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

Lecture 13: Inference in Quantitative Genomics

Jason Mezey
jgm45@cornell.edu
March 10, 2020 (T) 8:40-9:55
Announcements

- COVID-19 DISRUPTIONS (!!):
  - We will be making decisions as officially University policies & other information become available
  - NYC - students cannot come to class until further notice, we are setting up a zoom option for you to join (!!)
  - Computer labs: we are still figuring this out… Stay tuned (!!)
  - We WILL be assigning another homework this week (#4) - stay tuned…
• Today, we will continue our introduction to Quantitative Genomics!

• Specifically, we will introduce inference for the core Quantitative Genomics model (=linear regression)!
Conceptual Overview

- System
- Experiment
- Question
- Sample
- Inference
- Prob. Models
- Statistics
- Assumptions
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)
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 statistical model IV

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

• Specifically, we will consider a regression model

• For the moment, let’s consider a regression model with normal error:

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

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

• 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 the following:

\[ Pr(Y, X) = Pr(Y|X) \]
Review: 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: Linear regression II

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

\[ Y = \beta_0 + X\beta_1 + \epsilon \quad \epsilon \sim N(0, \sigma_\epsilon^2) \]
Review: Linear regression III

- 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.,
Review: The genetic probability model I

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

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

\[
\begin{array}{c|ccc}
1 & A_1A_2 \\
-1 & A_1A_1 & A_2A_2 \\
\hline
& -1 & 0 & 1
\end{array}
\]

- 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 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 \( A1A_j \), where the probability is normally distributed around the value determined by the \( X \)'s and \( \beta \)'s
  \[ \epsilon \sim N(0, \sigma_\epsilon^2) \]
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!)
As an example, consider the following of a “purely additive” case (= no dominance):

\[ y_i = \mu + X_{ia} \beta_a + X_{id} \beta_d + \varepsilon_i \]

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 \( \mu \)).

\[ \mu = 2, \beta_a = 5, \beta_d = 0, \sigma^2_\varepsilon = 1 \]
Review: Genetic example II

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

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

- A case of NO genetic effect:

\[ \beta_\mu = 2, \beta_a = 0, \beta_d = 0, \sigma_\varepsilon^2 = 1 \]
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:

\[ H_0 : \text{Cov}(X_a, Y) = 0 \cap \text{Cov}(X_d, Y) = 0 \]

\[ H_A : \text{Cov}(X_a, Y) \neq 0 \cup \text{Cov}(X_d, Y) \neq 0 \]

\[ \text{OR} \]

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

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

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

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

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

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

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

- 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) :
  \[ df(M) = 3 - 1 = 2 \quad \text{and} \quad df(E) = n - 3 \]

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

- Finally, we calculate our (LRT!) statistic, the F-statistic with degrees of freedom [2, n-3]:
  \[ F_{[2,n-3]} = \frac{MSM}{MSE} \]
That’s it for today

• Next lecture: quantitative genomics I (!!)