}$$

- We will define two (types) or random variables (* = state does not matter):

$$Y : (*, \Omega_P) \rightarrow \mathbb{R}$$

$$X : (\Omega_g, *) \rightarrow \mathbb{R}$$

- Note that the probability model induces a (joint) probability distribution on this random vector (these random variables):

$$Pr(Y, X)$$

Review: The statistical model III

- The goal of quantitative genomics and genetics is to identify cases of the following relationship:

$$Pr(Y \cap X) = Pr(Y, X) \neq Pr(Y)Pr(X)$$

- Remember that, regardless of the probability distribution of our random vector, we can define the expectation:

$$\mathrm{E}\left[Y, X\right] = \left[\mathrm{E}Y, \mathrm{E}X\right]$$

- and the variance:

$$Var\left[Y, X\right] = \begin{bmatrix} Var(Y) & Cov(Y, X) \\ Cov(Y, X) & Var(X) \end{bmatrix}$$

- The goal of quantitative genomics can be rephrased as assessing the following relationship:

$$Cov(Y, X) \neq 0$$

Review: The statistical model IV

- We are going to consider a parameterized model to represent the probability model of *X* and *Y* (that is the true statistical model of genetics!!!)
- Specifically, we will consider a *regression model*
- For the moment, let's consider a regression model with normal error:

$$Y = \beta_0 + X\beta_1 + \epsilon$$

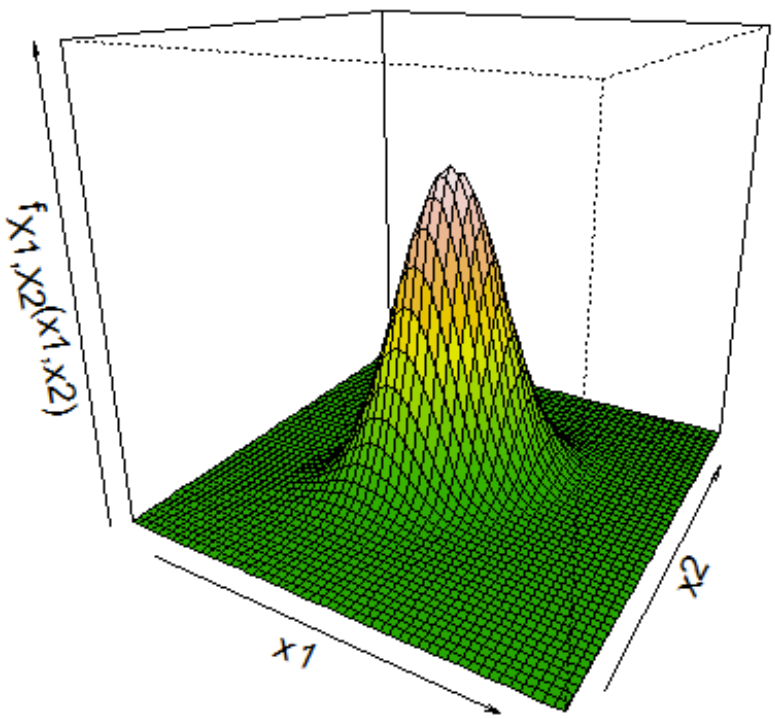
$$\epsilon \sim N(0, \sigma_\epsilon^2)$$

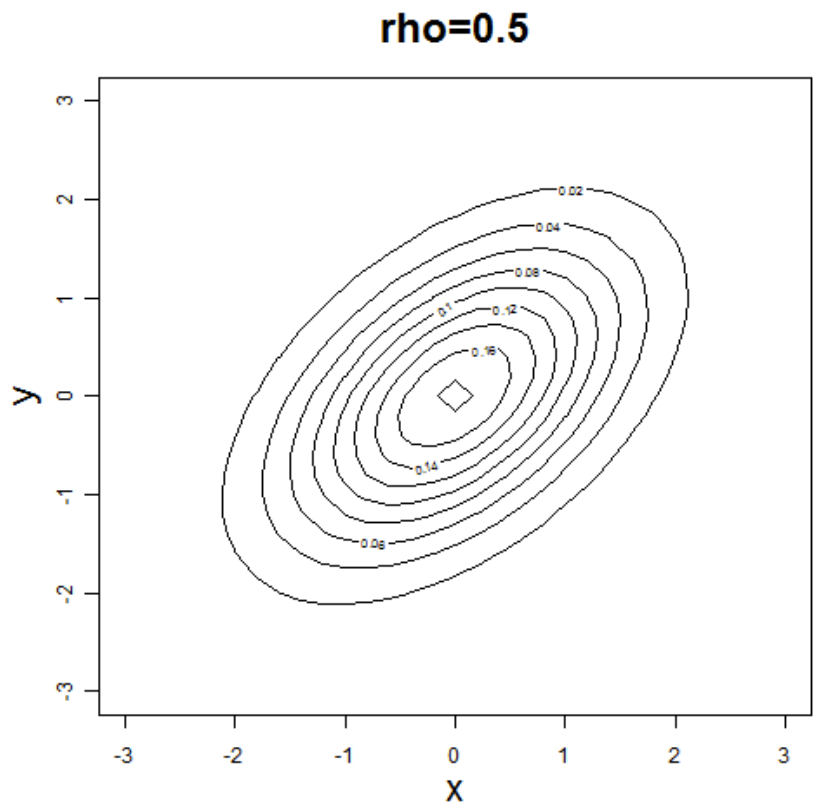
- Note that in this model, we consider Y to be the *dependent* or *response* variable and X to be the *independent* variable (what are the parameters!?)
- Also note implicitly assumes the following:

$$Pr(Y, X) = Pr(Y|X)$$

Review: Linear regression is a bivariate distribution

- We've seen bivariate (multivariate) distributions before:

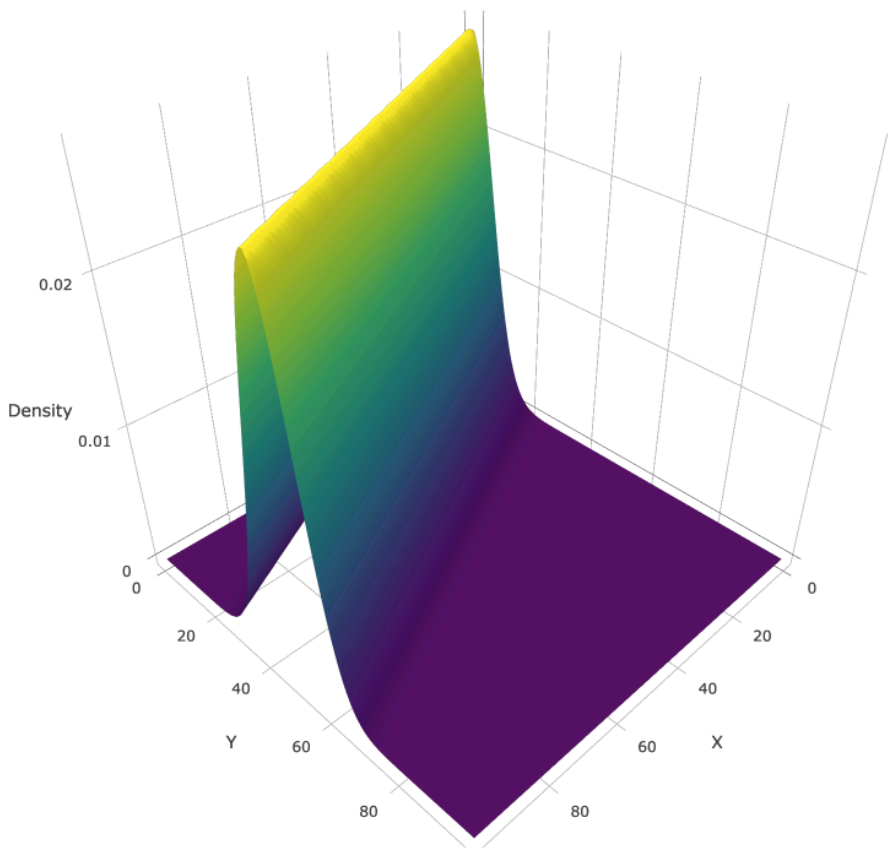


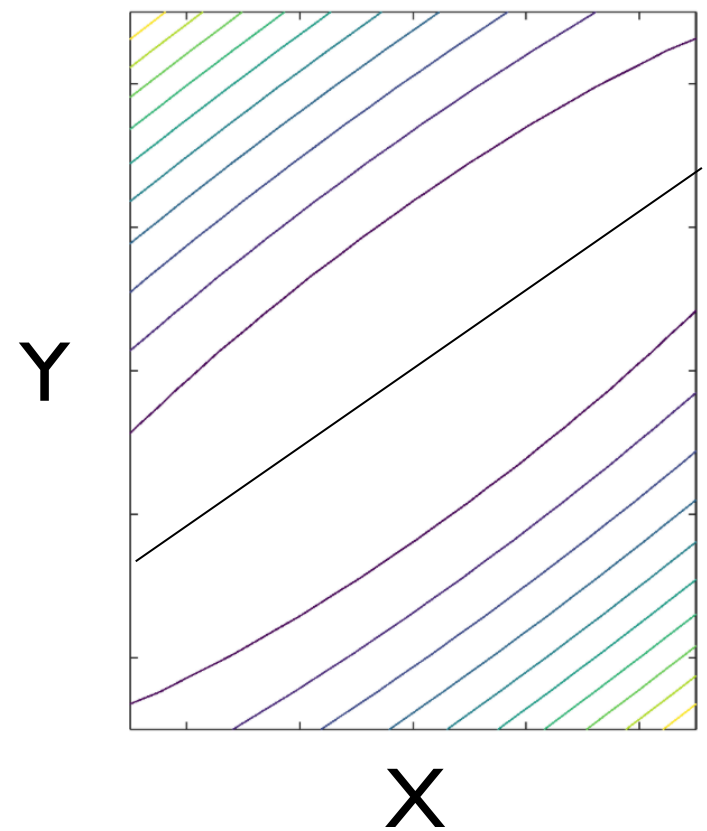


Review: Linear regression I

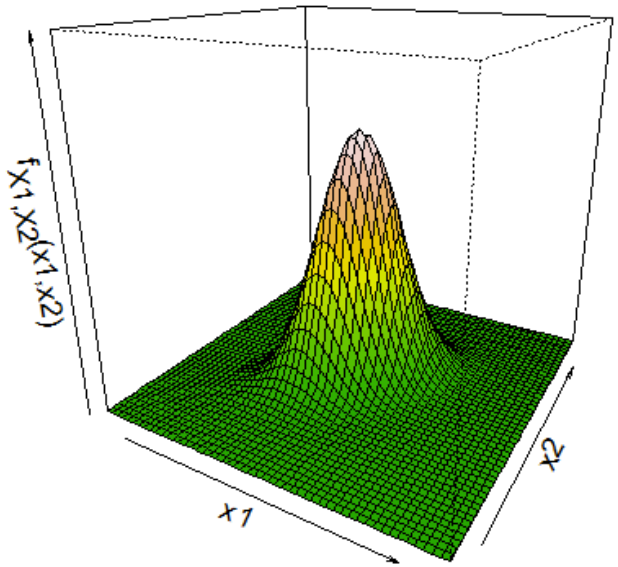
- Let's review the structure of a linear regression (not necessarily a genetic model):

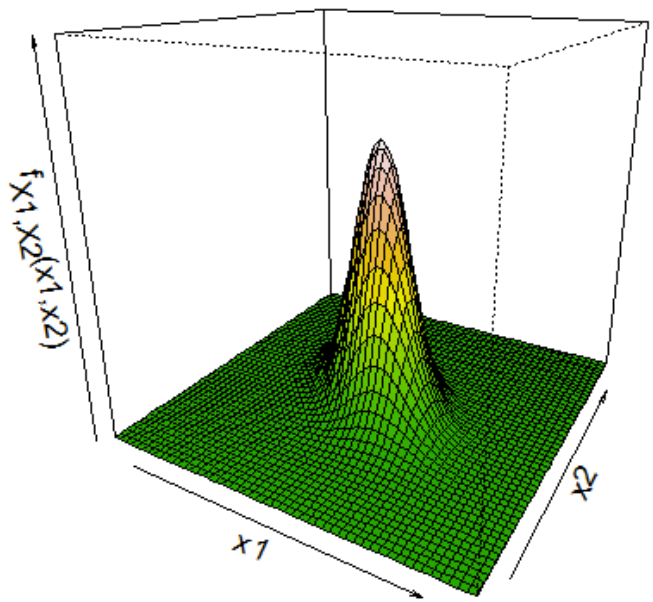
$$Y = \beta_0 + X\beta_1 + \epsilon \qquad \epsilon \sim N(0, \sigma_\epsilon^2)$$

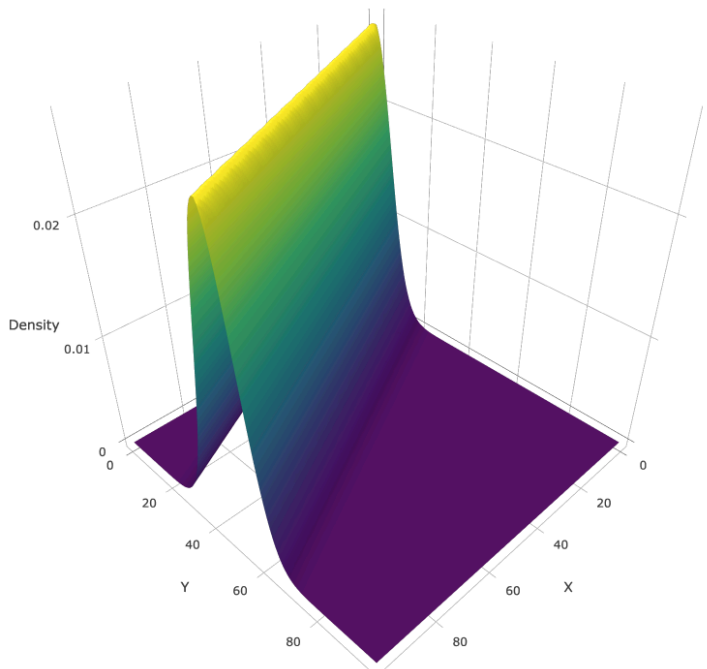


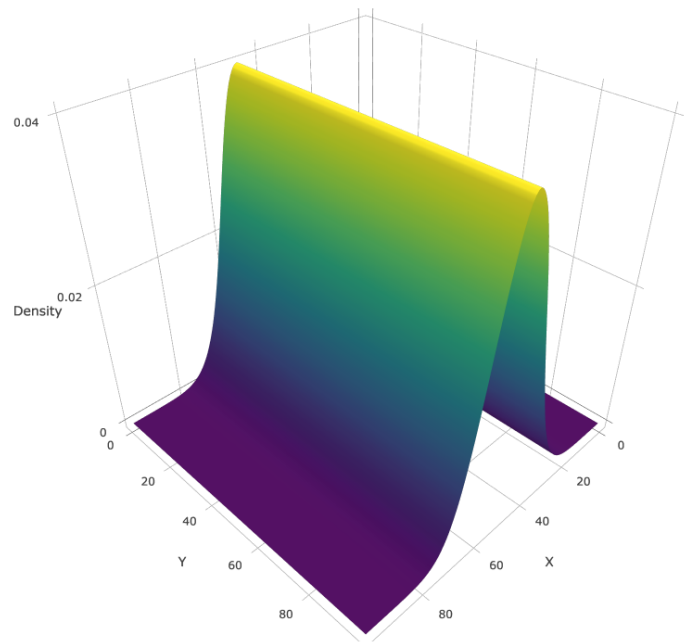


Review: Linear regression II









The genetic probability model I

- Remember that we define the random variables we need for our genetic model by

$$Y : (*, \Omega_P) \to \mathbb{R}$$

$$X : (\Omega_g, *) \to \mathbb{R}$$

- Where we have three possible genotypes:

$$\Omega_g = \{A_1A_1, A_1A_2, A_2A_2\}$$

- The quantitative genetic model is a "multiple" regression model with the following TWO independent ("dummy") *X* variables:

$$X_a(A_1A_1) = -1, X_a(A_1A_2) = 0, X_a(A_2A_2) = 1$$

$$X_d(A_1A_1) = -1, X_d(A_1A_2) = 1, X_d(A_2A_2) = -1$$

1		A_1A_2	
-1	A_1A_1		A_2A_2
	-1	0	1

- and the following "multiple" regression equation:

$$Y = \beta_\mu + X_a\beta_a + X_d\beta_d + \epsilon$$

$$\epsilon \sim N(0, \sigma^2_\epsilon)$$

The genetic probability model II

- The probability distribution of this model, is therefore:

$$Pr(Y|X) \sim N(\beta_\mu + X_a\beta_a + X_d\beta_d, \sigma^2_\epsilon)$$

- Which has four parameters:

$$\beta_\mu, \beta_a, \beta_d, \sigma^2_\epsilon$$

- The three β parameters are required to model the three separate genotypes (A1A1, A1A2, A2A2)

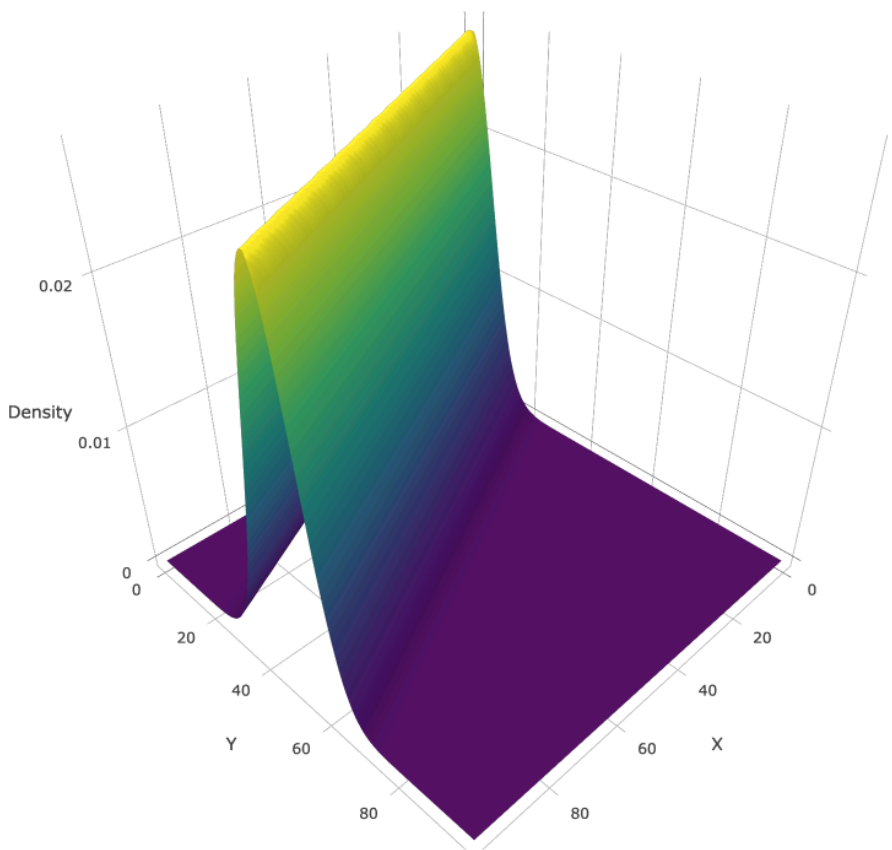
- The ϵ can be thought of as a random variable that describes the probability an individual will have a specific value of *Y*, conditional on the genotype AiAj, where the probability is normally distributed around the value determined by the *X*'s and β 's

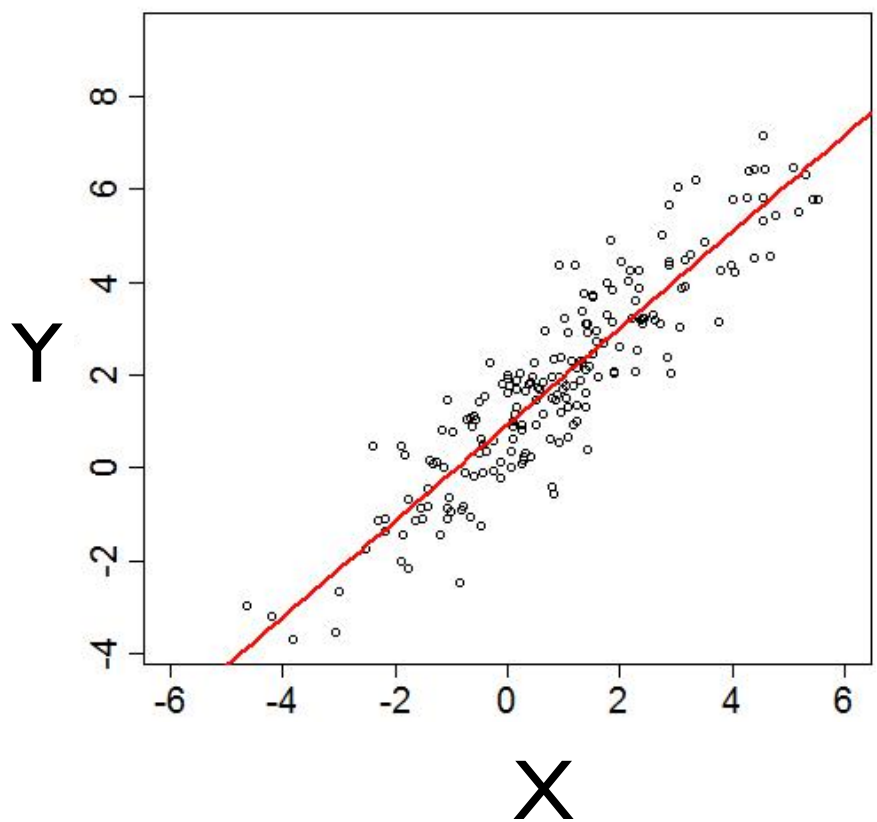
$$\epsilon \sim N(0, \sigma^2_\epsilon)$$

Linear regression III

- The linear regression model allows calculation of the (interval) probability of observations (!!)

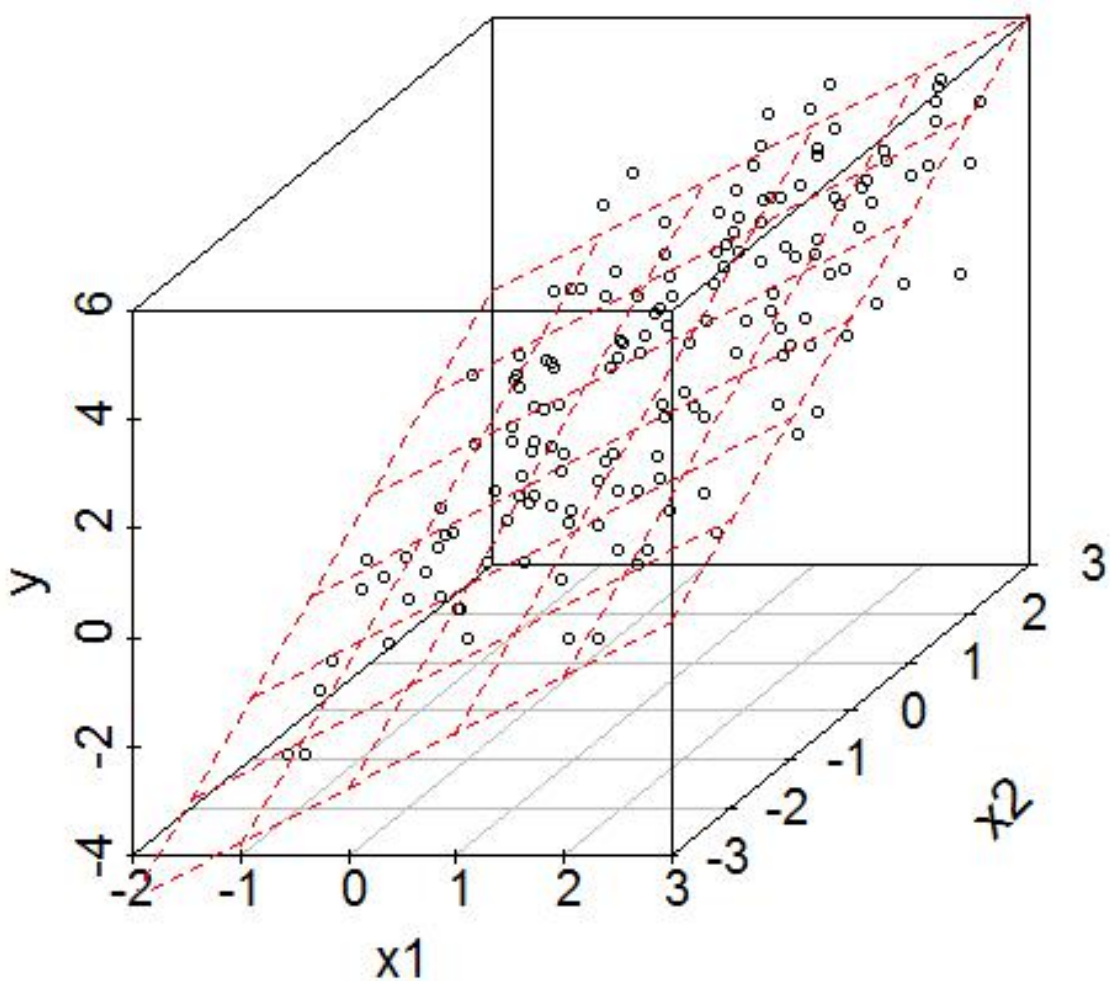
$$Y = \beta_0 + X\beta_1 + \epsilon \qquad \epsilon \sim N(0, \sigma_\epsilon^2)$$





Linear regression IV

- A *multiple regression* model has the same structure, with a single dependent variable *Y* and more than one independent variable X_i, X_j, e.g.,

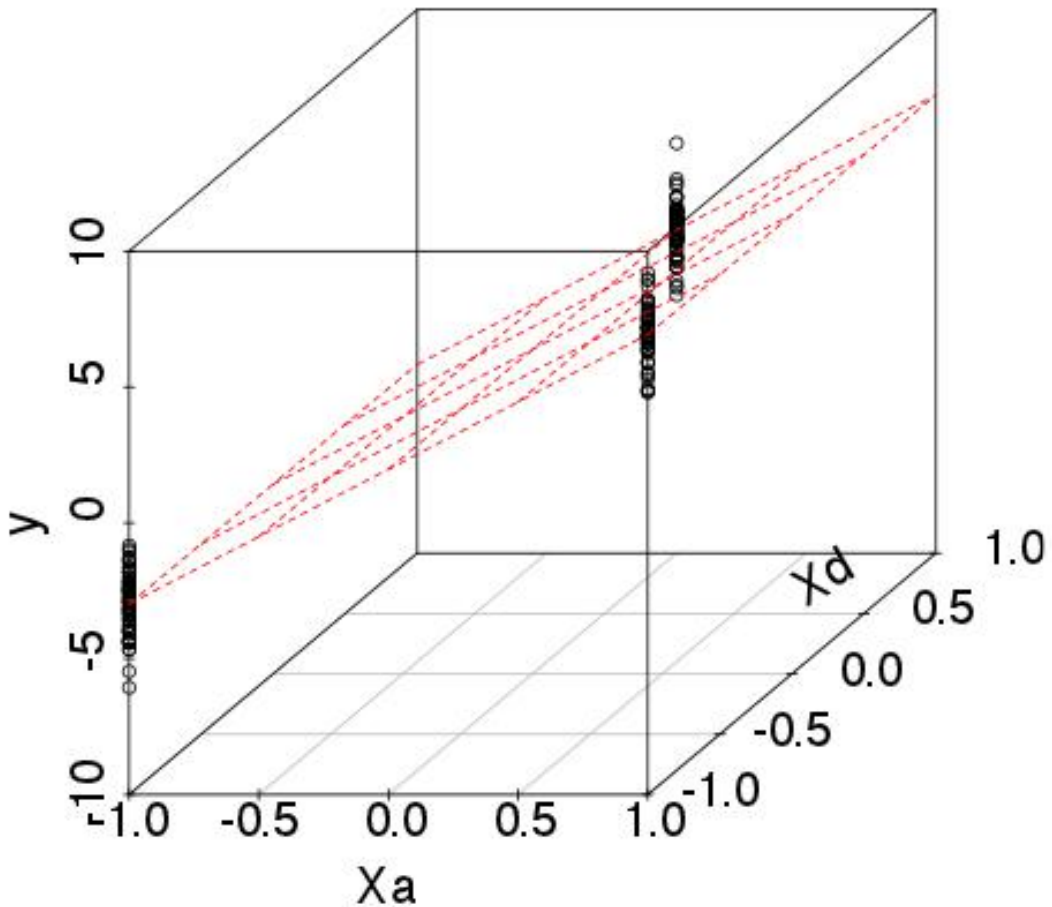


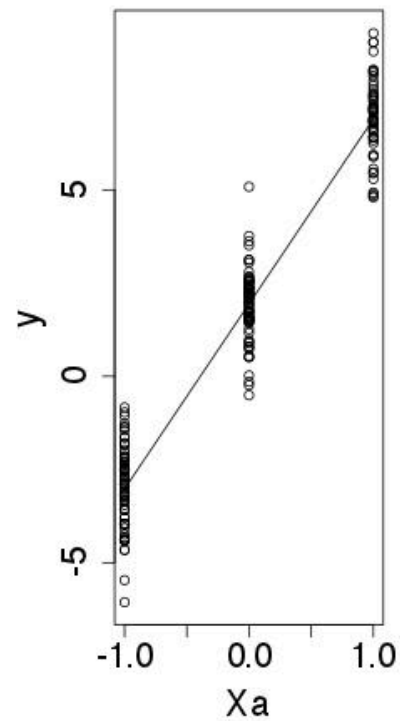
The genetic probability model III

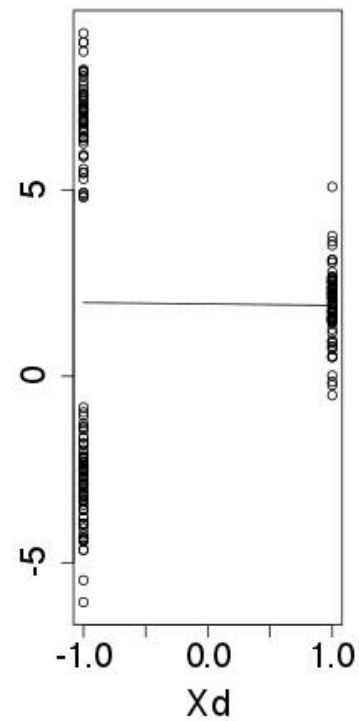
- Note that, while somewhat arbitrary, the advantage of the *X*a and *X*d coding is the parameters β_a and β_d map directly on to relationships between the genotype and phenotype that are important in genetics:
  - If $\beta_a \neq 0, \beta_d = 0$ then this is a "purely" additive case
  - If $\beta_a = 0, \beta_d \neq 0$ then this is only over- or under-dominance (homozygotes have equal effects on phenotype)
  - If both are non-zero, there are both additive and dominance effects
  - If both are zero, there is no effect of the genotype on the phenotype (the genotype is not causal!)

Genetic example 1

- As an example, consider the following of a "purely additive" case (= no dominance): $\beta_\mu = 2, \beta_a = 5, \beta_d = 0, \sigma^2_\epsilon = 1$



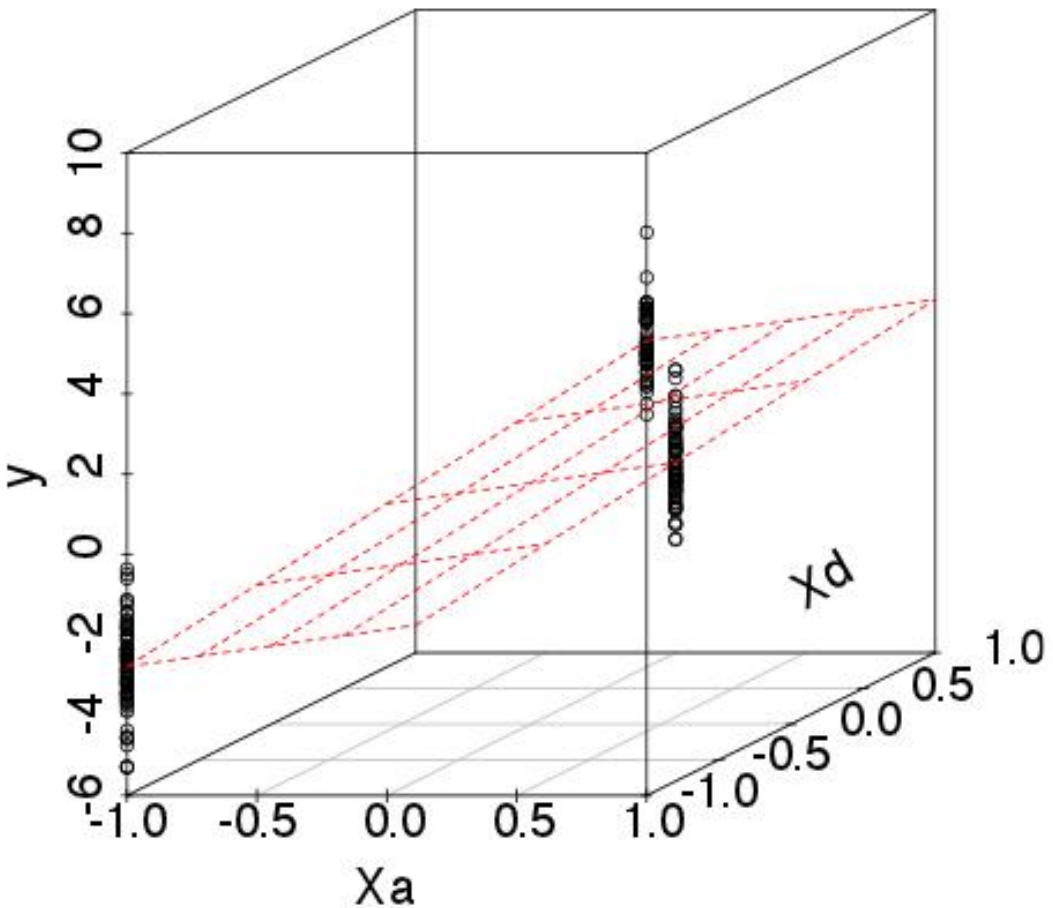


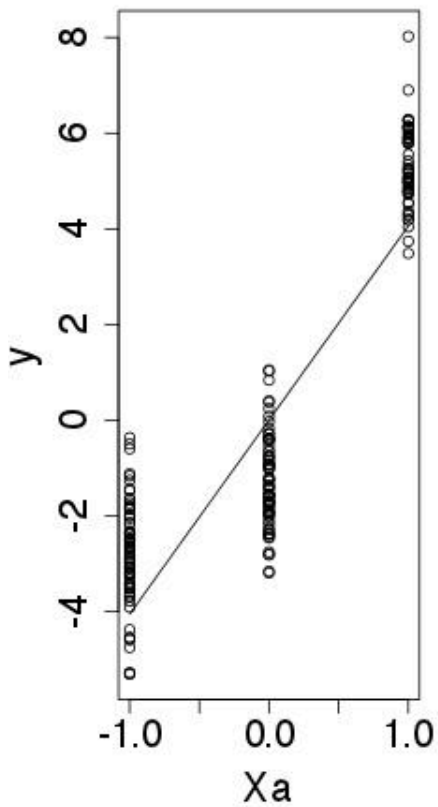


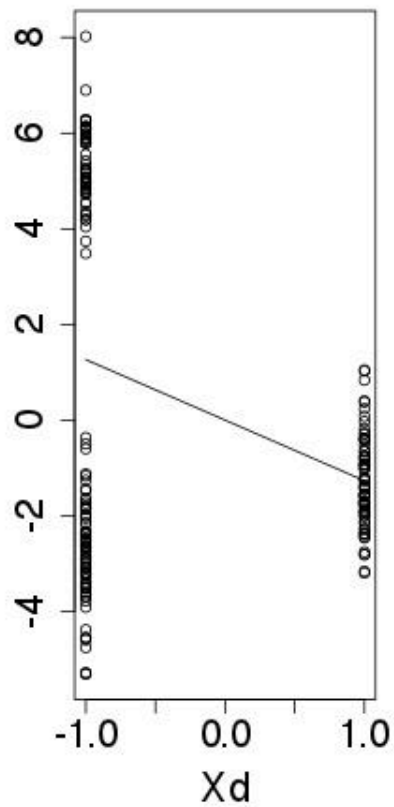
Genetic example II

- An example of "dominance" (= not a "pure additive" case):

$$\beta_\mu = 0, \beta_a = 4, \beta_d = -1, \sigma^2_\epsilon = 1$$



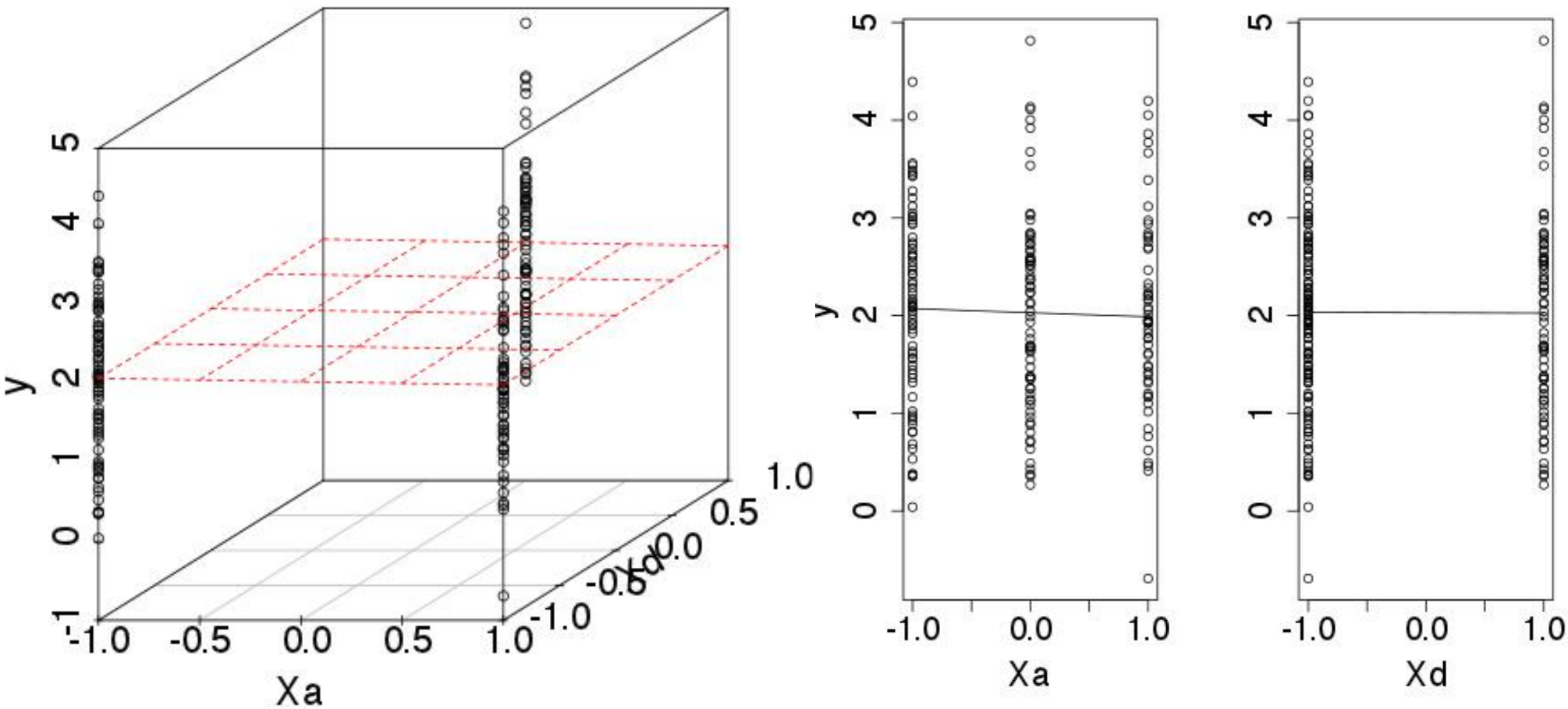




Review: Genetic example III

- A case of NO genetic effect:

$$\beta_\mu = 2, \beta_a = 0, \beta_d = 0, \sigma^2_\epsilon = 1$$



Quantitative genetic formalism

- For those of you who have been exposed to classic quantitative genetics, you have seen a different notation for this model:

$$P = G + E$$

  - *P* is the **phenotypic value** - the value of the aspect measured
  - *G* is the **genotypic value** - the expected value of the phenotype conditional on the genotype
  - *E* is the **environmental value** - the value of the phenotype that we cannot explain given the genotype

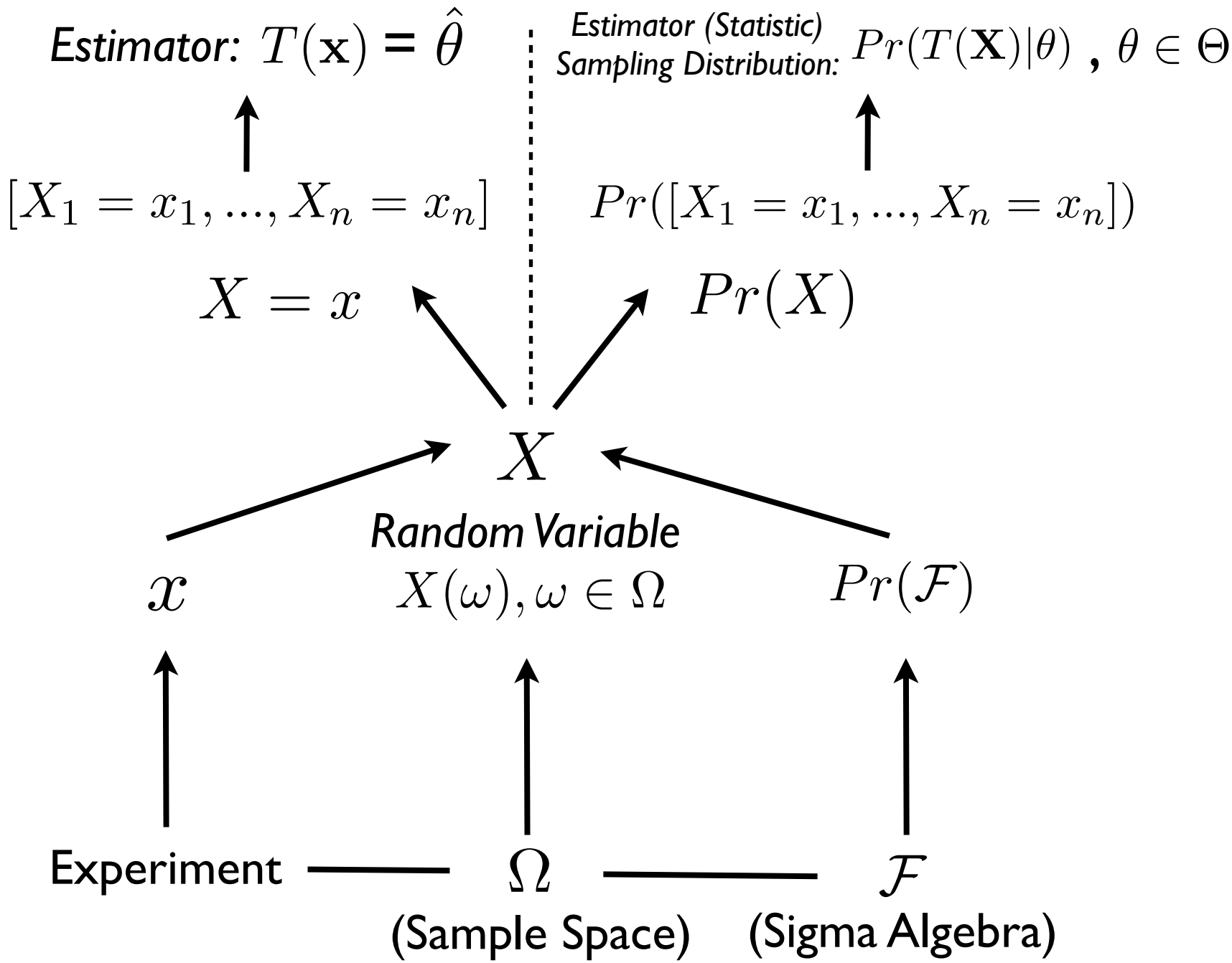
- These translate as follows for our one locus case (although note the formalism extends to any multiple locus case):

$$Y = P$$

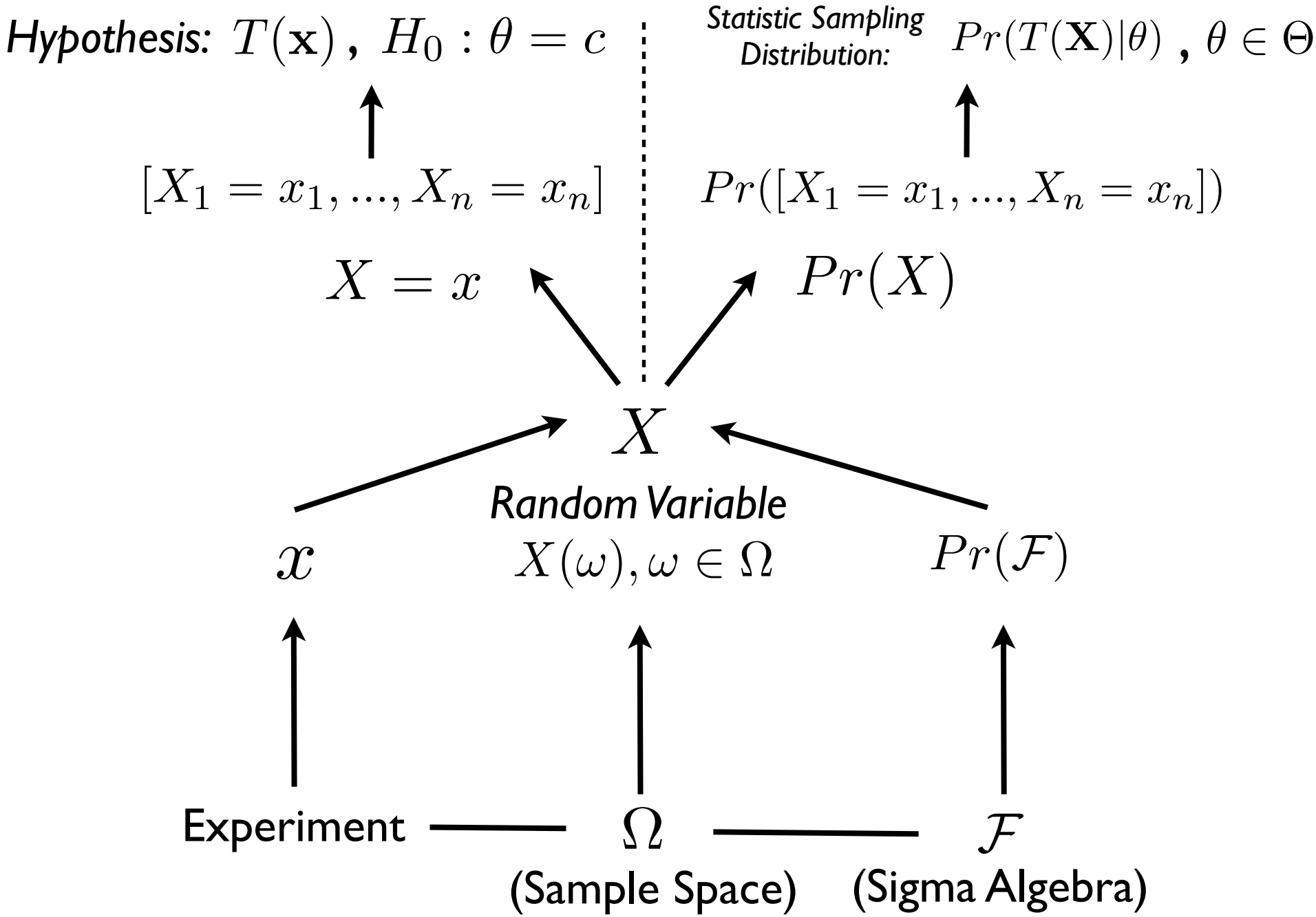
$$G = \mathrm{E}P = \mathrm{E}Y = \beta_\mu + X_a\beta_a + X_d\beta_d$$

$$\epsilon = E$$

Estimators


Estimator: $T(\mathbf{x}) = \hat{\theta}$
Estimator (Statistic)
Sampling Distribution: $Pr(T(\mathbf{X})|\theta)$, $\theta \in \Theta$
$[X_1 = x_1, ..., X_n = x_n]$
$Pr([X_1 = x_1, ..., X_n = x_n])$
$X = x$
$Pr(X)$
X
Random Variable
$X(\omega), \omega \in \Omega$
x
$Pr(\mathcal{F})$
Experiment
Ω
$\mathcal{F}$
(Sample Space)
(Sigma Algebra)

Hypothesis Tests



Genetic inference I

- For our model focusing on one locus:

$$Y = \beta_\mu + X_a\beta_a + X_d\beta_d + \epsilon$$

$$\epsilon \sim N(0, \sigma^2_\epsilon)$$

- We have four possible parameters we could estimate:

$$\theta = \left[\beta_\mu, \beta_a, \beta_d, \sigma^2_\epsilon\right]$$

- However, for our purposes, we are only interested in the genetic parameters and testing the following null hypothesis:

$$H_0 : Cov(X_a, Y) = 0 \cap Cov(X_d, Y) = 0$$
$$H_A : Cov(X_a, Y) \neq 0 \cup Cov(X_d, Y) \neq 0$$

OR

$$H_0 : \beta_a = 0 \cap \beta_d = 0$$
$$H_A : \beta_a \neq 0 \cup \beta_d \neq 0$$

Genetic inference II

- Recall that inference (whether estimation or hypothesis testing) starts by collecting a sample and defining a statistic on that sample
- In this case, we are going to collect a sample of *n* individuals where for each we will measure their *phenotype* and their *genotype* (i.e. at the locus we are focusing on)
- That is an individual *i* will have phenotype y_i and genotype $g_i = A_j A_k$ (where we translate these into x_a and x_d)
- Using the phenotype and genotype we will construct both an *estimator* (a statistic!) and we will additionally construct a *test statistic*
- Remember that our regression probability model defines a sampling distribution on our sample and therefore on our estimator and test statistic (!!)

Matrix Basics

$$\mathbf{v} = \vec{v} = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \qquad \mathbf{M}_1 = \bar{M}_1 = \begin{bmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{bmatrix} \qquad \mathbf{M}_2 = \bar{M}_2 = \begin{bmatrix} a & d \\ b & e \\ c & f \end{bmatrix}$$

We will also follow statistics convention where the first subscript will index rows and the second will index columns (note this is usually reversed in mathematics literature).

Matrix sum: $\mathbf{M}_1 + \mathbf{M}_1 = \begin{bmatrix} m_{11} + m_{11} & m_{12} + m_{12} \\ m_{21} + m_{21} & m_{22} + m_{22} \end{bmatrix}$

Matrix transpose: $\mathbf{M}_2^{\mathrm{T}} = \begin{bmatrix} a & b & c \\ d & e & f \end{bmatrix}$

Scalar times a matrix: $c\mathbf{M}_1 = \begin{bmatrix} cm_{11} & cm_{12} \\ cm_{21} & cm_{22} \end{bmatrix}$

Matrix multiplication:

$$\mathbf{M}_1\mathbf{M}_1 = \begin{bmatrix} m_{11}m_{11} + m_{12}m_{21} & m_{11}m_{12} + m_{21}m_{22} \\ m_{21}m_{11} + m_{22}m_{21} & m_{21}m_{12} + m_{22}m_{22} \end{bmatrix} \quad \mathbf{M}_2\mathbf{M}_1 = \begin{bmatrix} am_{11} + dm_{21} & am_{12} + dm_{22} \\ bm_{11} + em_{21} & bm_{12} + em_{22} \\ cm_{11} + fm_{21} & cm_{12} + fm_{22} \end{bmatrix}$$

$$\mathbf{v}\mathbf{v}^{\mathrm{T}} = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}\begin{bmatrix} v_1 & v_2 \end{bmatrix} = \begin{bmatrix} v_1v_1 & v_1v_2 \\ v_2v_1 & v_2v_2 \end{bmatrix}, \quad \mathbf{v}^{\mathrm{T}}\mathbf{v} = \begin{bmatrix} v_1 & v_2 \end{bmatrix}\begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = v_1v_1 + v_2v_2$$

If the following holds: $\mathbf{v}_1^{\mathrm{T}}\mathbf{v}_2 = \begin{bmatrix} v_1 & v_2 \end{bmatrix}\begin{bmatrix} v_3 \\ v_4 \end{bmatrix} = 0$ then $\mathbf{v}_1$ and $\mathbf{v}_2$ are orthogonal.

The identity matrix is defined as follows: $\mathbf{I} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, i.e. diagonal elements are “1” and all other elements are “0”.

The inverse of a matrix $\mathbf{M}^{-1}$ has a structure such that is satisfies the following relationship (for a “square”, *k* x *k* matrix): $\mathbf{M}\mathbf{M}^{-1} = \mathbf{I}$ and $\mathbf{M}^{-1}\mathbf{M} = \mathbf{I}$.

That's it for today

- Next lecture, we will continue our discussion of inference for Genetic Models (!!)