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

Lecture 17: Logistic regression I

Jason Mezey
jgm45@cornell.edu
March 27, 2018 (T) 8:40-9:55
Announcements

• MIDTERM EXAM (!!!) will be available on CMS by 11AM TODAY (!!) - see next slides

• Reminders:
  • No lecture Thurs. (March 29, = cancelled) and No lecture next week (April 3, April 5 = Ithaca Spring Break)
  • No office hours this week (March 29 = cancelled) and No office hours next week (April 5 = Spring Break)
  • No computer lab this week Thurs. (March 29 = cancelled) and no lab next week (April 5 = Ithaca Spring Break)

• All lectures, office hours, computer labs resume the week of April 9!
Announcements

Quantitative Genomics and Genetics - Spring 2018
BTRY 4830/6830; PBSB 5201.01

Midterm - available online 11:59PM, Tues., March 27

For midterm exam, due before 11:59PM, Fri., March 30

PLEASE NOTE THE FOLLOWING INSTRUCTIONS:

1. You are to complete this exam alone. The exam is open book, so you are allowed to use any books or information available online, your own notes and your previously constructed code, etc. HOWEVER YOU ARE NOT ALLOWED TO COMMUNICATE OR IN ANY WAY ASK ANYONE FOR ASSISTANCE WITH THIS EXAM IN ANY FORM (the only exceptions are Manisha, Zijun, and Dr. Mezey). As a non-exhaustive list this includes asking classmates or ANYONE else for advice or where to look for answers concerning problems, you are not allowed to ask anyone for access to their notes or to even look at their code whether constructed before the exam or not, etc. You are therefore only allowed to look at your own materials and materials you can access on your own. In short, work on your own! Please note that you will be violating Cornell’s honor code if you act otherwise.

2. Please pay attention to instructions and complete ALL requirements for ALL questions, e.g. some questions ask for R code, plots, AND written answers. We will give partial credit so it is to you advantage to attempt every part of every question.

3. A complete answer to this exam will include R code answers in Rmarkdown, where you will submit your .Rmd script and associated .pdf file. Note there will be penalties for scripts that fail to compile (!!). Also, as always, you do not need to repeat code for each part (i.e., if you write a single block of code that generates the answers for some or all of the parts, that is fine, but do please label your output that answers each question!!). You should include all of your plots and written answers in this same .Rmd script with your R code.

4. The exam must be uploaded on CMS before 11:59PM Fri., March 30. It is your responsibility to make sure that it is in uploaded by then and no excuses will be accepted (power outages, computer problems, Cornell’s internet slowed to a crawl, etc.). Remember: you are welcome to upload early! We will deduct points for being late for exams received after this deadline (even if it is by minutes!!).
Summary of lecture 17

- Last lecture, we completed our discussion of covariate modeling with our discussion of population structure
- Today, we will begin our discussion of logistic regression
Review: linear regression

- So far, we have considered a linear regression is a reasonable model for the relationship between genotype and phenotype (where this implicitly assumes a normal error provides a reasonable approximation of the phenotype distribution given the genotype):

\[
Y = \beta\mu + X_a\beta_a + X_d\beta_d + \epsilon \quad \epsilon \sim N(0, \sigma_\epsilon^2)
\]
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 $Y(\text{control}) = 0$. 
Let’s contrast the situation, let’s contrast data we might model with a linear regression model versus case / control data:
Case / Control Phenotypes II

- Let’s contrast the situation, let’s contrast data we might model with a linear regression model versus case / control data:
Logistic regression I

- Instead, we’re going to consider a logistic regression model
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 our approach to inference is the same with this model
Logistic regression III

- To begin, let’s consider the structure of a regression model:
  \[ Y = \text{logistic}(\beta_\mu + X_a \beta_a + X_d \beta_d) + \epsilon_l \]

- 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:
  \[ E(Y_i|X_i) = \text{logistic}(\beta_\mu + X_{i,a} \beta_a + X_{i,d} \beta_d) \]

- We can similarly write for a population using matrix notation (where the \( X \) matrix has the same form as we have been considering!):
  \[ E(Y|X) = \text{logistic}(X\beta) \]

- In fact the two major differences are in the form of the error and the logistic function
Logistic regression: error term I

• Recall that for a linear regression, the error term accounted for the difference between each point and the expected value (the linear regression line), which we assume follow a normal, but for a logistic regression, we have the same case but the value has to make up the value to either 0 or 1 (what distribution is this?):
Logistic regression: error term II

• 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 $Y = 0$

$$
\epsilon_i = -E(Y_i | X_i) = -E(Y | A_i A_j) = -\text{logistic}(\beta_\mu + X_{i,a} \beta_a + X_{i,d} \beta_d)
$$

• For $Y = 1$

$$
\epsilon_i = 1 - E(Y_i | X_i) = 1 - E(Y | A_i A_j) = 1 - \text{logistic}(\beta_\mu + X_{i,a} \beta_a + X_{i,d} \beta_d)
$$
Logistic regression: error term II

- 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 \( Y = 0 \)
    \[
    \epsilon_i = -E(Y_i | X_i) = -E(Y | A_i A_j) = -\text{logistic}(\beta_\mu + X_i,a \beta_a + X_i,d \beta_d)
    \]

  - For \( Y = 1 \)
    \[
    \epsilon_i = 1 - E(Y_i | X_i) = 1 - E(Y | A_i A_j) = 1 - \text{logistic}(\beta_\mu + X_i,a \beta_a + X_i,d \beta_d)
    \]

- For a distribution that takes two such values, a reasonable distribution is therefore the Bernoulli distribution with the following parameter

  \[
  \epsilon_i = Z - E(Y_i | X_i)
  \]
Logistic regression: error term II

• 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 $Y = 0$

    \[
    \epsilon_i = -E(Y_i | X_i) = -E(Y | A_i A_j) = -\text{logistic}(\beta_\mu + X_i,a \beta_a + X_i,d \beta_d)
    \]

  • For $Y = 1$

    \[
    \epsilon_i = 1 - E(Y_i | X_i) = 1 - E(Y | A_i A_j) = 1 - \text{logistic}(\beta_\mu + X_i,a \beta_a + X_i,d \beta_d)
    \]

• For a distribution that takes two such values, a reasonable distribution is therefore the Bernoulli distribution with the following parameter

  \[
  \epsilon_i = Z - E(Y_i | X_i)
  \]

  \[
  Pr(Z) \sim \text{bern}(p) \quad p = \text{logistic}(\beta_\mu + X_a \beta_a + X_d \beta_d)
  \]
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!):

\[
\epsilon_i = Z - E(Y_i | X_i)
\]

\[
Pr(Z) \sim \text{bern}(p)
\]

\[
p = \text{logistic}(\beta_\mu + X_a \beta_a + X_d \beta_d)
\]
Logistic regression: link function I

- 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 ANDXd!!):

\[
E(Y_i \mid X_i) = \text{logistic}(\beta_\mu + X_{i,a}\beta_a + X_{i,d}\beta_d)
\]

\[
E(Y_i \mid X_i) = \frac{e^{\beta_\mu + X_{i,a}\beta_a + X_{i,d}\beta_d}}{1 + e^{\beta_\mu + X_{i,a}\beta_a + X_{i,d}\beta_d}}
\]
Calculating the components of an individual II

- 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 \( E(Y_i|X_i) \) and the error term for i:

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

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

\[ 0 = 0.1 - 0.1 \]
Calculating the components of an individual III

- For example, say we have an individual i that has genotype A1A1 and phenotype Yi = 1
- 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:

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

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

\[
1 = 0.1 + 0.9
\]
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 \( \text{E}(Y_i|X_i) \) and the error term for i:

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

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

\[
0 = 0.6 - 0.6
\]
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 $X_d = -1$
- 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 $i$:
  \[ Y_i = \frac{e^{\beta_\mu + x_i,a \beta_a + x_i,d \beta_d}}{1 + e^{\beta_\mu + x_i,a \beta_a + x_i,d \beta_d}} + \epsilon_i \]
  \[ 0 = \frac{e^{0.2 + (1)2.2 + (-1)0.2}}{1 + e^{0.2 + (1)2.2 + (-1)0.2}} + \epsilon_i \]
  \[ 0 = 0.9 - 0.9 \]
For the entire probability distributions 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:

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

- 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:

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

\[
p = 0.1
\]
For the entire probability distributions 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

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

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

For example:

\[
\epsilon_i = -0.6 \quad \epsilon_i = 0.4
\]

\[
p = 0.6
\]
For the entire probability distributions 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

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

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

For example:

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

$$p = 0.9$$
That’s it for today

• See you April 10!