Arjun Biddanda
Statistical Genetics Reading Group

Problem Set 3
6/1/2025

Winner's Curse

One core finding in statistics relevant to GWAS is what is known as the Winner's Curse. This asserts that by restricting to effect-sizes above a specific threshold c, the actual underlying effect-size $\hat{Z}$ that is estimated will be estimated with some bias - where the degree of bias is tied to the p-value threshold that is used. We illustrate this using data where the underlying true effects are also shown (from a simulation of causal variants) under the polygenic model.

The datasets are under `data/pset3/` for this exercise.

1. Using different p-value thresholds, estimate the extent of absolute bias $B = \frac{1}{N_{sig}} \sum \hat{\beta} - \beta$[1]. Show visually that for stronger thresholds the effect-sizes are mis-estimated.

2. One common correction for biased effect-sizes is using the likelihood of effect-sizes *conditional* on selection as a step for de-biasing.

   The conditional likelihood is:

   $$L(|z| > c|\mu) = \frac{\phi(z-\mu)}{\Phi(-c+\mu) + \Phi(-c-\mu)},$$

   where ϕ is the normal density function and Φ is the normal cumulative density function. The conditional maximum likelihood estimator $\hat{\mu}$ serves as a de-biased estimator of the true effect-size. Using $\hat{z} = \frac{\hat{\beta}}{\hat{s}_\beta}$, plot the de-biased estimates of β against $\hat{\beta}$ for all test with $p < 10^{-4}$.

3. ♦♦♦ One other method for solving the de-biasing is using Empirical Bayesian (EB) methods. This assumes that each standardized estimate $z \sim \mathcal{N}(\mu, 1)$. The fundamental idea is to use an approximation to the *posterior* mean proposed by Efron 2009:

   $$\mathbb{E}[\mu|z] = z + \frac{d}{dz} \log p(z),$$

   where $p(z)$ is the marginal density function. By approximating with its empirical counterpart — $\log p(z) \approx \log \tilde{p}(z)$ — we can arrive at a reasonable estimator of μ. To get at $\tilde{p}$, we will use the following procedure.

   - Bin all realized z-scores into B equally spaced bins from $[min(Z), max(Z)]$. Keep track of the midpoints M of each bin in the range.
   - Generate K unit B-spline basis functions with knots at each of the M midpoints in the range.
   - Fit a poisson generalized model for the bin counts against all K spline functions evaluated at the knots. The fitted regression function at z is the estimand of $\log \tilde{p}(z)$.
   - Estimate $\mathbb{E}[\mu|z] = \hat{z} + \frac{d}{dz} \log \tilde{p}(\hat{z})$ using numerical differentiation of the fit regression function.

   Evaluate how this estimator performs from a bias perspective for accounting for the winner's curse for both datasets. Compare with the conditional likelihood approach.

Heritability

All data in this section are in `data/pset3/multi_pheno`.

[1] Calculate this separately for signals with positive effects and negative effects to avoid signs simply canceling out

Haseman-Elston Regression

One approach to estimating heritability is to consider a moment-based estimator:

$$\mathbb{E}[Y_i Y_j] = h^2 \mathbb{E}[X_i^T X_j]$$

where $\mathbb{E}[Y] = 0$, or that each phenotype is centered on 0. If genotypes are similarly centered such that:

$$G \in 0, 1, 2$$
$$X = \frac{G - 2\mu_G}{\sigma_G}$$

such that $\sigma_G = \sqrt{2p(1-p)}$ and μ_G is the population allele frequency. The regression is performed using all pairs of individuals $i < j$. Run this regression estimate for all traits in the dataset using simple linear regression.

Linear Mixed Models & REML

One way to model these quantitative phenotypes via likelihood is:

$$y \sim \mathcal{MVN}(0, \sigma_g^2 \mathbf{K} + \sigma_e^2 \mathbf{I}),$$

where $\mathbf{K}_{ij} = x_i^T x_j$, where x are the normalized genotypes as above. If we reparameterize the likelihood using $\eta^2 = \frac{\sigma_g^2}{\sigma + {}_g^2 + \sigma_e^2}$, then:

$$l(\sigma_e^2, \eta^2) = -\frac{1}{2}\log(\sigma_e^2) - \frac{1}{2n}\log\det\left(\frac{\eta^2}{m}\mathbf{K} + \mathbf{I}\right) - \frac{1}{2n\sigma_e^2}\mathbf{y}^T\left(\frac{\eta^2}{m}\mathbf{K} + \mathbf{I}\right)^{-1}\mathbf{y}$$

Using this likelihood definition - estimate $\hat{h^2} = \frac{\hat{\eta}^2}{\hat{\eta}^2 + 1}$ for each trait. Restrict $\sigma_e^2 = 1.0$ and estimate $\hat{\eta}^2 \in \{10^{-2}, 10^1\}$. The following hints may be helpful:

- For the term using the determinant - naively computing the determinant for large matrices takes quite some time ($O(n^3)$). To avoid this, you may want to use the following relationship

  $$\log\det \mathbf{A} \approx \text{trace}\log(\mathbf{A})$$

  where the finale log is the matrix-logarithm.

- For the matrix inverse term - you can use the properties of eigendecomposition

  $$\mathbf{A}^-1 = \mathbf{Q}\Lambda^{-1}\mathbf{Q}^T$$

  where $\Lambda_{ii}^{-1} = \frac{1}{\lambda_i}$ on the diagonal.