Lecture 11: Quantitative Genomics II

Jason Mezey
jgm45@cornell.edu
March 7, 2019 (Th) 10:10-11:25
Announcements

- Homework #5 will be posted by Sat (March 9) and you will have > 1 week to complete (Due: 11:59PM, Tues., March 19)

- Reminder: more posts coming (e.g., matrix basics!), computer labs today / tomorrow!
Summary of lecture 11

• Last lecture, we began our introduction of linear regression

• Today we will introduce estimation and hypothesis testing for the (genetic) linear regression, which will provide a rigorous introduction to the statistical foundation of Genome-wide Association Study (GWAS) analysis

• Next lecture, we will begin our introduction to GWAS
Conceptual Overview

Genetic System

Does A1 -> A2 affect Y?

Sample or experimental pop

Measured individuals (genotype, phenotype)

Reject / DNR

Regression model

Pr(Y|X)

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:

  \[ P(X = x) = P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = P_X(x) \text{ or } f_X(x) \]

  \[ \hat{p} = \frac{1}{n} \sum_{i=1}^{n} x_i \] (8)

  \[ \hat{\mu} = \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \] (9)

- \[ A_1 \rightarrow A_2 \Rightarrow \Delta Y \mid Z \] (10)

- 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):
  \[ Y : (*, \Omega_P) \rightarrow \mathbb{R} \]
  \[ X : (\Omega_g, *) \rightarrow \mathbb{R} \]

- 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):
  \[ Cov(Y, X) \neq 0 \]
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!!!):

$$Y = \beta_0 + X\beta_1 + \epsilon$$

$$\epsilon \sim N(0, \sigma^2_\epsilon)$$

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

\[ Y = \beta_0 + X\beta_1 + \epsilon \quad \epsilon \sim N(0, \sigma^2_\epsilon) \]
Linear regression I
Linear regression II

- The linear regression model allows calculation of the (interval) probability of observations (!!)

\[ Y = \beta_0 + X\beta_1 + \epsilon \quad \epsilon \sim N(0, \sigma^2) \]
A multiple regression model has the same structure, with a single dependent variable $Y$ and more than one independent variable $X_i, X_j$, e.g.,
The genetic probability model 1

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

\[ \begin{align*}
X_a(A_1A_1) &= -1, \; X_a(A_1A_2) = 0, \; X_a(A_2A_2) = 1 \\
X_d(A_1A_1) &= -1, \; X_d(A_1A_2) = 1, \; X_d(A_2A_2) = -1
\end{align*} \]

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- and the following "multiple" regression equation:

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

\[ \epsilon \sim N(0, \sigma_\epsilon^2) \]
The genetic probability model II

- 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 \( \text{AiA}_j \), where the probability is normally distributed around the value determined by the \( X \)'s and \( \beta \)'s

\[ \epsilon \sim N(0, \sigma_\epsilon^2) \]
The genetic probability model III

- Let’s consider a specific example where we are interested modeling the relationship between a genotype and a phenotype (such as height) where the latter is well approximated by a normal distribution.

- For this case, the (unknown) conditions of the experiment define the true values of the parameters (unknown to us!), which we will say are the following (note these are the same for all individuals in the population since they are parameters of the probability distribution):

  \[ \beta_\mu = 0.3, \beta_a = -0.2, \beta_d = 1.1, \sigma_\epsilon^2 = 1.1 \]

- Consider an individual \( i \) with \( g_i = A1A2 \) such that we have:

  \[ X_a(A1A2) = 0, \ X_d(A1A2) = 1 \]

- If this individual has a phenotype value \( y_i = 2.1 \) then we have the epsilon value \( \epsilon_i = 0.7 \) where the probability of this particular value (i.e. the interval surrounding this value) is defined by \( \epsilon \sim N(0, \sigma_\epsilon^2) \)

  \[ 2.1 = 0.3 + (0)(-0.2) + (1)(1.1) + 0.7 \]
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!)
As an example, consider the following of a “purely additive” case (= no dominance):

\[ y_i = \mu + X_{i,a}^a + x_{i,d}^d + \varepsilon \]

An intuitive way to consider this model, is to plot the phenotype \( Y \) on the Y-axis against the genotypes \( A_1A_1, A_1A_2, A_2A_2 \) on the X-axis for a sample (see class). We can represent all the individuals in our sample as points that are grouped in the three categories \( A_1A_1, A_1A_2, A_2A_2 \) and note that the true model would include points distributed in three normal distributions, with the means defined by the three classes \( A_1A_1, A_1A_2, A_2A_2 \).

If we were to then re-plot these points in two plots, \( Y \) versus \( X_a \) and \( Y \) versus \( X_d \), the first would look like the original plot, and the second would put the points in two groups (see class). The multiple linear regression equation (20, 21) defines 'two' regression lines (or more accurately a plane) for these latter two plots, where the slopes of the lines are \( a \) and \( d \) (see class). Note that \( \mu \) is where these two plots (the plane) intersect the Y-axis but with the way we have coded \( X_a \) and \( X_d \), this is actually an estimate of the overall mean of the population (hence the notation \( \mu \)).

\[
\begin{align*}
\mu &= 2, \\
\beta_a &= 5, \\
\beta_d &= 0, \\
\sigma^2 &= 1
\end{align*}
\]
Genetic example II

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

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

- 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 \textbf{phenotypic value} - the value of the aspect measured

• \( G \) is the \textbf{genotypic value} - the expected value of the phenotype conditional on the genotype

• \( E \) is the \textbf{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:
  \[ Y = \beta_\mu + X_a \beta_a + X_d \beta_d + \epsilon \]
  \[ \epsilon \sim N(0, \sigma^2_\epsilon) \]

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

- 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 $g_i = A_j A_k$ (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:

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

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

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

\[
MLE(\hat{\beta}) =
\begin{bmatrix}
\hat{\beta}_{\mu} \\
\hat{\beta}_{a} \\
\hat{\beta}_{d}
\end{bmatrix}
\]
That’s it for today

• Next lecture: genetic linear regression hypothesis testing and introduction to GWAS!