Quantitative Genomics and Genetics

BTRY 4830/6830; PBSB.52201.03

Lecture 22: Intro to Logistic Regression II

Jason Mezey

April 16, 2024 (T) 8:40-9:55

Announcements 1

- All homeworks have been graded and your midterm will be graded / available next week
- The final (required!) work for the class:
  - Final exam
  - One more computer lab (of three!) this week (Fri, April 19)
  - Class project, assigned today (Apr 16 - see following!)

Announcements II

- Your FINAL EXAM (!!):
  - SAME format as the midterm (take-home, open book, work alone!)
  - Will be made available May 11 and will be DUE by 11:59pm May 18 (ie Cornell, Ithaca exam week
  - Will be designed to take 2-3 hours IF you prepare (e.g., your understanding, code, etc.) ahead of time!
  - You will have to perform a GWAS analysis by applying a LINEAR regression with and without covariates AND a LOGISTIC regression with and without covariates (plus Manhattan plots, QQ plots, and some written answers) - That's it (!!)

Announcements III

- Last class topics covered in lecture:
  - Logistic regression (this week!)
  - Basics of mixed models (optional!)
  - (Brief) intro to Bayesian statistics (optional!)
  - Basics of pedigree, inbred line, evolutionary analysis (optional!)
- Last class topics covered in labs
  - Logistic regression (this week: April 19) - REQUIRED (!!)
  - Mixed models + review (next week: April 26) - optional!
  - MCMC algorithms for Bayesian inference + review (following: May 3) - optional!

BIOCB 4830/6830 & PBSB.5201.01
Quantitative Genomics and Genetics Spring 2024

Project - posted April 16

Due 11:59PM May 7

1 Introduction and instructions

The goal of the class project is for you to demonstrate what you have learned by performing a GWAS analysis on real data. To accomplish this, assume that you have been provided data by a collaborator who wants to identify positions of causal polymorphisms (loci). You will perform an in-depth analysis and write a report for your collaborator that explains your methods and results.

Instructions: While we provide some general guidelines for how to proceed below, the techniques you use to analyze the data and how you construct your report will be up to you. Do however note the following instructions (PLEASE READ THESE CAREFULLY!!):

(1) Your project must be uploaded by 11:59PM, May 7 - if it is late for any reason, standard grading policies apply.

(2) You are allowed to work together with other students in the class to analyze these data. However, note that turning in a report that describes exactly the same analyses as a fellow student is not a good strategy for getting a good grade. Also note that you must write your own report.

(3) This is an 'open book' assignment, such that you are allowed to use any resources online, ChatGPT, in books, etc. You may also ask third-party (i.e. people not in the class) for suggestions on what analyses to perform but you cannot have a third-party do any of the analyses (or write any code for you!).

(4) You are also allowed to use any software or programming language that you would like as part of your analysis. However, we expect that some of the tasks will be performed in R (also note that you are welcome to use any packages, functions, etc. in R).

(5) Your final project will include at most three files a SINGLE report file (ideally a .pdf), a SINGLE file including all of your R code (ideally an .rmd file!) and / or commands or scripts you used to run other software packages, and IF YOU WANT a SINGLE, a pdf or html conversion of your .rmd. That is, for your R code, the best way to maximize your grade is to have well commented code that we can run from the command line. If you use other software for some of the tasks, a reasonable approach is to include commented out descriptions in your

code that provides details on how you ran the software, e.g. what parameters did you use, etc.

(6) The report file must be no more than 8 pages (single-sided), with NO MORE than 5 pages of text and NO MORE than 3 pages of figures / tables.

(7) For your report, you must describe what you did in detail (a good guide is have you provided enough detail such that someone reading your report could replicate what you have done?). You also need to describe the results you have obtained from your analysis. You may also wish to include some text to describe interpretations and conclusions that may be of interest to your collaborator, including statistical and possibly, biological interpretations. For your Figures and Tables, note that clarity and clear labels is a strategy for maximizing your grade.

(8) We will grade on two broad criteria: 1. the overall quality of the analyses / report, 2. the amount of effort put into your project. Note that 'effort' does not mean run many analyses without thinking carefully about why you are running them or how they fit together to provide a clear picture of results. A guide maximizing your grade on effort is to think carefully about how to produce the best possible report that you can and then put in as many hours as you wish to devote to the project accomplishing this objective (your effort level will be clear to us).

2 The experiment and data

The experiment: About a decade ago, the large scale human genomics resources Genetic European Variation in Health and Disease (gEUVADIS) was made available (now a part of larger genomics consortium efforts but still relevant / relevant data!) - see the following links for relevant descriptions and information:

`http://www.internationalgenome.org/data-portal/data-collection/geuvadis/`

`https://www.nature.com/articles/nature12531`

with a samples from 4 different European populations. Each of these individuals were part of the 1000 Genomes project and their genomes were sequenced and analyzed to identify SNP genotypes. For expression profiling, lympoblastoid cell lines (LCL) were generated from each sample and mRNA levels were quantified through RNA sequencing.

Each of these gene expression measurements may be thought of as a phenotype and one can do a GWAS analysis on each individually, which is called an 'expression Quantitative Trait Locus' or 'eQTL' analysis, an unnecessarily fancy name for a GWAS when the phenotype is gene expression!

What you have been provided is a small subset of these data that are publicly available. Specifically, you have been provided 50,000 of the SNP genotypes for 344 samples from the CEU (Utah residents with European ancestry), FIN (Finns), GBR (British) and, TSI (Toscani) population. For these same individuals, you have also been provided the expression levels of five genes. You have also been provided information on the population and gender of each of these individuals, and information regarding the position of each gene and SNP in the genome.

The data: These have been provided to you in five total files: 'phenotypes.csv','genotypes.csv', 'covars.csv', 'gene_info.csv','SNP_info.csv' (within a 'data_files.zip').

'phenotypes.csv' contains the phenotype data for 344 samples and 5 genes.

'genotypes.csv' contains the SNP data for 344 samples and 50000 genotypes.

'covars.csv' contains the population origin and gender information for the 344 samples.

'gene_info.csv' contains information about each gene that was measured. The 'chromosome' column indicates the chromosome where the gene is located, 'start' marks the position in the chromosome where the region of the gene begins and 'end' marks the position where the region ends, 'symbol' contains the common gene name of the measured transcript and 'probe' contains the ids of the transcripts that match with the column names of the phenotype data.

'SNP_info.csv' contains the additional information on the genotypes and has four columns. The 1st column contains the chromosome number of each SNP, the 2nd column contains the physical position of the SNP on the chromosome, the 3rd column contains the abbreviation used to the 'rsID' = the name of each SNP in order.

3 Your assignment and hints for getting started

Your GWAS assignment is to find the position of as many causal polymorphisms as possible for the five expressed genes using the data (note that each 'hit' will potentially indicate an eQTL). You may / should use any and as many analysis approaches as you think that are useful to accomplish this goal. In your report, you will need to describe in detail what you did, why you did it, and describe results in a manner that your 'non-statistical' collaborator will be able to understand, e.g. explain your terms, provide interpretations, etc.

A few hints:

- Apply the applicable steps of a 'minimum GWAS' analysis.
- In your report, justify why you applied each individual step and statistical approach.
- In your report, provide a summary of your results and what they mean.
- You may want to consider going to various resources online (e.g. genecards, UCSC genome browser, dbSNP, many others) to incorporate biological information into your interpretation and hypotheses concerning what you may have found.
- Ask Beulah, Sam, and Jason for thoughts and ideas!

Good luck!

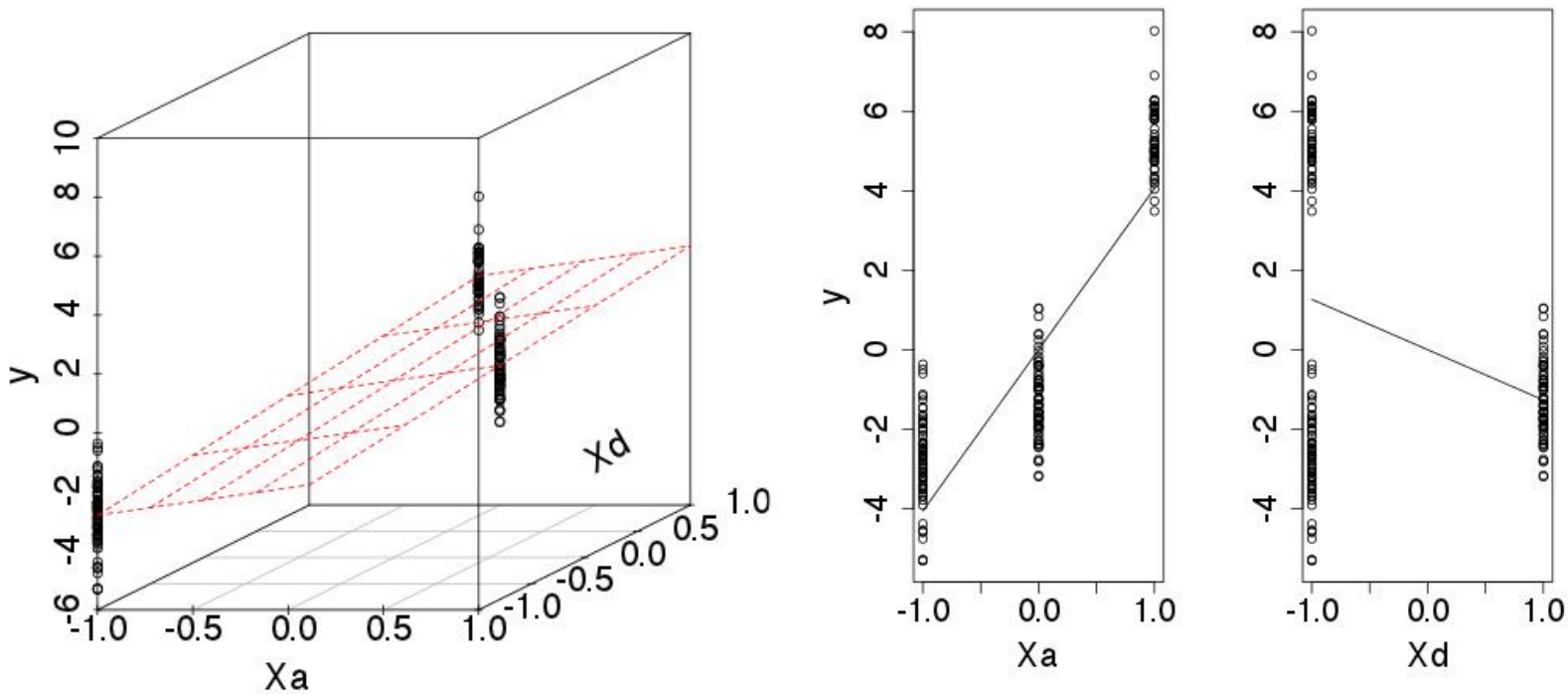
Summary of lecture 22: Logistic Regression II

- Last lecture, we began our discussion of the last major (non-optional!) topic: logistic regression
- Today we will continue our introduction!

Review: Linear regression

- 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):

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

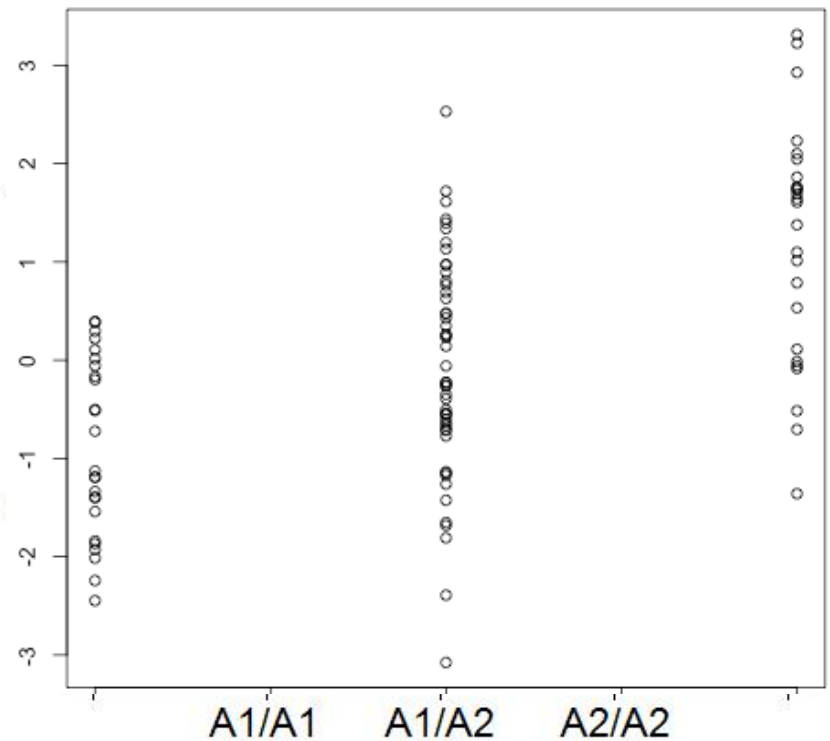


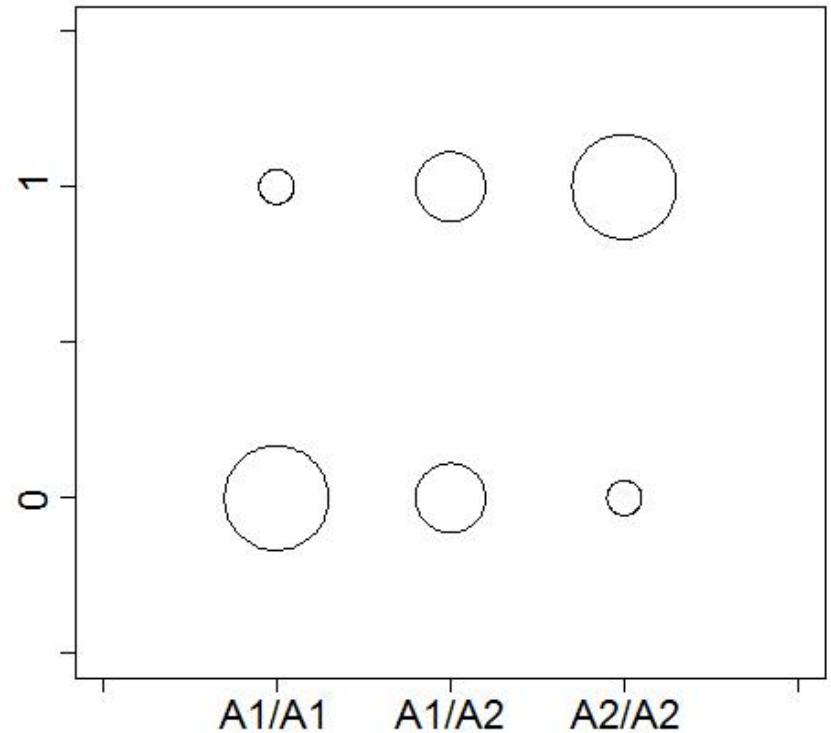
Review: Case / Control Phenotypes I

- 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

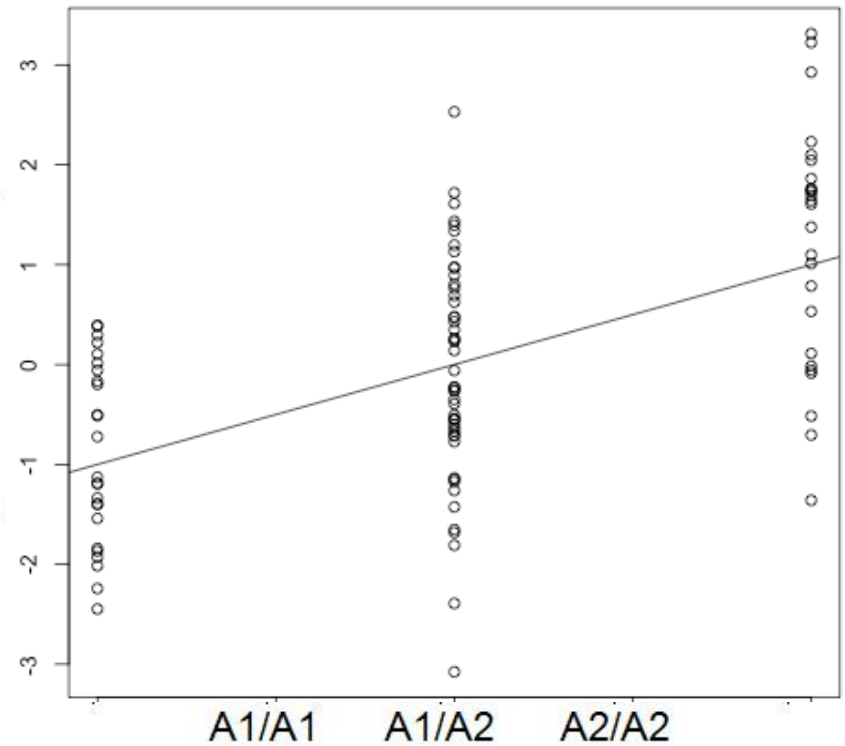
- Let's contrast the situation, let's contrast data we might model with a linear regression model versus case / control data:

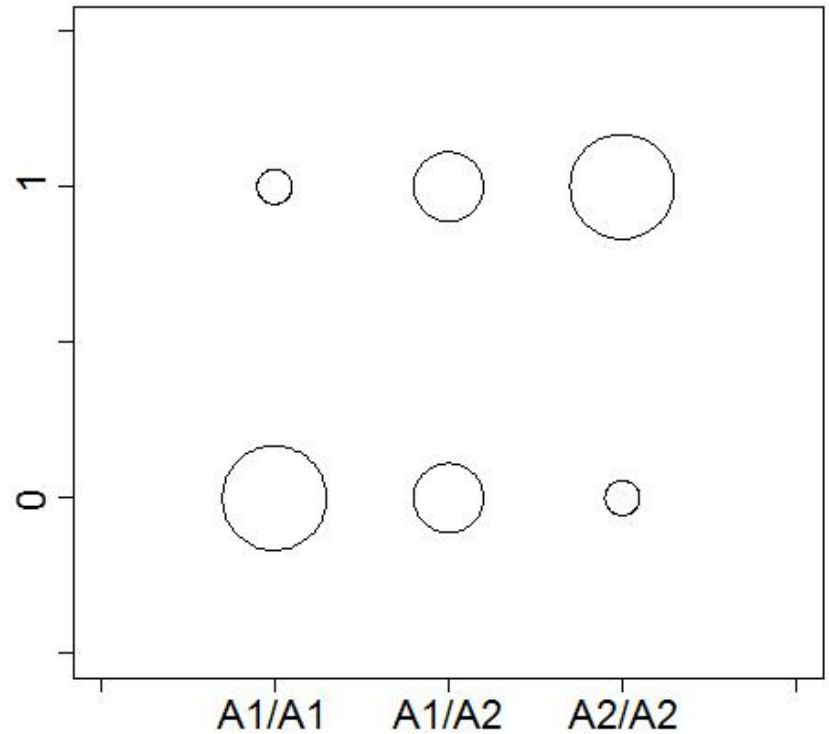




Case / Control Phenotypes II

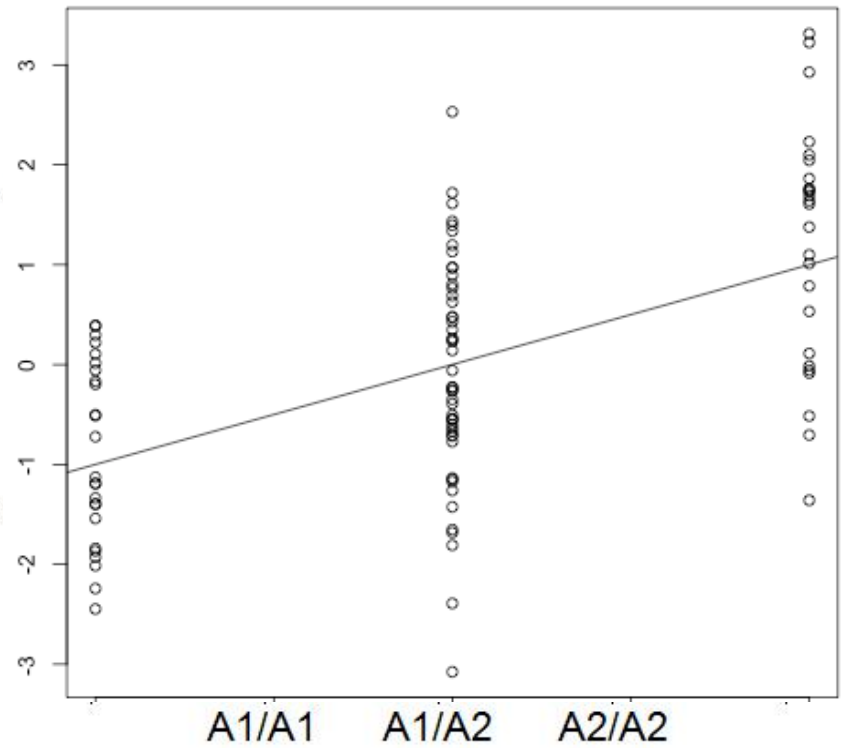
- Let's contrast the situation, let's contrast data we might model with a linear regression model versus case / control data:

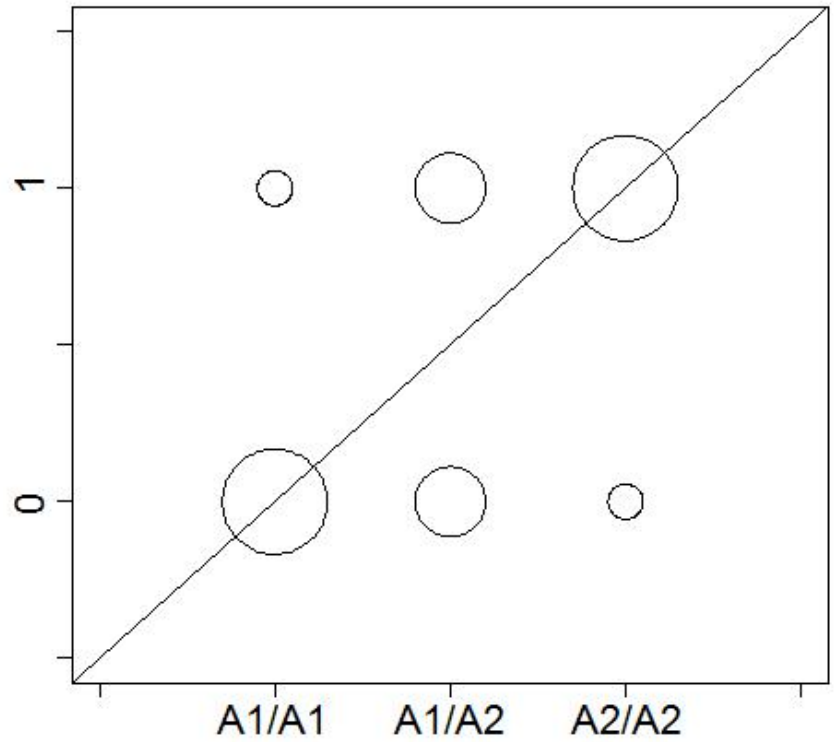




Case / Control Phenotypes II

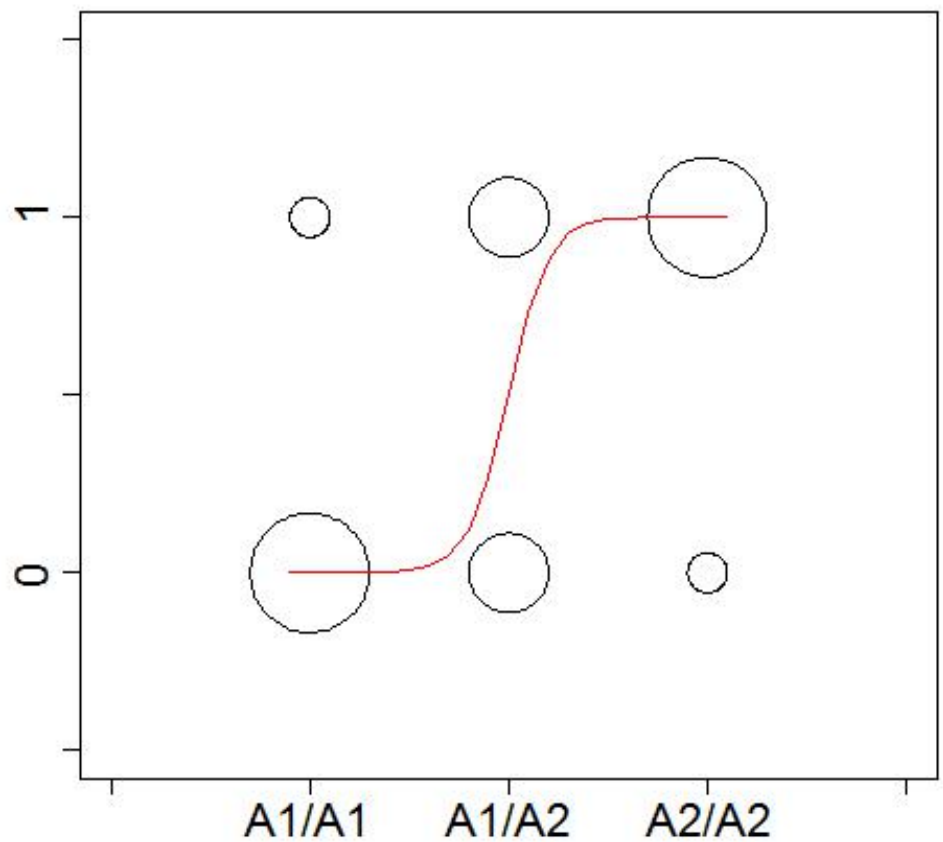
- Let's contrast the situation, let's contrast data we might model with a linear regression model versus case / control data:





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!