COMP 565 Winter 2023 Assignment 1

This assignment was released on January 6 and is worth 8% of your total grade and due at **23:59 on January 20, 2023**

Question 1 [4%] Estimation of variance

Assuming standardized phenotype and genotype matrix and following an additive model, the phenotype for N individuals is a linear combination of M SNPs plus the environment:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

Let's assume the following properties hold true:

$$\begin{aligned}
\mathbb{E}[\beta_j] = 0 \quad \forall j \in \{1, \ldots, M\}, \quad \mathrm{Var}[\boldsymbol{\beta}] = \mathbb{E}[\boldsymbol{\beta}\boldsymbol{\beta}^\top] = \frac{\sigma_\beta^2}{M}\mathbf{I}_M \\
\mathbb{E}[\epsilon_i] = 0 \quad \forall i \in \{1, \ldots, N\}, \quad \mathrm{Var}[\boldsymbol{\epsilon}] = \sigma_\epsilon^2 \mathbf{I}_N \\
\tilde{\beta}_j = \frac{1}{N}\mathbf{x}_j^\top \mathbf{y}
\end{aligned}$$

and $\sigma_{\boldsymbol{\epsilon}}^2$ is known.

Using the same technique we learned from LD score regression [1] in Lecture 2, show that the ordinary least-squared estimate for σ_β^2 is:

$$\hat{\sigma}_\beta^2 = \frac{\sum_{j=1}^M l_j(\tilde{\beta}_j^2 - \frac{\sigma_\epsilon^2}{N})}{\sum_{j=1}^M l_j^2/M} \tag{1}$$

where $\tilde{\beta}_j = \frac{1}{N}\mathbf{x}_j^\top \mathbf{y}$, $l_j = \sum_{k=1}^M r_{jk}^2$ and $r_{jk} = \frac{1}{N}\mathbf{x}_j^\top \mathbf{x}_k$. You may assume that there is no population confounder.

Provide the mathematical derivation for Eq. (1).

Submit your derivation in LaTex-compiled PDF file named `COMP565_A1_variance_derivation.pdf` in myCourses.

Question 2 [4%] Implementing LD score regression

For a phenotype of interest, we have collected the marginal statistics $\tilde{\boldsymbol{\beta}}$ for $M = 4268$ SNPs and the $M \times M$ LD matrix $\mathbf{R}$ (i.e., pairwise SNP-SNP Pearson correlation). The marginal statistics are based on $N = 1000$ individuals. Download the marginal statistics and LD matrix from here:

`https://drive.google.com/file/d/119Wmw9ockQNssHel3CZ88L2GhWqvW8ZJ/view?usp=sharing`

For this question, you may also assume there is no population stratification in this dataset. Both phenotype and genotype were standardized.

Implement the very basic LD score regression algorithm with a programming language of your choice (preferably Python or R) to estimate the heritability of the phenotype.

What's your estimate of the heritability?

Submit your code with name `COMP565_A1_ldsr.py` or `COMP565_A1_ldsr.R` on MyCourses.

References

[1] Brendan K Bulik-Sullivan, Po-Ru Loh, Hilary K 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 genomewide association studies. *Nature Genetics*, 47(3):291–295, February 2015.