Effective sample sizes for different GWAS sampling schemes

Alexander Strudwick Young

1 Setup

We consider a fixed budget with which we can genotype and phenotype $n = 6k, k \in \mathbb{N}$, individuals. We consider three different options:

1. n unrelated individuals
2. n/2 independent sibling pairs
3. n/3 independent parent-offspring trios

We consider an infinite, outbred, random-mating population, where different individuals are independent in the unrelated case (1), and different families are independent in scenarios (2) and (3). We consider a phenotype with infinitesimal genetic architecture and with heritability h^2 and no environmental effects.

Let $\mathbf{g}_i$ be the mean-normalized genotype vector for family i :

$$\mathbf{g}_i = \begin{bmatrix} g_{i1} \\ g_{i2} \\ g_{p(i)} \\ g_{m(i)} \end{bmatrix}$$

where g_{ij} is the genotype of sibling j in family i, and $g_{p(i)}$ and $g_{m(i)}$ are, respectively, the genotypes of the father and mother in family i. Assumming that the allele frequency is f, we have

$$\mathrm{Var}\left(\mathbf{g}_i\right) = f(1-f) \begin{bmatrix} 2 & 1 & 1 & 1 \\ 1 & 2 & 1 & 1 \\ 1 & 1 & 2 & 0 \\ 1 & 1 & 0 & 2 \end{bmatrix}$$

Let $\mathbf{y}_i$ be the phenotype vector for family i in the same order as the genotype vector, and let the phenotype have variance one. Given the above assumptions,

$$\mathrm{Var}\left(\mathbf{y}_i \mid \mathbf{g}_i\right) = \begin{bmatrix} 1 & h^2/2 & h^2/2 & h^2/2 \\ h^2/2 & 1 & h^2/2 & h^2/2 \\ h^2/2 & h^2/2 & 1 & 0 \\ h^2/2 & h^2/2 & 0 & 1 \end{bmatrix}$$

2 Sampling variances

We compute the variance of generalized least-squares (GLS) estimators for the 3 different scenarios.

2.1 Unrelated sample

For n unrelated samples, the estimation reduces to ordinary least-squares (OLS). The sample variance from performing ordinary least-squares (OLS) on n unrelated individuals is

$$\mathrm{Var}\left(\hat{\beta}_{\mathrm{unrel}}\right) = \frac{1}{2nf(1-f)}.$$

2.2 Sibling sample

We consider the sampling variance from the generalized least-squares estimator applied to siblings pairs from $n/2$ independent families. Let

$$\mathbf{g}_{\mathrm{sib},i} = \begin{bmatrix} g_{i1} \\ g_{i2} \end{bmatrix}$$

and let $\mathbf{y}_{\mathrm{sib}\ ,i}$ be the corresponding phenotype vector. Then the generalized least-squares estimator is

$$\hat{\beta}_{\mathrm{sib}} = \left(\sum_{i=1}^{n/2} \mathbf{g}_{\mathrm{sib},i}^T \mathrm{Var}\left(\mathbf{y}_{\mathrm{sib},i} \mid \mathbf{g}_{\mathrm{sib},i}\right)^{-1} \mathbf{g}_{\mathrm{sib},i} \right)^{-1} \sum_{i=1}^{n/2} \mathbf{g}_{\mathrm{sib},i}^T \mathrm{Var}\left(\mathbf{y}_{\mathrm{sib},i} \mid \mathbf{g}_{\mathrm{sib},i}\right)^{-1} \mathbf{y}_{\mathrm{sib},i},$$

and its sampling variance is

$$\mathrm{Var}\left(\hat{\beta}_{\mathrm{sib}}\right) = \left(\sum_{i=1}^{n/2} \mathbf{g}_{\mathrm{sib},i}^T \mathrm{Var}\left(\mathbf{y}_{\mathrm{sib},i} \mid \mathbf{g}_{\mathrm{sib},i}\right)^{-1} \mathbf{g}_{\mathrm{sib},i} \right)^{-1}.$$

We now compute the large-sample variance. First, we compute

$$\mathrm{Var}\left(\mathbf{y}_{\mathrm{sib},i} \mid \mathbf{g}_{\mathrm{sib},i}\right)^{-1} = \begin{bmatrix} 1 & h^2/2 \\ h^2/2 & 1 \end{bmatrix}^{-1} = \frac{1}{1-(h^2/2)^2} \begin{bmatrix} 1 & -h^2/2 \\ -h^2/2 & 1 \end{bmatrix}.$$

Therefore, for large n,

$$\sum_{i=1}^{n/2} \mathbf{g}_{\text{sib},i}^T \operatorname{Var}\left(\mathbf{y}_{\text{sib},i} \mid \mathbf{g}_{\text{sib},i}\right)^{-1} \mathbf{g}_{\text{sib},i} \approx \frac{nf(1-f)\left(2-h^2/2\right)}{1-\left(h^2/2\right)^2}.$$

We thereby obtain the approximate large-sample variance from $n/2$ sibling pairs:

$$\operatorname{Var}\left(\hat{\beta}_{\text{sib}}\right) \approx \frac{1-\left(h^2/2\right)^2}{\left(2-h^2/2\right)nf(1-f)}.$$

We can thus compute the relative effective sample size compared to n unrelated individuals:

$$\frac{\operatorname{Var}\left(\hat{\beta}_{\text{sib}}\right)}{\operatorname{Var}\left(\hat{\beta}_{\text{unrel}}\right)} \approx \frac{2\left(1-\left(h^2/2\right)^2\right)}{\left(2-h^2/2\right)} = 1 + \frac{h^2\left(1-h^2\right)}{4-h^2}$$

2.3 Trio sample

Following the same procedure as above, we have that

$$\operatorname{Var}\left(\hat{\beta}_{\text{trio}}\right) = \left(\sum_{i=1}^{n/3} \mathbf{g}_{\text{trio},i}^T \operatorname{Var}\left(\mathbf{y}_{\text{trio},i} \mid \mathbf{g}_{\text{trio},i}\right)^{-1} \mathbf{g}_{\text{trio},i}\right)^{-1},$$

where

$$\mathbf{g}_{\text{trio},i} = \begin{bmatrix} g_{i1} \\ g_{p(i)} \\ g_{m(i)} \end{bmatrix},$$

and

$$\operatorname{Var}\left(\mathbf{y}_{\text{trio},i} \mid \mathbf{g}_{\text{trio},i}\right)^{-1} = \begin{bmatrix} 1 & h^2/2 & h^2/2 \\ h^2/2 & 1 & 0 \\ h^2/2 & 0 & 1 \end{bmatrix}^{-1} \tag{1}$$

$$= \frac{1}{1-2\left(h^2/2\right)^2} \begin{bmatrix} 1 & -h^2/2 & -h^2/2 \\ -h^2/2 & 1-\left(h^2/2\right)^2 & \left(h^2/2\right)^2 \\ -h^2/2 & \left(h^2/2\right)^2 & 1-\left(h^2/2\right)^2 \end{bmatrix}. \tag{2}$$

Therefore, for large n,

$$\sum_{i=1}^{n/3} \mathbf{g}_{\text{trio},i}^T \operatorname{Var}\left(\mathbf{y}_{\text{trio},i} \mid \mathbf{g}_{\text{trio},i}\right)^{-1} \mathbf{g}_{\text{trio},i} \approx \frac{2nf(1-f)\left[1-h^2\left(1+h^2/2\right)/3\right]}{1-2\left(h^2/2\right)^2}.$$

We thereby obtain the approximate large-sample sampling variance from $n/3$ trios:

$$\mathrm{Var}\left(\hat{\beta}_{\mathrm{trio}}\right) \approx \frac{1-2\left(h^2/2\right)^2}{2nf(1-f)\left[1-h^2\left(1+h^2/2\right)/3\right]}$$

We can thus compute the relative effective sample size compared to n unrelated individuals:

$$\frac{\mathrm{Var}\left(\hat{\beta}_{\mathrm{trio}}\right)}{\mathrm{Var}\left(\hat{\beta}_{\mathrm{unrel}}\right)} \approx \frac{1-2\left(h^2/2\right)^2}{1-h^2\left(1+h^2/2\right)/3} = 1+\frac{h^2\left(1-h^2\right)}{3-h^2\left(1+h^2/2\right)}$$

2.4 Direct effect estimates

Using the results from Young et al. 2022[1], we have that the variance of the direct effect estimate from genetic differences between $n/2$ sibling pairs is

$$\mathrm{Var}(\hat{\delta}_{\mathrm{sib}}) = \frac{2-h^2}{nf(1-f)}. \tag{3}$$

If we use imputation of missing parental genotypes using phased data as outlined in Young et al.[1], then the variance is

$$\mathrm{Var}(\hat{\delta}_{\mathrm{sibimp}}) = \frac{3(1-(h^2/2)^2)}{(2+h^2/2)nf(1-f)}. \tag{4}$$

If we use $n/3$ unrelated trios to estimate the direct effect in a regression controlling for parental genotypes, we obtain

$$\mathrm{Var}(\hat{\delta}_{\mathrm{trio}}) = \frac{3}{nf(1-f)}. \tag{5}$$

The effective sample sizes compared to the OLS estimator of the standard GWAS population effect applied to n unrelated samples are therefore:

$$\frac{\mathrm{Var}(\hat{\beta}_{\mathrm{unrel}})}{\mathrm{Var}(\hat{\delta}_{\mathrm{sib}})} = \frac{1}{2(2-h^2)}; \tag{6}$$

$$\frac{\mathrm{Var}(\hat{\beta}_{\mathrm{unrel}})}{\mathrm{Var}(\hat{\delta}_{\mathrm{sibimp}})} = \frac{(2+h^2/2)}{6(1-(h^2/2)^2)}; \tag{7}$$

$$\frac{\mathrm{Var}(\hat{\beta}_{\mathrm{unrel}})}{\mathrm{Var}(\hat{\delta}_{\mathrm{trio}})} = \frac{1}{6}. \tag{8}$$

3 Simulations

To check the derivations for the relative effective sample size of estimating GWAS population effects using the three different scenarios, we simulated mean-normalized genotype data, $\mathbf{g}_i$, on $n = 600,000$ families. We did this by simulating a single variant independently for the father and mother by drawing each parental allele independently from a Bernoulli(0.5) distribution. For each of the two siblings in each family, we generated the sibling genotype by choosing one of the paternal alleles and one of the maternal alleles independently and at random, simulating meiosis.

For $h^2 = 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0$, we computed the sampling variances of the GLS estimators by the following formulae

$$\mathrm{Var}(\hat{\beta}_{\mathrm{unrel}}) = \frac{1}{\sum_{i=1}^{n} g_{i1}^2}; \tag{9}$$

$$\mathrm{Var}\left(\hat{\beta}_{\mathrm{sib}}\right) = \left(\sum_{i=1}^{n/2} \begin{bmatrix} g_{i1} & g_{i2} \end{bmatrix} \begin{bmatrix} 1 & h^2/2 \\ h^2/2 & 1 \end{bmatrix}^{-1} \begin{bmatrix} g_{i1} \\ g_{i2} \end{bmatrix}\right)^{-1}; \tag{10}$$

$$\mathrm{Var}\left(\hat{\beta}_{\mathrm{trio}}\right) = \left(\sum_{i=1}^{n/3} \begin{bmatrix} g_{i1} & g_{p(i)} & g_{m(i)} \end{bmatrix} \begin{bmatrix} 1 & h^2/2 & h^2/2 \\ h^2/2 & 1 & 0 \\ h^2/2 & 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} g_{i1} \\ g_{p(i)} \\ g_{m(i)} \end{bmatrix}\right)^{-1}. \tag{11}$$

We then compared the ratios $\mathrm{Var}(\hat{\beta}_{\mathrm{unrel}})/\mathrm{Var}\left(\hat{\beta}_{\mathrm{sib}}\right)$ and $\mathrm{Var}(\hat{\beta}_{\mathrm{unrel}})/\mathrm{Var}\left(\hat{\beta}_{\mathrm{trio}}\right)$ to the theoretical values derived above.

4 Sib differences with imputed genotypes

For a sibling pair from family i, the model is:

$$Y_{i1} = \delta g_{i1} + \alpha g_{\mathrm{par}(i)} + \epsilon_{i1}; \tag{12}$$

$$Y_{i2} = \delta g_{i2} + \alpha g_{\mathrm{par}(i)} + \epsilon_{i2}. \tag{13}$$

We assume that sibling genotypes are mean-normalized.

The sib-difference regression can proceed by OLS regression of Y_{i1} onto $(g_{i1} - g_{i2})$. This will give an unbiased estimate of δ. We can rewrite the equation in terms of the regressions onto $(g_{i1} - g_{i2})$ and $(g_{i1} + g_{i2})$ (which are uncorrelated with each other):

$$Y_{i1} = \delta(g_{i1} - g_{i2}) + (\delta/2 + 2\alpha/3)(g_{i1} + g_{i2}) + \epsilon'_{i1}, \tag{14}$$

for some ϵ'_{i1} uncorrelated with $(g_{i1} - g_{i2})$ and $(g_{i1} + g_{i2})$. We can then ask what we would obtain if we instead performed regression onto the difference in imputed sibling genotypes:

$$\frac{\mathrm{Cov}(Y_{i1}, \hat{g}_{i1} - \hat{g}_{i2})}{\mathrm{Var}(\hat{g}_{i1} - \hat{g}_{i2})} = \delta \frac{\mathrm{Cov}(g_{i1} - g_{i2}, \hat{g}_{i1} - \hat{g}_{i2})}{\mathrm{Var}(\hat{g}_{i1} - \hat{g}_{i2})} + (\delta/2 + 2\alpha/3) \frac{\mathrm{Cov}(g_{i1} + g_{i2}, \hat{g}_{i1} - \hat{g}_{i2})}{\mathrm{Var}(\hat{g}_{i1} - \hat{g}_{i2})} + \frac{\mathrm{Cov}(\epsilon'_{i1}, \hat{g}_{i1} - \hat{g}_{i2})}{\mathrm{Var}(\hat{g}_{i1} - \hat{g}_{i2})}. \tag{15}$$

We can thus assess the bias that comes from using imputed sibling genotypes by performing two regressions:

1. Slope of regression of $g_{i1} - g_{i2}$ on $\hat{g}_{i1} - \hat{g}_{i2}$, which gives bias coming from the sib-difference component;

2. Slope of regression of $g_{i1} + g_{i2}$ on $\hat{g}_{i1} - \hat{g}_{i2}$, which gives bias coming from the sib-sum component, i.e. the component that the sib-difference should be orthogonal with as it captures parental effects/stratification.

This can be done by obtaining the true sibling genotypes from WGS data and forming $g_{i1} - g_{i2}$ and $g_{i1} + g_{i2}$ and regressing these values onto the difference in sib imputed genotypes, $\hat{g}_{i1} - \hat{g}_{i2}$.

From these slopes, we could express the expected result of sib-difference regression as a linear combination of δ and α.

References

[1] Young, A. I., et al. Mendelian imputation of parental genotypes improves estimates of direct genetic effects. *Nature Genetics*, **54**(6):897–905 2022. ISSN 1546-1718. doi:10.1038/s41588-022-01085-0.