Supplementary Note to Estimating Genetic Correlation from GWAS Summary Statistics

Brendan Bulik-Sullivan

September 13, 2014

Contents

1 Definitions

Let y_1 and y_2 denote phenotypes defined for individuals in a hypothetical population of infinite size (or more precisely, for individuals drawn from a distribution). Let g denote a set of additively-coded SNPs, and let g_1 and g_2 denote the best linear predictors of y_1 and y_2 that can be constructed (at infinite sample size) from the SNPs in S^1. Then we can write

$$\begin{aligned} y_1 &= g_1 + \epsilon_1; \\ y_2 &= g_2 + \epsilon_2, \end{aligned}$$

where ϵ_i denotes the residual, which is uncorrelated (in the population) with g_i by the definition of a projection. Note that so far this construction is applicable to arbitrary phenotypes[2].

Definition 1.1. *The narrow-sense (or additive)* ***heritability*** *of phenotype y_i explained by the SNPs in g, denoted $h^2_g(y_i)$ is defined*

$$h^2_g(y_i) := \mathrm{Cor}[g_i, y_i]^2, \tag{1.1}$$

where Cor *denotes the correlation between random variables, (alternatively, the correlation in a hypothetical population of infinite size), not the empirical correlation in some finite sample.*

Definition 1.2. *The [additive]* ***genetic covariance*** *between y_1 and y_2 among SNP set g, denoted $\rho_g(y_1, y_2)$ is defined*

$$\rho_g(y_1, y_2) := \frac{\mathrm{Cov}[g_1, g_2]}{\sqrt{\mathrm{Var}[y_1]\mathrm{Var}[y_2]}}. \tag{1.2}$$

Definition 1.3. *The [additive]* ***genetic correlation*** *between y_1 and y_2 among SNP set g, denoted $r_g(y_1, y_2)$ is defined*

$$r_g(y_1, y_2) := \frac{\rho_g}{\sqrt{h^2_g(y_1)h^2_g(y_2)}}. \tag{1.3}$$

Note that these definitions make sense even when either or both phenotypes are binary, and we refer to the specialization of definition 1.1 to a binary phenotype as the ***heritability of the observed phenotype***.

There are two challenges when dealing with binary phenotypes. The first is inferential: often studies of binary phenotypes will over-sample cases in order to increase power. Some work is required in order to obtain valid estimates of the parameters of a population with, say, 1% cases from an ascertained sample with 50% cases. Ascertainment is addressed in section **??**. The second challenge is definitional: the heritability of the observed phenotype depends strongly on the prevalence of the phenotype. For example, consider two liability threshold phenotypes y_1 and y_2, determined by the same underlying liability ψ, but with different thresholds. That is, $y_i := \mathbf{1}[\psi > \tau_i]$ for $i = 1, 2$. Suppose $h^2_g(\psi) = 1$, $\tau_1 = 0$ and $\tau_2 = 1.96$ (meaning the population prevalence of y_1 is 50% and the

[1] Formally, the g_i are constructed by projecting the phenotypes onto the vector space of functions $\{0, 1, 2\}^M \to \mathbb{R}$, where $M := |g|$. As a result g_i may account for a large proportion of the variance in phenotype, even if in truth the way in which the phenotype is determined from genotype and environmental factors is completely non-additive.

[2] Well, measurable finite-variance phenotypes. But this is no restriction at all on the genetic component, and hardly any restriction at all on the environmental component.

population prevalence of y_2 is 5%), then the heritability of the observed phenotype y_1 is 0.64, and the heritability of the observed phenotype y_2 is 0.23.

Sometimes it is desirable to compare the heritabilities or genetic covariances of phenotypes with different prevalences on an even footing, and this is the primary application of liability-scale heritability and liability scale genetic covariance. We note that one need not take the liability threshold model literally[3] in order for the conversion to the liability scale to be a useful procedure. One can view this conversion simply as a tool for comparing phenotypes with different prevalences that is inspired by – but not dependent on – the liability threshold model. An interpretation of LD Score regression under the liability threshold model is provided in section **??**.

There is no need to specify a scale when discussing genetic correlation. Genetic correlation on the observed scale is the same as genetic correlation on the liability scale is the same as genetic correlation in an ascertained sample.

2 Models

2.1 Quantitative Traits

Suppose we sample two cohorts for two phenotypes, y_1 and y_2, with sample sizes N_1 and N_2, such that N_s individuals are shared between cohorts and phenotyped for both traits. We model phenotypes as generated by the equations

$$y_1 = Y\beta + \delta;$$
$$y_2 = Z\gamma + \epsilon,$$

where Y and Z are matrices of genotypes normalized to mean zero and variance one[4], with dimensions $N_1 \times M$ and $N_2 \times M$, respectively; β and γ are $M \times 1$ vectors of per-normalized genotype effect sizes, and δ and ϵ are vectors of environmental or non-additive genetic effects, with dimensions $N_1 \times 1$ and $N_2 \times 1$, respectively. In this model, Y and Z are unobserved matrices of *all* SNPs, unlike Yang, *et al* [8], we model the effects of SNPs that are not genotyped as well as those that are.

Now we introduce randomness: we model all of $Y, Z, \beta, \gamma, \delta$ and ϵ as random variables. Suppose that the $2M \times 1$ vector (β, γ) follows a distribution with mean zero and variance-covariance matrix

$$\mathrm{Var}[(\beta,\gamma)] = \frac{1}{M}\begin{pmatrix} h_1^2 I & \rho_g I \\ \rho_g I & h_2^2 I \end{pmatrix},$$

and the $2N \times 1$ vector (δ, ϵ) follows a distribution with mean zero and variance-covariance matrix

$$\mathrm{Var}[(\delta,\epsilon)] = \begin{pmatrix} (1-h_1^2)I & \rho_e I \\ \rho_e I & (1-h_2^2)I \end{pmatrix}.$$

[3] We also note that the liability threshold model is (trivially) completely general. Let y denote an arbitrary binary phenotype with prevalence K, and set $\tau := \Phi^-1(1-K)$, where Φ is the cdf of the standard normal distribution. We can construct a liability for y as follows: if individual i is a case, draw a liability ψ_i from a truncated standard normal with left truncation point τ. If individual i is a control, draw a liability ψ_i from a truncated standard normal with right truncation point τ. Then $y = \mathbf{1}[\psi > \tau]$.

[4] We ignore the distinction between normalizing and centering in the population and in the sample, since this introduces only $\mathscr{O}(1/N)$ error.

Finally suppose that each row (individual) of Y and Z represents an *i.i.d.* draw from a distribution with covariance matrix (LD matrix) R (except of course the N_s rows that are duplicated in Y and Z). We will write $\mathbb{E}[Y_{ij}Y_{ik}] = R_jk =: r_jk$. Note that since we assume normalized genotypes, R is both the covariance matrix and correlation matrix. Additionally, note that under this model, $\mathrm{Var}[y_1] = \mathrm{Var}[y_2] = 1$. Let $\rho := \rho_g + \rho_e$ denote the covariance between y_1 and y_2, which is also the correlation between y_1 and y_2, since both have variance one.

The assumption that all β is drawn with equal variance for all SNPs is, of course, not reasonable. We only make this assumption here for notational simplicity. In this paper, we use MAF- and LD-partitioned LD Score regression for estimation, as described in references [3, 1]. This technique minimizes confounding under models where $\mathrm{Var}[\beta]$ is correlated with MAF or LD Score.

Proposition 2.1. *Under this model, the expected genetic covariance between phenotypes is* ρ_g*, justifying our use of the notation* ρ_g*.*

Proof. Let X denote an $1 \times M$ vector of normalized, centered genotypes for an arbitrary individual. Under the model, the additive genetic component of y_1 for this individual is $\sum_j X_j\beta_j$, and the additive genetic component of y_2 for this individual is $\sum_j X_j\gamma_j$. Thus, the genetic covariance between y_1 and y_2 is

$$\mathrm{Cov}\left[\sum_{j=1}^{M} X_j\beta_j, \sum_{j=1}^{M} X_j\gamma_j\right],$$

We can simplify this covariance with some algebra:

$$\begin{aligned}
\mathrm{Cov}\left[\sum_{j=1}^{M} X_j\beta_j, \sum_{j=1}^{M} X_j\gamma_j\right] &= \mathbb{E}\left[\left(\sum_{j=1}^{M} X_j\beta_j\right)\left(\sum_{j=1}^{M} X_j\gamma_j\right)\right] \\
&= \mathbb{E}\left[\sum_{j=1}^{M}\sum_{k=1}^{M} X_jX_j\beta_j\gamma_k\right] \\
&= \mathbb{E}\left[\sum_{j=1}^{M} X_j^2\beta_j\gamma_j\right] \\
&= \sum_{j=1}^{M} \mathbb{E}[X_j^2]\mathbb{E}[\beta_j\gamma_j] \\
&= \rho_g.
\end{aligned}$$

□

2.2 Liability Threshold Model

Suppose unobserved liability is generated following the usual model for quantitative traits:

$$\psi_i = \sum_{j=1}^{M} X_{ij}\beta_j + \epsilon_i.$$

The observed binary phenotype has population prevalence K, and is generated from the unobserved liability via the liability threshold model:

$$y_i := \mathbf{1}[\psi_i > \tau],$$

where $\tau := \Phi^{-1}(1-K)$ is the liability threshold, and Φ denotes the cdf of the standard normal distribution.

3 Derivations

3.1 Non-Ascertained Samples

3.1.1 Genetic Covariance

Suppose we directly genotype SNP j. We estimate effect sizes using least-squares:

$$\begin{aligned}\hat{\beta}_j &:= \frac{1}{N_1} Y_j^\intercal y_1;\\ \hat{\gamma}_j &:= \frac{1}{N_2} Z_j^\intercal y_2,\end{aligned}$$

where Y_j and Z_j denote the genotypes of all individuals at SNP j and have dimensions $N_1 \times 1$ and $N_2 \times 1$, respectively.

Proposition 3.1. *Under the model described in 2.1,*

$$\mathbb{E}[\hat{\beta}_j \hat{\gamma}_j] = \frac{\rho_g}{M}\ell_j + \frac{N_s \rho}{N_1 N_2}. \tag{3.1}$$

Proof. By the law of total expectation,

$$\mathbb{E}[\hat{\beta}_j \hat{\gamma}_j] = \mathbb{E}[\mathbb{E}[\hat{\beta}_j \hat{\gamma}_j \,|\, Y, Z]]$$

First,

$$\begin{aligned}\mathbb{E}[\hat{\beta}_j \hat{\gamma}_j \,|\, Y, Z] &= \frac{1}{N_1 N_2}\mathbb{E}[Y_j^\intercal y_1 y_2^\intercal Z_j]\\ &= \frac{1}{N_1 N_2} Y_j^\intercal \mathbb{E}[(Y\beta + \delta)(Z\gamma + \epsilon)] Z_j\\ &= \frac{1}{N_1 N_2} Y_j^\intercal \left(Y\mathbb{E}[\beta\gamma^\intercal]Z + \mathbb{E}[\delta^\intercal \epsilon]\right) Z_j\\ &= \frac{1}{N_1 N_2}\left(\frac{\rho_g}{M} Y_j^\intercal Y Z_j^\intercal Z + \rho_e Y_j^\intercal Z_j\right).\end{aligned}$$

To remove the conditioning on Y and Z, we need only compute

$$\frac{1}{N_1 N_2}\mathbb{E}[Y_j^\intercal Z_j] = \frac{N_s}{N_1 N_2},$$

and

$$\frac{1}{N_1 N_2}\mathbb{E}[Y_j^\intercal Y Z_j^\intercal Z] = \ell_j + \frac{M N_s}{N_1 N_2}.$$

Thus,

$$\mathbb{E}[\hat{\beta}_j\hat{\gamma}_j] = \frac{\rho_g}{M}\ell_j + \frac{N_s\rho}{N_1N_2}.$$

Note that this expression does not contain any terms in which both the quantities N_s and ℓ_j appear. In the special case where there are no overlapping samples shared between the two cohorts, this expression simplifies to

$$\mathbb{E}[\hat{\beta}\hat{\gamma}] = \frac{\rho_g}{M}\ell_j.$$

□

3.1.2 Regression Weights

We can improve the efficiency of the LD Score regression by computing the conditional variance $\text{Var}[\hat{\beta}\hat{\gamma} \mid \ell_j]$ and weighting the regression by the reciprocal of this variance. In order to compute this conditional variance, we need further assumptions: in addition to the assumptions from 2.1, assume that the phenotypes follow a multivariate normal distribution[5].

If phenotypes are normally distributed, then by the central limit theorem, $\hat{\beta}$ and $\hat{\gamma}$ are jointly normally distributed with expectation zero. Thus,

$$\begin{aligned}\text{Var}[\hat{\beta}_j\hat{\gamma}_j \mid Y, Z] &= \mathbb{E}[\hat{\beta}^2\hat{\gamma}^2] \\ &= \text{Var}[\hat{\beta}]\text{Var}[\hat{\gamma}] + 2\mathbb{E}[\hat{\beta}\hat{\gamma}]^2 \\ &= \left(\frac{h_1^2\ell_j}{M} + \frac{1-h_1^2}{N_1}\right)\left(\frac{h_2^2\ell_j}{M} + \frac{1-h_2^2}{N_2}\right) + 2\left(\frac{\rho_g\ell_j}{M} + \frac{\rho N_s}{N_1N_2}\right). \end{aligned} \tag{3.2}$$

Note that we only assume normality in order to compute regression weights. If (quantitative) phenotypes are not normally distributed, this will not affect the expectation of the LD Score regression estimates (see 3.1.1, which makes no distributional assumptions about β and γ beyond first and second moments), but will increase the standard error, because in this case the regression weights will not be optimal for correcting for heteroskedasticity. We never assume homoskedasticity when computing standard errors or *p*-values (we use a block jackknife, which is robust to heteroskedasticity), so non-normality of the phenotypes also does not bias our inference. We derive regression weights for ascertained studies of binary phenotypes in section **??**. Note that if the phenotypes are normally distributed, then $\hat{\beta}_j\hat{\gamma}_j$ follows a product-normal distribution, which is not in the exponential family, so this is not a GLM.

3.2 Ascertainment

In this section, we derive the LD Score regression estimators of heritability and genetic covariance for ascertained case/control samples (which was addressed only via simulation in [2]). The fact that this estimator works is *not* a consequence of the equivalence between LD Score regression and HE regression (see section 4), and the fact that HE regression works in ascertained case/control

[5] For instance, it is sufficient but not necessary to assume that β, γ, δ and ϵ are multivariate normal, and that N_1 and N_2 are sufficiently large that we can invoke the central limit theorem. More generally, the phenotypes will be approximately normal if δ and ϵ are normal and if β and γ are reasonably polygenic. If there are few casual SNPs, then the conditional variance may be larger.

samples [4], because case/control ascertainment induces LD between causal SNPs in the ascertained samples. HE regression accounts for this LD by using a GRM computed from sample genotypes. It is not clear *a priori* that the LD Score regression approach of using population LD as an estimate of sample LD is valid when the sample is ascertained. However, this turns out to be fine, though we do not address this issue directly. Our proof strategy is first to note that GWAS summary statistics can be written in terms of the sample allele frequencies in cases and the sample allele frequencies in controls. Since the sample allele frequency in cases is a consistent estimator of the population allele frequency in cases, and likewise for the sample allele frequency in controls, we can write the large-N limit of our GWAS summary statistics in terms of the population allele frequencies (see section 3.2.1). Since the population allele frequencies depend on population LD rather than ascertained sample LD, LD Score regression with population LD yields a consistent estimators of heritability and genetic covariance.

This section is structured as follows: in 3.2.2 and 3.2.3, we show that using estimates of population LD is a valid, even with ascertained samples, and we derive the usual factors for converting heritabilities and genetic covariances from ascertained samples into estimates of the population heritabilities and genetic covariances.

In 3.2.4, we deal with the special case of the liability threshold model, and derive LD Score regression estimators of the heritability of liability, and genetic covariance between liability and other phenotypes.

3.2.1 Case/Control Test Statistics

Consider a study of size N where the sample prevalence of phenotype y is P. Let $N_{eff} := NP(1-P) = N_0N_1/N$ denote the effective sample size. We compute Z-statistics

$$Z_j := \frac{\sqrt{N_{eff}}(\hat{p}_1 - \hat{p}_0)}{\sqrt{\hat{p}_j(1-\hat{p}_j)}}, \tag{3.3}$$

and χ^2-statistics[6]

$$\chi^2_j := Z^2_j, \tag{3.4}$$

where $\hat{p}_j$ denotes allele frequency in the entire sample[7], $\hat{p}_1$ denotes allele frequency among cases in the sample and $\hat{p}_0$ denotes allele frequency among controls in the sample. We aim to derive an estimator of heritability from $\mathbb{E}[\chi^2_j \,|\, \ell_j]$ and an estimator of genetic covariance from $\mathbb{E}[Z_j \,|\, \ell_j]$ in samples where $P \neq K$ under various models of genetic architecture. First, we need a lemma, which allows us to write our χ^2-statistics in terms of population allele frequencies in the large-N limit.

Lemma 3.1. *In an ascertained study with sample size N, sample prevalence P and population prevalence K, the expected Z-statistic of a SNP j conditional on its population allele frequencies in cases and controls is*

$$\mathbb{E}[Z_j \,|\, p_0, p_1] = \tag{3.5}$$

and the expected χ^2 statistic is

$$\mathbb{E}[\chi^2_j \,|\, p_0, p_1] = +1. \tag{3.6}$$

[6] This is the equal to N times the squared correlation between phenotype and genotype, *i.e.,* the Armitage Trend Test (ATT).

[7] Note that if j has nonzero effect size, the expected value of $\hat{p}_j$ is not equal to p_j unless $P = K$.

Proof. Note that the case where we condition on p_0 and p_1 (*i.e.,* when the only randomness is from sampling and genetic architecture is nonrandom), is the usual case considered in power analyses for GWAS, so we can even obtain the asymptotic distributions of the Z and χ^2 statistics from standard results on Wald statistics. First,

$$Z_j \,|\, p_0, p_1 \sim N()$$

so

$$\mathbb{E}[Z_j \,|\, p_0, p_1] =$$

and χ_j^2 follows a noncentral χ^2 distribution with one degree of freedom and non-centrality parameter

$$\text{NCP} =$$

Since the expected value of a noncentral χ^2 distribution with k degrees of freedom and noncentrality parameter λ is $k + \lambda$,

$$\mathbb{E}[\chi_j^2 \,|\, p_0, p_1] = +1$$

□

3.2.2 Heritability of the Observed Phenotype

Next, we remove the conditioning on p_0 and p_1 by noting that p_0 and p_1 are fixed conditional on β_{loc}, and can be approximated using 3.10 and 3.11. Thus, the inner term of the numerator from **??** is

$$\begin{aligned}\mathbb{E}[(\hat{p}_1 - \hat{p}_0)^2 \,|\, \beta_{loc}] &\approx \frac{p_j^2 \phi(\tau)^2 \alpha_j^2}{K^2(1-K)^2} + \frac{p_j(1-p_j) + \dfrac{p_j\phi(\tau)\alpha_j}{K}\left(1 - 2p_j - \dfrac{p_j\phi(\tau)\alpha_j}{K}\right)}{N_1} \\ &\quad + \frac{p_j(1-p_j) + \dfrac{p_j\phi(\tau)\alpha_j}{K}\left(2p_j - 1 - \dfrac{p_j\phi(\tau)\alpha_j}{K}\right)}{N_0}. \\ &\approx \frac{p_j^2 \phi(\tau)^2 \alpha_j^2}{K^2(1-K)^2} + \mathscr{O}(1/N). \end{aligned} \tag{3.7}$$

Next, we remove the conditioning on p_0 and p_1. Let $C := \phi(\tau)P(1-K)+K(1-P))/(K(1-K))$. Using the approximations from 3.10 and 3.11,

$$\tilde{p}_j \,|\, \beta_{loc} \approx p_j \left(1 + C\alpha_j\right),$$

and

$$\begin{aligned}\mathbb{E}[\tilde{p}_j(1-\tilde{p}_j) \,|\, \beta_{loc}] &= p_j(1 + C\alpha_j)(1 - p_j - p_jC\alpha_j) \\ &= p_j\left(1 - p_j + C\alpha_j(1 - 2p_j - p_jC\alpha_j).\right).\end{aligned}$$

Note that we have already computed $p_0(1-p_0) \,|\, \beta_{loc}$ and $p_1(1-p_1) \,|\, \beta_{loc}$ in 3.7. Thus, the inner term of the denominator of **??** is

$$\begin{aligned}\mathbb{E}[\hat{p}_j(1-\hat{p}_j)\,|\,\beta_{loc}] &= p_j\left(1-p_j+C\alpha_j(1-2p_j-p_jC\alpha_j)\right)\\ &\quad - \frac{(1-P)^2p_j(1-p_j)+\dfrac{p_j\phi(\tau)\alpha_j}{K}\left(2p_j-1-\dfrac{p_j\phi(\tau)\alpha_j}{K}\right)}{N_0}\\ &\quad - \frac{P^2p_j(1-p_j)+\dfrac{p_j\phi(\tau)\alpha_j}{K}\left(1-2p_j-\dfrac{p_j\phi(\tau)\alpha_j}{K}\right)}{N_1}\\ &\approx p_j\left(1-p_j+C\alpha_j(1-2p_j-p_jC\alpha_j)\right)+\mathscr{O}(1/N). \end{aligned} \tag{3.8}$$

We now introduce randomness into β. For this section, model the entries of β as *i.i.d.* draws from a distribution with expectation zero and variance h^2/M. The heritability of liability (in the population) under this model is h^2; $\mathbb{E}[\alpha_j]=0$; $\mathbb{E}[Z_j]=0$, and

$$\mathbb{E}[\alpha_j^2]=\frac{(1-p_j)h^2}{p_jM}\ell_j.$$

The expected numerator of the χ^2-statistic is therefore

$$\begin{aligned}N_{eff}\mathbb{E}[(\hat{p}_1-\hat{p}_0)^2] &= p_j(1-p_j)\left(\frac{N_{eff}\phi(\tau)^2h^2\ell_j}{K^2(1-K)^2M}\left(1+\frac{N(1-K)^2}{N_0N_1K^2}\right)+\frac{NN_{eff}}{N_0N_1}\right)\\ &\approx p_j(1-p_j)\left(\frac{N_{eff}\phi(\tau)^2h^2\ell_j}{MK^2(1-K)^2}+1\right),\end{aligned}$$

where the approximation is justified by the fact that $N/N_0N_1 \ll 1$, which is a reasonable approximation when P is not so far from 1/2 (a balanced study). For brevity, let c denote

$$c := \frac{P(1-P)\phi(\tau)^2}{K^2(1-K)^2},$$

which is the factor used to convert between liability scale heritability, h^2 and observed scale heritability, $h^2_{obs} := ch^2$.

The expected denominator of the χ^2-statistic is

$$\begin{aligned}\mathbb{E}[\hat{p}_j(1-\hat{p}_j)] &= p_j(1-p_j)-\mathscr{O}(\ell_j/M)+\mathscr{O}(1/N)\\ &\approx p_j(1-p_j).\end{aligned}$$

Thus,

$$\begin{aligned}\mathbb{E}[\chi^2_j] &\approx c\frac{Nh^2}{M}\ell_j+1\\ &= \frac{Nh^2_{obs}}{M}\ell_j+1.\end{aligned} \tag{3.9}$$

3.2.3 Genetic Covariance with the Observed Phenotype

In this section, we derive a genetic covariance estimator that works when both studies are ascertained to oversample cases and may include overlapping samples. This derivation does not cover more complicated ascertainment schemes (*e.g.,* attempting to estimate genetic covariance between T2D and BMI from a T2D GWAS consisting of low-BMI cases and high-BMI controls).

3.2.4 Heritability and Genetic Covariance of Liability

For phenotypes generated according to the liability threshold model, we can estimate not only the heritability of the observed phenotype (genetic covariance between the observed phenotype and other phenotypes), but also the heritability of the unobserved liability (genetic covariance between unobserved liability and other phenotypes).

The derivation is the same as in 3.2.2 and 3.2.3, except with one extra step: we need to write $\mathbb{P}[y_i = 1 \,|\, G_{ij} = 1]$ in terms of the heritability of liability.

Let $\alpha_j := \mathbb{E}[\psi \,|\, G_{ij} = 1] = p_j^{-\frac{1}{2}}(1-p_j)^{\frac{1}{2}} \sum_{\{k \,|\, r_{jk} \neq 0\}} r_{jk}\beta_k$ denote the marginal per-normalized genotype effect size of SNP j on liability. This is the per-normalized-genotype effect size that one would obtain from regressing liability against the genotype at SNP j at infinite sample size.

Then if $\phi(x, \mu, \sigma^2)$ denotes the density of a normal distribution with expectation μ and variance σ^2 evaluated at x,

$$\begin{aligned}
\mathbb{P}[y_i = 1 \,|\, G_{ij} = 1] &= \int_{\tau}^{\infty} \phi(x, \alpha_j, 1-\alpha_j^2)dx \\
&= \int_{\tau-\alpha_j}^{\infty} \phi(x(1-\alpha_j^2)^{\frac{1}{2}}; 0, 1)dx \\
&= (1-\alpha_j^2)^{-\frac{1}{2}} \int_{(\tau-\alpha_j)(1-\alpha_j^2)^{\frac{1}{2}}}^{\infty} \phi(u; 0, 1)du \\
&= (1-\alpha_j^2)^{-\frac{1}{2}}(1-\Phi(\tau')),
\end{aligned}$$

where $\tau' := (\tau - \alpha_j)(1-\alpha_j^2)^{\frac{1}{2}}$. Similarly,

$$\begin{aligned}
\mathbb{P}[y_i = 0 \,|\, G_{ij} = 1] &= 1 - \mathbb{P}[y_i = 1 \,|\, G_{ij} = 1] \\
&= 1 - (1-\alpha_j^2)^{-\frac{1}{2}}(1-\Phi(\tau')).
\end{aligned}$$

We approximate $\Phi(\tau')$ with a first-order Taylor expansion[8] around τ:

$$\begin{aligned}
\Phi(\tau') &\approx \Phi(\tau) + \phi(\tau)(\tau' - \tau) \\
&= 1 - K + \phi(\tau)(\tau' - \tau).
\end{aligned}$$

We also make the approximation that

$$(1-\alpha_j^2)^{\frac{1}{2}} \approx 1.$$

Thus we have

$$\begin{aligned}
p_1 \,|\, \beta_{loc} &= \frac{p_j}{K}(1-\alpha_j^2)^{-\frac{1}{2}}(1-\Phi(\tau')) \\
&\approx p_j\left(1 + \frac{\phi(\tau)\alpha_j}{K}\right),
\end{aligned} \tag{3.10}$$

and

$$\begin{aligned}
p_0 \,|\, \beta_{loc} &= \frac{p_j}{1-K}\left(1 - (1-\alpha_j^2)^{-\frac{1}{2}}(1-\Phi(\tau'))\right) \\
&\approx p_j\left(1 - \frac{\phi(\tau)\alpha_j}{1-K}\right).
\end{aligned} \tag{3.11}$$

[8] This is a reasonable approximation for small α_j, *e.g.,* for polygenic traits and away from loci with huge effects.

We can plug these results into the expressions from 3.2.2 in order to obtain an estimator of the heritability of liability:

WRITE MEEEE

3.2.5 Regression Weights

Lorem Ipsum Dolor Sic Amet

4 LD Score Regression is Haseman-Elston Regression

In this section we re-interpret LD Score regression as an approximation to Haseman-Elston (HE) regression [5] that is valid for samples of unrelated individuals. In section 4.1, we prove equivalence between a modified LD Score estimator of heritability and HE regression, and and in section 4.2 we prove equivalence between and modified LD Score estimator of genetic covariance and HE regression. In section 4.3, we describe the interpretation of standard LD Score regression as an approximation to HE regression, and some of the advantages of this approximation, in particular, the ability to model all SNPs, not just genotyped SNPs, and robustness to population stratification.

4.1 Heritability

Suppose we condition on the matrix of sample genotypes X (which is a matrix encoding normalized and centered genotypes at *all* SNPs), but otherwise follow the univariate version of the model in 2.1 (*i.e.,* the model in [2]). Then the variance-covariance matrix of the $N \times 1$ vector y of phenotypes is

$$\mathrm{Var}[y] = h_g^2 A + (1 - h_g^2) I,$$

where $A := X^\mathsf{T} X/M$ is the sample GRM. This means that $\mathbb{E}[y_h y_i] = h_g^2 A_{hi}/M$, so we can estimate genetic covariance by regressing $y_h y_i$ against A_{hi} and multiplying by the slope by M. Precisely,

$$\hat{h}^2_{g,HE} := \frac{M\widehat{\mathrm{Cov}}[A_{hi}, y_h y_i]}{\widehat{\mathrm{Var}}[A_{hi}]}, \tag{4.1}$$

where $\widehat{\mathrm{Var}}$ and $\widehat{\mathrm{Cov}}$ denote sample variance and sample covariance, respectively. This estimator is called Haseman-Elston (HE) regression [5].

There is another LD Score-based estimator of heritability, the aggregate estimator:

$$\hat{h}^2_{g,agg} := \frac{1}{\bar{\ell}_{sample}} \sum_{j=1}^{M} \chi_j^2 \tag{4.2}$$

$$\approx \frac{1}{\bar{\ell}} \sum_{j=1}^{M} (\chi_j^2 - 1) \tag{4.3}$$

where $\bar{\ell} := \mathrm{Mean}_j(\ell_j)$. We should call attention to a notational choice: we can choose to deal with the $\mathscr{O}(1/N)$ upward bias in sample r^2 by subtracting one from χ^2 and using $\bar{\ell}$, the mean population LD Score, or by not subtracting one from χ^2 and using $\bar{\ell}_{sample}$, the mean LD Score in our (finite) sample. Recall from the supplementary note of [2] that we have the relationship

$$\bar{\ell}_{sample} \approx \bar{\ell} + \frac{M}{N} \tag{4.4}$$

in an unstructured sample. Since for this section we are assuming that we have access to raw genotypes, we deal with $\bar{\ell}_{sample}$. We discuss the implications of approximating sample LD Score with an estimate of population LD Score from a sequenced reference panel in 4.3.

Proposition 4.1. *The aggregate estimator is an unbiased estimate of heritability, and is equivalent to the LD Score regression estimator with the intercept constrained to zero*[9] *(if using sample LD Score) or constrained to one (if using population LD Score).*

Proof. This follows immediately from the supplementary note of [2]. □

Note that computing these estimators requires raw genotypes at all SNPs, which is generally only possible with sequence data. Often geneticists substitute a GRM computed using only genotyped SNPs. The distinction is unimportant for this derivation; for now, fix a set of observed SNPs $j = 1, \dots, M_g$. We discuss the distinction between using all SNPs and only genotyped SNPs in section 4.3.

Proposition 4.2. *Let $\hat{r}_{jk}$ denote the sample correlation between SNPs j and k in matrix X Let*

$$\hat{\ell}_j := \sum_{j=1}^{M_g} \hat{r}_{jk}^2$$

denote sample LD Score in YX with the sum taken only over the M_g observed SNPs. Let $\bar{\ell}_X$ denote the mean value of $\hat{\ell}_j$ among genotyped SNPs $j = 1, \dots, M_g$. If we use this version of $\bar{\ell}$ for the denominator of the aggregate estimator, then the aggregate estimator is equal to the HE estimator.

Proof. The numerator of the HE estimator is

$$\begin{aligned} M\widehat{\mathrm{Cov}}[A_{hi}, y_h y_i] &= \frac{M}{N^2} \sum_{h=1}^{N} \sum_{i=1}^{N} A_{hi} y_h y_i \\ &= \frac{1}{N^2} \sum_{h=1}^{N} \sum_{i=1}^{N} \sum_{j=1}^{M} X_{hi} X_{ij} y_h y_i \\ &= \frac{1}{N^2} \sum_{j=1}^{M} \left(\sum_{h=1}^{N} \sum_{i=1}^{N} X_{hj} y_1 X_{ij} y_2 \right) \\ &= \sum_{j=1}^{M} \chi_j^2. \end{aligned}$$

[9] Note that a consequence of constraining the intercept is that the aggregate estimator, unlike the LD Score regression estimator, is not immune to population stratification or cryptic relatedness.

The denominator of the HE estimator is

$$\begin{aligned}
\widehat{\mathrm{Var}}[A_{hi}] &= \frac{1}{N^2}\sum_{h=1}^{N}\sum_{i=1}^{N} A_{hi}^2 \\
&= \frac{1}{MN^2}\sum_{h=1}^{N}\sum_{i=1}^{N}\left(\sum_{j=1}^{M} X_{hj}X_{ij}\right)^2 \\
&= \frac{1}{MN^2}\sum_{h=1}^{N}\sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{k=1}^{M} X_{hj}X_{hk}X_{ij}X_{ik} \\
&= \frac{1}{M}\sum_{j=1}^{M}\left(\sum_{k=1}^{M}\hat{r}_{jk}^2\right), \\
&= \frac{1}{M}\sum_{j=1}^{M}\hat{\ell}_j \\
&= \bar{\ell}_X.
\end{aligned}$$

Since the numerator of the HE estimator is equal to the numerator of the aggregate estimator and the denominator of the HE estimator is equal to the denominator of the aggregate estimator, the conclusion follows.

□

4.2 Genetic Covariance

The proof of the equivalence between the LD Score aggregate estimator of genetic covariance and the HE regression estimator of genetic covariance is a direct bivariate analogue of the proof of the equivalence of the LD Score aggregate heritability estimator in the HE regression heritability estimator from 4.1; however, we present the proof in full detail for reference. For notational simplicity, we assume no sample overlap in this section.

Suppose we condition on Y and Z (which are matrices encoding genotypes at *all* SNPs), but otherwise follow the model in 2.1. Then the variance-covariance matrix of the $(N_1 + N_2) \times 1$ vector (y_1, y_2) of phenotypes is

$$\mathrm{Var}[(y_1, y_2)] = \frac{1}{M}\begin{pmatrix} h_1^2 YY^{\mathsf{T}} & \rho_g YZ^{\mathsf{T}} \\ \rho_g ZY^{\mathsf{T}} & h_2^2 ZZ^{\mathsf{T}} \end{pmatrix} + \begin{pmatrix} (1-h_1^2)I & 0 \\ 0 & (1-h_2^2)I \end{pmatrix}.$$

Let $A := Y^{\mathsf{T}}Z/M$. This means that $\mathbb{E}[y_{1h}y_{2i}] = \rho_g A_{hi}$, so we can estimate genetic covariance by regressing $y_{1h}y_{2i}$ against A_{hi} and multiplying the slope by M. This is a bivariate analogue of HE) regression first introduced (to the best of our knowledge) in [6]. The estimator is

$$\hat{\rho}_{g,HE} := \frac{M\widehat{\mathrm{Cov}}[A_{hi}, y_{1h}y_{2i}]}{\widehat{\mathrm{Var}}[A_{hi}]}, \tag{4.5}$$

where $\widehat{\mathrm{Var}}$ and $\widehat{\mathrm{Cov}}$ denote sample variance and sample covariance, respectively.

There is another LD Score-based estimator of genetic correlation, the aggregate estimator:

$$\hat{\rho}_{g,agg} := \frac{1}{\bar{\ell}} \sum_{j=1}^{M} \hat{\beta}_j \hat{\gamma}_j, \tag{4.6}$$

where $\bar{\ell} := \text{Mean}_j(\ell_j)$.

Proposition 4.3. *The aggregate estimator is an unbiased estimate of genetic covariance, and is equivalent to the LD Score regression estimator with the intercept constrained to zero*[10]*.*

Proof. For unbiasedness,

$$\begin{aligned} \mathbb{E}[\hat{\rho}_{g,agg}] &= \frac{1}{\bar{\ell}} \sum_{j=1}^{M} \mathbb{E}[\hat{\beta}_j \hat{\gamma}_j] \\ &= \frac{1}{\bar{\ell}} \sum_{j=1}^{M} \frac{\rho_g}{M} \ell_j \\ &= \rho_g. \end{aligned}$$

If we constrain the LD Score regression intercept to zero, then the LD Score regression slope is equal to the slope of the line connecting the origin to the point $(\bar{\ell}, \text{Mean}_j[\hat{\beta}_j, \hat{\gamma}_j])$, which is $\text{Mean}_j[\hat{\beta}_j, \hat{\gamma}_j])/\bar{\ell}$. The LD Score regression estimate of genetic covariance is M times this slope, which is equal to the aggregate estimate. □

Note again that computing these estimators requires raw genotypes at all SNPs, which is generally only possible with sequence data. Often geneticists substitute a GRM computed using only genotyped SNPs. The distinction is unimportant for this derivation; for now, fix a set of observed SNPs $j = 1, \ldots, M_g$. We discuss the distinction between using all SNPs and only genotyped SNPs in section 4.3.

Proposition 4.4. *Let $\hat{r}_{Y,jk}$ denote the sample correlation between SNPs j and k in matrix Y, and likewise for $\hat{r}_{Z,jk}$. Let*

$$\hat{\ell}_{YZ,j} := \sum_{j=1}^{M_g} \hat{r}_{Y,jk} \hat{r}_{Z,jk}$$

denote sample LD Score in Y and Z with r_{jk}^2 estimated as $\hat{r}_{Y,jk}\hat{r}_{Z,jk}$, (which is an unbiased estimate in the absence of sample overlap) and the sum taken only over the M_g genotyped SNPs. Let $\bar{\ell}_{YZ}$ denote the mean value of $\hat{\ell}_{YZ,j}$ among genotyped SNPs $j = 1, \ldots, M_g$. If we use this version of $\bar{\ell}$ for the denominator of the aggregate estimator, then the aggregate estimator is equal to the HE estimator.

[10] Note that a consequence of constraining the intercept to zero is that the aggregate estimator, unlike the LD Score regression estimator, is not immune to sample overlap.

Proof. The numerator of the HE estimator is

$$\begin{aligned}
M\widehat{\mathrm{Cov}}[A_{hi}, y_{ih}y_{2i}] &= \frac{M}{N_1N_2}\sum_{h=1}^{N_1}\sum_{i=1}^{N_2} A_{hi}y_{1h}y_2 \\
&= \frac{1}{N_1N_2}\sum_{h=1}^{N_1}\sum_{i=1}^{N_2}\sum_{j=1}^{M} Y_{hj}Z_{ij}y_{1h}y_{2i} \\
&= \frac{1}{N_1N_2}\sum_{j=1}^{M}\left(\sum_{h=1}^{N_1}\sum_{i=1}^{N_2} Y_{hj}y_{1h}Z_{ij}y_{2i}\right) \\
&= \sum_{j=1}^{M}\hat{\beta}_j\hat{\gamma}_j.
\end{aligned}$$

The denominator of the HE estimator is

$$\begin{aligned}
\widehat{\mathrm{Var}}[A_{hi}] &= \frac{1}{N_1N_2}\sum_{h=1}^{N_1}\sum_{i=1}^{N_2} A_{hi}^2 \\
&= \frac{1}{MN_1N_2}\sum_{h=1}^{N_1}\sum_{i=1}^{N_2}\left(\sum_{j=1}^{M} Y_{hj}Z_{ij}\right)^2 \\
&= \frac{1}{MN_1N_2}\sum_{h=1}^{N_1}\sum_{i=1}^{N_2}\sum_{j=1}^{M}\sum_{k=1}^{M} Y_{hj}Z_{hk}Y_{ij}Z_{ik} \\
&= \frac{1}{M}\sum_{j=1}^{M}\left(\sum_{k=1}^{M}\hat{r}_{Y,jk}\hat{r}_{Z,jk}\right), \\
&= \frac{1}{M}\sum_{j=1}^{M}\hat{\ell}_{YZ,j} \\
&= \bar{\ell}_{YZ}.
\end{aligned}$$

Since the numerator of the HE estimator is equal to the numerator of the aggregate estimator and the denominator of the HE estimator is equal to the denominator of the aggregate estimator, the conclusion follows.

□

Corollary 4.1. *The HE regression estimator of genetic correlation is equivalent to the LD Score regression estimate obtained from regressing*[11] $\hat{\beta}_j\hat{\gamma}_j$ *against* $\hat{\ell}_{YZ,j}$ *with the intercept constrained to zero.*

Proof. This follows from propositions 4.3 and 4.2. □

[11] Since this is not the optimal weighting for LD Score regression, it follows that the aggregate estimator should be more efficient than HE regression, so long as the LD Scores are estimated with reasonable accuracy.

4.3 Local LD

Finally, we can interpret ordinary LD Score regression with LD Scores estimated from a sequenced reference panel as a modification of HE regression with a GRM computed from all SNPs. We state this interpretation in terms of the heritability estimators; the corresponding statements for the genetic covariance estimators are directly analogous. The approximation consists of observing that although the equality $\widehat{\mathrm{Var}}[A_{hi}] = \bar{\ell}_X$ holds exactly only when $\bar{\ell}_X$ is computed using a whole-genome window (*i.e.,* including LD across chromosomes), all real LD is local, and all non-local LD results from finite sample bias in $\hat{r}^2$ or sample structure, so we can write

$$\hat{\ell}_j \approx \ell_j + \frac{M}{N} + a, \tag{4.7}$$

where $\hat{\ell}_j$ denotes the sample LD Score for SNP j, and a is a term that depends on sample structure [2]. This approximation justifies replacing $\hat{\ell}_j$ and $\bar{\ell}_X$ with an unbiased estimate of population LD Score (ℓ_j) from a sequenced reference panel. This approximation step provides several advantages:

1. The computational complexity of LD Score regression is much lower than HE regression. HE regression requires time $\mathscr{O}(M_g N^2)$ for computing the GRM and $\mathscr{O}(N^2)$ for the regression. LD Score regression requires time $\mathscr{O}(MN_{ref})$ for computing LD Scores (times a constant factor $B \approx 5000$, the mean window size measured in number of SNPs), where N_{ref} is the sample size of the sequenced reference panel, $\mathscr{O}(MN)$ for computing summary statistics[12], and time $O(M_g)$ for the regression. Thus, HE regression is quadratic in N, and LD Score regression is linear[13].

2. Using a sequenced reference panel for estimating LD Scores allows us to model the effects of all SNPs in the reference panel, not just those SNPs that are directly genotyped or well-imputed in our GWAS sample.

3. The LD Score regression intercept, which results from the approximation 4.7 is a term that depends on sample structure, confers robustness to population stratification and cryptic relatedness (or sample overlap, in the case of the genetic covariance estimator). In contrast, HE regression does not know the difference between "real" LD and spurious LD induced by sample structure (though of course if one has the genotypes, it is easy to fix this problem by including principal components as covariates [7]).

[12] It almost doesn't seem fair to count the time it takes to compute summary statistics towards the time complexity of LD Score regression, since computing summary statistics is the first thing that every GWAS consortium does.

[13] If instead of local LD, we had used genome-wide LD, the complexity of estimating LD Scores would have been $\mathscr{O}(M^2N)$, which is intimidating.

5 Supplementary Figures

6 Supplementary Tables

6.1 Simulations with Sample Overlap

Parameter	True Value	Estimate
h^2_1	0.40	0.41 (0.12)
h^2_2	0.60	0.60 (0.12)
ρ_g	0.34	0.33 (0.10)
r_g	0.70	0.66 (0.15)

We simulated two GWAS with quantitative phenotypes, using genotypes from the 4,292 individuals in the WTCCC1 bipolar disorder cohort for the first GWAS and genotypes from the 4,482 individuals in the WTCCC1 coronary artery disease cohort for the second GWAS. These cohorts contain 2,713 overlapping individuals. Additive genetic effect sizes were drawn from a bivariate point-normal distribution with 10% of SNPs causal. The environmental correlation between the two phenotypes was 0.3. The true values of each parameter are displayed in the column labeled true value. Means and standard deviations of the LD Score regression estimates of these parameters across 100 simulation replicates are displayed in the column labeled estimate. These simulations confirm that LD Score regression estimates of genetic covariance are not biased by sample overlap, as proved in 3.1.1.

6.2 Simulations with One Quantitative Trait and One Binary Trait

Prevalence	QT $\hat{h}^2$	CC $\hat{h}^2_{liab}$	$\hat{r}_g$
0.01	0.72 (0.1)	0.59 (0.04)	0.51 (0.4)
0.05	0.72 (0.12)	0.59 (0.07)	0.45 (0.17)
0.2	0.72 (0.11)	0.6 (0.08)	0.46 (0.14)
0.5	0.73 (0.11)	0.59 (0.08)	0.42 (0.17)

This table displays results from simulations with one quantitative trait and one binary trait. In order to simulate case/control ascertainment, we used simulated genotypes with an LD block LD structure ($r^2 = 0$ or $r^2 = 1$). In all simulations, the sample size for the QT GWAS was 1000, the binary trait GWAS had 1000 cases and 1000 controls, and 500 individuals appeared in both studies.

From left to right, the columns are prevalence of the binary trait, estimated heritability of the quantitative trait, estimated heritability of the binary trait (on the liability scale) and estimated genetic correlation. Estimates are averages taken across 100 simulations per prevalence. Standard deviations (in parentheses) are calculated as the empirical standard deviation across 100 simulations.

These simulations confirm that LD Score regression gives consistent estimates of genetic correlation between an ascertained binary phenotype and a quantitative phenotype, even for binary phenotypes with low prevalence (1%) and even with sample overlap.

6.3 Simulations with Parallel LD- and MAF-Dependence

LD Score	h^2(5-50%)	ρ_g(5-50%)	r_g(5-50%)
Truth	0.83	0.42	0.5
Naive	0.36 (0.08)	0.18 (0.06)	0.5 (0.13)
30 Bins	0.81 (0.12)	0.41 (0.08)	0.51 (0.09)
60 Bins	0.81 (0.12)	0.41 (0.09)	0.51 (0.09)

This table displays simulations with MAF- and LD-dependent genetic architecture where the MAF- and LD- dependence was the same for both phenotypes and genetic correlation did not vary with MAF or LD. Precisely, effect sizes were drawn from a normal distribution so that the magnitude of per-allele effect sizes were uncorrelated with MAF and variants with LD Score below 100 were 4× enriched for heritability.

In all simulations, the sample size was 2062 individuals with full sample overlap between studies; the causal SNPs were best-guess imputed 1000 Genomes SNPs on chromosome 2, and the SNPs retained for the LD Score regression were HapMap 3 SNPs.

Estimates are averages across 100 simulations. Standard deviations (in parentheses) are calculated as the empirical standard deviation across 100 simulations.

LD Scores were estimated using in-sample LD and a 1cM window. The naive LD Score is simply $\sum r^2$ as in [2]. The 30 bins LD Score is a per-allele LD Score binned on a MAF by LD Score grid with MAF breaks at 0.05, 0.1, 0.2, 0.3 and 0.4 and LD Score breaks at 35, 75, 150 and 400. The 60 bins LD Score is a per-allele LD Score binned on a MAF by LD Score grid with MAF breaks at 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4 and 0.45 and LD Score breaks at 30, 60, 120, 200 and 300

These simulations demonstrate that naive LD Score regression can give accurate genetic correlation estimates even in situations where the heritability and genetic covariance estimates are badly biased, so long as genetic correlation does not depend on MAF or LD. In addition, these simulations demonstrate that MAF- and LD-binned LD Score regression can give accurate estimates of heritability and genetic covariance even for genetic architectures with MAF- and LD-dependence.

6.4 Simulations with Antiparallel LD- and MAF-Dependence

LD Score	h_1^2(5-50%)	h_2^2(5-50%)	ρ_g(5-50%)	r_g(5-50%)
Truth	0.83	0.89	0.33	0.38
Naive	0.36 (0.08)	1.13 (0.08)	0.32 (0.07)	0.5 (0.09)
30 Bins	0.8 (0.14)	0.93 (0.12)	0.34 (0.11)	0.39 (0.1)
60 Bins	0.79 (0.14)	0.93 (0.12)	0.33 (0.11)	0.39 (0.1)

This table displays simulations with MAF- and LD-dependent genetic architecture where the MAF- and LD- dependence was in opposite directions for each phenotype, and genetic correlation did not vary with MAF or LD. Precisely, per-allele effect sizes for the first phenotype were drawn from a normal distribution so that the variance of per-allele effect sizes were uncorrelated with MAF, and variants with LD Score below 100 were 4× enriched for heritability. Per-allele effect sizes for the second phenotype were drawn from a normal distribution so that the variance of per-allele effect size followed $\sqrt{p(1-p)}$, where p is MAF, and variants with LD Score above 100 were 4× enriched for heritability. Otherwise, the parameters of these simulations were the same as in 6.4 These simulations demonstrate that naive LD Score regression can give approximately accurate genetic correlation estimates even in situations where the heritability and genetic covariance estimates are badly biased, so long as genetic correlation does not depend on MAF or LD. In addition, these simulations demonstrate that MAF- and LD-binned LD Score regression can give accurate estimates of heritability and genetic covariance even for genetic architectures with MAF- and LD-dependence.

6.5 Simulations with LD- and MAF-Dependent Genetic Correlation

LD Score	h_1^2(5-50%)	h_2^2(5-50%)	ρ_g(5-50%)	r_g(5-50%)
Truth	0.83	0.89	0.41	0.48
Naive	0.38 (0.07)	1.15 (0.08)	0.5 (0.06)	0.75 (0.06)
30 Bins	0.81 (0.1)	0.94 (0.13)	0.42 (0.1)	0.48 (0.09)
60 Bins	0.81 (0.12)	0.94 (0.13)	0.42 (0.1)	0.48 (0.09)

This table displays simulations with MAF- and LD-dependent genetic architecture where the MAF- and LD- dependence was in opposite directions for each phenotype, and additionally, genetic correlation varied with MAF or LD. The parameters of these simulations were the same as 6.5, except that genetic correlation was 0.2 for variants with LD Score less than 100 and 0.8 for variants with LD Score greater than 100.

These simulations show that if genetic correlation (as opposed to just heritability or covariance) varies with MAF or LD, then all estimates from naive LD Score regression maybe be badly biased; however, MAF- and LD-binned LD Score regression give approximately correct results, at the cost of slightly higher standard error.

6.6 Comparison of Standard Error Estimates to Empirical Standard Deviation across Simulations

LD Score	$\widehat{se}(\hat{h}^2)$	$sd(\hat{h}^2)$	$\widehat{se}(\hat{\rho}_g)$	$sd(\hat{\rho}_g)$	$\widehat{se}(\hat{r}_g)$	$sd(\hat{r}_g)$
Naive	0.11 (0.01)	0.08	0.09 (0)	0.06	0.1 (0.03)	0.13
30 Bins	0.11 (0.01)	0.12	0.09 (0.01)	0.08	0.1 (0.01)	0.09
60 Bins	0.11 (0.01)	0.12	0.09 (0.01)	0.09	0.1 (0.01)	0.09

This table compares the block jackknife standard errors from ldsc (denoted $\widehat{se}$, which represents the mean standard error estimate across 100 simulation replicates) in the simulations from 6.4 to the empirical standard deviations of the parameter estimates (denoted sd) across 100 simulation replicates. The block jackknife standard errors closely match the empirical standard deviations. This confirms that block jackknife standard error estimates are approximately unbiased even with locally correlated error terms, so long as the block size is sufficiently large.

6.7 700 Genetic Correlations

6.8 Summary Statistic Metadata

References

[1] Brendan Bulik-Sullivan. Inferring Genetic Architecture with LD Score Kernel Regression. *In Preparation*, 2014.

[2] Brendan Bulik-Sullivan, Po-Ru Loh, Hilary Finucane, Stephan Ripke, Jian Yang, Nick Patterson, Mark J Daly, Alkes L Price, and Benjamin M Neale. Ld score regression distinguishes confounding from polygenicity in genome-wide association studies. *bioRxiv*, 2014.

[3] Hilary K Finucane and Brendan Bulik-Sullivan. Partitioning heritability with ld score regression. *In preparation*, 2014.

[4] David Golan and Saharon Rosset. Narrowing the gap on heritability of common disease by direct estimation in case-control gwas. *arXiv preprint arXiv:1305.5363*, 2013.

[5] JK Haseman and RC Elston. The investigation of linkage between a quantitative trait and a marker locus. *Behavior Genetics*, 2(1):3–19, 1972.

[6] Alkes L Price, Agnar Helgason, Gudmar Thorleifsson, Steven A McCarroll, Augustine Kong, and Kari Stefansson. Single-tissue and cross-tissue heritability of gene expression via identity-by-descent in related or unrelated individuals. *PLoS Genetics*, 7(2):e1001317, 2011.

[7] Alkes L Price, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, 38(8):904–909, 2006.

[8] Jian Yang, Beben Benyamin, Brian P McEvoy, Scott Gordon, Anjali K Henders, Dale R Nyholt, Pamela A Madden, Andrew C Heath, Nicholas G Martin, Grant W Montgomery, et al. Common snps explain a large proportion of the heritability for human height. *Nature Genetics*, 42(7):565–569, 2010.