Supplementary Note

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

James J. Lee, Matt McGue, William G. Iacono, Carson C. Chow

Here we supply mathematical details justifying the results reported in the main text. The standard linear model of quantitative genetics is

$$y = X\alpha + e, \tag{1}$$

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. Now recall that the heritability of a phenotype is the proportion of its variance attributable to differences in the breeding (genetic) values $X\alpha \in \mathbb{R}^n$. Because of standardization, the variance is the expected squared element of $X\alpha$. From these definitions, the heritability of the phenotype attributable to the average effects of the p SNPs is

$$h^2 = (1/n)\alpha' X' X \alpha, \tag{2}$$

although LD Score regression uses the definition $h^2_{\text{LE}} = \alpha'\alpha$. The quantities, h^2 and h^2_{LE}, coincide if all causal SNPs 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 h^2 and h^2_{LE} 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

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

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 subpopulations varying in allele frequencies and exposures to environmental causes. We also note that

$$\mathbb{E}_n(X_{ij}X_{ik}) = \frac{1}{n}\sum_i X_{ij}X_{ik} \to \frac{1}{n}X'X = \Gamma. \tag{4}$$

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

$$l_j = \gamma_j'\gamma_j. \tag{5}$$

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

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 $\hat{\beta}_j$ 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

$$\mathbb{E}_n(\hat{\beta}_j^2) = \frac{1}{n^2}\mathbb{E}_n\left(x_j'X\alpha\alpha'X'x_j + x_j'X\alpha e'x_j + x_j'e\alpha'X'x_j + x_j'ee'x_j\right). \tag{6}$$

To evaluate Equation (6), we must compute the multivariate fourth moment of X, which is in general difficult to estimate given current GWAS population sizes. However, the multivariate fourth moment for a zero-mean random variable can be written as

$$\mathbb{E}_n(x_ax_bx_cx_d) = \mathbb{E}_n(x_ax_b)\mathbb{E}_n(x_cx_d) + \mathbb{E}_n(x_ax_c)\mathbb{E}_n(x_bx_d) + \mathbb{E}_n(x_ax_d)\mathbb{E}_n(x_bx_c) + c_4(x_a, x_b, x_c, x_d) \tag{7}$$

where c_4 is the multivariate fourth cumulant, which captures any four-body interactions beyond pairwise correlations. Barring undue multiple-site selection pressure, c_4 should be small and can be neglected. Note that the adequacy of this assumption for our purposes is borne out by the simulation results reported in the main text. Applying (7) with c_4 set to zero to the first term

of Equation (6) and recalling that $(x_j'X)_k = \sum_{i=1}^n X_{ij}X_{ik}$, yields

$$\begin{aligned}
&\mathbb{E}_n(x_j'X\alpha\alpha'X'x_j) \\
&= \sum_{k,k',i,i'} \alpha_k\alpha_{k'}\mathbb{E}_n\left(X_{ij}X_{ik}X_{i'j}X_{i'k'}\right) \\
&\approx \sum_{k,k',i,i} \alpha_k\alpha_{k'}[\mathbb{E}_n(X_{ij}X_{ik})\mathbb{E}_n(X_{i'j}X_{i'k'}) + \mathbb{E}_n(X_{ij}X_{i'j})\mathbb{E}_n(X_{ik}X_{i'k'}) + \mathbb{E}_n(X_{ij}X_{i'k'})\mathbb{E}_n(X_{ik}X_{i'j})]
\end{aligned} \tag{8}$$

which is straightforward to evaluate. The first term of (8) yields

$$\begin{aligned}
\sum_{k,k',i,i'} \alpha_k\alpha_{k'}\mathbb{E}_n(X_{ij}X_{ik})\mathbb{E}_n(X_{i'j}X_{i'k'}) &= n^2\sum_{k,k'} \alpha_k\alpha_{k'}\Gamma_{jk}\Gamma_{jk'} \\
&= n^2\beta_j^2,
\end{aligned} \tag{9}$$

where we have employed the identity

$$\beta_j = \sum_k \Gamma_{jk}\alpha_k = \gamma_j'\alpha$$

implied by the path-tracing rules (Wright, 1934) and which can also be seen by noting that $(X'X)\hat{\alpha} = X'y = n\hat{\beta}$ and $X'X = n\Gamma$.

The second term of (8) gives

$$\begin{aligned}
\sum_{k,k',i,i'} \alpha_k\alpha_{k'}\mathbb{E}_n(X_{ij}X_{i'j})\mathbb{E}_n(X_{ik}X_{i'k'}) &= \sum_{k,k',i} \alpha_k\alpha_{k'}\mathbb{E}_n(X_{ij}^2)\mathbb{E}_n(X_{ik}X_{ik'}) \\
&= n\sum_{k,k'} \alpha_k\alpha_{k'}\Gamma_{kk'} \\
&= n\alpha'\Gamma\alpha \qquad (10) \\
&= nh^2. \qquad (11)
\end{aligned}$$

If $\sum_{k\neq k} \Gamma_{kk'}\alpha_k\alpha_{k'}$ is negligible then Γ is approximately the identity matrix and $h^2 = h_{\text{LE}}^2$. Note that this is the quantity distinguishing the two definitions of heritability mentioned above. If all causal sites ($\alpha_k \neq 0$) are in LE ($\Gamma_{kk'} = 0$), then the approximation is exact. Hereon, we will set $h^2 = h_{\text{LE}}^2$. In the third term of (8), the expectation factor $\mathbb{E}_n(X_{ij}X_{i'k'})$ is zero except when $i = i'$, and thus the third term can be evaluated to

$$\begin{aligned}
\sum_{k,k',i,i'} \alpha_k\alpha_{k'}\mathbb{E}_n(X_{ij}X_{i'k'})\mathbb{E}_n(X_{ik}X_{i'j}) &= \sum_{k,k',i} \alpha_k\alpha_{k'}\mathbb{E}_n(X_{ij}X_{ik'})\mathbb{E}_n(X_{ik}X_{ij}) \\
&= n\sum_{k,k'} \alpha_k\alpha_{k'}\Gamma_{jk'}\Gamma_{jk} \\
&= n\beta_j^2. \qquad (12)
\end{aligned}$$

Together they yield

$$\sum_{k,k',i,i'} \alpha_k \alpha_{k'} \mathbb{E}_n \left(X_{ij} X_{ik} X_{i'j} X_{i'k'} \right) \approx (n^2 + n)\beta_j^2 + nh^2. \tag{13}$$

The second, third, and fourth expectation terms in Equation (6) can be evaluated similarly. Factors of the form $\mathbb{E}_n(X_{ij}e_i)$ can be replaced with v_j—the confounding-induced covariance between SNP j and the non-genetic residual. The result is

$$\begin{aligned}
&\sum_k 2n^2 \alpha_k \Gamma_{jk} v_j + \sum_k 2n\alpha_k v_j \Gamma_{kj} + \sum_k 2n\alpha_k v_k \\
&\quad + n^2 v_j^2 + n v_j^2 + n \left[1 - h^2 - 2\mathrm{Cov}_n \left(\sum_k X_{ik}\alpha_k, e_i \right) \right] \\
&= 2n^2 \beta_j v_j + 2n\beta_j v_j + n^2 v_j^2 + n v_j^2 + n(1 - h^2).
\end{aligned} \tag{14}$$

Substituting (13) and (14) back into Equation (6) and assigning $\chi_j^2 = n\hat{\beta}_j^2$ gives

$$\begin{aligned}
\mathbb{E}_n(\chi_j^2) &= (n+1)\beta_j^2 + 1 + 2(n+1)\beta_j v_j + (n+1)v_j^2 \\
&\approx n\beta_j^2 + n v_j^2 + 2n\beta_j v_j + 1 \\
&= n(\gamma_j \alpha)^2 + n v_j^2 + 2n\beta_j v_j + 1.
\end{aligned} \tag{15}$$

If we let θ_j be the angle between γ_j and α_j then

$$\mathbb{E}_n(\chi_j^2) = n\alpha' \alpha l_j \cos^2 \theta_j + n v_j^2 + 2n v_j |\alpha_j| \sqrt{l_j} \cos \theta_j \beta_j + 1. \tag{16}$$

Now consider how the chi-square statistic is related to the heritability, in the manner of Bulik-Sullivan et al. (2015a). Their formulation, in our notation, is

$$\mathbb{E}_p(\chi_j^2 \,|\, l_j) \approx \frac{n}{p} l_j h^2 + n v_j^2 + 1. \tag{17}$$

In the absence of confounding ($v = 0$), a direct comparison to (16) shows that in order for (17) to hold,

$$h^2 = \alpha' \alpha p \mathbb{E}_p(\cos^2 \theta_j) \tag{18}$$

which is not true in general. This clearly demonstrates that LD Score regression is not a reliable estimator for heritability. However, if the γ_j are uniformly distributed at random with respect to α (i.e. LD score is not correlated with effect size), then we have $\mathbb{E}_p(\cos^2 \theta_j) = 1/p$ and the slope is a good estimator of heritability.

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

Here we show that even when the slope is an unreliable estimator of heritability, the intercept can still be a good estimator of confounding. The LD Score prediction for confounding as seen from

setting $l_j = 0$ in (17) is given by $1 + nv_j^2$. In the event that v_j^2 itself has a linear l_j dependence then the intercept would be $1 + nv_0^2$ where $v_0^2 = v_j^2\Big|_{l_j=0}$. The question then is whether this estimate of confounding would still hold if the slope of the LD Score regression was no longer an unbiased estimate of the heritability.

From (15) and the linearity of regression, we clearly see that the intercept is always $1 + nv_0$ as long as 1) the regression of β_j^2 on l_j has intercept zero, 2) the regression of $\beta_j v_j$ on l_j has intercept zero, and 3) the regression of v_j^2 on l_j has intercept v_0^2.

Condition 1) should almost always hold to an excellent approximation. The near linearity of the (χ_j^2, l_j) scatterplot has been found in all empirical applications of LD Score regression so far, strongly supporting the linearity of the (β_j^2, l_j) regression as well. 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 v_0^2. Condition 2) is trivially satisfied if $\beta_j v_j$ has zero mean and is uncorrelated with l_j. This is expected since it is unlikely that there would be a correlation over SNPs between the regression coefficient induced by true genetic effects and the extent of confounding with non-genetic factors affecting the phenotype. It is also unlikely that there be a correlation within any bin of SNPs with the same LD Score. Condition 3) implies that the intercept for a regression of v_j^2 as a function of l_j is the same as v_j^2 evaluated at $l_j = 0$. This is certainly true if v_j^2 is an affine function of l_j. Mild nonlinearities can bias the intercept but this is also alleviated by over weighting the smallest LD Scores.

There is a subtlety in that the quantity $1+v_0^2$, which the intercept is expected to equal under fairly general conditions, is not necessarily the average chi-square statistics of all null SNPs. It is instead a linear extrapolation to the average chi-square statistic of SNPs that are null by virtue of tagging few other SNPs of any kind. Whether we can expect this to equal the average chi-square statistic of all null SNPs is an issue considered in the main text.

An example of confounding is depicted in Figure 1 of the main text. This example can be represented by

$$\text{offspring SNP } j \xleftarrow{\Gamma_{jk}} \text{offspring SNP } k \xrightarrow{\alpha_k} \text{offspring trait} \tag{19}$$

contributing to β_j and the confounding path

$$\text{offspring trait} \xleftarrow{\lambda} \text{parent trait} \xleftarrow{\alpha_k} \text{parent SNP } k \xleftarrow{\Gamma_{jk}} \text{parent SNP } j \xrightarrow{1/2} \text{offspring SNP } j \tag{20}$$

making a contribution to v_j of the same sign. Note we have used λ to denote the direct causal effect of parent on offspring phenotype. This type of confounding is rather peculiar in that it only inflates the regression coefficients of SNPs truly associated with the trait by an amount that

depends on their unconfounded coefficients (Lee, 2012). That is, those SNPs with the largest values of

$$\beta_j^2 = \left(\sum_k \Gamma_{jk} \alpha_k \right)^2$$

will also have the largest values of

$$\beta_j v_j \propto \frac{1}{2} \lambda \left(\sum_k \Gamma_{jk} \alpha_k \right)^2 \quad \text{and} \quad v_j^2 \propto \left(\frac{1}{2} \lambda \sum_k \Gamma_{jk} \alpha_k \right)^2.$$

This is a form of l_j-dependent confounding because $(2\beta_j v_j + v_j^2)$ grows with the LD Score l_j, just as does β_j^2. But the corollary is that if the genuine association $\sum_k \Gamma_{jk} \alpha_k$ is null, then the spurious association $(2\beta_j v_j + v_j^2)$ is null as well because $\sum_k \Gamma_{jk} \alpha_k$ is a factor in both of these latter terms.

3 Bivariate LD Score regression as an estimator of genetic correlations

The linear model for two traits is given by

$$\begin{aligned} y_1 &= X_1 \alpha_1 + e_1, \\ y_2 &= X_2 \alpha_2 + e_2. \end{aligned} \tag{21}$$

The genetic correlation is defined as $r := \alpha_1' \alpha_2 / \sqrt{h_1^2 \, h_2^2}$. We will use r_{LDSC} to denote the genetic correlation as it is estimated by bivariate LD Score correlation and examine when it gives an unbiased estimate of r. The dependent variable in bivariate LD Score regression is 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 the small c_4 approximation to evaluate the expectation and obtain

$$\begin{aligned} \mathbb{E}_n(Z_{1j} Z_{2j}) \approx n\beta_{1j}\beta_{2j} + n\beta_{1j} v_{2j} + n\beta_{2j} v_{1j} + n v_{1j} v_{2j} \\ + \text{Cov}_n \left(\sum_k X_{ik} \alpha_{1k}, e_{2i} \right) + \text{Cov}_n \left(\sum_k X_{ik} \alpha_{2k}, e_{1i} \right) + \rho. \end{aligned} \tag{22}$$

where $\rho := \rho_g + \rho_e$, $\rho_g := \alpha_1'\alpha_2 = \alpha_2'\alpha_1$ is the genetic covariance, and $\rho_e := \mathbb{E}_n(e_{i1}e_{i2})$ is the environmental covariance. The last three terms in (22) 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

$$\begin{aligned}\mathbb{E}_n(Z_{1j}Z_{2j}) &\approx n\beta_{1j}\beta_{2j} + n\beta_{1j}v_{2j} + n\beta_{2j}v_{1j} + nv_{1j}v_{2j} \\ &= n\gamma_j'\alpha_1\gamma_j'\alpha_2 + n\beta_{1j}v_{2j} + n\beta_{2j}v_{1j} + nv_{1j}v_{2j}. \end{aligned} \tag{23}$$

from which we can infer the genetic correlation by regression on LD Score.

The analogous expression from Bulik-Sullivan et al. (2015b) is

$$\mathbb{E}_p(Z_{1j}Z_{2j} \,|\, l_j) \approx \frac{n}{p}\alpha_1'\alpha_2\gamma_j'\gamma_j := \frac{n}{p}\rho_{g,\mathrm{LDSC}}l_j \tag{24}$$

and the estimate of genetic correlation is given by

$$r_{\mathrm{LDSC}} = \frac{\rho_{g,\mathrm{LDSC}}}{\sqrt{h^2_{1,\mathrm{LDSC}}\, h^2_{2,\mathrm{LDSC}}}}. \tag{25}$$

where $h^2_{i,\mathrm{LDSC}}$ is estimated from univariate LD Score regression. In the regression of $Z_{1j}Z_{2j}$ on l_j to obtain $\rho_{g,\mathrm{LDSC}}$, 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.

The expectation value of (23) over all SNPs and (24) are equivalent if

$$\rho_{g,\mathrm{LDSC}} \equiv \mathbb{E}_p\left(\frac{1}{p}\alpha_1'\alpha_2\gamma_j'\gamma_j\right) = \mathbb{E}_p(\gamma_j'\alpha_1\gamma_j'\alpha_2), \tag{26}$$

which is not generally true. As in the univariate case above, the righthand side of Equation (26) can be rewritten as

$$\mathbb{E}_p(\gamma_j'\alpha_1\gamma_j'\alpha_2) = |\alpha_1||\alpha_2|l_j\mathbb{E}_p(\cos\theta_j^1\cos\theta_j^2), \tag{27}$$

where $\cos\theta_j^k$ is the unit-vector projection of α_k onto γ_j. From (27) 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 see that this bias cancels from the numerator and denominator of Equation (25). 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 the projection of α_1 and α_2 onto the γ basis. The expectation over SNPs in (27) is equivalent to taking the dot product between $\cos\theta_j^1$, which is the unit vector projection of α_1 onto γ_j, and $\cos\theta_j^2$.

We first assume that the γ_j form an orthogonal basis and there is no correlation between l_j and $\cos\theta_j$. (In reality, they are nearly orthogonal since $\gamma_j'\gamma_k$ vanishes for $|j-k|$ sufficiently

distant.) Then the angle between α_1 and α_2 is preserved in the new basis, and we have the condition

$$\begin{aligned}\mathbb{E}_p(\gamma_j'\alpha_1\gamma_j'\alpha_2) &= |\alpha_1||\alpha_2|\sqrt{\mathbb{E}_p(\cos^2\theta_j^1)\mathbb{E}_p(\cos^2\theta_j^2)}\cos\theta_{12}l_j,\\ &= \alpha_1'\alpha_2\sqrt{\mathbb{E}_p(\cos^2\theta_j^1)\mathbb{E}_p(\cos^2\theta_j^2)}l_j,\end{aligned}$$

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

$$\rho_{g,\mathrm{LDSC}} \approx \alpha_1'\alpha_2 p\sqrt{\mathbb{E}_p(\cos^2\theta_j^1)\mathbb{E}_p(\cos^2\theta_j^2)}. \tag{28}$$

Inserting (28) and (18) into (25) then gives

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

which is an unbiased estimator of the genetic correlation.

If the γ vectors do not form an orthogonal basis, then LD Score regression will not produce an unbiased estimator of genetic correlation. 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

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

We now decompose $\Lambda^2 = \lambda^2 I + \Delta$, $\operatorname{Tr}\Delta = 0$, where λ^2 represents the average correlation of γ_j and α and Δ represents the deviation from orthogonality. This yields

$$p\mathbb{E}_p(\gamma_j'\alpha_1\gamma_j'\alpha_2) = \lambda^2\alpha_1'\alpha_2 + \alpha_1' S\Delta S'\alpha_2, \tag{29}$$

The genetic correlation is then given by

$$\begin{aligned}&\frac{\lambda^2\alpha_1'\alpha_2 + \alpha_1' S\Delta S'\alpha_2}{\sqrt{\lambda^2\alpha_1'\alpha_1 + \alpha_1' S\Delta S'\alpha_1}\sqrt{\lambda^2\alpha_2'\alpha_2 + \alpha_2' S\Delta S'\alpha_2}}\\ &= \frac{\alpha_1'\alpha_2}{\sqrt{\alpha_1'\alpha_1\alpha_2'\alpha_2}}\frac{1 + \alpha_1' S\Delta S'\alpha_2/\lambda^2\alpha_1'\alpha_2}{\sqrt{1 + \alpha_1' S\Delta S'\alpha_1/\lambda^2\alpha_1'\alpha_1}\sqrt{1 + \alpha_2' S\Delta S'\alpha_2/\lambda^2\alpha_2'\alpha_2}}\end{aligned} \tag{30}$$

The λ^2 bias is effectively cancelled by the same biases in the heritability estimates but the Δ bias will remain. The effect of the Δ bias is also most pronounced when the genetic correlation is low. It vanishes if the genetic correlation is unity.

We now address the accuracy of bivariate LD Score regression in the case of confounding by parent genotype. As shown above the bivariate LD Score regression makes use of three separate regressions:

1. the regression of $Z_{1j}Z_{2j} := n\hat{\beta}_{1j}\hat{\beta}_{2j}$ on l_j;
2. the regression of $\chi^2_{1j} := n\hat{\beta}^2_{1j}$ on l_j; and
3. the regression of $\chi^2_{2j} := n\hat{\beta}^2_{2j}$ on l_j.

The numerator of the estimated genetic correlation is the slope of regression 1. The denominator is the square root of slope 2 times slope 3. Thus, if l_j is independent of all terms in Equation (23) other than β^2_{j1}, β^2_{j2}, and $\beta_{1j}\beta_{2j}$ (the standard assumption), then bivariate LD Score regression estimates the quantity

$$r_{\text{LDSC}} = \frac{\text{Cov}_p(\beta_{1j}\beta_{2j}, l_j)}{\sqrt{\text{Cov}_p(\beta^2_{1j}, l_j)\text{Cov}_p(\beta^2_{2j}, l_j)}}. \tag{31}$$

If Equation (31) is an unbiased estimator of the genetic correlation, then all we need to show is that the output of bivariate LD Score regression continues to be given by Equation (31) even when the type of confounding represented by the path in (20) is operating.

Introducing the subscripts 1 and 2 to denote the trait, we note again that those SNPs with the largest values of

$$\beta^2_{1j} = \left(\sum_k \Gamma_{jk}\alpha_{1k}\right)^2$$

will also have the largest values of

$$\beta_{1j}v_{1j} \propto \frac{1}{2}\lambda_1\left(\sum_k \Gamma_{jk}\alpha_{1k}\right)^2 \quad \text{and} \quad v^2_{1j} \propto \left(\frac{1}{2}\lambda_1\sum_k \Gamma_{jk}\alpha_{1k}\right)^2.$$

Similarly, those SNPs with the largest values of $\beta_{1j}\beta_{2j}$ will have the largest values of $\beta_{2j}v_{1j}$, $\beta_{1j}v_{2j}$, and $v_{1j}v_{2j}$. This means that the rank-correlation matrix of the variables

$$(l_j, \beta^2_{1j}, \beta^2_{2j}, \beta_{1j}\beta_{2j})$$

in the absence of the path in (20) will be the same as that of

$$(l_j, \beta^2_{1j} + 2\beta_{1j}v_{1j} + v^2_{1j}, \beta^2_{2j} + 2\beta_{2j}v_{2j} + v^2_{2j}, \beta_{1j}\beta_{2j} + \beta_{1j}v_{2j} + \beta_{2j}v_{1j} + v_{1j}v_{2j})$$

in the case when the path is active.

The estimator of the genetic correlation becomes

$$
\begin{aligned}
r_{\text{LDSC}} &= \frac{\text{Cov}_p(\beta_{1j}\beta_{2j} + \cdots, l_j)}{\sqrt{\text{Cov}_p(\beta_{1j}^2 + \cdots, l_j)\text{Cov}_p(\beta_{2j}^2 + \cdots, l_j)}} \\
&= \frac{\left(1 + \frac{\lambda_1}{2} + \frac{\lambda_2}{2} + \frac{\lambda_1\lambda_2}{4}\right)\text{Cov}_p(\beta_{1j}\beta_{2j}, l_j)}{\sqrt{\left(1 + \lambda_1 + \frac{\lambda_1^2}{4}\right)\text{Cov}_p(\beta_{1j}^2, l_j)\left(1 + \lambda_2 + \frac{\lambda_2^2}{4}\right)\text{Cov}_p(\beta_{2j}^2, l_j)}} \\
&= \frac{\text{Cov}_p(\beta_{1j}\beta_{2j}, l_j)}{\sqrt{\text{Cov}_p(\beta_{1j}^2, l_j)\text{Cov}_p(\beta_{2j}^2, l_j)}},
\end{aligned}
$$

in agreement with Equation (31).

References

Bulik-Sullivan, B., Finucane, H. K., Anttila, V., Gusev, A., Day, F. R., Loh, P.-R., … Neale, B. M. (2015b). 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. (2015a). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics*, *47*(3), 291–295. doi:10.1038/ng.3211

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

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

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

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