

Quantitative Genomics and Genetics

BTRY 4830/6830; PBSB.5201.03

Lecture 23: Intro to Logistic Regression III

Jason Mezey

April 18, 2024 (Th) 8:40-9:55

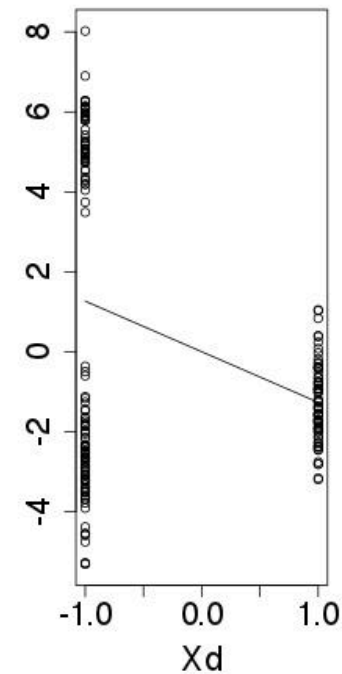
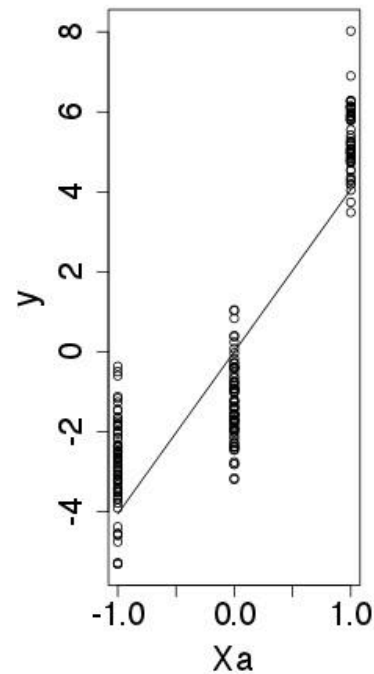
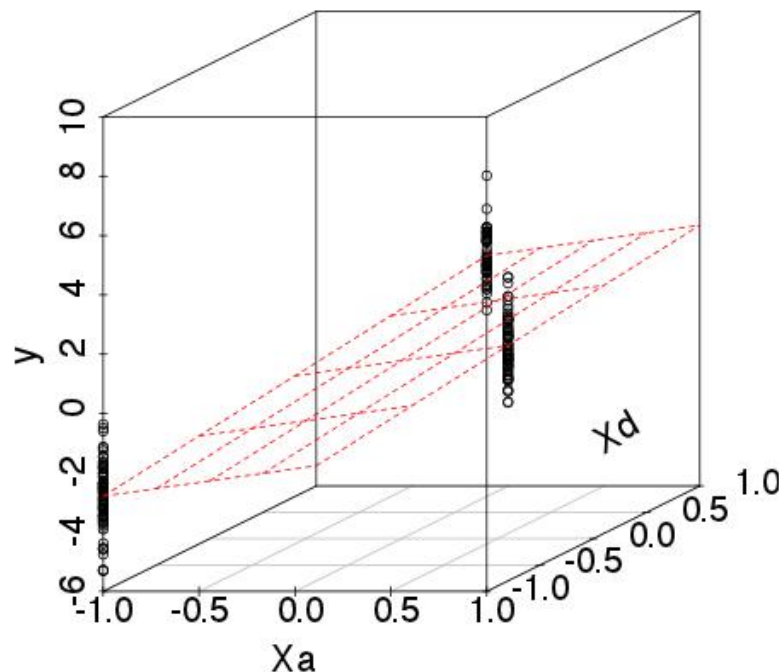
Summary of lecture 23: Logistic Regression II

- 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):

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



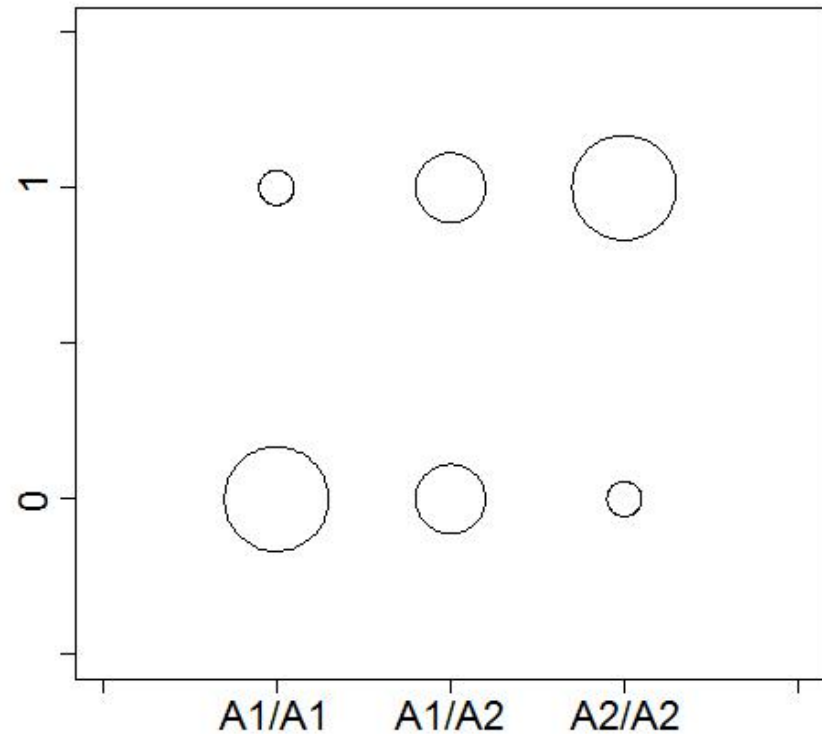
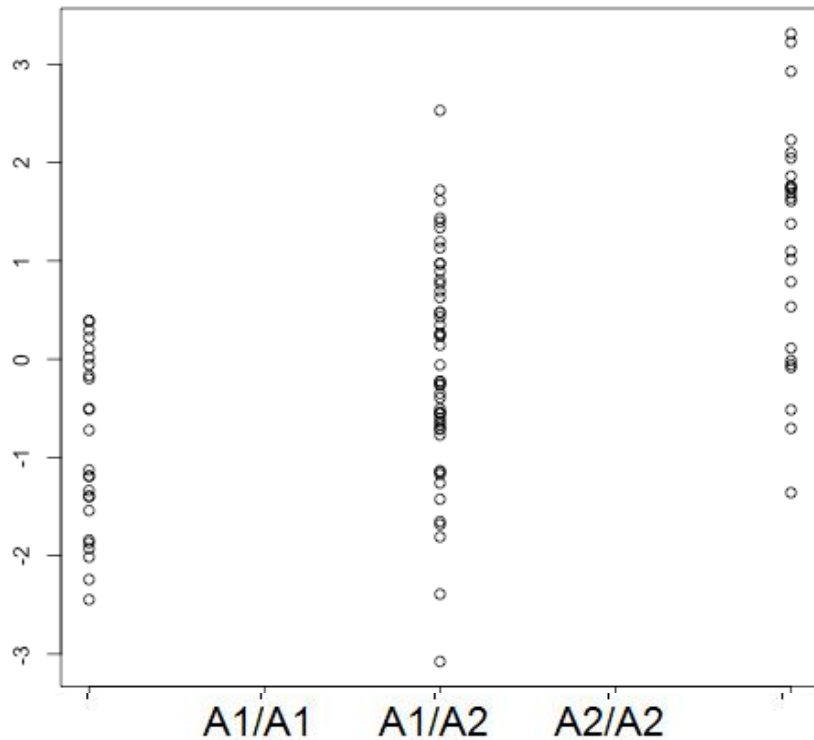
Review: Case / Control

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

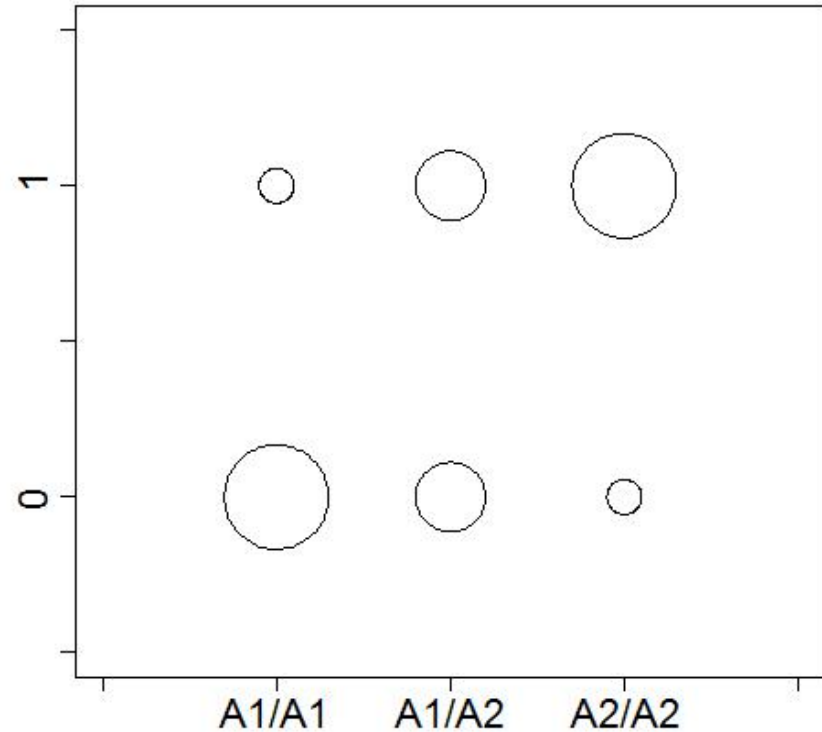
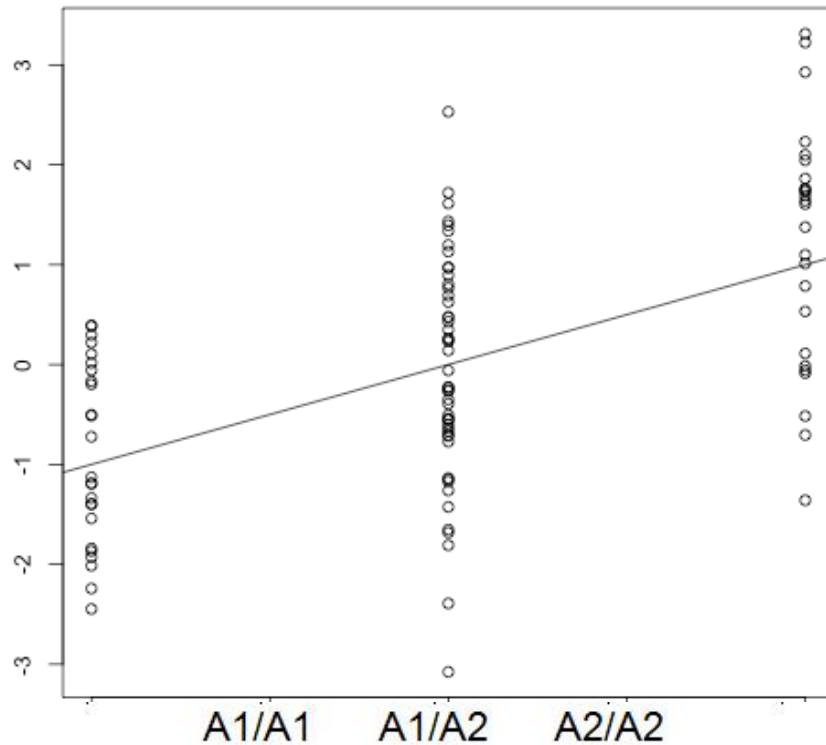
- Let's contrast the situation, let's contrast data we might model with a linear regression model versus case / control data:



Review: Case / Control

Phenotypes II

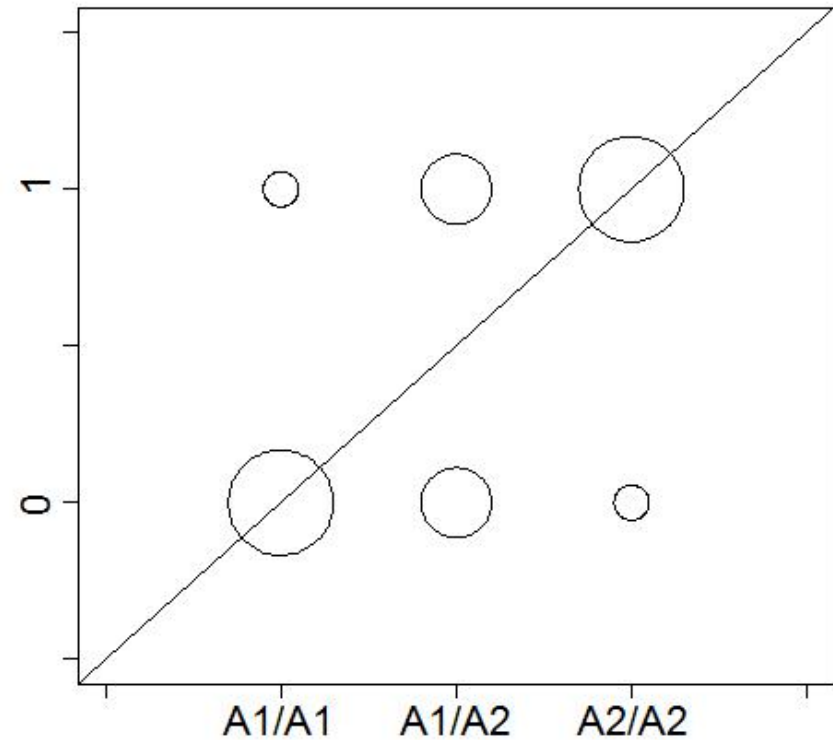
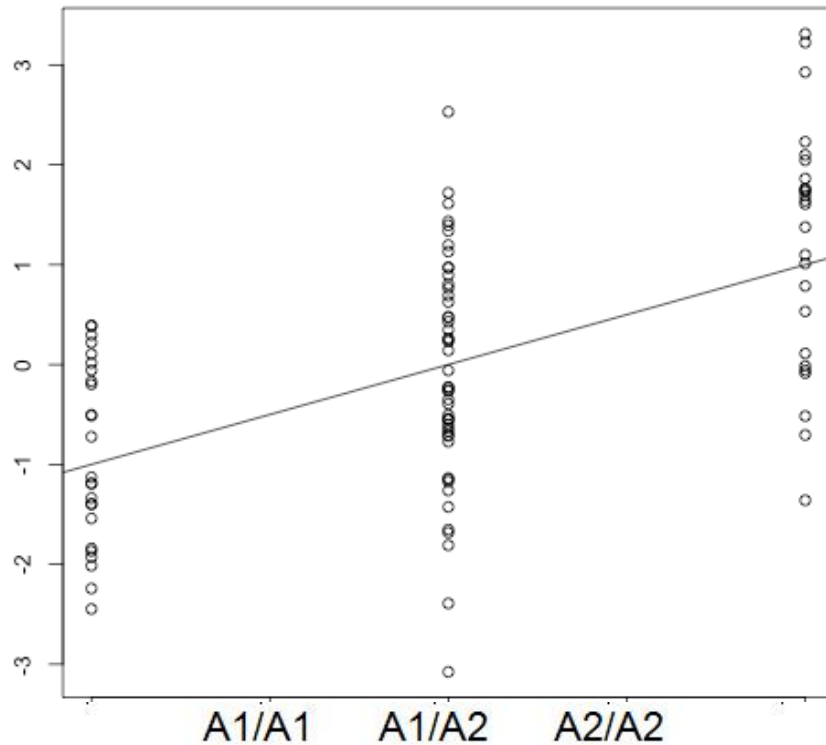
- Let's contrast the situation, let's contrast data we might model with a linear regression model versus case / control data:



Review: Case / Control

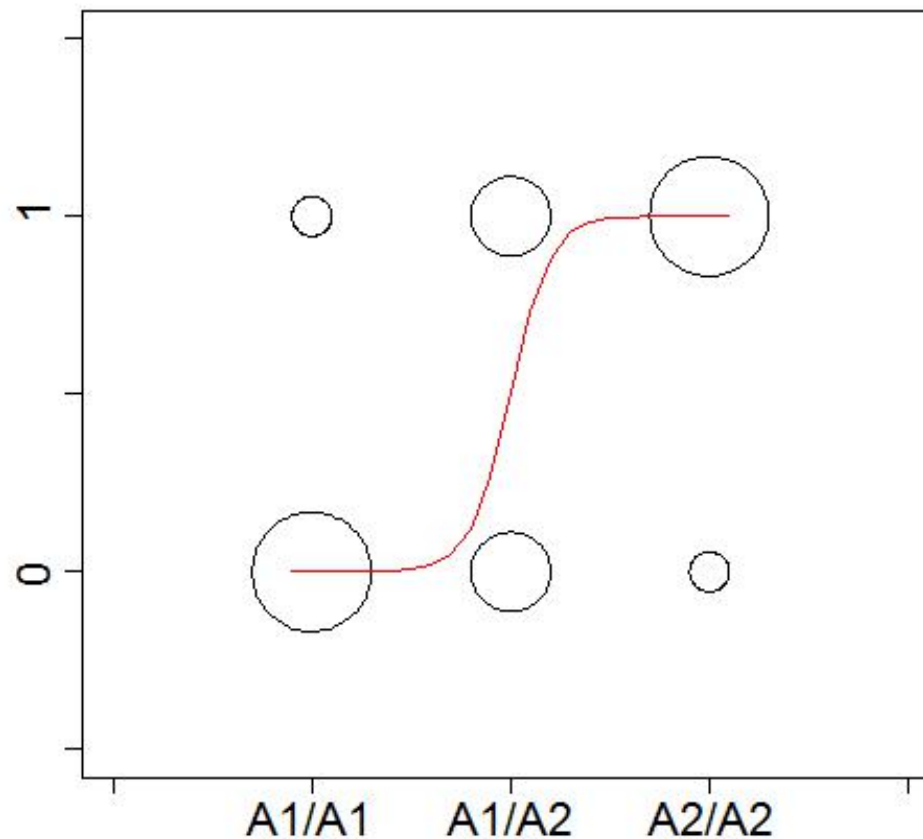
Phenotypes II

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



Review: Logistic regression I

- 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 our approach to inference is the same with this model!

Logistic regression III

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

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

- 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:

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

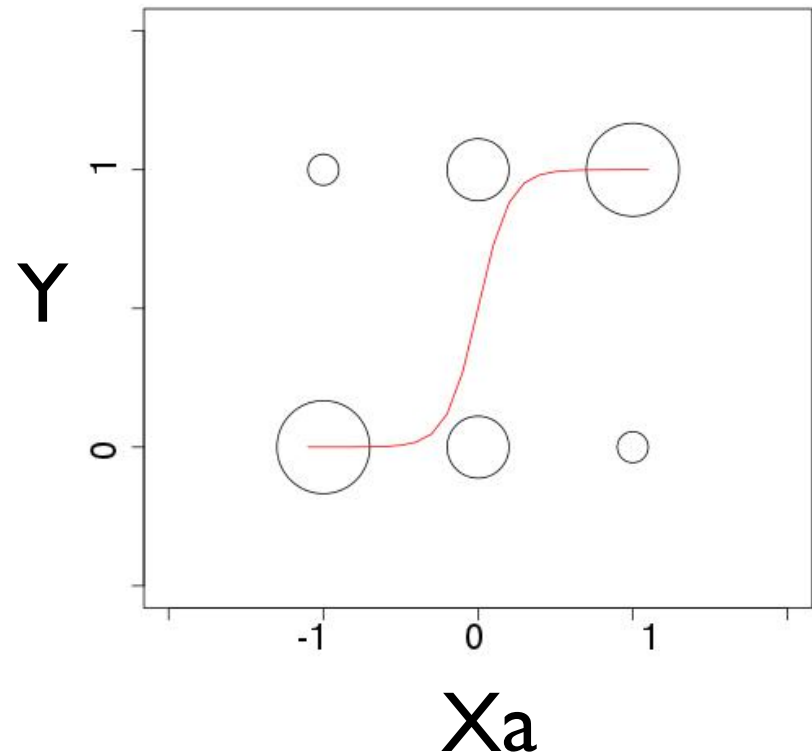
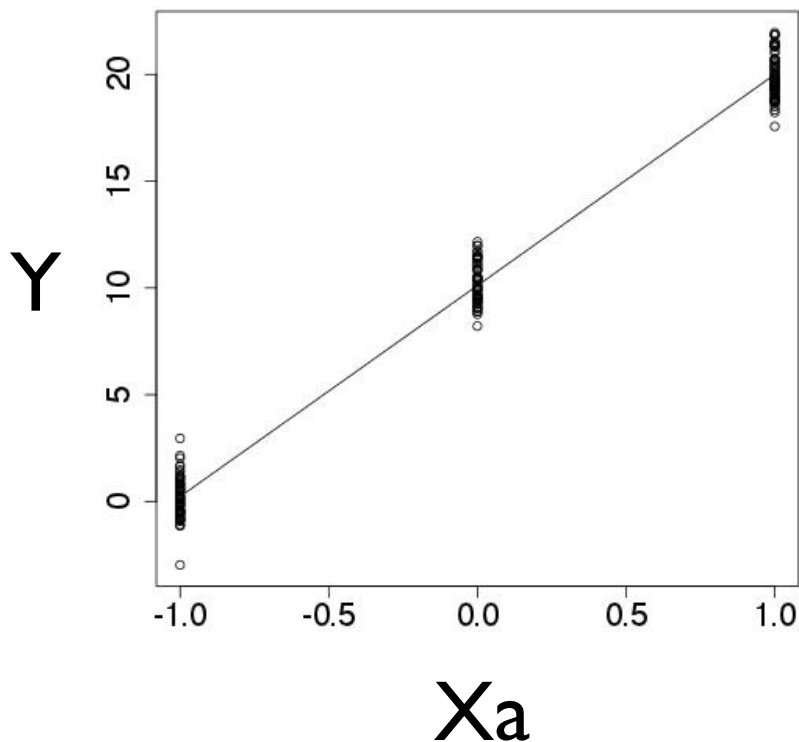
- We can similarly write for a population using matrix notation (where the X matrix has the same form as we have been considering!):

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

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

Logistic regression: error term I

- 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 $Y = 0$

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

- For $Y = 1$

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

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

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

$$\text{Pr}(Z) \sim \text{bern}(p) \quad p = \text{logistic}(\beta_\mu + X_a\beta_a + X_d\beta_d)$$

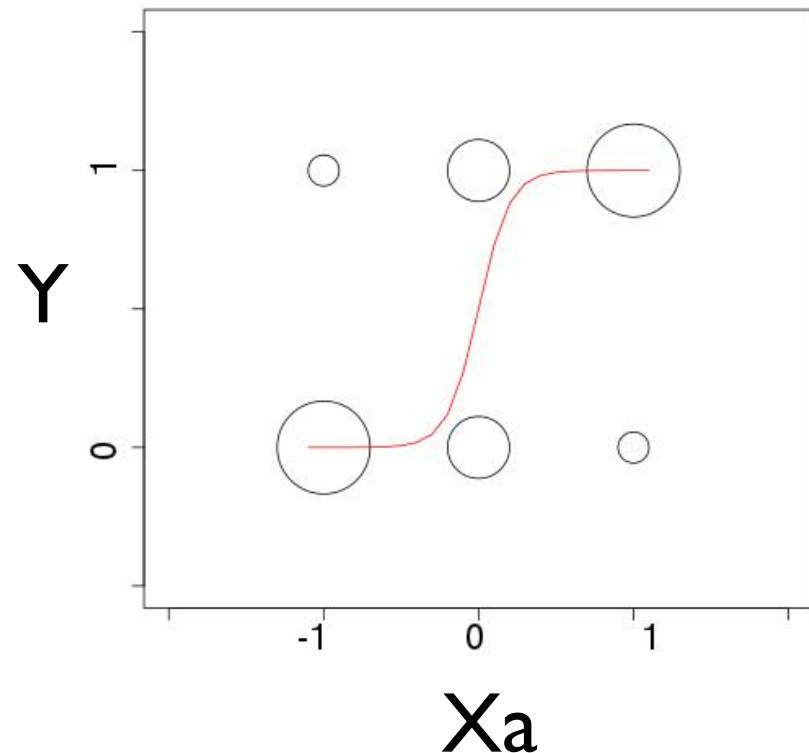
Logistic regression: error term III

- 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!):

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

$$Pr(Z) \sim \text{bern}(p)$$

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



The error term I

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

$$\epsilon_i | (Y_i = 0) = -E(Y_i | X_i) \quad \epsilon_i | (Y_i = 1) = 1 - E(Y_i | X_i)$$

- For the entire distribution of the population, recall that

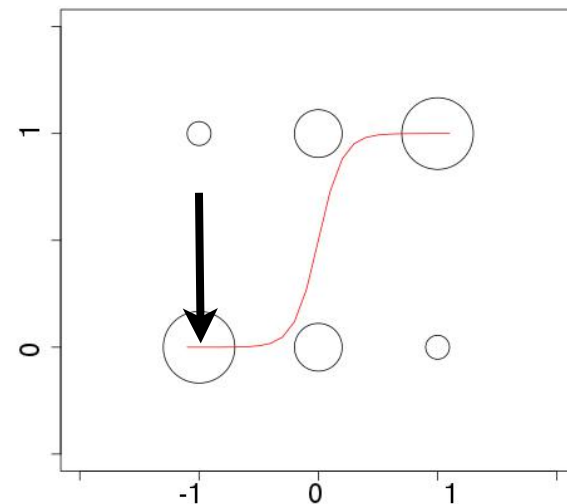
$$Pr(\epsilon_i) \sim \text{bern}(p | X) - E(Y | X)$$

$$p = E(Y | X)$$

For example:

$$\epsilon_i = -0.1 \quad \epsilon_i = 0.9$$

$$p = 0.1$$



The error term II

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

$$\epsilon_i | (Y_i = 0) = -E(Y_i | X_i) \quad \epsilon_i | (Y_i = 1) = 1 - E(Y_i | X_i)$$

- For the entire distribution of the population, recall that

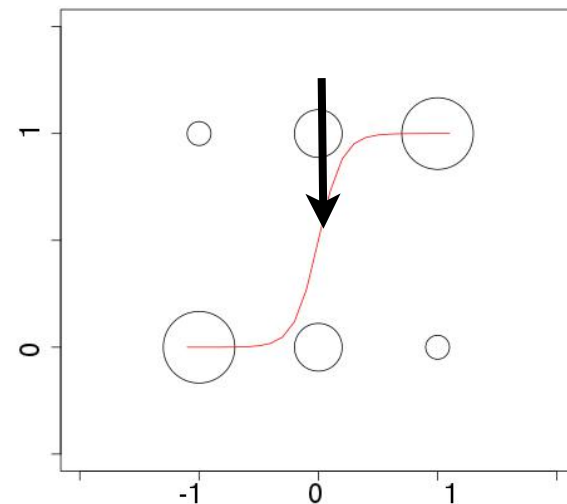
$$Pr(\epsilon_i) \sim \text{bern}(p | X) - E(Y | X)$$

$$p = E(Y | X)$$

For example:

$$\epsilon_i = -0.6 \quad \epsilon_i = 0.4$$

$$p = 0.6$$



The error term III

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

$$\epsilon_i | (Y_i = 0) = -E(Y_i | X_i) \quad \epsilon_i | (Y_i = 1) = 1 - E(Y_i | X_i)$$

- For the entire distribution of the population, recall that

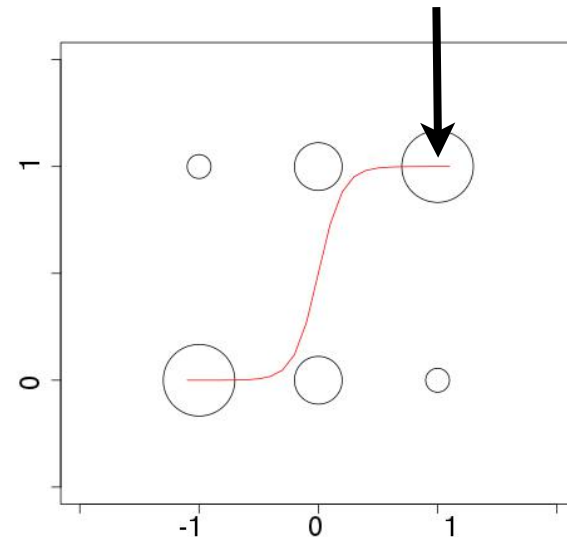
$$Pr(\epsilon_i) \sim \text{bern}(p | X) - E(Y | X)$$

$$p = E(Y | X)$$

For example:

$$\epsilon_i = -0.9 \quad \epsilon_i = 0.1$$

$$p = 0.9$$

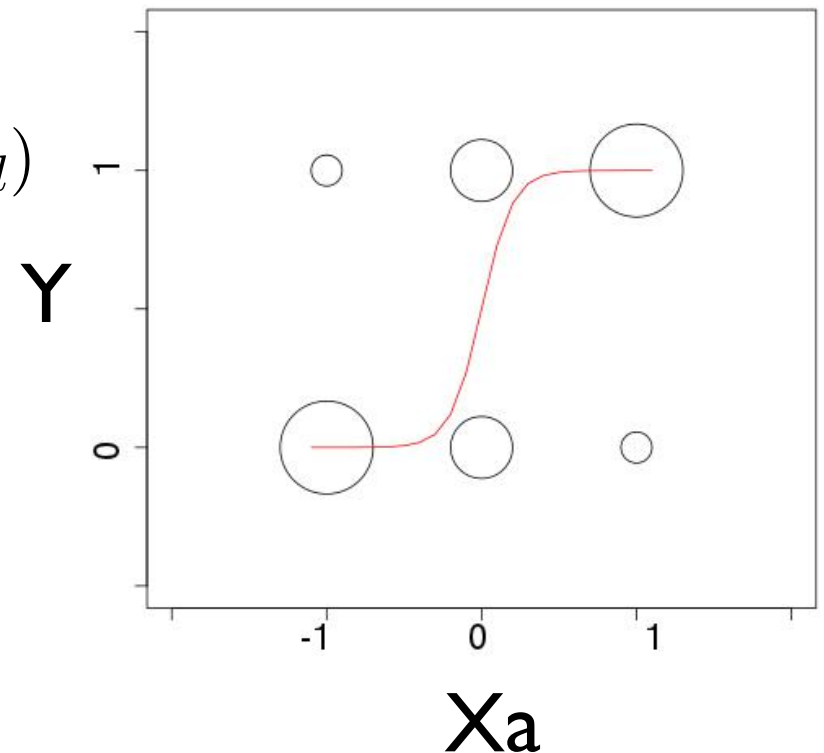


Logistic regression: link function

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

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

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



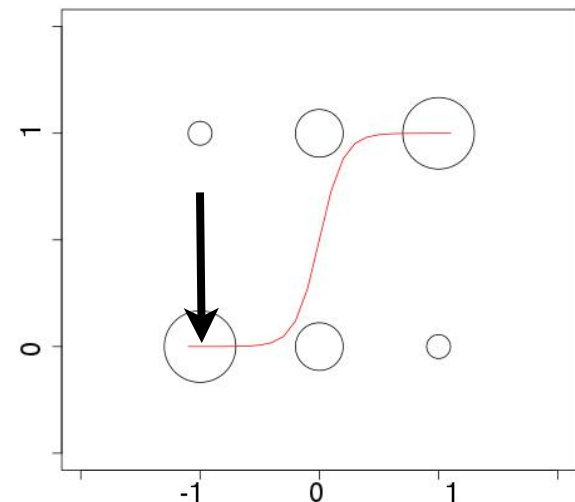
Calculating the components of an individual I

- For example, say we have an individual i that has genotype A|A| and phenotype $Y_i = 0$
- We know $X_a = -1$ and $X_d = -1$
- 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 $E(Y_i|X_i)$ and the error term for i :

$$Y_i = \frac{e^{\beta_\mu + x_{i,a}\beta_a + x_{i,d}\beta_d}}{1 + e^{\beta_\mu + x_{i,a}\beta_a + x_{i,d}\beta_d}} + \epsilon_i$$
$$0 = \frac{e^{0.2 + (-1)2.2 + (-1)0.2}}{1 + e^{0.2 + (-1)2.2 + (-1)0.2}} + \epsilon_i$$
$$0 = 0.1 - 0.1$$



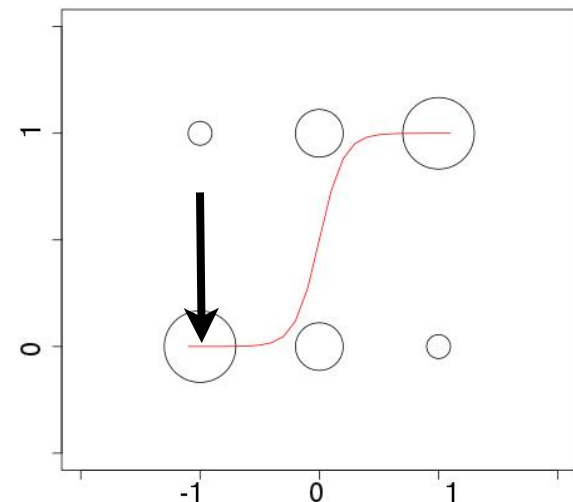
Calculating the components of an individual II

- For example, say we have an individual i that has genotype A|A| and phenotype $Y_i = 1$
- We know $X_a = -1$ and $X_d = -1$
- 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 $E(Y_i|X_i)$ 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 $Y_i = 0$
- We know $X_a = 0$ and $X_d = 1$
- Say we also know that for the population, the true parameters (which we will not know in practice! We need to infer them!) are:

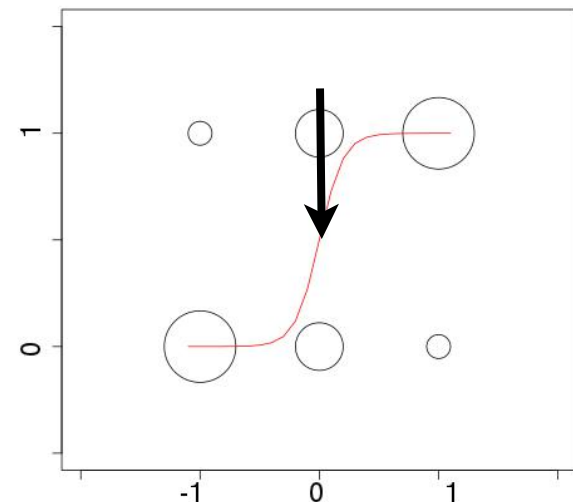
$$\beta_\mu = 0.2 \quad \beta_a = 2.2 \quad \beta_d = 0.2$$

- We can then calculate the $E(Y_i|X_i)$ and the error term for i :

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

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

$$0 = 0.6 - 0.6$$



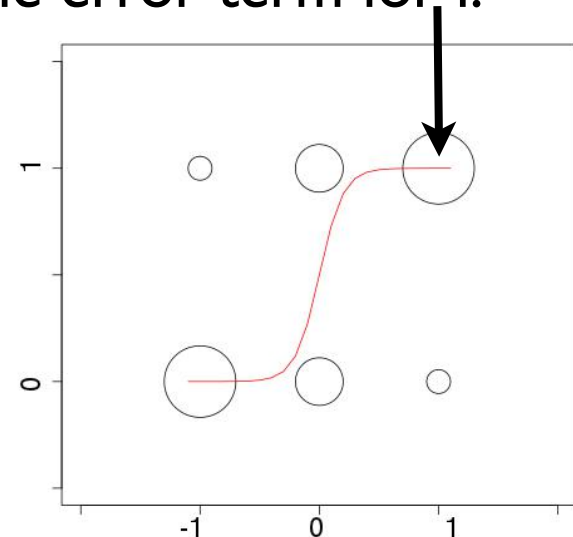
Calculating the components of an individual IV

- For example, say we have an individual i that has genotype A2A2 and phenotype $Y_i = 0$
- We know $X_a = 1$ and $X_d = -1$
- 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 $E(Y_i|X_i)$ 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 (R code)

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

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

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

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:

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

- 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 β 's. Since we have a vector of three β '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 β 's. Since we have a vector of three β '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 (Update Step): 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!):

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

$$\mathbf{x} = \begin{bmatrix} 1 & x_{1,a} & x_{1,d} \\ 1 & x_{2,a} & x_{2,d} \\ \vdots & \vdots & \ddots \\ 1 & x_{n,a} & x_{n,d} \end{bmatrix}$$

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

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

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

$$\beta^{[t]} = \begin{bmatrix} \beta_{\mu}^{[t]} \\ \beta_a^{[t]} \\ \beta_d^{[t]} \end{bmatrix}$$

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

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

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

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 too 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 $\ln(0)$!!):

$$\Delta D = |D[t+1] - D[t]| \quad \Delta D < 10^{-6}$$

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

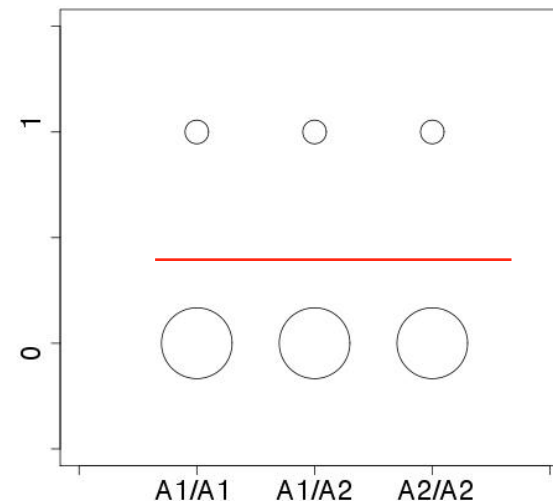
$$D = 2 \sum_{i=1}^n \left[y_i \ln \left(\frac{y_i}{\frac{e^{\beta_\mu^{[t] \text{ or } [t+1]} + x_{i,a} \beta_a^{[t] \text{ or } [t+1]} + x_{i,d} \beta_d^{[t] \text{ or } [t+1]}}}{1 + e^{\beta_\mu^{[t] \text{ or } [t+1]} + x_{i,a} \beta_a^{[t] \text{ or } [t+1]} + x_{i,d} \beta_d^{[t] \text{ or } [t+1]}}}}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

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

Logistic hypothesis testing I

- Recall that our null and alternative hypotheses are:

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

$$H_A : \beta_a \neq 0 \cup \beta_d \neq 0$$

- We will use the LRT for the null (0) and alternative (1):

$$LRT = -2\ln\Lambda = -2\ln\frac{L(\hat{\theta}_0|\mathbf{y})}{L(\hat{\theta}_1|\mathbf{y})} \quad LRT = -2\ln\Lambda = 2l(\hat{\theta}_1|\mathbf{y}) - 2l(\hat{\theta}_0|\mathbf{y})$$

- For our case, we need the following:

$$l(\hat{\theta}_1|\mathbf{y}) = l(\hat{\beta}_\mu, \hat{\beta}_a, \hat{\beta}_d|\mathbf{y})$$

$$l(\hat{\theta}_0|\mathbf{y}) = l(\hat{\beta}_\mu, 0, 0|\mathbf{y})$$

Logistic hypothesis testing II

- For the alternative, we use our MLE estimates of our logistic regression parameters we get from our IRLS algorithm and plug these into the log-likelihood equation

$$l(\hat{\theta}_1 | \mathbf{y}) = \sum_{i=1}^n \left[y_i \ln(\gamma^{-1}(\hat{\beta}_\mu + x_{i,a}\hat{\beta}_a + x_{i,d}\hat{\beta}_d)) + (1 - y_i) \ln(1 - \gamma^{-1}(\hat{\beta}_\mu + x_{i,a}\hat{\beta}_a + x_{i,d}\hat{\beta}_d)) \right]$$
$$\gamma^{-1}(\beta_\mu + x_{i,a}\beta_a + x_{i,d}\beta_d) = \frac{e^{\beta_\mu + x_{i,a}\beta_a + x_{i,d}\beta_d}}{1 + e^{\beta_\mu + x_{i,a}\beta_a + x_{i,d}\beta_d}}$$

- For the null, we plug in the following parameter estimates into this same equation

$$l(\hat{\theta}_0 | \mathbf{y}) = \sum_{i=1}^n \left[y_i \ln(\gamma^{-1}(\hat{\beta}_{\mu,0} + x_{i,a} * 0 + x_{i,d} * 0)) + (1 - y_i) \ln(1 - \gamma^{-1}(\hat{\beta}_{\mu,0} + x_{i,a} * 0 + x_{i,d} * 0)) \right]$$

- where we use the same IRLS algorithm to provide estimates of $\hat{\beta}_{\mu,0}$ by running the algorithm EXACTLY the same with $\hat{\beta}_{\mu,0}$ EXCEPT we set $\hat{\beta}_a = 0, \hat{\beta}_d = 0$ and we do not update these!

Logistic hypothesis testing III

- To calculate our p-value, we need to know the distribution of our LRT statistic under the null hypothesis
- There is no simple form for this distribution for any given n (contrast with F-statistics!!) but we know that as n goes to infinite, we know the distribution is i.e. ($n \rightarrow \infty$):

$$LRT = -2\ln\Lambda = 2l(\hat{\theta}_1|\mathbf{y}) - 2l(\hat{\theta}_0|\mathbf{y})$$

$$LRT \rightarrow \chi_{df}^2$$

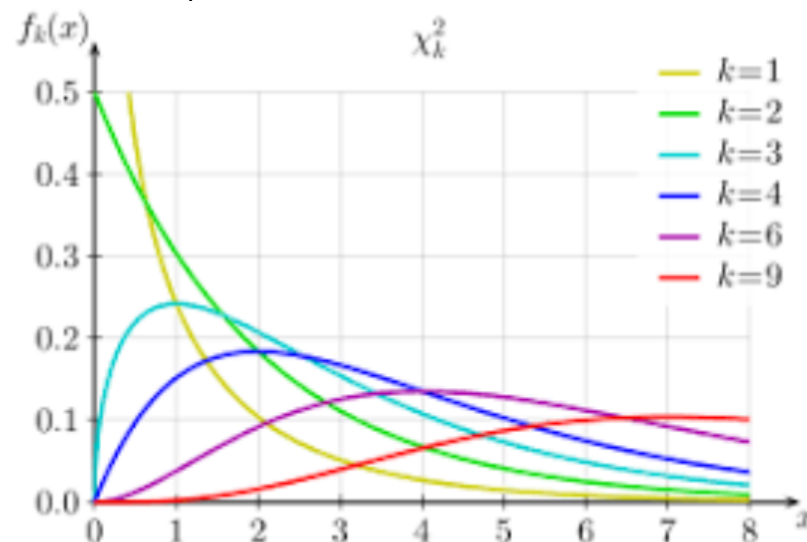
- What's more, it is a reasonably good assumption that under our (not all!!) null, this LRT is (approximately!) a chi-square distribution with 2 degrees of freedom (d.f.) assuming n is not too small!

Logistic Regression p-value

- To calculate our p-value, we need to know the distribution of our LRT statistic under the null hypothesis
- There is no simple form for this distribution for any given n (contrast with F-statistics!!) but we know that as n goes to infinite, we know the distribution is i.e. ($n \rightarrow \infty$):

$$LRT = -2\ln\Lambda = 2l(\hat{\theta}_1|\mathbf{y}) - 2l(\hat{\theta}_0|\mathbf{y})$$

$$LRT \rightarrow \chi_{df}^2$$



Modeling logistic covariates I

- Therefore, if we have a factor that is correlated with our phenotype and we do not handle it in some manner in our analysis, we risk producing false positives AND/OR reduce the power of our tests!
- The good news is that, assuming we have measured the factor (i.e. it is part of our GWAS dataset) then we can incorporate the factor in our model as a *covariate*:

$$Y = \gamma^{-1}(\beta_{\mu} + X_a\beta_a + X_d\beta_d + X_z\beta_z)$$

- The effect of this is that we will estimate the covariate model parameter and this will account for the correlation of the factor with phenotype (such that we can test for our marker correlation without false positives / lower power!)

Modeling logistic covariates II

- For our a logistic regression, our LRT (logistic) we have the same equations:

$$LRT = -2\ln\Lambda = 2l(\hat{\theta}_1|\mathbf{y}) - 2l(\hat{\theta}_0|\mathbf{y})$$

$$l(\hat{\theta}_1|\mathbf{y}) = \sum_{i=1}^n \left[y_i \ln(\gamma^{-1}(\hat{\beta}_\mu + x_{i,a}\hat{\beta}_a + x_{i,d}\hat{\beta}_d + x_{i,z}\hat{\beta}_z)) + (1 - y_i) \ln(1 - \gamma^{-1}(\hat{\beta}_\mu + x_{i,a}\hat{\beta}_a + x_{i,d}\hat{\beta}_d + x_{i,z}\hat{\beta}_z)) \right]$$
$$l(\hat{\theta}_0|\mathbf{y}) = \sum_{i=1}^n \left[y_i \ln(\gamma^{-1}(\hat{\beta}_\mu + x_{i,z}\hat{\beta}_z)) + (1 - y_i) \ln(1 - \gamma^{-1}(\hat{\beta}_\mu + x_{i,z}\hat{\beta}_z)) \right]$$

- Using the following estimates for the null hypothesis and the alternative making use of the IRLS algorithm (just add an additional parameter!):

$$\hat{\theta}_0 = \{ \hat{\beta}_\mu, \hat{\beta}_a = 0, \hat{\beta}_d = 0, \hat{\beta}_z \}$$

$$\hat{\theta}_1 = \{ \hat{\beta}_\mu, \hat{\beta}_a, \hat{\beta}_d, \hat{\beta}_z \}$$

- Under the null hypothesis, the LRT is still distributed as a Chi-square with 2 degree of freedom (why?):

$$LRT \rightarrow \chi_{df=2}^2$$

Summary I: logistic (no covariates)

- Test the null hypothesis: $H_0 : \beta_a = 0 \cap \beta_d = 0$ vs $H_A : \beta_a \neq 0 \cup \beta_d \neq 0$
- Step 1: use IRLS algorithm to get $MLE(\hat{\beta}) = \hat{\beta}_\mu$ which is the MLE under H_0 (i.e., $\hat{\theta}_0$) by using \mathbf{x} matrix with one column that is all ones!
- Step 2: substitute this MLE into:

$$l(\hat{\theta}_0|\mathbf{y}) = \sum_{i=1}^n \left[y_i \ln \left(\frac{e^{\hat{\beta}_\mu}}{1 + e^{\hat{\beta}_\mu}} \right) + (1 - y_i) \left(1 - \frac{e^{\hat{\beta}_\mu}}{1 + e^{\hat{\beta}_\mu}} \right) \right]$$

- Step 3: use IRLS algorithm to get $MLE(\hat{\beta}) = [\hat{\beta}_\mu, \hat{\beta}_a, \hat{\beta}_d]$ which is the MLE under H_A (i.e., $\hat{\theta}_1$) by using \mathbf{x} matrix with first column that is all ones, second column with $x_{i,a}$'s and third column with the $x_{i,d}$'s)
- Step 4: substitute these MLE into:

$$l(\hat{\theta}_1|\mathbf{y}) = \sum_{i=1}^n \left[y_i \ln \left(\frac{e^{\hat{\beta}_\mu + x_{i,a}\hat{\beta}_a + x_{i,d}\hat{\beta}_d}}{1 + e^{\hat{\beta}_\mu + x_{i,a}\hat{\beta}_a + x_{i,d}\hat{\beta}_d}} \right) + (1 - y_i) \left(1 - \frac{e^{\hat{\beta}_\mu + x_{i,a}\hat{\beta}_a + x_{i,d}\hat{\beta}_d}}{1 + e^{\hat{\beta}_\mu + x_{i,a}\hat{\beta}_a + x_{i,d}\hat{\beta}_d}} \right) \right]$$

- Step 5: use results from step 2 and step 4 to calculate:

$$LRT = -2\ln\Lambda = 2l(\hat{\theta}_1|\mathbf{y}) - 2l(\hat{\theta}_0|\mathbf{y})$$

- Use LRT and appropriate function in R (which?) to calculate p-value under chi-square $df = 2!$

Summary 2: logistic (covariates)

- Test the null hypothesis: $H_0 : \beta_a = 0 \cap \beta_d = 0$ vs $H_A : \beta_a \neq 0 \cup \beta_d \neq 0$
- Step 1: use IRLS algorithm to get $MLE(\hat{\beta}) = [\hat{\beta}_\mu, \hat{\beta}_z]$ which is the MLE under H_0 (i.e., $\hat{\theta}_0$) by using \mathbf{x} matrix with one column that is all ones!

- Step 2: substitute this MLE into:

$$l(\hat{\theta}_0|\mathbf{y}) = \sum_{i=1}^n \left[y_i \ln \left(\frac{e^{\hat{\beta}_\mu + x_{i,z} \hat{\beta}_z}}{1 + e^{\hat{\beta}_\mu + x_{i,z} \hat{\beta}_z}} \right) + (1 - y_i) \left(1 - \frac{e^{\hat{\beta}_\mu + x_{i,z} \hat{\beta}_z}}{1 + e^{\hat{\beta}_\mu + x_{i,z} \hat{\beta}_z}} \right) \right]$$

- Step 3: use IRLS algorithm to get $MLE(\hat{\beta}) = [\hat{\beta}_\mu, \hat{\beta}_a, \hat{\beta}_d, \hat{\beta}_z]$ which is the MLE under H_A (i.e., $\hat{\theta}_1$) by using \mathbf{x} matrix with first column that is all ones, second column with $x_{i,a}$'s and third column with the $x_{i,d}$'s)

- Step 4: substitute these MLE into:

$$l(\hat{\theta}_1|\mathbf{y}) = \sum_{i=1}^n \left[y_i \ln \left(\frac{e^{\hat{\beta}_\mu + x_{i,a} \hat{\beta}_a + x_{i,d} \hat{\beta}_d + x_{i,z} \hat{\beta}_z}}{1 + e^{\hat{\beta}_\mu + x_{i,a} \hat{\beta}_a + x_{i,d} \hat{\beta}_d + x_{i,z} \hat{\beta}_z}} \right) + (1 - y_i) \left(1 - \frac{e^{\hat{\beta}_\mu + x_{i,a} \hat{\beta}_a + x_{i,d} \hat{\beta}_d + x_{i,z} \hat{\beta}_z}}{1 + e^{\hat{\beta}_\mu + x_{i,a} \hat{\beta}_a + x_{i,d} \hat{\beta}_d + x_{i,z} \hat{\beta}_z}} \right) \right]$$

- Step 5: use results from step 2 and step 4 to calculate:

$$LRT = -2 \ln \Lambda = 2l(\hat{\theta}_1|\mathbf{y}) - 2l(\hat{\theta}_0|\mathbf{y})$$

- Use LRT and appropriate function in R (which?) to calculate p-value under chi-square $df = 2!$

That's it for today

- Next lecture we will finish our discussion of GLMs and begin our discussion of mixed models!