Quantitative Genomics and Genetics

BTRY 4830/6830; PBSB.5201.03

Lecture 20: Minimal GWAS analysis & Logistic Regression 1

Jason Mezey
April 13, 2023 (Th) 8:05-9:20

Announcements

- Your project has been assigned (!!)
- We will have (zoom) office hours this coming Mon (April 17) from 12:30-2:30 (great time to talk about the project…)
- For your final (same format as midterm!) you will do a GWAS analysis by doing a linear regression with and without covariates AND a logistic regression with and without covariates (!!)
- In computer lab today (Thurs., April 13) and tomorrow (Fri., April 14) your TAs will go over how to perform a linear regression with covariates (and provide code for how to do a PCA for population structure!)

Summary of lecture 20: Minimal GWAS & Logistic Regression 1

- Last lecture, we discussed population structure (and accounting for population structure with a covariate!)
- Today, we will also discuss "minimal steps" to perform in a GWAS analysis (!!)
- We will begin also begin our discussion of the last major (non-optional!) topic: logistic regression

Minimal GWAS 1

- You have now reached a stage when you are ready to perform a real GWAS data on your own (please note that there is more to learn and analyzing GWAS well requires that you jump in and analyze!!)
- Our final concept to allow you to do this are *minimal GWAS steps*, i.e. a list of analyses you should always do when analyzing GWAS data (you now know how to do most of these, a few you will have to do additional work to figure out)
- While these minimal steps are fuzzy (=they do not apply in every situation!) they provide a good guide to how you should think about analyzing your GWAS data (in fact, no matter how experienced you become, you will always consider these steps!)

Minimal GWAS II

- The minimal steps are as follows:
  - (1) Make sure you understand the data and are clear on the components of the data
  - (2) Check the phenotype data
  - (3) Check and filter the genotype data
  - (4) Perform a GWAS analysis and diagnostics
  - (5) Present your final analysis and consider other evidence
- Note I: the software PLINK (google it!) is a very useful tool for some (but not all) of these steps (but you can do everything in R!)
- Note II: GWAS analysis is not "do this and you are done" - it requires that you consider the output of each step (does it make sense? what does it mean in this case?) and that you use this information to iteratively change your analysis / try different approaches to get to your goal (what is this goal!?)

Minimal GWAS III: check data

- Look at the files (!!) using a text editor (if they are too large to do this - you will need another approach)
- Make sure you can identify: phenotypes, genotypes, covariates, and that you know what all other information indicates, i.e. indicators of the structure of the data, missing data, information that is not useful, etc. (also make sure you do not have any strange formatting, etc. in your file that will mess up your analysis!)
- Make sure you understand how phenotypes are coded and what they represent (how are they collected? are they the same phenotype?) and the structure of the genotype data (are they SNPs? are there three states for each?) - ideally talk to your collaborator about this (!!)

Minimal GWAS IV: phenotype data

- Plot your phenotype data (histogram!)
- Check for odd phenotypes or outliers (remove if applicable)
- Make sure it conforms to a distribution that you expect and can model (!!) - this will determine which analysis techniques you can use
    - e.g. if the data is continuous, is it approximately normal (or can be transformed to normal?)
    - e.g. if it has two states (see logistic regression this week!), make sure you have coded the two states appropriately and know what they represent (are there enough in each category to do an analysis?
    - e.g. what if your phenotype does not conform to either?

Minimal GWAS V: genotype data (“human centric rules"…)

- Make sure you know how many states you have for your genotypes and that they are coded appropriately
- Filter your genotypes (fuzzy rules! e.g., for humans = not all organisms!):
  - Remove individuals with >10% missing data across all genotypes (also remove individuals without phenotypes!)
  - Remove genotypes with >5% missing data across the entire individual
  - Remove genotypes with MAF < 5%
  - Remove individuals that fail a test of Hardy-Weinberg equilibrium (where appropriate!)
  - Remove individuals that fail transmission, sex chromosome test, etc.
- Perform a Principal Component Analysis (PCA) to check for clustering of n individuals (population structure!) or outliers, i.e. use the covariance matrix among individuals after scaling genotypes (by mean and sd) and look at the loadings of each individual on the PCs (you may have to “thin” the data!)

Minimal GWAS VI: GWAS analysis

- Perform an association analysis considering the association of each marker one at a time (always do this not matter how complicated your experimental design!)

- Apply as many individual analyses as you find informative (e.g. perform individual GWAS each with different statistical covariate sets)

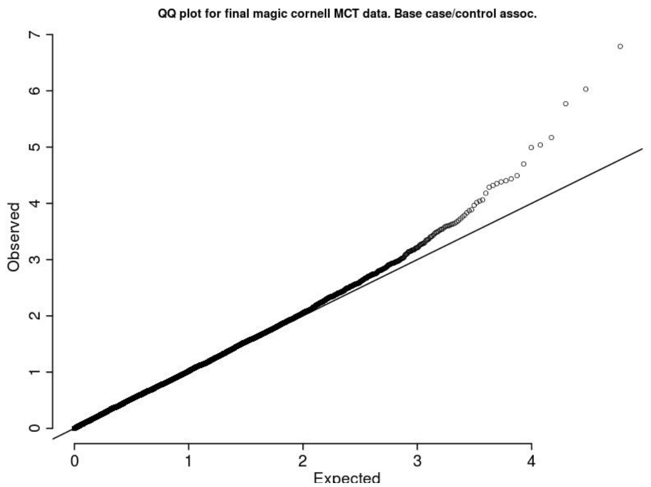
- CHECK QQ PLOTS FOR EACH INDIVIDUAL GWAS ANALYSIS and use this information to indicate if your analysis can be interpreted as indicating the positions of causal polymorphisms (if not, try more analyses, different filtering, etc. = experience is key!)

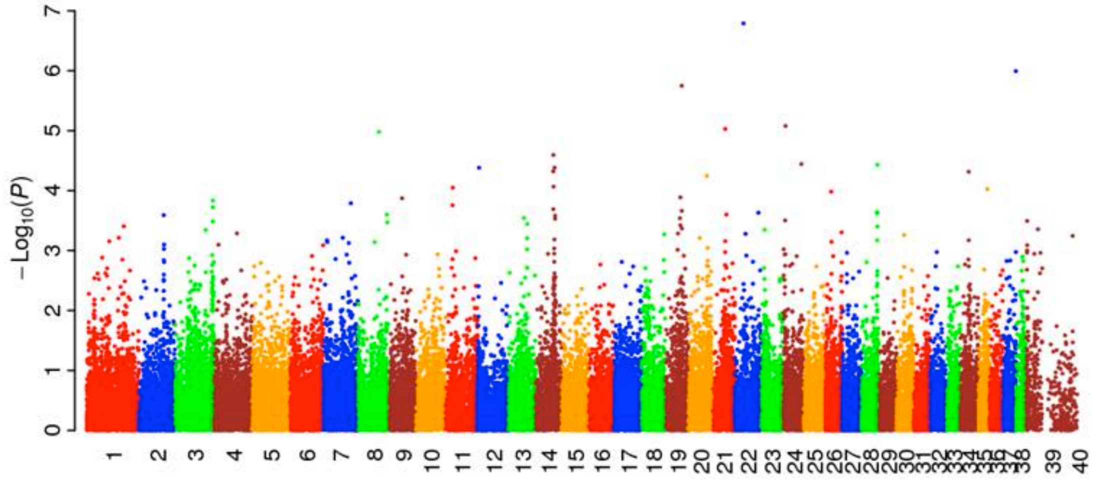
- For significant markers (multiple test correction!) do a "local" Manhattan plot and visualize the LD among the markers (r^2 or D' if possible but just a correlation of you Xa can work) to determine if anything might be amiss

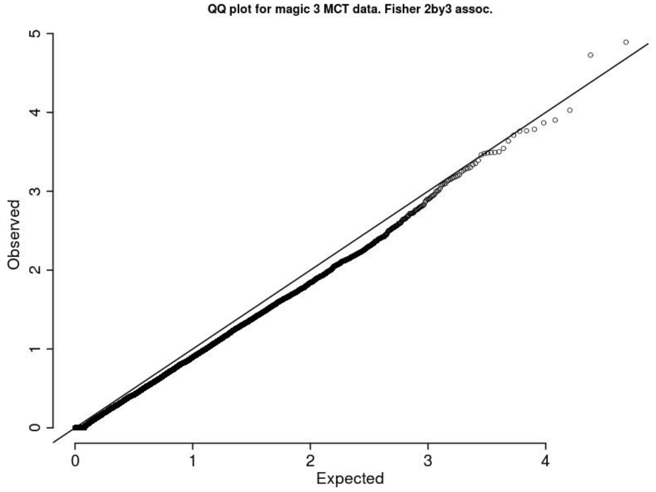
- Compare significant "hits" among different analyses (what might be causing the differences if there are any?)

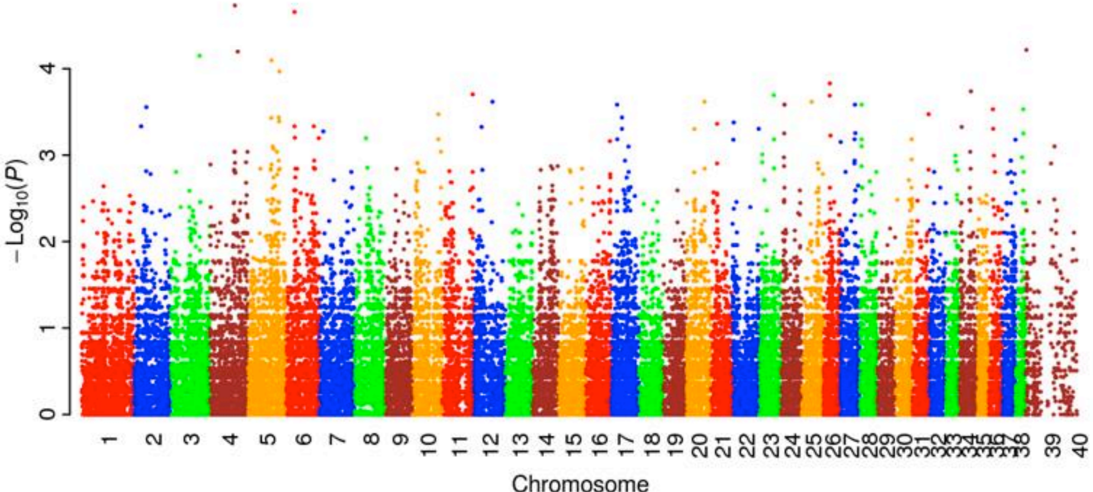
Comparing results of multiple analyses of the same GWAS data I

- I’ve run my initial analyses using covariate models and produced the following (now what!?):









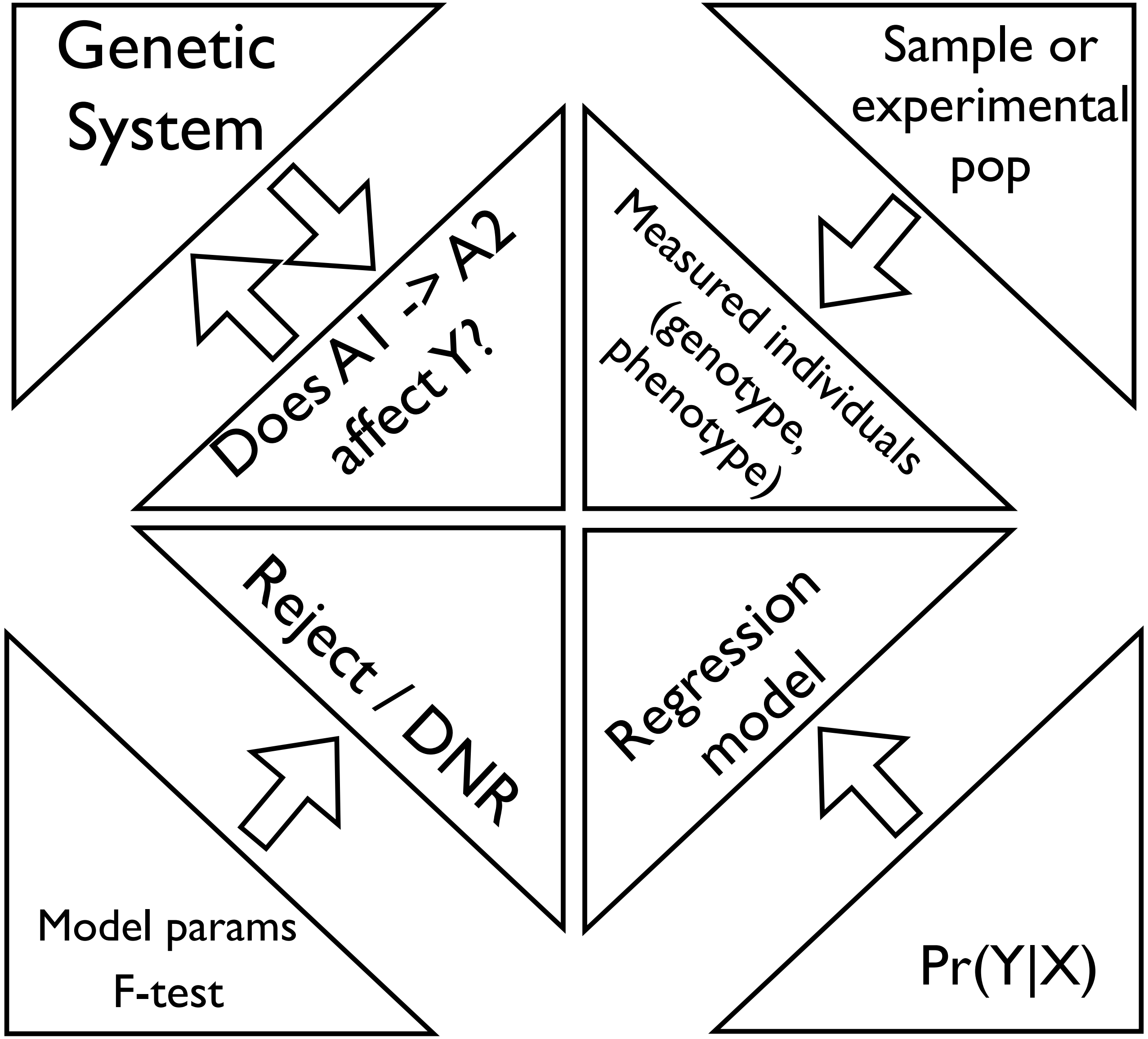
Comparing results of multiple analyses of the same GWAS data II

- The best case is that the same markers (SNPs) pass a multiple test correction regardless of the testing approach used, i.e. the result is robust to testing approach.
- In cases where this does not happen (most) it becomes helpful to understand why test results could be different:
  - Are particular covariate models altering the results if included/excluded? Why might this be?
  - Are some tests more powerful than others or depend on certain assumptions being true?
  - Does it depend on how you partition the data (e.g. batch effects)?
- This can help narrow down the set of tests you feel are the most informative. In general, a good publishing strategy is limiting yourself to tests that both give you significant results that you believe!

Minimal GWAS VII: present results

- List ALL of the steps (methods!) you have taken to analyze the data such that someone could replicate what you did from your description (!!), i.e. what data did you remove? what intermediate analyses did you do? how did you analyze the data? if you used software what settings did you use?
- Plot a Manhattan and QQ plot (at least!)
- Present your hits (many ways to do this)
- Consider other information available from other sources (databases, literature) to try to determine more about the possible causal locus, i.e. are there good candidate loci, control regions, known genome structure, gene expression or other types of data, pathway information, etc.

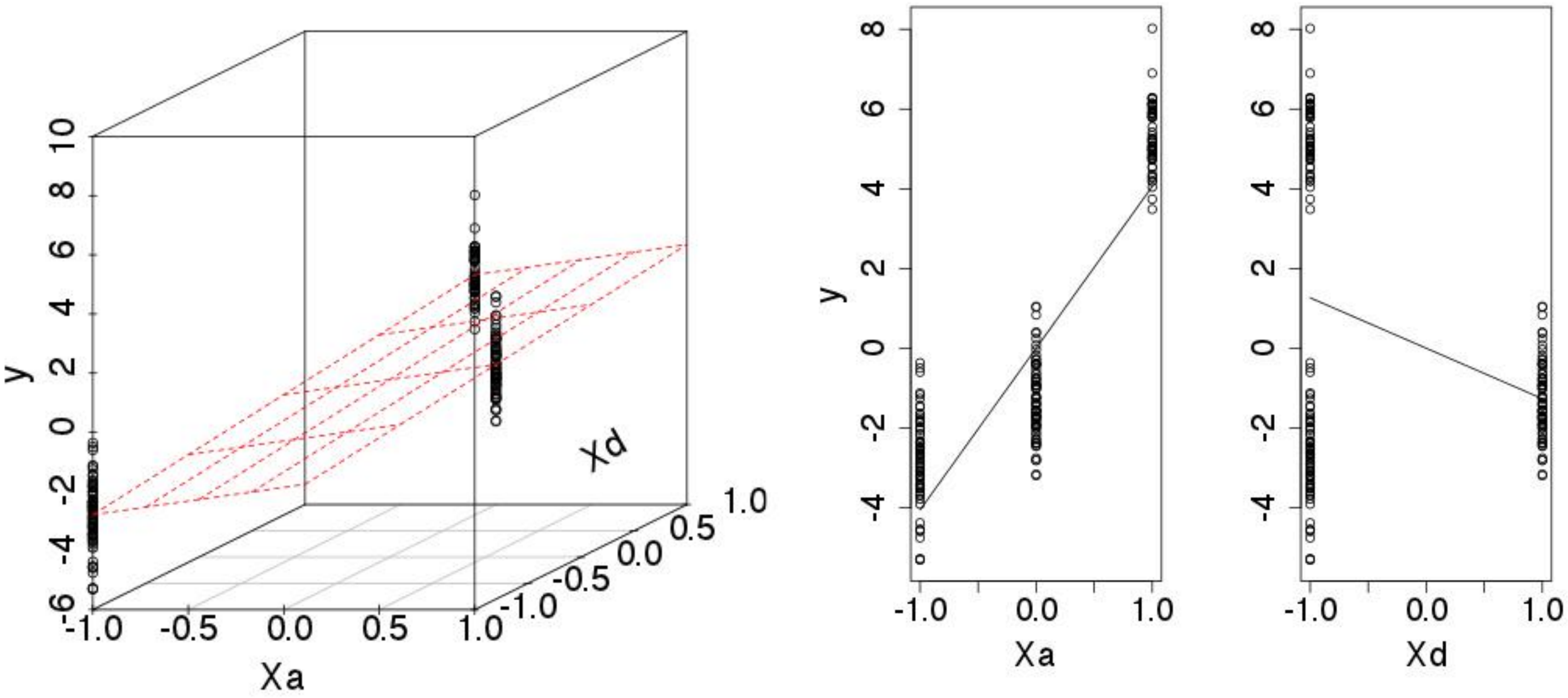
Conceptual Overview



Linear regression review

- So far, we have considered a linear regression is a reasonable model for the relationship between genotype and phenotype (where this implicitly assumes a normal error provides a reasonable approximation of the phenotype distribution given the genotype):

$$Y = \beta_\mu + X_a\beta_a + X_d\beta_d + \epsilon \qquad \epsilon \sim N(0, \sigma_\epsilon^2)$$



Case / Control Phenotypes I

- While a linear regression may provide a reasonable model for many phenotypes, we are commonly interested in analyzing phenotypes where this is NOT a good model

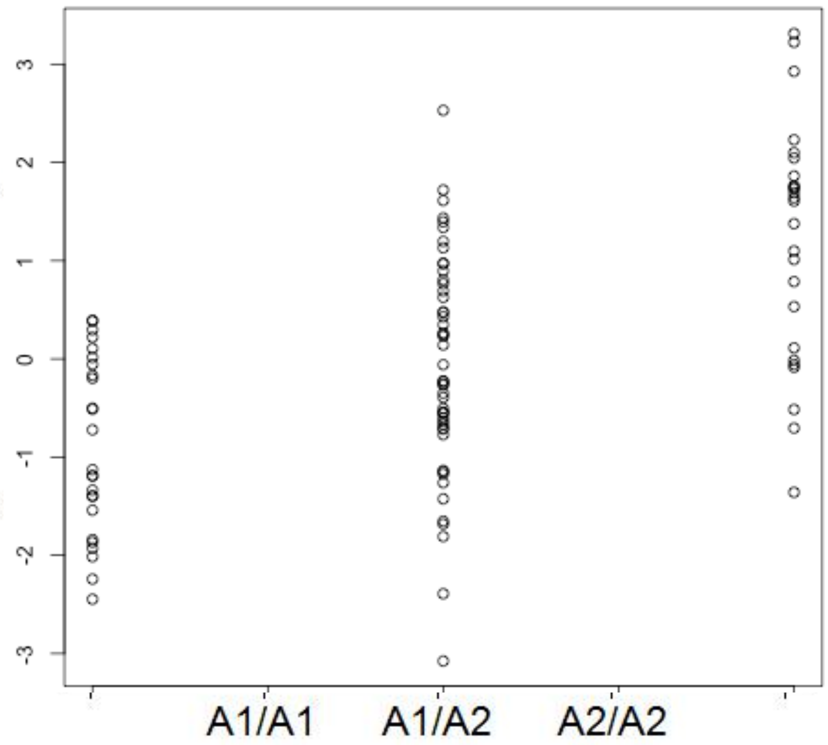
- As an example, we are often in situations where we are interested in identifying causal polymorphisms (loci) that contribute to the risk for developing a disease, e.g. heart disease, diabetes, etc.

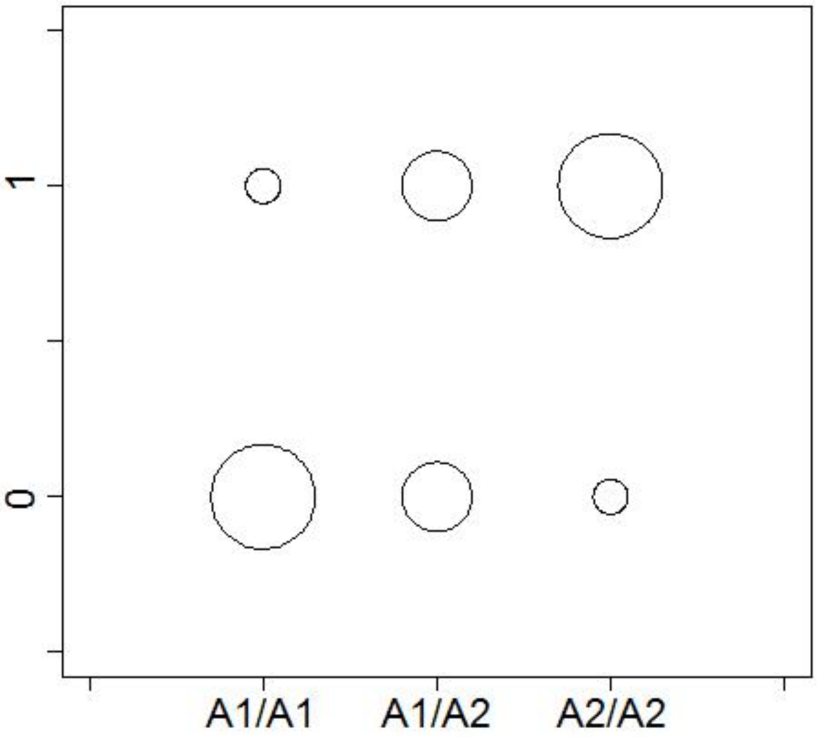
- In this case, the phenotype we are measuring is often “has disease” or “does not have disease” or more precisely “case” or “control”

- Recall that such phenotypes are properties of measured individuals and therefore elements of a sample space, such that we can define a random variable such as $Y(\text{case}) = 1$ and $Y(\text{control}) = 0$

Case / Control Phenotypes II

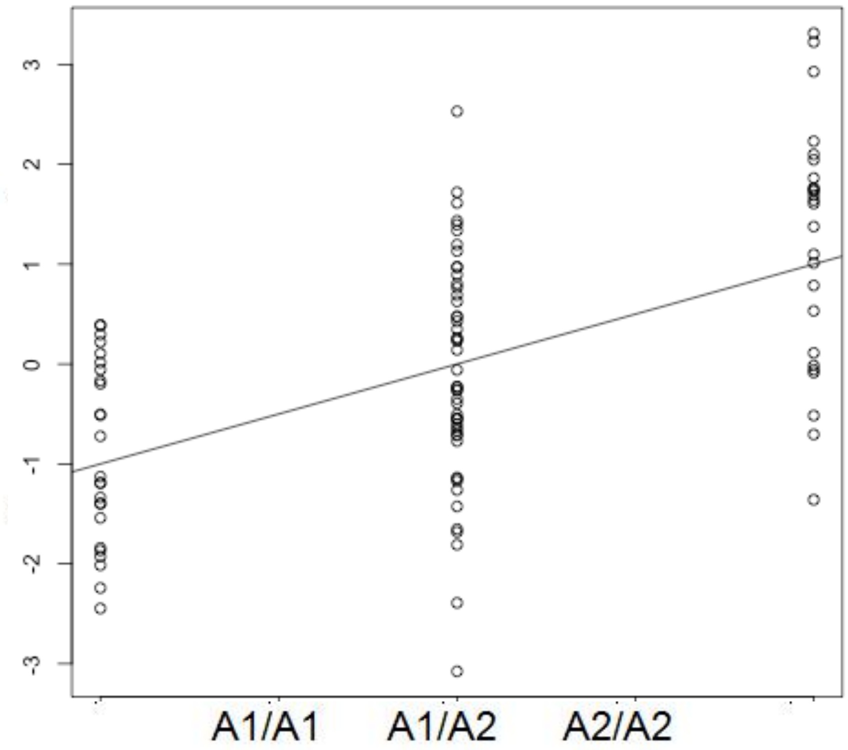
- Let’s contrast the situation, let’s contrast data we might model with a linear regression model versus case / control data:

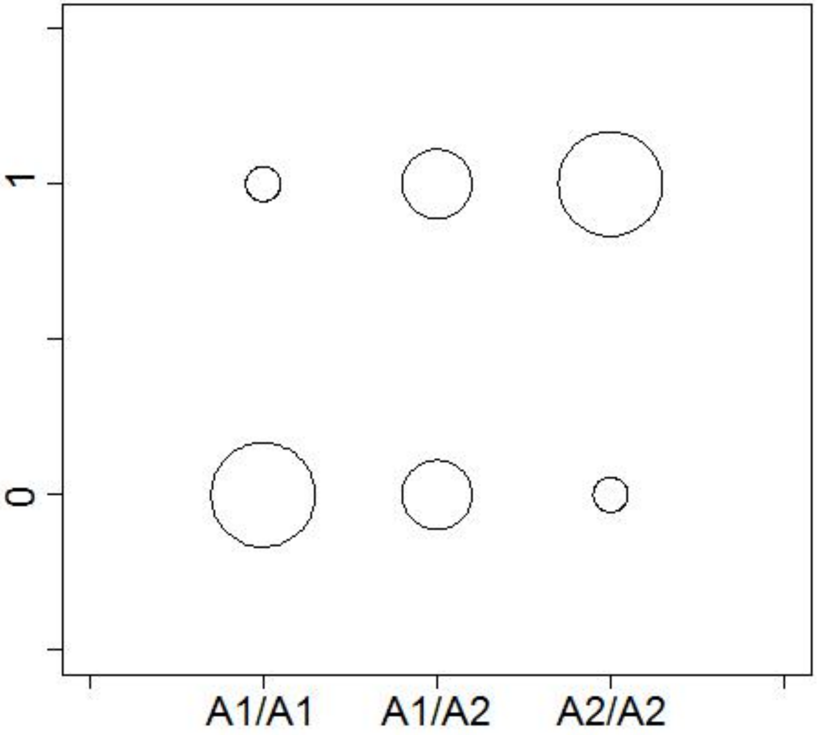




Case / Control Phenotypes II

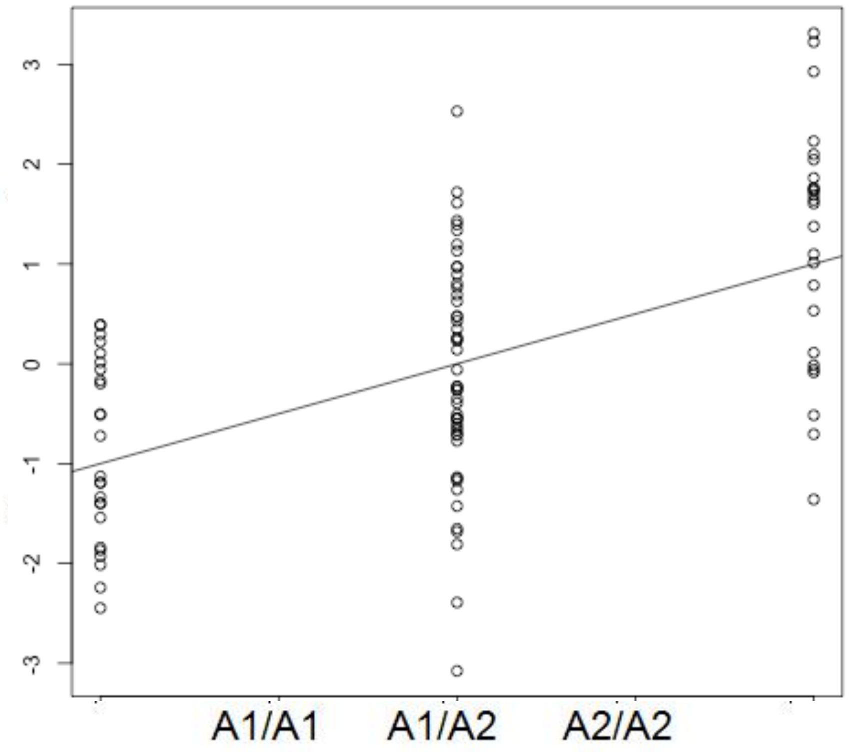
- Let’s contrast the situation, let’s contrast data we might model with a linear regression model versus case / control data:

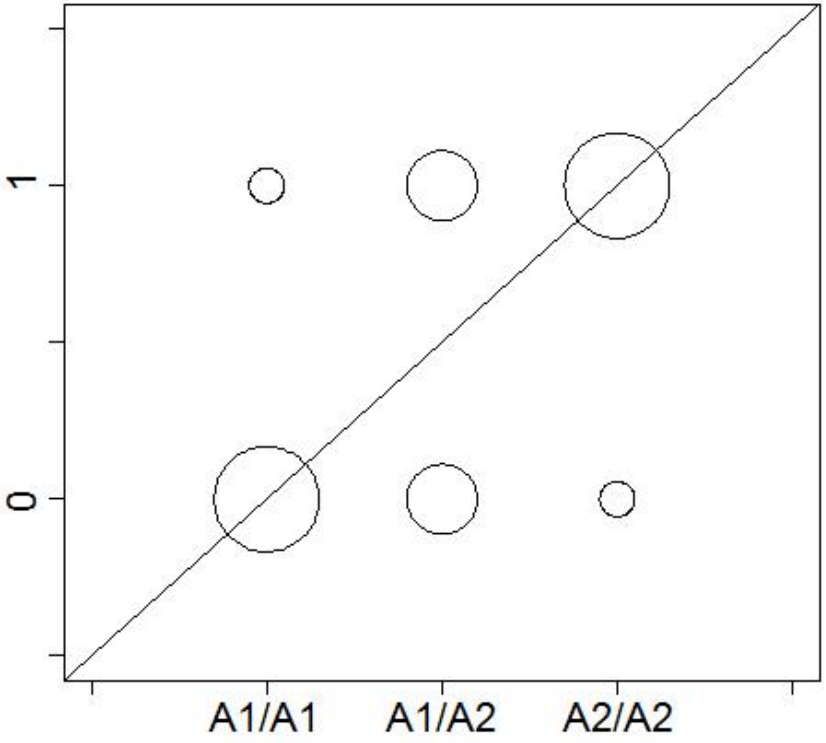




Case / Control Phenotypes II

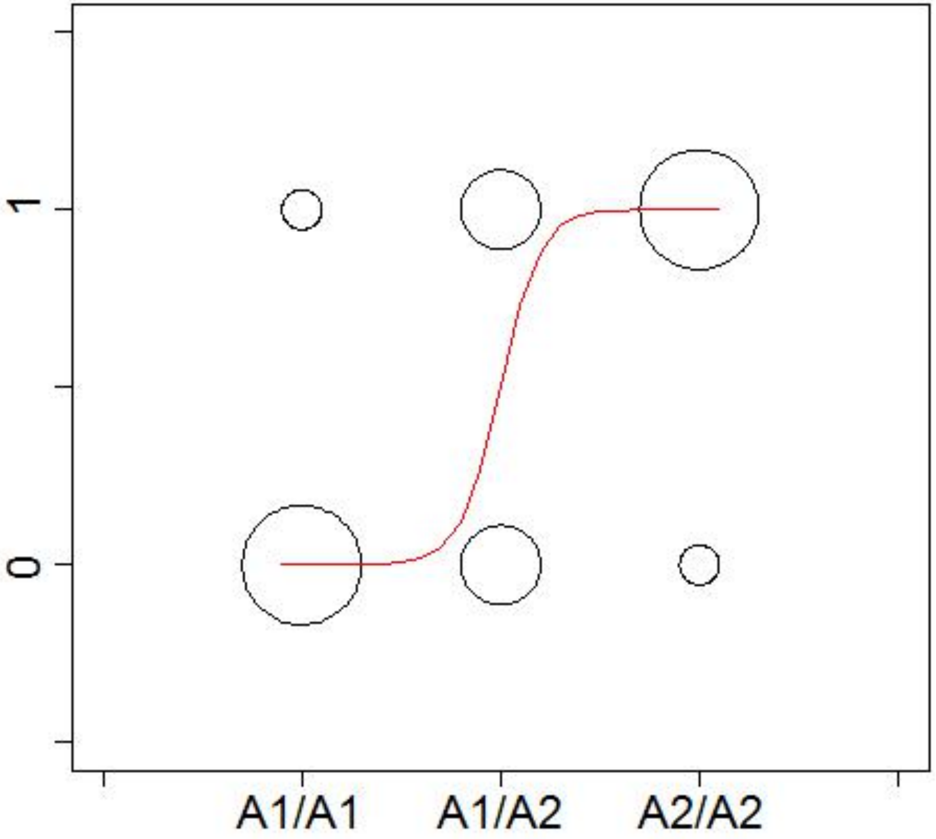
- Let’s contrast the situation, let’s contrast data we might model with a linear regression model versus case / control data:





Logistic regression I

- Instead, we're going to consider a logistic regression model



Logistic regression II

- It may not be immediately obvious why we choose regression “line” function of this “shape”
- The reason is mathematical convenience, i.e. this function can be considered (along with linear regression) within a broader class of models called Generalized Linear Models (GLM) which we will discuss next lecture
- However, beyond a few differences (the error term and the regression function) we will see that the structure and out approach to inference is the same with this model!

Logistic regression III

- To begin, let's consider the structure of a regression model:

$$Y = logistic(\beta_\mu + X_a\beta_a + X_d\beta_d) + \epsilon_l$$

- We code the "X's" the same (!!) although a major difference here is the "logistic" function as yet undefined

- However, the expected value of Y has the same structure as we have seen before in a regression:

$$\mathrm{E}(Y_i|X_i) = logistic(\beta_\mu + X_{i,a}\beta_a + X_{i,d}\beta_d)$$

- We can similarly write for a population using matrix notation (where the X matrix has the same form as we have been considering!):

$$\mathrm{E}(\mathbf{Y}|\mathbf{X}) = logistic(\mathbf{X}\beta)$$

- In fact the two major differences are in the form of the error and the logistic function

That’s it for today

- Next lecture we will continue our discussion of logistic regression!