BTRY 4830/6830: Quantitative Genomics and Genetics

Lecture 16: Logistic regression II: GWAS for case / control phenotypes

Jason Mezey
jgm45@cornell.edu
Oct. 21, 2014 (T) 8:40-9:55
Announcements

• Homework #5 will be available tomorrow

• I will have reduced (!!) office hours this week: Thurs. 4-5PM
Summary of lecture 16

• In previous lectures, have began our introduction to logistic regression

• Today we will complete our discussion of how to perform inference (and a GWAS!) with a logistic regression
Conceptual Overview

Genetic System

Does A1 -> A2 affect Y?

Sample or experimental pop

Measured individuals (genotype, phenotype)

Regression model

Reject / DNR

Model params F-test

Pr(Y|X)
Review: Case / Control Phenotypes

• 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: linear vs. logistic

- Recall that for a linear regression, the regression function was a line and the error term accounted for the difference between each point and the expected value (the linear regression line), which we assume follow a normal distribution.

- For a logistic regression, we use the logistic function and the error term makes up the value to either 0 or 1:
Review: calculating the components of an individual I

- Recall that an individual with phenotype $Y_i$ is described by the following equation:

$$Y_i = E(Y_i|X_i) + \epsilon_i$$

$$Y_i = \gamma^{-1}(Y_i|X_i) + \epsilon_i$$

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

- To understand how an individual with a phenotype $Y_i$ and a genotype $X_i$ breaks down in this equation, we need to consider the expected (predicted!) part and the error term (we will do this separately)
Review: calculating the components of an individual II

- 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 \( i \):

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

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

  \[
  0 = 0.1 - 0.1
  \]
Review: calculating the components of an individual III

- For example, say we have an individual $i$ that has genotype $A1A1$ 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 $i$:
  \[ Y_i = \frac{e^{\beta_{\mu} + x_{i,a}\beta_{a} + x_{i,d}\beta_d}}{1 + e^{\beta_{\mu} + x_{i,a}\beta_{a} + x_{i,d}\beta_d}} + \epsilon_i \]
  \[ 1 = \frac{e^{0.2 + (-1)2.2 + (-1)0.2}}{1 + e^{0.2 + (-1)2.2 + (-1)0.2}} + \epsilon_i \]
  \[ 1 = 0.1 + 0.9 \]
Review: calculating the components of an individual IV

- 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 $i$:

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

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

$$
0 = 0.6 - 0.6
$$
Review: calculating the components of an individual $V$

- For example, say we have an individual $i$ that has genotype A2A2 and phenotype $Y_i = 0$
- We know $X_a = 1$ and $Xd = -1$
- Say we also know that for the population, the true parameters (which we will not know in practice! We need to infer them!) are:
  \[
  \beta_\mu = 0.2 \quad \beta_a = 2.2 \quad \beta_d = 0.2
  \]
- We can then calculate the $E(Y_i|X_i)$ and the error term for $i$:
  \[
  Y_i = \frac{e^{\beta_\mu + x_{i,a} \beta_a + x_{i,d} \beta_d}}{1 + e^{\beta_\mu + x_{i,a} \beta_a + x_{i,d} \beta_d}} + \epsilon_i
  \]
  \[
  0 = \frac{e^{0.2 + (1)2.2 + (-1)0.2}}{1 + e^{0.2 + (1)2.2 + (-1)0.2}} + \epsilon_i
  \]
  \[
  0 = 0.9 - 0.9
  \]
Review: the error term I

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

\[
\epsilon_i | (Y_i = 0) = -E(Y_i \mid X_i) \quad \epsilon_i | (Y_i = 1) = 1 - E(Y_i \mid X_i)
\]

- For the entire distribution of the population, recall that

\[
Pr(\epsilon_i) \sim bern(p \mid X) - E(Y \mid X)
\]

\[
p = E(Y \mid X)
\]

For example:

\[
\epsilon_i = -0.1 \quad \epsilon_i = -0.9 \\
p = 0.1
\]
Review: the error term II

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

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

- For the entire distribution of the population, recall that

  $$Pr(\epsilon_i) \sim bern(p \mid X) - E(Y \mid X)$$

  $$p = E(Y \mid X)$$

  For example:

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

  $$p = 0.6$$
Review: the error term III

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

\[
\epsilon_i \mid (Y_i = 0) = -E(Y_i \mid X_i) \quad \epsilon_i \mid (Y_i = 1) = 1 - E(Y_i \mid X_i)
\]

- For the entire distribution of the population, recall that

\[
Pr(\epsilon_i) \sim bern(p \mid X) - E(Y \mid X)
\]

\[
p = E(Y \mid X)
\]

For example:

\[
\epsilon_i = -0.9 \quad \epsilon_i = 0.1
\]

\[
p = 0.9
\]
Comments

- 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)
- We are dealing with only thee Xa points in our genetic model, why does the function have a value for all values of X?
- For an entire sample, we can use matrix notation as follows:

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

\[
E(y|x) = \gamma^{-1}(x\beta) = \begin{bmatrix}
\frac{e^{\beta \mu + x_1, a\beta a + x_1, d\beta d}}{1 + e^{\beta \mu + x_1, a\beta a + x_1, d\beta d}} \\
\vdots \\
\frac{e^{\beta \mu + x_n, a\beta a + x_n, d\beta d}}{1 + e^{\beta \mu + x_n, a\beta a + x_n, d\beta d}}
\end{bmatrix}
\]
Inference in logistic regression

- 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 and although we will not construct an F-test for logistic regression hypothesis testing, we will construct an LRT!)

- Just as with linear regression to construct a LRT, we need the MLE of the (beta) parameters of the logistic regression.
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
IRLS algorithm I

- **algorithm** - a sequence of instructions for taking an input and producing an output

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

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

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

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

• For the step (1), we can assign any starting values, since the algorithm is “convex” (although for other algorithms, we need to be careful in how we assign our starting values!)

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

• For step (2), we will update this vector at each “iteration” using the following equation:

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

\[
W_{ii} = \gamma^{-1}(\beta^{[t]} + x_{i,a} \beta^{[t]}_a + x_{i,d} \beta^{[t]}_d) (1 - \gamma^{-1}(\beta^{[t]} + x_{i,a} \beta^{[t]}_a + x_{i,d} \beta^{[t]}_d))
\]

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

• Note an alternative way of writing these equations:

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

• For step (3), we decide when to stop the algorithm using the “deviance criterion”:

\[
\Delta D = |D[t+1] - D[t]|
\]

\[
D = 2 \sum_{i=1}^{n} \left[ y_i ln \left( \frac{y_i}{\gamma^{-1}(\beta^{[t]}_\mu + x_{i,a} \beta^{[t]}_a + x_{i,d} \beta^{[t]}_d)} \right) + (1 - y_i) ln \left( \frac{1 - y_i}{1 - \gamma^{-1}(\beta^{[t]}_\mu + x_{i,a} \beta^{[t]}_a + x_{i,d} \beta^{[t]}_d)} \right) \right]
\]

\[
\beta^{[t+1]} \approx \beta^{[t]} \quad \text{when} \quad \Delta D < 10^{-6}
\]
Hypothesis testing: LRT I

• Recall that our null and alternative hypotheses are:

\[ H_0 : \beta_a = 0 \cap \beta_d = 0 \quad H_A : \beta_a \neq 0 \cup \beta_d \neq 0 \]

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

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

• Under the null, this LRT is (approximately!) a chi-square distribution with 2 degrees of freedom (d.f.) or more accurately:

\[ n \to \infty \quad LRT \to \chi^2_{df} \]
Hypothesis testing: LRT II

- For the left and right terms of the LRT we need to plug our parameter estimates under the alternative and null into the log-likelihood equation:

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

- 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

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

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

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

- where we use the same IRLS algorithm to provide estimates of \( \hat{\beta}_{\mu,0} \)
Performing a GWAS

• 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 in R?)
That’s it for today

• Next lecture: we will complete our discussion of logistic regression by introducing the broader GLM framework

• We will also begin our discussion of covariates!