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

Lecture I I: Quantitative Genomics I I

Jason Mezey
jgm45@cornell.edu
March 7, 2017 (T) 8:40-9:55
Announcements

- Lecture 10 will be posted today
- We will post some supplementary materials (i.e., matrix cheat sheet)
- Homework #4 is due March 15 - NOT tomorrow
Summary of lecture 11

- Last lecture, we introduced the quantitative genetic model = a (linear) regression
- Today, we will continue our discussion, including how to perform inference with this model (estimation and hypothesis testing)
Conceptual Overview

Genetic System

Does A1 $\rightarrow$ A2 affect Y?

Sample or experimental pop

Measured individuals (genotype, phenotype)

Regression model

Pr(Y|X)

Reject / DNR

Model params

F-test
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 under specified conditions

• Formally, we may represent this as follows:

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

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

- As with any statistical experiment, we need to begin by defining our sample space.

- In the most general sense, our sample space is:

\[ \Omega = \{ \text{Possible Individuals} \} \]

- More specifically, each individual in our sample space can be quantified as a pair of sample outcomes so our sample space can be written as:

\[ \Omega = \{ \Omega_g \cap \Omega_P \} \]

- Where \( \Omega_g \) is the genotype sample space at a locus and \( \Omega_P \) is the phenotype sample space.

- Note that genotype \( g_i = A_j A_k \) is the set of possible genotypes, where for a diploid, with two alleles:

\[ \Omega_g = \{ A_1 A_1, A_1 A_2, A_2 A_2 \} \]

- For the phenotype, this can be any type of measurement (e.g. sick or healthy, height, etc.)
Review: the statistical model

- Next, we need to define the probability model on the sigma algebra of the sample space \( (\mathcal{F}_{g,P}) \):

\[
Pr(\mathcal{F}_{g,P})
\]

- Which defines the probability of each possible genotype and phenotype pair:

\[
Pr\{g, P\}
\]

- We will define two (types) or random variables (* = state does not matter):

\[
Y : (\ast, \Omega_P) \rightarrow \mathbb{R}
\]

\[
X : (\Omega_g, \ast) \rightarrow \mathbb{R}
\]

- Note that the probability model induces a (joint) probability distribution on this random vector (these random variables):

\[
Pr(Y, X)
\]
Review: looking ahead (the goal...)

- The goal of quantitative genomics and genetics is to identify cases of the following relationship where when performing the “perfect” genotype manipulation experiment we have:

\[ Pr(Y \cap X) = Pr(Y, X) \neq Pr(Y)Pr(X) \]

- Remember that, regardless of the probability distribution of our random vector, we can define the expectation:

\[ E[Y, X] = [EY, EX] \]

- and the variance:

\[ Var[Y, X] = \begin{bmatrix} Var(Y) & Cov(Y, X) \\ Cov(Y, X) & Var(X) \end{bmatrix} \]

- The goal of quantitative genomics can be rephrased as assessing the following relationship in the “perfect” experimental framework (although in practice, we will do this by assessing the following relationship in an uncontrolled setting):

\[ Cov(Y, X) \neq 0 \]
Review: linear regression

- We are going to consider a regression model a parameterized model to represent the probability model of $X$ and $Y$ (that is the true statistical model of genetics!!!):

\[ Y = \beta_0 + X \beta_1 + \epsilon \]

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

- Note that in this model, we consider $Y$ to be the dependent or response variable and $X$ to be the independent variable (what are the parameters!?)

- Also note implicitly assumes:

\[ Pr(Y, X) = Pr(Y|X) \]

- That is, that $X$ is effectively “fixed” when considering an individual, although note that $X$ still varies (has a probability)
Linear regression is a bivariate distribution

- We’ve seen bivariate (multivariate) distributions before:
Linear regression I

Let’s review the structure of a linear regression (not necessarily a genetic model):

\[ Y = \beta_0 + X \beta_1 + \epsilon \quad \epsilon \sim N(0, \sigma^2_\epsilon) \]
**Review: genetic probability model**

- The quantitative genetic model is a multiple regression model with the following independent ("dummy") variables:

\[
X_a(A_1 A_1) = -1, \ X_a(A_1 A_2) = 0, \ X_a(A_2 A_2) = 1
\]
\[
X_d(A_1 A_1) = -1, \ X_d(A_1 A_2) = 1, \ X_d(A_2 A_2) = -1
\]

<table>
<thead>
<tr>
<th>1</th>
<th>(A_1 A_2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1</td>
<td>(A_1 A_1)</td>
</tr>
<tr>
<td>-1</td>
<td>0</td>
</tr>
<tr>
<td>-1</td>
<td>1</td>
</tr>
</tbody>
</table>

- and the following "multiple" regression equation:

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

\[
\epsilon \sim N(0, \sigma_\epsilon^2)
\]
Review: the genetic statistical model

- The probability distribution of this model, is therefore:

\[ Pr(Y|X) \sim N(\beta_\mu + X_a\beta_a + X_d\beta_d, \sigma_\epsilon^2) \]

- Which has four parameters:

\[ \beta_\mu, \beta_a, \beta_d, \sigma_\epsilon^2 \]

- The three \( \beta \) parameters are required to model the three separate genotypes (A1A1, A1A2, A2A2)

- The \( \epsilon \) can be thought of as a random variable that describes the probability an individual will have a specific value of \( Y \), conditional on the genotype A1A1, where the probability is normally distributed around the value determined by the \( X \)'s and \( \beta \)'s

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

\[ Y = P \]

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

\[ \epsilon = E \]
The genetic probability model

- 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!)
Genetic example 1

- As an example, consider the following of a “purely additive” case (= no dominance): \( \beta_\mu = 2, \beta_a = 5, \beta_d = 0, \sigma_\epsilon^2 = 1 \)
An example of “dominance” (= not a “pure additive” case):

$$\beta_\mu = 0, \beta_a = 4, \beta_d = -1, \sigma^2_\epsilon = 1$$
Genetic example III

- A case of NO genetic effect:

\[ \beta_\mu = 2, \beta_a = 0, \beta_d = 0, \sigma^2_\epsilon = 1 \]
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:
  \[ H_0 : Cov(X_a, Y) = 0 \cap Cov(X_d, Y) = 0 \]
  \[ H_A : Cov(X_a, Y) \neq 0 \cup Cov(X_d, Y) \neq 0 \]
  OR
  \[ H_0 : \beta_a = 0 \cap \beta_d = 0 \]
  \[ H_A : \beta_a \neq 0 \cup \beta_d \neq 0 \]
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:
  \[ H_0 : \text{Cov}(X, Y) = 0 \]

- 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:

\[ \hat{y}_i = \beta_\mu + x_{i,a}\beta_a + x_{i,d}\beta_d \]

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

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

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

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

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

- 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 V

- 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)
That’s it for today

- Next lecture: population genetic principles critical for genome-wide association studies (GWAS)!