Summer Presentation(1):

LD Score Regression

Name: Chenxiao Tian

Date: 2022/7/11

OUTLINE

- 1. Motivation
- 2. Method:
- (1) Unstructured Case,
- (2) Population Stratification Case,
- (3)Conditional Variance of the χ^2-Statistic and its application in improving the estimating of the confounding term aNF$_{ST}$ from summary statistics
- 3. Simulations
- (1) Simulations with polygenic genetic architectures
- (2) Simulations with confounding
- (3) Simulations with confounding and polygenicity
- 4. Real Data
- 5. Conclusion

MOTIVATION

- In GWAS research, polygenicity (i.e. several minor gene effects) and the deviation caused by interference factors (such as recessive Association cryptic correlation, population stratification, etc.) will cause the distribution of test statistics to be increased.
- But we can't tell whether the higher statistics come from polygenesis or interference factors, so through LD score expression, we can quantitatively analyze the impact of each part by studying the relationship between test statistics and linkage disequilibrium score.

Method: Unstructured Case

- 1.1 Model

$$\phi = X\beta + \epsilon$$

Assume 1:A model where all three variables on the right side are random:

ϕ: N $\times$ 1 vector of (quantitative) phenotypes

X: N $\times$ M matrix of genotypes normalized to mean zero and variance one

β: M $\times$ 1 vector of per-normalized-genotype effect sizes

ϵ: N $\times$ 1 vector of environmental effects

$E[\epsilon]=0$, $Var[\epsilon] = (1 - h_g^2)I$,

$E[\beta]=0$, $Var[\beta]=(h_g^2 /M)I$,

Implicit Assume 1: Here we not only just assume i.i.d. per-normalized genotype effect sizes for genotyped SNPs, we also make this assumption for all SNPs.

Implicit Assume 2: X normalization to variance hides an implicit assumption that rare SNPs have larger effect sizes.

Method: Unstructured Case

- Assume 2: We assume that the genotype at variant j for individual i is independent of other individuals' genotypes.
- Assume 3: We do incorporate linkage disequilibrium into the model: define $r_{jk} := E[X_{ij}X_{ik}]$, which does not depend on i.
- Assume 4: We assume that X, β and ϵ are mutually independent.
- Remark: We will relax the assumption that environmental effects are independent of genotype when we model population stratification

Method: Unstructured Case

- 1.2. Relationship between LD and χ^2-Statistics.

Least-squares estimates

For each variant j = 1,...,M, the least-squares estimates of effect size is:

$$\hat{\beta}_j := X_j^{\mathsf{T}} \phi / N$$

Here Xj denotes the N $\times$ 1 vector of genotypes at variant j.

χ^2-Statistics

$$\chi_j^2 := N \hat{\beta}_j^2.$$

LD Score of variant j

$$\ell_j := \sum_{k=1}^{M} r_{jk}^2.$$

Method: Unstructured Case

- 1.3 Proposition 1:
- We compute the E[χ_j^2] with the expectation taken over random X, ß, and ϵ.

Proposition 1. *Define the* ***LD Score*** *of variant* j *as*

$$\ell_j := \sum_{k=1}^{M} r_{jk}^2. \tag{1.2}$$

Under the model described in §1.1, *the expected* χ^2*-statistic of variant* j *is*

$$\mathbb{E}[\chi_j^2] \approx \frac{N h_g^2}{M} \ell_j + 1. \tag{1.3}$$

Method: Unstructured Case

$$\phi = X\beta + \epsilon \qquad \hat{\beta}_j := X_j^T \phi / N \qquad \chi_j^2 := N\hat{\beta}_j^2.$$

$$r_{jk} := \mathbb{E}[X_{ij}X_{ik}], \qquad \ell_j := \sum_{k=1}^{M} r_{jk}^2. \qquad \mathbb{E}[\chi_j^2] \approx \frac{Nh_g^2}{M}\ell_j + 1.$$

- Proof:
- Step 1: By the Law of Total Variance, We translate it into the computing of conditional variance. Notice that we assume that X, β,ϵ are independent and random:

Proof. Since $\mathbb{E}[\hat{\beta}_j] = 0$, observe that $\mathbb{E}[\chi_j^2] = N \cdot \mathrm{Var}[\hat{\beta}_j]$. We will obtain the variance of $\hat{\beta}_j$ via the law of total variance:

$$\begin{aligned} \mathrm{Var}[\hat{\beta}_j] &= \mathbb{E}[\mathrm{Var}[\hat{\beta}_j \mid X]] + \mathrm{Var}[\mathbb{E}[\hat{\beta}_j \mid X]] \\ &= \mathbb{E}[\mathrm{Var}[\hat{\beta}_j \mid X]], \end{aligned} \tag{1.4}$$

where the second line follows from the fact that $\mathbb{E}[\hat{\beta}_j \mid X] = 0$, irrespective of X.

Method: Unstructured Case

$$\phi = X\beta + \epsilon \qquad \hat{\beta}_j := X_j^\mathsf{T}\phi/N \qquad \chi_j^2 := N\hat{\beta}_j^2.$$

$$r_{jk} := \mathbb{E}[X_{ij}X_{ik}], \qquad \ell_j := \sum_{k=1}^{M} r_{jk}^2. \qquad \mathbb{E}[\chi_j^2] \approx \frac{Nh_g^2}{M}\ell_j + 1.$$

- Proof:

Remember that we assume E[ϵ]=0 , Var[ϵ] = (1 − h_g^2)I, E[β]=0, Var[β]=(h_g^2 /M)I,

also by the assume that X, β,ϵ are independent and random, but under the fixed condition X, X is constant, by the line property of Var(AX)=A Var(X)A^T;

And by the assume that X: N × M matrix of genotypes normalized to mean zero and variance one, we have:

$$\begin{aligned} \text{(1.5)} \qquad \mathrm{Var}[\hat{\beta}_j \mid X] &= \frac{1}{N^2}\mathrm{Var}[X_j^\mathsf{T}\phi \mid X] \\ &= \frac{1}{N^2}X_j^\mathsf{T}\mathrm{Var}[\phi \mid X]X_j \\ &= \frac{1}{N^2}\left(\frac{h_g^2}{M}X_j^\mathsf{T}XX^\mathsf{T}X_j + N(1-h_g^2)\right). \end{aligned}$$

Method: Unstructured Case

- Proof:

$$\phi = X\beta + \epsilon \qquad \hat{\beta}_j := X_j^\intercal \phi / N \qquad \chi_j^2 := N\hat{\beta}_j^2.$$

$$r_{jk} := \mathbb{E}[X_{ij}X_{ik}], \qquad \ell_j := \sum_{k=1}^{M} r_{jk}^2. \qquad \mathbb{E}[\chi_j^2] \approx \frac{Nh_g^2}{M}\ell_j + 1.$$

- By the assume that $r_{jk} := E[X_{ij}X_{ik}]$ does not depend on i with some direct computing and liner property of expectation, (also by the assume that the genotype at variant j for individual i is independent of other individuals' genotypes.). We obtain the approximation sign by hiding terms of order $O(1/N^2)$ and smaller:

We can write the term on the left in terms of more familiar quantities as

$$\frac{1}{N^2} X_j^\intercal X X^\intercal X_j = \sum_{k=1}^{M} \tilde{r}_{jk}^2, \tag{1.6}$$

where $\tilde{r}_{jk} := \frac{1}{N}\sum_{i=1}^{N} X_{ij}X_{ik}$ denotes the sample correlation between additively-coded genotypes at variants j and k. Since

$$\mathbb{E}[\tilde{r}_{jk}^2] \approx r_{jk}^2 + (1 - r_{jk}^2)/N, \tag{1.7}$$

(where the approximation sign hides terms of order $\mathcal{O}(1/N^2)$ and smaller; one can obtain this approximation via *e.g.*, the δ-method),

Method: Unstructured Case

- Proof:

$$\phi = X\beta + \epsilon \qquad \hat{\beta}_j := X_j^\intercal \phi / N \qquad \chi_j^2 := N\hat{\beta}_j^2.$$

$$r_{jk} := \mathbb{E}[X_{ij}X_{ik}], \qquad \ell_j := \sum_{k=1}^{M} r_{jk}^2. \qquad \mathbb{E}[\chi_j^2] \approx \frac{Nh_g^2}{M}\ell_j + 1.$$

- Finally, we put all them together to get the final approximation equality:, here notice that N is large enough to assume again $1-\frac{1}{N} \approx 0$:

(1.8)
$$\mathbb{E}\left[\sum_{k=1}^{M} \tilde{r}_{jk}^2\right] \approx \ell_j + \frac{M - \ell_j}{N}.$$

Thus,

(1.9)
$$\begin{aligned} \mathbb{E}[\chi_j^2] &\approx \frac{N(1 - 1/N)h_g^2}{M}\ell_j + 1 \\ &\approx \frac{Nh_g^2}{M}\ell_j + 1, \end{aligned}$$

Values of N (study sample size) considered in the main text generally fall between 10^4 and 10^5, so the approximation $1 - 1/N \approx 1$ is appropriate. □

LD Score with Population Stratification

- 2.1. Model of Population Structure.
- (1)Let X be a matrix of normalized genotypes X consisting of N/2 samples from population 1 and N/2 samples from population 2:

(Here we will use the notation i ∈ Pm for m ∈ {1, 2} to denote that individual i is a member of population m)

(2)Subject to the following constraints:

$$\mathrm{Var}[X_{ij}] = 1,\ \mathbb{E}[X_{ij} \,|\, i \in P_1] = f_j \text{ and } \mathbb{E}[X_{ij} \,|\, i \in P_2] = -f_j.$$

(3) Where the drift term f is modeled as:

We model the drift term f as $f \sim N(0, F_{ST}V)$, where V is a correlation matrix[3] and F_{ST} is Wright's F_{ST} [2]. We postpone discussion of the off-diagonal entries

LD Score with Population Stratification

- Assume 1: If $\ell_{j,m}$ denotes the LD Score of variant j in population m, we assume that $\ell_{j,1} \approx \ell_{j,2} =: \ell_j$.
- Reason for Assume 1： Assuming approximately equal LD Scores in both populations is certainly not reasonable for very large values of FST.
- However, typical values of FST for human populations are ≈ 0.1 for populations from different continents, FST ≈ 0.01 for populations on the same continent, and FST < 0.01 for subpopulations within the same country.
- Assume 2：
- In particular, we assume that the diagonal entries of V are all equal, or at least uncorrelated with LD Score.

LD Score with Population Stratification

Compute of the Variance of rmix,jk:

(1)Suppose j and k are unlinked variants such that rjk,1 = rjk,2 = 0 and fj is independent of fk.

(2)Let rmix,jk denote the correlation between SNPs j and k in such a mixture of populations. Conditional on f：

$$
\begin{aligned}
(2.1) \qquad \mathbb{E}[r_{mix,jk} \mid f] &= \mathbb{E}[X_{ij} X_{ik} \mid f] \\
&= \frac{1}{2}\left(\mathbb{E}[X_{ij} X_{ik} \mid f, i \in P_1] + \mathbb{E}[X_{ij} X_{ik} \mid f, i \in P_2]\right) \\
&= f_j f_k.
\end{aligned}
$$

(3) If we take the expectation over random fj and fj , then E[rmix,jk]=0, since fj and fk are independent with expectation zero.

(4)We can use equation 2.1 to compute the variance, also since that E[rmix,jk]=0, Var[rmix,jk] = E[r^2_{mix},jk], we have:

$$
\begin{aligned}
(2.2) \qquad \mathrm{Var}[r_{mix,jk}] &= \mathrm{Var}[\mathbb{E}[r_{mix,k} \mid f]] + \mathbb{E}[\mathrm{Var}[r_{mix,jk} \mid f]] \\
&= \mathbb{E}[f_j^2 f_k^2] + 0 \\
&= \mathbb{E}[f_j^2]\mathbb{E}[f_k^2] \\
&= F_{ST}^2.
\end{aligned}
$$

LD Score with Population Stratification

- The approximation of the sample LD Score (Assume: Note that we have ignored the case where j and k are linked and Vjk is not 0)

(1.7) $$\mathbb{E}[\tilde{r}_{jk}^2] \approx r_{jk}^2 + (1 - r_{jk}^2)/N,$$

(where the approximation sign hides terms of order $\mathcal{O}(1/N^2)$ and smaller; one can obtain this approximation via *e.g.*, the δ-method),

Observe that since $\mathbb{E}[r_{mix,jk}] = 0$, $\mathrm{Var}[r_{mix,jk}] = \mathbb{E}[r_{mix,jk}^2]$. By equation 1.7, in a finite sample,

(2.3) $$\mathbb{E}[\tilde{r}_{mix,jk}^2] \approx F_{ST}^2 + (1 - F_{ST}^2)/N.$$

Thus, the sample LD Score is approximately

(2.4) $$\begin{aligned} \mathbb{E}[\tilde{\ell}_j] &\approx \ell_j + MF_{ST}^2 + \frac{M(1 - F_{ST}^2)}{N} \\ &\approx \ell_j + MF_{ST}^2 + \frac{M}{N}. \end{aligned}$$

LD Score with Population Stratification

- About the Assume: Note that we have ignored the case where j and k are linked and Vjk is not 0.

Note that we have ignored the case where j and k are linked and $V_{jk} \neq 0$. In this case, $\mathbb{E}[f_j^2 f_k^2] = F_{ST}^2 + 2F_{ST}^2 V_{jk}^2$ (from the formula for the double second moments of a multivariate normal distribution). Even if for some variants j, the number of variants k such that $V_{jk} > 0$ is $\approx 10^3$, this will make a negligible difference in $\mathbb{E}[\tilde{\ell}_j]$, because $\sum_{k:V_{jk}>0} 2F_{ST}^2 V_{jk}^2 < 2000 F_{ST}^2 \ll M F_{ST}^2$ when $M \approx 10^7$.

Thus, the sample LD Score is approximately

$$\begin{aligned} \mathbb{E}[\tilde{\ell}_j] &\approx \ell_j + M F_{ST}^2 + \frac{M(1 - F_{ST}^2)}{N} \\ &\approx \ell_j + M F_{ST}^2 + \frac{M}{N}. \end{aligned} \tag{2.4}$$

LD Score with Population Stratification

- 2.3. Model of Stratified Phenotype.
- For the environmental term, we addtionally add an environmental stratification term, where S is a fixed constant and take other variables are random:

2.3. **Model of Stratified Phenotype.** To model population stratification, we model phenotypes as generated by the equation

$$\phi = X\beta + S + \epsilon, \tag{2.5}$$

where X is as described in §2.1, β is as described in §1.1 and where S is an environmental stratification[4] term defined by

$$S_i := \begin{cases} \sigma_s/2, & i \in P_1 \\ -\sigma_s/2, & i \in P_2. \end{cases} \tag{2.6}$$

Finally, ϵ is as described in §1.1, except $\mathrm{Var}[\epsilon] = (1 - h_g^2 - \sigma_s^2)$, which assures that the variance of ϕ in the population is 1[5]. We compute χ^2-statistics as defined in §1.1. In this section, we compute $\mathbb{E}[\chi_j^2]$ with the expectation taken over random X, β, ϵ, f but with S fixed to ensure population stratification.

LD Score with Population Stratification

2.4. Relationship between LD and Stratified χ^2-Statistics.

Proposition 2. *Under the model described in §2.3, the expected χ^2-statistic of variant j is*

$$\mathbb{E}[\chi_j^2] = \frac{Nh_g^2}{M}\ell_j + 1 + aNF_{ST}, \tag{2.7}$$

*where a is the expectation of squared difference in mean phenotypes between population*1 *and population* 2.

$$a := \mathbb{E}[(\bar{\phi}_1 - \bar{\phi}_2)^2].$$

LD Score with Population Stratification

- Step 1: Also by the law of total variance, but since we use different model where has the term S to describe the confounding from population stratification, the left term $E(\widehat{\beta_j}|X)$ unfortunately is not 0. So we have to both compute them:

Proof. Since $\mathbb{E}[\hat{\beta}_j] = 0$, observe that $\mathbb{E}[\chi_j^2] = N \cdot \mathrm{Var}[\hat{\beta}_j]$. We will obtain the variance of $\hat{\beta}_j$ via the law of total variance:

$$\mathrm{Var}[\hat{\beta}_j] = \mathbb{E}[\mathrm{Var}[\hat{\beta}_j \mid X]] + \mathrm{Var}[\mathbb{E}[\hat{\beta}_j \mid X]]. \tag{2.8}$$

Note that one can calculate f from X, so by conditioning on X we also implicitly condition on f. Unlike in equation 1.4, $\mathbb{E}[\hat{\beta}_j \mid X] \neq 0$, because of confounding from population stratification. The inner portion of the first term on the right side of equation 2.8 is the same as in equation 1.5,

LD Score with Population Stratification

- Step 2:The inner portion of the left term on the right side of equation 2.8 isthe same as in equation 1.5:

(1.5)
$$\begin{aligned}\mathrm{Var}[\hat{\beta}_j \mid X] &= \frac{1}{N^2}\mathrm{Var}[X_j^\intercal \phi \mid X] \\ &= \frac{1}{N^2} X_j^\intercal \mathrm{Var}[\phi \mid X] X_j \\ &= \frac{1}{N^2}\left(\frac{h_g^2}{M} X_j^\intercal X X^\intercal X_j + N(1-h_g^2)\right).\end{aligned}$$

Thus, the sample LD Score is approximately

(2.4)
$$\begin{aligned}\mathbb{E}[\tilde{\ell}_j] &\approx \ell_j + M F_{ST}^2 + \frac{M(1-F_{ST}^2)}{N} \\ &\approx \ell_j + M F_{ST}^2 + \frac{M}{N}.\end{aligned}$$

(2.9)
$$\mathrm{Var}[\hat{\beta}_j \mid X] = \frac{1}{N^2}\left(\frac{h_g^2}{M} X_j^\intercal X X^\intercal X_j + N(1-h_g^2)\right).$$

We can take the expectation over random X (and therefore over random f) using the result from equation 2.4. Thus,

(2.10)
$$\begin{aligned}\mathbb{E}[\mathrm{Var}[\hat{\beta}_j \mid X]] &= \frac{1}{N^2}\left(\frac{h_g^2}{M}\mathbb{E}[X_j^\intercal X X^\intercal X_j] + N(1-h_g^2)\right) \\ &\approx \frac{h_g^2}{M}\ell_j + h_g^2 F_{ST}^2 + \frac{1}{N}.\end{aligned}$$

LD Score with Population Stratification

- Step 3: Then the inner portion of the right term on the right side of equation 2.8 is:

$$(2.11) \qquad \begin{aligned} \mathbb{E}[\hat{\beta}_j \mid X] &= \frac{1}{N}\mathbb{E}[X_j^\intercal X\beta + X_j^\intercal S + X_j^\intercal \epsilon] \\ &= \frac{1}{N}X_j^\intercal S \\ &= f\sigma_s. \end{aligned}$$

Since f has variance F_{ST}, $\mathrm{Var}[f\sigma_s] = \sigma_s^2 F_{ST}$. Thus,

LD Score with Population Stratification

Step 4: Put the previous computation results of the both terms of the right side of the equation together, we finally have that:

Since f has variance F_{ST}, $\mathrm{Var}[f\sigma_s] = \sigma_s^2 F_{ST}$. Thus,

$$\mathbb{E}[\chi_j^2] = N \cdot \mathrm{Var}[\hat{\beta}_j] \tag{2.12}$$

$$\frac{Nh_g^2}{M}\ell_j + 1 + NF_{ST}(\sigma_s^2 + h_g^2 F_{ST}).$$

LD Score with Population Stratification

- Step 5:If we let $\overline{\phi_m}$ denote the mean phenotype in population m $\in$ {1, 2},
- Let the expected squared mean difference in phenotype between populations be：

$$a := \mathbb{E}[(\bar{\phi}_1 - \bar{\phi}_2)^2].$$

- Then we have the following new interpretation of a：

$$\begin{aligned} &= \sigma_s^2 + \sum_{j=1}^{M} \left[\mathbb{E}[\beta_j^2] \left(\sum_{i \in P_1} \mathbb{E}[X_{ij}^2 \,|\, i \in P_1] + \sum_{i \in P_2} \mathbb{E}[X_{ij}^2 \,|\, i \in P_2] \right) \right] \\ &= \sigma_s^2 + h_g^2 F_{ST}. \end{aligned}$$

$$\mathbb{E}[\chi_j^2] = \frac{N h_g^2}{M} \ell_j + 1 + a N F_{ST}, \tag{2.14}$$

Estimating the confounding term aNFST from summary statistics and Variance

- Method 1：We can directly regress χ^2 against LD Score, then the intercept minus one is an estimate of aNFST ：

$$\mathbb{E}[\chi_j^2] = \frac{Nh_g^2}{M}\ell_j + 1 + aNF_{ST}, \tag{2.14}$$

- Weakness：the variance of χ^2 increases with LD Score,

Estimating the confounding term aNFST from summary statistics and Variance

- Method 2：We can improve the efficiency of this estimator by weighting the regression by the reciprocal of the conditional variance function Var[$\chi_j^2 \mid \ell$]。

Note that the regression weights do not affect the expectation of the parameter estimates, only the standard error.

Assume 1：We need stronger assumptions in order to derive the conditional variance: we assume that N is large and ß ~ N(0, h2 g I), and ϵ ~ N(0,(1 − h2 g) I).

Then we have that：

$$
\begin{aligned}
(3.1) \qquad \hat{\beta}_j &= \frac{1}{N}(X_j^{\mathsf{T}} X\beta + X_j^{\mathsf{T}}\epsilon) \\
&\sim N(0, h_g^2 \ell_j / M + h_g^2/N) + N(0, (1-h_g^2)/N) \\
&\sim N(0, h_g^2 \ell_j / M + 1/N),
\end{aligned}
$$

Estimating the confounding term aNFST from summary statistics and Variance

- Thus, $\chi_j^2 = N\hat{\beta}_j^2$ follows a scaled χ^2 distribution with scale factor $Nh_g^2\ell j/M$, and the conditional variance function is：

$$\mathrm{Var}[\chi_j^2 \mid \ell_j] = \left(1 + \frac{Nh_g^2}{M}\ell_j\right)^2. \tag{3.2}$$

This is the correct conditional variance function for GWAS with no confounding bias. Since most published GWAS have taken steps to control for population stratification, The most likely use case will be a GWAS with at most a small amount of population stratification.

Simulations Results

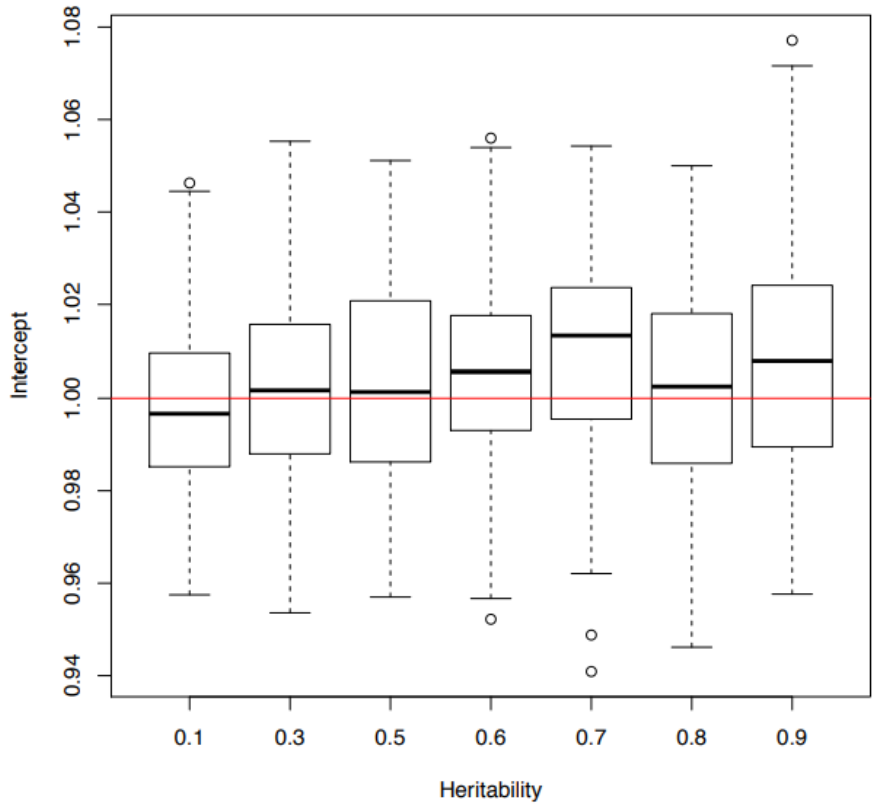
- 1.Simulations with polygenic genetic architectures （The ideal case）
- Goal： To verify the relationship between LD and χ^2 statistics：
- Method： We assigned per-allele effect sizes drawn from the distribution N(0, h2/(2Mp(1 − p))) to varying numbers of causal variants and for varying heritabilities in an approximately unstructured cohort of 1,000 Swedes.
- Result： Polygenic Traits vs few causal variants：
- When there were few causal variants, the LD Score regression estimates were still unbiased meaning that this approach is best suited to polygenic traits， but the standard errors became very large.

Simulations Results:

Polygenic Traits vs few causal variants case

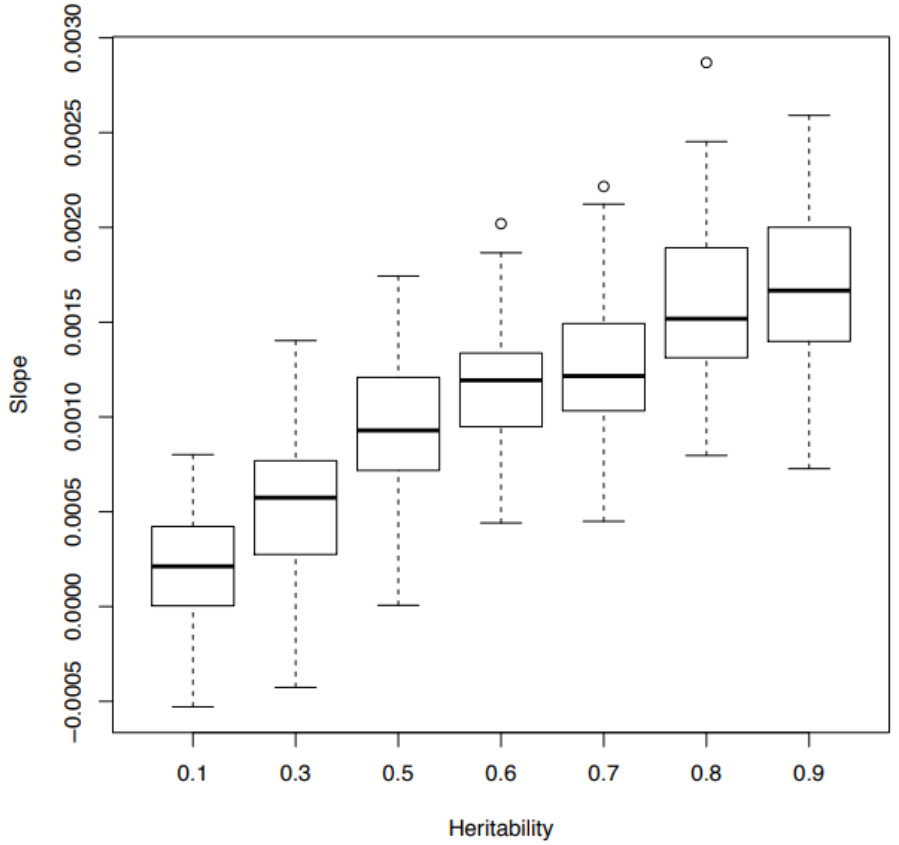
Supplemental Figures

Supplementary Figure 1: Intercepts from simulations with varying heritability



The *x*-axis displays different heritabilities specified for simulations, and the *y*-axis displays LD Score regression intercepts from 100 simulation replicates for each value of heritability. The red line shows the expected LD Score regression intercept in the absence of confounding bias. For all simulations, 1% of SNPs were causal.

Supplementary Figure 2: Slopes from simulations with varying heritability

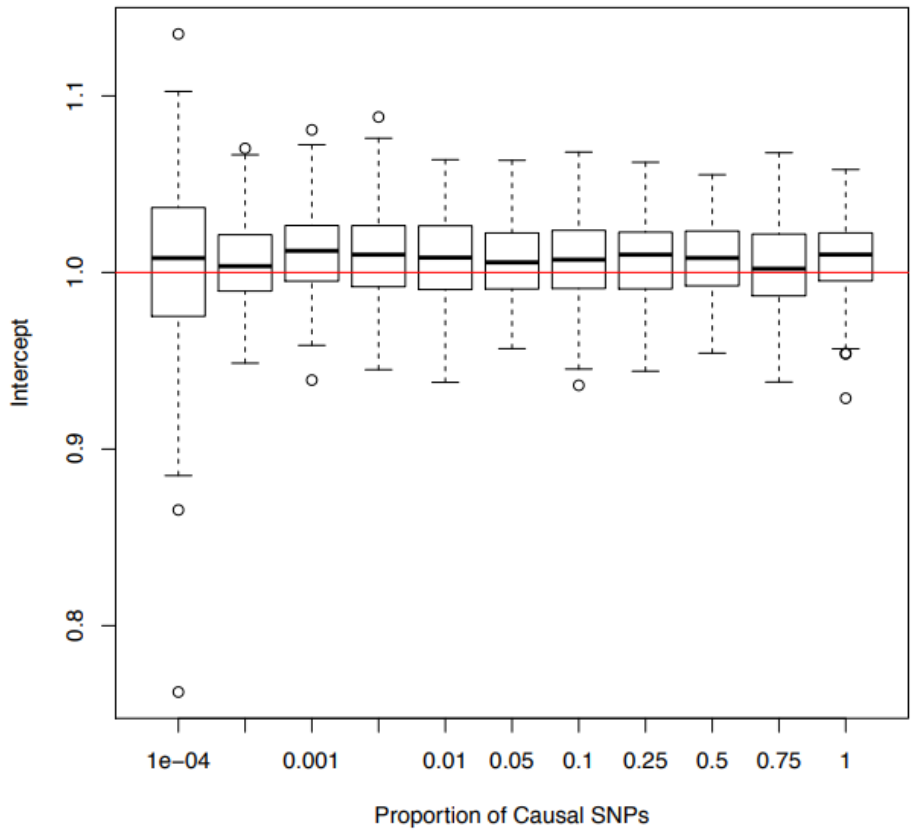


The *x*-axis displays different heritabilities specified for simulations, and the *y*-axis displays LD Score regression slopes from 100 simulation replicates for each value of heritability. For all simulations, 1% of SNPs were causal.

Simulations Results:

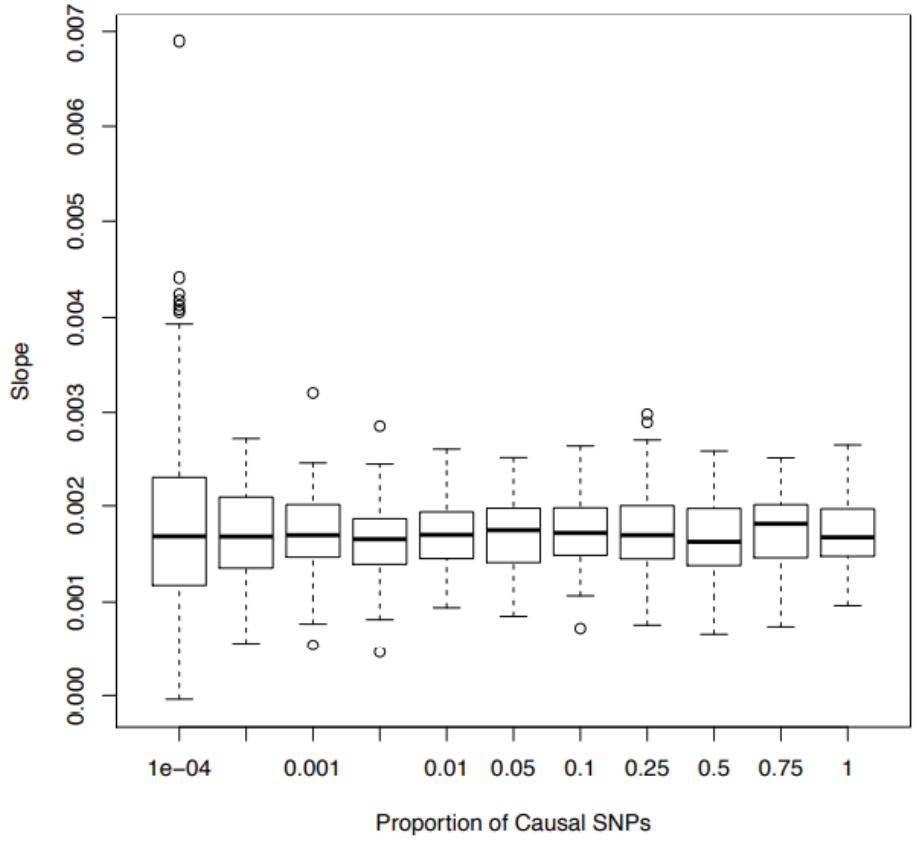
Polygenic Traits vs few causal variants case

Supplementary Figure 3: Intercepts from simulations with various proportions of causal SNPs



The *x*-axis displays different proportions of causal SNPs specified for simulations, and the *y*-axis displays LD Score regression intercepts from 100 simulation replicates for each value of the proportion of causal SNPs. For all simulations, the heritability was 0.9.

Supplementary Figure 4: Slopes from simulations with various proportions of causal SNPs



The *x*-axis displays different proportions of causal SNPs specified for simulations, and the *y*-axis displays LD Score regression slopes from 100 simulation replicates for each value of the proportion of causal SNPs. For all simulations, the heritability was 0.9.

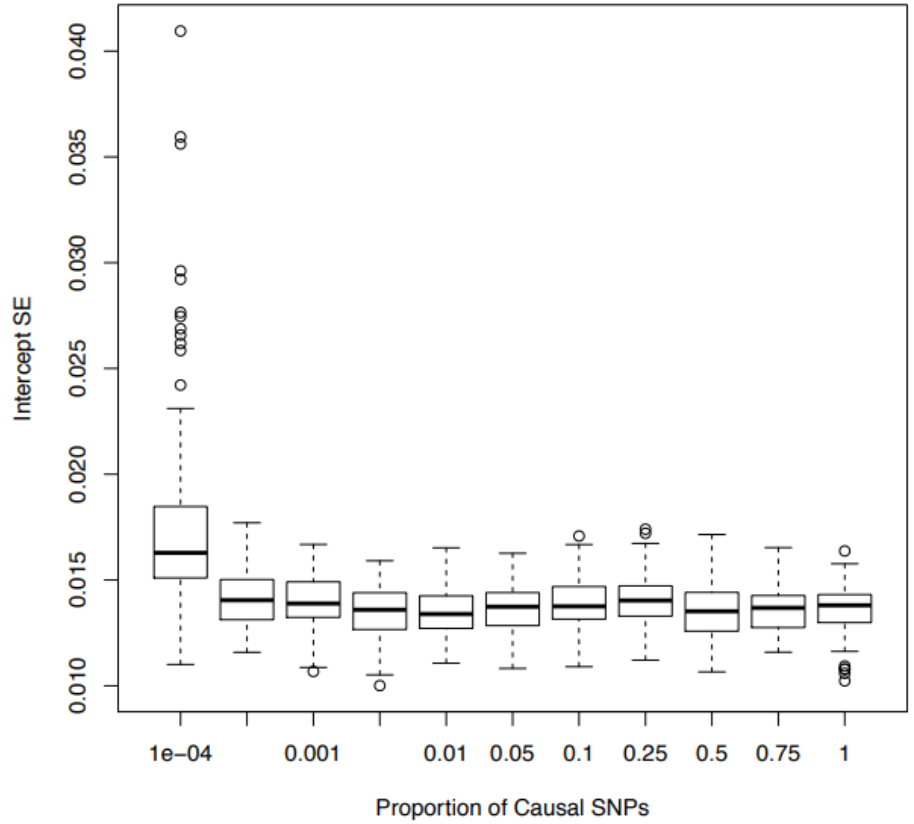
Simulations Results：

Polygenic Traits vs few causal variants

case

- When causal SNPs proportion is low， the SE is large；

Supplementary Figure 5: Estimated standard error from simulations with various proportions of causal SNPs



The *x*-axis displays different proportions of causal SNPs specified for simulations, and the *y*-axis displays block jackknife estimates of the standard error of the intercept from each of 100 simulation replicates for each proportion of causal SNPs. For all simulations, the heritability was 0.9.

Simulations with confounding

- Background：The model assumes that there is no systematic correlation between FST and LD Score. This assumption may be violated in practice as a result of linked selection (positive selection and background selection).
- Goal and the Problem：
- Problem：If there were a positive correlation between LD Score and FST, the LD Score regression intercept would underestimate the contribution of population stratification to the inflation in χ^2 statistics.
- Goal：To quantify the bias that this might introduce into the LD Score regression intercept, we performed a series of simulations with real population stratification.

Simulations with confounding

Method：We obtained unimputed genotypes for Psychiatric Genomics Consortium (PGC) controls from seven European cohorts genotyped on the same array：

Supplementary Table 2: Descriptions of cohorts for simulations with population stratification

Abbreviation	Origin	Principal Investigator	Controls
clo3	Cardiff, UK	Walters, J	945
cou3	UK	O'Donovan, M	544
egcu	Estonia	Esko, T	1,177
swe5	Sweden	Sullivan, PF	2,617
swe6	Sweden	Sullivan, PF	1,219
umeb	Umeå, Sweden	Adolfsson, R	584
umes	Umeå, Sweden	Adolfsson, R	713

Supplementary table 1 describes the seven PGC Schizophrenia control cohorts used for simulation with population stratification. All cohorts were genotyped on the Illumina Omni Express array; only unaffected individuals (controls) and directly genotyped SNPs post-QC (between approximately 600,000 and 700,000 SNPs, depending on cohort) were retained for simulations. In total genotypes for 9,135 individuals were incorporated into the simulations with pure population stratification

Simulations with confounding

Method：To simulate population stratification on a continental scale, we assigned case or control status on the basis of cohort membership and then computed association statistics for each pair of cohorts.

Method：To simulate population stratification on a national scale, we computed the top three principal components within each cohort and then computed association statistics using each of these principal components as a phenotype.

Results：The observed correlations between F_{ST} and LD Score in all simulations were negligible.

We also note that, in simulations with population stratification, the slope of the LD Score regression was slightly greater than zero on average, likely as a result of linked selection.

Simulations with confounding

- The observed correlations between FST and LD Score in all simulations were negligible.

Supplementary Table 3b. Correlation between LD Score and FST in simulations with continental-scale population stratification

Population 1	Population 2	Signed R-squared
cou3	clo3	5.00e-05
egcu	clo3	8.28e-04
egcu	cou3	7.41e-04
swe5	clo3	1.61e-04
swe5	cou3	2.02e-04
swe5	egcu	4.29e-04
swe6	clo3	1.81e-04
swe6	cou3	-1.12e-05
swe6	egcu	4.32e-04
swe6	swe5	5.95e-05
umeb	clo3	1.07e-04
umeb	cou3	5.05e-05
umeb	egcu	2.55e-04
umeb	swe5	1.92e-06
umeb	swe6	-6.60e-05
umes	clo3	5.22e-06
umes	cou3	6.79e-09
umes	egcu	7.26e-05
umes	swe5	-2.19e-06
umes	swe6	-5.23e-06
umes	umeb	-7.52e-06
	Mean	**2.51e-4**

Column descriptions. Population 1 and population 2 are the two populations involved in the simulations. Signed R-squared is the squared Pearson correlation coefficient between F_{ST} (between the two populations in the simulation) and LD Score multiplied by the negative one if the (non-squared) correlation is negative.

Supplementary Table 4b. Correlation between LD Score and FST in simulations with national-scale population stratification

Population	PC	Signed R-Squared
clo3	1	-2.96e-04
clo3	2	1.85e-05
clo3	3	5.03e-06
cou3	1	2.71e-05
cou3	2	4.75e-05
cou3	3	4.73e-05
egcu	1	3.74e-05
egcu	2	1.17e-04
egcu	3	3.23e-05
swe5	1	1.94e-04
swe5	2	8.46e-05
swe5	3	8.79e-08
swe6	1	3.37e-05
swe6	2	1.30e-04
swe6	3	4.49e-05
umeb	1	4.49e-05
umeb	2	1.44e-05
umeb	3	2.25e-05
umes	1	1.44e-04
umes	2	2.18e-05
umes	3	9.09e-05
	Mean:	**4.10e-05**

Column descriptions. Population is the population and PC is the principal component used to simulate population stratification in the simulations and population 2. Signed R-squared is the squared Pearson correlation coefficient between F_{ST} (between the two populations in the simulation) and LD Score multiplied by the negative one if the (non-squared) correlation is negative.

Simulations with confounding

- In simulations with population stratification, the slope of the LD Score regression was slightly greater than zero on average, likely as a result of linked selection.

Supplementary Table 3c Heritability and intercept for a confounded GWAS with continental-scale population stratification

Population 1	Population2	Heritability	Intercept	Lambda
cou3	clo3	0.140	1.397	1.531
egcu	clo3	0.063	1.454	1.509
swe6	clo3	0.033	1.476	1.502
swe5	clo3	0.034	1.475	1.502
umeb	clo3	0.027	1.480	1.484
umes	clo3	0.009	1.493	1.487
swe5	cou3	0.044	1.468	1.506
umes	cou3	0.007	1.495	1.429
egcu	cou3	0.063	1.454	1.504
swe6	cou3	-0.096	1.570	1.399
umeb	cou3	0.026	1.481	1.418
swe5	egcu	0.048	1.465	1.502
umes	egcu	0.027	1.480	1.456
swe6	egcu	0.051	1.463	1.503
umeb	egcu	0.051	1.463	1.443
swe6	swe5	0.031	1.477	1.483
umes	swe5	0.002	1.498	1.465
umeb	swe5	0.007	1.495	1.431
umeb	swe6	-0.213	1.656	1.208
umes	swe6	-0.001	1.500	1.461
umes	umeb	-0.005	1.504	1.498
	Mean	**0.017**	**1.488**	**1.463**

This table puts the slopes from the simulations with continental-scale population stratification on an interpretable scale by transforming all parameters to the scale of a GWAS with 100,000 samples and mean chi-square of 1.5, where all inflation in the mean chi-square comes from population stratification. All estimates of h2(1kG) use M=15 million. For comparison, the aggregate LD Score estimator of h2(1kG), $\widehat{h^2} = \frac{M(\overline{\chi^2}-1)}{N\bar{\ell}}$, which is representative of heritability estimators that are highly susceptible to population stratification, would give a heritability estimate of **0.68** in all cases, assuming mean LD Score = 110. The reason why the LD Score regression slope is not equal to zero is likely because linked selection introduces a small correlation between LD Score and F_{ST}. The conclusion is that in a worst-case scenario (pure population stratification), LD Score regression misattributes on average a small proportion of stratification to heritability, but nevertheless performs many times better than existing estimators (upward bias of 0.017 for LD Score regression vs. approximately 0.68 for other methods).

Supplementary Table 4c: Heritability and intercept for a confounded GWAS with national-scale population stratification

Population	PC	Heritability	Intercept	Lambda
clo3	1	-0.178	1.630	1.347
clo3	2	0.048	1.465	1.404
clo3	3	0.031	1.477	1.471
cou3	1	0.144	1.395	1.505
cou3	2	0.254	1.313	1.450
cou3	3	0.229	1.332	1.467
egcu	1	0.035	1.474	1.506
egcu	2	0.068	1.450	1.494
egcu	3	0.043	1.468	1.389
swe5	1	0.039	1.472	1.473
swe5	2	0.076	1.444	1.491
swe5	3	0.024	1.483	1.469
swe6	1	0.017	1.488	1.502
swe6	2	0.036	1.474	1.488
swe6	3	0.047	1.466	1.487
umeb	1	0.031	1.477	1.501
umeb	2	0.014	1.490	1.491
umeb	3	0.042	1.469	1.498
umes	1	0.047	1.466	1.505
umes	2	0.037	1.473	1.510
umes	3	0.082	1.440	1.490
	Mean	**0.056**	**1.459**	**1.473**

This table is similar to supplementary table 2c it puts the slopes from the simulations with national-scale population stratification on an interpretable scale by transforming all parameters to the scale of a GWAS with 100,000 samples and mean chi-square of 1.5 where all inflation in the mean chi-square comes from population stratification along the relevant principal component. As in supplementary table 4, the aggregate estimator would give a h2(1kG) estimate of 0.68, which is similar to the result that one would obtain with Haseman-Elston regression or linear mixed models (using M=15 million). The conclusions are similar to supplementary table 4.

Simulations with confounding and polygenicity

- Goal: To simulate a more realistic scenario where both polygenicity and bias contribute simultaneously to the inflation of test statistics,
- Method:
- We obtained the genotypes for approximately 22,000 individuals throughout Europe from the Wellcome Trust Case Control Consortium.
- We simulated polygenic phenotypes by drawing causal SNPs only from the first halves of chromosomes.
- All SNPs on the second halves of chromosomes were not causal. So In this setup, the mean $\chi 2$ statistic among SNPs on the second halves of chromosomes measures the average contribution of stratification bias.)
- We included an environmental stratification component aligned with the first principal component of the genotype data, representing northern versus southern European ancestry.

Simulations with confounding and polygenicity

- Results:
- In all simulation replicates, the LD Score regression intercept was approximately equal to the mean χ^2 statistic among null SNPs (Supplementary Table 5), which demonstrates that LD Score regression can partition the inflation in test statistics, even in the presence of both bias and polygenicity.

Supplementary Table 5: Simulations with both bias and polygenicity

Bias	Intercept (SD)	Null $\bar{\chi}^2$ (SD)	Null $\bar{\chi}^2$/Intercept (SD)
Relatedness	1.46 (0.02)	1.45 (0.02)	1.00 (0.00)
Stratification	1.53 (0.17)	1.48 (0.15)	0.97 (0.01)

Column descriptions. The column labeled bias identifies the source of bias, either cryptic relatedness (from the Framingham Heart Study) or population stratification (from introducing an environmental stratification term correlated with the first PC of the WTCCC2 data). Intercept is LD Score regression intercept, with the standard deviation (SD) across five simulations in parentheses. Null $\bar{\chi}^2$ is the mean χ^2 among SNPs on the opposite halves of chromosomes from causal SNPs, with SD across five simulations in parentheses. Since null SNPs are not in LD with causal SNPs, the mean χ^2 among null SNPs precisely quantifies the mean inflation in χ^2-statistics that results from bias. Null $\bar{\chi}^2$/Intercept is equal to the mean χ^2 among null SNPs divided by the LD Score regression intercept, with the SD across five simulations in parentheses. Null $\bar{\chi}^2$/Intercept should be approximately equal to one if the LD Score regression intercept is accurately estimating the mean inflation in test statistics that results from bias.

Real data

- Method: Finally, we applied LD Score regression to summary statistics from GWAS representing more than 20 different phenotypes.
- Results:
- (1)For all studies, the slope of the LD Score regression was significantly greater than zero.
- (2)For all studies, the LD Score regression intercept was substantially less than λ_{GC}.
- Conclusion:
- (1) Suggesting that polygenicity accounts for a majority of the increase in the mean χ^2 statistic
- (2) Confirming that correcting test statistics by dividing by λ_{GC} is unnecessarily conservative.

Real data

- As an example, we show the LD Score regression for the most recent schizophrenia GWAS, restricted to ~70,000 European-ancestry individuals

- The low intercept of 1.07 indicates at most a small contribution of bias and that the mean χ^2statistic of 1.613 results mostly from polygenicity.

Conclusion

- (1)In conclusion, we have developed LD Score regression, a method to distinguish between inflated test statistics from confounding bias and polygenicity.
- (2)Application of LD Score regression to over 20 complex traits confirms that polygenicity accounts for the majority of inflation in test statistics for GWAS results.
- (3)We propose that the LD Score regression intercept provides a more robust quantification of the extent of the confounding bias from inflation than λ_{GC}.
- (4)This approach can be used to generate a correction factor for GWAS that retains more power than λ_{GC}, especially with large sample sizes.

Thank you!