# Quantitative Genomics and Genetics

## BTRY 4830/6830; PBSB.5201.03

*Lecture 21: Logistic Regression II*

Jason Mezey

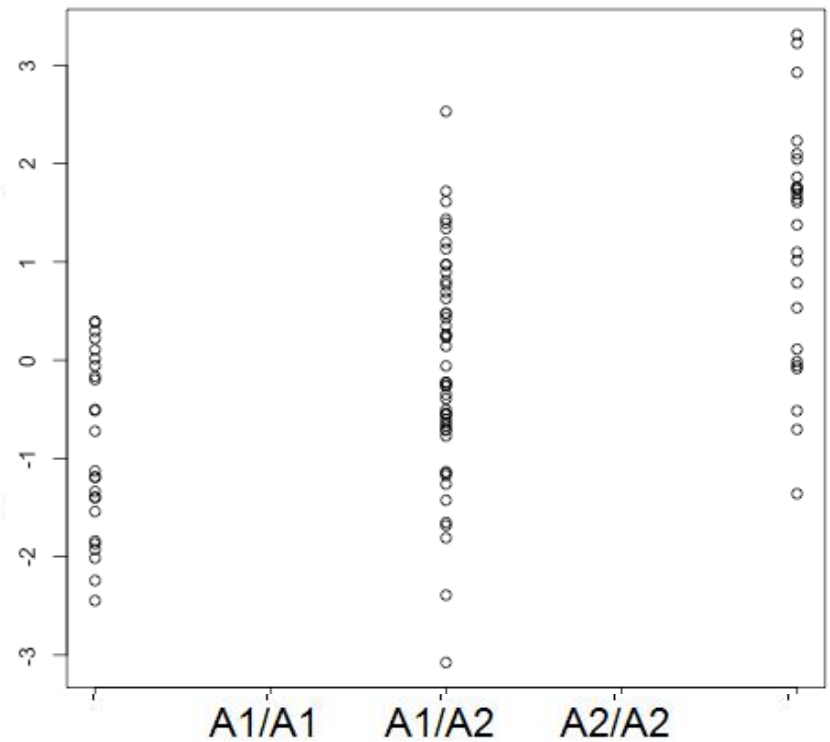
April 18, 2023 (T) 8:05-9:20

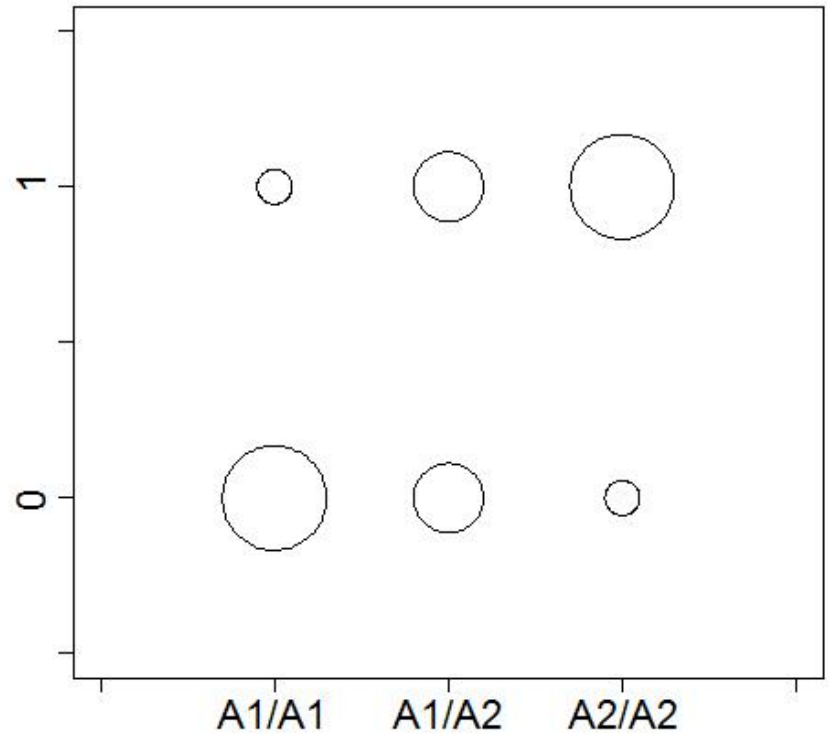
# Review: Case / Control Phenotypes I

- 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 $Y(\text{control}) = 0$

# Review: Case / Control Phenotypes II

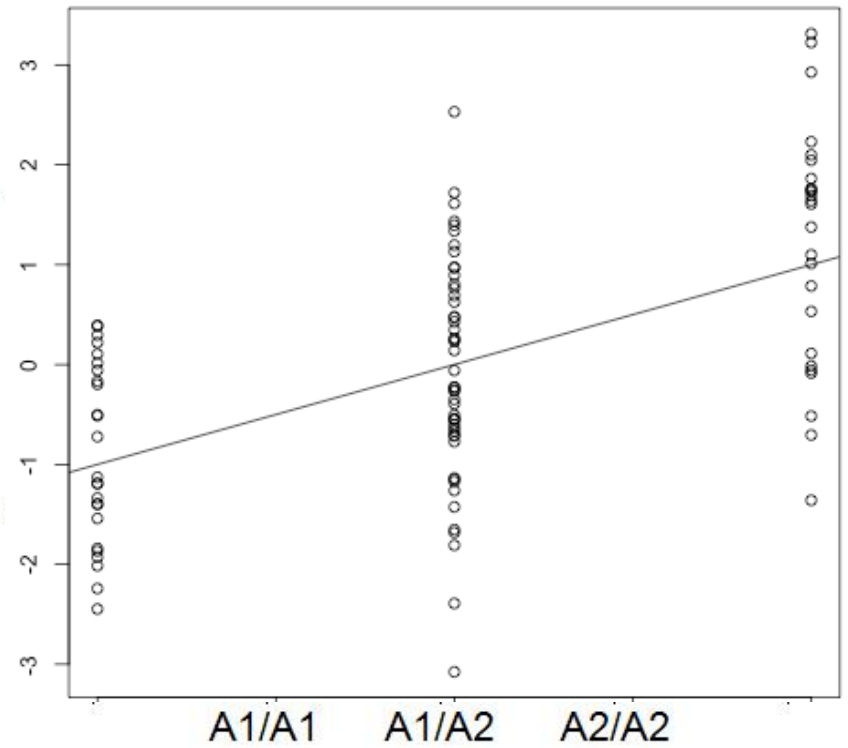
- Let's contrast the situation, let's contrast data we might model with a linear regression model versus case / control data:

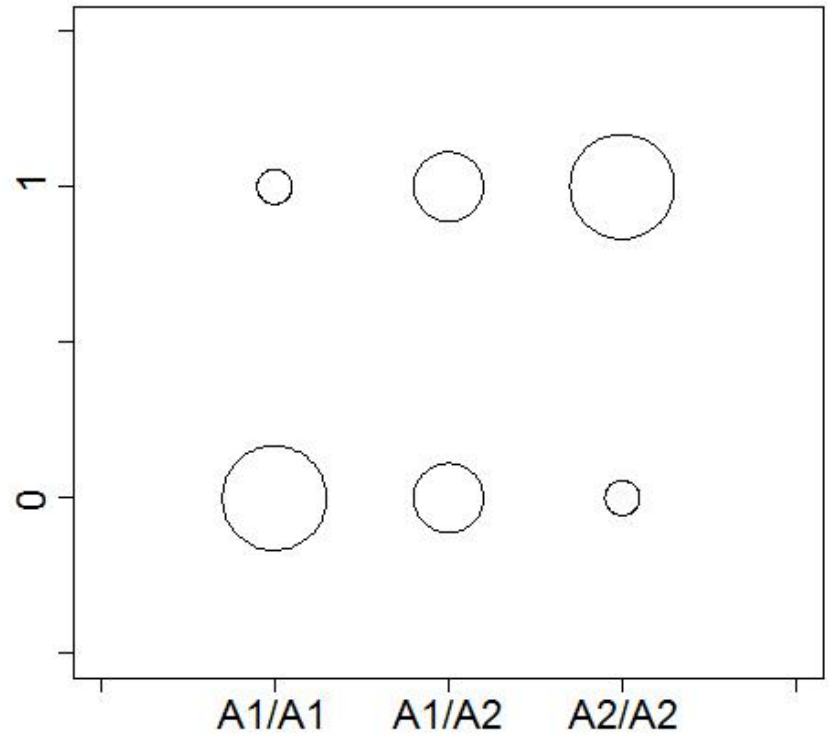


# Review: Case / Control Phenotypes II

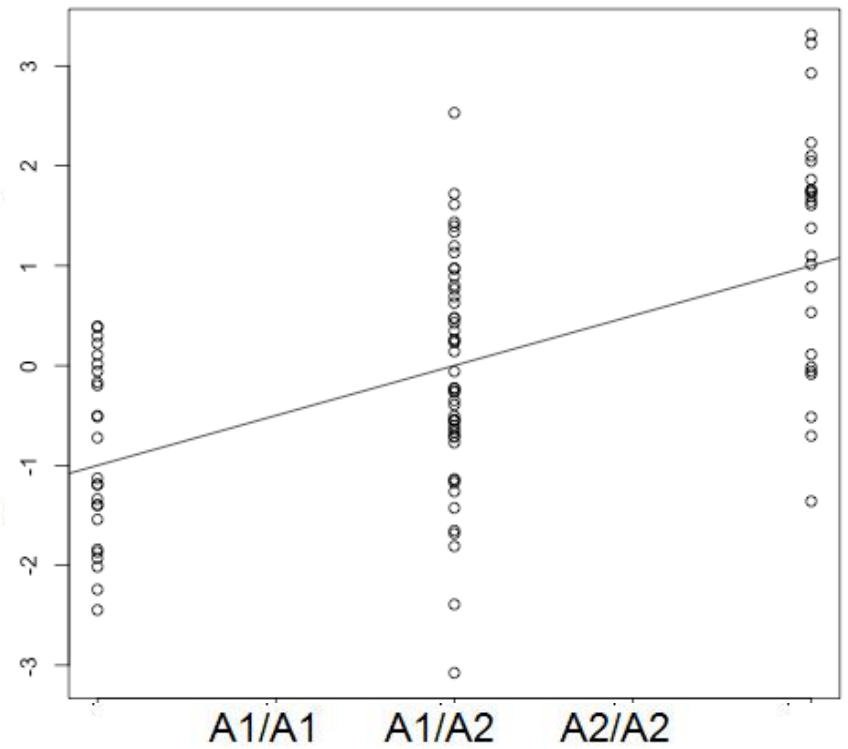
- Let's contrast the situation, let's contrast data we might model with a linear regression model versus case / control data:

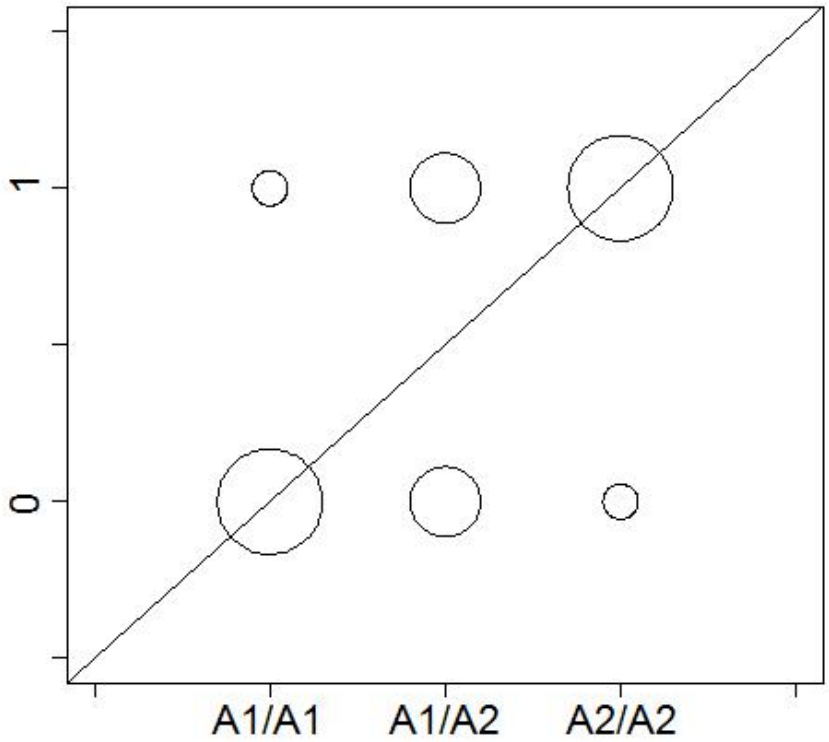


# Review: Case / Control Phenotypes II

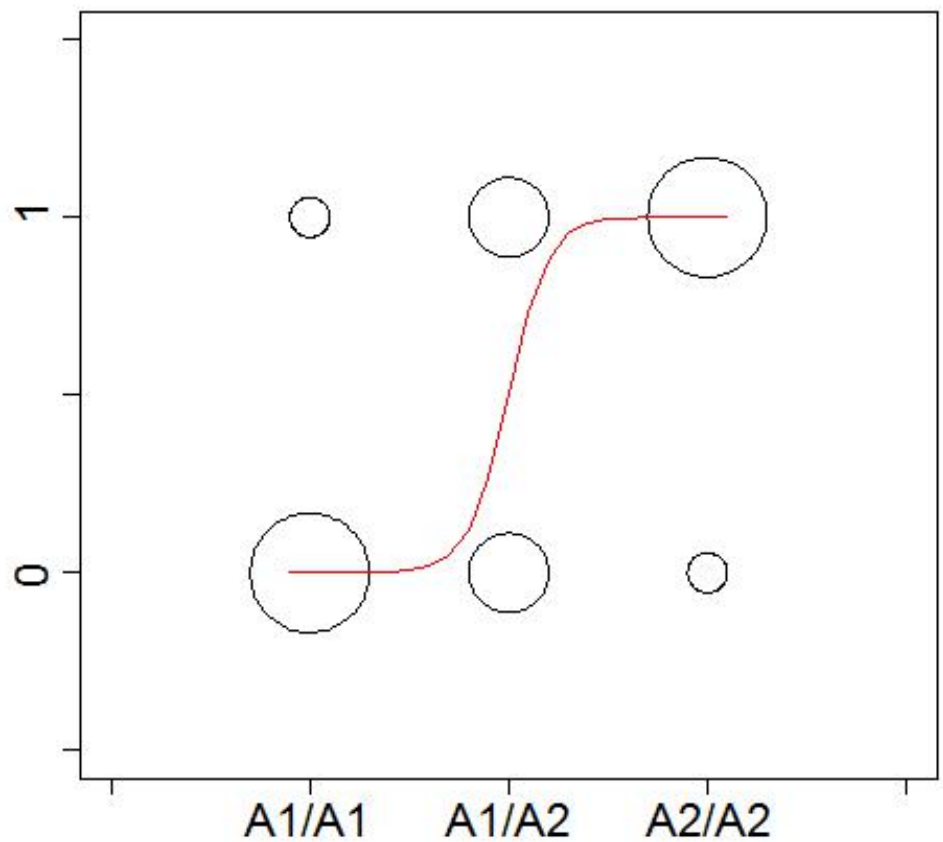
- Let's contrast the situation, let's contrast data we might model with a linear regression model versus case / control data:



# Review: Logistic regression 1

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



# Review: 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!

# Review: 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:

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

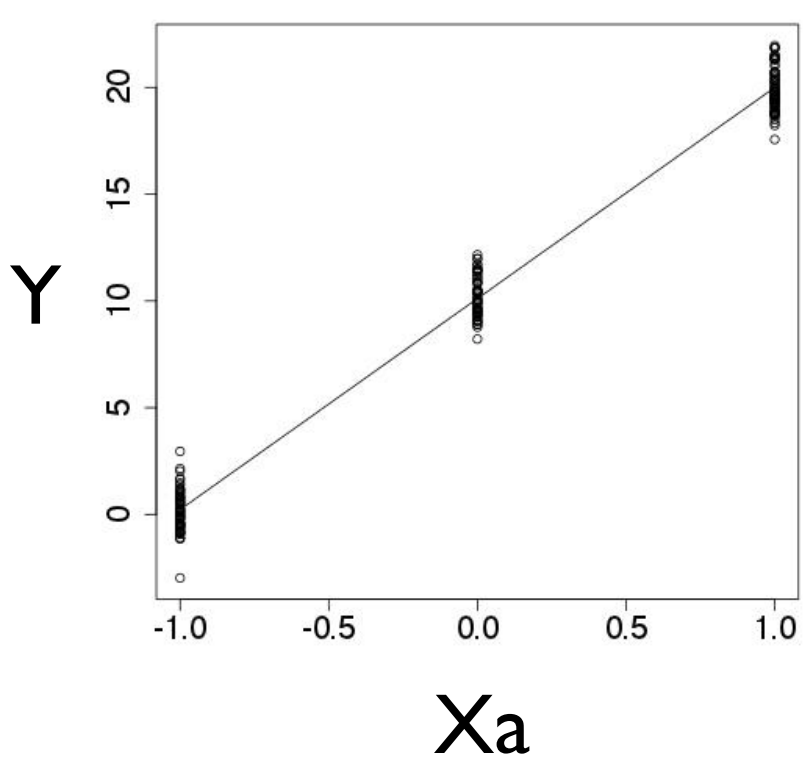
- We can similarly write for a population using matrix notation (where the X matrix has the same form as we have been considering!):

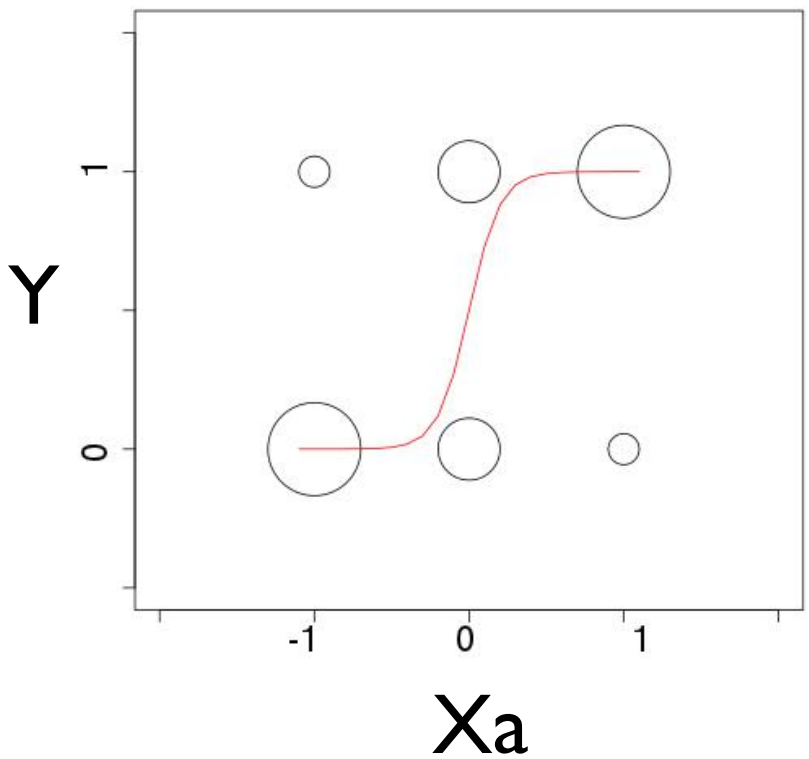
$$\mathrm{E}(\mathbf{Y}|\mathbf{X}) = logistic(\mathbf{X}\beta)$$

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

# Logistic regression: error term 1

- Recall that for a linear regression, the error term accounted for the difference between each point and the expected value (the linear regression line), which we assume follow a normal, but for a logistic regression, we have the same case but the value has to make up the value to either 0 or 1 (what distribution is this?):



# Logistic regression: error term II

- For the error on an individual i, we therefore have to construct an error that takes either the value of "1" or "0" depending on the value of the expected value of the genotype
  - For Y = 0

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

  - For Y = 1

$$\epsilon_i = 1 - E(Y_i|X_i) = 1 - E(Y|A_iA_j) = 1 - logistic(\beta_\mu + X_{i,a}\beta_a + X_{i,d}\beta_d)$$

- For a distribution that takes two such values, a reasonable distribution is therefore the Bernoulli distribution with the following parameter

$$\epsilon_i = Z - E(Y_i|X_i)$$

$$Pr(Z) \sim bern(p) \qquad p = logistic(\beta_\mu + X_a\beta_a + X_d\beta_d)$$

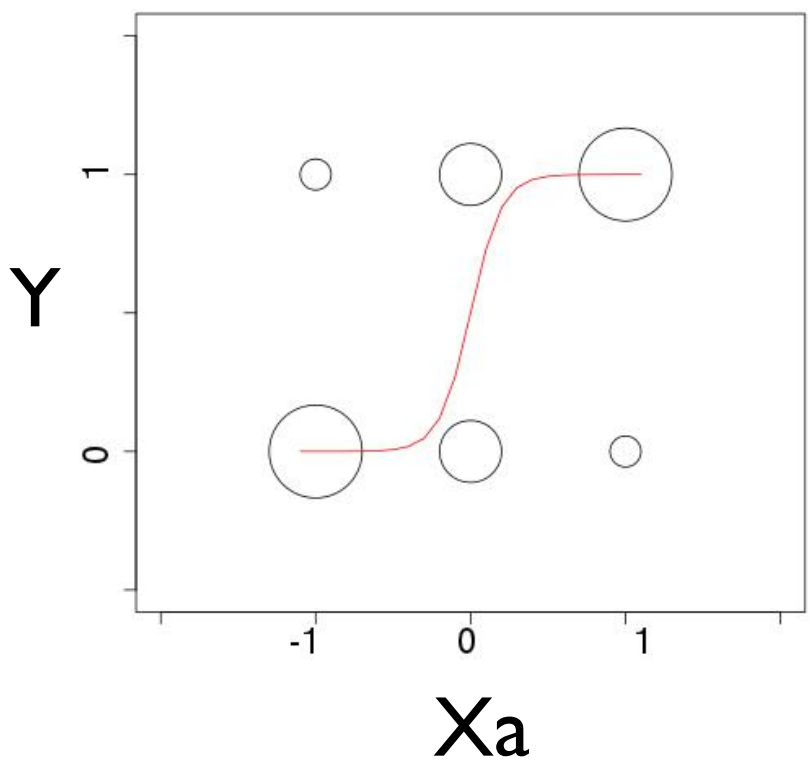
# Logistic regression: error term III

- This may look complicated at first glance but the intuition is relatively simple
- If the logistic regression line is near zero, the probability distribution of the error term is set up to make the probability of Y being zero greater than being one (and vice versa for the regression line near one!):

$$\epsilon_i = Z - E(Y_i|X_i)$$

$$Pr(Z) \sim bern(p)$$

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



# The error term I

- Recall that the error term is either the negative of E(Yi | Xi) when Yi is zero and 1- E(Yi | Xi) when Yi is one:

$$\epsilon_i|(Y_i = 0) = -\mathrm{E}(Y_i|X_i) \qquad \epsilon_i|(Y_i = 1) = 1 - \mathrm{E}(Y_i|X_i)$$

- For the entire distribution of the population, recall that

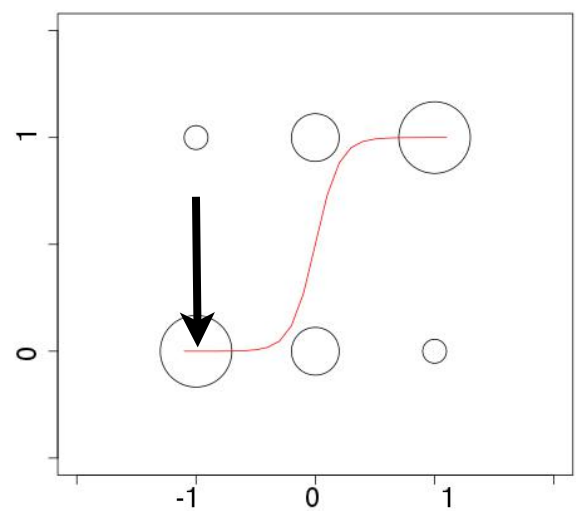
$$Pr(\epsilon_i) \sim bern(p|X) - \mathrm{E}(Y|X)$$

$$p = \mathrm{E}(Y|X)$$

For example:

$$\epsilon_i = -0.1 \qquad \epsilon_i = 0.9$$

$$p = 0.1$$



# The error term II

- Recall that the error term is either the negative of E(Yi | Xi) when Yi is zero and 1- E(Yi | Xi) when Yi is one:

$$\epsilon_i|(Y_i = 0) = -\mathrm{E}(Y_i|X_i) \qquad \epsilon_i|(Y_i = 1) = 1 - \mathrm{E}(Y_i|X_i)$$

- For the entire distribution of the population, recall that

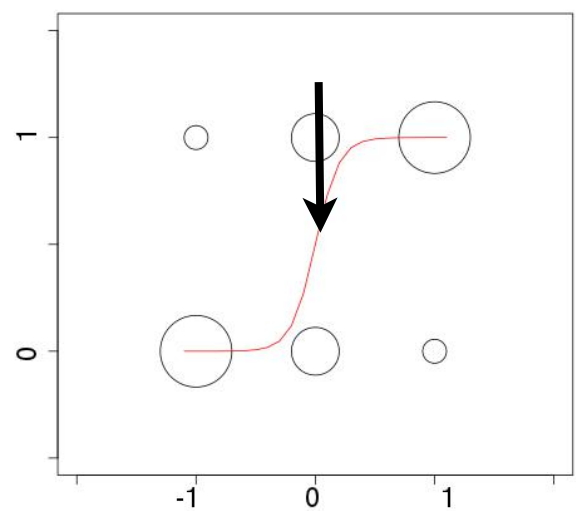
$$Pr(\epsilon_i) \sim bern(p|X) - \mathrm{E}(Y|X)$$

$$p = \mathrm{E}(Y|X)$$

For example:

$$\epsilon_i = -0.6 \qquad \epsilon_i = 0.4$$

$$p = 0.6$$



# The error term III

- Recall that the error term is either the negative of E(Yi | Xi) when Yi is zero and 1- E(Yi | Xi) when Yi is one:

$$\epsilon_i|(Y_i = 0) = -\mathrm{E}(Y_i|X_i) \qquad \epsilon_i|(Y_i = 1) = 1 - \mathrm{E}(Y_i|X_i)$$

- For the entire distribution of the population, recall that

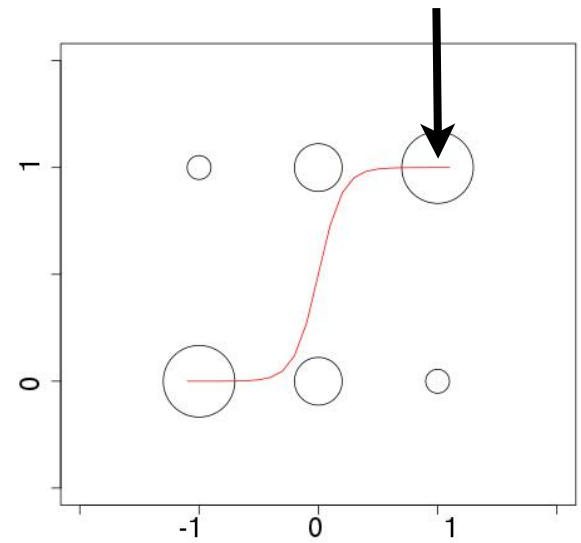
$$Pr(\epsilon_i) \sim bern(p|X) - \mathrm{E}(Y|X)$$

$$p = \mathrm{E}(Y|X)$$

For example:

$$\epsilon_i = -0.9 \qquad \epsilon_i = 0.1$$

$$p = 0.9$$

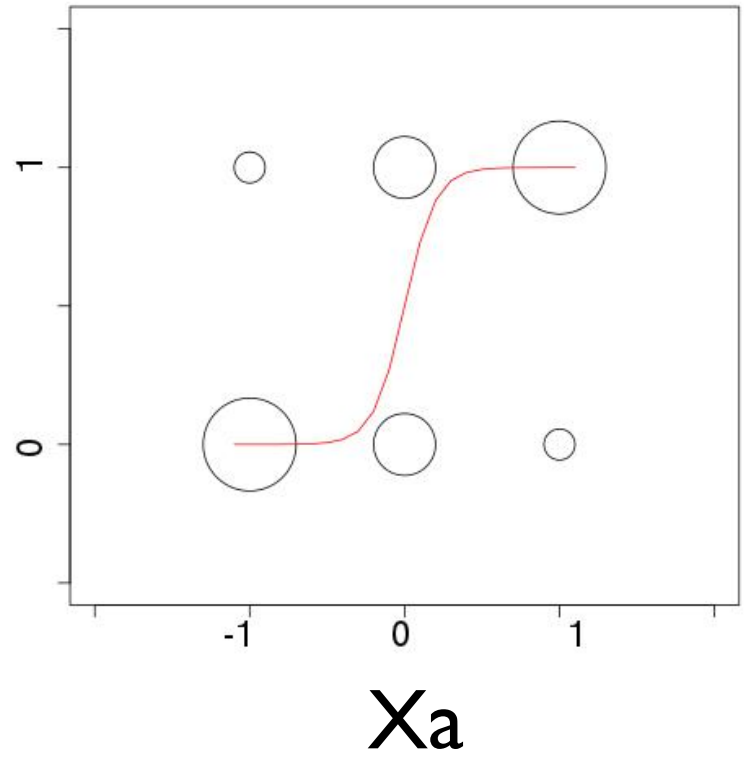


# Logistic regression: link function

- Next, we have to consider the function for the regression line of a logistic regression (remember below we are plotting just versus Xa but this really is a plot versus Xa AND Xd!!):

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

$$\mathrm{E}(Y_i|X_i) = \frac{e^{\beta_\mu + X_{i,a}\beta_a + X_{i,d}\beta_d}}{1 + e^{\beta_\mu + X_{i,a}\beta_a + X_{i,d}\beta_d}}$$



# Calculating the components of an individual I

- For example, say we have an individual i that has genotype A1A1 and phenotype Yi = 0
- We know Xa = -1 and Xd = -1
- Say we also know that for the population, the true parameters (which we will not know in practice! We need to infer them!) are:

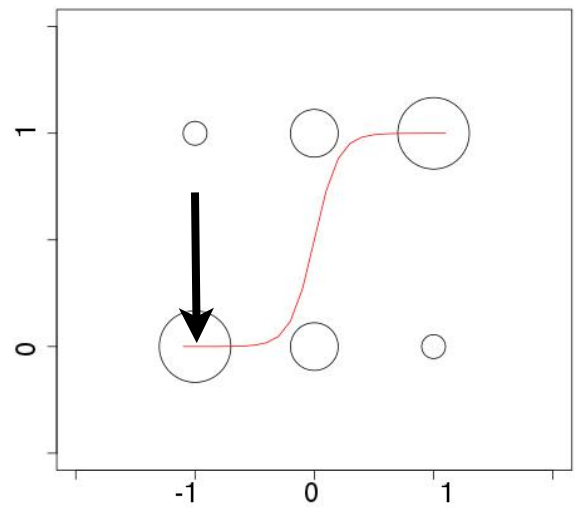
$$\beta_\mu = 0.2 \quad \beta_a = 2.2 \quad \beta_d = 0.2$$

- We can then calculate the E(Yi|Xi) and the error term for i:

$$Y_i = \frac{e^{\beta_\mu + x_{i,a}\beta_a + x_{i,d}\beta_d}}{1 + e^{\beta_\mu + x_{i,a}\beta_a + x_{i,d}\beta_d}} + \epsilon_i$$

$$0 = \frac{e^{0.2 + (-1)2.2 + (-1)0.2}}{1 + e^{0.2 + (-1)2.2 + (-1)0.2}} + \epsilon_i$$

$$0 = 0.1 - 0.1$$



# Calculating the components of an individual II

- For example, say we have an individual i that has genotype A1A1 and phenotype Yi = 1
- We know Xa = -1 and Xd = -1
- Say we also know that for the population, the true parameters (which we will not know in practice! We need to infer them!) are:

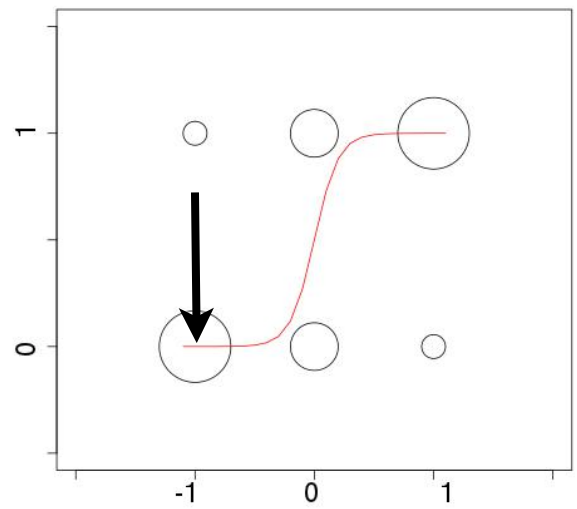
$$\beta_\mu = 0.2 \quad \beta_a = 2.2 \quad \beta_d = 0.2$$

- We can then calculate the E(Yi|Xi) and the error term for i:

$$Y_i = \frac{e^{\beta_\mu + x_{i,a}\beta_a + x_{i,d}\beta_d}}{1 + e^{\beta_\mu + x_{i,a}\beta_a + x_{i,d}\beta_d}} + \epsilon_i$$

$$1 = \frac{e^{0.2+(-1)2.2+(-1)0.2}}{1 + e^{0.2+(-1)2.2+(-1)0.2}} + \epsilon_i$$

$$1 = 0.1 + 0.9$$



# Calculating the components of an individual III

- For example, say we have an individual i that has genotype A1A2 and phenotype Yi = 0
- We know Xa = 0 and Xd = 1
- Say we also know that for the population, the true parameters (which we will not know in practice! We need to infer them!) are:

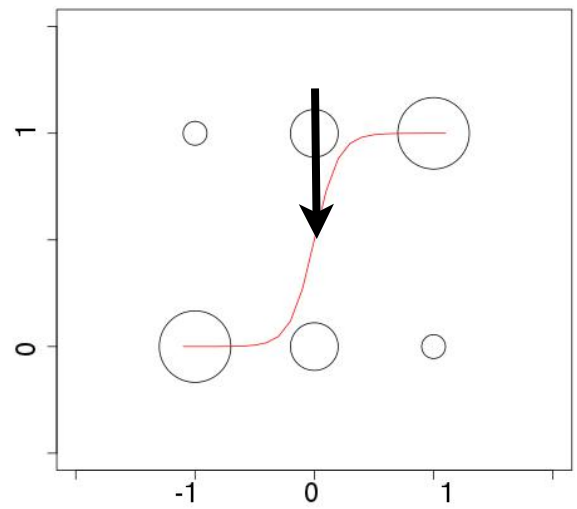
$$\beta_\mu = 0.2 \quad \beta_a = 2.2 \quad \beta_d = 0.2$$

- We can then calculate the E(Yi|Xi) and the error term for i:

$$Y_i = \frac{e^{\beta_\mu + x_{i,a}\beta_a + x_{i,d}\beta_d}}{1 + e^{\beta_\mu + x_{i,a}\beta_a + x_{i,d}\beta_d}} + \epsilon_i$$

$$0 = \frac{e^{0.2+(0)2.2+(1)0.2}}{1 + e^{0.2+(0)2.2+(1)0.2}} + \epsilon_i$$

$$0 = 0.6 - 0.6$$



# Calculating the components of an individual IV

- For example, say we have an individual i that has genotype A2A2 and phenotype Yi = 0
- We know Xa = 1 and Xd = -1
- Say we also know that for the population, the true parameters (which we will not know in practice! We need to infer them!) are:

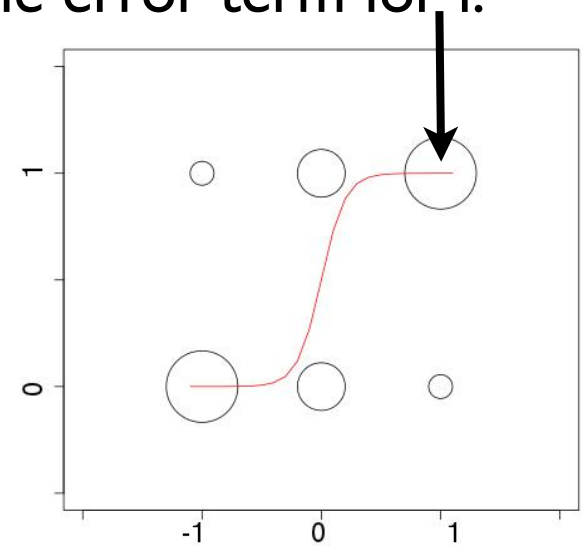
$$\beta_\mu = 0.2 \quad \beta_a = 2.2 \quad \beta_d = 0.2$$

- We can then calculate the E(Yi|Xi) and the error term for i:

$$Y_i = \frac{e^{\beta_\mu + x_{i,a}\beta_a + x_{i,d}\beta_d}}{1 + e^{\beta_\mu + x_{i,a}\beta_a + x_{i,d}\beta_d}} + \epsilon_i$$

$$0 = \frac{e^{0.2+(1)2.2+(-1)0.2}}{1 + e^{0.2+(1)2.2+(-1)0.2}} + \epsilon_i$$

$$0 = 0.9 - 0.9$$



# Notation

- Remember that while we are plotting this versus just Xa, the true plot is versus BOTH Xa and Xd (harder to see what is going on)
- For an entire sample, we can use matrix notation as follows:

$$\mathrm{E}(\mathbf{Y}|\mathbf{X}) = \gamma^{-1}(\mathbf{X}\beta) = \frac{e^{\mathbf{X}\beta}}{1+e^{\mathbf{X}\beta}} = \frac{1}{1+e^{-\mathbf{X}\beta}}$$

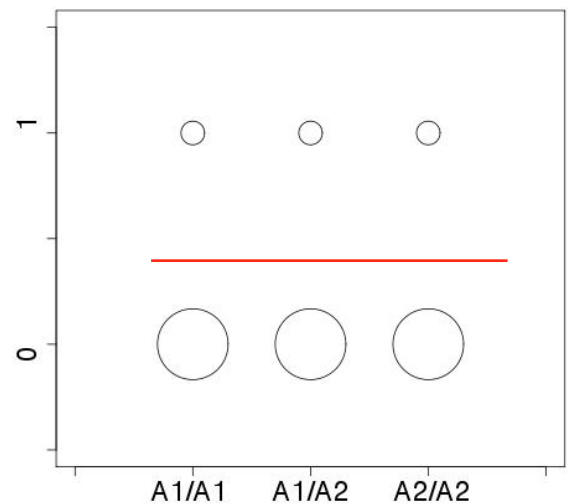
$$\mathrm{E}(\mathbf{y}|\mathbf{x}) = \gamma^{-1}(\mathbf{x}\beta) = \begin{bmatrix} \frac{e^{\beta\mu+x_{1,a}\beta_a+x_{1,d}\beta_d}}{1+e^{\beta\mu+x_{1,a}\beta_a+x_{1,d}\beta_d}} \\ \vdots \\ \frac{e^{\beta\mu+x_{n,a}\beta_a+x_{n,d}\beta_d}}{1+e^{\beta\mu+x_{n,a}\beta_a+x_{n,d}\beta_d}} \end{bmatrix}$$

# Inference

- Recall that our goal with using logistic regression was to model the probability distribution of a case / control phenotype when there is a causal polymorphism
- To use this for a GWAS, we need to test the null hypothesis that a genotype is not a causal polymorphism (or more accurately that the genetic marker we are testing is not in LD with a causal polymorphism!):

$$\beta_\mu = c \quad \beta_a = 0 \quad \beta_d = 0$$

$$H_0 : \beta_a = 0 \cap \beta_d = 0$$




- To assess this null hypothesis, we will use the same approach as in linear regression, i.e. we will construct a LRT = likelihood ratio test (recall that an F-test is an LRT!)
- We will need MLE for the parameters of the logistic regression for the LRT

# MLE of logistic regression parameters

- Recall that an MLE is simply a statistic (a function that takes the sample as an input and outputs the estimate of the parameters)!
- In this case, we want to construct the following MLE:

$$MLE(\hat{\beta}) = MLE(\hat{\beta}_\mu, \hat{\beta}_a, \hat{\beta}_d)$$

- To do this, we need to maximize the log-likelihood function for the logistic regression, which has the following form (sample size n):

$$l(\beta) = \sum_{i=1}^{n} [y_i ln(\gamma^{-1}(\beta_\mu + x_{i,a}\beta_a + x_{i,d}\beta_d)) + (1 - y_i) ln(1 - \gamma^{-1}(\beta_\mu + x_{i,a}\beta_a + x_{i,d}\beta_d))$$

- Unlike the case of linear regression, where we had a "closed-form" equation that allows us to plug in the Y's and X's and returns the beta values that maximize the log-likelihood, there is no such simple equation for a logistic regression
- We will therefore need an *algorithm* to calculate the MLE

# Algorithm Basics

- **algorithm** - a sequence of instructions for taking an input and producing an output
- We often use algorithms in estimation of parameters where the structure of the estimation equation (e.g., the log-likelihood) is so complicated that we cannot
  - Derive a simple (closed) form equation for the estimator
  - Cannot easily determine the value the estimator should take by other means (e.g., by graphical visualization)
- We will use algorithms to "search" for the parameter values that correspond to the estimator of interest
- Algorithms are not guaranteed to produce the correct value of the estimator (!!), because the algorithm may "converge" (=return) the wrong answer (e.g., converges to a "local" maximum or does not converge!) and because the compute time to converge to exactly the same answer is impractical for applications

# IRLS algorithm I

- For logistic regression (and GLM's in general!) we will construct an algorithm to find the parameters that correspond to the maximum of the log-likelihood:

$$l(\beta) = \sum_{i=1}^{n} \left[ y_i ln(\gamma^{-1}(\beta_\mu + x_{i,a}\beta_a + x_{i,d}\beta_d)) + (1 - y_i) ln(1 - \gamma^{-1}(\beta_\mu + x_{i,a}\beta_a + x_{i,d}\beta_d)) \right.$$

- For logistic regression (and GLM's in general!) we will construct an Iterative Re-weighted Least Squares (IRLS) algorithm, which has the following structure:

1. Choose starting values for the $\beta$'s. Since we have a vector of three $\beta$'s in our case, we assign these numbers and call the resulting vector $\beta^{[0]}$.
2. Using the re-weighting equation (described next slide), update the $\beta^{[t]}$ vector.
3. At each step $t > 0$ check if $\beta^{[t+1]} \approx \beta^{[t]}$ (i.e. if these are approximately equal) using an appropriate function. If the value is below a defined threshold, stop. If not, repeat steps 2,3.

# Step 1: IRLS algorithm

1. Choose starting values for the $\beta$'s. Since we have a vector of three $\beta$'s in our case, we assign these numbers and call the resulting vector $\beta^{[0]}$.

- These are simply values of the vector that we assign (!!)
- In one sense, these can be anything we want (!!) although for algorithms in general there are usually some restrictions and / or certain starting values that are "better" than others in the sense that the algorithm will converge faster, find a more "optimal" solution etc.
- In our case, we can assign our starting values as follows:

$$\beta^{[0]} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

# Step 2: IRLS algorithm

2. Using the re-weighting equation (described next slide), update the $\beta^{[t]}$ vector.

- At step 2, we will update (= produce a new value of the vector) using the following equation (then do this again and again until we stop!):

$$\beta^{[t+1]} = \beta^{[t]} + [\mathbf{x}^{\mathrm{T}}\mathbf{W}\mathbf{x}]^{-1}\mathbf{x}^{\mathrm{T}}(\mathbf{y} - \gamma^{-1}(\mathbf{x}\beta^{[t]})$$

$$\mathbf{x} = \begin{bmatrix} 1 & x_{1,a} & x_{1,d} \\ 1 & x_{2,a} & x_{2,d} \\ \vdots & \vdots & \ddots \\ 1 & x_{n,a} & x_{n,d} \end{bmatrix} \qquad \gamma^{-1}(\beta_\mu^{[t]} + x_{i,a}\beta_a^{[t]} + x_{i,d}\beta_d^{[t]}) = \frac{e^{\beta_\mu^{[t]} + x_{i,a}\beta_a^{[t]} + x_{i,d}\beta_d^{[t]}}}{1 + e^{\beta_\mu^{[t]} + x_{i,a}\beta_a^{[t]} + x_{i,d}\beta_d^{[t]}}}$$

$$\gamma^{-1}(\mathbf{x}\beta^{[t]}) = \frac{e^{\mathbf{x}\beta^{[t]}}}{1 + e^{\mathbf{x}\beta^{[t]}}}$$

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad \beta^{[t]} = \begin{bmatrix} \beta_\mu^{[t]} \\ \beta_a^{[t]} \\ \beta_d^{[t]} \end{bmatrix}$$

$$W_{ii} = \gamma^{-1}(\beta_\mu^{[t]} + x_{i,a}\beta_a^{[t]} + x_{i,d}\beta_d^{[t]})(1 - \gamma^{-1}(\beta_\mu^{[t]} + x_{i,a}\beta_a^{[t]} + x_{i,d}\beta_d^{[t]}))$$

$$W_{ii} = \frac{e^{\beta_\mu^{[t]} + x_{i,a}\beta_a^{[t]} + x_{i,d}\beta_d^{[t]}}}{1 + e^{\beta_\mu^{[t]} + x_{i,a}\beta_a^{[t]} + x_{i,d}\beta_d^{[t]}}} \left(1 - \frac{e^{\beta_\mu^{[t]} + x_{i,a}\beta_a^{[t]} + x_{i,d}\beta_d^{[t]}}}{1 + e^{\beta_\mu^{[t]} + x_{i,a}\beta_a^{[t]} + x_{i,d}\beta_d^{[t]}}}\right)$$

$$(W_{ij} = 0 \text{ for } i \neq j)$$

# Step 3: IRLS algorithm

3. At each step $t > 0$ check if $\beta^{[t+1]} \approx \beta^{[t]}$ (i.e. if these are approximately equal) using an appropriate function. If the value is below a defined threshold, stop. If not, repeat steps 2,3.

- At step 3, we "check" to see if we should stop the algorithm and, if we decide not to stop, we go back to step 2
- If we decide to stop, we will assume the final values of the vector are the MLE (it may not be exactly the true MLE, but we will assume that it is close if we do not stop the algorithm to early!), e.g. $\beta^{[t+1]} \approx \beta^{[t]}$
- There are many stopping rules, using change in Deviance is one way to construct a rule (note the issue with ln(0)!!:

$$\triangle D = |D\,[t+1] - D\,[t]\,| \qquad \triangle D < 10^{-6}$$

$$D = 2\sum_{i=1}^{n}\left[y_i ln\left(\frac{y_i}{\gamma^{-1}(\beta_\mu^{[t]\text{or}[t+1]} + x_{i,a}\beta_a^{[t]\text{or}[t+1]} + x_{i,d}\beta_d^{[t]\text{or}[t+1]})}\right) + (1-y_i) ln\left(\frac{1-y_i}{1-\gamma^{-1}(\beta_\mu^{[t]\text{or}[t+1]} + x_{i,a}\beta_a^{[t]\text{or}[t+1]} + x_{i,d}\beta_d^{[t]\text{or}[t+1]})}\right)\right]$$

$$D = 2\sum_{i=1}^{n}\left[y_i ln\left(\frac{y_i}{\frac{e^{\beta_\mu^{[t]\text{or}[t+1]} + x_{i,a}\beta_a^{[t]\text{or}[t+1]} + x_{i,d}\beta_d^{[t]\text{or}[t+1]}}}{1+e^{\beta_\mu^{[t]\text{or}[t+1]} + x_{i,a}\beta_a^{[t]\text{or}[t+1]} + x_{i,d}\beta_d^{[t]\text{or}[t+1]}}}}\right) + (1-y_i) ln\left(\frac{1-y_i}{1-\frac{e^{\beta_\mu^{[t]\text{or}[t+1]} + x_{i,a}\beta_a^{[t]\text{or}[t+1]} + x_{i,d}\beta_d^{[t]\text{or}[t+1]}}}{1+e^{\beta_\mu^{[t]\text{or}[t+1]} + x_{i,a}\beta_a^{[t]\text{or}[t+1]} + x_{i,d}\beta_d^{[t]\text{or}[t+1]}}}}\right)\right]$$

# That's it for today

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