On the Relationships Between Population Characteristics and QTL Mapping

Benjamin Yakir

Department of Statistics
The Hebrew University

IMS, Singapore, 2009

Outline of the talk

Models

- A model for the response.
- A model for pedigrees.
- A model for the population.

Score statistics

- Association and linkage.
- The noncentrality parameter.

Population covariance

- The noncentrality parameter in the normal case.
- The structure of the covariance and the resulting effect.

Genetic model for the trait

The phenotype

- Y = the response in a GLM regression.
- $\theta = h(\eta)$ = the natural parameter. $\eta(\beta)$ = a linear predictor.
- $\ell(\beta) = \sum_{i=1}^{n}[Y_i\theta_i - \psi(\theta_i)]/\phi$ = the conditional likelihood of the sample, given the predictors.

The genotype

- Q = the genotype at the investigated locus.
- G = the genetic background, independent of Q.
- $\eta_i = \beta Q_i + z_i + G_i$, for z deterministic predictors.
- The genetic predictors are correlated.
- Q and G share the same correlation structure.

A model for pedigrees

The pedigree graph

- A pedigree is composed of founders and non-founders.
- Founder gametes = Independence + Hardy-Weinberg.
- Segregation to non-founders = Mendel rules.

The sample

- The sample is a sub-collection of pedigree members.
- Y and Q are measured over the sample. Correlation tested.
- $\mathbb{E}(Q_i|\nu) = \mathbb{E}(Q_i)$, $\mathbb{V}\mathrm{ar}(Q_i) = V(0)$.
- $\mathbb{C}\mathrm{ov}(Q_i, Q_j|\nu) = \hat{\Phi}_{ij} V(0)$.
- $\hat{\Phi}_{ij}$ = identity by decent (IBD) between i and j.
  ν = inheritance vector.

A population model

A coalescent tree

- The gametes of the sample coalesce to the most recent common ancestor (MRCA).
- Mutations emerge as a Poisson process over the tree.
- A mutation $=$ an ergodic transformation of Q.

The sample

- The sample corresponds to the leaves of the tree.
- The marginal distributions of genotypes are fixed.
- $\mathbb{C}\mathrm{ov}(Q_i, Q_j|\mu) = V(m_{ij})$, $V(m) \downarrow 0$.
- $m_{ij} = \#\{\text{mutations}\}$ between i and j and their MRCA.

A population-pedigree model

The graph

- The population model describes the production of the founders' gametes.
- The pedigree graph describes the inheritance of these gametes to the sample.

The sample genotypes

Conditional on the mutation process and the inheritance vector:

$$\begin{aligned}
\mathbb{E}(Q_i|\nu,\mu) &= \mathbb{E}(Q_i)\ , \\
\mathbb{V}\mathrm{ar}(Q_i|\nu,\mu) &= V(0)\ , \\
\mathbb{C}\mathrm{ov}(Q_i,Q_j|\nu,\mu) &= \hat{\Phi}_{ij}V(0) + [1-\hat{\Phi}_{ij}]V(m_{f(i),f(j)}) \\
&= \hat{\Phi}_{ij}[V(0) - V(m_{f(i),f(j)})] + V(m_{f(i),f(j)})\ .
\end{aligned}$$

The likelihood function

Joint log-likelihood

$$\ell(\beta; Y, Q) = \log \mathbb{E}\left[e^{\ell(\beta)} \middle| Y, Q\right]$$

Association (conditional on Y)

$$\ell(\beta; Q \mid Y) = \log \mathbb{E}\left[e^{\ell(\beta)} \middle| Y, Q\right] - \log \mathbb{E}\left[e^{\ell(\beta)} \middle| Y\right]$$

Linkage (conditional on Y)

$$\ell(\beta; \nu \mid Y) = \log \mathbb{E}\left[e^{\ell(\beta)} \middle| Y, \nu\right] - \log \mathbb{E}\left[e^{\ell(\beta)} \middle| Y\right]$$

(ν can be replaced by μ or by the pair ν, μ).

Score statistic for association

- Score = derivative of the log-likelihood, evaluated at $\beta = 0$.
- If $\beta = 0$ then (Y, G) and Q are independent.
- Let $\epsilon_i = \epsilon(Y_i, z_i, G_i) = \left[Y_i - \psi'(h(\eta_i(0))\right] h'(\eta_i(0))/\phi$.

Then

$$\begin{aligned}\dot{\ell}(0; Q \mid Y) &= \sum_{i=1}^{n} \mathbb{E}(\epsilon_i | Y)\{Q_i - \mathbb{E}(Q_i)\} \\ &= \left\langle \mathbb{E}(\epsilon | Y), Q - \mathbb{E}(Q) \right\rangle ,\end{aligned}$$

and

$$\mathbb{V}\mathrm{ar}\big(\dot{\ell}(0; Q \mid Y)\big) = \mathrm{tr}\big(\mathrm{AB}\big) ,$$

for $\mathrm{A} = \mathbb{V}\mathrm{ar}(\mathbb{E}[\epsilon | Y])$, $\mathrm{B} = \mathbb{V}\mathrm{ar}(Q)$.

Score statistic for linkage

- $\dot{\ell}(0; \nu \mid Y) \equiv 0$, since $\mathbb{E}(Q|\nu) = \mathbb{E}(Q)$.
- The local behavior of the log-likelihood is determined by the second derivative:

$$\begin{aligned}\ddot{\ell}(\beta_0; \nu \mid Y) &= 2 \sum_{i=2}^{n} \sum_{j=1}^{i-1} \mathbb{E}[\epsilon_i \epsilon_j | \, Y] \big\{ \mathbb{E}[Q_i Q_j | \nu] - \mathbb{E}[Q_i Q_j] \big\} \\ &= \big\langle \mathrm{vec}(\mathbb{E}[\epsilon \epsilon' | Y]), \mathrm{vec}(\mathbb{V}\mathrm{ar}[Q|\nu] - \mathbb{V}\mathrm{ar}[Q]) \big\rangle\end{aligned}$$

- The same argument applies to the population and the population-pedigree model when ν is replaced by μ or by ν and μ, respectively.

Statistical power

The noncentrality parameter

- The noncentrality parameter = expectation of the test statistic (squared).
- A major factor in the determination of the power.
- Local alternatives + score $\Rightarrow$ ncp $\propto \mathbb{V}\mathrm{ar}\big(\dot{\ell}(0; Q \mid Y)\big)/n$.

Normal response

- Assume normal family, regular linear regression.
- $\mathbb{V}\mathrm{ar}(G) = \sigma_g^2 \cdot \mathbb{V}\mathrm{ar}(Q) = \sigma_g^2\, \mathrm{B}$.
- $\mathbb{V}\mathrm{ar}(Y) = \Sigma = \sigma_e^2\, \mathrm{I} + \sigma_g^2\, \mathrm{B}$.
- $\mathrm{A} = \mathbb{V}\mathrm{ar}(\mathbb{E}[\epsilon|Y]) = \Sigma^{-1}$.

Facts about matrices

Eigenvalues

- λ_i, $1 \leq i \leq n$, eigenvalues of B.
- B non-negative definite $\Rightarrow \lambda_i \geq 0$.
- B, Σ share eigenvectors $\Rightarrow \mathrm{tr}(\mathrm{AB}) = \sum_{i=1}^{n} \left\{\lambda_i/[\sigma_e^2 + \sigma_g^2 \lambda_i]\right\}$.
- If B_1, B_2 non-negative definite $\Rightarrow$
  $\lambda_i(\mathrm{B}_1) \leq \lambda_i(\mathrm{B}_1 + \mathrm{B}_2) \leq \lambda_i(\mathrm{B}_1) + \lambda_1(\mathrm{B}_2)$

Rank-one modifications

- $\mathrm{B} = \mathrm{B}_0 + bb'$, $\Sigma = \Sigma_0 + \sigma_g^2 bb'$.
- $\mathrm{A} = \Sigma^{-1}$, $\mathrm{A}_0 = \Sigma_0^{-1}$.
- $\mathrm{tr}(\mathrm{AB}) = \mathrm{tr}(\mathrm{A}_0\mathrm{B}_0) + [1 - \sigma_g^2 b'\mathrm{A}_0\mathrm{B}_0\mathrm{A}_0 b]/[1 + \sigma_g^2 b'\mathrm{A}_0 b]$.

Population covariance structure

Covariance

- $V_{ij} = \mathrm{E}\big[V(m_{ij})\big]$ = covariance between i and j.
- A decreasing function of the time to the MRCA of i and j.
- $V_{ii} = V(0)$ = the variance.

A sum of rank-one matrices

- i and j coalesced by time t, h has not $\Rightarrow V_{ih} = V_{jh}$.
- τ_t = time between consecutive coalescent events.
- $v_t = \mathrm{E}\big[V\big(m(2\sum_{j=1}^{t-1}\tau_j)\big)\big] - \mathrm{E}\big[V\big(m(2\sum_{j=1}^{t}\tau_j)\big)\big]$.
- $1_{t,k}$ = indicator of a coalescent set.
- $\mathrm{B} = \mathrm{I}\cdot v_T + \sum_{t=2}^{T}\big\{v_{t-1}\big[\sum_{i=1}^{K_t}(1_{t,i})(1_{t,i})'\big]\big\}$.

The effect of population covariance

A bound on the trace

- $\epsilon > 0$, given.
- $T(\epsilon) \Leftrightarrow \sum_{t=T(\epsilon)}^{T} \left\{ v_t \max_{1 \le i \le K_t}(n_{t,i}) \right\} \le \epsilon$.

$$\Rightarrow \mathrm{tr}(\mathrm{AB}) \le \mathrm{rank}\left(\sum_{t=2}^{T(\epsilon)} \left\{ v_{t-1} \left[\sum_{i=1}^{K_t} (1_{t,i})(1_{t,i})' \right] \right\} \right) + \epsilon n$$

Corollary

- Assume $v_t \le c\tau_t$, for all t.
- For the Kingman coalescent $T = n$ and $T(\epsilon) = c\sqrt{n \log n}$.
- ncp of population model is $o(1)$.
- ncp of population-pedigree model = ncp pedigree model with IBD probabilities decreased by $1 - \mathrm{E}[V(m_{f(i),f(j)})]/V(0)$.