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

Lecture 15: Covariates and QQ plots

Jason Mezey
jgm45@cornell.edu
March 30, 2017 (Th) 8:40-9:55
Announcements

• NO computer lab today

• NO CLASS (or computer lab or office hours) NEXT WEEK (!!) = Spring Break

• Midterm is due 11:59PM, Fri. March 31
Conceptual Overview

Genetic System

Does A1 -> A2 affect Y?

Sample or experimental pop

Measured individuals (genotype, phenotype)

Regression model

Reject / DNR

Pr(Y|X)

Model params

F-test
Review: Measuring LD

- We often see LD among a set of contiguous markers, using either r-squared or D’, with the “triangle, half-correlation matrices” where darker squares indicating higher LD (values of these statistics, e.g. LD in a “zoom-in” plot:}

![Image of a graph showing LD measurements with various markers and gene names like RCOR3, TRAF5, C1orf97, RD3, SLC30A1, NEK2, and LPGAT1. Labels include genes like ZNFBI, ADO, EGRV, Bk, Ik, dV, IYLV, IYLB, IYLY, IYL, IYLI, IYL, ZNFBI, ADO, EGRV, Bk, Ik, dV, IYLV, IYLB, IYLY, IYL, RCOR3, TRAF5, C1orf97, RD3, SLC30A1, NEK2, and LPGAT1.]
Review: Measuring LD II

- There are many statistics used to represent LD but we will present the two most common.

- For the first, define the correlation:

$$r = \frac{Pr(A_i, B_k) - Pr(A_i)Pr(B_k)}{\sqrt{Pr(A_i)(1 - Pr(A_i))}\sqrt{Pr(B_k)(1 - Pr(B_k))}}$$

- As a measure of LD, we will consider this squared:

$$r^2 = \frac{(Pr(A_i, B_k) - Pr(A_i)Pr(B_k))^2}{(Pr(A_i)(1 - Pr(A_i)))(Pr(B_k)(1 - Pr(B_k)))}$$

- Note that this is always between one and zero!
Review: Measuring LD III

• A “problem” with r-squared is that when the MAF of A or B is small, this statistic is small

• For the second measure of LD, we will define a measure $D'$ that is not as dependent on MAF:

\[
D = Pr(A_i, B_k) - Pr(A_i)Pr(B_k)
\]

\[
D' = \frac{D}{\min(Pr(A_1B_2), Pr(A_2, B_1))} \text{if } D > 0
\]

\[
D' = \frac{D}{\min(Pr(A_1B_1), Pr(A_2, B_2))} \text{if } D < 0
\]

• Note that this is always between -1 and 1 (!!)
Review: Manhattan Plot

-Log P

Chromosome

Pi-ta
Quantile-Quantile (QQ) plots

- We will now introduce an essential tool for detecting the most problematic covariates (and can be used to diagnose many other problems!): a Quantile-Quantile (QQ) plot.
- While the definition of a QQ-plot is complex, you will see that how we generate a QQ-plot is easy!
- We will demonstrate the value of a QQ plot for detecting the often problematic variable: population structure.
- In general, whenever you perform a GWAS, you should construct a QQ plot (!!) and always include a QQ plot in your publication.
Quantile-Quantile (QQ) plots II

- Consider a random variable with a continuous probability distribution

- **quantile** - regular, equally spaced intervals of a random variable that divide the random variable into units of equal distribution

- A Quantile-Quantile (QQ) plot (in general) plots the observed quantiles of one distribution versus another OR plots the observed quantiles of a distribution versus the quantiles of the ideal distribution

- We will use a QQ plot to plot our the quantile distribution of observed p-values (on the y-axis) versus the quantile distribution of expected p-values (what distribution is this!?)
Quantile-Quantile (QQ) plots III

• How to construct a QQ plot for a GWAS:
  • If you performed N tests, take the -log (base 10) of each of the p-values and put them in rank order from smallest to largest
  • Create a vector of N values evenly spaces from 1 to 1 / N (how do we do this?), take the -log of each of these values and rank them from smallest to largest
  • Take the pair of the smallest of values of each of these lists and plot a point on an x-y plot with the observed -log p-value on the y-axis and the spaced -log value on the x-axis
  • Repeat for the next smallest pair, for the next, etc. until you have plotted all N pairs in order
Quantile-Quantile (QQ) plots III

• In an ideal GWAS case where there ARE NO causal polymorphisms, your QQ plot will be a line:

![QQ plot where null is true in every case](image)

• The reason is that we will observe a uniform distribution of p-values from such a case and in our QQ we are plotting this observed distribution of p-value versus the expected distribution of p-values: a uniform distribution (where both have been -log transformed)

• Note that if you GWAS analysis is correct but you do not have enough power to detect positions of causal polymorphisms, this will also be your result (!!), i.e. it is a way to assess whether you can detect any hits in your GWAS (!!)
• In an ideal GWAS case where there ARE causal polymorphisms, your QQ plot will be a line with a tail (!!):

![QQ plot where there are some true associations](image)

-\log(\text{expected p-values})

-\log(\text{observed p-values})

• This happens because most of the p-values observed follow a uniform distribution (i.e. they are not in LD with a causal polymorphism so the null hypothesis is correct!) but the few that are in LD with a causal polymorphism will produce significant p-values (extremely low = extremely high -\log(p-values)) and these are in the “tail”

• This is ideally how you want your QQ-plot to look - if it does, you are in good shape!
In practice, you can find your QQ plot looks different than either the “null GWAS” case or the “ideal GWAS” case, for example:

- This indicates that something is wrong (!!!!) and if this is the case, you should not interpret any of your significant p-values as indicating locations of causal polymorphisms (!!!!)

- Note that this means that you need to find an analysis strategy such that the result of your GWAS produces a QQ plot that does NOT look like this (note that this takes experience and many tools to do consistently!)

- Also note that unaccounted for covariates can cause this issue and the most frequent culprit is unaccounted for population structure
Introduction to covariates I

• Recall that in a GWAS, we are considering the following regression model and hypotheses to assess a possible association for every marker with the phenotype

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

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

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

• Also recall that with these hypotheses we are actually testing:

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

\[ H_A : Cov(Y, X_a) \neq 0 \cup Cov(Y, X_d) \neq 0 \]
Introduction to covariates II

- Let’s consider these two cases:
- For the first, the marker is not correlated with a causal polymorphism but the factor is correlated with BOTH the phenotype and the marker such that a test of the marker using our framework will produce a false positive (!!):

\[ \text{Cov}(Y, X_z) \neq 0, \quad H_0 : \beta_a = 0 \cap \beta_d = 0 \]
\[ \text{Cov}(X_a, X_z) \neq 0, \quad H_A : \beta_a \neq 0 \cup \beta_d \neq 0 \]
\[ Y = \beta_\mu + X_a\beta_a + X_d\beta_d + \epsilon \]

- For the second, the marker is not correlated with a causal polymorphism and while the factor is correlated with the phenotype but not the marker, a test of the marker in our framework will model the effect of the factor in our error term:

\[ \text{Cov}(Y, X_z) \neq 0 \]
\[ \text{Cov}(X_a, X_z) = 0 \]
\[ Y = \beta_\mu + X_a\beta_a + X_d\beta_d + \epsilon_{X_z} \]
\[ \epsilon_{X_z} = X_z\beta_z + \epsilon \]
\[ \epsilon \sim N(0, \sigma_\epsilon^2) \]
Modeling covariates I

• Therefore, if we have a factor that is correlated with our phenotype and we do not handle it in some manner in our analysis, we risk producing false positives AND/OR reduce the power of our tests!

• The good news is that, assuming we have measured the factor (i.e. it is part of our GWAS dataset) then we can incorporate the factor in our model as a covariate(s):

\[ Y = \beta_\mu + X_a\beta_a + X_d\beta_d + X_{z,1}\beta_{z,1} + X_{z,2}\beta_{z,2} + \epsilon \]

• The effect of this is that we will estimate the covariate model parameter and this will account for the correlation of the factor with phenotype (such that we can test for our marker correlation without false positives / lower power!)
Modeling covariates II

• How do we perform inference with a covariate in our regression model?

• We perform MLE the same way (!!) our X matrix now simply includes extra columns, one for each of the additional covariates, where for the linear regression we have (and for the logistic regression we use the IRLS algorithm!):

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

• We perform hypothesis testing the same way (!!) with a slight difference: our LRT includes the covariate in both the null hypothesis and the alternative, but we are testing the same null hypothesis:

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

\[
H_A : \beta_a \neq 0 \cup \beta_d \neq 0
\]
Modeling covariates III

• How do we perform inference with a covariate in our lines regression model?

• We perform MLE the same way (!!) our $X$ matrix now simply includes extra columns, one for each of the additional covariates, where for the linear regression we have:

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

• We perform hypothesis testing the same way (!!) with a slight difference: our LRT includes the covariate in both the null hypothesis and the alternative, but we are testing the same null hypothesis:

$$H_0 : \beta_a = 0 \cap \beta_d = 0$$

$$H_A : \beta_a \neq 0 \cup \beta_d \neq 0$$
Modeling covariates IV

• First, determine the predicted value of the phenotype of each individual under the null hypothesis (how do we set up $\mathbf{x}$?):

$$\hat{y}_i, \hat{\theta}_0 = \hat{\beta}_\mu + \sum_{j=1} x_{i,z,j} \hat{\beta}_{z,j}$$

• Second, determine the predicted value of the phenotype of each individual under the alternative hypothesis (set up $\mathbf{x}$?):

$$\hat{y}_i, \hat{\theta}_1 = \hat{\beta}_\mu + x_{i,a} \hat{\beta}_a + x_{i,d} \hat{\beta}_d + \sum_{j=1} x_{i,z,j} \hat{\beta}_{z,j}$$

• Third, calculate the “Error Sum of Squares” for each:

$$SSE(\hat{\theta}_0) = \sum_{i=1}^n (y_i - \hat{y}_i, \hat{\theta}_0)^2 \quad SSE(\hat{\theta}_1) = \sum_{i=1}^n (y_i - \hat{y}_i, \hat{\theta}_1)^2$$

• Finally, we calculate the F-statistic with degrees of freedom $[2, n-3]$ (why two degrees of freedom?):

$$F_{[2, n-3]}(\mathbf{y}, \mathbf{x}) = \frac{SSE(\hat{\theta}_0) - SSE(\hat{\theta}_1)}{\frac{2}{n-3} SSE(\hat{\theta}_1)}$$
Modeling covariates V

- Thus, for testing the null hypothesis in a linear regression, we can construct an F-test using a slightly different formula:

\[
SSE(\hat{\theta}_0) = \sum_{i=1}^{n} (y_i - \hat{y}_{i,\hat{\theta}_0})^2
\]

\[
SSE(\hat{\theta}_1) = \sum_{i=1}^{n} (y_i - \hat{y}_{i,\hat{\theta}_1})^2
\]

\[
F_{[2,n-3]}(y, x_a, x_d) = \frac{\frac{SSE(\hat{\theta}_0) - SSE(\hat{\theta}_1)}{2}}{\frac{SSE(\hat{\theta}_1)}{n-3}}
\]

- Note that our previous formula for an F-statistic can be represented this way as well (!!)

- Once you calculate this F-statistic, you compare to an F-distribution (under the null) with 2 and n-3 degrees of freedom to test the null hypothesis
Modeling covariates VI

• Say you have GWAS data (a phenotype and genotypes) and your GWAS data also includes information on a number of covariates, e.g. male / female, several different ancestral groups (different populations!!), other risk factors, etc.

• First, you need to figure out how to code the \( Xz \) in each case for each of these, which may be simple (male / female) but more complex with others (where how to code them involves fuzzy rules, i.e. it depends on your context!!)

• Second, you will need to figure out which to include in your analysis (again, fuzzy rules!) but a good rule is if the parameter estimate associated with the covariate is large (=significant individual p-value) you should include it!

• There are many ways to figure out how to include covariates (again a topic in itself!!)
Example of a covariate impacting a QQ plot: population structure

• “Population structure” or “stratification” is a case where a sample includes groups of people that fit into two or more different ancestry groups (fuzzy def!)

• Population structure is often a major issue in GWAS where it can cause lots of false positives if it is not accounted for in your model

• Intuitively, you can model population structure as a covariate if you know:
  • How many populations are represented in your sample
  • Which individual in your sample belongs to which population

• QQ plots are good for determining whether there may be population structure

• “Clustering” techniques are good for detecting population structure and determining which individual is in which population (=ancestry group)
People geographically separate through migration and then the set of alleles present in the population evolves (=changes) over time.
Why might (unaccounted for) structure be a problem in a GWAS?

- Even if you had a case where there were NO causal polymorphisms for a phenotype, you can get false positives if:
  - If you have more than one population in your sample (that you do not model with a covariate)
  - If these populations differ in MAF at a subset of measured genotypes
  - If these populations differ in the mean value of the phenotype
- In such a case, every genotype where an MAF is different between the populations would be expected to produce a low p-value (=biological false positives!) - this is often detectable in a QQ plot (!!!)
- Note: if you can “learn” (or know) the population information for your data, you can model this as a covariate and you (may) be able to correct this problem
QQ plots can detect population structure impacts in GWAS
That’s it for today

- NO CLASS NEXT WEEK - see you April 11 (!!)