Sib-pair regression GWAS

Yengo L. (included in Robison et al. 2018, bioRxiv)

The model

We implemented a sibling regression analysis that consists in regressing height (adjusted for age and sex) difference between siblings onto minor allele counts difference between siblings. This approach is similar to the QFAM procedure implemented in Plink when families are only composed of sibling pairs (Figure).

We considered the following model

$$y_{ik} = \mu + bx_{ik} + v_k + e_{ik} \tag{1}$$

where y_{ik} is the phenotype of sibling i in family k, x_{ik} is the minor allele count at a particular SNP in sibling i from family k, b is the allele substitution effect at the tested SNP, $v_k \sim \mathcal{N}(0, \sigma_f^2)$ is a random family effect specific to family k, such that σ_f^2 is a between-family variance and $e_{ik} \sim \mathcal{N}(0, \sigma_e^2)$ is the residual term. We assume that the e_{ik}'s are all independent.

We denote n_k as the number of siblings in family k. Therefore the number of siblings pairs in that family is $n_k(n_k - 1)/2$. We also denote

$$\bar{y}_k = \frac{1}{n_k}\sum_{i=1}^{n_k} y_{ik},\ \bar{x}_k = \frac{1}{n_k}\sum_{i=1}^{n_k} x_{ik}, \text{ and } \bar{e}_k = \frac{1}{n_k}\sum_{i=1}^{n_k} e_{ik}. \tag{2}$$

Using Equation (1), we can write that

$$\bar{y}_k = \mu + b\bar{x}_k + v_k + \bar{e}_k. \tag{3}$$

As in QFAM, our inference is based on the difference between y_{ik} and $\bar{y}_k$:

$$y_{ik} - \bar{y}_k \ = \ b\,(x_{ik} - \bar{x}_k) + (e_{ik} - \bar{e}_k). \tag{4}$$

Conditionally on allele counts (hereafter denoted $\mathbf{x}$),

$$\begin{aligned} \text{var}(y_{ik} - \bar{y}_k|\mathbf{x}) &= \text{var}(e_{ik} - \bar{e}_k) = \text{var}(e_{ik}) + \text{var}(\bar{e}_k) - 2\text{cov}(e_{ik}, \bar{e}_k) = \sigma_e^2 + \frac{\sigma_e^2}{n_k} - \frac{2\sigma_e^2}{n_k} \\ &= \sigma_e^2\left(1 - \frac{1}{n_k}\right) \end{aligned} \tag{5}$$

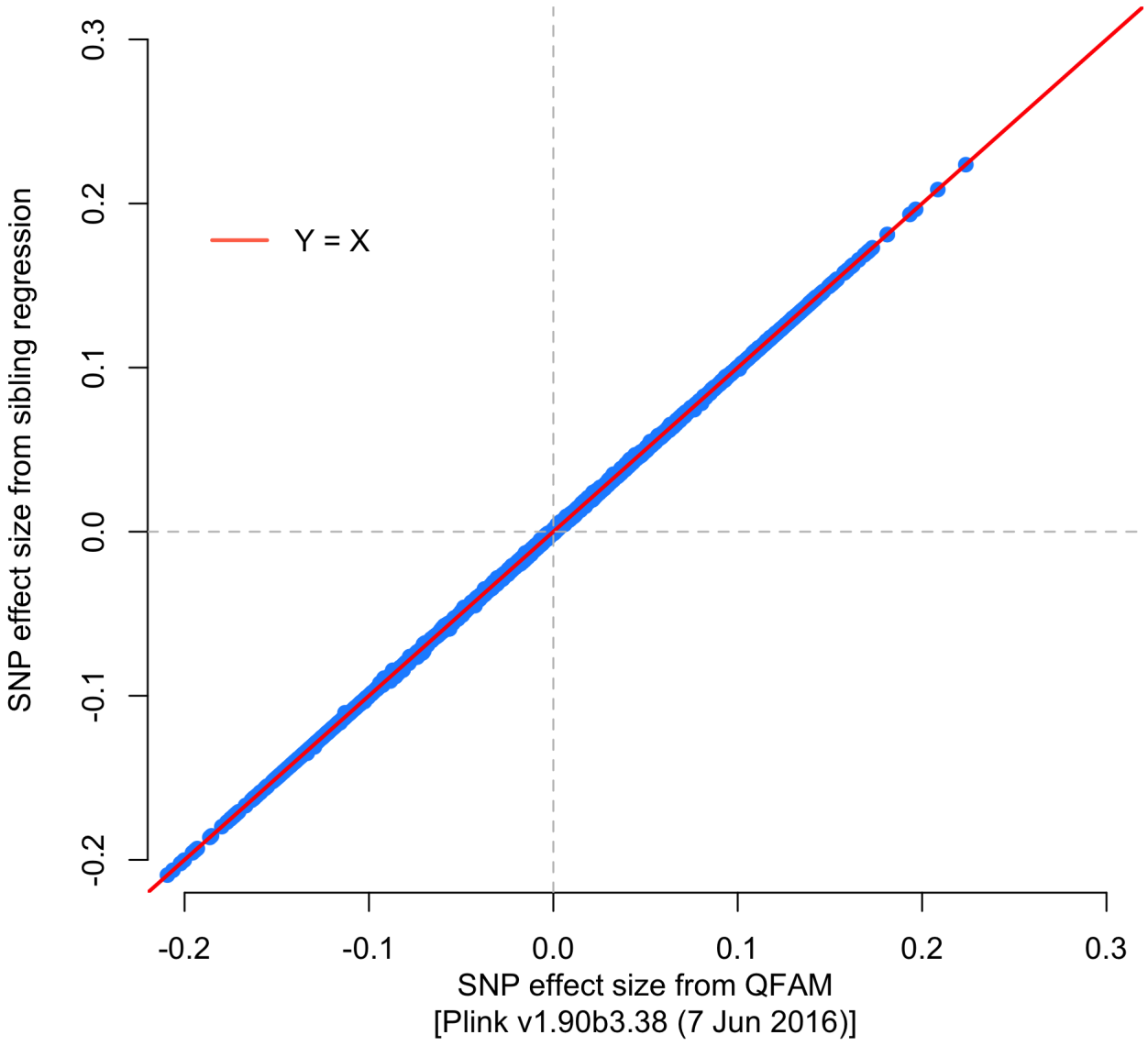

Figure 1: **Comparison of estimated SNP effects from Plink's QFAM procedure vs. sibling regression.** SNP effects are calculated in 18,514 sibling pairs from the Robinson et al. (2015) study. The correlation and regression slope between SNP effects from these two analyses both equal 1.

Similarly, we can show that $\text{cov}\left(y_{ik} - \bar{y}_k, y_{i'k} - \bar{y}_k | \mathbf{x}\right) = -\sigma_e^2/n_k$.

We define $\widehat{b}_k$ as the within-family ordinary least-squares (OLS) estimator of b:

$$\widehat{b}_k = \frac{\sum_{i=1}^{n_k}\left(y_{ik} - \bar{y}_k\right)\left(x_{ik} - \bar{x}_k\right)}{\sum_{i=1}^{n_k}\left(x_{ik} - \bar{x}_k\right)^2} \tag{6}$$

From equation (4), we can easily show that $\mathbb{E}[\widehat{b}_k|\mathbf{x}] = b$. Also,

$$\begin{aligned}
\text{var}(\widehat{b}_k|\mathbf{x}) &= \sigma_e^2\left(1 - \frac{1}{n_k}\right) \times \frac{\sum_{i=1}^{n_k}\left(x_{ik} - \bar{x}_k\right)^2}{\left[\sum_{i=1}^{n_k}\left(x_{ik} - \bar{x}_k\right)^2\right]^2} - \left(\sigma_e^2/n_k\right) \times \frac{\sum_{i=1}^{n_k}\sum_{i'=1,i'\neq i}^{n_k}\left(x_{ik} - \bar{x}_k\right)\left(x_{i'k} - \bar{x}_k\right)}{\left[\sum_{i=1}^{n_k}\left(x_{ik} - \bar{x}_k\right)^2\right]^2} \\
&= \frac{\sigma_e^2}{\sum_{i=1}^{n_k}\left(x_{ik} - \bar{x}_k\right)^2} - \left(\sigma_e^2/n_k\right) \times \left(\frac{\sum_{i=1}^{n_k}\sum_{i'=1}^{n_k}\left(x_{ik} - \bar{x}_k\right)\left(x_{i'k} - \bar{x}_k\right)}{\left[\sum_{i=1}^{n_k}\left(x_{ik} - \bar{x}_k\right)^2\right]^2}\right) \\
&= \frac{\sigma_e^2}{\sum_{i=1}^{n_k}\left(x_{ik} - \bar{x}_k\right)^2} - \left(\sigma_e^2/n_k\right) \times \underbrace{\left(\frac{\sum_{i=1}^{n_k}\left(x_{ik} - \bar{x}_k\right)}{\sum_{i=1}^{n_k}\left(x_{ik} - \bar{x}_k\right)^2}\right)^2}_{=0} \\
&= \frac{\sigma_e^2}{\sum_{i=1}^{n_k}\left(x_{ik} - \bar{x}_k\right)^2}
\end{aligned} \tag{7}$$

We now consider N_F families. We denote $\widehat{b}$ as the OLS estimator of b over all families defined as

$$\widehat{b} = \frac{\sum_{k=1}^{N_F}\sum_{i=1}^{n_k}\left(y_{ik} - \bar{y}_k\right)\left(x_{ik} - \bar{x}_k\right)}{\sum_{k=1}^{N_F}\sum_{i=1}^{n_k}\left(x_{ik} - \bar{x}_k\right)^2} = \sum_{k=1}^{N_F}\omega_k\widehat{b}_k, \tag{8}$$

where

$$\omega_k = \frac{\sum_{i=1}^{n_k}\left(x_{ik} - \bar{x}_k\right)^2}{\sum_{k=1}^{N_F}\sum_{i=1}^{n_k}\left(x_{ik} - \bar{x}_k\right)^2}. \tag{9}$$

It follows that $\mathbb{E}[\widehat{b}|\mathbf{x}] = b$ and, assuming independence between families,

$$\begin{aligned}
\text{var}(\widehat{b}|\mathbf{x}) &= \sum_{k=1}^{N_F}\omega_k^2\text{var}(\widehat{b}_k|\mathbf{x}) = \sigma_e^2 \times \sum_{k=1}^{N_F}\left\{\frac{\left(\sum_{i=1}^{n_k}\left(x_{ik} - \bar{x}_k\right)^2\right)^2}{\left(\sum_{k=1}^{N_F}\sum_{i=1}^{n_k}\left(x_{ik} - \bar{x}_k\right)^2\right)^2} \times \frac{1}{\sum_{i=1}^{n_k}\left(x_{ik} - \bar{x}_k\right)^2}\right\} \\
&= \sigma_e^2 \times \sum_{k=1}^{N_F}\left\{\frac{\sum_{i=1}^{n_k}\left(x_{ik} - \bar{x}_k\right)^2}{\left(\sum_{k=1}^{N_F}\sum_{i=1}^{n_k}\left(x_{ik} - \bar{x}_k\right)^2\right)^2}\right\} = \sigma_e^2 \times \frac{\sum_{k=1}^{N_F}\sum_{i=1}^{n_k}\left(x_{ik} - \bar{x}_k\right)^2}{\left(\sum_{k=1}^{N_F}\sum_{i=1}^{n_k}\left(x_{ik} - \bar{x}_k\right)^2\right)^2} \\
&= \frac{\sigma_e^2}{\sum_{k=1}^{N_F}\sum_{i=1}^{n_k}\left(x_{ik} - \bar{x}_k\right)^2}.
\end{aligned} \tag{10}$$

An estimator of the residual variance can obtained using the following equation

$$\widehat{\sigma_e^2} = \frac{\sum_{i=1}^{n_k} \left[(y_{ik} - \bar{y}_k) - \widehat{b}\,(x_{ik} - \bar{x}_k) \right]^2}{\left(\sum_{k=1}^{N_F} n_k \right) - N_F}. \tag{11}$$

Significance of associations were assessed using a asymptotic Wald-test and p-values were calculated as $p = 1 - \mathrm{pdf}_{\chi_1^2}\left[\widehat{b}^2/\mathrm{var}(\widehat{b}|\mathbf{x})\right]$, where $\mathrm{pdf}_{\chi_1^2}(.)$ denotes the cumulative distribution function of χ^2 distribution with 1 degree of freedom.

Within-family regression coefficients from the Robinson et al. (2015) sibling pair data and the UK Biobank sibling pair data were combined using inverse variance weighting meta-analysis.