Random-effects models incorporating genomic relationships among individuals

Jing Hua Zhao, Jian’an Luan, Stephen Sharp

MRC Epidemiology Unit & Institute of Metabolic Science
University of Cambridge School of Clinical Medicine

E-mail: jinghua.zhao@mrc-epid.cam.ac.uk
Outline

1. Background
2. Models and Data
3. Results
4. Discussion
Outline

1. Background

2. Models and Data

3. Results

4. Discussion
Background

Traditionally, correlations for pairs of relatives in a pedigree are assumed to be known. In contrast, individuals in a population sample are considered to be unrelated.

Whole genome data available in a genomewide association study (GWAS) allows for a genomic relationship matrix (GRM) to be estimated for both types of data. Software to furnish this include GCTA, SNPRelate, PLINK, among others.

It is important to understand the implications for data analysis when a GRM is involved.
Outline

1 Background

2 Models and Data

3 Results

4 Discussion
Statistical Models

Linear mixed model \( y = X\beta + Zu + \epsilon \)

- \( y \) – a vector of observations
- \( X \) – matrix of known covariates
- \( \beta \) – a vector of unknown regression coefficients
- \( Z \) - a known matrix of either continuous or dummy variables
- \( u \) – a vector of unknown random effects
- \( \epsilon \) - a vector of (unobservable random) errors

A key assumption about \( u \) and \( \epsilon \) is that they are normally distributed with \( E[u, \epsilon] = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \), \( \text{Var}[u, \epsilon] = \begin{bmatrix} G & 0 \\ 0 & R \end{bmatrix} \) so that \( y \sim N(X\beta, V) \) with \( V = R + ZGZ \). The model extends linear regression \( y = X\beta + \epsilon \) and counterparts for a binary outcome are available.
Parameter estimation

Parameter estimation for linear and logistic regression can be derived from ordinary least squares (OLS) or maximum likelihood (ML). For mixed models it is often from restricted maximum likelihood (REML).

Usually for data containing pedigrees, the models can be expressed for each pedigree, i.e., the variance-covariance matrix for the whole study sample is block diagonal.

However, it is now necessary to accommodate nonzero correlations among all individuals.
The polygenic model

Given whole genome data involving many single nucleotide polymorphisms (SNPs), regression models of covariate ($X$) and standardised genotypes ($W$) on outcome ($y$) is now

$$y = X\beta + Wu + \epsilon$$

$$V(y) = W^T W \sigma_u^2 + I \sigma^2.$$ The model is often expressed simply as a polygenic model

$$y = X\beta + g + \epsilon$$

with $V(y) = A \sigma_g^2 + I \sigma^2$ with $g$ being a polygene. The heritability is defined as $h^2 = \sigma_g^2 / (\sigma_g^2 + \sigma^2)$

These are also naturally defined for a binary outcome.
A simple pedigree

The figure shows that individuals are from a single pedigree (left) or unrelated (right); only those in dark had genotype data and used in the analysis.
The familial relationship matrix (FRM)

<table>
<thead>
<tr>
<th></th>
<th>309</th>
<th>908</th>
<th>1649</th>
<th>7098</th>
<th>10418</th>
<th>11361</th>
<th>11879</th>
<th>12004</th>
<th>12785</th>
<th>19113</th>
<th>19235</th>
<th>20334</th>
<th>22301</th>
</tr>
</thead>
<tbody>
<tr>
<td>309</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>908</td>
<td>0.250</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1649</td>
<td>0.500</td>
<td>0.125</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7098</td>
<td>0.500</td>
<td>0.500</td>
<td>0.250</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10418</td>
<td>0.250</td>
<td>0.500</td>
<td>0.125</td>
<td>0.500</td>
<td>1.0000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11361</td>
<td>0.250</td>
<td>0.125</td>
<td>0.125</td>
<td>0.250</td>
<td>0.125</td>
<td>1.0000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11879</td>
<td>0.000</td>
<td>0.500</td>
<td>0.000</td>
<td>0.000</td>
<td>0.500</td>
<td>0.000</td>
<td>1.0000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12004</td>
<td>0.500</td>
<td>0.250</td>
<td>0.250</td>
<td>0.500</td>
<td>0.250</td>
<td>0.250</td>
<td>0.0000</td>
<td>1.0000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12785</td>
<td>0.250</td>
<td>0.500</td>
<td>0.125</td>
<td>0.500</td>
<td>0.125</td>
<td>0.500</td>
<td>0.2500</td>
<td>1.0000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16612</td>
<td>0.250</td>
<td>0.500</td>
<td>0.125</td>
<td>0.500</td>
<td>0.125</td>
<td>0.500</td>
<td>0.2500</td>
<td>0.5000</td>
<td>1.0000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18645</td>
<td>0.250</td>
<td>0.125</td>
<td>0.125</td>
<td>0.250</td>
<td>0.125</td>
<td>0.125</td>
<td>0.0000</td>
<td>0.5000</td>
<td>0.1250</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19113</td>
<td>0.500</td>
<td>0.250</td>
<td>0.250</td>
<td>0.500</td>
<td>0.250</td>
<td>0.500</td>
<td>0.0000</td>
<td>0.5000</td>
<td>0.2500</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19235</td>
<td>0.250</td>
<td>0.125</td>
<td>0.125</td>
<td>0.250</td>
<td>0.125</td>
<td>0.500</td>
<td>0.0000</td>
<td>0.2500</td>
<td>0.1250</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20334</td>
<td>0.250</td>
<td>0.0625</td>
<td>0.5000</td>
<td>0.1250</td>
<td>0.0625</td>
<td>0.0625</td>
<td>0.0625</td>
<td>0.0000</td>
<td>0.1250</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22301</td>
<td>0.250</td>
<td>0.125</td>
<td>0.125</td>
<td>0.250</td>
<td>0.125</td>
<td>0.125</td>
<td>0.0000</td>
<td>0.5000</td>
<td>0.1250</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Same values in the table indicate same degree of relationship.
We have seen that all elements in the difference matrix were nonzero.
Comparison of relationship matrices from different sources

Now we compare these from different software. Denote the relationship matrices to be $A_{FRM}$, $A_{GCTA}$, $A_{SNPRelate}$, $A_{PLINK}$ and $\|A\|_1 = \sum_i \sum_j |A_{ij}|$ the entrywise norm of matrix $A$, the following table was obtained.

<table>
<thead>
<tr>
<th></th>
<th>FRM</th>
<th>GCTA</th>
<th>SNPRelate</th>
<th>PLINK</th>
<th>$|.|_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>FRM</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td>65.25</td>
</tr>
<tr>
<td>GCTA</td>
<td>6.80</td>
<td>0</td>
<td></td>
<td></td>
<td>62.19</td>
</tr>
<tr>
<td>SNPRelate</td>
<td>6.13</td>
<td>2.57</td>
<td>0</td>
<td></td>
<td>62.75</td>
</tr>
<tr>
<td>PLINK</td>
<td>5.63</td>
<td>3.90</td>
<td>2.49</td>
<td>0</td>
<td>65.25</td>
</tr>
</tbody>
</table>

We have $\|A_{FRM}\|_1=65.25$, $\|A_{FRM} - A_{GCTA}\|_1=6.80$, and so on. From these it appears PLINK was overall the closest to those based on family structure. Entries in $A_{PLINK}$ and $A_{SNPRelate}$ were also close since they employed the same algorithm.
The aim of our investigation was

- To evaluate the evidence of genetic association from linear (logistic) regression and mixed models;
- To compare polygenic effects when $A$ can be derived from family data as well as whole genome data;
- To assess the impact on longitudinal data and meta-analysis involving correlations over time or study sample, e.g., age-related heteroscedasticity of polygenic effect, one- (individual participant data) versus two-step meta-analysis.

We will use FOUR datasets to investigate the three scenarios.
The European Prospective Investigation in Cancer and nutrition (EPIC-Norfolk)

It is a population-based cohort study of 25,663 men and women of European descent aged 39-79 years recruited in Norfolk, UK between 1993 and 1997.

A GWAS includes 3,850 individuals who had data from Affymetrix 500K assays based on a case-cohort design where 1,284 participants had BMI greater than 30 and 2,566 participants who were randomly selected from the cohort. In the final sample, there were 3,552 individuals among whom 1,135 were obese cases and 2,417 cohort samples.
The Framingham Heart Study (FHS)

The study began in 1948 with the recruitment of adults from the town of Framingham, Massachusetts.

Data available for Genetic Analysis Workshop 16 (GAW16) were 7,130 individuals from the original cohort (373), the first generation cohort (2,760) and the third generation cohort (3,997). A total of 6,848 individuals in 962 families with Affymetrix 500K SNP genotypes were used here.
The MRC National Survey of Health and Development (NSHD)

It is a birth cohort study of men and women born in England, Scotland and Wales in one week in 1946 who have been followed up ever since.

BMI variables were available longitudinally at 12 time-points on 2,452 individuals aged from 2 to 63 years (1,225 men, 1,227 women). The genotypes at 147,949 SNPs were available on the Metabochips.

The Metabochip was designed based on findings from GWASs for type 2 diabetes, coronary artery disease and myocardial infarction, as well as for related traits such as body mass index, glucose and insulin levels, lipid levels, and blood pressure, containing 200,000 SNPs for part of genomic regions with additional loci for fine-mapping.
It is collaboration amongst eight of the ten countries in the EPIC cohorts to investigate genelifestyle interaction on 12,403 incident type 2 diabetes cases occurring between 1991 and 2007 and a random subcohort of 16,835 individuals.

A GWAS was carried out on 9,431 individuals using Illumina660WQuad GeneChip. Countries with Illumina660WQuad data were France, Germany, Italy, Netherlands, Spain, Sweden and UK, involving 24 centers.
Outline

1. Background
2. Models and Data
3. Results
4. Discussion
The mixed models were based on imputed genotypes according to HapMap. The heritability for BMI was 0.20 with relatively large standard error of 0.13. The comparisons between linear regression (N=2,414) and logistic regression (N=3,549) with mixed models for 96 SNPs (Locke et al. in preparation) (no rs4787491) found in the imputed data are shown in the following figure.

Clearly, when correlation among observations is ignored the statistical significance appeared stronger for both linear and logistic regressions.
SNP-trait association

QQ plots for SNP-BMI (left) SNP-obesity (right) association.
GRM and relationship matrix based on designated familial structure were both obtained. \( h^2 = 0.47 \) with adjustment for age with standard errors 0.02 and 0.03, respectively. Although the large difference in loglikelihoods suggested that those based on the family structures was desirable, the actual estimates were very close.

<table>
<thead>
<tr>
<th>Family structure</th>
<th>Genomic data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variance components</td>
<td>SE</td>
</tr>
<tr>
<td>( \sigma_g^2 )</td>
<td>10.84</td>
</tr>
<tr>
<td>( \sigma_e^2 )</td>
<td>12.07</td>
</tr>
<tr>
<td>( \sigma_P^2 )</td>
<td>22.91</td>
</tr>
<tr>
<td>( h^2 = \sigma_g^2 / \sigma_P^2 )</td>
<td>0.47</td>
</tr>
<tr>
<td>( l )</td>
<td>-13492.10</td>
</tr>
<tr>
<td>( l_0 )</td>
<td>-13731.54</td>
</tr>
<tr>
<td>( \chi^2 )</td>
<td>478.87</td>
</tr>
</tbody>
</table>
The $h^2$'s varied markedly with age, ranging from 0.0 (95% CI: 0-0.082) at age 2 to 0.205 (95% CI: 0.088-0.322) at age 20 and then declining steadily during adult life. Semi-parametric regression yielded $F=14.72$, $P=0.000505$, $R^2_{adj}=0.857$, showing strong evidence of age-related change in heritability.
As expected, the heritability estimates were more reasonable and more precise as sample sizes got larger. Possibly due to the same reasons, the meta-analysis based directly on heritabilities led to nearer estimate ($h^2 = 0.31$, SE=0.12) to the one-step meta-analysis ($N = 9,212$) than indirectly based on variance components.

<table>
<thead>
<tr>
<th></th>
<th>Individual participant data</th>
<th>Meta-analysis by variance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\sigma^2_g$</td>
<td>$\sigma^2_e$</td>
</tr>
<tr>
<td>$\sigma^2_g$</td>
<td>7.91</td>
<td>0.77</td>
</tr>
<tr>
<td>$\sigma^2_e$</td>
<td>12.08</td>
<td>0.74</td>
</tr>
<tr>
<td>$\sigma^2_p$</td>
<td>19.99</td>
<td>0.30</td>
</tr>
<tr>
<td>$h^2$</td>
<td>0.40</td>
<td>0.04</td>
</tr>
</tbody>
</table>
Outline

1 Background

2 Models and Data

3 Results

4 Discussion
SNP-trait association

The commonly reported SNP-trait associations seem anticonservative compared to those obtained from models incorporating genetic relatedness.

Results based on the 96 reported SNPs would resemble genomewide associations.

The finding is in line with literature on the magnitude of the association either in the case-control setting or population stratification, i.e., spurious genotype-phenotype association due to difference in allele frequencies in different populations.
Impact of GRM estimation

We saw equivalence of genomic versus specific relationship as in linkage analysis and association on FEV1/FVC in FHS. It is possible that the relationship according to family structure is true leading to slightly better REML estimates nevertheless there was no substantial difference in the heritability estimates.

Sampling errors in estimating GRM do not appear to be a concern in estimating the polygenic effect, although likelihood ratio-based explained proportion of variation statistics ($R^2$) will also be less reliable.

Polygenic model allows population sample and family data to be used seamlessly in the same model, avoiding the need for meta-analysis or joint modelling of related and unrelated individuals as done in earlier literature.
Correlations between individuals have implications for both longitudinal data analysis and meta-analysis.

An on-going collaboration in the GIANT consortium requested contribution of variance components. Our results showed that meta-analysis on variance components will be less accurate. Meta-analysis on individual heritabilities may still be susceptible to bias.

Furthermore, even with whole genome data and easily measure trait such as BMI in Europeans, we see variations in heritability estimates so the amount of variation across various ethnicities will be even larger.
Issues

We highlight the following areas which require further work,

- One step implementation involving longitudinal or clustered data.
- Consortium meta-analysis involving summary statistics and relatedness.
- Modelling survival outcome in a case-cohort design.

As an example, given the proximity of $A_{PLINK}$ to $A_{family}$, it was not positive definite and therefore failed to be associated with REML estimates in neither GCTA or R. Robust algorithm coupled with efficient implementations in R or preferably in C/C++ is necessary and possibly in a multicore and/or parallel environment.
Acknowledgements

The EPIC Norfolk Study is funded by program grants from the Medical Research Council UK and Cancer Research UK, and by additional support from the European Union, Stroke Association, British Heart Foundation, Department of Health, Food Standards Agency, and the Wellcome Trust.

The Genetic Analysis Workshops are supported by NIH grant R01 GM031575 from the National Institute of General Medical Sciences, and additionally by a NIMH K08 award (KO8 MH074057).

Work on NSHD was supported by the Medical Research Council [MC_U106179472; U106179472], the British Heart Foundation [RG/10/12/28456] and the Wellcome Trust [088869/B/09/Z].

The InterAct study received funding from the European Union (Integrated Project LSHM-CT-2006-037197 in the Framework Programme 6 of the European Community).
Books


Papers

Papers