LD Score regression distinguishes confounding from polygenicity in genome-wide association studies

Bulik-Sullivan et al (2015a)

Timothy Mak

Journal club, 9/1/2019

Core idea

- If we examine $\mathcal{X}^2$ statistics from a GWAS, there should be three "components":
  - A *local* component which is proportional to the amount of LD the SNP has. (Number of SNPs it can potentially tag.)
  - A *global* component reflecting the amount of population stratification
  - A *random* component unrelated to LD or population stratification
- Therefore it makes sense to consider a linear regression of the statistics on the amount of LD a SNP has, in order to estimate the relative contribution of the three components.
- It so happens that the slope of the linear regression also represents an estimate of the SNP heritability.

Some preliminaries

Let:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$
$$\mathbb{E}(y_i) = \mathbb{E}(x_{ij}) = 0$$
$$Var(y_i) = Var(x_{ij}) = 1$$
$$\beta_j \overset{iid}{\sim} N(0, \sigma^2_\beta)$$
$$\epsilon_i \overset{iid}{\sim} N(0, \sigma^2_\epsilon)$$

Then:

$$\text{Heritability} = h^2 = \frac{Var\left(\sum_j x_{ij}\beta_j\right)}{Var(y_i)}$$
$$= \sigma^2_\beta E\left(\sum_j x^2_{ij}\right) = M\sigma^2_\beta$$

Some preliminaries

In the following, we further assume, by standardization:

$$\mathbf{1}^T \mathbf{X}_j = \mathbf{1}^T \mathbf{y} = 0$$
$$\mathbf{X}_j^T \mathbf{X}_j = \mathbf{y}^T \mathbf{y} = N$$

Moreover, we define:

$$\begin{aligned}
\tilde{r}_{jk} &= \mathbf{X}_j^T \mathbf{X}_k / N \\
\tilde{r}_j^{(y)} &= \mathbf{X}_j^T \mathbf{y} / N \\
z_j &= \tilde{r}_j^{(y)} / SE(\tilde{r}_j^{(y)}) \\
&= \tilde{r}_j^{(y)} / \sqrt{1/N} \\
\mathbb{E}(\mathcal{X}_j^2) &= z_j^2
\end{aligned}$$

Therefore...

$$\begin{aligned}\tilde{r}_j^{(y)} &= \frac{\mathbf{X}_j^T(\mathbf{X}\beta + \epsilon)}{N} \\ &= \tilde{\mathbf{r}}_j^T\beta + \mathbf{X}_j^T\epsilon/N \\ &= N\,\mathbb{E}(\tilde{\mathbf{r}}_j^T\beta\beta^T\tilde{\mathbf{r}}_j) + 2\,\mathbb{E}(\tilde{\mathbf{r}}_j^T\beta\mathbf{X}_j^T\epsilon) + \mathbb{E}(\mathbf{X}_j^T\epsilon\epsilon^T\mathbf{X}_j)/N\end{aligned}$$

Taking expectations over β and ϵ,

$$\mathbb{E}(\mathcal{X}_j^2) = N\sigma_\beta^2 E\left(\sum_k \tilde{r}_{jk}^2\right) + \sigma_\epsilon^2$$

Therefore...

Under Hardy-Weinberg equilibrum (i.e. no population stratification),

$$\begin{aligned} E(\tilde{r}_{jk}^2) &= r_{jk}^2 + (1 - r_{jk}^2)/N + \mathcal{O}(1/N^2) \\ &\approx r_{jk}^2 + 1/N \end{aligned}$$

(Let me work this out later...)
Hence,

$$\mathbb{E}(\mathcal{X}_j^2) \approx \frac{Nh^2 l_j}{M} + 1$$

where

$$l_j = \sum_j r_{jk}^2$$

if we have $h^2 = M\sigma_\beta^2 = 1 - \sigma_\epsilon^2$

Introducing population stratification

Those F statistics

Remember F for the inbreeding coefficient?
If the frequency of an allele A is p, then the frequency of the genotype AA is:

$$Pr(AA) = pF + (1 - F)p^2 = p^2 + p(1 - p)F$$

If X and Y denote the two alleles from the same individual, then

$$\begin{aligned} Cov(X, Y) &= \mathbb{E}(X = A, Y = A) - \mathbb{E}(X = A)\,\mathbb{E}(Y = A) \\ &= Pr(AA) - p^2 = p(1 - p)F \end{aligned}$$

Since $Var(X) = Var(Y) = p(1 - p)$, another interpretation of F is that it is the correlation of the two alleles in an individual (for a particular gene).

The F_{ST} statistic

The idea of F_{ST} is that it is the equivalent of F when we examine two alleles from two different sub-populations in a population. Again, let p denote the overall frequency of the allele, and X, Y denote two alleles taken from the same sub-population, then

$$F_{ST} = Corr(X, Y) = \frac{Cov(X, Y)}{p(1-p)}$$

Intuitively, the larger F_{ST} is, the greater the *between* sub-population variation in frequency relatively to the total variation, and hence the greater the population stratification.

The LD score setup

In a GWAS setting, let us let F_{ST} to be the average F_{ST} across all SNPs. A simple model is to assume there are two sub-populations of equal size, such that the overall frequency for SNP j is p_j, and that the frequencies for the two sub-populations are $p_{j1} = p_j + g_j$ and $p_{j2} = p_j - g_j$ respectively. Let X_1, X_2 denote a pair sampled alleles from the same sub-population, and $D = 1, 2$ denote the event of sampling from either the first/second population. We have:

$$Pr(D = 1) = Pr(D = 2) = 0.5$$
$$\mathbb{E}(X_h|D = 1) = p_j + g_j$$
$$\mathbb{E}(X_h|D = 2) = p_j - g_j$$

The LD score setup

$$
\begin{aligned}
\mathbb{E}(X_1 X_2) &= Pr(D=1)\,\mathbb{E}(X_1 X_2 | D=1) + Pr(D=2)\,\mathbb{E}(X_1 X_2 | D=2) \\
&= \left((p_j + g_j)^2 + (p_j - g_j)^2\right)/2 \\
&= p_j^2 + g_j^2 \\
Cov(X_1, X_2) &= \mathbb{E}(X_1 X_2) - \mathbb{E}(X_1)\,\mathbb{E}(X_2) \\
&= g_j^2 \\
F_{ST\,j} &= \frac{g_j^2}{p_j(1-p_j)}
\end{aligned}
$$

The LD score setup

In the LD score paper, they denote $\sqrt{F_{STj}}$ by f_j, and let

$$f_j \sim N(0, F_{ST})$$

Moreover, let $Corr(f_j, f_k)$ be V_{jk}.

The LD score setup

Finally, they let instead:

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{S} + \epsilon$$

where $\mathbf{S} = (S_1, S_2, \ldots, S_n)^T$, and

$$S_i = \begin{cases} \sigma_s & \text{if } D_i = 1 \\ -\sigma_s & \text{if } D_i = 2 \end{cases}$$

i.e. they suppose that the phenotype has different means in the two sub-populations. (There's a typo in the manuscript. It should be σ_s rather than $\sigma_s/2$.)

Finally, they continue to require $var(y_i) = 1 = h^2 + \sigma_s^2 + \sigma_\epsilon^2$.

This means...

$$\begin{aligned}
\tilde{r}_j^{(y)} &= \frac{\mathbf{X}_j^T(\mathbf{X}\beta + \mathbf{S} + \epsilon)}{N} \\
&= \tilde{\mathbf{r}}_j^T\beta + \mathbf{X}_j^T\mathbf{S}/N + \mathbf{X}_j^T\epsilon/N \\
\mathbb{E}(\mathcal{X}_j^2) &= n\,\mathbb{E}(\tilde{\mathbf{r}}_j^T\beta\beta^T\tilde{\mathbf{r}}_j) + \mathbb{E}(\mathbf{X}_j^T\mathbf{S}\mathbf{S}^T\mathbf{X}_j)/N + \mathbb{E}(\mathbf{X}_j^T\epsilon\epsilon^T\mathbf{X}_j)/N \\
\mathbb{E}(\mathbf{X}_j^T\mathbf{S}\mathbf{S}^T\mathbf{X}_j) &= \frac{1}{2}\left(\mathbb{E}(\mathbf{X}_j^T\mathbf{S}\mathbf{S}^T\mathbf{X}_j|D_i = 1) + \mathbb{E}(\mathbf{X}_j^T\mathbf{S}\mathbf{S}^T\mathbf{X}_j|D_i = 2)\right)
\end{aligned}$$

It turns out that:

$$\begin{aligned}
\mathbb{E}(\mathbf{X}_j^T\mathbf{S}\mathbf{S}^T\mathbf{X}_j) = \mathbb{E}(\mathbf{X}_j^T\mathbf{S}\mathbf{S}^T\mathbf{X}_j|D_i) &= N^2\sigma_s^2F_{ST} + N\sigma_s^2(1 - F_{ST}) \\
&\approx N^2\sigma_s^2F_{ST}
\end{aligned}$$

(The derivation in the manuscript is not entirely correct...)

Finally...

In the manuscript we have:

$$N\,\mathbb{E}(\tilde{\mathbf{r}}_j^T \boldsymbol{\beta}\boldsymbol{\beta}^T \tilde{\mathbf{r}}_j) \approx \frac{Nh^2}{M} l_j + 1 + Nh^2 F_{ST}$$

A key point to take from the derivation is that although $\mathbb{E}(\tilde{\mathbf{r}}) \approx \mathbf{r}$, $\mathbb{E}(\tilde{\mathbf{r}^2}) > \mathbf{r^2}$ if $\mathbf{r^2}$ denotes the within sub-population LD.

However, the inflation due to population stratification is not dependent on the LD score, and hence only affects the intercept. Note that this inflation is also not dependent on σ_s.

Another point is, again, the derivation in the manuscript is not correct, although the final result is valid.

Not covered...

- Variance estimates
- Meta-analysis

Extension to genetic correlations

Instead of considering $\mathbb{E}(X_j^2) = N\,\mathbb{E}(\tilde{r}_j^{(y)^2})$, consider

$$\sqrt{N1N2}\,\mathbb{E}(\tilde{r}_j^{(y1)}\tilde{r}_j^{(y2)}) \approx \frac{\sqrt{N_1 N_2}\rho_g}{M} l_j + \frac{N_s \rho}{\sqrt{N_1 N_2}} + \rho_g F_{ST}^2 \sqrt{N_1 N_2} + \frac{N_s^2 F_{ST} \sigma_s^2}{\sqrt{N_1 N_2}}$$

(Bulik-Sullivan et al, 2015b; Yengo et al, 2018)

Unfortunately, although the equation is correct, the derivation is again incorrect, partly because they relied on some of the results from Bulik-Sullivan et al (2015a).

LD score for ascertained (case/control) samples

- Bulik-Sullivan et al (2015a) showed by simulations that LD score regression also gives unbiased estiamtes when working with summary statistics from case-control data. But there's no theoretical derivation.
- However, Bulik-Sullivan (2015) showed that LD score regression is in fact very similar to doing Haseman-Elston regression on summary statistics, which is valid for ascertained samples.
- I believe there's also another way to show that it works, but it's still only in my head.

References

- Bulik-Sullivan et al (2015a). *LD Score regression distinguishes confounding from polygenicity in genome-wide association studies*. Nature Genetics
- Bulik-Sullivan et al (2015b). *An atlas of genetic correlations across human diseases and traits*. Nature Genetics
- Bulik-Sullivan (2015). *Relationship between LD Score and Haseman-Elston Regression*. BioRxiv
- Yengo et al (2018). *Meta-analysis of genome-wide association studies for height and body mass index in 700000 individuals of European ancestry*. Human Molecular Genetics