Quantitative Genomics and Genetics
BTRY 4830/6830; PBSB.5201.03

Lecture 16: Introduction to Genetics III

Jason Mezey
jgm45@cornell.edu
April 8, 2021 (Th) 8:05-9:20
Announcements

• Homework #5 is posted

  • LAST HOMEWORK (!!)

  • Due April 11:59PM (ET) April 16 (Fri)

• Critical material presented today and in computer labs (and some Tues. / Thurs. lectures)

• Note that a “Matrix Basics” Supplemental Reading (#3) has been posted
Summary of lecture 16

- Today we will discuss INFERENCES for the genetic linear regression model!
Conceptual Overview

Genetic System

Does A1 -> A2 affect Y?

Sample or experimental pop

Measured individuals (genotype, phenotype)

Regression model

Pr(Y|X)

Reject / DNR

Model params F-test
Review: Causal Mutation

- **causal mutation** - a position in the genome where an experimental manipulation of the DNA would produce an effect on the phenotype under specifiable conditions

- Formally, we may represent this as follows:

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

- Note: that this definition considers “under specifiable” conditions” so the change in genome need not cause a difference under every manipulation (just under broadly specifiable conditions)

- Also note the symmetry of the relationship

- Identifying these is the core of quantitative genetics/genomics (why do we want to do this!?)

- What is the perfect experiment?

- Our experiment will be a statistical experiment (sample and inference!)
Review: The genetic probability model I

- The quantitative genetic model is a multiple regression model:

\[ Y = \beta_\mu + X_a \beta_a + X_d \beta_d + \epsilon \]

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

- with the following independent ("dummy") variables:

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

\[ X_d(A_1A_1) = -1, \ X_d(A_1A_2) = 1, \ X_d(A_2A_2) = -1 \]

\[
\begin{array}{c|cc}
1 & A_1A_2 \\
-1 & A_1A_1 & A_2A_2 \\
\hline
& -1 & 0 & 1
\end{array}
\]
Review: The genetic probability model IV

- Note that, while somewhat arbitrary, the advantage of the Xa and Xd coding is the parameters $\beta_a$ and $\beta_d$ map directly on to relationships between the genotype and phenotype that are important in genetics:

- If $\beta_a \neq 0, \beta_d = 0$ then this is a “purely” additive case

- If $\beta_a = 0, \beta_d \neq 0$ then this is only over- or under-dominance (homozygotes have equal effects on phenotype)

- If both are non-zero, there are both additive and dominance effects

- If both are zero, there is no effect of the genotype on the phenotype (the genotype is not causal!)
Review: Genetic example 1

- As an example, consider the following of a “purely additive” case (= no dominance): \( \beta_\mu = 2, \beta_a = 5, \beta_d = 0, \sigma_\varepsilon^2 = 1 \)
Review: Genetic example II

- An example of “dominance” (= not a “pure additive” case):

\[ \beta_\mu = 0, \beta_a = 4, \beta_d = -1, \sigma^2_\epsilon = 1 \]
A case of NO genetic effect:

\[ \beta_\mu = 2, \beta_a = 0, \beta_d = 0, \sigma_\epsilon^2 = 1 \]
Quantitative genetic formalism

- For those of you who have been exposed to classic quantitative genetics, you have seen a different notation for this model:

\[ P = G + E \]

- \( P \) is the **phenotypic value** - the value of the aspect measured

- \( G \) is the **genotypic value** - the expected value of the phenotype conditional on the genotype

- \( E \) is the **environmental value** - the value of the phenotype that we cannot explain given the genotype

- These translate as follows for our one locus case (although note the formalism extends to any multiple locus case):

\[ Y = P \]

\[ G = EP = EY = \beta_{\mu} + X_a\beta_a + X_d\beta_d \]

\[ \epsilon = E \]
Genetic inference I

- 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:
  \[ H_0 : Cov(X_a, Y) = 0 \cap Cov(X_d, Y) = 0 \]
  \[ H_A : Cov(X_a, Y) \neq 0 \cup Cov(X_d, Y) \neq 0 \]
  \[ H_0 : \beta_a = 0 \cap \beta_d = 0 \]
  \[ H_A : \beta_a \neq 0 \cup \beta_d \neq 0 \]
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 (!!!)
Genetic inference III

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

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

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

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

\[y = x\beta + \epsilon\]
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 \( n \) (!!)
Genetic hypothesis testing II

- 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:

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

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

- Third, we need the “Mean Squared Model” (MSM) and the “Mean Square Error” (MSE) with degrees of freedom (df):
  \[
  df(M) = 3 - 1 = 2 \quad \text{and} \quad df(E) = n - 3
  \]
  \[
  MSM = \frac{SSM}{df(M)} = \frac{SSM}{2} \quad \text{MSE} = \frac{SSE}{df(E)} = \frac{SSE}{n - 3}
  \]

- 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.:

![F Distribution PDF](image)

- 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 (!?)
That’s it for today

• Next lecture: quantitative genomics I (!!)