Quantitative Genomics and Genetics

BTRY 4830/6830; PBSB.5201.03

Lecture 13: Introduction to Quantitative Genetic Inference

Jason Mezey

March 9, 2023 (Th) 8:05-9:20

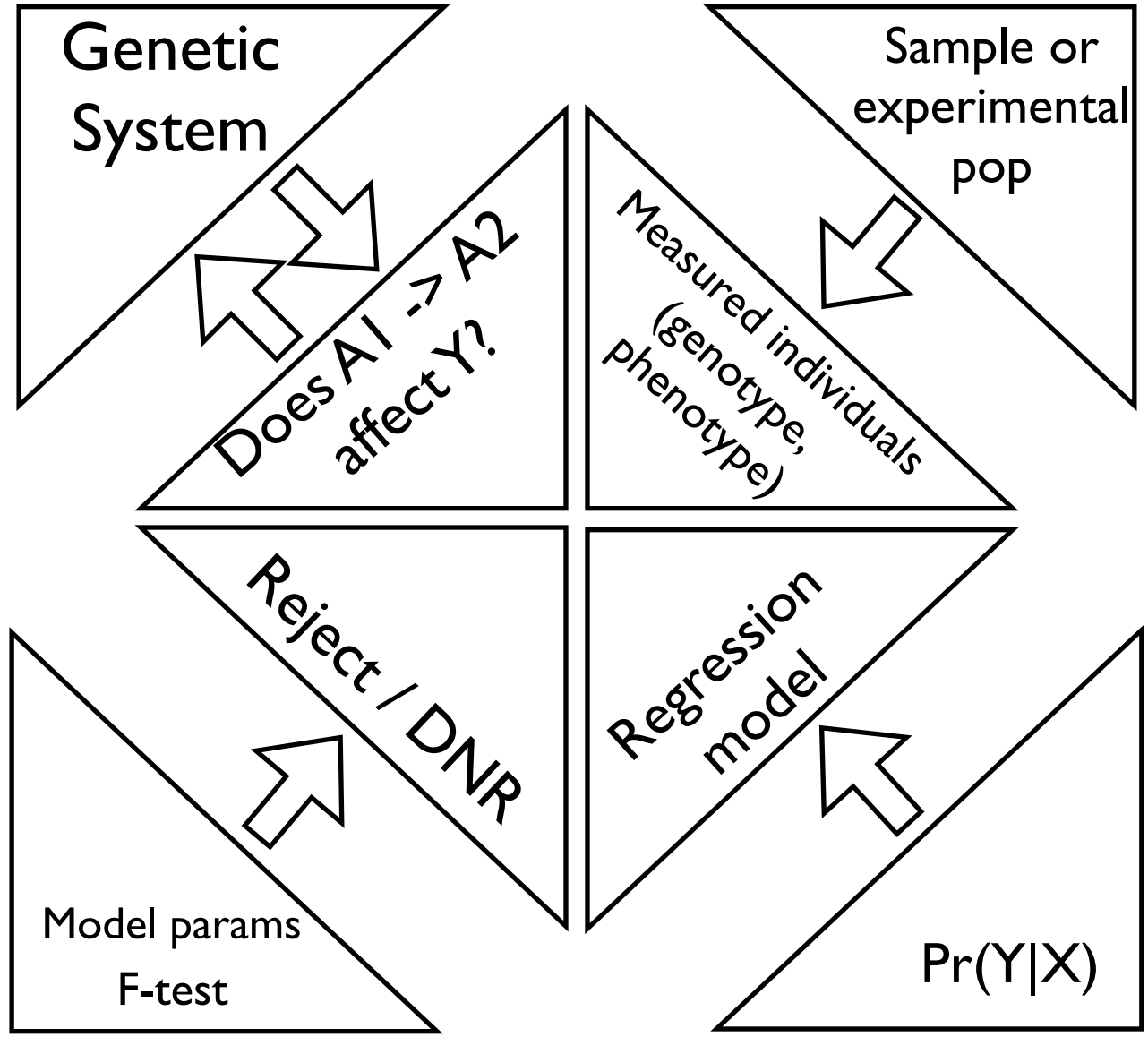
Announcements

- Homework #4 will be assigned soon…
- FOR THOSE IN NYC… Lab TODAY (Thurs., March 9) will NOT take place 4-5PM (usual time) - we are looking into finding a room 5:30-6:30 OR we will schedule a zoom lab WE WILL LET YOU KNOW BY PIAZZA MESSAGE THIS AFTERNOON

Summary of lecture 13: Introduction to Genetic Inference

- Last lecture, we began discussion of genetic modeling
- Today, we will discuss genetic inference!

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!)

Review: 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.)

Review: 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:

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

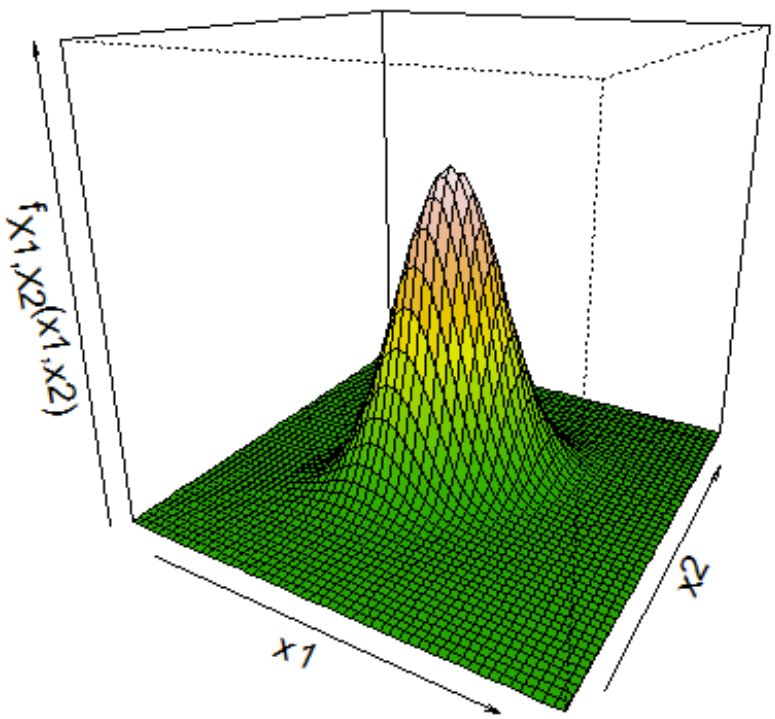
$$\epsilon \sim N(0, \sigma_\epsilon^2)$$

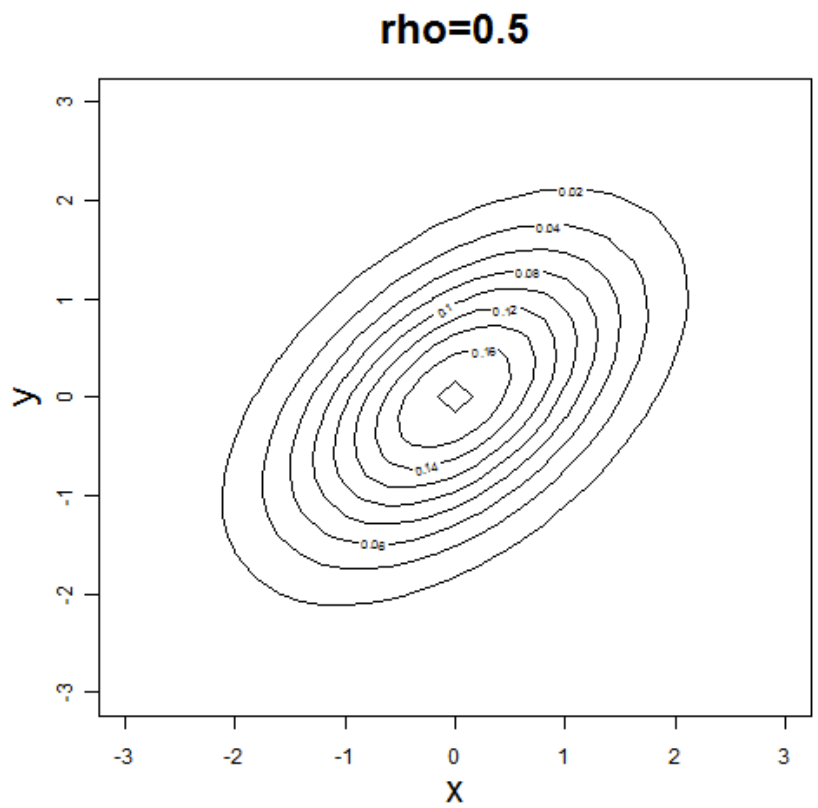
- Note that in this model, we consider Y to be the *dependent* or *response* variable and X to be the *independent* variable (what are the parameters!?)
- Also note implicitly assumes the following:

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

Review: Linear regression is a bivariate distribution

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

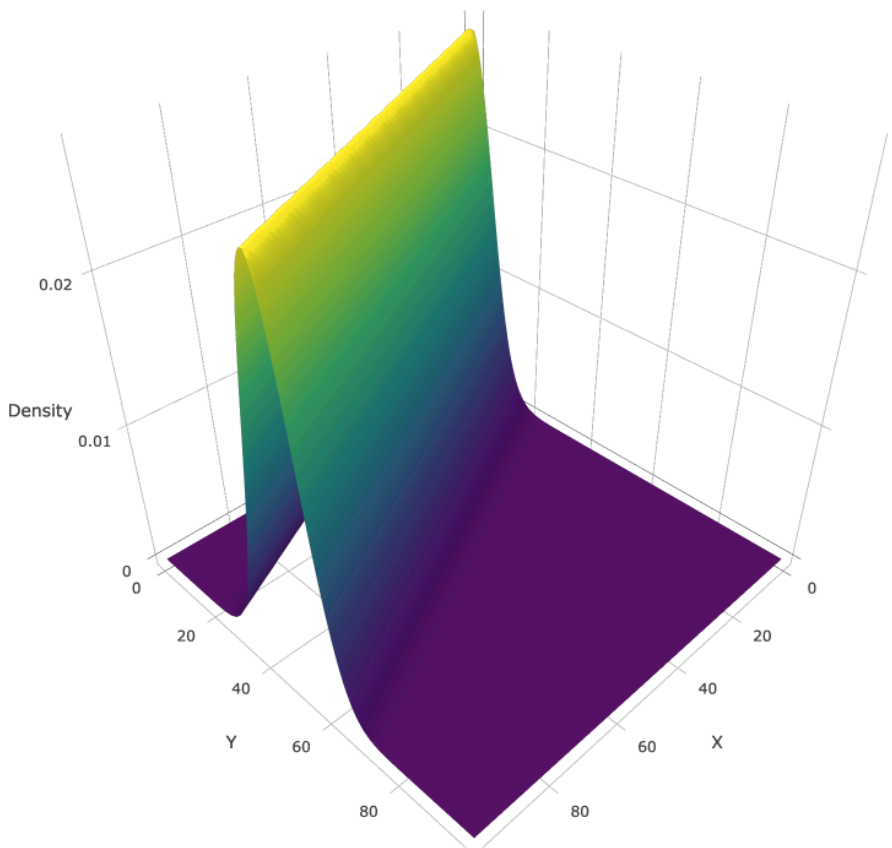


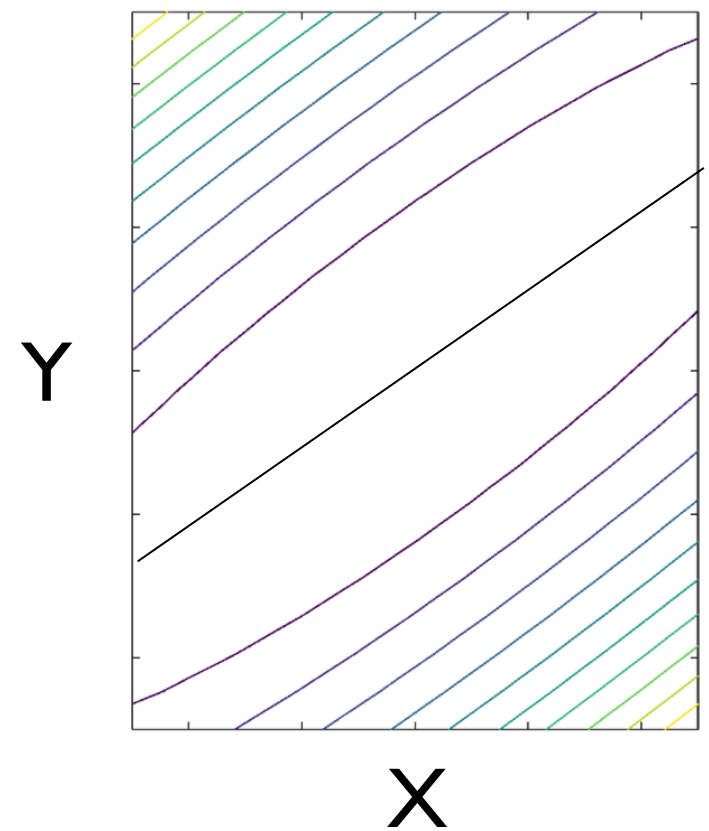


Review: Linear regression I

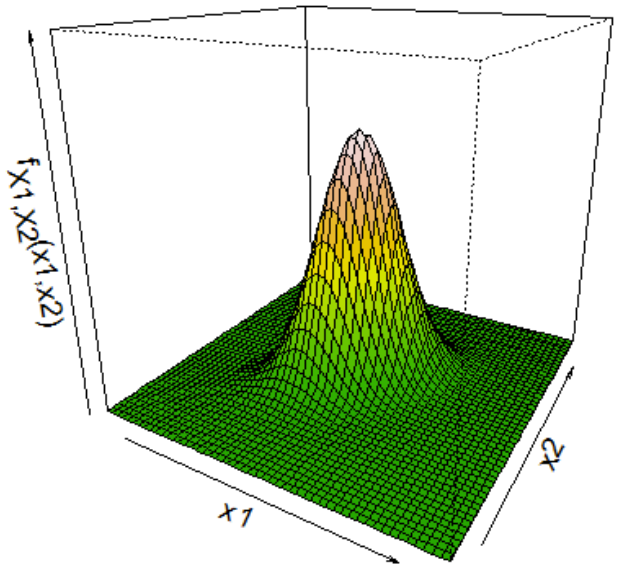
- Let's review the structure of a linear regression (not necessarily a genetic model):

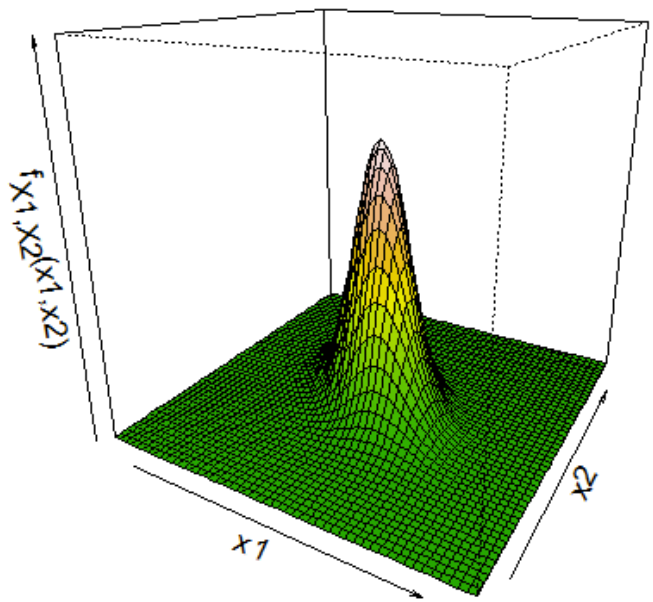
$$Y = \beta_0 + X\beta_1 + \epsilon \qquad \epsilon \sim N(0, \sigma_\epsilon^2)$$

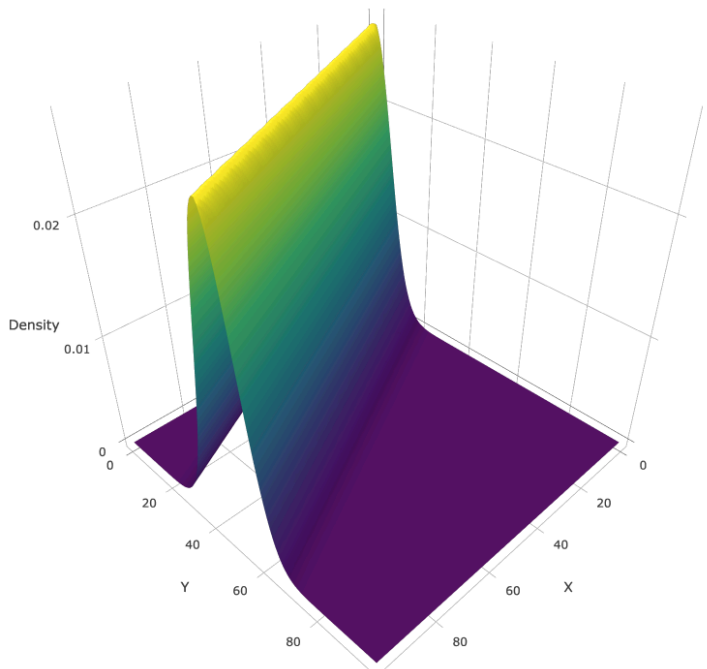


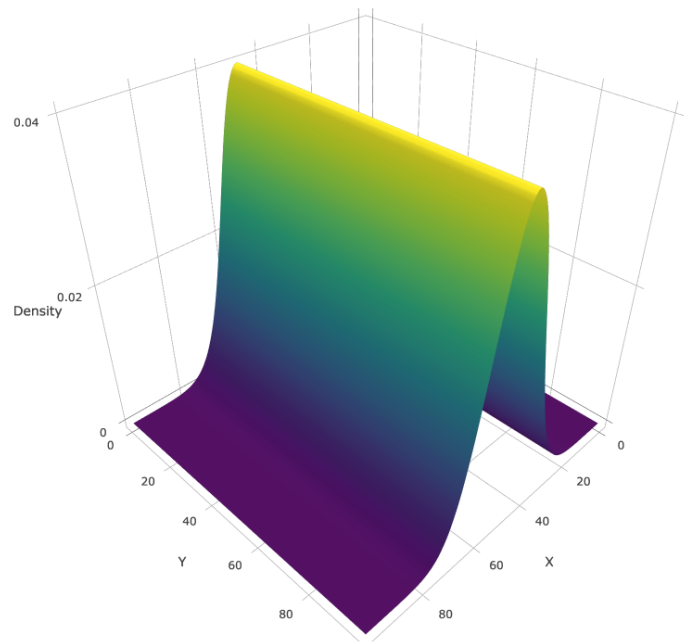


Review: Linear regression II









The genetic probability model I

- The quantitative genetic model is a multiple regression model with the following independent ("dummy") 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

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

The genetic probability model III

- Let's consider a specific example where we are interested modeling the relationship between a genotype and a phenotype (such as height) where the latter is well approximated by a normal distribution

- For this case, the (unknown) conditions of the experiment define the true values of the parameters (unknown to us!), which we will say are the following (note these are the same for all individuals in the population since they are parameters of the probability distribution):

$$\beta_\mu = 0.3, \beta_a = -0.2, \beta_d = 1.1, \sigma_\epsilon^2 = 1.1$$

- Consider an individual *i* with g_i = A1A2 such that we have:

$$X_a(A1A2) = 0, X_d(A1A2) = 1$$

- If this individual has a phenotype value y_i = 2.1 then we have the epsilon value $\epsilon_i = 0.7$ where the probability of this particular value (i.e. the interval surrounding this value) is defined by $\epsilon \sim N(0, \sigma_\epsilon^2)$

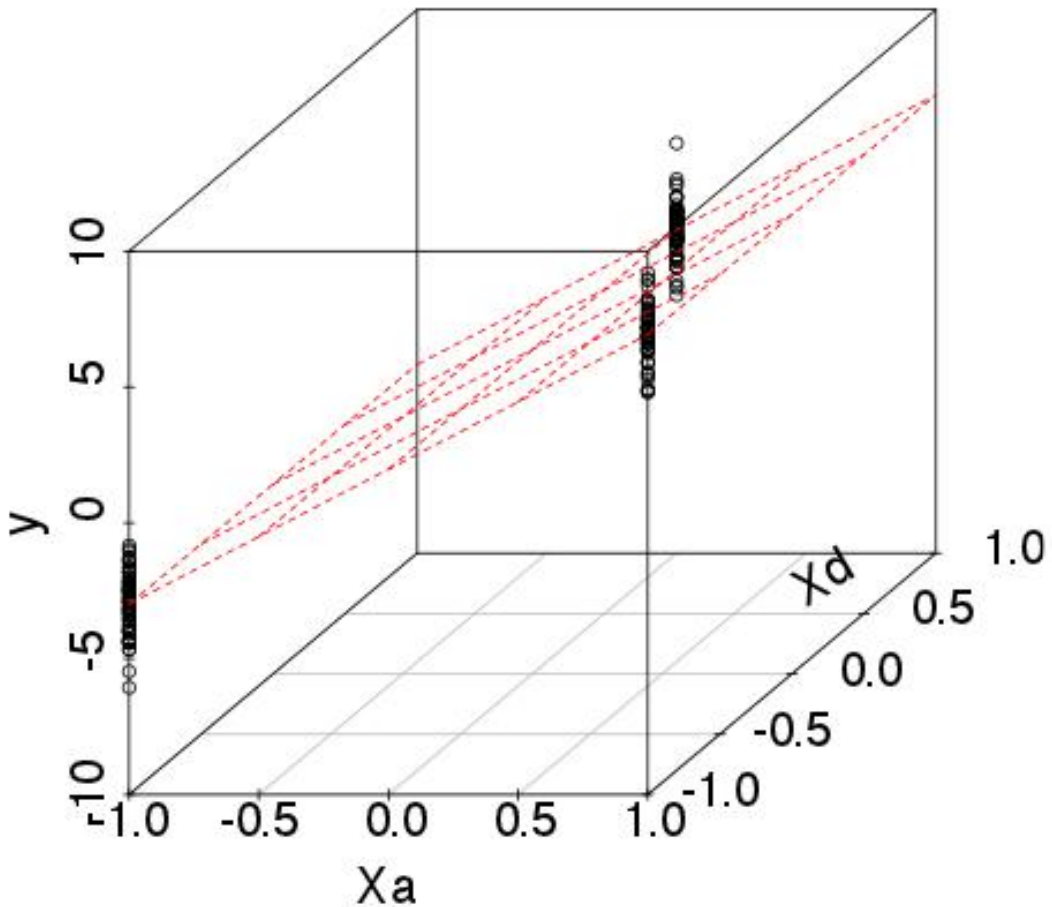
$$2.1 = 0.3 + (0)(-0.2) + (1)(1.1) + 0.7$$

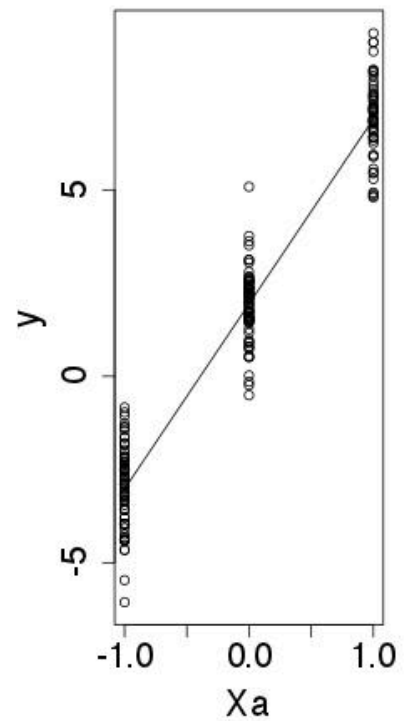
The genetic probability model IV

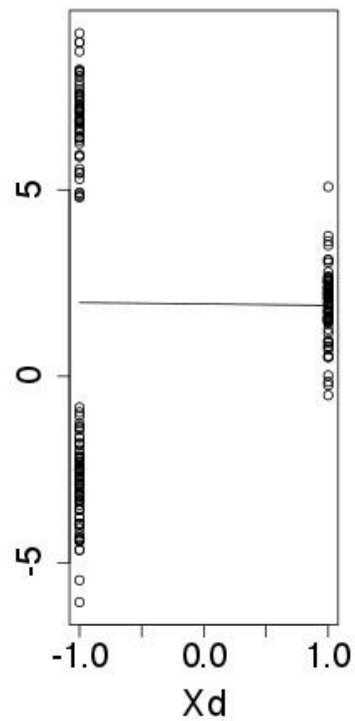
- Note that, while somewhat arbitrary, the advantage of the *X*a and *X*d coding is the parameters β_a and β_d map directly on to relationships between the genotype and phenotype that are important in genetics:
  - If $\beta_a \neq 0, \beta_d = 0$ then this is a "purely" additive case
  - If $\beta_a = 0, \beta_d \neq 0$ 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!)

Review: Genetic example 1

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



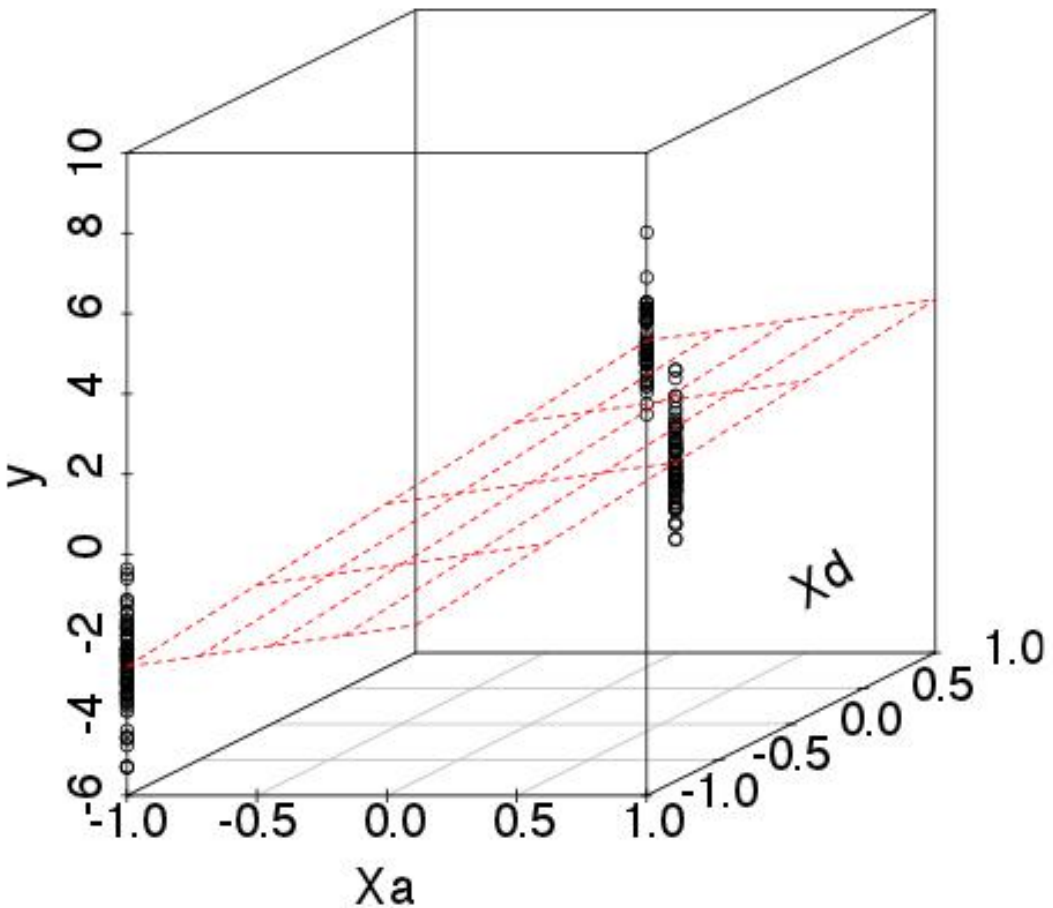


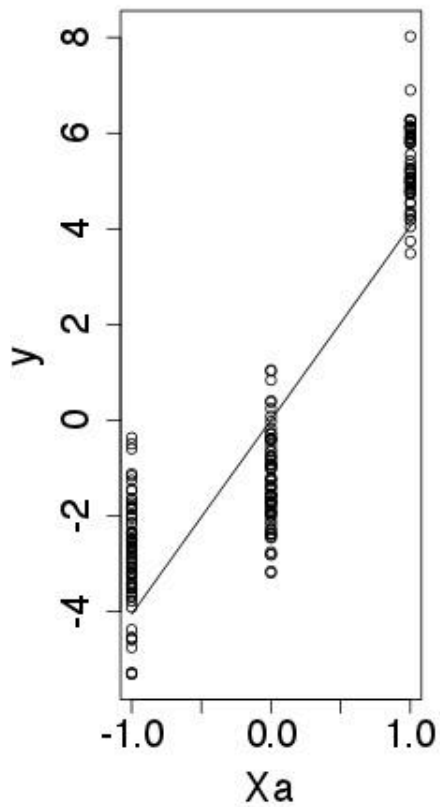


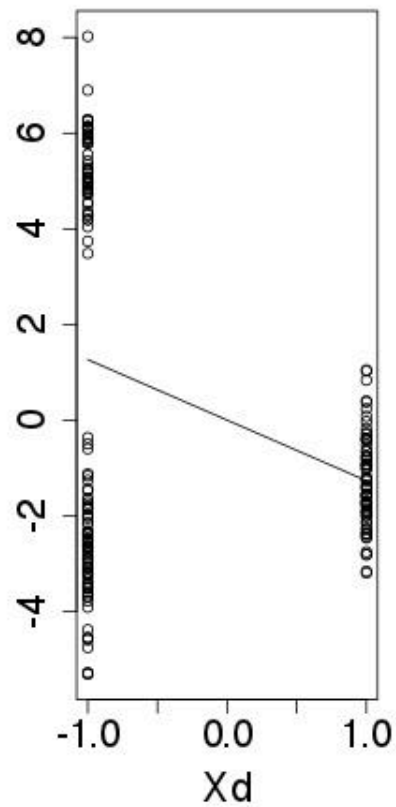
Review: Genetic example II

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

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



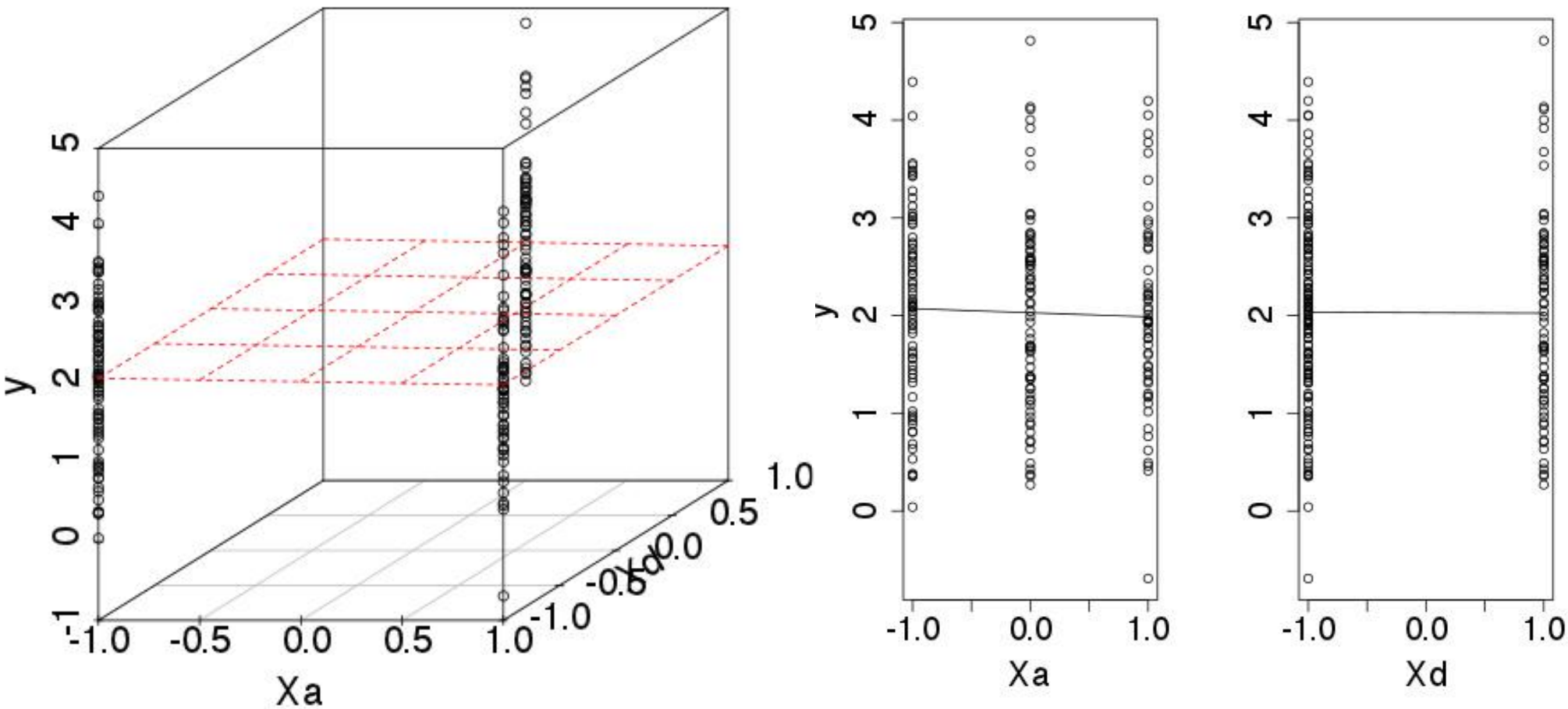




Review: Genetic example III

- 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):

$$Y = P$$

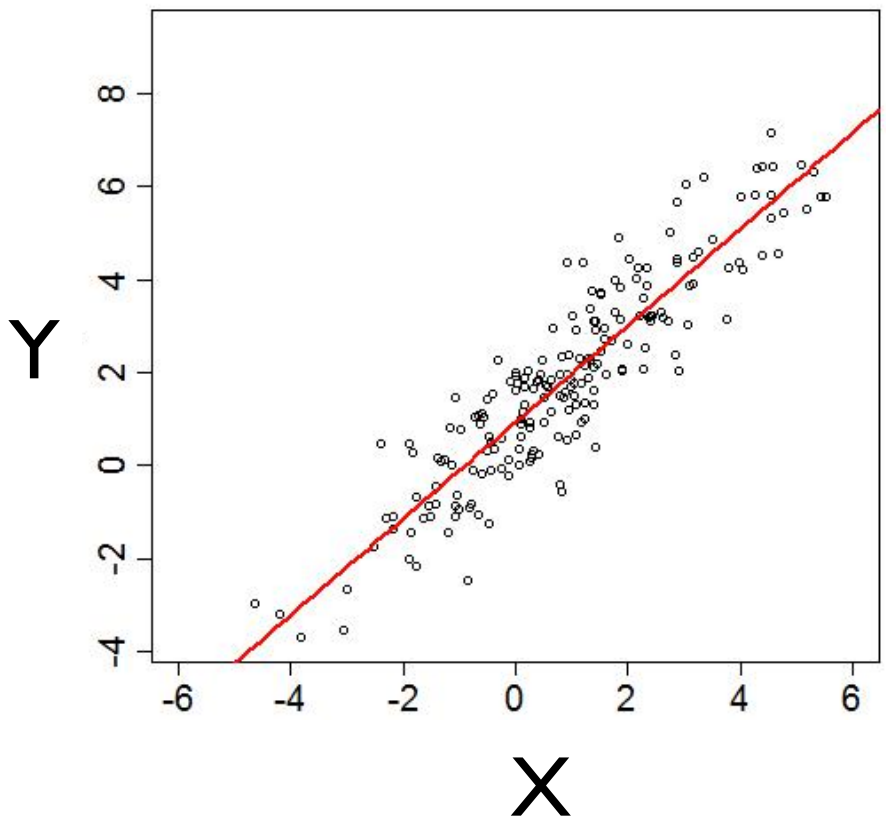
$$G = \mathrm{E}P = \mathrm{E}Y = \beta_\mu + X_a\beta_a + X_d\beta_d$$

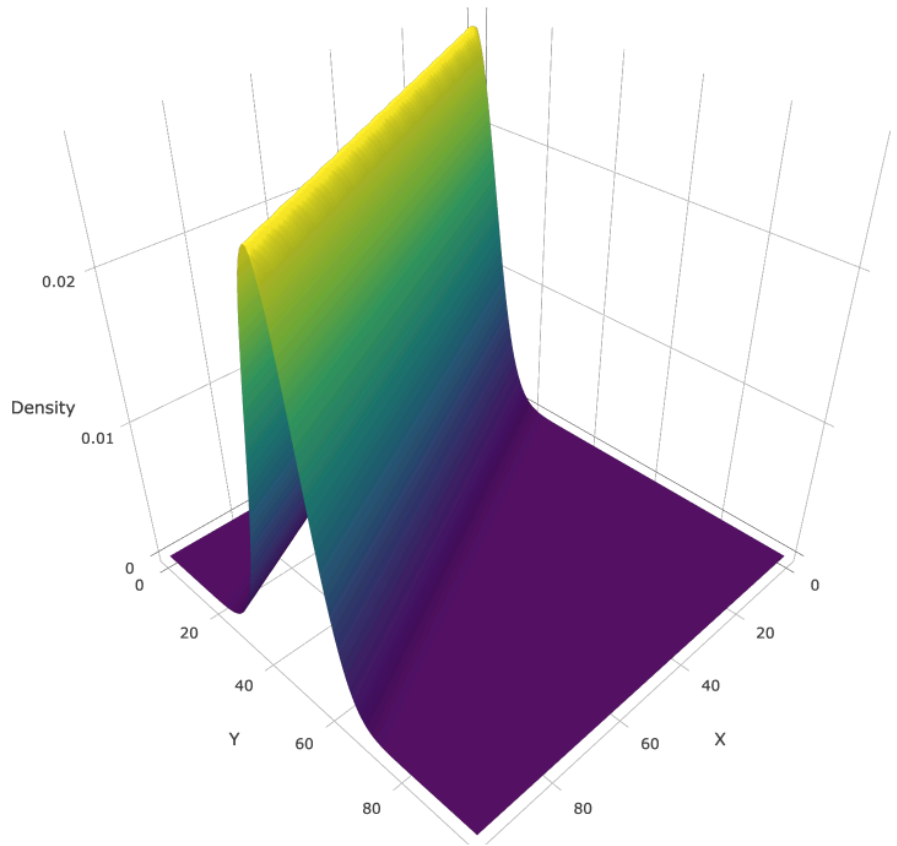
$$\epsilon = E$$

Linear regression III

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

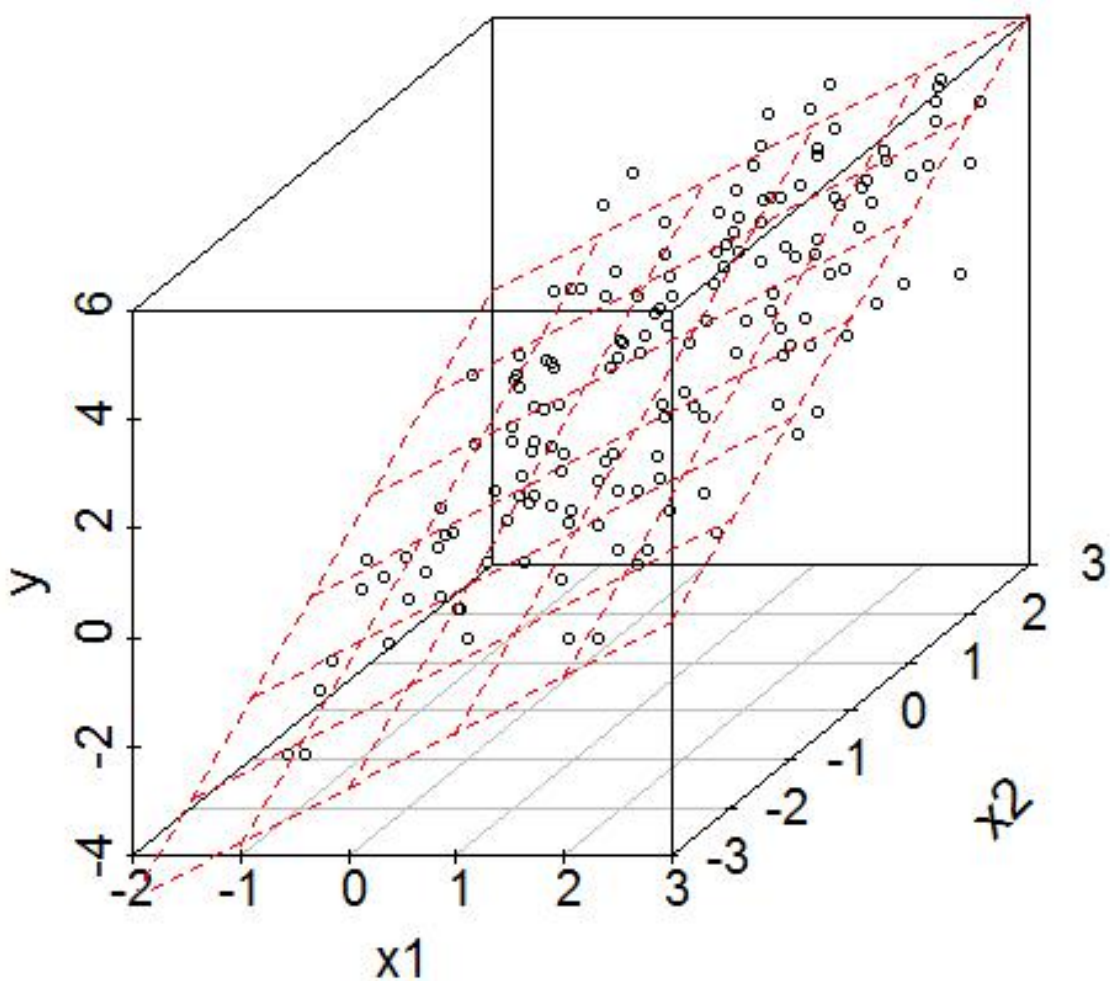
$$Y = \beta_0 + X\beta_1 + \epsilon \qquad \epsilon \sim N(0, \sigma_\epsilon^2)$$





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.,



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 (!!)

That's it for today

- Next lecture (Tues, March 14), we complete our initial discussion of genetic inference and begin our discussion of GWAS!