Lecture 12: Genetic association testing and introduction to genome-wide association studies

Jason Mezey
jgm45@cornell.edu
March 12, 2019 (T) 10:10-11:24
Announcements

• Homeworks:
  • Homework #3 graded
  • Homework #4 key up this week
  • Homework #5 assigned (due: 11:59PM, March 20)

• Office hours:
  • This Fri. March 15, 2-4PM
  • This coming Mon. (March 18), 2-4PM

• New Schedule posted
  • For homework due dates: the date on the homework (or announcement updates) takes precedence
  • No lecture: March 25, April 2, April 4
  • Midterm: week of April 8 (after Spring break!)

• Supplemental materials on CMS: take a look!
Summary of lecture 12

• Last lecture, we introduced the genetic linear regression model

• Today we will discuss estimation and more importantly hypothesis testing for this model

• We will also provide a rigorous introduction to the statistical foundation of Genome-wide Association Study (GWAS) analysis (!!!)
Conceptual Overview

Genetic System

Does A1 → A2 affect Y?

Sample or experimental pop

Pr(Y|X)

Model params
F-test

Measured individuals (genotype, phenotype)

Regression model

Reject / DNR
Review: genetic system

• Our goal in quantitative genetics / genomics is to identify loci (positions in the genome) that contain causal mutations / polymorphisms / alleles

• **causal mutation** or **polymorphism** - a position in the genome where an experimental manipulation of the DNA produces an effect on the phenotype on average or under specified conditions

• Formally, we may represent this as follows:

\[
A_1 \rightarrow A_2 \Rightarrow \Delta Y | Z
\]

• Our experiment will be a statistical experiment (sample and inference!)
Review: genetic inference

- For our model focusing on one locus:
  \[ Y = \beta_\mu + X_a\beta_a + X_d\beta_d + \epsilon \]
  \[ \epsilon \sim N(0, \sigma_\epsilon^2) \]

- We have four possible parameters we could estimate:
  \[ \theta = [\beta_\mu, \beta_a, \beta_d, \sigma_\epsilon^2] \]

- However, for our purposes, we are only interested in the genetic parameters and testing the following null hypothesis:
  \[ H_0 : \text{Cov}(X_a, Y) = 0 \cap \text{Cov}(X_d, Y) = 0 \]
  \[ H_A : \text{Cov}(X_a, Y) \neq 0 \cup \text{Cov}(X_d, Y) \neq 0 \]
  OR
  \[ H_0 : \beta_a = 0 \cap \beta_d = 0 \]
  \[ H_A : \beta_a \neq 0 \cup \beta_d \neq 0 \]
Review: genetic inference II

- Recall that inference (whether estimation or hypothesis testing) starts by collecting a sample and defining a statistic on that sample.

- In this case, we are going to collect a sample of $n$ individuals where for each we will measure their phenotype and their genotype (i.e. at the locus we are focusing on).

- That is an individual $i$ will have phenotype $y_i$ and genotype $g_i = A_jA_k$ (where we translate these into $x_a$ and $x_d$).

- Using the phenotype and genotype we will construct both an estimator (a statistic!) and we will additionally construct a test statistic.

- Remember that our regression probability model defines a sampling distribution on our sample and therefore on our estimator and test statistic (!!).
Review: genetic inference III

- For notation convenience, we are going to use vector / matrix notation to represent a sample:

\[ y_i = \beta_\mu + x_{i,a}\beta_a + x_{i,d}\beta_d + \epsilon_i \]

\[
\begin{bmatrix}
  y_1 \\
  y_2 \\
  \vdots \\
  y_n
\end{bmatrix}
= \begin{bmatrix}
  \beta_\mu + x_{1,a}\beta_a + x_{1,d}\beta_d + \epsilon_1 \\
  \beta_\mu + x_{2,a}\beta_a + x_{2,d}\beta_d + \epsilon_2 \\
  \vdots \\
  \beta_\mu + x_{n,a}\beta_a + x_{n,d}\beta_d + \epsilon_n
\end{bmatrix}
\]

\[
\begin{bmatrix}
  y_1 \\
  y_2 \\
  \vdots \\
  y_n
\end{bmatrix}
= \begin{bmatrix}
  1 & x_{1,a} & x_{1,d} \\
  1 & x_{2,a} & x_{2,d} \\
  \vdots & \vdots & \vdots \\
  1 & x_{n,a} & x_{n,d}
\end{bmatrix}
\begin{bmatrix}
  \beta_\mu \\
  \beta_a \\
  \beta_d
\end{bmatrix}
+ \begin{bmatrix}
  \epsilon_1 \\
  \epsilon_2 \\
  \vdots \\
  \epsilon_n
\end{bmatrix}
\]

\[ \mathbf{y} = \mathbf{x}\beta + \epsilon \]
Genetic estimation I

- We will define a MLE for our parameters:
  \[ \beta = [\beta_{\mu}, \beta_a, \beta_d] \]

- Recall that an MLE is simply a statistic (a function that takes a sample in and outputs a number that is our estimate)

- In this case, our statistic will be a vector valued function that takes in the vectors that represent our sample
  \[ T(y, x_a, x_d) = \hat{\beta} = [\hat{\beta}_{\mu}, \hat{\beta}_a, \hat{\beta}_d] \]

- Note that we calculate an MLE for this case just as we would any case (we use the likelihood of the fixed sample where we identify the parameter values that maximize this function)

- In the linear regression case (just as with normal parameters) this has a closed form:
  \[ MLE(\hat{\beta}) = (x^T x)^{-1} x^T y \]
Genetic estimation II

Let’s look at the structure of this estimator:

\[ y = x\beta + \epsilon \]

\[
\begin{bmatrix}
    y_1 \\
    y_2 \\
    \vdots \\
    y_n
\end{bmatrix} =
\begin{bmatrix}
    1 & x_{1,a} & x_{1,d} \\
    1 & x_{2,a} & x_{2,d} \\
    \vdots & \vdots & \vdots \\
    1 & x_{n,a} & x_{n,d}
\end{bmatrix}
\begin{bmatrix}
    \beta_\mu \\
    \beta_a \\
    \beta_d
\end{bmatrix}
+ 
\begin{bmatrix}
    \epsilon_1 \\
    \epsilon_2 \\
    \vdots \\
    \epsilon_n
\end{bmatrix}
\]

\[ MLE(\hat{\beta}) = (x^T x)^{-1} x^T y \]

\[ MLE(\hat{\beta}) = \begin{bmatrix} \hat{\beta}_\mu \\ \hat{\beta}_a \\ \hat{\beta}_d \end{bmatrix} \]
Genetic hypothesis testing I

- We are going to test the following hypothesis:
  \[ H_0 : \beta_a = 0 \cap \beta_d = 0 \]
  \[ H_A : \beta_a \neq 0 \cup \beta_d \neq 0 \]

- To do this, we need to construct the following test statistic (for which we know the distribution!):
  \[ T(y, x_a, x_d | H_0 : \beta_a = 0 \cap \beta_d = 0) \]

- Specifically, we are going to construct a likelihood ratio test (LRT)

- This is calculated using the same structure that we have discussed (i.e. ratio of likelihoods that take values of parameters maximized under the null and alternative hypothesis)

- In the case of a regression (not all cases!) we can write the form of the LRT for our null in an alternative (but equivalent!) form

- In addition, our LRT has an exact distribution for all sample sizes \( n \) (!!)
We now have everything we need to construct a hypothesis test for:

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

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

This is equivalent to testing the following:

\[ H_0 : Cov(X, Y) = 0 \]

For a linear regression, we use the F-statistic for our sample:

\[
F_{[2, n-3]}(y, x_a, x_d) = \frac{MSE}{MSE}
\]

We then determine a p-value using the distribution of the F-statistic under the null:

\[
pval(F_{[2, n-3]}(y, x_a, x_d))
\]
Genetic hypothesis testing III

- To construct our LRT for our null, we will need several components, first the predicted value of the phenotype for each individual:

\[
\hat{y}_i = \hat{\beta}_\mu + x_{i,a}\hat{\beta}_a + x_{i,d}\hat{\beta}_d
\]

- Second, we need the “Sum of Squares of the Model” (SSM) and the “Sum of Squares of the Error” (SSE):

\[
SSM = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \quad SSE = \sum_{n=1}^{n} (y_i - \hat{y}_i)^2
\]

- Third, we need the “Mean Squared Model” (MSM) and the “Mean Square Error” (MSE) with degrees of freedom (df) such that:

\[
df(M) = 3 - 1 = 2 \quad df(E) = n - 3
\]

\[
MSM = \frac{SSM}{df(M)} = \frac{SSM}{2} \quad MSE = \frac{SSE}{df(E)} = \frac{SSE}{n - 3}
\]

- Finally, we calculate our (LRT!) statistic, the F-statistic with degrees of freedom [2, n-3]:

\[
F_{[2,n-3]} = \frac{MSM}{MSE}
\]
Genetic hypothesis testing IV

In general, the F-distribution (continuous random variable!) under the H0 has variable forms that depend on d.f.:

Note when calculating a p-value for the genetic model, we consider the value of the F-statistic we observe or more extreme towards positive infinite (!!) using the F-distribution with $[2,n=3]$ d.f.

However, also this is actually a two-tailed test (what is going on here (!?)
Genetic hypothesis testing

- An F-statistic is a Likelihood Ratio Test (LRT) statistic after a simple (monotonic) transformation

  \[ F\text{-statistic} = f(\Lambda) \]

- Note that an F-statistic has an exact pdf under many conditions (note that we do not always produce a LRT that has an exact pdf that we can state easily)

- Also note that a t-test is actually an F-statistic (and therefore a transformed LRT) for a case where we are comparing the means of just two groups (when might this apply in genetic testing!?), similarly for a t-test of the slope of a regression)
Side-topic: Alternative (ANOVA) formulation I

- Note that we can construct an equivalent formulation to our linear regression using an ANOVA coding.
- ANOVA stands for ANalysis Of VAriance and, despite the name, it is really a test of whether “means” of groups are different.
- A genetic ANOVA model is the same as our linear regression, except the “dummy” variables are coded differently (everything else is the same!)
Side-topic: Alternative (ANOVA) formulation II

- Remember the independent (dummy) variable coding for a regression is:

\[ X_\mu(A_1A_1) = 1, X_\mu(A_1A_2) = 1, X_\mu(A_2A_2) = 1 \]
\[ X_a(A_1A_1) = -1, X_a(A_1A_2) = 0, X_a(A_2A_2) = 1 \]
\[ X_d(A_1A_1) = -1, X_d(A_1A_2) = 1, X_d(A_2A_2) = -1 \]

- The ANOVA coding is the following:

\[ X_{A_1A_1}(A_1A_1) = 1, X_{A_1A_1}(A_1A_2) = 0, X_{A_1A_1}(A_2A_2) = 0 \]
\[ X_{A_1A_2}(A_1A_1) = 0, X_{A_1A_2}(A_1A_2) = 1, X_{A_1A_2}(A_2A_2) = 0 \]
\[ X_{A_2A_2}(A_1A_1) = 0, X_{A_2A_2}(A_1A_2) = 0, X_{A_2A_2}(A_2A_2) = 1 \]

- The models corresponding to a linear regression and ANOVA are:

\[ Y = X_\mu \beta_\mu + X_a \beta_a + X_d \beta_d + \epsilon \]
\[ Y = X_{A_1A_1} \beta_{A_1A_1} + X_{A_1A_2} \beta_{A_1A_2} + X_{A_2A_2} \beta_{A_2A_2} + \epsilon \]
Side-topic: Alternative (ANOVA) formulation III

- For the ANOVA formulation, the parameters are:
  \[ \theta = [\beta_{A_1A_1}, \beta_{A_1A_2}, \beta_{A_2A_2}] \]

- And we test the null hypothesis:
  \[ H_0 : \beta_{A_1A_1} = \beta_{A_1A_2} = \beta_{A_2A_2} \]
  \[ H_A : \beta_{A_jA_k} \neq \beta_{A_lA_m} \quad jk \neq lm \]

- Note that estimation (MLE) and the hypothesis test (F-test) construction are the same (=same equations)!!

- Why would we use an ANOVA formulation (what is the difference)?
We now know how to assess the null hypothesis as to whether a polymorphism has a causal effect on our phenotype.

Occasionally we will assess this hypothesis for a single genotype.

In quantitative genomics, we generally do not know the location of causal polymorphisms in the genome.

We therefore perform a hypothesis test of *many genotypes throughout the genome*.

This is a genome-wide association study (GWAS).
Quantitative genomic analysis II

• Analysis in a GWAS raises (at least) two issues we have not yet encountered:
  • An analysis will consist of many hypothesis tests (not just one)
  • We often do not test the causal polymorphism (usually)
  • Note that this latter issue is a bit strange (!?) - how do we assess causal polymorphisms if we have not measured the causal polymorphism?
  • Also note that causal genotypes will begin to be measured in our GWAS with next-generation sequencing data (but the issue will still be present!)
Correlation among genotypes

• If we test a (non-causal) genotype that is correlated with the causal genotype AND if correlated genotypes are in the same position in the genome THEN we can identify the genomic position of the casual genotype (!!)

• This is the case in genetic systems (why!?)

• Do we know which genotype is causal in this scenario?
Linkage Disequilibrium (LD)

- Mapping the position of a causal polymorphism in a GWAS requires there to be LD for genotypes that are both physically linked and close to each other AND that markers that are either far apart or on different chromosomes to be in equilibrium.

- Note that disequilibrium includes both linkage disequilibrium AND other types of disequilibrium (!!), e.g. gametic phase disequilibrium.
That’s it for today

- Next lecture: continued discussion of population genetics, GWAS statistical, analysis, and interpretation issues!