Quantitative Genomics and Genetics

BTRY 4830/6830; PBSB.5201.03

Lecture 23: Mixed Models

Jason Mezey
April 25, 2023 (T) 8:05-9:20

Summary of lecture 23: Mixed Models

- Last lecture, we largely completed our discussion of logistic regression
- Today, we will complete our logistic regression discussion with a quick review and by briefly introducing the broader family of models that linear and logistic regression belong to: generalized linear models!
- We will also (briefly) introduce Mixed Models!

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

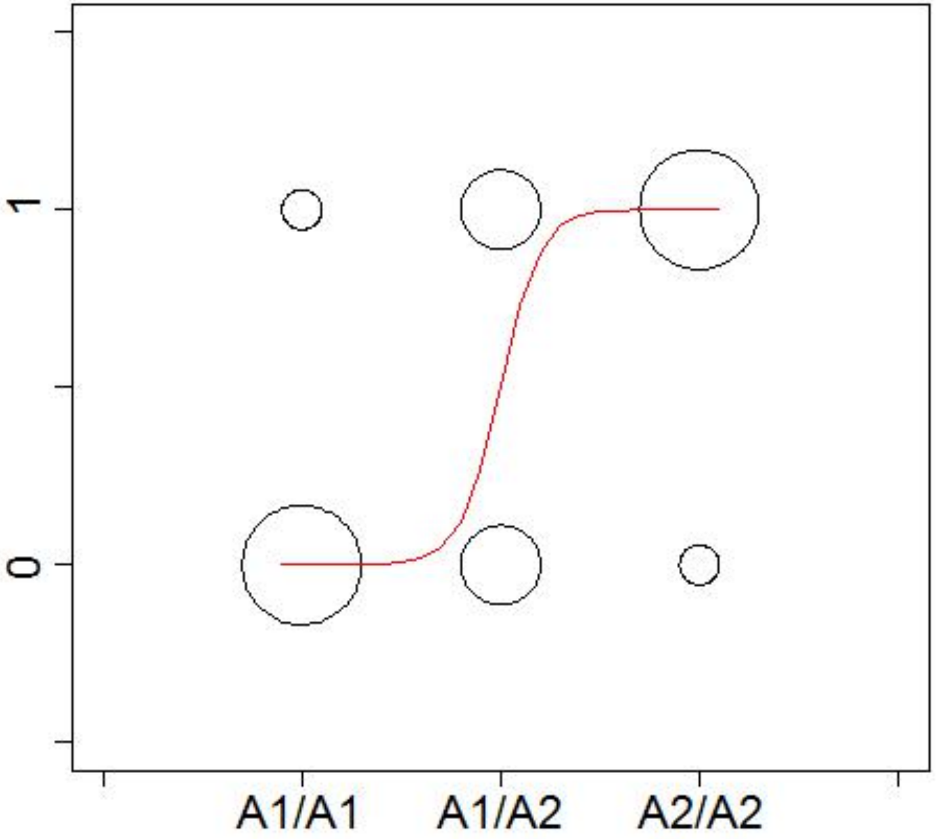
- As an example, we are often in situations where we are interested in identifying causal polymorphisms (loci) that contribute to the risk for developing a disease, e.g. heart disease, diabetes, etc.

- 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: 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 out approach to inference is the same with this model!

Review: Logistic regression: error term III

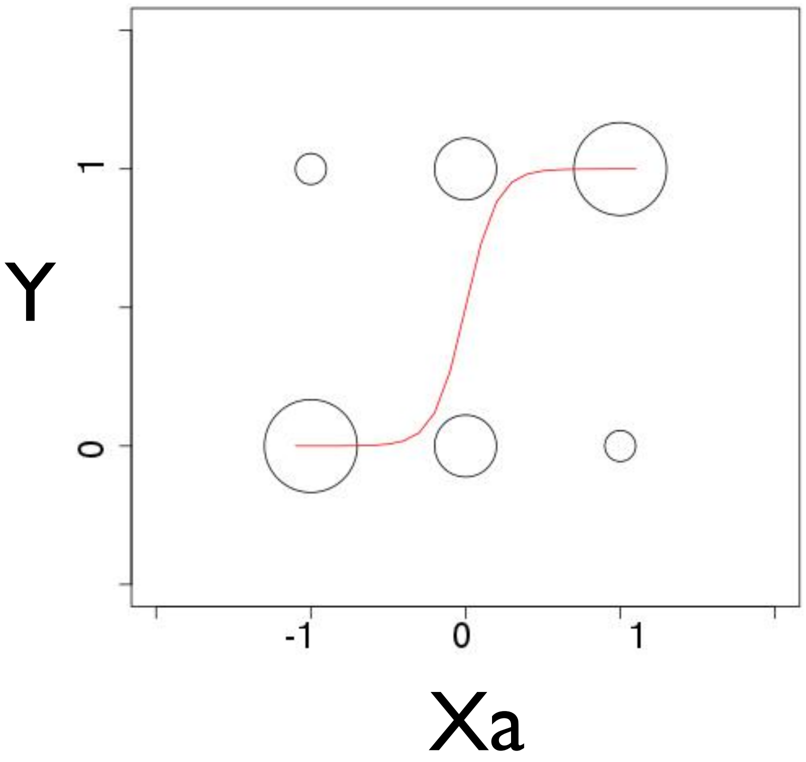
- This may look complicated at first glance but the intuition is relatively simple

- 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!):

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

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

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

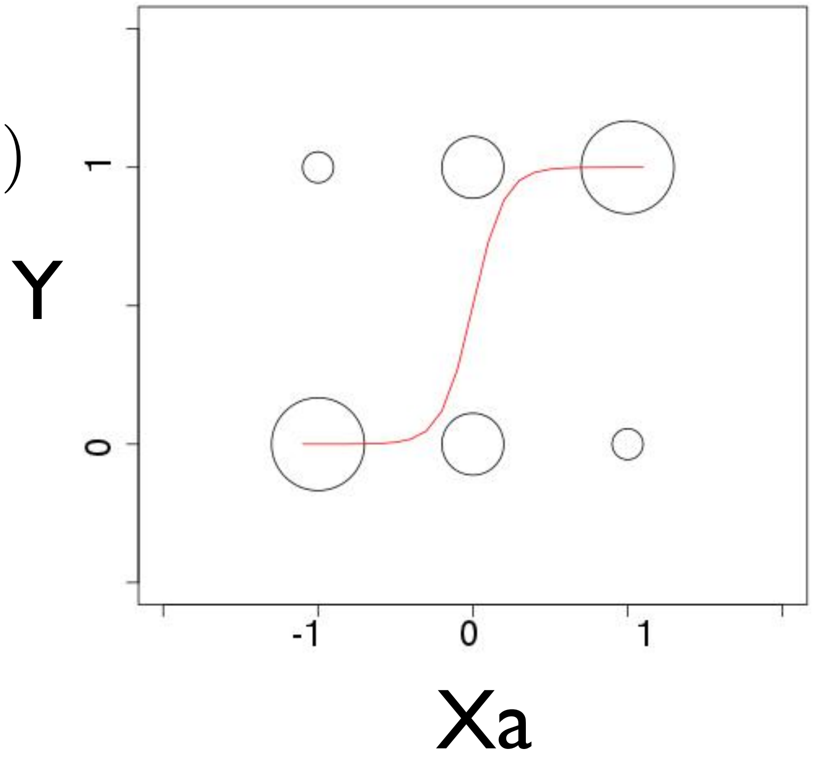


Review: Logistic regression: link function

- 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!!):

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

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



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:

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

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

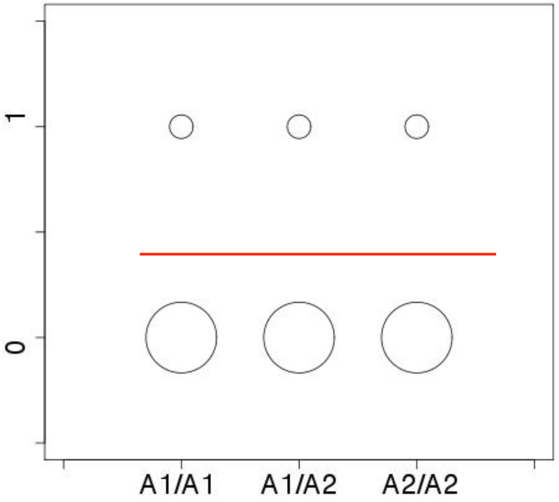
Inference

- 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

Review: logistic MLE

- 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):

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

- 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

Review: 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:

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

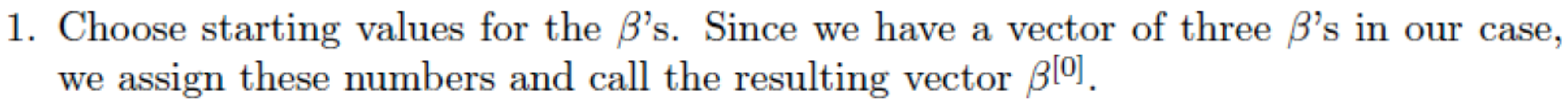
- For logistic regression (and GLM's in general!) we will construct an Iterative Re-weighted Least Squares (IRLS) algorithm, which has the following structure:

1. Choose starting values for the 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.

Review: 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:

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

Review: 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!):

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

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

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

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

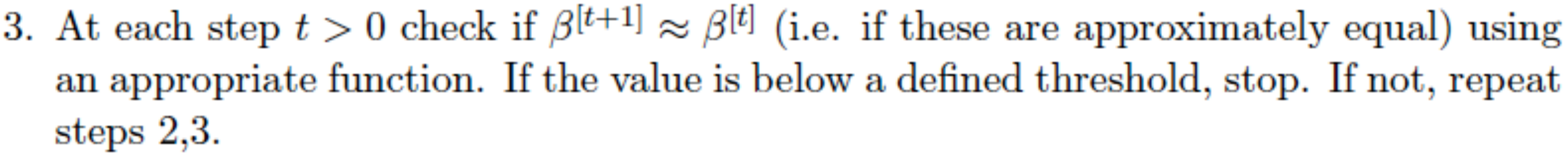
$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad \beta^{[t]} = \begin{bmatrix} \beta_{\mu}^{[t]} \\ \beta_{a}^{[t]} \\ \beta_{d}^{[t]} \end{bmatrix}$$

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

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

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

Review: 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)!!:

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

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

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

Review: 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):

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

- For our case, we need the following:

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

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

Review: 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-like equation

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

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

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

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

- where we use the same IRLS algorithm to provide estimates of 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!

Review: 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. ($n \to \infty$):

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

$$LRT \to \chi^2_{df}$$

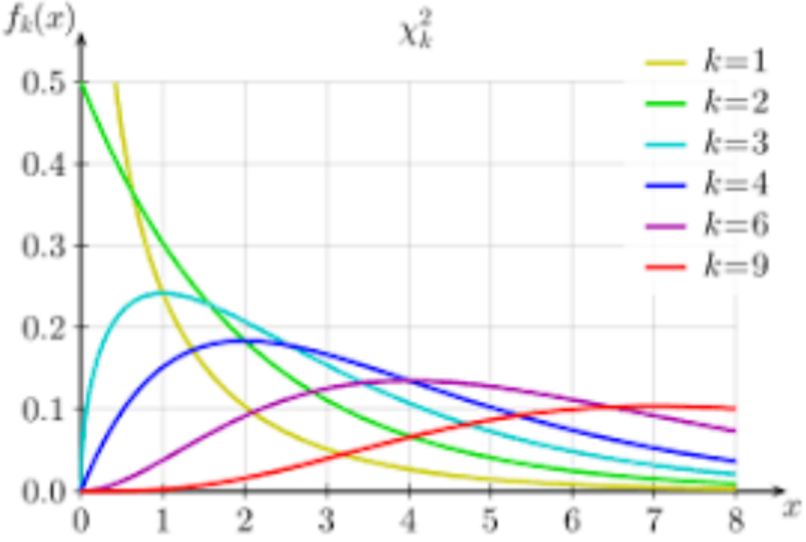
- What's more, it is a reasonably good assumption that under our (not all!!) null, this LRT is (approximately!) a chi-square distribution with 2 degrees of freedom (d.f.) assuming n is not too small!

Review: 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. ($n \to \infty$):

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

$$LRT \to \chi^2_{df}$$



Review: 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!)

Review: logistic covariates II

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

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

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

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

- 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?):

$$LRT \to \chi^2_{df=2}$$

Summary 1: 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 H0 (i.e., $\hat{\theta}_0$) by using **x** matrix with one column that is all ones!)

- Step 2: substitute this MLE into:

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

- Step 3: use IRLS algorithm to get $MLE(\hat{\beta}) = [\hat{\beta}_\mu, \hat{\beta}_a, \hat{\beta}_d]$ which is the MLE under H0 (i.e., $\hat{\theta}_1$) by using **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:

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

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

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

- 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 H0 (i.e., $\hat{\theta}_0$) by using **x** matrix with one column that is all ones!)

- Step 2: substitute this MLE into:

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

- 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 H0 (i.e., $\hat{\theta}_1$) by using **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:

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

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

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

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

Introduction to Generalized Linear Models (GLMs) I

- We have introduced linear and logistic regression models for GWAS analysis because these are the most versatile framework for performing a GWAS (there are many less versatile alternatives!)
- These two models can handle our genetic coding (in fact any genetic coding) where we have discrete categories (although they can also handle *X* that can take on a continuous set of values!)
- They can also handle (the sampling distribution) of phenotypes that have normal (linear) and Bernoulli error (logistic)
- How about phenotypes with different error (sampling) distributions? Linear and logistic regression models are members of a broader class called Generalized Linear Models (GLMs), where other models in this class can handle additional phenotypes (error distributions)

Introduction to Generalized Linear Models (GLMs) II

- To introduce GLMs, we will introduce the overall structure first, and second describe how linear and logistic models fit into this framework

- There is some variation in presenting the properties of a GLM, but we will present them using three (models that have these properties are considered GLMs):

  - The probability distribution of the response variable Y conditional on the independent variable X is in the exponential family of distributions

    $$Pr(Y|X) \sim expfamily$$

  - A link function relating the independent variables and parameters to the expected value of the response variable (where we often use the inverse!!)

    $$\gamma : \mathrm{E}(\mathbf{Y}|\mathbf{X}) \rightarrow \mathbf{X}\beta,$$

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

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

  - The error random variable ϵ has a variance which is a function of ONLY $\mathbf{X}\beta$

Exponential family I

- The exponential family is includes a broad set of probability distributions that can be expressed in the following `natural' form:

$$Pr(Y) \sim e^{\frac{Y\theta - b(\theta)}{\phi} + c(Y,\phi)}$$

- As an example, for the normal distribution, we have the following:

$$\theta = \mu, \phi = \sigma^2, b(\theta) = \frac{\theta^2}{2}, c(Y,\phi) = -\frac{1}{2}\left(\frac{Y^2}{\phi} + log(2\pi\phi)\right)$$

- Note that many continuous and discrete distributions are in this family (normal, binomial, poisson, lognormal, multinomial, several categorical distributions, exponential, gamma distribution, beta distribution, chi-square) but not all (examples that are not!?) and since we can model response variables with these distributions, we can model phenotypes with these distributions in a GWAS using a GLM (!!)

- Note that the normal distribution is in this family (linear) as is Bernoulli or more accurately Binomial (logistic)

Exponential family II

- Instead of the `natural' form, the exponential family is often expressed in the following form:

$$Pr(Y) \sim h(Y)s(\theta)e^{\sum_{i=1}^{k} w_i(\theta)t_i(Y)}$$

- To convert from one to the other, make the following substitutions:

$$k = 1, h(Y) = e^{c(Y,\phi)}, s(\theta) = e^{-\frac{b(\theta)}{\phi}}, w(\theta) = \frac{\theta}{\phi}, t(Y) = Y$$

- Note that the dispersion parameter is now no longer a direct part of this formulation

- Which is used depends on the application (i.e., for glm's the `natural' form has an easier to use form + the dispersion parameter is useful for model fitting, while the form on this slide provides advantages for other types of applications)

GLM link function

- A "link" function is just a function (!!) that acts on the expected value of *Y* given *X*:
- This function is defined in such a way such that it has a useful form for a GLM although there are some general restrictions on the form of this function, the most important is that they need to be monotonic such that we can define an inverse:

$$Y = f(X) \qquad f^{-1}(Y) = X$$

- For the logistic regression, we have selected the following link function, which is a logit function (a "canonical link") where the inverse is the logistic function (but note that others are also used for binomial response variables):

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

- What is the link function for a normal distribution?

GLM error function

- The variance of the error term in a GLM must be function of ONLY the independent variable and beta parameter vector:

$$Var(\epsilon) = f(\mathbf{X}\beta)$$

- This is the case for a linear regression (note the variance of the error is constant!!):

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

$$Var(\epsilon) = f(\mathbf{X}\beta) = \sigma_\epsilon^2$$

- As an example, this is the case for the logistic regression (note the error changes depending on the value of X!!):

$$Var(\epsilon) = \gamma^{-1}(\mathbf{X}\beta)(1 - \gamma^{-1}(\mathbf{X}\beta))$$

$$Var(\epsilon_i) = \gamma^{-1}(\beta_\mu + X_{i,a}\beta_a + X_{i,d}\beta_d)(1 - \gamma^{-1}(\beta_\mu + X_{i,a}\beta_a + X_{i,d}\beta_d)$$

Inference with GLMs

- We perform inference in a GLM framework using the same approach, i.e. MLE of the beta parameters using an IRLS algorithm (just substitute the appropriate link function in the equations, etc.)

- We can also perform a hypothesis test using a LRT (where the sampling distribution as the sample size goes to infinite is chi-square)

- In short, what you have learned can be applied for most types of regression modeling you will likely need to apply (!!)

(Brief) introduction to mixed models I

- A *mixed model* describes a class of models that have played an important role in early quantitative genetic (and other types) of statistical analysis before genomics (if you are interested, look up variance component estimation)

- These models are now used extensively in GWAS analysis as a tool for model covariates (often population structure!)

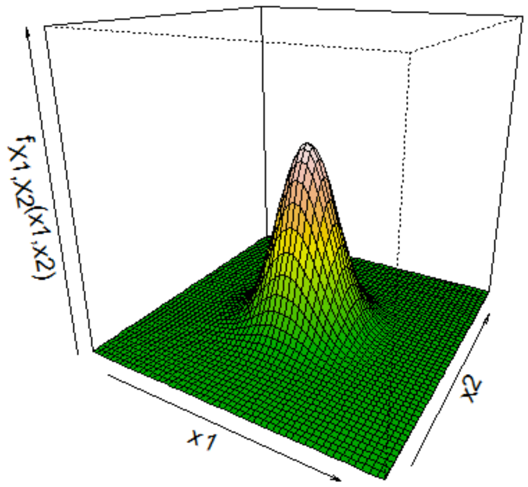
- These models considered effects as either "fixed" (they types of regression coefficients we have discussed in the class) and "random" (which just indicates a different model assumption) where the appropriateness of modeling covariates as fixed or random depends on the context (fuzzy rules!) - you will generally not have to deal with these issues in GWAS

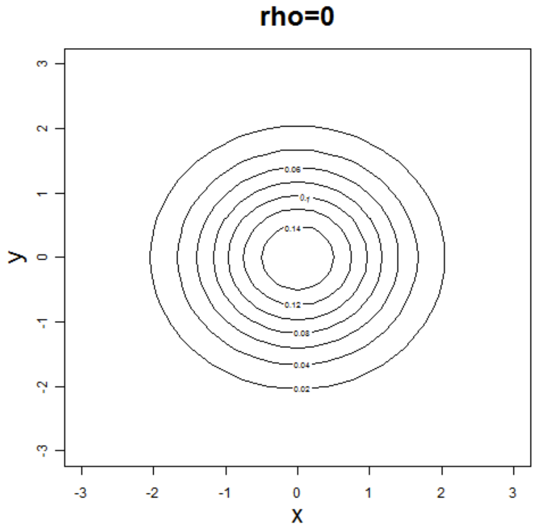
Introduction to mixed models II

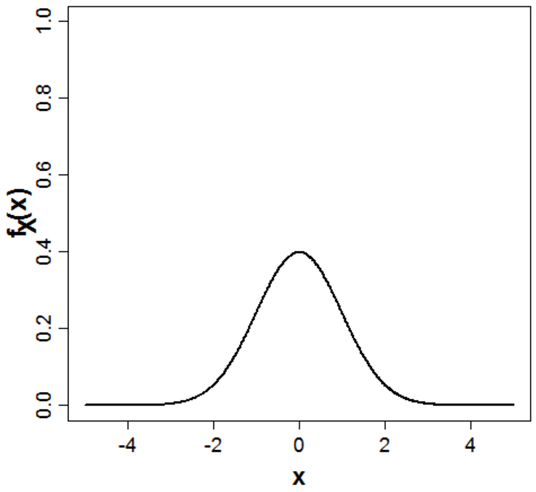
- Recall that for a linear regression of sample size n, we model the distributions of n total yi phenotypes using a linear regression model with normal error:

$$y_i = \beta_\mu + X_{i,a}\beta_a + X_{i,d}\beta_d + \epsilon_i \qquad \epsilon_i \sim N(0, \sigma^2_\epsilon)$$

- A reminder about how to think about / visualize multivariate (bivariate) normal distributions and marginal normal distributions:







- We can therefore consider the entire sample of yi and their associated error in an equivalent multivariate setting:

$$\mathbf{y} = \mathbf{x}\beta + \epsilon \qquad \epsilon \sim multiN(\mathbf{0}, \mathbf{I}\sigma^2_\epsilon)$$

Introduction to mixed models III

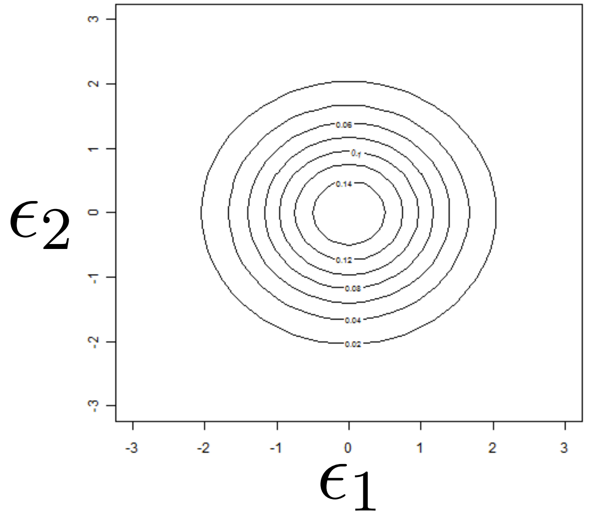
- Recall our linear regression model has the following structure:

$$y_i = \beta_\mu + X_{i,a}\beta_a + X_{i,d}\beta_d + \epsilon_i \quad \epsilon_i \sim N(0, \sigma_\epsilon^2)$$

- For example, for *n*=2:

$$y_1 = \beta_\mu + X_{1,a}\beta_a + X_{1,d}\beta_d + \epsilon_1$$

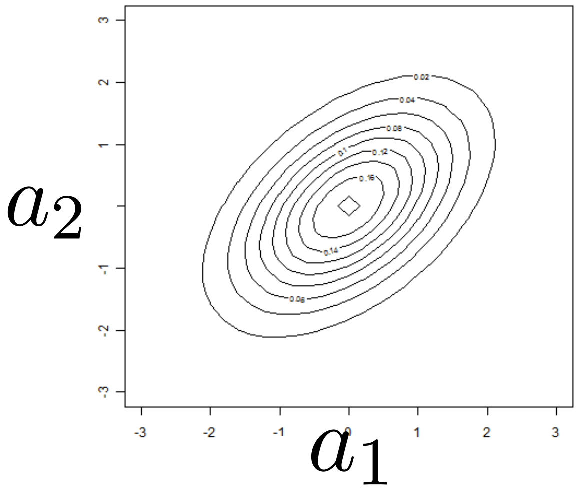
$$y_2 = \beta_\mu + X_{2,a}\beta_a + X_{2,d}\beta_d + \epsilon_2$$



- What if we introduced a correlation?

$$y_1 = \beta_\mu + X_{1,a}\beta_a + X_{1,d}\beta_d + a_1$$

$$y_2 = \beta_\mu + X_{2,a}\beta_a + X_{2,d}\beta_d + a_2$$



Introduction to mixed models IV

- The formal structure of a mixed model is as follows:

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{a} + \epsilon$$

$$\epsilon \sim multiN(\mathbf{0}, \mathbf{I}\sigma^2_\epsilon) \qquad \mathbf{a} \sim multiN(\mathbf{0}, \mathbf{A}\sigma^2_{\mathbf{a}})$$

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{i,a} & X_{i,d} \\ 1 & X_{i,a} & X_{i,d} \\ 1 & X_{i,a} & X_{i,d} \\ \vdots & \vdots & \vdots \\ 1 & X_{i,a} & X_{i,d} \end{bmatrix} \begin{bmatrix} \beta_\mu \\ \beta_a \\ \beta_d \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & ... & ... & ... & 1 \end{bmatrix} \begin{bmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \\ \mathbf{a}_3 \\ \vdots \\ \mathbf{a}_n \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

- Note that **X** is called the "design" matrix (as with a GLM), **Z** is called the "incidence" matrix, the **a** is the vector of random effects and note that the **A** matrix determines the correlation among the ai values where the structure of **A** is provided from external information (!!)

Introduction to mixed models V

- The matrix **A** is an *nxn* covariance matrix (what is the form of a covariance matrix?)
- Where does **A** come from? This depends on the modeling application...
- In GWAS, the random effect is usually used to account for population structure OR relatedness among individuals
    - For population structure, a matrix is constructed from the covariance (or similarity) among individuals based on their genotypes
    - For relatedness, we use estimates of identity by descent, which can be estimated from a pedigree or genotype data

Introduction to mixed models VI

- We perform inference (estimation and hypothesis testing) for the mixed model just as we would for a linear regression (!!)

- Note that in some applications, people might be $\sigma^2_\epsilon, \sigma^2_a$ interested in estimating the variance components but for GWAS, we are generally interested in regression parameters for our genotype (as before!): β_a, β_d

- For a GWAS, we will therefore determine the MLE of the genotype association parameters and use a LRT for the hypothesis test, where we will compare a null and alternative model (what is the difference between these models?)

Mixed models: inference I

- To estimate parameters, we will use the MLE, so we are concerned with the form of the likelihood equation

$$L(\beta, \sigma_a^2, \sigma_\epsilon^2|\mathbf{y}) = \int_{-\infty}^{\infty} Pr(\mathbf{y}|\beta, \mathbf{a}, \sigma_\epsilon^2) Pr(\mathbf{a}|\mathbf{A}\sigma_a^2) d\mathbf{a}$$

$$L(\beta, \sigma_a^2, \sigma_\epsilon^2|\mathbf{y}) = |\mathbf{I}\sigma_\epsilon^2|^{-\frac{1}{2}} e^{-\frac{1}{2\sigma_\epsilon^2}[\mathbf{y}-\mathbf{X}\beta-\mathbf{Za}]^{\mathrm{T}}[\mathbf{y}-\mathbf{X}\beta-\mathbf{Za}]} |\mathbf{A}\sigma_a^2|^{-\frac{1}{2}} e^{-\frac{1}{2\sigma_a^2}\mathbf{a}^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{a}}$$

$$l(\beta, \sigma_a^2, \sigma_\epsilon^2|\mathbf{y}) \propto -\frac{n}{2} ln\sigma_\epsilon^2 - -\frac{n}{2} ln\sigma_a^2 - \frac{1}{2\sigma_\epsilon^2} \left[\mathbf{y} - \mathbf{X}\beta - \mathbf{Za}\right]^{\mathrm{T}} \left[\mathbf{y} - \mathbf{X}\beta - \mathbf{Za}\right] - \frac{1}{2\sigma_a^2} \mathbf{a}^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{a}$$

- Unfortunately, there is no closed form for the MLE since they have the following form:

$$MLE(\hat{\beta}) = (\mathbf{X}\hat{\mathbf{V}}^{-1}\mathbf{X}^{\mathrm{T}})^{-1}\mathbf{X}^{\mathrm{T}}\hat{\mathbf{V}}^{-1}\mathbf{Y}$$

$$MLE(\hat{\mathbf{V}}) = f(\mathbf{X}, \hat{\mathbf{V}}, \mathbf{Y}, \mathbf{A})$$

$$\mathbf{V} = \sigma_a^2\mathbf{A} + \sigma_\epsilon^2\mathbf{I}$$

Mixed models: inference II

- We therefore need an algorithm to find the MLE for the mixed model

- We will discuss the use of an EM (Expectation-Maximization) algorithm for this purpose, which is an algorithm with good theoretical and practical properties, e.g. hill-climbing algorithm, guaranteed to converge to a (local) maximum, it is a stable algorithm, etc.

- We do not have time to introduce these properties in detail so we will just show the steps / equations you need to implement this algorithm (such that you can implement it yourself = see computer lab this week!)

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
- In general: 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

Mixed models: EM algorithm

1. At step $[t]$ for $t = 0$, assign values to the parameters: $\beta^{[0]} = \left[\beta_\mu^{[0]}, \beta_a^{[0]}, \beta_d^{[0]}\right], \sigma_a^{2,[0]}, \sigma_\epsilon^{2,[0]}$. These need to be selected such that they are possible values of the parameters (e.g. no negative values for the variance parameters).

2. Calculate the expectation step for $[t]$:

$$\mathbf{a}^{[t]} = \left(\mathbf{Z}^{\mathrm{T}}\mathbf{Z} + \mathbf{A}^{-1}\frac{\sigma_\epsilon^{2,[t-1]}}{\sigma_a^{2,[t-1]}}\right)^{-1}\mathbf{Z}^{\mathrm{T}}(\mathbf{y} - \mathbf{x}\beta^{[t-1]})$$

$$V_{\mathbf{a}}^{[t]} = \left(\mathbf{Z}^{\mathrm{T}}\mathbf{Z} + \mathbf{A}^{-1}\frac{\sigma_\epsilon^{2,[t-1]}}{\sigma_a^{2,[t-1]}}\right)^{-1}\sigma_\epsilon^{2,[t-1]}$$

3. Calculate the maximization step for $[t]$:

$$\beta^{[t]} = (\mathbf{x}^{\mathrm{T}}\mathbf{x})^{-1}\mathbf{x}^{\mathrm{T}}(\mathbf{y} - \mathbf{Z}\mathbf{a}^{[t]})$$

$$\sigma_a^{2,[t]} = \frac{1}{n}\left[\mathbf{a}^{[t]}\mathbf{A}^{-1}\mathbf{a}^{[t]} + tr(\mathbf{A}^{-1}V_{\mathbf{a}}^{[t]})\right]$$

$$\sigma_\epsilon^{2,[t]} = -\frac{1}{n}\left[\mathbf{y} - \mathbf{x}\beta^{[t]} - \mathbf{Z}\mathbf{a}^{[t]}\right]^{\mathrm{T}}\left[\mathbf{y} - \mathbf{x}\beta^{[t]} - \mathbf{Z}\mathbf{a}^{[t]}\right] + tr(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}V_{\mathbf{a}}^{[t]})$$

where tr is a trace function, which is equal to the sum of the diagonal elements of a matrix.

4. Iterate steps 2, 3 until $(\beta^{[t]}, \sigma_a^{2,[t]}, \sigma_\epsilon^{2,[t]}) \approx (\beta^{[t+1]}, \sigma_a^{2,[t+1]}, \sigma_\epsilon^{2,[t+1]})$ (or alternatively $lnL^{[t]} \approx lnL^{[t+1]}$).

Mixed Model 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):

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

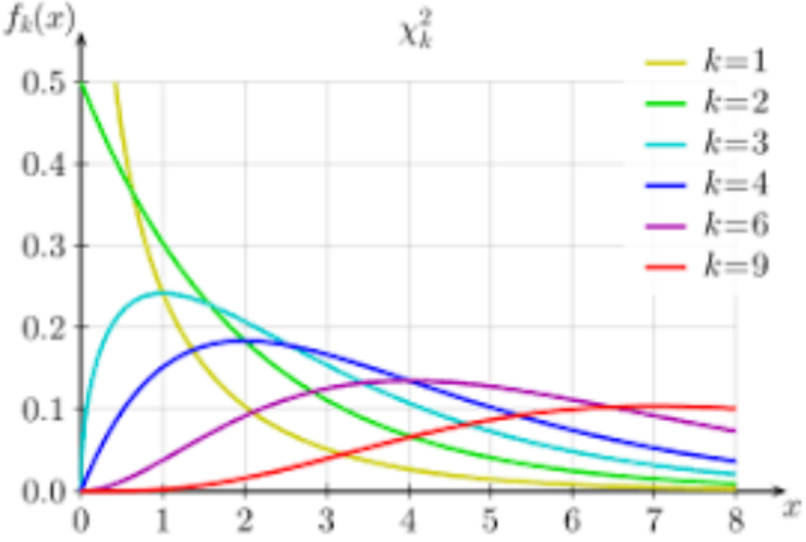
- To do this, run the EM algorithm twice, once for the null hypothesis (again what is this?) and once for the alternative (i.e. all parameters unrestricted) and then substitute the parameter values into the log-likelihood equations and calculate the LRT

Mixed Model 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. ($n \to \infty$):

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

$$LRT \to \chi^2_{df}$$



Mixed models: inference V

- In general, a mixed model is an advanced methodology for GWAS analysis but is proving to be an extremely useful technique for covariate modeling

- There is software for performing a mixed model analysis (e.g. R-package: lrgpr, EMMAX, FAST-LMM, TASSEL, etc.)

- Mastering mixed models will take more time than we have to devote to the subject in this class, but what we have covered provides a foundation for understanding the topic

Construction of **A** matrix I

- The matrix **A** is an *nxn* covariance matrix (what is the form of a covariance matrix?)
- Where does **A** come from? This depends on the modeling application...
- In GWAS, the random effect is usually used to account for population structure OR relatedness among individuals
  - For relatedness, we use estimates of identity by descent, which can be estimated from a pedigree or genotype data
  - For population structure, a matrix is constructed from the covariance (or similarity) among individuals based on their genotypes

Construction of **A** matrix II

$$Data = \left[\begin{array}{cccccc|ccc|}
\cline{7-9}
z_{11} & \dots & z_{1k} & y_{11} & \dots & y_{1m} & x_{11} & \dots & x_{1N} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
z_{n1} & \dots & z_{nk} & y_{n1} & \dots & y_{nm} & x_{11} & \dots & x_{nN} \\
\cline{7-9}
\end{array}\right]$$

- Calculate the *n*xn (*n*=sample size) covariance matrix for the individuals in your sample across all genotypes - this is a reasonable **A** matrix!

- There is software for calculating **A** and for performing a mixed model analysis (e.g. EMMAX, FAST-LMM, etc.)

- Mastering mixed models will take more time than we have to devote to the subject in this class, but what we have covered provides a foundation for understanding the topic

That’s it for today

- Next OPTIONAL lectures: Bayesian Statistics (!!)