LD Score regression as an estimator of confounding and genetic correlations in genome-wide association studies

James J. Lee[1*]
Carson C. Chow[2*]

[1]Department of Psychology
University of Minnesota Twin Cities
75 East River Parkway
Minneapolis, MN 55455, USA
(612) 625-4980

[2]Mathematical Biology Section
Laboratory of Biological Modeling, NIDDK
National Institutes of Health
10 Center Drive
Bethesda, MD 20892, USA
(301) 402-8250

[*]To whom correspondence should be addressed;
E-mail: leex2293@umn.edu, carsonc@niddk.nih.gov.

ARTICLE

RUNNING HEAD: LD Score regression

The authors declare no conflict of interest.

Abstract

In order to infer that a single-nucleotide polymorphism (SNP) either affects a phenotype or is linkage disequilibrium with a causal site, we must have some assurance that any SNP-phenotype correlation is not the result of confounding with some environmental variable that also affects the trait. Here we provide a mathematical analysis of LD Score regression, a recently developed method for using summary statistics from genome-wide association studies (GWAS) to ensure that confounding does not inflate the number of false positives. We do not treat the effects of genetic variation as a random variable and thus are able to obtain results about the unbiasedness of this method. We demonstrate that LD Score regression can produce estimates of confounding at null SNPs that are either unbiased or overly conservative under fairly general conditions. This robustness holds even in cases now thought to be unfavorable, such as a correlation over SNPs between LD Scores and the degree of confounding. LD Score regression is thus an even stronger technique for causal inference than foreseen by its developers. Additionally, we demonstrate that LD Score regression can produce unbiased estimates of the genetic correlation, even when its estimates of the genetic covariance and the two univariate heritabilities are substantially biased.

Key Words: causal inference; heritability; population stratification; quantitative genetics

1 Introduction

The goal of genome-wide association studies (GWAS) is to find loci in the genome where variation affects a phenotype. However, this must be accomplished from observed correlations, and inferring causation from correlation is a famously perilous endeavor (Freedman, 1999; Pearl, 2009). GWAS has been fortunate in that it offers a variety of methods to check whether confounding effects have produced spurious correlations between genetic and phenotypic variation. These methods have led to a strong consensus that confounding has a minimal impact on GWAS results (Goldstein, 2011; Visscher, Brown, McCarthy, & Yang, 2012; Lee, 2012; Lee, Vattikuti, & Chow, 2016).

One of the newer methods used to check the causal status of GWAS associations is known as LD Score regression (Bulik-Sullivan et al., 2015b), which can be applied to summary statistics assembled from the contributions of different research groups and thus does not require access to individual-level data. This ingenious technique relies on the simple linear regression of assayed single-nucleotide polymorphism (SNP) j's association chi-square statistic on

$$l_j = \sum_k \Gamma_{jk}^2, \tag{1}$$

the sum over all SNPs of each SNP's squared correlation with the focal SNP j. This latter quantity is called SNP j's "LD Score." Empirically, the regression curve relating chi-square statistics to LD Scores is always very close to an upwardly sloping straight line. This result is explicable because a SNP tagging more of its neighbors—and, thus, having a higher LD Score—is more likely to tag one or more causal sites affecting the phenotype. The lowest possible LD Score of a SNP is in fact one, which is obtained when a SNP is in perfect linkage equilibrium (LE) with all other SNPs. A hypothetical SNP with an LD Score of zero, then, fails to tag the causal effect of any SNP in the genome—including

whatever effect the SNP itself may have. Therefore, if the intercept of LD Score regression departs upward from unity (the theoretical expectation of the chi-square distribution with one degree of freedom), the departure must be due to confounding, poor quality control, overlapping samples in the meta-analysis, or other artifacts. This simple and insightful method of estimating the distribution of truly null SNPs should in most cases lead to a much better global correction of the association statistics than the overly conservative genomic control (Devlin & Roeder, 1999).

The slope obtained from LD Score regression could in principle also provide an estimate of the trait's heritability, although the developers do not recommend this particular use of the method. We will show in detail why LD Score regression is not a reliable estimator of heritability below.

Another use of LD Score regression is the estimation of genetic correlations (Bulik-Sullivan et al., 2015a). The dependent variable in this case is not the chi-square statistic from the GWAS of a single trait but rather the product of two Z statistics, each taken from a GWAS of a distinct trait. In principle, this use offers a means of determining whether a trait-trait correlation (as opposed to a SNP-trait correlation) is attributable to the presence of confounders affecting both traits. If the genetic correlation is statistically and quantitatively significant, then we can be sure that the total phenotypic correlation is not attributable solely to confounders that are entirely environmental in nature. Many interesting relationships have been confirmed or discovered by bivariate LD Score regression, including a high genetic correlation ($\sim$0.70) between years of education and age at first childbirth (Barban et al., 2016) and a moderate one ($\sim$0.35) between years of education and intracranial volume (Okbay et al., 2016).

In the classical era of quantitative genetics, genetic correlations were most commonly estimated with twin data. Rather large samples of twinships are required for precise esti-

mates with this design, and in some cases the estimates are not as robust against modeling assumptions as estimates of univariate heritabilities (Beauchamp, Cesarini, Johannesson, Lindqvist, & Apicella, 2011). For these reasons a welcome development in quantitative genetics has been the advent of GWAS, which can now reach sample sizes in the millions. The appearance of robustness offered by GWAS can be illusory, however, if estimates of genetic correlations are themselves subject to confounding. One can devise estimators of the genetic correlation that might be biased by environmental confounders that affect both phenotypes and happen to be correlated with genetic variation (Palla & Dudbridge, 2015; Okbay et al., 2016). An attractive feature of LD Score regression in this respect is that that its control of confounding extends not just to the evidence of association at individual SNPs but also to its genome-wide estimates of genetic correlations. This is important because, again, it is precisely the issue of a phenotypic correlation's underlying causal nature that can call for an accurate estimate of the genetic correlation.

As appealing as the intuition behind LD Score regression may be, the mathematical justifications of this method given so far in the literature raise questions because of their assumption that the effects of genetic variants can be treated as a random variable. This assumption is a useful convenience for computations, but it is not biological; the effects of genetic polymorphisms should be invariant, and it is genotypes and phenotypic residuals that vary between individuals (Lee & Chow, 2014; de los Campos, Sorensen, & Gianola, 2015). The assumption also precludes a quantitative treatment of the method's accuracy. Here we refrain from this assumption of random genetic effects and instead treat the effects as a vector of arbitrary fixed constants. Hence we are able to obtain precise expressions of the quantities estimated by LD Score regression, which can be compared with the quantities of actual interest to determine when they coincide. Here is a preview of our results:

1. If the per-SNP heritability contributed by SNP j and its correlated neighbors is not related to SNP j's LD Score, then the slope of LD Score regression provides an unbiased estimate of heritability. For evolutionary reasons, however, per-SNP heritability is typically smaller near SNPs with higher LD Scores (Gazal et al., 2017). LD Score regression is therefore not a reliable way to estimate the heritability of a trait (or, by extension, the genetic covariance between two traits).

2. Here is the most novel and important conclusion of our analysis. The intercept of LD Score regression reflects a useful measure of confounding in the GWAS even if there is a relationship between LD Scores and the correlations of SNPs with environmental confounders. The developers of LD Score regression warn that in this case the intercept will not accurately estimate the contribution of confounding to the GWAS statistics (Bulik-Sullivan et al., 2015b). If linearity characterizes the relevant relationships, however, then it is the *conditional* extent of confounding at just those SNPs neither affecting the trait nor in LD with any causal SNPs that contributes to the intercept. This is the only piece of information needed to correct the association statistics of null SNPs so that they follow the proper Type 1 error rate with respect to the hypothesis of no causality.

3. LD Score regression provides an accurate estimate of the genetic correlation between two traits, even if neither trait's heritability is well estimated.

We now substantiate these claims.

2 Materials and methods

Consider a meta-analytic sample of n individuals and p biallelic SNPs. The standard linear model of quantitative genetics is

$$y = X\alpha + e, \tag{2}$$

where $y \in \mathbb{R}^n$ is the vector of standardized phenotypes, $\alpha \in \mathbb{R}^p$ is a vector of fixed constants equaling the average effects of gene substitution (Fisher, 1941; Lee & Chow, 2013), $e \in \mathbb{R}^n$ is the vector of non-genetic residuals, and $X \in \mathbb{R}^{n \times p}$ is the matrix of standardized genotypes. From these definitions the heritability of the phenotype attributable to the average effects of the p SNPs is $h^2 = (1/n)\alpha' X' X \alpha$, although LD Score regression uses the definition $h^2 = \alpha'\alpha$. These two definitions coincide if all causal sites are in linkage equilibrium (LE). As a result of LD induced by assortative mating and natural selection, this condition will often fail to be satisfied, but the resulting discrepancy between is likely to be small (Tenesa, Rawlik, Navarro, & Canela-Xandri, 2016). Henceforth we will mostly ignore the distinction between these two quantities (and similar distinctions that arise in the consideration of the genetic correlation).

We consider two different types of averages: 1) the expectation over individuals and 2) the empirical average over some attribute of SNPs, such as their GWAS association statistics, represented by the symbols $\mathbb{E}_n$ and $\mathbb{E}_p$ respectively. With this convention, X

and e are random variables with the properties

$$\begin{aligned}
\mathbb{E}_n(e_i) &= 0, \\
\mathbb{E}_n(X_{ij}) &= 0, \\
\mathbb{E}_n(X_{ij}^2) &= 1, \\
\mathbb{E}_n(X_{ij}X_{ik}) &= \Gamma_{jk}, \\
\mathbb{E}_n(X_{ij}e_i) &= v_j.
\end{aligned} \tag{3}$$

The last condition represents confounding due to a correlation between SNP j and e. Note that our representation of confounding as a correlation between a SNP and the non-genetic residual e is extremely general, including as a special case the sampling of the individuals from different geographically defined subpopulations varying in allele frequencies and exposures to environmental causes. We will use γ_j to denote the jth column (row) of $\Gamma \in \mathbb{R}^{p \times p}$, such that the jth LD Score is equal to $l_j = \gamma_j' \gamma_j$.

Different populations, such as Europeans and East Asians, are characterized by different values of Γ. We assume throughout this work that the individuals studied in the GWAS can be regarded as members of the same population as the reference sample used to estimate Γ.

Any contribution to the chi-square statistic of a given SNP from a causal site not included in the computation of its LD Score will introduce some form of bias. Such omissions from Equation (1) might occur because the windows used in practice to compute LD Scores are too short or because some causal sites have properties that lead to their exclusion from the reference sample (rare alleles or being a type of polymorphism other than a SNP). Although it may be worthwhile to analyze these and other limitations, we do not do so here.

3 Results

3.1 The slope of univariate LD Score regression as an estimator of heritability

Although Bulik-Sullivan et al. (2015b) do not encourage using the slope of the (χ_j^2, l_j) regression as a heritability estimator, it is useful to see in further detail why, not least because we will reuse our primary result later. Let x_j be the jth column of X. In the regression of the GWAS phenotype on a single SNP j, the estimated marginal (univariate) regression coefficient is $\hat{\beta}_j = (1/n)x_j'y = (1/n)y'x_j$. Note that in the absence of confounding, the average effects of gene substitution can be estimated by the multivariate regression coefficient $\hat{\alpha} = (X'X)^{-1}X'y$ (Fisher, 1941; Lee & Chow, 2013). Squaring gives

$$\begin{aligned}\hat{\beta}_j^2 &= \frac{1}{n^2}x_j'yy'x_j \\ &= \frac{1}{n^2}x_j'(X\alpha + e)(X\alpha + e)'x_j,\end{aligned}$$

which has the expected value over random sampling of individuals

$$\mathbb{E}_n(\hat{\beta}_j^2) = \frac{1}{n^2}\mathbb{E}_n\left(x_j'X\alpha\alpha'X'x_j + x_j'X\alpha e'x_j + x_j'e\alpha'X'x_j + x_j'ee'x_j\right). \tag{4}$$

The problem with evaluating Equation (4) is that the fourth moment of the genotypes is required and it is generally not known. However, if we assume that higher-order cumulants of the genotype distribution are small compared to the second cumulants, then the distribution governing the genotypes can be approximated with a multivariate normal distribution. We can then use Wick's theorem (sometimes called Isserlis's theorem), which states that if $(X_1, \ldots, X_{2n})$ follows a zero-mean multivariate normal distribution, then

$$\mathbb{E}_n(X_1X_2\cdots X_{2n}) = \sum\prod \mathbb{E}_n(X_iX_j),$$

where the notation $\sum\prod$ means summing over all distinct ways of partitioning $X_1, \ldots, X_{2n}$ into pairs such as $X_i X_j$ and each summand is the product of pair expectations.

Applying Wick's theorem to the first expectation term of Equation (4) yields

$$\begin{aligned}
&\sum_{k,k',i,i'} \alpha_k \alpha_{k'} \mathbb{E}_n \left(X_{ij} X_{ik} X_{i'j} X_{i'k'} \right) \\
&\approx \sum_{i,i',k,k'} \alpha_k \alpha_{k'} [\mathbb{E}_n(X_{ij}X_{ik})\mathbb{E}_n(X_{i'j}X_{i'k'}) + \mathbb{E}_n(X_{ij}X_{i'j})\mathbb{E}_n(X_{ik}X_{i'k'}) \\
&+ \mathbb{E}_n(X_{ij}X_{i'k'})\mathbb{E}_n(X_{ik}X_{i'j})]] \\
&= n^2 \sum_{k,k'} \Gamma_{jk}\alpha_k \Gamma_{jk'}\alpha_{k'} + \sum_{i,i',k,k'} \mathbb{E}_n(X_{ij}X_{i'j})\alpha_k \mathbb{E}_n(X_{ik}X_{i'k'})\alpha_{k'} + \sum_{k,k'} n\Gamma_{jk'}\alpha_{k'}\Gamma_{kj}\alpha_k \\
&= n^2\beta_j^2 + \sum_{i,k,k'} \mathbb{E}_n(X_{ij}X_{ij})\mathbb{E}_n(X_{ik}X_{ik'})\alpha_k\alpha_{k'} + n\beta_j^2 \\
&\approx (n^2+n)\beta_j^2 + nh^2,
\end{aligned}$$

where we have applied Equation (3) and the identity $\beta_j = \gamma_j'\alpha$. The latter is true in the large-n limit or by the path-tracing rules (Wright, 1934). The last line assumes that $\sum_{k\neq k'} \Gamma_{kk'}\alpha_k\alpha_{k'}$, the term distinguishing the two definitions of heritability, is small; recall our assumption that these two definitions lead to numerically close values.

Similarly, the expected sum of the second, third, and fourth terms in Equation (4) is

$$\begin{aligned}
&\sum_k 2n^2\alpha_k\Gamma_{jk}v_j + \sum_k 2n\alpha_k v_j \Gamma_{kj} + \sum_k 2n\alpha_k v_k \\
&\quad + n^2 v_j^2 + n v_j^2 + n\left[1 - h^2 - 2\mathrm{Cov}_n\left(\sum_k X_{ik}\alpha_k, e_i\right)\right] \\
&= 2n^2\beta_j v_j + 2n\beta_j v_j + n^2 v_j^2 + n v_j^2 + n(1-h^2).
\end{aligned}$$

Substituting all terms back into Equation (4) and assigning $\chi_j^2 = n\hat{\beta}_j^2$ gives

$$\begin{aligned}
\mathbb{E}_n(\chi_j^2) &\approx (n+1)\beta_j^2 + 1 + 2(n+1)\beta_j v_j + (n+1)v_j^2 \\
&\approx n\beta_j^2 + n v_j^2 + 2n\beta_j v_j + 1 \\
&= n\alpha'\alpha\cos^2\theta_j l_j + n v_j^2 + 2n|\alpha|\cos\theta_j\sqrt{l_j}v_j + 1. \qquad (5)
\end{aligned}$$

Here we have used

$$\beta_j^2 = (\gamma_j'\alpha)^2 = \gamma_j'\gamma_j\alpha'\alpha\cos^2\theta_j,$$

where θ_j is the angle between γ_j and α. Hence, the square of the estimated marginal regression coefficient equals the sum of the following quantities:

1. the square of the regression coefficient induced by any true average effects of gene substitution;
2. the square of the bias induced by confounding;
3. twice the cross-product of the true coefficient and the bias; and
4. sampling noise with a variance equal to $1/n$.

We now consider the conditions under which the slope of the (χ_j^2, l_j) regression is proportional to h^2. We can compute this explicitly by using Equation (5) in the formula for the regression coefficient. However, a more informative way is to compare to the analogous expression in Bulik-Sullivan et al. (2015b), which in our notation is

$$\mathbb{E}_p(\chi_j^2 \,|\, l_j) \approx \frac{n}{p} l_j h_{\text{LDSC}}^2 + n\mathbb{E}_p(v_j^2) + 1. \tag{6}$$

Our placement of the subscript LDSC on h^2 emphasize that this factor in the regression slope might not necessarily equal h^2. In the case of $v = 0$, the equivalence of (6) to the average of (5) over all SNPs implies

$$h_{\text{LDSC}}^2 = \alpha'\alpha p\mathbb{E}_p(\cos^2\theta_j), \tag{7}$$

which gives a biased estimate of heritability unless $\mathbb{E}_p(\cos^2\theta_j) = 1/p$. This condition can hold if the γ_j are uniformly distributed with respect to α. Thus, the slope of LD Score

regression is proportional to the heritability if the average effects of gene substitution and LD Scores are uncorrelated.

The requirement of this null correlation for an unbiased estimate of h^2 is quite reasonable. Regressing χ_j^2 on l_j to estimate the heritability depends on a constant average per-SNP heritability regardless of LD. If average per-SNP heritability declines in higher-LD regions, say, then the estimated heritability must fall short of the true heritability. This sensitivity to LD is a feature shared with the heritability-estimation method GREML (Speed, Hemani, Johnson, & Balding, 2012; Lee & Chow, 2014; Yang et al., 2015; Chen, 2016).

However, a negative correlation between LD and heritability tagged per SNP is expected. Mutations with larger effects on a given trait will tend to be selectively disfavored as a result of stabilizing selection or deleterious pleiotropic side effects. Such mutations will thus rarely drift to high allele frequencies, and SNPs where one allele is rare tend to have smaller LD Scores. The empirical evidence to date clearly bears out this evolutionary prediction (Kemper, Visscher, & Goddard, 2012; Yang et al., 2015; Gazal et al., 2017). In this case of SNPs with higher LD Scores tagging less heritability, the slope of the (χ_j^2, l_j) regression leads to $h^2_{\text{LDSC}} < h^2$, an underestimation of the true heritability.

3.2 The intercept of univariate LD Score regression as an estimator of confounding

A far more important use of LD Score regression is the estimation and correction of confounding (or any other bias that can inflate the association statistics, such as overestimation of the effective sample size). If the intercept of LD Score regression is truly equal to the average chi-square statistic of null SNPs that neither affect the phenotype nor tag any causal sites, then dividing all of the GWAS chi-square statistics by the intercept

should restore the average chi-square statistic of null SNPs to the theoretically proper value of unity and bring the Type 1 error rate close to the targeted level.

We will now show that the intercept can be used for the correction of confounding in the situation where LD Scores and SNP-environment correlations are related (i.e., l_j and v_j^2 are correlated).

At $l_j = 0$, Equation (5) gives

$$\mathbb{E}_n(\chi_j^2 \,|\, l_j = 0) = nv_j^2 + 1. \tag{8}$$

Thus, if $\mathbb{E}_p(\chi_j^2 \,|\, l_j)$ is found to be linear in l_j (which is empirically observed), then the intercept of the (χ_j^2, l_j) regression will equal the average of Equation (8) over some set of SNPs. If v_j^2 is independent of l_j, then the intercept will equal the average over all SNPs, $n\mathbb{E}_p(v_j^2) + 1$. Otherwise, we can write for the intercept

$$n\mathbb{E}_p(v_j^2 \,|\, l_j = 0) + 1. \tag{9}$$

There are of course no SNPs with an actual LD Score of zero. Suppose, however, that the dependence of v_j^2 on l_j is linear. The most likely case of l_j dependence is a direct (non-genetic) effect of parent on offspring phenotype, such as when highly educated parents can help even their adopted offspring become highly educated in turn (Sacerdote, 2007). In this case v_j^2 indeed depends linearly on l_j to the extent that β_j^2 does so, because v_j^2 is equal to β_j^2 up to an attenuating constant (Lee, 2012). SNPs with $l_j = 1$ can tag at most one causal SNP, SNPs with $l_j = 2$ can tag more than one, and so on. The extrapolation $\mathbb{E}_p(v_j^2 \,|\, l_j = 0)$ in (9) is thus very close to the average squared confounding at null SNPs (i.e., SNPs that tag no causal SNPs), and this is all that is needed to rescale the chi-square statistics of these SNPs to bring their average in line with the theoretical value under the null hypothesis (i.e., no causality or LD with a causal site). Some impact of l_j dependence on factoring out the contribution of $2n\beta_j v_j + nv_j^2$ to the chi-square statistics of non-null

SNPs simply leaves us with more or less statistical power to detect such SNPs without affecting the Type 1 error rate.

It is certainly possible to create gross violations of linearity in simulations (Bulik-Sullivan et al., 2015b, Supplementary Fig. 7). For example, if we depopulate high-LD regions of causal SNPs, then the (β_j^2, l_j) regression curve can be non-monotonic, rising at first and then declining as l_j increases. In this case the slope of LD Score regression can be negative and the intercept greater than unity even in the absence of confounding ($v = 0$). However, no empirical application of LD Score regression has ever uncovered any situation remotely resembling this hypothetical one. Nevertheless it is a salutary practice to inspect the actual (χ_j^2, l_j) scatterplot for any evidence of pathology.

A mild degree of nonlinearity might have some effect on the intercept if the SNPs with largest LD Scores deviate from the linear trend extrapolated from the SNPs with the smallest LD Scores. For this reason it is fortunate that in practice LD Score regression is a weighted regression where the SNPs with the smallest LD Scores receive the largest weights. The purpose of this weighting is to address heteroskedasticity and non-independence; if the (χ_j^2, l_j) regression curve is perfectly linear, then the effect of this weighting is to improve the standard errors. If the curve is nonlinear, then an additional effect is to bring the entire regression line closer to the linear extrapolation from the SNPs with the smallest LD Scores and the intercept thereby closer to the average chi-square statistic of truly null SNPs.

This conclusion regarding the extraordinary robustness of LD Score regression as a safeguard against confounding is a novel result of our analysis. Bulik-Sullivan et al. (2015b) went to some lengths to show that LD Scores are uncorrelated with F_{ST} (a measure of population differentiation in allele frequencies) at various geographical scales within Europe. This is very convincing evidence in support of the assumption that confounding

is uncorrelated with LD Scores—at least when the confounding takes the form of "population stratification," the sampling of the individuals in the study from geographically distinct subpopulations differing in both allele frequencies and exposure to environmental factors. But even if confounding is correlated with LD Scores, what we find is that the intercept of LD Score regression can still be used to ensure that null SNPs have an average chi-square statistic of close to unity.

With all of these considerations in mind, we turn to the recent work of de Vlaming, Johannesson, Magnusson, Ikram, and Visscher (2017). These authors found that a very large $\mathbb{E}_p(v_j^2)$ in their simulations leads to an intercept falling short of $\mathbb{E}_p(v_j^2)$ itself and also an overestimate of h^2. These *in silico* results are rather puzzling because they were not replicated by Bulik-Sullivan et al. (2015b) despite apparently similar simulation settings. One possibility is that SNPs with larger LD Scores tend to exhibit higher F_{ST} in the cohorts available to de Vlaming et al. (2017), perhaps because of higher-quality imputation leading to more accurate estimates of allele-frequency differences. This would lead to both $h^2_{\text{LDSC}} > h^2$ and $\mathbb{E}_p(v_j^2) > \mathbb{E}_p(v_j^2 \,|\, l_j = 0)$. (Note that it is only the smaller of the two quantities in the latter inequality that is needed for the restoration of the Type 1 error rate.) Whatever the problem may be, evidence for it can be seen in the (χ_j^2, l_j) scatterplot, which shows a nonlinearity in the leftmost simulated data points that we have never observed in real empirical data. It is also worth noting that the problems in these simulations only arise when population stratification is quite extreme, leading to an intercept greater than 1.5 with rather small sample sizes. In this regime Wick's theorem may no longer provide a good approximation, although we think this unlikely to be the explanation of the simulation results. In any event intercepts of this magnitude have not yet been observed in actual GWAS.

3.3 Bivariate LD Score regression as an estimator of genetic correlations

We now consider LD Score regression as an estimator of the genetic correlation between the two traits

$$\begin{aligned} y_1 &= X_1\alpha_1 + e_1, \\ y_2 &= X_2\alpha_2 + e_2. \end{aligned} \tag{10}$$

We will use r_{LDSC} to denote the genetic correlation as it is estimated by bivariate LD Score correlation—which is not necessarily the same as the true genetic correlation $r := \alpha_1'\alpha_2/\sqrt{h_1^2\, h_2^2}$. Nevertheless, previous studies have found these two quantities to be consistently close (Bulik-Sullivan et al., 2015a; Shi, Mancuso, Spendlove, & Pasaniuc, 2017), and our goal now is to explain this robustness.

The dependent variable in bivariate LD Score regression is now the product of SNP j's two Z statistics,

$$\begin{aligned} n\hat{\beta}_{1j}\hat{\beta}_{2j} &= \frac{1}{n}x_j' y_1 y_2' x_j \\ &= \frac{1}{n}x_j'(X\alpha_1 + e_1)(X\alpha_2 + e_2)' x_j, \end{aligned}$$

which has the expected value

$$\mathbb{E}_n(Z_{1j}Z_{2j}) = \frac{1}{n}\mathbb{E}_n\left(x_j' X\alpha_1\alpha_2' X' x_j + x_j' X\alpha_1 e_2' x_j + x_j' e_1\alpha_2' X' x_j + x_j' e_1 e_2' x_j\right).$$

As before, we can use Wick's theorem to evaluate the expectation and obtain

$$\begin{aligned} \mathbb{E}_n(Z_{1j}Z_{2j}) \approx\ & n\beta_{1j}\beta_{2j} + n\beta_{1j}v_{2j} + n\beta_{2j}v_{1j} + nv_{1j}v_{2j} \\ & + \text{Cov}_n\left(\sum_k X_{ik}\alpha_{1k}, e_{2i}\right) + \text{Cov}_n\left(\sum_k X_{ik}\alpha_{2k}, e_{1i}\right) + \rho. \end{aligned}$$

where $\rho_g := \alpha_1'\alpha_2 = \alpha_2'\alpha_1$ is the genetic covariance, $\rho_e = \mathbb{E}_n(e_{i1}e_{i2})$ is the environmental covariance, and $\rho := \rho_g + \rho_e$. The last three terms arise from the coincidence of the person indices in the summations and thus become smaller with decreasing sample overlap. They vanish if the samples are independent. Henceforth we ignore these overlap-dependent terms. We are then left with

$$\begin{aligned}\mathbb{E}_n(Z_{1j}Z_{2j}) &\approx n\beta_{1j}\beta_{2j} + n\beta_{1j}v_{2j} + n\beta_{2j}v_{1j} + nv_{1j}v_{2j} \\ &= n\gamma_j'\alpha_1\gamma_j'\alpha_2 + n\beta_{1j}v_{2j} + n\beta_{2j}v_{1j} + nv_{1j}v_{2j}. \end{aligned} \tag{11}$$

In LD Score regression (regression of $Z_{1j}Z_{2j}$ on l_j), the slope is naively expected to be proportional to the genetic covariance. In the absence of confounding and sample overlap, the intercept is zero since the expected product of two independent and null-distributed Z statistics is zero. Any upward departure of the intercept from zero in this case is indicative of confounders affecting both traits, just as an upward departure from unity is analogously indicative of confounders affecting the focal trait in the univariate case.

As in the univariate case, we can compute the circumstances under which the regression slope is proportional to the genetic covariance explicitly using Equation (11) in the formula for the regression coefficient, but it is more informative to compare directly to the analogous expression from Bulik-Sullivan et al. (2015a),

$$\mathbb{E}_p(Z_{1j}Z_{2j} \,|\, l_j) \approx \frac{n}{p}\rho_{g,\mathrm{LDSC}}l_j. \tag{12}$$

Assume that $\beta_{1j}v_{2j}$, $\beta_{1j}v_{1j}$, and $v_{1j}v_{1j}$ are all uncorrelated with l_j; a total absence of confounding, $v_1 = v_2 = 0$, meets this assumption. We have found that the robustness of bivariate LD Score regression holds in certain importance cases of l_j dependence, such as a direct effect of parental phenotype discussed by Lee (2012), but these details are beyond

the scope of this work. The output of bivariate LD Score regression is then

$$r_{\text{LDSC}} = \frac{\rho_{g,\text{LDSC}}}{\sqrt{h^2_{1,\text{LDSC}}\, h^2_{2,\text{LDSC}}}}. \tag{13}$$

The average of (11) over all SNPs and (12) are equivalent if

$$\rho_{g,\text{LDSC}} \equiv \mathbb{E}_p\left(\frac{1}{p}\alpha_1'\alpha_2\gamma_j'\gamma_j\right) = \mathbb{E}_p(\gamma_j'\alpha_1\gamma_j'\alpha_2), \tag{14}$$

which we will show is not generally true. As in the univariate case above, the righthand side of Equation (14) can be rewritten as

$$\mathbb{E}_p(\gamma_j'\alpha_1\gamma_j'\alpha_2) = |\alpha_1||\alpha_2|\mathbb{E}_p(\cos\theta_j^1\cos\theta_j^2 l_j), \tag{15}$$

where $\cos\theta_j^k$ is the unit-vector projection of α_k onto γ_j. The average over SNPs in (15) is equivalent to taking the unit-vector projections of α_1 onto the γ_j in turn, doing the same with α_2, and taking the l_j-weighted dot product of the two results. From (15) we can see two sources of bias, which can be interpreted geometrically. The first is the nontrivial correlation between γ_j and α_k as in the univariate case and manifested as nonuniformity in $\cos\theta_j^k$. We will shortly see, however, that this bias cancels from the numerator and denominator of Equation (13). The second source of bias is that the γ_j vectors do not form an orthogonal basis over SNP space, which then distorts the angle between α_1 and α_2 after projecting onto the γ basis.

We will proceed as if the γ_j are indeed an orthogonal basis. In reality, they are nearly orthogonal; if the SNPs are numbered in order, then $\gamma_j'\gamma_k$ will be virtually zero for $|j-k|$ sufficiently large. Then the angle between α_1 and α_2 is preserved in the new basis, and we have the condition

$$\mathbb{E}_p(\cos\theta_j^1\cos\theta_j^2 l_j) = \sqrt{\mathbb{E}_p(\cos^2\theta_j^1)\mathbb{E}_p(\cos^2\theta_j^2)}\cos\theta_{12} l_j,$$

where θ_{12} is the angle between α_1 and α_2. We can then obtain

$$\rho_{g,\text{LDSC}} \approx \alpha_1' \alpha_2 p \sqrt{\mathbb{E}_p(\cos^2 \theta_j^1)\mathbb{E}_p(\cos^2 \theta_j^2)}. \tag{16}$$

Inserting (16) and (7) into (13) then gives

$$\begin{aligned} r_{\text{LDSC}} &\approx \frac{\alpha_1' \alpha_2 p \sqrt{\mathbb{E}_p(\cos^2 \theta_j^1)\mathbb{E}_p(\cos^2 \theta_j^2)}}{\sqrt{\alpha_1' \alpha_1 p \mathbb{E}_p(\cos^2 \theta_j^1) \alpha_2' \alpha_2 p \mathbb{E}_p(\cos^2 \theta_j^2)}} \\ &= \frac{\alpha_1' \alpha_2}{\sqrt{\alpha_1' \alpha_1 \alpha_2' \alpha_2}}, \end{aligned}$$

which is an unbiased estimator of the genetic correlation.

If, on the other hand, it is unacceptable to treat the γ vectors as an orthogonal basis, then LD Score regression will not produce an unbiased estimator of genetic correlation—at least when this quantity is defined as $\alpha_1' \alpha_2 / \sqrt{h_1^2 h_2^2}$. We can estimate the bias by considering the eigenvalue decomposition $S' \Gamma S = \Lambda$, where S is the orthonormal matrix with columns of eigenvectors and Λ is the diagonal matrix of eigenvalues. We then have

$$\begin{aligned} p\mathbb{E}_p(\gamma_j' \alpha_1 \gamma_j' \alpha_2) &= \sum_j \gamma_j' \alpha_1 \gamma_j' \alpha_2 \\ &= \alpha_1' \Gamma \Gamma' \alpha_2 \\ &= \alpha_1' S S' \Gamma S S' \Gamma' S S' \alpha_2 \\ &= \alpha_1' S \Lambda^2 S' \alpha_2. \end{aligned}$$

We now decompose $\Lambda^2 = \lambda^2 I + \Delta$ and obtain

$$p\mathbb{E}_p(\gamma_j' \alpha_1 \gamma_j' \alpha_2) = \lambda^2 \alpha_1' \alpha_2 + \alpha_1' S \Delta S' \alpha_2, \tag{17}$$

where λ^2 represents the average correlation of γ_j and α and Δ represents the deviation from orthogonality.

4 Discussion

The regression of GWAS association statistics on LD Scores partitions the statistics into a part that covaries with LD Scores (the slope) and a part that does not (the intercept). Polygenic causal signal contributes to the first part by necessity, whereas confounding and other biases spuriously inflating the statistics need not—and typically do not—make any such contribution. This insight lies at the heart of the LD Score regression, the outstanding invention of Bulik-Sullivan et al. (2015b).

The reason that the slope of LD Score regression cannot be used to estimate the heritability of a trait (or the genetic covariance between two traits) is that per-SNP heritability (genetic covariance) will itself vary as a function of LD Score, such that naive estimates based on LD Score regression will typically fall short of the target quantities.

In order for the intercept to equal the average squared covariance between SNP and residual ("environment") present in the GWAS (which can then be factored out from the association statistics), LD Scores must be uncorrelated over SNPs with squared SNP-residual covariance. In the framework of Bulik-Sullivan et al. (2015b), this is equivalent to the absence of a correlation between LD Scores and the F_{ST} characterizing the two subpopulations. There may be such a correlation, however, in certain cases such as when the phenotype of the parent affects the phenotype of the offspring through some environmental mechanism. Remarkably we found that LD Score regression remains a robust means of correcting the association statistics, for in such a case the intercept approaches the average squared confounding at just those SNPs that are neither causal themselves nor in LD with any causal sites—that is, at precisely those SNPs where otherwise an excess of false positives might occur.

These conclusions depend importantly on the linearity of the relationship between LD

Scores and the GWAS chi-square statistics (product of Z statistics). This is essentially because without linearity there is no guarantee that the intercept of a particular simple least-squares regression equals the conditional expected value of the dependent variable characterizing observations with a zero value of the independent variable. In real-data applications of LD Score regression to date, the (χ_j^2, l_j) scatterplots have always borne out approximate linearity, and they should continue to be inspected in future applications. When users follow the developers' recommendations for weighting of the SNPs in the regression, those SNPs with smaller LD Scores will receive larger weights, which in the case of nonlinearity brings the intercept closer to the conditional expected chi-square statistic of null SNPs.

Despite the inability of LD Score regression to estimate the heritability (genetic covariance) without bias, the method is able to estimate the genetic correlation quite accurately. Our argument on this point will be valid if the genetic correlation depends primarily on direct overlap of the causal sites affecting the two traits—and negligibly on SNPs in LD with more potential causal sites thereby being more likely to tag one site affecting trait 1 and a distinct site affecting trait 2, with the signs of the alleles coupled with the reference allele at the tagging SNP showing a consistency across the genome. This tagging of distinct sites with appropriately coupled alleles contributes to the second term of the genetic covariance in Equation (17), which is not a multiplicative bias and therefore cannot be canceled by any division in the calculation of the genetic correlation. Such a genome-wide pattern seems quite implausible; for example, if it is to create a misleading nonzero r_{LDSC} when r is in fact zero, it amounts to causal sites that affect the two traits occurring in the same genes and regulatory elements, with the appropriate coupling of alleles, but never coinciding. Furthermore, one might argue that this biologically implausible scenario does not necessarily invalidate r_{LDSC} as an estimator of r when the latter is defined

correctly. We have adopted the definition $r := \alpha_1' \alpha_2 / \sqrt{h_1^2 \, h_2^2}$ because this seems most consistent with the definition of heritability given in the original LD Score regression paper (Bulik-Sullivan et al., 2015b, Supplementary Note, p. 1), but other authors have included contributions from LD and consistent coupling of allele signs to the definition of r (Lynch & Walsh, 1998).

A use of LD Score regression that we did not study in this work is the functional partition of heritability between different parts of the genome (Finucane et al., 2015). Simulation studies conducted by the authors suggest that this use is also quite robust, and this is probably the result of a similar cancellation of bias from numerator and denominator.

In a field already marked by remarkable progress toward the goal of elucidating the causal relationship between its variables of interest without undue hindrance by confounding, LD Score regression adds a powerful new tool that allows whatever confounding there may be in a GWAS to be estimated and removed. In addition, it is a robust estimator of the genetic correlation, which is valuable in its own right because of its relevance to the causal nature of the phenotypic correlation (Duffy & Martin, 1994). It is fascinating to speculate about why the inference of causation of correlation has proven to be so eminently possible in genetics when it has been elusive in so many other scientific fields (Lee, 2012; Plomin, DeFries, Knopik, & Neiderhiser, 2016). Whatever the reasons, researchers in genetics can be grateful that Nature seems to be willing to oblige their curiosity.

Acknowledgements

This research was supported in part by the Intramural Research Program of the NIH, The National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK).

References

Barban, N., Jansen, R., de Vlaming, R., Vaez, A., Mandemakers, J., Tropf, F. C., ... Mills, M. C. (2016). Genome-wide analysis identifies 12 loci influencing human reproductive behavior. *Nature Genetics*, *48*(12), 1462–1472. doi:10.1038/ng.3698

Beauchamp, J. P., Cesarini, D., Johannesson, M., Lindqvist, E., & Apicella, C. (2011). On the sources of the height-intelligence correlation: New insights from a bivariate ACE model with assortative mating. *Behavior Genetics*, *41*(2), 242–252. doi:10.1007/s10519-010-9376-7

Bulik-Sullivan, B., Finucane, H. K., Anttila, V., Gusev, A., Day, F. R., Loh, P.-R., ... Neale, B. M. (2015a). An atlas of genetic correlations across human diseases and traits. *Nature Genetics*, *47*(11), 1236–1241. doi:10.1038/ng.3406

Bulik-Sullivan, B., Loh, P.-R., Finucane, H. K., Ripke, S., Yang, J., Schizophrenia Working Group of the Psychiatric Genomics Consortium, ... Neale, B. M. (2015b). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics*, *47*(3), 291–295. doi:10.1038/ng.3211

Chen, G.-B. (2016). On the reconciliation of missing heritability for genome-wide association studies. *European Journal of Human Genetics*, *24*(12), 1810–1816. doi:10.1038/ejhg.2016.89

de los Campos, G., Sorensen, D., & Gianola, D. (2015). Genomic heritability: What is it? *PLoS Genetics*, *11*(5), e1005048. doi:10.1371/journal.pgen.1005048

de Vlaming, R., Johannesson, M., Magnusson, P. K. E., Ikram, M. A., & Visscher, P. M. (2017). Equivalence of LD-score regression and individual-level-data methods. *bioRxiv*. doi:10.1101/211821

Devlin, B. & Roeder, K. (1999). Genomic control for association studies. *Biometrics*, *55*(4), 997–1004. doi:10.1111/j.0006-341X.1999.00997.x

Duffy, D. L. & Martin, N. G. (1994). Inferring the direction of causation in cross-sectional twin data: Theoretical and empirical considerations. *Genetic Epidemiology*, *11*(6), 483–502. doi:10.1002/gepi.1370110606

Finucane, H. K., Bulik-Sullivan, B., Gusev, A., Trynka, G., Reshef, Y., Loh, P.-R., ... Price, A. L. (2015). Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature Genetics*, *47*(11), 1228–1235. doi:10.1038/ng.3404

Fisher, R. A. (1941). Average excess and average effect of a gene substitution. *Annals of Eugenics*, *11*, 53–63.

Freedman, D. (1999). From association to causation: Some remarks on the history of statistics. *Statistical Science*, *14*(3), 243–258.

Gazal, S., Finucane, H. K., Furlotte, N. A., Loh, P.-R., Palamara, P. F., Liu, X., ... Price, A. L. (2017). Linkage disequilibrium-dependent architecture of human complex traits shows action of negative selection. *Nature Genetics*, *49*(10), 1421–1427. doi:10.1038/ng.3954

Goldstein, D. B. (2011). The importance of synthetic associations will only be resolved empirically. *PLoS Biology*, *9*(1), e1001008.

Kemper, K. E., Visscher, P. M., & Goddard, M. E. (2012). Genetic architecture of body size in mammals. *Genome Biology*, *13*(4), 244. doi:10.1186/gb4016

Lee, J. J. (2012). Correlation and causation in the study of personality (with discussion). *European Journal of Personality*, *26*(4), 372–412. doi:10.1002/per.1863

Lee, J. J. & Chow, C. C. (2013). The causal meaning of Fisher's average effect. *Genetics Research*, *95*(2–3), 89–109. doi:10.1017/S0016672313000074

Lee, J. J. & Chow, C. C. (2014). Conditions for the validity of SNP-based heritability estimation. *Human Genetics*, *133*(8), 1011–1022. doi:10.1007/s00439-014-1441-5

Lee, J. J., Vattikuti, S., & Chow, C. C. (2016). Uncovering the genetic architectures of quantitative traits. *Computational and Structural Biotechnology Journal*, *14*, 28–34. doi:10.1016/j.csbj.2015.10.002

Lynch, M. & Walsh, B. (1998). *Genetics and the analysis of quantitative traits*. Sunderland, MA: Sinauer.

Okbay, A., Beauchamp, J. P., Fontana, M. A., Lee, J. J., Pers, T. H., Rietveld, C. A., ... Benjamin, D. J. (2016). Genome-wide association study identifies 74 loci associated with educational attainment. *Nature*, *533*(7604), 539–542. doi:10.1038/nature17671

Palla, L. & Dudbridge, F. (2015). A fast method that uses polygenic scores to estimate the variance explained by genome-wide marker panels and the proportion of variants affecting a trait. *American Journal of Human Genetics*, *97*(2), 250–259. doi:10.1016/j.ajhg.2015.06.005

Pearl, J. (2009). *Causality: Models, reasoning, and inference* (2nd ed.). New York: Cambridge University Press.

Plomin, R., DeFries, J. C., Knopik, V. S., & Neiderhiser, J. M. (2016). Top 10 replicated findings from behavioral genetics. *Perspectives on Psychological Science*, *11*(1), 3–23. doi:10.1177/1745691615617439

Sacerdote, B. (2007). How large are the effects from changes in family environment? A study of Korean American adoptees. *Quarterly Journal of Economics*, *122*(1), 119–157. doi:10.1162/qjec.122.1.119

Shi, H., Mancuso, N., Spendlove, S., & Pasaniuc, B. (2017). Local genetic correlation gives insights into the shared genetic architecture of complex traits. *American Journal of Human Genetics*, *101*(5), 737–751. doi:10.1016/j.ajhg.2017.09.022

Speed, D., Hemani, G., Johnson, M. R., & Balding, D. J. (2012). Improved heritability estimation from genome-wide SNPs. *American Journal of Human Genetics*, *91*(6), 1011–1021. doi:10.1016/j.ajhg.2012.10.010

Tenesa, A., Rawlik, K., Navarro, P., & Canela-Xandri, O. (2016). Genetic determination of height-mediated mate choice. *Genome Biology*, *16*(1), 269. doi:10.1186/s13059-015-0833-8

Visscher, P. M., Brown, M. A., McCarthy, M. I., & Yang, J. (2012). Five years of GWAS discovery. *American Journal of Human Genetics*, *90*(1), 7–24. doi:10.1016/j.ajhg.2011.11.029

Wright, S. (1934). The method of path coefficients. *Annals of Mathematical Statistics*, *5*(3), 161–215.

Yang, J., Bakshi, A., Zhu, Z., Hemani, G., Vinkhuyzen, A. A. E., Lee, S. H., ... Visscher, P. M. (2015). Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nature Genetics*, *47*(10), 1114–1120. doi:10.1038/ng.3390