Lecture 20: Logistic Regression 1

Jason Mezey
April 12, 2022 (T) 8:05-9:20
<table>
<thead>
<tr>
<th>Date</th>
<th>Topic</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>April 5</td>
<td><strong>No Class!!</strong></td>
<td></td>
</tr>
<tr>
<td>April 7</td>
<td><strong>No Class!!</strong></td>
<td></td>
</tr>
<tr>
<td>April 12</td>
<td>Genome-Wide Association Studies (GWAS): logistic regression I</td>
<td>12</td>
</tr>
<tr>
<td>April 14</td>
<td><strong>Project Assigned</strong></td>
<td>Genome-Wide Association Studies (GWAS): logistic regression II (IRLS algorithm)</td>
</tr>
<tr>
<td>April 19</td>
<td><strong>No Class!!</strong></td>
<td></td>
</tr>
<tr>
<td>April 21</td>
<td><strong>No Class!!</strong></td>
<td></td>
</tr>
<tr>
<td>April 26</td>
<td>Advanced topics IV: Mixed Models</td>
<td></td>
</tr>
<tr>
<td>April 28</td>
<td><strong>BAYESIAN STATISTICS</strong></td>
<td>Bayesian inference I: introduction &amp; inference basics and linear models</td>
</tr>
<tr>
<td>May 3</td>
<td>Bayesian inference II: MCMC algorithms</td>
<td></td>
</tr>
<tr>
<td>May 5</td>
<td>PEDIGREE AND INBRED LINE DESIGNS</td>
<td>Basics of linkage analysis &amp; Inbred line analysis</td>
</tr>
<tr>
<td>May 10</td>
<td><strong>CLASSIC QUANTITATIVE GENOMICS</strong></td>
<td>Additive genetic variance and heritability</td>
</tr>
<tr>
<td></td>
<td>Project Due</td>
<td></td>
</tr>
</tbody>
</table>
• Last lecture, we discussed some (optional topic!) concepts in population genetics relevant for understanding GWAS including a discussion of population structure and PCA (and how PCs are often used as a covariate in GWAS analysis)

• Today, we will begin our discussion of the last major (non-optional!) topic: 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)
Linear regression review

- 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 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 $Y(\text{control}) = 0$. 
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:
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 overall 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:

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

• We can similarly write for a population using matrix notation (where the X matrix has the same form as we have been considering!):

\[ \mathbb{E}(Y|X) = \text{logistic}(X\beta) \]

• In fact the two major differences are in the form of the error and the logistic function.
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?):
That’s it for today

• Next lecture: logistic regression II (including the IRLS algorithm!)