Quantitative Genomics and Genetics

BTRY 4830/6830; PBSB.52201.03

Lecture 24: Intro to Logistic Regression IV

Jason Mezey

April 23, 2024 (T) 8:40-9:55

Announcements

- All homeworks and midterm have been graded (!!)
- Last work for the class: Project and Final
  - For project (due by 11:59PM, May 7!) typo in instructions: "SNP_info.csv" has three columns not four
  - Final will be same format as midterm (available May 11 and due by 11:59PM, May 18!) and you will do a GWAS analysis with a linear regression with and without covariates AND a logistic regression with and without covariates (!!)
- Reminder: last two computer labs are optional
  - This week: more logistic regression with and without covariates (probably worth attending!)
  - Next week: examples of EM algorithm for mixed models and MCMC algorithm for Bayesian inference

Summary of lecture 24: Logistic Regression III

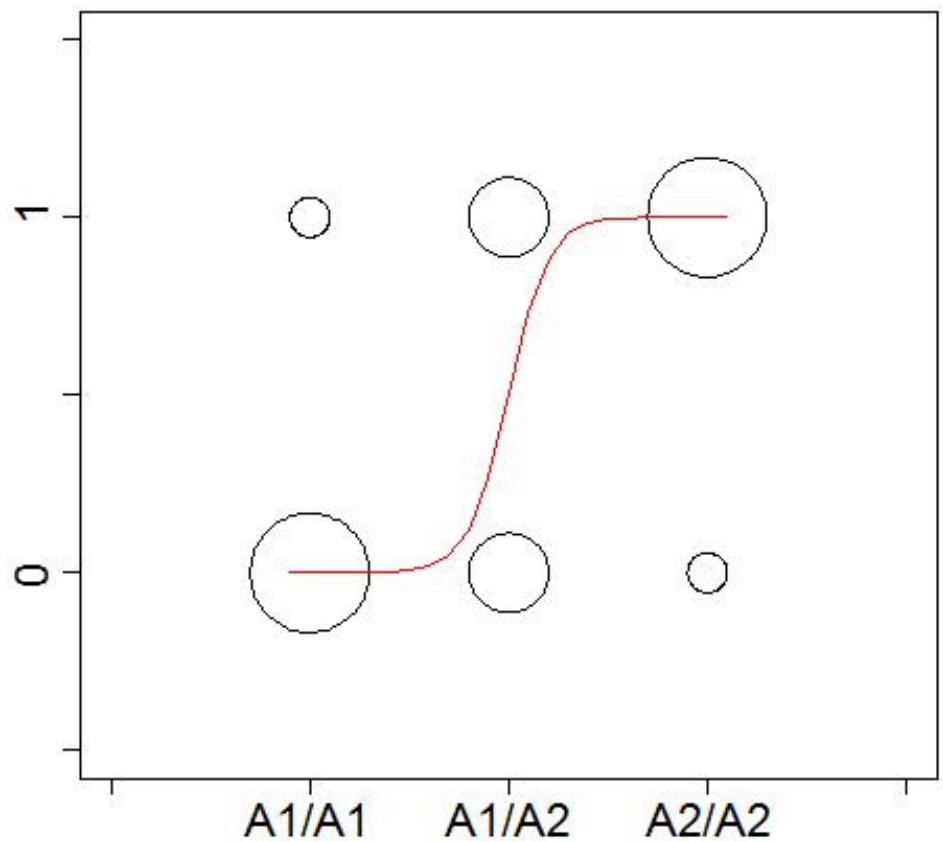
- Last lecture, we continued our discussion of the last major (non-optional!) topic: logistic regression
- Today we will complete our discussion!

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 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

Review: Logistic regression

- 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

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

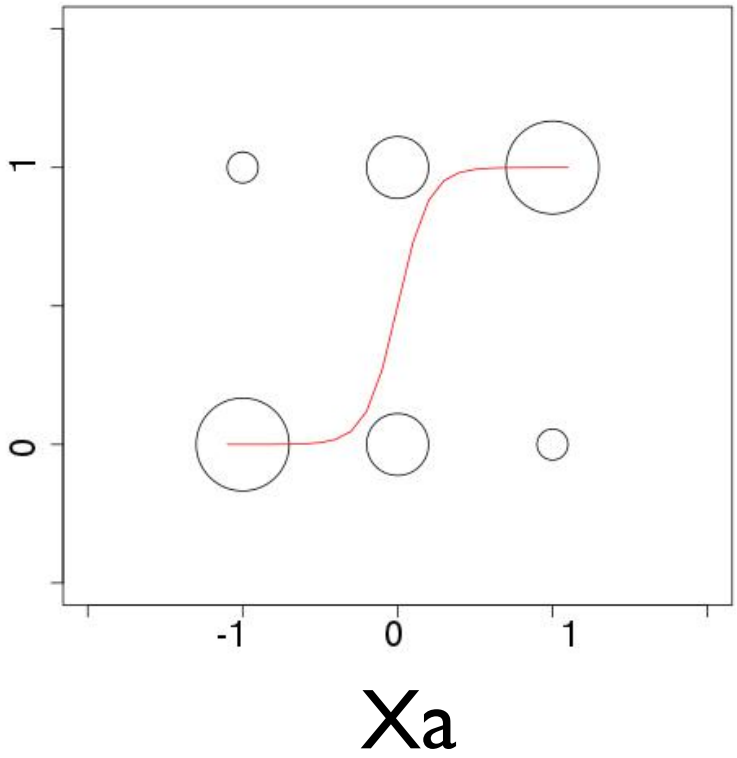
$$Pr(Z) \sim bern(p) \qquad p = logistic(\beta_\mu + X_a\beta_a + X_d\beta_d)$$

Review: Logistic regression

- 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!!):

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

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



Review: Notation (for R coding)

- 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:

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

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

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 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!):

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

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

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

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

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

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

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

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

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)!!:

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

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

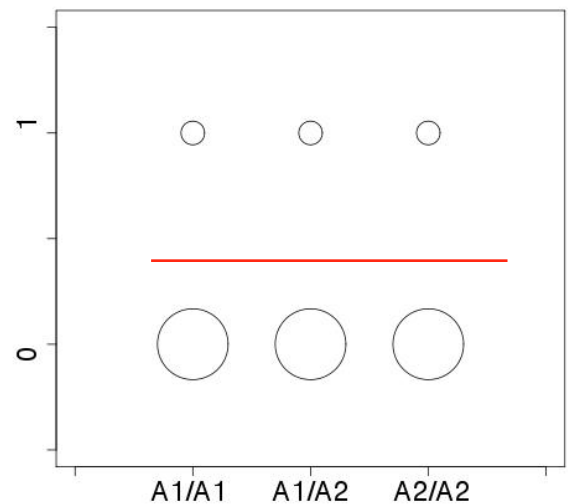
$$D = 2\sum_{i=1}^{n}\left[y_i ln\left(\frac{y_i}{\frac{e^{\beta_\mu^{[t]\text{or}[t+1]} + x_{i,a}\beta_a^{[t]\text{or}[t+1]} + x_{i,d}\beta_d^{[t]\text{or}[t+1]}}}{1+e^{\beta_\mu^{[t]\text{or}[t+1]} + x_{i,a}\beta_a^{[t]\text{or}[t+1]} + x_{i,d}\beta_d^{[t]\text{or}[t+1]}}}}\right) + (1-y_i) ln\left(\frac{1-y_i}{1-\frac{e^{\beta_\mu^{[t]\text{or}[t+1]}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

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

Reminder (!!): linear model estimation / testing

- First, determine the predicted value of the phenotype of each individual under the null hypothesis (how do we set up **x**?):

$$\hat{y}_{i,\hat{\theta}_0} = \hat{\beta}_{\mu,\hat{\theta}_0} + \sum_{j=1} x_{i,z,j}\hat{\beta}_{z,\hat{\theta}_0,j}$$

- Second, determine the predicted value of the phenotype of each individual under the alternative hypothesis (set up **x**?):

$$\hat{y}_{i,\hat{\theta}_1} = \hat{\beta}_{\mu,\hat{\theta}_1} + x_{i,a}\hat{\beta}_{a,\hat{\theta}_1} + x_{i,d}\hat{\beta}_{d,\hat{\theta}_1} + \sum_{j=1} x_{i,z,j}\hat{\beta}_{z,\hat{\theta}_1,j}$$

- Third, calculate the "Error Sum of Squares" for each:

$$SSE(\hat{\theta}_0) = \sum_{i=1}^{n}(y_i - \hat{y}_{i,\hat{\theta}_0})^2 \qquad SSE(\hat{\theta}_1) = \sum_{i=1}^{n}(y_i - \hat{y}_{i,\hat{\theta}_1})^2$$

- Finally, we calculate the F-statistic with degrees of freedom [2, n-3] (why two and n-#params degrees of freedom?):

$$F_{[2,n-\#(\hat{\theta}_1)]}(\mathbf{y}, \mathbf{x_a}, \mathbf{x_d}) = \frac{\frac{SSE(\hat{\theta}_0)-SSE(\hat{\theta}_1)}{2}}{\frac{SSE(\hat{\theta}_1)}{n-\#(\hat{\theta}_1)}}$$

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:

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

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

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!

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. ($n \rightarrow \infty$):

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

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

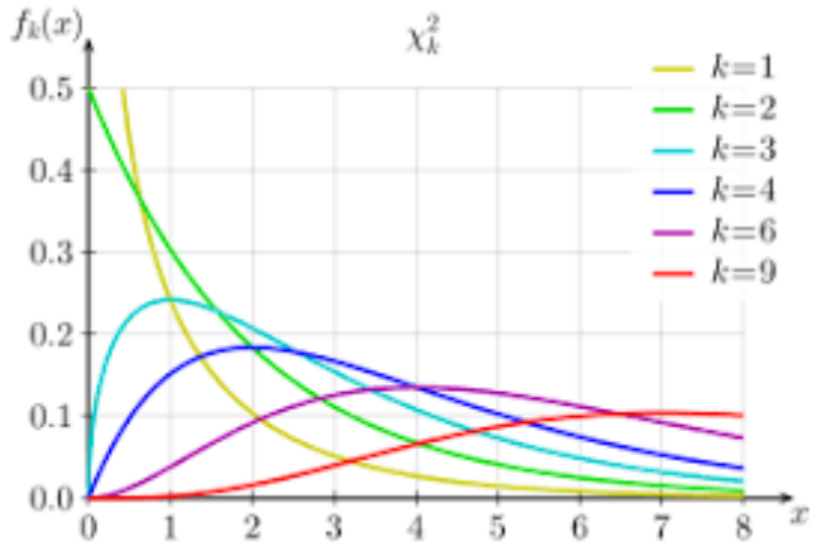
- What's more, it is a reasonably good assumption that under our (not all!!) null, this LRT is (approximately!) a chi-square distribution with 2 degrees of freedom (d.f.) assuming n is not too small!

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. ($n \rightarrow \infty$):

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

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



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:

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

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

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

- 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?):

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

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 HA (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 HA (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 (!!)

That's it for today

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