Lecture 16: Logistic Regression III and Haplotype Testing

Jason Mezey
jgm45@cornell.edu
April 12, 2016 (T) 8:40-9:55
Announcements I

- Homework #6 posted (last one!), this is due 11:59PM Fri., April 15
- Homework #5 now graded and available
- Project available April 14 (more details to come!)
- Final will be on **DURING THE FIRST WEEK OF EXAMS:** available Mon., May 16 and due Thurs., May 19 (!!!)
- Midterm is graded:
  - Available after class today
  - How did I do? See next slide...
Announcements II

- If you got < 80, please email me to let me know what your grade type so we can make sure you are on track...

Histogram of Midterm grades: Ithaca–NYC combined
Today, we will complete our discussion of logistic regression by discussing the broader class of models that includes both linear and logistic regressions: generalized linear models.

We will also briefly discuss haplotypes and how we use these in a GWAS.

We will discuss covariates next lecture.
Conceptual Overview

Genetic System

Does $A_1 \rightarrow A_2$ affect $Y$?

Sample or experimental pop

Measured individuals
(genotype, phenotype)

Regression model

Pr($Y|X$)

Reject / DNR

Model params
F-test
• Now we have all the critical components for performing a GWAS with a case / control phenotype!

• The procedure (and goals!) are the same as before, for a sample of \( n \) individuals where for each we have measured a case / control phenotype and \( N \) genotypes, we perform \( N \) hypothesis tests

• To perform these hypothesis tests, we need to run our IRLS algorithm for EACH marker to get the MLE of the parameters under the alternative (= no restrictions on the beta’s!) and use these to calculate our LRT test statistic for each marker

• We then use these \( N \) LRT statistics to calculate \( N \) p-values by using a chi-square distribution (how do we do this is R?)
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!!)
    \[
    \gamma : E(Y|X) \rightarrow X\beta.
    \]
    \[
    \gamma(E(Y|X)) = X\beta
    \]
    \[
    E(Y|X) = \gamma^{-1}(X\beta)
    \]

  - The error random variable $\epsilon$ has a variance which is a function of ONLY $X\beta$
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 \( X \):

\[
Y = f(X) \quad f^{-1}(Y) = X
\]

- 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:

\[
\gamma(\mathbb{E}(Y|X)) = \ln \left( \frac{e^{X\beta}}{1+e^{X\beta}} \right) \quad \mathbb{E}(Y|X) = \gamma^{-1}(X\beta) = \frac{e^{X\beta}}{1 + e^{X\beta}}
\]

- 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):

\[
\gamma(\mathbb{E}(Y|X)) = \ln \left( \frac{e^{X\beta}}{1+e^{X\beta}} \right)
\]

- 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:

\[ \text{Var}(\epsilon) = f(X\beta) \]

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

\[ \epsilon \sim N(0, \sigma_\epsilon^2) \]

\[ \text{Var}(\epsilon) = f(X\beta) = \sigma_\epsilon^2 \]

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

\[ \text{Var}(\epsilon) = \gamma^{-1}(X\beta)(1 - \gamma^{-1}(X\beta)) \]

\[ \text{Var}(\epsilon_i) = \gamma^{-1}(\beta_\mu + X_{i,a}\beta_a + X_{i,d}\beta_d)(1 - \gamma^{-1}(\beta_\mu + X_{i,a}\beta_a + X_{i,d}\beta_d)) \]
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 (!!!)
Haplotype testing I

- We have just extended our GWAS framework to handle additional phenotypes.
- We can also extend our GWAS framework to handle genotypes defined using a different approach.
- In this case, let’s consider using haplotype alleles in our testing framework.
- Note that a haplotype collapses genetic marker information but in some cases, testing using haplotypes is more effective than testing one genetic marker at a time.
Haplotype testing II

- **Haplotype** - a series of ordered, linked alleles that are inherited together

- For the moment, let’s consider a haplotype to define a “function” that takes a set of alleles at several loci A, B, C, D, etc. and outputs a haplotype allele:

  \[ h = f(A_i, B_j, ...) \]

- For example, if these loci are each a SNP with the following alleles (A,G), (A,T),(G,C),(G,C) we could define the following haplotype alleles:

  \[ h_1 = (A, A, C, C') \quad h_2 = (G, T, G, G) \]
Haplotype testing III

- Note that how we define haplotype alleles is somewhat arbitrary but in general, we define a haplotype for a set of genetic markers (loci) that are physically linked that are frequently occur in a population.

- How many markers is somewhat arbitrary, e.g., we often define sets that match observed patterns of LD.

- How many haplotype alleles we define is also somewhat arbitrary, where we define haplotype alleles that have appreciable frequency in the population.

- For example, four the four loci with alleles (A,G), (A,T), (G,C), (G,C), how many haplotype alleles could we define?

- However, it could be that only the following two combinations have relatively “high” allele frequencies (say >0.05 = arbitrary!)

  \[ h_1 = (A, A, C, C) \quad h_2 = (G, T, G, G) \]

- In such a case, we can collapse the many alleles into just a few!
Haplotype testing IV

- As an example of haplotype allele collapsing, say for our case of four loci (A,G), (A,T),(G,C),(G,C), we have lots of LD (!!) such that there are only 4 alleles in the population (i.e. all other combinations have frequency of zero!):


- Let’s also say that the frequencies of the third and fourth of these in the population are < 0.01

- In this case, we can define just two haplotype alleles that collapse the other alleles as follows (where * means “any” genetic marker allele):

\[ h_1 = (A, A, *, C) \quad h_2 = (G, T, *, G) \]

\[ h_1 = h_1^* \cup h_3^* \quad h_2 = h_2^* \cup h_4^* \]

- NOTE: we are therefore loosing information using this approach!!
GWAS with haplotypes I

- Once we have defined haplotype alleles, we can proceed with a GWAS using our framework (just substitute haplotype alleles and genotypes for genetic marker alleles and genotypes!)

- For example, in a case where we only have two haplotype alleles, we can code our independent variables for our regression model as follows:

  \[ X_a(h_1h_1) = -1, X_a(h_1h_2) = 0, X_a(h_2h_2) = 1 \]
  \[ X_d(h_1h_1) = -1, X_d(h_1h_2) = 1, X_d(h_2h_2) = -1 \]

- All other aspects remain the same (although what is the effect on our interpretation of where the causal polymorphism is located?)
GWAS with haplotypes II

- Given that we are losing information by using a haplotype testing approach in a GWAS, why might we want to use this approach?

- As one example consider the following case of haplotypes in a population:

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$A_1$</td>
<td>$B_1$</td>
<td>$(C_1)^*$</td>
</tr>
<tr>
<td>$A_1$</td>
<td>$B_2$</td>
<td>$(C_1)^*$</td>
</tr>
<tr>
<td>$A_2$</td>
<td>$B_1$</td>
<td>$(C_1)^*$</td>
</tr>
<tr>
<td>$A_1$</td>
<td>$B_1$</td>
<td>$(C_1)^*$</td>
</tr>
<tr>
<td>$A_2$</td>
<td>$B_2$</td>
<td>$(C_2)^*$</td>
</tr>
<tr>
<td>$A_2$</td>
<td>$B_1$</td>
<td>$(C_2)^*$</td>
</tr>
<tr>
<td>$A_1$</td>
<td>$B_2$</td>
<td>$(C_2)^*$</td>
</tr>
<tr>
<td>$A_2$</td>
<td>$B_2$</td>
<td>$(C_2)^*$</td>
</tr>
</tbody>
</table>
Advantages of haplotype testing

• In some cases (system and sample dependent!), the haplotype is a better “tag” of the causal polymorphism than any of the surrounding markers

• In such a case, the \( \text{Corr}(X_h, X) > \text{Corr} (X', X) \) and therefore has a higher probability of correctly rejecting the null hypothesis

• Another “advantage” is by putting together markers, we are performing less total tests in our GWAS (in what sense is this an advantage!?)
Disadvantages of haplotype testing

- Collapsing to haplotypes may produce a better tag but it also may not (!!), i.e. sometimes (in fact often!) individual genetic markers are better tags of the causal polymorphism.

- Another disadvantage is resolution, since we absolutely cannot resolve the position of the causal polymorphism to a position smaller than the range of the haplotype alleles, i.e. large haplotypes can have smaller resolution.

- If we had measured the causal polymorphism in our data, should we use haplotype testing (i.e. in the future, the importance of haplotype testing may decrease).
Should I apply haplotype testing in my GWAS?

- Yes! but apply both an individual marker testing approach (always!) as well as a haplotype test (optional)

- The reason is that we never know the true answer in our GWAS (as with any statistical analysis!) so it doesn’t hurt us to explore our dataset with as many techniques as we want to apply

- In fact, this will be a continuing theme of the class, i.e. keep analyzing GWAS with as many methods as you find useful

- However, since we never know the right answer for certain, if we get conflicting results, which one do we interpret as “correct”!?
Where do haplotypes come from?

- A deep discussion of the origin of haplotypes (remember: a fuzzy definition!) is another subject that is in the realm of population genetics and therefore we cannot discuss this in detail in this class (again: I encourage you to take a class on population genetics!)

- However, we can get an intuition about where haplotypes come from by remembering that the origin of new haplotype alleles are mutations and that new haplotype alleles can be produced by recombination.

- In fact, these two processes also underlie the amount of LD in the population and therefore what blocks of alleles are inherited as a haplotype (and we therefore use them to define haplotypes using system specific criteria).
Defining haplotypes

- We could spend multiple lectures on how people define haplotypes for given systems and the algorithms used for this purpose (so we will just briefly mention the main concepts here)

- To define haplotypes, we need to “phase” measured genotype markers, decide on the number of genotype markers to put together into a haplotype block, and decide how many haplotype alleles to consider

- Remember: there are no universal rules for doing this (system dependent!)
Phasing haplotypes

• To get a sense of the phasing problem, consider a case where we have two markers that are right next to each other on a chromosome and we know we want to put them together in a haplotype block.

• Say one marker is (A,T) and the other marker is (G,C) and we are considering a diploid individual who is a heterozygote for both of these markers, which of the marker alleles are physically linked in this individual?

• Figuring this out for individuals in a sample is the phasing problem and there are many algorithms for accomplishing this goal (note that in the future, technology may make this a non-issue...).
Deciding on how many genotypes to include in a haplotype block

- Again, while there is no set rule, how we decide on genotypes to include in a haplotype block depends on LD

- The general rule: if we have a set of markers in high LD with each other but low LD with other markers, we use this as a guide for defining the haplotype block
Deciding on how many haplotype alleles to consider

- Again, there are no set rules for how many haplotype alleles to define, but in general, we define a set where the frequency in a population is above some MAF threshold (which depends on the system).
- With a MAF cutoff of say 0.05, this generally limits us to 2-5 haplotype alleles (e.g. in humans!)
- There are however cases where we might want to consider rarer haplotypes (what are some of these?)
Haplotype GWAS wrap-up

• Haplotypes are a physical and sampling consequence of how genetic systems work (just like LD!)

• Definitions of haplotype blocks and haplotype alleles depend on the system and context (fuzzy definition)

• Regardless of how we define them, once we have haplotype alleles, we can use them as we would genetic markers in our GWAS analysis framework

• While optional, it is never a bad idea to perform a haplotype analysis of your GWAS in addition to your single marker analysis (ALWAYS do a single marker analysis)
That’s it for today

- Next lecture: we will discuss covariates, including one of the most important (population structure)!
- We will also discuss a “minimum checklist” that you should perform in your GWAS analysis (!!)