Relationship between LD Score and Haseman-Elston Regression

Brendan Bulik-Sullivan*[1,2,3]

[1]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA
[2]Stanley Center for Psychiatric Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA
[3]Analytic and Translational Genetics Unit, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts, USA.

April 19, 2015

Abstract

Estimating SNP-heritability from summary statistics using LD Score regression provides a convenient alternative to standard variance component models, because LD Score regression is computationally very fast and does not require individual genotype data. However, the mathematical relationship between variance component methods and LD Score regression is not clear; in particular, it is not known in general how much of an increase in standard error one incurs by working with summary data instead of individual genotypes.

In this paper, I show that in samples of unrelated individuals, LD Score regression with constrained intercept is essentially the same as Haseman-Elston (HE) regression, which is currently the state-of-the-art method for estimating SNP-heritability from ascertained case/control samples. Similar results hold for SNP-genetic correlation.

Introduction

I begin by reviewing three estimators of SNP-heritability that can be applied to GWAS data: HE regression, REML and LD Score regression. These estimators are described elsewhere, so I provide only a brief overview, with references to more detailed derivations.

Consider a model where the N-vector of phenotypes Y is generated as $y = X\beta + \epsilon$, where X is an $N \times M$ matrix of standardized and centered genotypes, β is a vector of SNP effect sizes of length M, and ϵ is a vector of length N of residuals (which includes genetic effects orthogonal to an additive model, environmental effects, measurement error, etc).

If we condition on the study genotype matrix X, the entries of β as *i.i.d.* draws from a distribution with mean zero and variance h_g^2/M, and the entries of ϵ as *i.i.d.* draws from a distribution with mean zero and variance $1 - h_g^2$, then

$$\mathbb{E}[y_h y_i \,|\, X] = h_g^2 A_{hi}, \tag{1}$$

[1] where A is a normalized identity-by-state matrix $A := XX^{\mathsf{T}}/M$. Matrix A is typically called the empirical kinship matrix or genetic relatedness matrix (GRM). Equation 1 shows that we can

[*] Address correspondence to BBS (`bulik@broadinstitute.org`)

estimate h_g^2 by regressing products of phenotypes $y_h y_i$ against GRM entries A_{hi} for $h < i$. This is called Haseman-Elston regression [2, 1, 3, 4]. The estimator has a closed form:

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

where the hats over variance and covariance denote the sample variance and covariance. The HE regression estimator is inefficient, because the datapoints are correlated; conditional on X, $y_h y_i$ is correlated with $y_h y_j$.

If we are willing to make distributional assumptions about β and ϵ, then we can do better: if β and ϵ follow a normal distribution (or if β is sufficiently polygenic that $X\beta$ is approximately normal by a central limit theorem argument), then y is distributed as $N(0, h_g^2 A + (1 - h_g^2)I)$, and we can estimate h_g^2 via maximum likelihood (REML) [5]. This approach is implemented in the software package `GCTA` [6] (URLs).

The LD Score regression estimator of heritability [7, 8] takes as input GWAS summary statistics and LD data instead of a GRM and phenotypes. The precise LD data required are LD Scores, defined for each SNP j as $\ell_j := \sum_j r_{jk}^2$, where the sum is taken over all other SNPs k. In practice, there is very little LD in human samples outside of small window, so LD Scores are typically estimated using a 1 centiMorgam (cM) window [7]. The GWAS summary data required are 1 degree-of-freedom χ^2 statistics. Precisely, let χ_j^2 denote the Armitage Trend Test (ATT) statistic of SNP j, $\chi_j^2 := N(X_j^\top y)^2$ [9]. Under the same model as above, we have the regression equation

$$\mathbb{E}[\chi_j^2] = \frac{N h_g^2}{M} \ell_j + 1 + Na, \tag{3}$$

where a is a term that quantifies the average inflation in χ^2 statistics from cryptic relatedness or population stratification [7]. We can therefore estimate heritability by regressing χ_j^2 against ℓ_j and multiplying the slope by M/N. If the value of the intercept term $1 + Na$ is known ahead of time; for example, if the χ^2 statistics were generated from data with relatives removed and PC covariates [10] such that $a \approx 0$, then we can improve the efficiency of the regression can be improved by constraining the intercept. We refer to this estimator as LD Score regression with constrained intercept. The standard error can also be improved by weighting to account for heteroskedasticity [7]. Finally, the datapoints in this regression are non-independent (due to LD), so it is necessary to use a correlation-robust standard error such as a block jackknife [7, 8].

Results

Derivation

The HE regression estimator is typically written as a function of the GRM. However, in samples of unrelated individuals, it is possible to re-write the HE regression estimator in terms of linkage disequilibrium. Starting from the definition of covariance, we can rewrite the numerator of the HE regression estimator as

$$\begin{aligned}
\widehat{\mathrm{Cov}}[A_{hi}, y_h y_i] &:= \frac{1}{N(N-1)} \sum_{h \neq i} A_{hi} y_h y_i && (4)\\
&= \frac{1}{MN(N-1)} \sum_{h \neq i} \sum_{j=1}^{M} X_{hi} X_{ij} y_h y_i && (5)\\
&= \frac{1}{MN(N-1)} \sum_{j=1}^{M} \left(\sum_{h \neq i} X_{hj} y_h X_{ij} y_i \right) && (6)\\
&= \frac{1}{MN(N-1)} \sum_{j=1}^{M} \left(\left(\sum_{i=1}^{N} X_{ij} y_i \right)^2 - \sum_{i=1}^{N} X_{ij}^2 y_i^2 \right) && (7)\\
&= \frac{1}{MN(N-1)} \sum_{j=1}^{M} \left(N\chi_j^2 - \sum_{i=1}^{N} X_{ij}^2 y_i^2 \right). && (8)
\end{aligned}$$

All of the preceding steps are exact and follow from the definitions of the quantities in question. I am not aware of a convenient way to simplify the term $(MN)^{-1} \sum_j \sum_i X_{ij}^2 y_i^2$; however, this term is the mean over a large number of individuals and a large number of SNPs, so the law of large numbers suggests that replacing this term with its expectation should yield a good approximation. If the marginal effect size $\sum_k r_{jk}\beta_j$ of SNP j is small, which is the typical case in GWAS, then X_{ij} and y_i will be close to uncorrelated. Even if some SNPs have large effect sizes, the average marginal variance explained will still only be $\approx h_g^2 \bar{\ell}/M$, which is much less than 1 (where $\bar{\ell}$ denotes mean LD Score [7]). Therefore, we approximate $\mathbb{E}[X_{ij}^2 Y_i^2] \approx 1$. Thus,

$$\begin{aligned}
\frac{1}{MN(N-1)} \sum_{j=1}^{M} \left(N\chi_j^2 - \sum_{i=1}^{N} X_{ij}^2 y_i^2 \right) &\approx \frac{1}{M(N-1)} \sum_{j=1}^{M} (\chi_j^2 - 1)\\
&= \frac{\bar{\chi}^2 - 1}{(N-1)}, && (9)
\end{aligned}$$

where $\bar{\chi}^2$ denotes mean χ^2.

Next, we need to express the denominator of the HE regression estimator in terms of linkage disequilibrium. Beginning from the definition,

$$\begin{aligned}
\widehat{\mathrm{Var}}[A_{hi}] &:= \frac{1}{N(N-1)}\sum_{h\neq i} A_{hi}^2 && (10)\\
&= \frac{1}{MN(N-1)}\sum_{h\neq i}\left(\sum_{j=1}^{M} X_{hj}X_{ij}\right)^2 && (11)\\
&= \frac{1}{MN(N-1)}\sum_{j=1}^{M}\sum_{k=1}^{M}\sum_{h\neq i} X_{hj}X_{hk}X_{ij}X_{ik} && (12)\\
&= \frac{1}{MN(N-1)}\sum_{j=1}^{M}\sum_{k=1}^{M}\left(\left(\sum_{i=1}^{N} X_{ij}X_{jk}\right)^2 - \sum_{i=1}^{N} X_{ij}^2X_{ik}^2\right) && (13)\\
&= \frac{1}{M(N-1)}\sum_{j=1}^{M}\sum_{k=1}^{M}\left(\hat{r}_{jk}^2 - \frac{1}{N}\sum_{i=1}^{N} X_{ij}^2X_{ik}^2\right), && (14)
\end{aligned}$$

where $\hat{r}_{jk}^2$ denotes the squared correlation between genotypes at SNPs j and k in our sample. The preceding steps are exact, and rely only on the definitions of the quantities in question. At this stage, we again approximate a term with its expectation. The squared sample correlation $\hat{r}_{jk}^2$ is an upwardly biased estimator the squared population correlation [11]. In fact, the bias is equal to the expectation of $\frac{1}{N}\sum_{i=1}^{N} X_{ij}^2X_{ik}^2$. This means that the term in parentheses in Equation 14 is an unbiased estimate of the squared population correlation r_{jk}^2. If we replace the term in parentheses with its expectation, we have

$$\begin{aligned}
\frac{1}{M(N-1)}\sum_{j=1}^{M}\sum_{k=1}^{M}\left(\hat{r}_{jk}^2 - \frac{1}{N}\sum_{i=1}^{N} X_{ij}^2X_{ik}^2\right) &\approx \frac{N}{M(N-1)}\sum_{j=1}^{M}\sum_{k=1}^{M} r_{jk}^2\\
&= \frac{N}{M(N-1)}\sum_{j=1}^{M}\ell_j\\
&= \frac{N\bar{\ell}}{(N-1)}, && (15)
\end{aligned}$$

where $\bar{\ell}$ denotes mean LD Score. A similar derivation for the denominator appears in [3].

By dividing Equation 14 by Equation 15, we obtain

$$\hat{h}_{HE}^2 \approx \frac{(\bar{\chi}^2-1)}{N\bar{\ell}}. \qquad (16)$$

The approximation sign hides the fact that we have twice replaced terms with their expectations. However both of the terms that we replaced with expectations are means over a large number of terms, so by the law of large numbers, this approximation should be good. We verify this via simulation later in the paper.

To see that Equation 16 is equivalent to LD Score regression with the intercept constrained to one and regression weights $1/\ell$, first observe that by definition, unweighted LD Score regression with intercept constrained to one gives the estimator

$$\hat{h}^2_{constrain} := \frac{M\sum_j(\chi^2_j - 1)\ell_j}{N\sum_j \ell_j^2}. \tag{17}$$

In general, weighting the regression of y_i on x_i by w_i is equivalent to unweighted regression of $y_i\sqrt{w_i}$ on $x_i\sqrt{w_i}$. Therefore, weighting LD Score regression with constrained intercept by $1/\ell$ is the same as regressing $(\chi^2_j - 1)/\sqrt{\ell_j}$ against $\ell_j/\sqrt{\ell_j} = \sqrt{\ell_j}$ with constrained intercept. This gives the estimator

$$\begin{aligned}\hat{h}^2_{constrain,w} &:= \frac{M\sum_j\left((\chi^2 - 1)/\sqrt{\ell_j})(\ell_j/\sqrt{\ell_j}\right)}{N\sum_j(\ell_j\sqrt{\ell_j})^2}. \\ &= \frac{(\bar{\chi}^2 - 1)}{N\bar{\ell}},\end{aligned} \tag{18}$$

which is identical to Equation 16.

A parallel derivation in Appendix A shows that the HE regression estimator of genetic covariance is equivalent to the LD Score regression estimator of genetic covariance with $1/\ell_j$ regression weights and in-sample LD Scores.

Fixed Effects and Covariates

Suppose we model phenotypes as $y = X\beta + F + \epsilon$, where F is a matrix of covariates, and y, X, β represent phenotypes, genotypes and effect sizes as before. Let y' denote y residualized on F, and let X' denote X residualized on F. Then the HE regression estimator of h^2 controlling for fixed effects F is obtained by applying HE regression to y' and X'. Similarly, we can incorporate fixed effects into the ATT χ^2 statistic by taking $\chi^2_j(F) := N(X_j'^{\mathsf{T}}y')^2$. By the Frisch-Waugh-Lovell theorem [13], this χ^2 statistic is equivalent to N times the squared standardized regression coefficient of X_j in the multivariate regression $y \sim X_j + F$. It then follows immediately from the previous section that HE regression with covariates is equivalent to LD Score regression with constrained intercept, $1/\ell$ regression weights, F-adjusted χ^2 statistics $\chi^2(F)$, and LD Scores computed from X'.

What are the properties of LD Scores computed from X'? If F is a matrix of covariates that are uncorrelated with genotype in the population, *i.e.,* covariates that are not heritable, then X' is equal to X in expectation, and LD Scores computed from X' are equal to LD Scores computed from X in expectation. If F is a matrix of heritable covariates, then X and X' will differ in expectation. For example, if F is a matrix of the first 10 principal components of X, and X is a structured sample, then X' will be X with most of the population structure removed [10], and LD Scores computed from X' will be equal to LD Scores from X, except with spurious LD due to population structure removed.

The above derivations do not make any assumptions about sample structure; for example, the sample could include related individuals or a mixture of individuals with different ancestry. In these cases, HE regression is equivalent to LD Score regression with in-sample LD Scores estimated using a genome-wide window (*i.e.,* by taking the sum $\sum_{j=1}^{M} r^2_{jk}$ over all M SNPs).

The standard implementation of LD Score regression approximates in-sample LD by using LD Scores estimated from an external reference panel (such as 1000 Genomes [12]) and a 1cM window for estimating LD Scores [7]. Using a 1cM window can be viewed as a form of regularization: by taking the sum over only SNPs in a 1cM window, we reduce the variance of the estimate compared to taking the sum over all SNPs, and in samples where there is no long-range LD, we introduce only a small amount of bias. In structured samples, there will be long-range LD due to population structure; however, this LD will mostly be removed by regressing the top PCs out of the genotype matrix. If the out-of-sample LD Scores computed with a 1cM window are a good approximation to in-sample LD Scores computed with a genome-wide window after residualizing the genotypes on all principal components included as covariates in the GWAS, then the relationship between LD Score regression and HE regression should hold. If in-sample LD Score is inflated due to population structure or the inclusion of related individuals in the sample, then out-of-sample LD and in-sample LD will differ, and the estimates from HE regression and the standard implementation of LD Score regression will not be the same.

Simulations

In order to check the approximations in the preceding derivations, I performed a series of simulations. I simulated phenotypes according to an infinitesimal model using 98,839 HapMap 3 [14] SNPs on chromosome 2 and an approximately unstructured sample of 2,062 Swedish controls from [15], chosen to be representative of a typical GWAS cohort of unrelated individuals. In all simulations, the true h_g^2 was 0.5 and effect sizes were drawn from a normal distribution. I performed 100 total simulations. I estimated heritability with five estimators: residual maximum likelihood (REML, as implemented in the software package `GCTA` [6]), HE regression, LD Score regression with unconstrained intercept and default weights (as implemented in `ldsc` [7]), LD Score regression with intercept constrained to 1 and default weights, and LD Score regression with intercept constrained to 1 and $1/\ell$ regression weights. For all LD Score regression estimates, I used in-sample LD estimated with a 1 cM window, following [7].

Figure 1 shows a scatterplot of simulation results from all five estimators. The squared correlation between the estimates from HE regression and LD Score regression with constrained intercept and $1/\ell$ weights in these simulations was 0.999.

	Mean	SD
REML	0.50	0.050
HE	0.51	0.065
LDSC, intercept	0.52	0.091
LDSC, no intercept	0.52	0.060
LDSC, $1/\ell$ weights	0.52	0.066

Table 1: **Comparison of Heritability Estimators.** This table displays the mean and standard deviation of the h_g^2 estimates from several estimators across 100 simulations of quantitative traits. The true value of h_g^2 was 0.5. As expected, all estimators are approximately unbiased. REML gives the lowest standard error, followed by LD Score regression with default weights and intercept constrained to 1.

Means and standard deviations across 100 simulations for all five estimators are displayed in Table 1. As expected, all estimators give approximately unbiased estimates. REML is the most efficient, followed by LD Score regression with default weights and constrained intercept. The worst performing estimator is LD Score regression with unconstrained intercept. Nevertheless; LD Score regression with unconstrained intercept has some advantages that may compensate for the increased standard error. Fitting an intercept protects from bias due to cryptic relatedness and

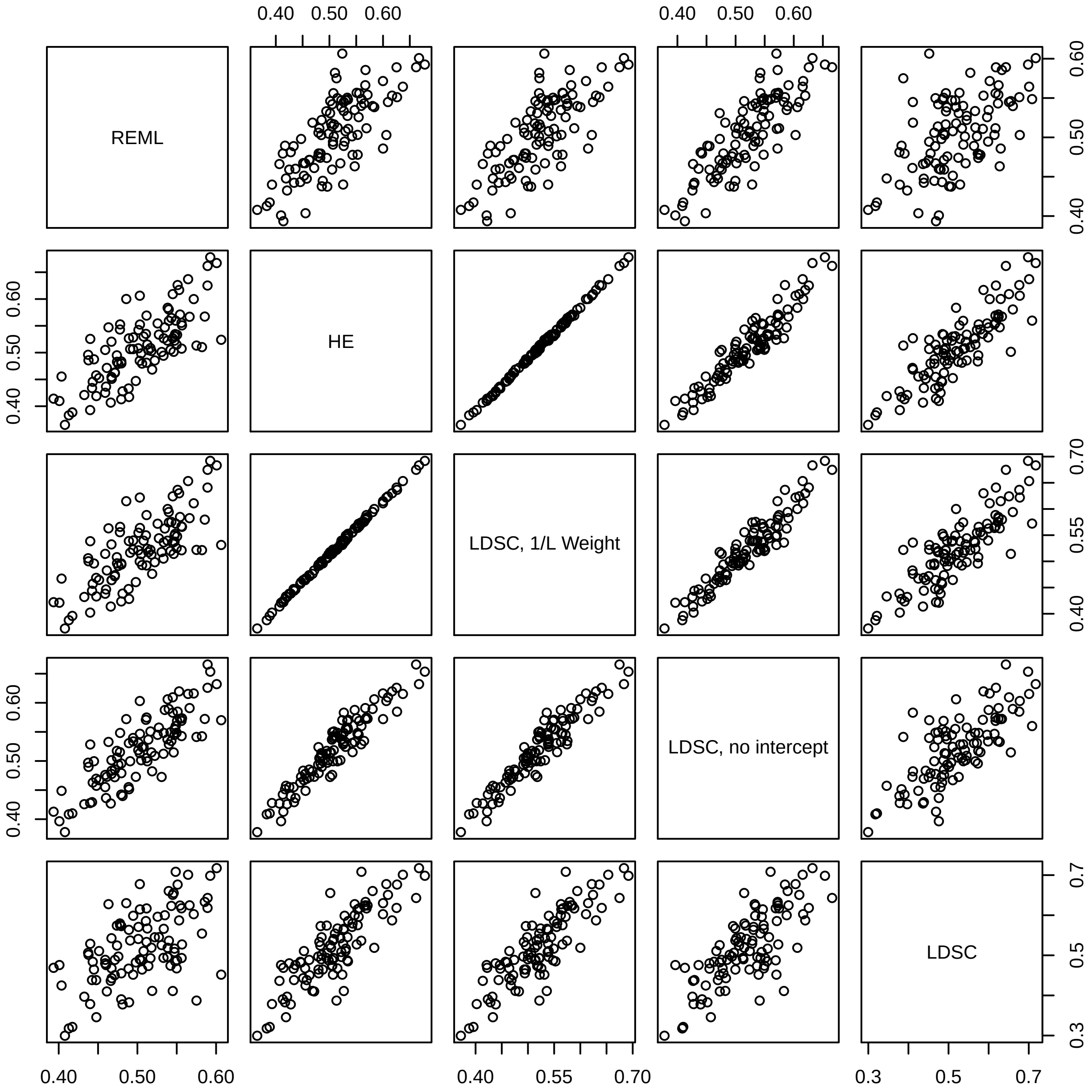

Figure 1: **HE Regression vs LD Score Regression.** Scatterplot displaying the relationships among the five heritability estimators across 100 simulation replicates. As predicted by the derivations, HE regression (HE) and LD Score regression with constrained intercept and $1/\ell$ regression weights were almost equivalent ($R^2 = 0.999$).

population stratification [7]; however, these advantages come at the cost of an increased standard error, due to the fact that LD Score regression with unconstrained intercept fits an extra parameter.

Discussion

I have derived an approximate equivalence between HE regression and LD Score regression with constrained intercept and $1/\ell$ regression weights. Although this equivalence is only approximate, mathematical arguments and simulations show that the approximation error is small. This provides a connection between standard kinship-based estimators of heritability and the LD Score regression estimators based on LD, and bounds the loss of precision incurred by working with summary statistics. In addition, several recent papers [3, 1] have shown that estimates of heritability from REML are biased downwards in ascertained case/control studies and recommended using HE regression instead. Since HE regression is approximately equivalent to LD Score regression, it follows that for case/control studies for which ancestry-matched LD Scores are available, LD Score regression should perform comparably to HE regression, but at lower computational cost ($\mathcal{O}(MN^2)$ time and $\mathcal{O}(N^2)$ space for HE regression vs $\mathcal{O}(MN)$ time and $\mathcal{O}(M+N)$ space for LD Score regression [7]).

URLs

1. `GCTA` software (REML):
`http://www.complextraitgenomics.com/software/gcta/`
2. `ldsc` software:
`github.com/bulik/ldsc`
3. Coffee:
`http://www.trianglecoffeeshop.com`

Acknowledgements

Thanks to H. Finucane, B. Neale, M, Daly, C. Arabica, P. Sullivan, P. Fontanillas, D. Posthuma and C. de Leeuw for helpful comments.

1 References

[1] David Golan, Eric S Lander, and Saharon Rosset. Measuring missing heritability: Inferring the contribution of common variants. *Proceedings of the National Academy of Sciences*, 111(49):E5272–E5281, 2014.

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

[3] Guo-Bo Chen. Estimating heritability of complex traits from genome-wide association studies using ibs-based haseman–elston regression. *Frontiers in genetics*, 5, 2014.

[4] Robert C Elston, Sarah Buxbaum, Kevin B Jacobs, and Jane M Olson. Haseman and elston revisited. *Genetic epidemiology*, 19(1):1–17, 2000.

[5] 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.

[6] Jian Yang, S Hong Lee, Michael E Goddard, and Peter M Visscher. Gcta: a tool for genome-wide complex trait analysis. *The American Journal of Human Genetics*, 88(1):76–82, 2011.

[7] 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. *Nature Genetics*, 2015.

[8] Brendan Bulik-Sullivan, Hilary K Finucane, Verneri Anttila, Alexander Gusev, Felix R Day, John RB Perry, Nick Patterson, Elise Robinson, Mark J Daly, Alkes L Price, et al. An atlas of genetic correlations across human diseases and traits. *bioRxiv*, page 014498, 2015.

[9] Peter Armitage. Tests for linear trends in proportions and frequencies. *Biometrics*, 11(3):375–386, 1955.

[10] 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.

[11] Ping Yin and Xitao Fan. Estimating r 2 shrinkage in multiple regression: A comparison of different analytical methods. *The Journal of Experimental Education*, 69(2):203–224, 2001.

[12] 1000 Genomes Project Consortium et al. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65, 2012.

[13] Ragnar Frisch and Frederick V Waugh. Partial time regressions as compared with individual trends. *Econometrica: Journal of the Econometric Society*, pages 387–401, 1933.

[14] International HapMap 3 Consortium et al. Integrating common and rare genetic variation in diverse human populations. *Nature*, 467(7311):52–58, 2010.

[15] Stephan Ripke, Colm O'Dushlaine, Kimberly Chambert, Jennifer L Moran, Anna K Kähler, Susanne Akterin, Sarah E Bergen, Ann L Collins, James J Crowley, Menachem Fromer, et al. Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nature genetics*, 45(10):1150–1159, 2013.

[16] Sang Hong Lee, Jian Yang, Michael E Goddard, Peter M Visscher, and Naomi R Wray. Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. *Bioinformatics*, 28(19):2540–2542, 2012.

Appendix A: Genetic Covariance

To begin, I will describe three estimators of genetic covariance that can be applied to GWAS data: HE regression, REML and LD Score regression. These estimators are derived elsewhere, so I provide only a brief overview, along with references to more detailed descriptions.

Consider a model where the vectors of phenotypes y_1 and y_2 are generated as $y_1 = Y\beta + \delta$ and $y_2 = Z\gamma + \epsilon$ where Y, Z are matrices of normalized and centered genotypes, β, γ are vectors of SNP effect sizes, and δ, ϵ are vectors of residuals (which includes genetic effects orthogonal to an additive model, environmental effects, measurement error, etc). Let N_1 denote the number of individual in matrix Y, N_2 the number of individuals in matrix Z, and N_s the number of individuals who appear in both matrices.

I model the entries of (β, γ) as *i.i.d.* draws from a distribution with mean zero and covariance matrix

$$\mathrm{Var}[(\beta_j, \gamma_j)] = \frac{1}{M}\begin{pmatrix} h_1^2 & \rho_g \\ \rho_g & h_2^2 \end{pmatrix}, \tag{19}$$

and the entries of (δ, ϵ) as *i.i.d.* draws from a distribution with mean zero and covariance matrix

$$\begin{pmatrix} (1-h_1^2)I & \rho_e \\ \rho_e & (1-h_2^2)I \end{pmatrix}, \tag{20}$$

where ρ_e denotes the environmental covariance. who have been phenotyped for both y_1 and y_2. If we condition on the study genotype matrices Y and Z, the covariance matrix of the vector (y_1, y_2) of phenotypes is

$$\mathrm{Var}[(y_1, y_2)] = \frac{1}{M}\begin{pmatrix} h_1^2 YY^\top & \rho_g YZ^\top \\ \rho_g ZY^\top & h_2^2 ZZ^\top \end{pmatrix} + \begin{pmatrix} (1-h_1^2)I & \rho_e I \\ \rho_e I & (1-h_2^2)I \end{pmatrix}, \tag{21}$$

Let $A := Y^\top Z/M$. 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}. This estimator is the HE regression estimator of genetic covariance, which is inefficient for the same reasons that the HE regression estimator of heritability is inefficient.

The closed from expression for the estimator is

$$\hat{\rho}_{g,HE} := \frac{\widehat{\mathrm{Cov}}[A_{hi}, y_{1h}y_{2i}]}{\widehat{\mathrm{Var}}[A_{hi}]}, \tag{22}$$

the variance and covariance are taken over all pairs (h, i) such that $h \neq i$. That is, if an individual i is one of the N_s individuals phenotyped for both traits, we do not include the term $y_{1i}y_{2i}$ in the regression. However, if $h \neq i$ but both h and i are among the N_s individuals phenotyped for both traits, we include both of the terms $y_{1h}y_{2i}$ and $y_{1i}y_{2h}$ in the regression. This is a slight difference from the single-phenotype case: if $y_1 = y_2$ then the terms $y_{1h}y_{2i}$ and $y_{1i}y_{2h}$ are identical, so it makes sense to include only one of these in the regression. If the phenotypes are not identical, then these two terms are distinct, so we would lose information by excluding one of them. There are $N_1N_2 - N_s$ terms in the regression.

If we are willing to assume that all distributions above are multivariate normal, then the distribution of the vector of phenotypes (y_1, y_2) is normal with mean zero and the variance equal to the matrix from Equation 21. The REML estimator of genetic covariance is obtained by maximizing the corresponding likelihood [16].

The LD Score regression estimator of genetic covariance [8] takes as input GWAS summary statistics and LD Scores instead of a GRM and phenotypes. Let $\hat{\beta}_j$ and $\hat{\gamma}_j$ denote the estimates of the effect size of SNP j on y_1 and y_2, respectively from marginal linear regression. Under the same model from above, we have the regression equation from [8]

$$\mathbb{E}[\hat{\beta}_j \hat{\gamma}_j] = \frac{\rho_g}{M}\ell_j + \frac{\rho N_s}{N_1 N_2}, \tag{23}$$

where ρ is the phenotypic correlation. We can therefore estimate genetic covariance by regressing $\hat{\beta}_j \hat{\gamma}_j$ against ℓ_j and multiplying the slope by M. If the value of the intercept term $\rho N_s / N_1 N_2$ is known ahead of time, the efficiency of the regression can be improved by constraining the intercept. The standard error can also be improved by weighting to account for heteroskedasticity. Finally, the datapoints in this regression are non-independent (due to LD), so it is necessary to use a correlation-robust estimator of the standard error [8].

Genetic Covariance HE Regression and LD Score Regression

As with the heritability estimators, there is a connection between the HE regression and LD Score regression estimators of genetic covariance. If we let S denote the set of individuals shared by both studies, then the numerator of the HE estimator is

$$\begin{aligned}
\widehat{\mathrm{Cov}}[A_{hi}, y_{ih} y_{2i}] &= \frac{1}{N_1 N_2 - N_s} \sum_{h,i} A_{hi} y_{1h} y_2 \\
&= \frac{1}{M(N_1 N_2 - N_s)} \sum_{h,i} \sum_{j=1}^{M} Y_{hj} Z_{ij} y_{1h} y_{2i} \\
&= \frac{1}{M(N_1 N_2 - N_s)} \sum_{j=1}^{M} \left(\left(\sum_{i=1}^{N_1} Y_{ij} y_{1j} \right) \left(\sum_{i=1}^{N_2} Y_{ij} y_{1j} \right) - \sum_{i \in S} Y_{ij}^2 y_{1i} y_{2i} \right) \\
&= \frac{N_1 N_2}{M(N_1 N_2 - N_s)} \sum_{j=1}^{M} \left(\hat{\beta}_j \hat{\gamma}_j - \sum_{i \in S} Y_{ij}^2 y_{1i} y_{2i} \right) \\
&\approx \frac{N_1 N_2}{M(N_1 N_2 - N_s)} \sum_{j=1}^{M} \left(\hat{\beta}_j \hat{\gamma}_j - \frac{\rho N_s}{N_1 N_2} \right). \qquad (24)
\end{aligned}$$

The approximation in the last line is valid in the regime of small effects, where Y_{ij}^2 and $y_{1i} y_{2i}$ are approximately uncorrelated, in which case $\mathbb{E}[Y_{ij}^2 y_{1i} y_{2i}] \approx \mathbb{E}[Y_{ij}^2]\mathbb{E}[y_{1i} y_{2i}] = \rho$.

The denominator of the HE regression estimator is

$$\begin{aligned}
\widehat{\mathrm{Var}}[A_{hi}] &= \frac{1}{N_1N_2 - N_s}\sum_{h,i} A_{hi}^2 \\
&= \frac{1}{M(N_1N_2 - N_s)}\sum_{h,i}\left(\sum_{j=1}^{M} Y_{hj}Z_{ij}\right)^2 \\
&= \frac{1}{M(N_1N_2 - N_s)}\sum_{h,i}\sum_{j=1}^{M}\sum_{k=1}^{M} Y_{hj}Z_{hk}Y_{ij}Z_{ik} \\
&= \frac{1}{M(N_1N_2 - N_s)}\sum_{j=1}^{M}\sum_{k=1}^{M}\left(\left(\sum_{i=1}^{N_1} Y_{ij}Y_{ik}\right)\left(\sum_{i=1}^{N_2} Z_{ij}Z_{ik}\right) - \sum_{i\in S} Y_{ij}^2Y_{ik}^2\right) \\
&\approx \frac{N_1N_2}{M(N_1N_2 - N_s)}\sum_{j=1}^{M}\sum_{k=1}^{M}\left(\hat{r}_{Y,jk}\hat{r}_{Z,jk} - \frac{N_s}{N_1N_2}\right), \qquad (25)
\end{aligned}$$

where $\hat{r}_{Y,jk}$ denotes the sample correlation between j and k in matrix Y and likewise for $\hat{r}_{Z,jk}$. The approximation in the second-to-last line results from observing that almost all pairs of SNPs j, k are in linkage equilibrium, so since our genotype matrix is normalized to mean zero and variance one, then the variance of their product $Y_{ij}Y_{ik}$ will be approximately equal to the product of their variances, which is one. Observe that $\hat{r}_{Y,jk}\hat{r}_{Z,jk} - N_s/(N_1N_2)$ is an unbiased estimator of r^2, so

$$\begin{aligned}
\frac{N_1N_2}{2M(N_1N_2 - N_s)}\sum_{j=1}^{M}\sum_{k=1}^{M}\left(\hat{r}_{Y,jk}\hat{r}_{Z,jk} - \frac{N_s}{N_1N_2}\right) &\approx \frac{N_1N_2}{2M(N_1N_2 - N_s)}\sum_{j=1}^{M}\ell_{jk} \\
&= \frac{N_1N_2}{2(N_1N_2 - N_s)}\overline{\ell}. \qquad (26)
\end{aligned}$$

Thus, we can rewrite the HE regression estimator as

$$\hat{h}^2_{HE} \approx \frac{M}{\overline{\ell}}\sum_j\left(\hat{\beta}_j\hat{\gamma}_j - \frac{\rho N_s}{N_1N_2}\right). \qquad (27)$$

This is equivalent to LD Score regression with the intercept constrained to $\rho N_s/(N_1N_2)$ and regression weights $1/\ell$. To see this, first observe that unweighted LD Score regression with intercept constrained to one yields the estimator

$$\hat{\rho}^2_{g,constrain} := \frac{M\sum_j(\hat{\beta}_j\hat{\gamma}_j - \rho N_s/(N_1N_2))\ell_j}{\sum_j \ell_j^2}. \qquad (28)$$

Weighting the regression by $1/\ell$ is the same as regressing $(\hat{\beta}_j\hat{\gamma}_j - \rho N_s/(N_1N_2))/\sqrt{\ell_j}$ against $\ell_j/\sqrt{\ell_j} = \sqrt{\ell_j}$. This yields the estimator

$$\begin{aligned}
\hat{\rho}_{g,w} &:= \frac{M\sum_j((\hat{\beta}_j\hat{\gamma}_j - \rho N_s)/\sqrt{\ell_j})(\ell_j/\sqrt{\ell_j})}{\sum_j(\ell_j\sqrt{\ell_j})^2}. \\
&= \frac{M}{\overline{\ell}}\sum_j\left(\hat{\beta}_j\hat{\gamma}_j - \frac{\rho N_s}{N_1N_2}\right), \qquad (29)
\end{aligned}$$

which is identical to equation 27. Using $1/\ell$ for regression weights is more efficient than unweighted LD Score regression, but still sub-optimal [8]).