Quantitative Genomics and Genetics
BTRY 4830/6830; PBSB.5201.01

Lecture 17: Logistic regression II

Jason Mezey
jgm45@cornell.edu
April 13, 2017 (T) 8:40-9:55
Announcements

• Project will be available later today (!!)
Conceptual Overview

Genetic System

Does A1 -> A2 affect Y?

Measured individuals (genotype, phenotype)

Reject / DNR

Regression model

Model params F-test

Sample or experimental pop

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

• For a logistic regression, we use the logistic function and the error term makes up the value to either 0 or 1:

![Graphs showing linear and logistic regression]
Review: calculating the components of an individual I

- Recall that an individual with phenotype Yi 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 Yi and a genotype Xi breaks down in this equation, we need to consider the expected (predicted!) part and the error term (we will do this separately
For example, say we have an individual i that has genotype A1A1 and phenotype Yi = 0

We know Xa = -1 andXd = -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(Yi|Xi) 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 $AIAI$ 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: the error term 1

- 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:

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

$$
p = 0.1
$$
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:

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

- 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:

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

$$p = 0.6$$
Review: 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:

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

\[
E(y | x) = \gamma^{-1}(x\beta) = \left[ \begin{array}{c}
\frac{e^{\beta_{\mu} + x_1, a \beta a + x_1, d \beta d}}{1 + e^{\beta_{\mu} + x_1, a \beta a + x_1, d \beta d}} \\
\vdots \\
\frac{e^{\beta_{\mu} + x_n, a \beta a + x_n, d \beta d}}{1 + e^{\beta_{\mu} + x_n, a \beta a + x_n, d \beta d}}
\end{array} \right]
\]
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 is R?)
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.
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):

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

- 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
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 \( \beta \)'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.
Step 1: IRLS algorithm

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 $\beta^{[0]}$.

   • 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: 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!):

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

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

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

$$
\beta^{[t]} = \begin{bmatrix}
\beta^{[t]}_\mu \\
\beta^{[t]}_a \\
\beta^{[t]}_d
\end{bmatrix}
$$

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

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

(W_{ij} = 0 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 $\ln(0)$!!:

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

$$D = 2 \sum_{i=1}^{n} \left[ y_i \ln \left( \frac{y_i}{\gamma^{-1}(\beta^{[t+1]}|x_i) + x_i, a \beta_a + x_i, d \beta_d} \right) + (1 - y_i) \ln \left( \frac{1 - y_i}{1 - \gamma^{-1}(\beta^{[t+1]}|x_i) + x_i, a \beta_a + x_i, d \beta_d} \right) \right]$$

$$D = 2 \sum_{i=1}^{n} \left[ y_i \ln \left( \frac{y_i}{e^{\beta^{[t+1]}|x_i} + x_i, a \beta_a + x_i, d \beta_d} \right) + (1 - y_i) \ln \left( \frac{1 - y_i}{1 - e^{\beta^{[t+1]}|x_i} + x_i, a \beta_a + x_i, d \beta_d} \right) \right]$$
Logistic hypothesis testing I

• Recall that our null and alternative hypotheses are:

\[ H_0 : \beta_a = 0 \cap \beta_d = 0 \]

\[ H_A : \beta_a \neq 0 \cup \beta_d \neq 0 \]

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

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

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

• For our case, we need the following:

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

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

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

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

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

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

• where we use the same IRLS algorithm to provide estimates of 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. \( n \to \infty \):

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

\[
LRT \to \chi^2_{df}
\]

• 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!
Logistic 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 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

    $$Pr(Y|X) \sim \text{exp family}$$

  • 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:

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

• 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:

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

- 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):

- 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:

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

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

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

\[ V \text{ar}(\epsilon) = f(X\beta) = \sigma^2_\epsilon \]

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

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

\[ V \text{ar}(\epsilon_i) = \gamma^{-1}(\beta_\mu + X_{i,a}\beta_a + X_{i,d}\beta_d)(1 - \gamma^{-1}(\beta_\mu + X_{i,a}\beta_a + X_{i,d}\beta_d)) \]
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

• See you on Tues.!