The American Journal of Human Genetics, Volume *109*

Supplemental information

Leveraging LD eigenvalue regression to improve the estimation of SNP heritability and confounding inflation

Shuang Song, Wei Jiang, Yiliang Zhang, Lin Hou, and Hongyu Zhao

Contents

1 Supplementary Figures

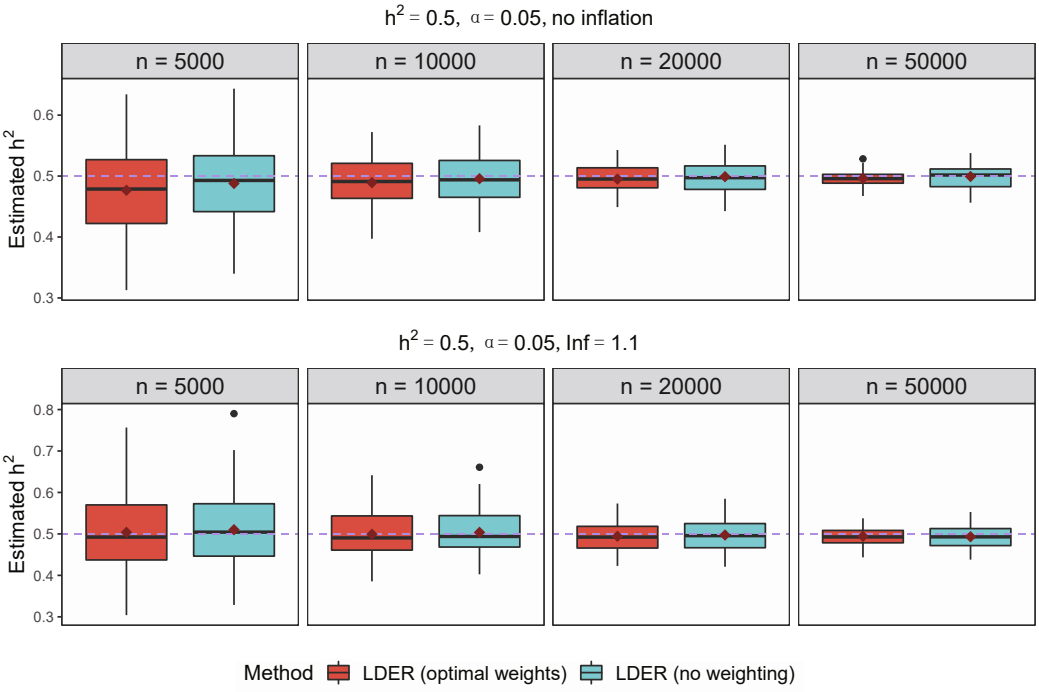
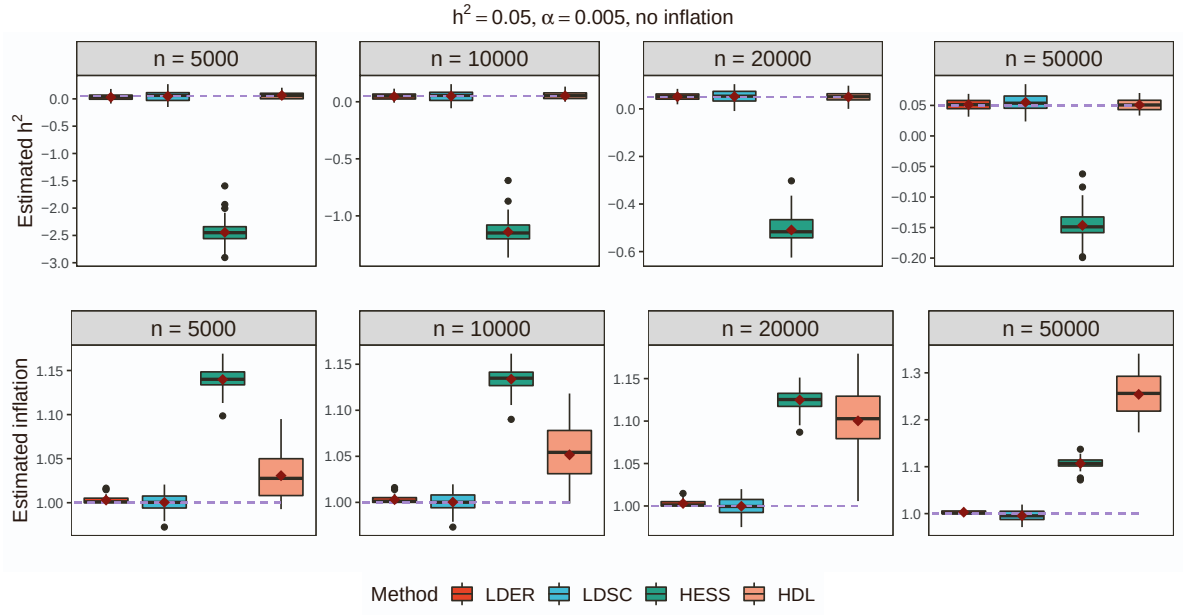
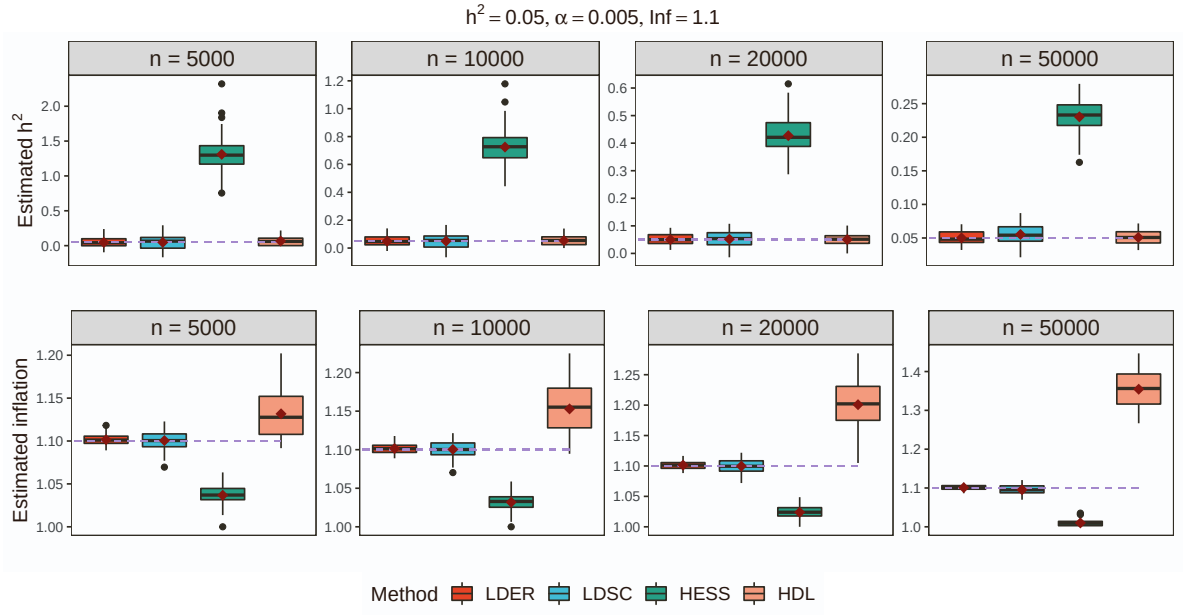
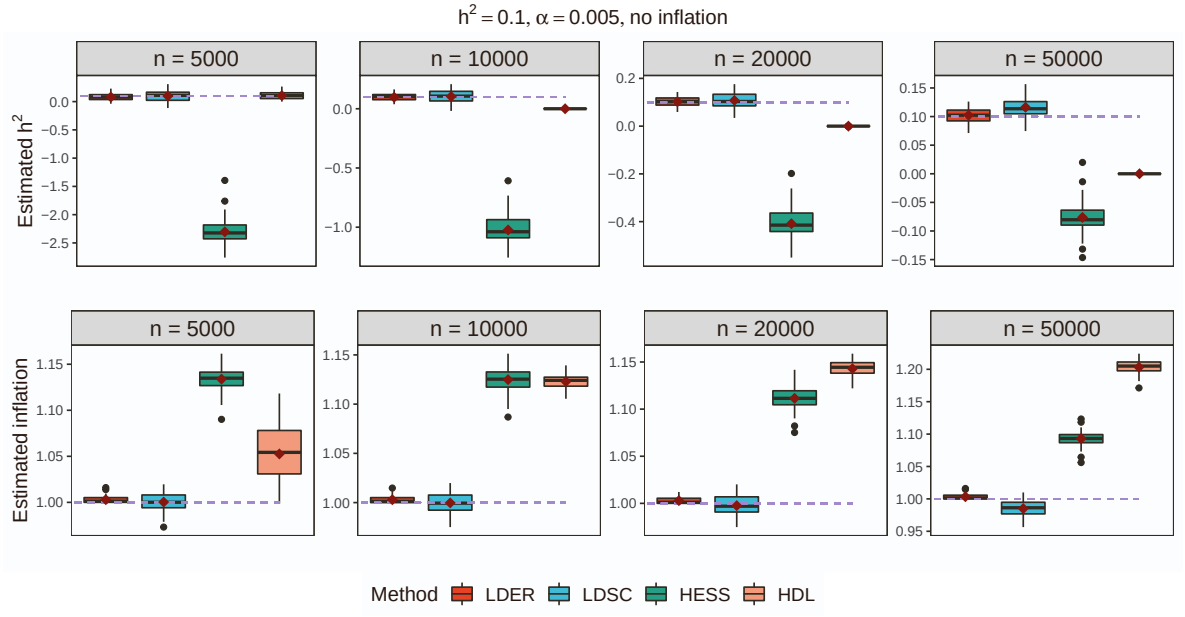
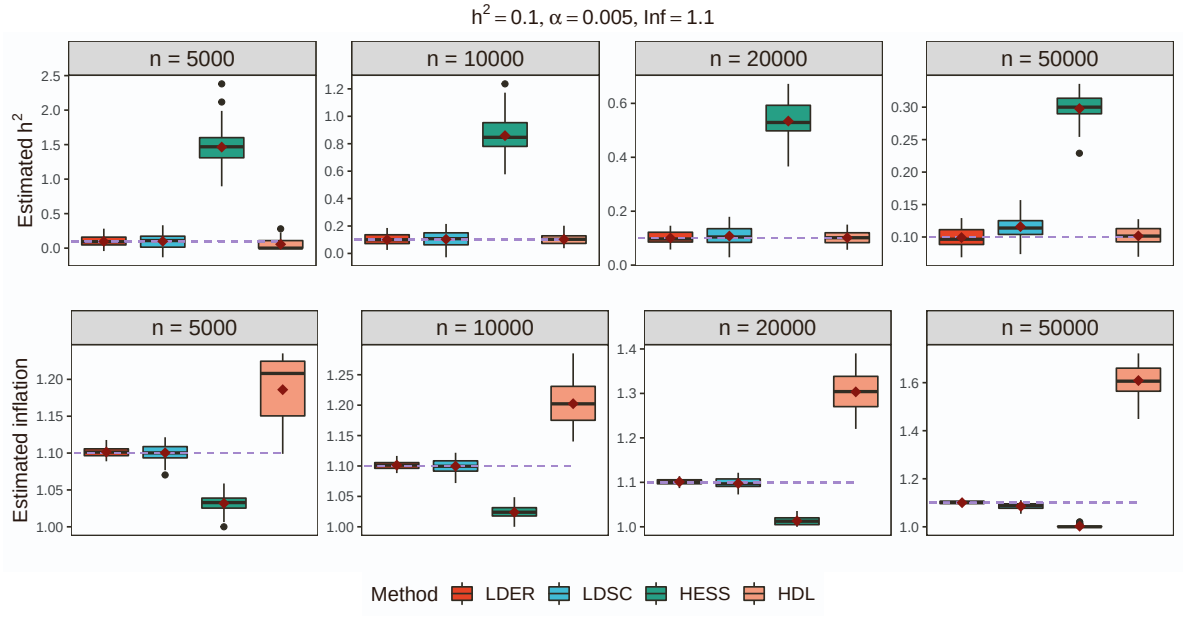


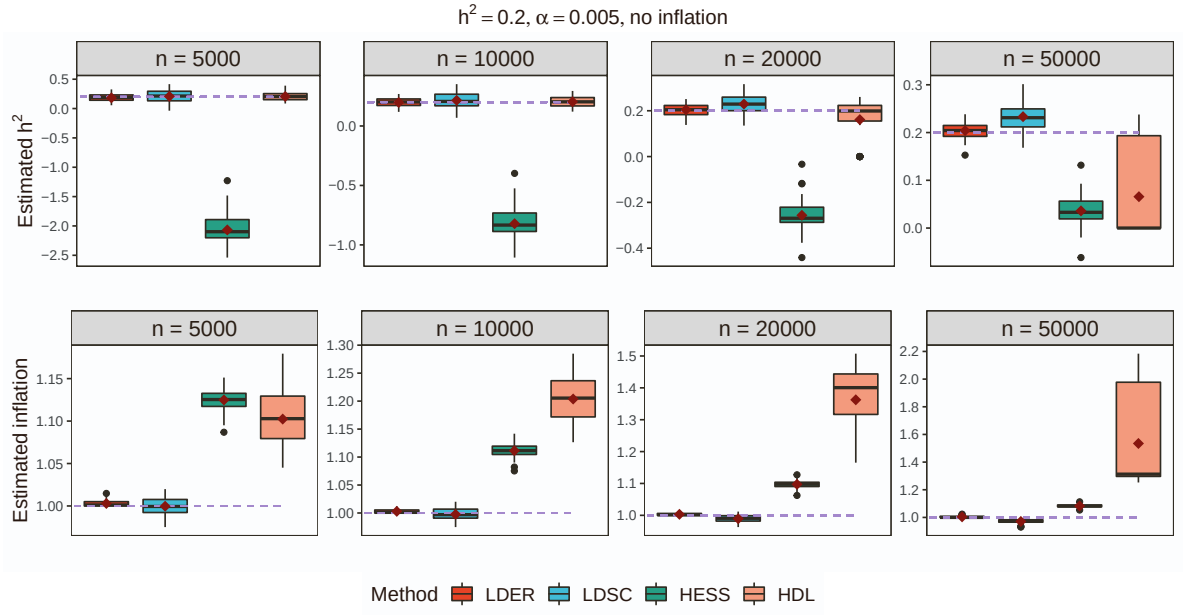
Figure S1: **Estimated heritability with different regression weights with LDER based on simulated GWAS summary statistics.** The sample sizes varied from $5,000$ to $50,000$. The number of SNPs was fixed at $100,000$. The proportion of causal SNPs was 5%. The effect sizes were sampled from a spike-and-slab distribution with heritability 0.5. The simulations were repeated for 50 times. Dashed lines represent the true value. Diamonds indicate means in boxplots. The colors of the boxes differentiate the estimation methods.

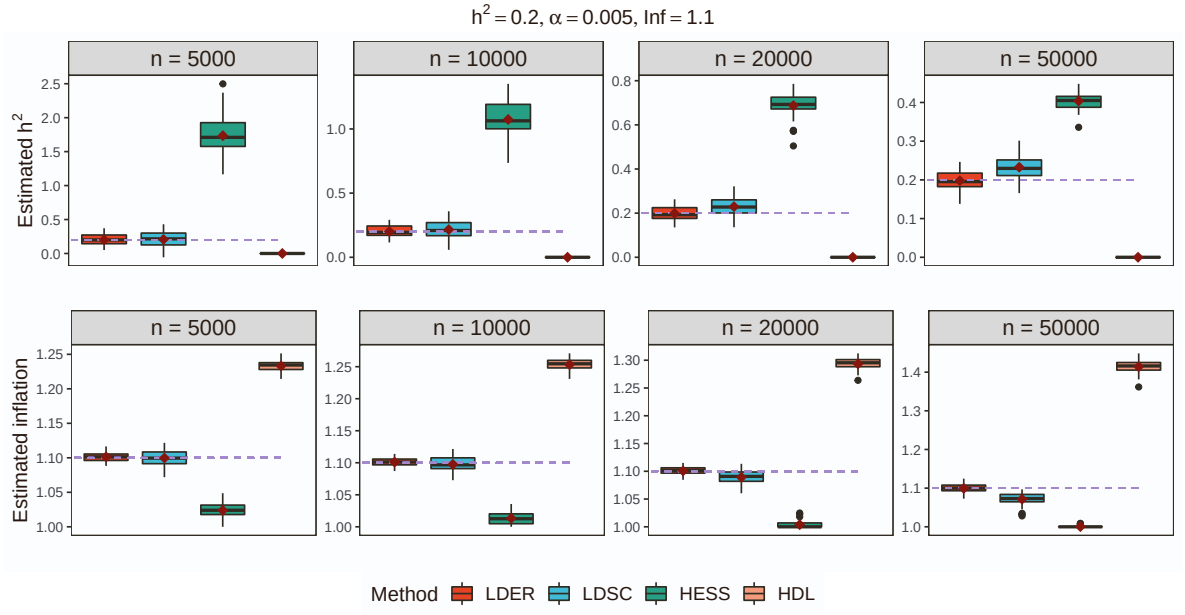

$h^2 = 0.05, \alpha = 0.005$, no inflation
n = 5000
n = 10000
n = 20000
n = 50000
Estimated h^2
Estimated inflation
Method LDER LDSC HESS HDL

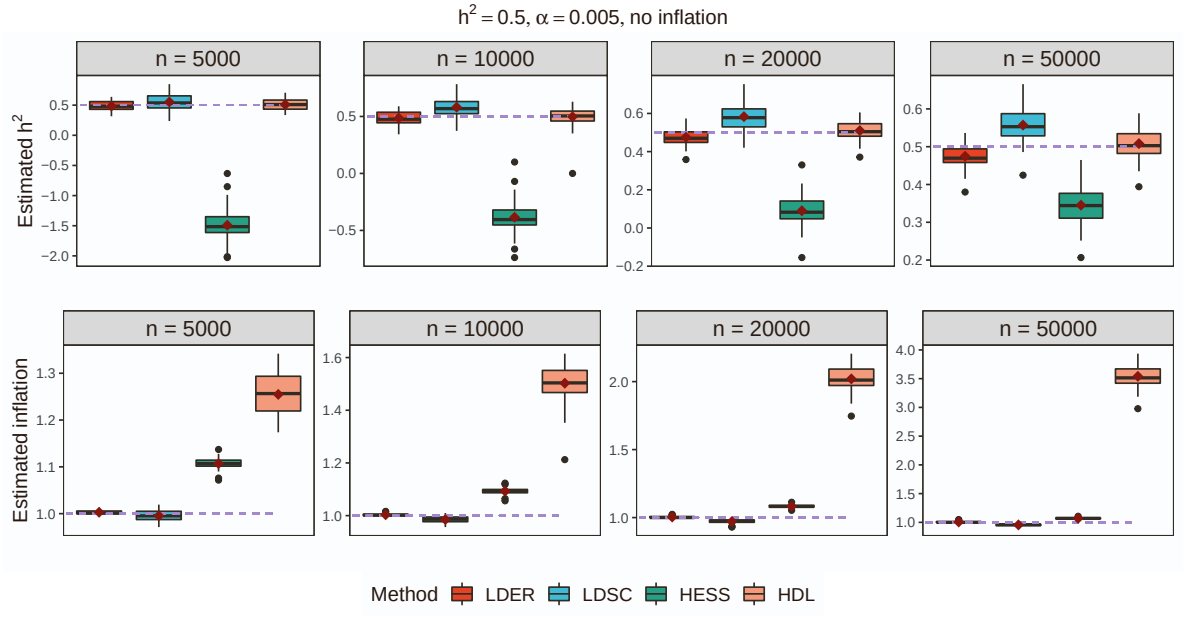

$h^2 = 0.05, \alpha = 0.005$, Inf = 1.1
n = 5000
n = 10000
n = 20000
n = 50000
Estimated h^2
Estimated inflation
Method LDER LDSC HESS HDL

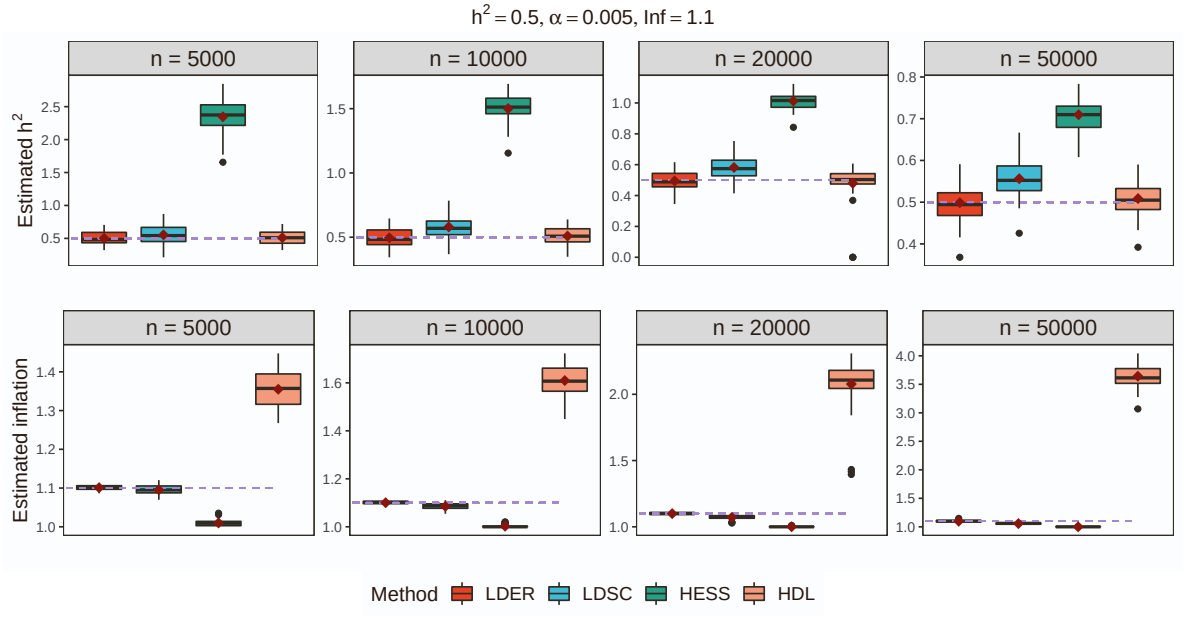

$h^2 = 0.1, \alpha = 0.005$, no inflation
n = 5000
n = 10000
n = 20000
n = 50000
Estimated h^2
Estimated inflation
Method LDER LDSC HESS HDL

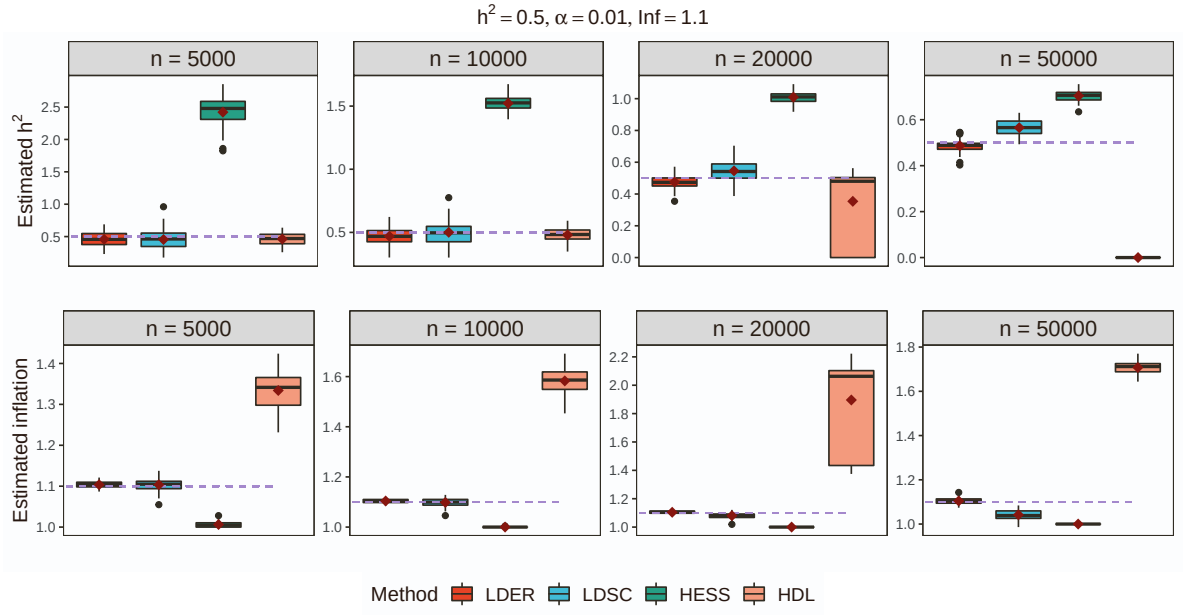

$h^2 = 0.1, \alpha = 0.005$, Inf = 1.1
n = 5000
n = 10000
n = 20000
n = 50000
Estimated h^2
Estimated inflation
Method LDER LDSC HESS HDL

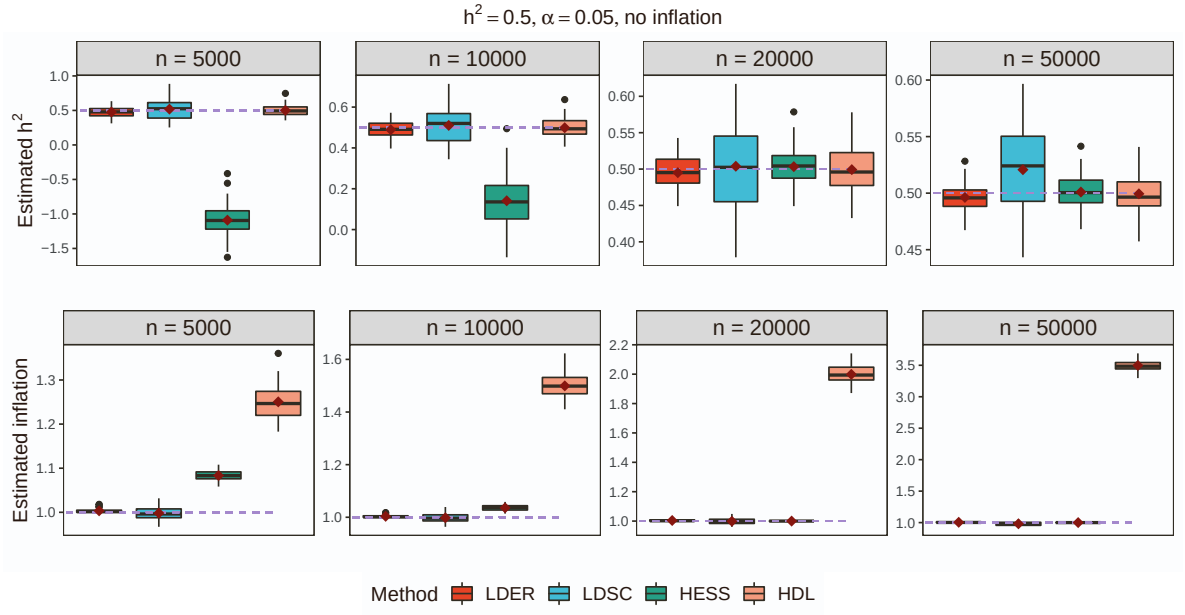

$h^2 = 0.2, \alpha = 0.005$, no inflation
n = 5000
n = 10000
n = 20000
n = 50000
Estimated h^2
Estimated inflation
Method LDER LDSC HESS HDL

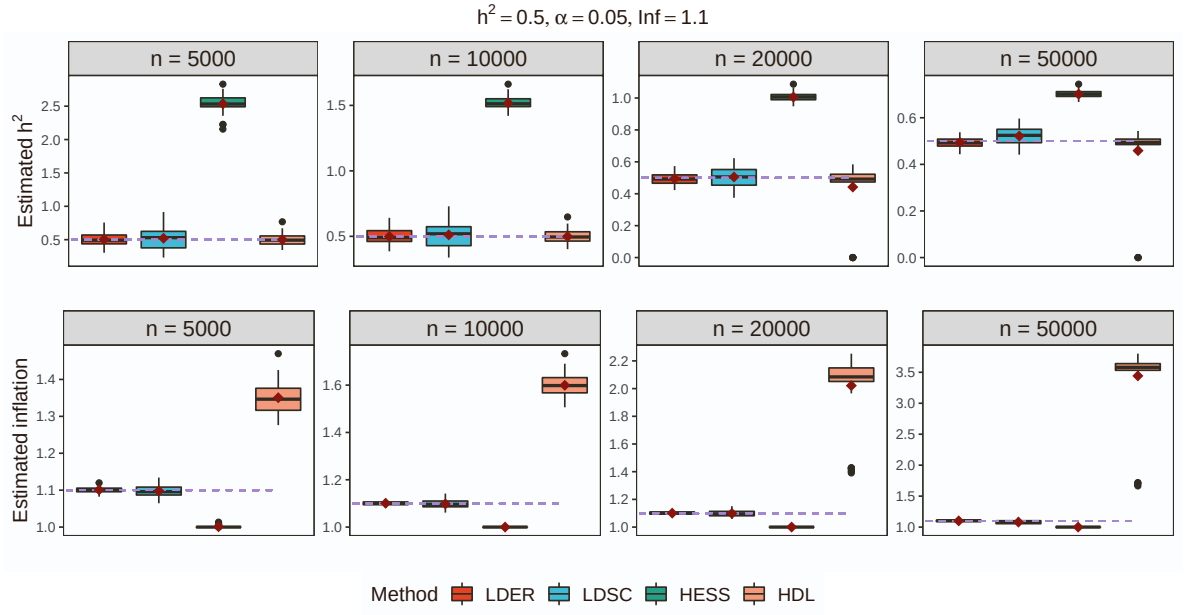

$h^2 = 0.2, \alpha = 0.005$, Inf = 1.1
n = 5000
n = 10000
n = 20000
n = 50000
Estimated h^2
Estimated inflation
Method LDER LDSC HESS HDL

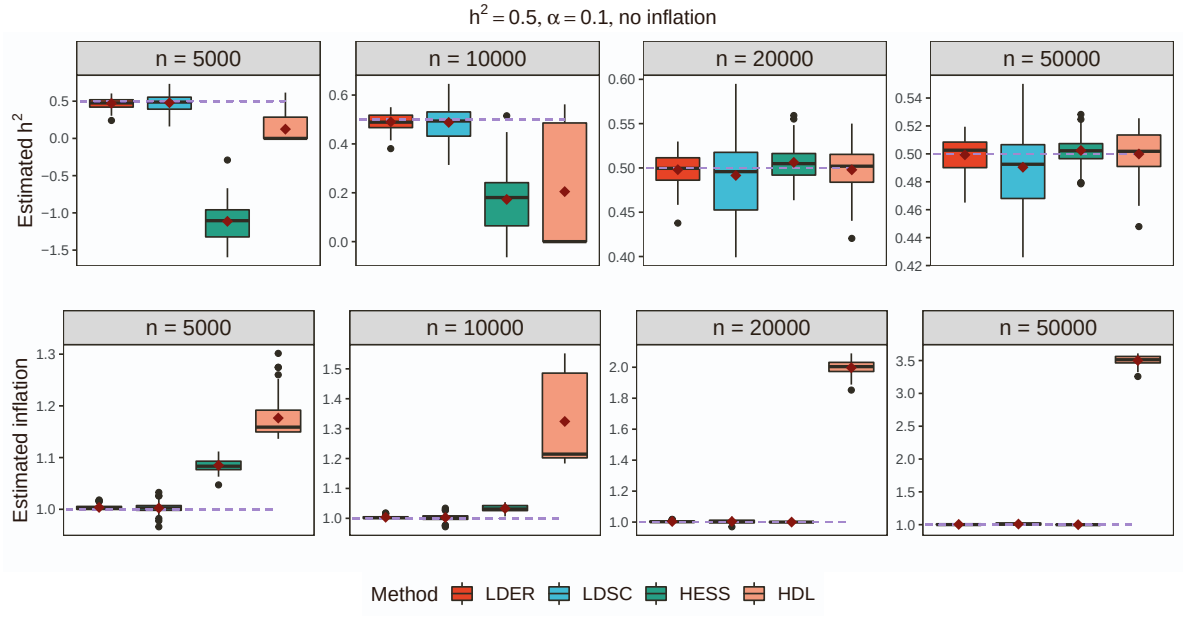

$h^2 = 0.5, \alpha = 0.005$, no inflation
n = 5000
n = 10000
n = 20000
n = 50000
Estimated h^2
Estimated inflation
Method LDER LDSC HESS HDL


$h^2 = 0.5, \alpha = 0.005$, Inf = 1.1
n = 5000
n = 10000
n = 20000
n = 50000
Estimated h^2
Estimated inflation
Method LDER LDSC HESS HDL


$h^2 = 0.5, \alpha = 0.01, \text{Inf} = 1.1$
n = 5000
n = 10000
n = 20000
n = 50000
Estimated h^2
Estimated inflation
Method LDER LDSC HESS HDL


$h^2 = 0.5, \alpha = 0.05, \text{no inflation}$
n = 5000
n = 10000
n = 20000
n = 50000
Estimated h^2
Estimated inflation
Method LDER LDSC HESS HDL


$h^2 = 0.5, \alpha = 0.05$, Inf = 1.1
n = 5000
n = 10000
n = 20000
n = 50000
Estimated h^2
Estimated inflation
Method LDER LDSC HESS HDL


$h^2 = 0.5, \alpha = 0.1$, no inflation
n = 5000
n = 10000
n = 20000
n = 50000
Estimated h^2
Estimated inflation
Method LDER LDSC HESS HDL

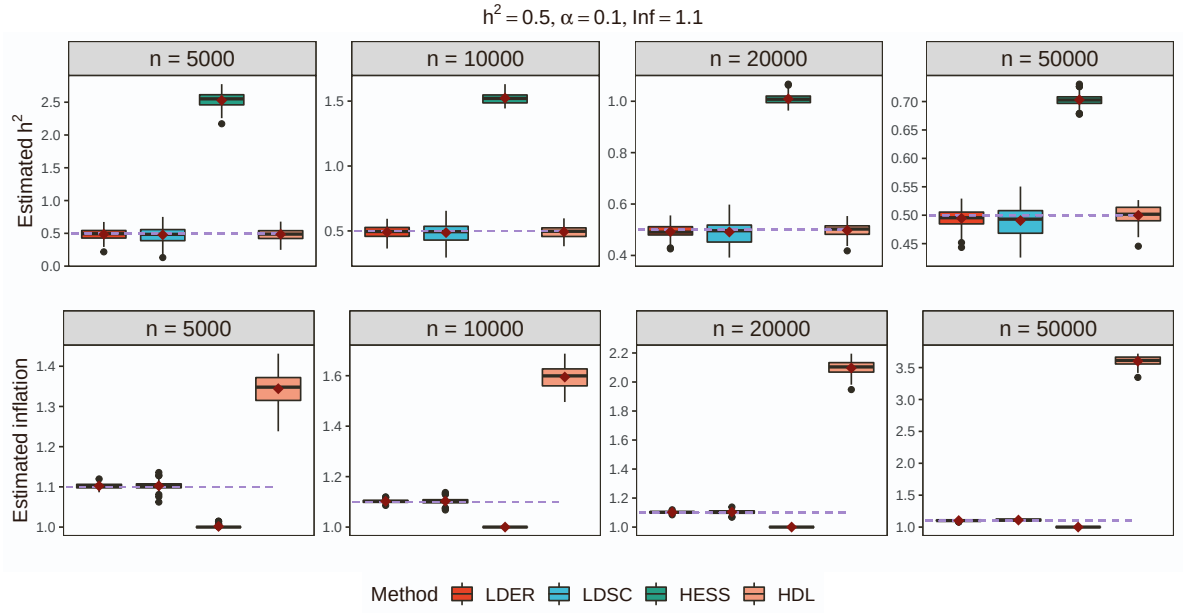


Figure S2: **Comparisons between LDER, LDSC, HESS, and HDL on the estimation of heritability and confounding inflation based on simulated GWAS summary statistics with varying sample sizes.** The number of SNPs was fixed at $100,000$. The proportion of causal SNPs was varied from 0.5% to 10%. The effect sizes were sampled from a spike-and-slab distribution with heritability from 0.05 to 0.5 and with no confounding effects or inflation factor 1.1. The simulations were repeated for 50 times. Dashed lines represent the true value. Diamonds indicate means in boxplots. The colors of the boxes differentiate the estimation methods.

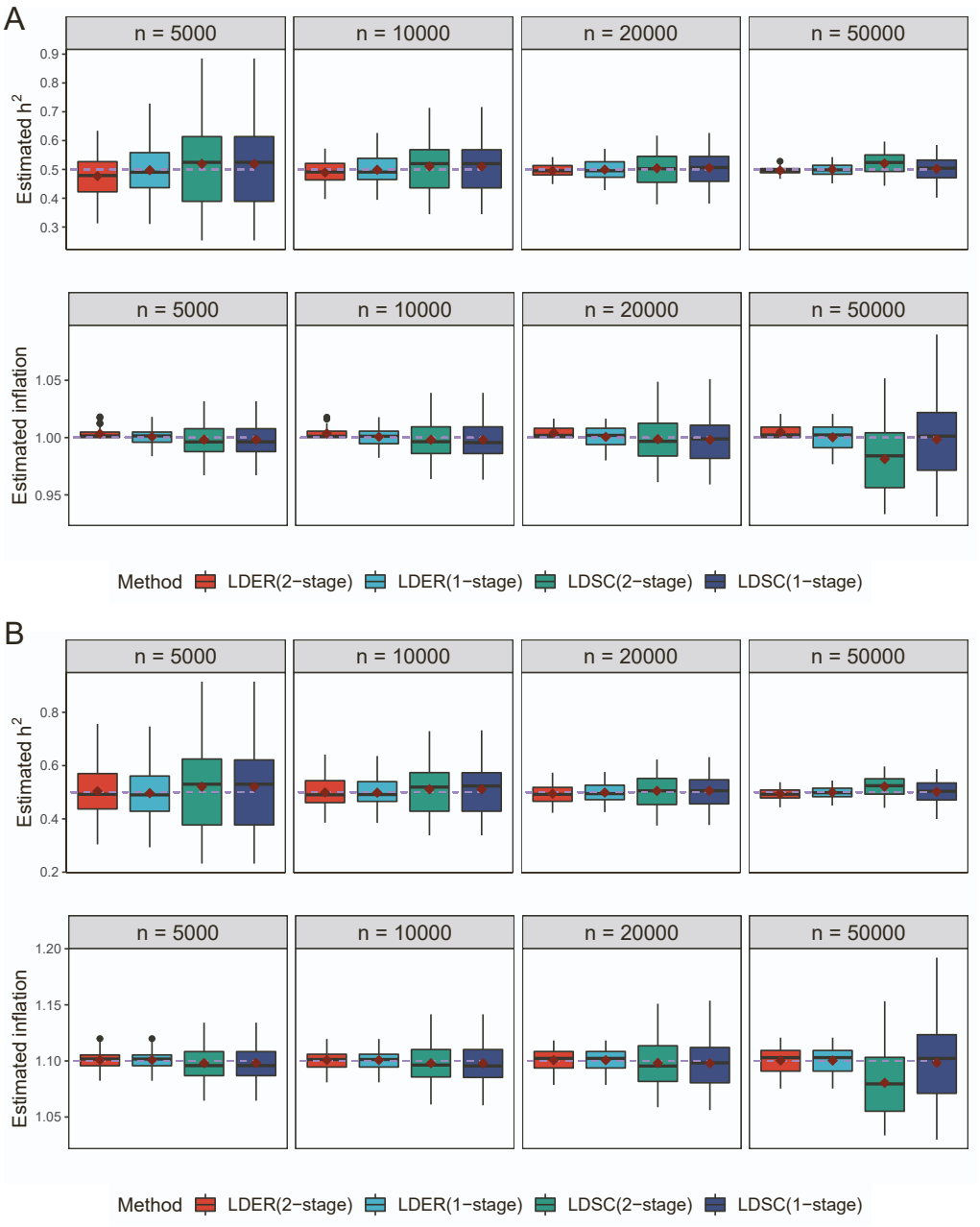


Figure S3: **Comparisons between the one and two-stage procedure of LDER and LDSC on the estimation of heritability and confounding inflation based on simulated GWAS summary statistics.** The sample sizes varied from $5,000$ to $50,000$. The number of SNPs was fixed at $100,000$. The proportion of causal SNPs was 5%. Effect sizes were sampled from a spike-and-slab distribution with heritability 0.5. The simulations were repeated for 50 times. Dashed lines represent the true value. Diamonds indicate means in boxplots. The colors of the boxes differentiate the estimation methods. (A) The confounding inflation factor was set to 1 (with no confounding inflation). (B) The confounding inflation factor was set to 1.1.

2 Supplementary Tables

Table S1: **Precision and accuracy of the heritability estimates with LDER, LDSC, HESS, and HDL.** The simulations were based on UKBB genotypes, and repeated for 50 times. The heritability was fixed at 0.5 and the proportion of causal SNPs was 0.5%. The highest precision and smallest RMSEs are highlighted in boldface.

(a) **The performance of LDER and LDSC with in-sample LD estimated by UKBB European samples.**

n	**Precision (1/SD)**				**RMSE**			
	LDER	LDSC	HESS	HDL	LDER	LDSC	HESS	HDL
5,000	15.74	10.73	**16.92**	NA*	**0.063**	0.089	0.084	0.500
10,000	**30.75**	24.08	25.11	NA	**0.043**	0.058	0.093	0.500
20,000	**43.05**	32.75	28.84	NA	**0.022**	0.034	0.069	0.500
50,000	65.21	48.00	47.23	**163.32**	**0.019**	0.022	0.167	0.462

* NA indicates the estimates are too close to zero, yielding infinite 1/SD.

(b) **The performance of LDER and LDSC with external LD estimated by 1000G European samples.**

n	**Precision (1/SD)**			**RMSE**		
	LDER	LDSC	HESS	LDER	LDSC	HESS
5,000	12.35	10.51	**17.50**	**0.089**	0.091	0.150
10,000	23.60	24.55	**25.85**	0.044	0.062	**0.040**
20,000	**34.79**	31.67	33.81	**0.024**	0.037	0.031
50,000	**57.84**	46.16	47.76	**0.021**	0.026	0.113

Table S2: **Precision and accuracy of the heritability estimates with one- and two-stage of LDER and LDSC.** The simulations were based on UKBB genotypes, and repeated for 50 times. The heritability was fixed at 0.5 and the proportion of causal SNPs was 1%. The highest precision and smallest RMSEs are highlighted in boldface.

(a) **The performance of LDER and LDSC with in-sample LD estimated by UKBB European samples.**

n	**Precision (1/SD)**				**RMSE**			
	LDER		LDSC		LDER		LDSC	
	2-stage	1-stage	2-stage	1-stage	2-stage	1-stage	2-stage	1-stage
5,000	**17.82**	16.82	12.51	12.43	**0.073**	0.074	0.095	0.082
10,000	**23.53**	21.31	17.82	17.47	**0.034**	0.050	0.053	0.057
20,000	**40.18**	37.38	31.14	29.69	**0.030**	0.033	0.041	0.037
50,000	73.79	**75.49**	53.77	41.53	**0.016**	0.027	0.022	0.030

(b) **The performance of LDER and LDSC with external LD estimated by 1000G European samples.**

n	**Precision (1/SD)**				**RMSE**			
	LDER		LDSC		LDER		LDSC	
	2-stage	1-stage	2-stage	1-stage	2-stage	1-stage	2-stage	1-stage
5,000	**14.07**	14.06	12.28	12.22	0.077	**0.074**	0.102	0.085
10,000	**20.05**	19.63	17.32	17.02	**0.047**	0.051	0.056	0.059
20,000	**33.60**	30.87	27.99	29.12	**0.032**	0.035	0.045	0.041
50,000	**64.16**	55.15	53.84	40.50	**0.017**	0.039	0.025	0.035

Table S3: **Simulations with both population stratification and polygenicity.** The $\overline{\chi^2}$ in even chromosomes is treated as the true confounding inflation factors. The SDs derived from 50 repeated simulations are shown in brackets.

n	Null $\overline{\chi^2}$	Inf_{LDER}	Inf_{LDSC}	Null $\overline{\chi^2}$ $/Inf_{LDER}$	Null $\overline{\chi^2}$ $/Inf_{LDSC}$
5,000	1.12 (0.00)	1.12 (0.00)	1.08 (0.00)	0.99 (0.00)	1.03 (0.00)
10,000	1.16 (0.01)	1.16 (0.01)	1.14 (0.00)	1.00 (0.00)	1.02 (0.00)
20,000	1.18 (0.01)	1.18 (0.01)	1.18 (0.01)	1.00 (0.01)	1.00 (0.01)
50,000	1.18 (0.03)	1.17 (0.02)	1.19 (0.03)	1.00 (0.02)	0.99 (0.01)

Table S4: **Comparisons of the estimated heritability on ten UKBB traits by LDER, LDSC, HESS, and HDL.** The heritability estimated by BOLT-LMM was regarded as the true value. The SEs were estimated by a delete-block jackknife procedure for LDER. The precision is defined as the reciprocal of the estimated SE reported by each method.

Method	LDER	LDSC	HESS	HDL
RMSE	**0.049**	0.057	0.580	0.200
Mean precision	86.12	38.53	**227.6**	NA*

* NA indicates the software failed to provide an estimate of the standard error in the estimated heritability.

Table S5 (.xlsx): **The 97 UKBB traits with significantly different heritability estimates (after Bonferroni correction) between LDER and LDSC.** The SEs in brackets were estimated with block-jackknife. The heritability estimates of binary traits has been transformed to liability scale.

Table S6 (.xlsx): **A numerical comparison between the estimates of heritability and confounding inflation by LDER and LDSC on 221 UKBB quantitative traits.** The SEs in brackets were estimated with block-jackknife.

Table S7 (.xlsx): **A numerical comparison between the estimates of heritability and confounding inflation by LDER and LDSC on 593 UKBB dichotomous traits.** The estimated heritability have been transformed to liability scale. The SEs in brackets were estimated with block-jackknife.

Table S8: **Runtime (minutes) comparison for LDER, LDSC, HESS, and HDL.** The LD information is calculated with $10,000$ UKBB individuals with $404,892$ autosomal variants. The time is the average based on 10 repeats of the simulation.

	LDER	LDSC	HESS	HDL
LD preparation	6	243	7	60
Estimation	1.8	0.4	0.2	3.2
Total (k traits)	$6 + 1.8k$	$243 + 0.4k$	$7.2k$	$60 + 3.2k$

3 Supplementary Methods

3.1 The two-stage procedure

In parallel to LDSC, we use a two-stage procedure to reduce the variance of the estimates. In the first stage, we estimate the regression intercept, and constrain the intercept to be no smaller than 1. In the second stage, we fix the intercept to the value derived from the first stage and estimate the heritability using all SNPs. In real data applications, we employ a linear shrinkage method for the LD matrix estimation, and set the eigenvalues smaller than $1e-06$ to 0. We notice that there were many projected z-values ($\tilde{z}_i^2$) with extremely small magnitudes, which were generated mainly due to the eigen-decomposition of inaccurately estimated LD matrices, and contributed to the downward bias of confounding inflations and further upward bias in heritability. Therefore, we estimate the LDER regression intercept with SNPs of $\tilde{z}_i^2$ smaller than 0.005 removed in the first stage of the estimation.

3.2 Derivation of regression weights

In order to derive the variance of $\tilde{z}_i^2$, we assume that n is large and $\boldsymbol{\beta} \sim N\left(0, \frac{h_g^2}{m}\mathbf{I}\right)$, and the i.i.d. random error term $\boldsymbol{e} \sim N\left(0, \left(1-h_g^2\right)\mathbf{I}\right)$. Then we have $\tilde{\boldsymbol{z}} \mid \boldsymbol{\beta} \sim N\left(\sqrt{n}\mathbf{D}^{\frac{1}{2}}\mathbf{U}^{\mathrm{T}}\boldsymbol{\beta}, \mathbf{I}\right)$. The expectation and covariance matrix of $\tilde{\boldsymbol{z}}$ can be calculated with $\mathbb{E}(\tilde{\boldsymbol{z}}) = \mathbb{E}\left[\mathbb{E}(\tilde{\boldsymbol{z}} \mid \boldsymbol{\beta})\right] = 0$, and $\mathrm{Cov}(\tilde{\boldsymbol{z}}) = \mathbb{E}[\mathrm{Cov}(\tilde{\boldsymbol{z}} \mid \boldsymbol{\beta})] + \mathrm{Cov}[\mathbb{E}(\tilde{\boldsymbol{z}} \mid \boldsymbol{\beta})] = \mathbf{I} + \frac{nh_g^2}{m}\mathbf{D}$. Thus, $\tilde{z}_i \sim N\left(0, 1 + nh_g^2 D_{ii}/m\right)$, where D_{ii} is the i-th eigenvalue of the LD matrix. The $\tilde{z}_i^2$ follows a scaled χ^2 distribution with scale factor $1 + nh_g^2 D_{ii}/m$, and the variance function is

$$\mathrm{Var}\left(\tilde{z}_i^2\right) = 2\left(1 + \frac{nh_g^2}{m}D_{ii}\right)^2. \tag{1}$$

Similarly, when the confounding inflation exists, we have

$$\mathrm{Var}\left(\tilde{z}_i^2\right) = 2\left(\lambda + \frac{nh_g^2}{m}D_{ii}\right)^2. \tag{2}$$

3.3 Variance of the estimated heritability and inflation factor

We now show that theoretically, the estimation of inflation will have larger variance as sample size goes larger for LDER and LDSC. The intuition behind is that the magnitude of the z-scores of the GWAS, which determines the magnitude of the dependent variable in the regression of LDER, increases with sample size. Please note that the "sample size" of the regression in LDER is determined by the number of SNPs, rather than the sample size of the GWAS. Moreover, we provide theoretical justification for estimates of least square regression on Equation $\mathbb{E}\left(\tilde{z}_i^2\right) = \lambda + \frac{nh_g^2}{m}D_{ii}$, and the same holds for LDSC. By regressing $\tilde{z}_i^2$ on D_{ii}, we have

$$\mathrm{Var}\left(\widehat{\frac{nh_g^2}{m}}\right) = \mathrm{Var}\left(\frac{\sum_{i=1}^m \left(D_{ii}-\bar{D}\right)\tilde{z}_i^2}{S_{dd}}\right) = \left(\frac{1}{S_{dd}}\right)^2 \sum_{i=1}^m \left[\left(D_{ii}-\bar{D}\right)^2 \mathrm{Var}\left(\tilde{z}_i^2\right)\right], \tag{3}$$

where $S_{dd} = \sum_{i=1}^m \left(D_{ii}-\bar{D}\right)^2$, and $\bar{D} = \frac{1}{m}\sum_{i=1}^m D_{ii} = \bar{l}$, where $\bar{l}$ is the average LD score, i.e., $\bar{l} = \sum_{j=1}^m l_j/m$. The second equation follows the independence among $\tilde{z}_i^2$s. Combining with Eqn (2),

we have

$$\operatorname{Var}\left(\widehat{h_g^2}\right) = 2\left(\frac{1}{S_{dd}}\right)^2 \sum_{i=1}^{m}\left[\left(D_{ii}-\bar{D}\right)\left(\frac{m}{n}\lambda + h_g^2 D_{ii}\right)\right]^2. \tag{4}$$

In addition,

$$\begin{aligned}\operatorname{Var}(\hat{\lambda}) &= \operatorname{Var}\left(\overline{\widetilde{\mathbf{z}}^2} - \frac{\widehat{nh_g^2}}{m}\bar{D}\right) = \operatorname{Var}\left(\overline{\widetilde{\mathbf{z}}^2}\right) + \bar{D}^2 \operatorname{Var}\left(\frac{\widehat{nh_g^2}}{m}\right) - 2\operatorname{Cov}\left(\overline{\widetilde{\mathbf{z}}^2}, \frac{\widehat{nh_g^2}}{m}\bar{D}\right) \\ &= \sum_{i=1}^{m}\left\{\left[\frac{1}{m^2} + \bar{D}^2\left(\frac{1}{S_{dd}}\right)^2\left(D_{ii}-\bar{D}\right)^2\right]\operatorname{Var}\left(\tilde{z}_i^2\right)\right\} - 2\operatorname{Cov}\left(\overline{\widetilde{\mathbf{z}}^2}, \frac{\widehat{nh_g^2}}{m}\bar{D}\right).\end{aligned} \tag{5}$$

We consider the covariance term:

$$\begin{aligned}\operatorname{Cov}\left(\overline{\widetilde{\mathbf{z}}^2}, \frac{\widehat{nh_g^2}}{m}\bar{D}\right) &= \operatorname{Cov}\left(\sum_{i=1}^{m}\frac{1}{m}\tilde{z}_i^2, \bar{D}\sum_{j=1}^{m}\frac{\left(D_{jj}-\bar{D}\right)\tilde{z}_j^2}{S_{dd}}\right) = \sum_{i=1}^{m}\sum_{j=1}^{m}\bar{D}\frac{\left(D_{jj}-\bar{D}\right)}{mS_{dd}}\operatorname{Cov}\left(\tilde{z}_i^2, \tilde{z}_j^2\right) \\ &= \sum_{i=1}^{m}\bar{D}\frac{\left(D_{ii}-\bar{D}\right)}{mS_{dd}}\operatorname{Var}\left(\tilde{z}_i^2\right),\end{aligned} \tag{6}$$

where the last equation follows the independence among $\tilde{z}_i$s. Thus, we have

$$\begin{aligned}\operatorname{Var}(\hat{\lambda}) &= \sum_{i=1}^{m}\left\{\left[\frac{1}{m^2} + \bar{D}^2\left(\frac{1}{S_{dd}}\right)^2\left(D_{ii}-\bar{D}\right)^2 - 2\bar{D}\frac{\left(D_{ii}-\bar{D}\right)}{mS_{dd}}\right]\operatorname{Var}\left(\tilde{z}_i^2\right)\right\} \\ &= \sum_{i=1}^{m}\left[\left(\frac{1}{m} - \bar{D}\frac{\left(D_{ii}-\bar{D}\right)}{S_{dd}}\right)^2\operatorname{Var}\left(\tilde{z}_i^2\right)\right] \\ &= 2\sum_{i=1}^{m}\left[\left(\frac{1}{m} - \bar{D}\frac{\left(D_{ii}-\bar{D}\right)}{S_{dd}}\right)^2\left(\lambda + \frac{nh_g^2}{m}D_{ii}\right)^2\right],\end{aligned} \tag{7}$$

which is an increasing function of n.