MATH6450E Project 2: Heritability Estimation based on Linear Mixed Model

FAN Min [*]
Department of Mathematics
Hong Kong University of Science and Technology
mfanab@connect.ust.hk

Abstract

One application of linear mixed model is to estimate heritability of phenotype based on genotype. In linear mixed model, covariates effects such as age and sex are represented as fixed effects part, while heritability is modeled as random effects. Heritability is estimated by the variance of random effects over total variance. In this project, we implement linear mixed model on GWAS.RData dataset to inference coefficients of fixedd effects, variance of random effects and error term, heritability and calculate standard errors using observed fisher information and delta methods.

1 Introduction

GWAS refers to Genome-wide association study. GWAS.RData collects $n = 5123$ individuals genotypes and phenotypes. $\mathbf{G} = [g_{im} \in \{0,1,2\}] \in \mathbb{R}^{n\times p}$ is the genotype matrix where $p = 319147$ and each column corresponds to a genetic marker. $\mathbf{y}$ is a $n \times 4$ phenotype matrix, where each column indicates one phenotype. Because each time we only care about one phenotype, without loss of generality, the phenotype (one colum of $\mathbf{y}$) we are considering is denoted as $\mathbf{y}$.

Linear mixed model is as follows

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{W}\mathbf{u} + \mathbf{e}. \tag{1}$$

$\mathbf{X}\beta$ is about fixed effects. $\mathbf{X} \in \mathbb{R}^{n\times(10+1)}$ includes the principal components scores corresponding to the first ten leading principal components and one column of ones, instead of covariates matrix in general setting. β is the vector of coefficients corresponding to fixed effects.

$\mathbf{W}\mathbf{u}$ is about random effects. $\mathbf{u} \sim \mathcal{N}(0, \sigma_u^2\mathbf{I})$ is the vector of random effect size. $\mathbf{W}$ is the standardized genotype matrix with zero mean and unit variance, that is,

$$w_{im} = \frac{g_{im} - 2p_m}{\sqrt{2p_m(1-p_m)p}}, \tag{2}$$

where p_m is the frequency of the reference allele.

$\mathbf{e} \sim \mathcal{N}(0, \sigma_e^2\mathbf{I})$.

The parameters $\theta = \left\{\beta, \sigma_u^2, \sigma_e^2\right\}$ can be estimated by maximum likelihood estimation. The heritability is calculated by $\hat{h}^2 = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2+\hat{\sigma}_e^2}$.

[*] https://github.com/ProteusFAN/MATH6450E/tree/master/Heritability

2 Method

2.1 Maximum Likelihood Estimation

We can easily have

$$P(\mathbf{y}|\beta, \sigma_u^2, \sigma_e^2) = \mathcal{N}(\mathbf{X}\beta, \sigma_u^2\mathbf{W}\mathbf{W}^{\mathrm{T}} + \sigma_e^2\mathbf{I}). \tag{3}$$

And the log-likelihood function is

$$\ell(\beta, \sigma_u^2, \sigma_e^2) = -\frac{1}{2}(n\log(2\pi\sigma_u^2) + \log(|\mathbf{K} + \delta\mathbf{I}|) + \frac{1}{\sigma_u^2}(\mathbf{y} - \mathbf{X}\beta)^{\mathrm{T}}(\mathbf{K} + \delta\mathbf{I})^{-1}(\mathbf{y} - \mathbf{X}\beta)), \tag{4}$$

where $\mathbf{K} = \mathbf{W}\mathbf{W}^{\mathrm{T}}$ and $\delta = \frac{\sigma_e^2}{\sigma_u^2}$.

By spectral decomposition, $\mathbf{K} = \mathbf{U}\mathbf{S}\mathbf{U}^{\mathrm{T}}$ and $\mathbf{I} = \mathbf{U}\mathbf{U}^{\mathrm{T}}$. Equation 4 becomes

$$\ell(\beta, \sigma_u^2, \delta) = -\frac{1}{2}\left(n\log(2\pi\sigma_u^2) + \log(|\mathbf{S} + \delta\mathbf{I}|) + \frac{1}{\sigma_u^2}(\mathbf{U}^{\mathrm{T}}\mathbf{y} - \mathbf{U}^{\mathrm{T}}\mathbf{X}\beta)^{\mathrm{T}}(\mathbf{S} + \delta\mathbf{I})^{-1}(\mathbf{U}^{\mathrm{T}}\mathbf{y} - \mathbf{U}^{\mathrm{T}}\mathbf{X}\beta)^{\mathrm{T}}\right) \tag{5}$$

$$= -\frac{1}{2}\left(n\log(2\pi\sigma_u^2) + \sum_{i=1}^{n}\log([\mathbf{S}]_{ii} + \delta) + \frac{1}{\sigma_u^2}\sum_{i=1}^{n}\frac{([\mathbf{U}^{\mathrm{T}}\mathbf{y}]_i - [\mathbf{U}^{\mathrm{T}}\mathbf{X}]_{i:}\beta)^2}{[\mathbf{S}]_{ii} + \delta}\right). \tag{6}$$

Taking the derivative of equation 6 w.r.t. β and setting it to zero, we can have

$$\hat{\beta} = [(\mathbf{U}^{\mathrm{T}}\mathbf{X})^{\mathrm{T}}(\mathbf{S} + \delta\mathbf{I})^{-1}(\mathbf{U}^{\mathrm{T}}\mathbf{X})]^{-1}(\mathbf{U}^{\mathrm{T}}\mathbf{X})^{\mathrm{T}}(\mathbf{S} + \delta\mathbf{I})^{-1}(\mathbf{U}^{\mathrm{T}}\mathbf{y}) \tag{7}$$

$$= \left[\sum_{i=1}^{n}\frac{1}{[\mathbf{S}]_{ii} + \delta}[\mathbf{U}^{\mathrm{T}}\mathbf{X}]_{i:}^{\mathrm{T}}[\mathbf{U}^{\mathrm{T}}\mathbf{X}]_{i:}\right]^{-1}\left[\sum_{i=1}^{n}\frac{1}{[\mathbf{S}]_{ii} + \delta}[\mathbf{U}^{\mathrm{T}}\mathbf{X}]_{i:}^{\mathrm{T}}[\mathbf{U}^{\mathrm{T}}\mathbf{y}]_i\right] \tag{8}$$

Substituting $\hat{\beta}$ in equation 6 and taking derivative w.r.t. σ_u^2, we can have

$$\hat{\sigma_u^2} = \frac{1}{n}\sum_{i=1}^{n}\frac{([\mathbf{U}^{\mathrm{T}}\mathbf{y}]_i - [\mathbf{U}^{\mathrm{T}}\mathbf{X}]_{i:}\hat{\beta})^2}{[\mathbf{S}]_{ii} + \delta} \tag{9}$$

Plugging $\hat{\sigma}_u^2$ and $\hat{\beta}$ into equation 6, we have

$$\ell(\delta) = -\frac{1}{2}\left(n\log(2\pi) + \sum_{i=1}^{n}\log([\mathbf{S}]_{ii} + \delta) + n + n\log\frac{1}{n}\sum_{i=1}^{n}\frac{([\mathbf{U}^{\mathrm{T}}\mathbf{y}]_i - [\mathbf{U}^{\mathrm{T}}\mathbf{X}]_{i:}\hat{\beta})^2}{[\mathbf{S}]_{ii} + \delta}\right), \tag{10}$$

which is a function only related to δ. We let $\log(\delta) \in [-10, 10]$, partition $[-10, 10]$ into 100 intervals evenly and apply Brent's method for each interval, whereby we find the maximum of equation 10 and corresponding $\hat{\delta}$.

We substitute $\hat{\delta}$ in equation 8 and get $\hat{\beta}$. Then putting $\hat{\delta}$ and $\hat{\beta}$ in equation 9, we can have $\hat{\sigma_u^2}$. Using the relationship among $\hat{\delta}$, $\hat{\sigma_u^2}$ and $\hat{\sigma_e^2}$, we can have $\hat{\sigma_e^2}$.

Heritability is estimated as

$$\hat{h}^2 = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_e^2}. \tag{11}$$

2.2 Standard Error based on Observed Fisher Information and Delta Method

Probability density function of random variable X w.r.t parameter θ is $f(X;\theta)$. And fisher information of one observation is defined as

$$\mathcal{I}(\theta) = \mathrm{E}\left[\left(\frac{\partial}{\partial\theta}\log f(X;\theta)\right)^2\middle|\,\theta\right], \tag{12}$$

$$= -\mathrm{E}\left[\frac{\partial^2}{\partial\theta^2}\log f(X;\theta)\middle|\,\theta\right], \tag{13}$$

When θ is a vector, the element of fisher information matrix based on one observation is

$$\begin{aligned} [\mathcal{I}(\theta)]_{i,j} &= \mathrm{E}\left[\frac{\partial}{\partial\theta_i} \log f(X;\theta) \frac{\partial}{\partial\theta_j} \log f(X;\theta) \middle| \theta \right], && (14) \\ &= -\mathrm{E}\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j} \log f(X;\theta) \middle| \theta \right]. && (15) \end{aligned}$$

One property of MLE is $\sqrt{n}(\hat{\theta}-\theta) \xrightarrow{d} \mathcal{N}(0, \frac{1}{\mathcal{I}(\theta)})$. In our case, $(\hat{\theta}-\theta) \xrightarrow{d} \mathcal{N}(0, \frac{1}{\mathcal{I}(\theta)})$ where $\mathcal{I}(\theta)$ is cumulative observed fisher information instead of fisher information based on only one observation.

As for heritability, we can use delta method. Let $\hat{h}^2 = g(\hat{\sigma}_u^2, \hat{\sigma}_e^2)$, where $g(x,y) = \frac{x}{x+y}$. We have $(\hat{h}^2 - h^2) \xrightarrow{d} \mathcal{N}(0, \nabla g(\sigma_u^2, \hat{\sigma}_e^2)^{\mathrm{T}} \mathcal{I}(\sigma_u^2, \hat{\sigma}_e^2)^{-1} \nabla g(\sigma_u^2, \hat{\sigma}_e^2))$.

3 Experiment and Result

3.1 Estimation

First, we give the result of estimators using maximum likelihood estimation as table 1 and 2 show.

Table 1: $\hat{\beta}$

Phenotype index	1	2	3	4
β_1	-98.7419	207.8310	-169.5053	-129.4608
β_2	-0.7463	-1.2053	-0.4977	-0.6524
β_3	-4.8848	-1.1517	-5.9902	2.0639
β_4	-2.7444	-0.1593	-3.2975	-0.3526
β_5	-4.0216	-0.6453	-4.8566	-1.0331
β_6	-0.6055	2.2048	-1.5982	-0.3666
β_7	-0.3871	-3.2558	2.5048	0.1385
β_8	0.8406	-0.5296	0.8713	1.3091
β_9	0.4326	-0.2600	0.9979	-1.1264
β_{10}	0.7049	2.6655	-0.0058	-0.5154
β_{11}	1.3878	-2.8899	2.3698	1.8129

Table 2: $\hat{\sigma}_u^2, \hat{\sigma}_e^2, \hat{h}^2$

Phenotype index	1	2	3	4
$\hat{\sigma}_u^2$	0.2180	0.3045	0.2854	0.1544
$\hat{\sigma}_e^2$	0.7715	0.6890	0.6890	0.8430
$\hat{h}^2$	0.2203	0.3065	0.2902	0.1548

3.2 Standard Error

Then we can calculate standards errors of estimators based on observed fisher information matrix and delta methods, as table 3 and 4 show.

4 Discussion

Using linear mixed model, we estimate fixed effects coefficients β, the variance of random effects σ_u^2, the variance of error term σ_e^2 and the heritability h^2 and compute their standard errors based on observed fisher information matrix.

There is one weird result. The standard error of β is too large and even the magnitude of $\mathrm{se}(\hat{\beta})$ is the same as or bigger than the magnitude of $\hat{\beta}$. It might indicate that the estimation of β is not accurate

Table 3: $\mathrm{se}(\hat{\beta})$

Phenotype index	1	2	3	4
$\mathrm{se}(\beta_1)$	230.6555	177.3722	171.4476	171.4921
$\mathrm{se}(\beta_2)$	1.6478	1.8320	1.7800	1.4887
$\mathrm{se}(\beta_3)$	1.4532	1.6002	1.5657	1.3388
$\mathrm{se}(\beta_4)$	1.3590	1.4800	1.4509	1.2669
$\mathrm{se}(\beta_5)$	1.3463	1.4640	1.4354	1.2580
$\mathrm{se}(\beta_6)$	1.2835	1.3854	1.3572	1.2110
$\mathrm{se}(\beta_7)$	1.2890	1.3895	1.3640	1.2140
$\mathrm{se}(\beta_8)$	1.2556	1.3363	1.3138	1.1840
$\mathrm{se}(\beta_9)$	1.2736	1.3405	1.3174	1.1857
$\mathrm{se}(\beta_{10})$	1.3740	1.2900	1.2757	1.1682
$\mathrm{se}(\beta_{11})$	3.1844	2.4377	2.3558	2.3739

Table 4: $\mathrm{se}(\hat{\sigma}_u^2), \mathrm{se}(\hat{\sigma}_e^2), \mathrm{se}(\hat{h}^2)$

Phenotype index	1	2	3	4
$\mathrm{se}(\hat{\sigma}_u^2)$	0.0750	0.0549	0.0554	0.0546
$\mathrm{se}(\hat{\sigma}_e^2)$	0.0734	0.0541	0.0540	0.0556
$\mathrm{se}(\hat{h}^2)$	0.0496	0.0505	0.0527	0.0512

enough. This might be caused by the fact that $\mathbf{X}$ is principal component score of phenotype matrix, instead of true covariates matrix.