The Abdus Salam
International Centre for Theoretical Physics

United Nations
Educational, Scientific and
Cultural Organization

IAEA
International Atomic Energy Agency

1863-20

Advanced School and Conference on Statistics and Applied Probability in Life Sciences

24 September - 12 October, 2007

Identifying Genetic Variants Associated with Disease: Case-Control Association Testing with Related Individuals

Mary Sara McPeek

Department of Statistics
University of Chicago
Chicago, IL 60637, USA

Identifying Genetic Variants Associated with Disease: Case-Control Association Testing with Related Individuals

Mary Sara McPeek

Department of Statistics
University of Chicago

Joint work with Catherine Bourgain and Tim Thornton

Case-control association testing: some preliminaries

- We consider a complex trait or disease (e.g. asthma, alcoholism), which we treat as binary (affected/unaffected).
- **Complex** trait: may be influenced by multiple genetic and non-genetic factors
- Goal: identify some of the genetic risk factors related to the trait.

Terminology

- A person who has the disease may be called an **affected** or a **case**.
- **Control** could be someone who does not have the disease or someone whose disease status is unknown (these types are treated differently in the analysis).

Terminology (continued)

- **SNP** (single nucleotide polymorphism) — site in genome with single base-pair change that distinguishes some individuals from others in same population, e.g. A**A**GGCTAA vs. A**T**GGCTAA

- The two different variants at a SNP are called **alleles**.

- **Genotype** is the pair of alleles of an individual at a SNP. Typically observe only number of copies of each allele held by individual, e.g. i copies of allele A, which is equivalent to 2-i copies of allele T, $i = 0, 1, 2$.

Case-control association testing

- For a given marker, compare the allele or genotype distributions of cases and controls.

- Null hypothesis is that there is no difference between case and control allele/genotype distributions at the given marker.

- E.g. consider classical χ^2 test for association with a biallelic marker, equivalent to test for independence in 2×2 table

	Case	Control
Allele 0	C_{00}	C_{01}
Allele 1	C_{10}	C_{11}

- This χ^2 test is valid when alleles are independent within and between individuals under the null hypothesis

Use of related individuals in case-control testing: how it arises, why it can be desirable

- Families sampled for a linkage study may be included in an association study.
  - **Linkage** is co-inheritance of trait with allele or genotype within a family.
  - Linkage analysis of families with multiple affected individuals may be used for coarse mapping of genetic variants.
  - Then population-based association used for finer-scale mapping.
- Sampling individuals from families with multiple affecteds may increase power to detect assocation with complex traits because of enrichment for genetic cases.
  - Complex diseases such as breast cancer and heart disease have both genetic and non-genetic causes.
  - Cases from families with multiple affecteds are more likely to have predisposing genetic variants.

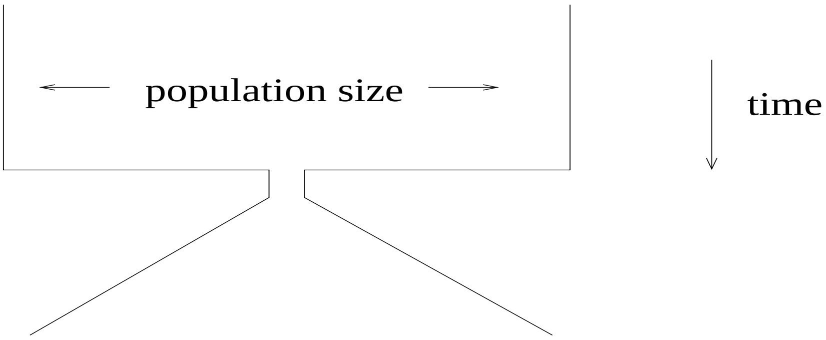
Use of related individuals in case-control testing: how it arises, why it can be desirable (cont.)

- Use of unaffected relatives of cases as controls can provide some measure of protection against potential problems of population substructure.
  - Frequency of disease may differ across sub-populations, frequency of allele or genotype may differ as well.
  - Sub-population may be a hidden covariate contributing to false detection of association.
  - Using related cases and controls makes this problem less likely.

- Founder populations, in which most or all individuals are related, can be particularly valuable for genetic studies.

Complex trait mapping in founder populations

- **Founder population:** a population in which a recent bottleneck has resulted in a large number of individuals all descended from a small number of founders



- Founder populations may be particularly useful for complex trait mapping because of
  - reduced genetic heterogeneity due to small number of founders
  - in some cases, reduced environmental heterogeneity
  - linkage disequilibrium may exist over greater distances ⇒ less dense map required to detect association

Statistical issues that arise with use of related individuals in case-control testing

- When some individuals in the samples are related, it creates dependence among the observations.

- Both case-control status and allele/genotype run in families

- Type I error:

  - application of standard methods can result in a dramatic increase in false detections (Newman et al. 2001; Bourgain et al. 2003).

- Power:

  - information on relatedness can be used to increase power

  - explicitly taking into account the fact that there is an enrichment for predisposing variants in affecteds with affected relatives can also increase power

- We develop quasi-likelihood (QL) methods for case-control association testing of genetic traits with related individuals. Properties include:
  - uses only first and second moments
  - applicable to any sample of related individuals
  - Type I error is corrected for the dependence
  - weights depending on case-control status and relationships of individuals are used to optimize power (maximize non-centrality parameter) within a linear class of statistics
  - allows us to leave unspecified some parts of the model of which we are ignorant $\Longrightarrow$ retains a major part of the appeal of the original case-control association test
  - computationally feasible, even in large complex inbred pedigrees with multiple inbreeding loops

- More generally, QL inference methods can be used to extend other types of classical population genetic inference to founder populations, e.g.
  - allele frequency estimation
  - Armitage test for case-control association
  - Hardy-Weinberg equilibrium test

- Agenda for remainder of talk:
  - General QL approach
  - 3 approaches for case-control association testing in related individuals:
    - correct the variance of the standard χ^2 statistic: $W_{\chi^2_{corr}}$
    - QL approach under a simple model for case-control differences: W_{QLS}
    - Improvement of power by more detailed consideration of properties of a genetic trait: M_{QLS}
  - Simulations and an example

Quasi-likelihood (QL) estimation

- Wedderburn (1974), Godambe (1960), Jarrett (1973), McCullagh and Nelder (1989), Heyde (1997)

- Let $X_{n\times 1}$ be random with $E(X) = \mu_{n\times 1}$ and Var$(X) = V_{n\times n}$, where
  - μ is a known, twice differentiable function of unknown parameter $\theta_{m\times 1}$,
  - V is a known differentiable function of θ (or sometimes known only up to unknown scale factor σ^2 not depending on θ), V invertible.

- Let $U(\theta) = D^T V^{-1}(X - \mu)$, where $D_{ij} = \partial\mu_i/\partial\theta_j$

- $U(\theta)$ is QL score function.

- QL estimator (QLE) $\widehat{\theta}$ of θ is a solution of $U(\theta) = 0$.

- Matrix $i_\theta = D^T V^{-1} D = Cov(U(\theta)) = -E(\partial U/\partial\theta)$ plays similar role to Fisher information.

- Under regularity conditions, $i_\theta^{1/2}(\widehat{\theta} - \theta)$ is asymptotically $N(0, I)$, where$(i_\theta^{1/2})^T i_\theta^{1/2} = i_\theta$.

QL estimation: a few more details

- In special case when both of the following hold:

  1. $\mu = D\theta$, where D is known, i.e. μ is a linear function of θ and
  2. $V = Ks(\theta)$, where K is a known invertible matrix and s is a possibly unknown scalar that may depend on θ,

  then the QL estimator for θ is same as generalized regression estimator $(D^T K^{-1} D)^{-1}(D^T K^{-1} X)$.

- However, for some of the problems we are interested in, both 1 and 2 fail to hold.

- More generally, $\widehat{\theta}$ is not linear in X and can be obtained by Newton-Raphson with Fisher scoring:

  $$\widehat{\theta}_{j+1} = \widehat{\theta}_j + (\widehat{D}_j^T \widehat{V}_j^{-1} \widehat{D}_j)^{-1} \widehat{D}_j^T \widehat{V}_j^{-1} (X - \widehat{\mu}_j),$$

  where $\widehat{D}_j = D|_{\theta=\widehat{\theta}_j}$, $\widehat{V}_j = V|_{\theta=\widehat{\theta}_j}$.

- QL estimating equation is of linear type $H(\theta)^T(X - \mu(\theta)) = 0$, where H and μ do not depend on X. QLE is asymptotically optimal among estimators obtained as solutions of linear estimating equations (under regularity conditions).

QL score test

- Simple null hypothesis
  - Suppose want to test null hypothesis $H_0 : \theta = \theta_0$ vs. alternative $H_A : \theta \neq \theta_0$.
  - Let ρ be the dimension of θ.
  - By analogy with usual likelihood score test, consider

$$W = U(\theta_0)^T i_{\theta_0}^{-1} U(\theta_0)$$

    assuming i_{θ_0} invertible.

  - Compare to a χ^2_ρ distribution asymptotically

QL Score test: composite null hypothesis

- Set $\theta = (r, a)$ and consider testing null hypothesis $H_0 : r = r_0$ vs. alternative $H_A : r \neq r_0$.

- Let ρ be the dimension of r.

- Let
$$U(r,a) = \left(\begin{array}{c} U_r(r,a) \\ U_a(r,a) \end{array} \right) = \left(\begin{array}{c} D_r^T V^{-1}(X-\mu) \\ D_a^T V^{-1}(X-\mu) \end{array} \right),$$
where $D_r = \partial\mu/\partial r$ and $D_a = \partial\mu/\partial a$.

- By analogy with usual likelihood score statistic for composite null, consider
$$W = U_r^T(r_0, \hat{a}_0) i^{rr}(r_0, \hat{a}_0) U_r(r_0, \hat{a}_0),$$
where
  - $i^{rr}(\theta)$ is the (r,r)th entry of i_θ^{-1}
  - $\hat{a}_0$ is the QLE of the nuisance parameter a when $r = r_0$, i.e. $a = \hat{a}_0$ is solution of $U_a(r_0, a) = 0$.

- Compare to a χ^2_ρ distribution asymptotically

Some terminology and notation

- **IBD:** identical by descent; a set of alleles is IBD if the alleles are inherited copies of the same ancestral allele

- **HBD:** homozygous by descent; an individual is said to be HBD at a locus if that individual's two alleles are IBD. This occurs when individuals are inbred.

- **kinship coefficient:** ϕ_{ij} is the probability that a randomly chosen pair of alleles, one each from individuals i and j are IBD at a given locus, conditional on the genealogy connecting i and j

  - For example, the kinship coefficient for siblings or for parent-offspring is .25, for first cousins it is .0625

- **inbreeding coefficient:** h_i is the probability that individual i is HBD at a given locus, conditional on the genealogy connecting i's parents; $h_i = \phi_{mf}$, where m and f are the parents of i

Case-control association testing with relatives Approach 1: Correct the variance of the standard χ^2 statistic

- For simplicity, consider a biallelic locus. All results generalize to multiple alleles.

- Let $X_j = \frac{1}{2}$(the number of alleles (0, 1, or 2) of type 1 held by individual j), $j = 1, \ldots n$

- Let D_r be case indicator vector with $D_{rj} = 1$ if j is a case, 0 if j is a control.

- The standard Pearson's χ^2 statistic for the test of allelic association can be written
$$W_{\chi^2} = \frac{n[\sum_{j \in \text{cases}}(X_j - \bar{X})]^2}{\frac{1}{2}\bar{X}(1-\bar{X})n_{\text{case}}n_{\text{con}}},$$
where $n_{\text{case}} = 1^T D_r$, and $n_{\text{con}} = n - n_{\text{case}}$.

- This statistic has the form $W = S^T[\widehat{\text{Var}_o}(S)]^{-1}S$, where $S = V^T X$ is linear in X.

- Standard χ^2 statistic has the form $W = S^T[\widehat{\text{Var}_o}(S)]^{-1}S$, where $S = V^T X$.

- Thus, $\text{Var}_o(S) = V^T \text{Var}_o(X) V$

- With unrelated outbred individuals, $\text{Var}_o(X) = \frac{1}{2}a(1-a)I_{n \times n}$, where $a = E(X)$ is the allele frequency under the null hypothesis. Can approximate a by $\bar{X}$ under the null.

- With related and possibly inbred individuals, $\text{Var}_o(X) = \frac{1}{2}a(1-a)L$, where $L_{ij} = 1 + h_i$ if $i = j$ and $2\phi_{ij}$ if $i \neq j$, with h_i and ϕ_{ij} denoting inbreeding and kinship coeffs, respectively.

- Then the corrected χ^2 statistic is $W_{\chi^2_{corr}} = W_{\chi^2}\,\gamma$, where

$$\gamma = \frac{n_{case} n_{con}}{n(D_r^T L D_r - 2\frac{n_{case}}{n} 1^T L D_r + (\frac{n_{case}}{n})^2 1^T L 1)}.$$

- Compare to χ^2_1 distribution under the null hypothesis of no association between the locus and the trait.

Case-control association testing with relatives Approach 2: QL approach under simple model

- The corrected χ^2 test works reasonably well, but can we increase power with minimal additional effort?

- Let $\mu = E(X)$, and consider the simple model: $\mu_i = a+r$ if i is a case, a if i is a control (constrain $0 < a < 1$, $0 < a + r < 1$)

- Null hypothesis is $H_0 : r = 0$ vs. alternative $H_A : r \neq 0$

- We have $\text{Var}_o(X) = \frac{1}{2}a(1-a)L$

- $D_r = \partial\mu/\partial r$ has $D_{rj} = 1$ if j is a case and 0 if j is a control

- $D_a = \partial\mu/\partial a = \mathbf{1}_{n\times 1}$

- Then we obtain QL score statistic $W_{QLS} =$

$$\frac{[D_r^T L^{-1}(X - \hat{a}_0\mathbf{1})]^2}{\hat{a}_0(1-\hat{a}_0)[D_r^T L^{-1} D_r - (D_r^T L^{-1}\mathbf{1})^2(\mathbf{1}^T L^{-1}\mathbf{1})^{-1}]}$$

where $\hat{a}_0 = (\mathbf{1}^T L^{-1}\mathbf{1})^{-1}\mathbf{1}^T L^{-1} X$.

Case-control association testing with relatives Approach 2: QL approach under simple model (cont.)

- Both $W_{\chi^2_{corr}}$ and W_{QLS} are of the form $W = S^T[\widehat{\text{Var}_o}(S)]^{-1}S$, where $S = V^T X$ is linear in X. Call this class of statistics $\mathcal{W}$.

- When the assumed model holds, i.e. $\mu_i = a + r$ if i is a case, a if i is a control, then W_{QLS} is optimal in the sense that it maximizes the non-centrality parameter.

- Thus, W_{QLS} should be more powerful than $W_{\chi^2_{corr}}$ under the assumed model.

- E.g., to compare the allele frequency in, say, Swedes vs. Han Chinese, with related individuals in the samples, W_{QLS} is much more powerful.

- Difficulty: our simple model does not take into account the fact that in complex diseases, cases with affected relatives are more likely to have a predisposing genetic variant than are cases without affected relatives.

- As a result, W_{QLS} can do worse than $W_{\chi^2_{corr}}$ for complex trait mapping, because it downweights related cases compared to cases with no relatives in the sample.

Case-control association testing with relatives
Approach 3: QL approach with better model

- Initial model (frequency $a + r$ in cases and a in controls, i.e. $\mu = a + rD_r$) was too simple to work well.

- We obtain slightly less simple model as follows:
  - Consider a two-allele trait model specified by an allele frequency a and penetrance parameters p_0, $p_1 = p_0 + \epsilon_1$, and $p_2 = p_0 + \epsilon_2$, where $p_i =$ P$\{$affected$|$ have i copies of allele$\}$.
  - Let $\nu_j = E(X_j | D_r)$, calculated under the two-allele model.
  - Let $\nu = E(X_j | D_{rj} = 1)$.
  - Let
    $$\tilde{D}_r = \lim_{\epsilon_1, \epsilon_2 \to 0} \frac{\nu_j - a}{\nu - a}.$$
  - Our model is $\mu = a + r\tilde{D}_r$.
  - Intuitively: $r \approx \nu - a$, where ν is frequency in cases (unconditional on affection statuses of relatives) and a is population frequency.

Case-control association testing with relatives Approach 3: QL approach with better model (cont.)

- Model $\mu = a + r\tilde{D}_r$, where $\tilde{D}_r = \lim_{\epsilon_1,\epsilon_2 \to 0} \frac{\nu_j - a}{\nu - a}$.
- $\tilde{D}_r$ turns out to have a simple form: $\tilde{D}_{rj} = L\delta$
- Here, L is the matrix defined before with $L_{ij} = 1 + h_i$ if $i = j$ and $2\phi_{ij}$ if $i \neq j$, with h_i and ϕ_{ij} denoting inbreeding and kinship coeffs, respectively.
- δ is the vector with jth entry = 1 if j is affected, $-K_p/(1 - K_p)$ if j is unaffected, and 0 if j's status is unknown (e.g. population-based control).
- K_p is the population prevalence of the trait; in practice, an estimate of K_p can be used, if available, or an arbitrary number could be used.
- The M_{QLS} is the QL score statistic based on this model: $M_{QLS} =$
$$\frac{[\delta^T(X - \hat{a}_0\mathbf{1})]^2}{\hat{a}_0(1 - \hat{a}_0)[\delta^T L\delta - (\delta^T\mathbf{1})^2(\mathbf{1}^T L^{-1}\mathbf{1})^{-1}]}$$
where $\hat{a}_0 = (\mathbf{1}^T L^{-1}\mathbf{1})^{-1}\mathbf{1}^T L^{-1} X$.

Case-control association testing with relatives
Approach 3: QL approach with better model (cont.)

- We do not believe the simple model.
- Nonetheless, the resulting statistic captures the property that in complex diseases, cases with affected relatives are more likely to carry a genetic variant predisposing to the trait than are cases without affected relatives.
- Misspecification of the population prevalence K_p has no effect on validity of the test.
- We use simulation studies to assess
  - power of the method for multilocus trait models (i.e. 2-allele model does not hold).
  - power when K_p is drastically misspecified

Simulations to assess power and Type I error

- 60 extended outbred pedigrees, each with 16 individuals in 3 generations, ascertained for multiple affecteds.

- 20 pedigrees with 4 affecteds each, 20 with 5, and 20 with 6

- Each individual from pedigree is included in study if at least half of his first-degree relatives are affected

- Control sample includes 200 unrelated unaffecteds in addition to the unaffected relatives in the pedigrees

- Models

  - Model I: 2 unlinked causal SNPs with epistasis; 2 penetrance params.

  - Model II: 2 unlinked causal SNPs with epistasis; 4 penetrance parameters

  - Model III: 3 unlinked causal SNPs with epistasis; 2 penetrance parameters

Power to Detect Association

(at .05 level, based on 5,000 simulated replicates)

	Estimated Power		
Model	$W_{\chi^2_{corr}}$	W_{QLS}	M_{QLS}
I-a	0.57	0.42	0.85
I-b	0.78	0.60	0.98
I-c	0.77	0.66	0.90
II-a	0.65	0.55	0.75
III-a	0.57	0.51	0.77
III-b	0.85	0.74	0.94

- Type I error verified at nominal level for each test.
- In each simulation
  - Use of the M_{QLS} drastically increases power.
  - More than 1 causal locus is involved in true model.
  - Assumptions used to derive M_{QLS} are false.
- M_{QLS} captures property that cases with affected relatives are enriched for predisposing variants.
- M_{QLS} seems to have high power under complex trait models.

Robustness of M_{QLS} to misspecification of K_p

Power of M_{QLS}

Assumed K_p	Multiple of True K_p	Estimated Power (s.e.)
0.039	1/2	0.78 (.006)
0.078	1	0.77 (.006)
0.156	2	0.79 (.006)
0.312	4	0.68 (.007)
0.390	5	0.46 (.007)

- Model has 3 unlinked causal SNPs with epistasis.

- True K_p is .078

- When assumed K_p is within a factor of 2 of true K_p, there is no change in power in this case.

- Power of M_{QLS} appears to be fairly robust to misspecification of K_p.

Example: Testing for Association With Alcoholism, Using Genome Screen Data

- Genetic Analysis Workshop (GAW) 14 data from Collaborative Study for the Genetics of Alcoholism (COGA)

- 143 pedigrees each with at least 3 affecteds

- 506 cases and 202 controls

- 10,081 autosomal SNPs analyzed

- In the M_{QLS}, set $K_p = .05$, an estimate from the National Institute on Alcohol Abuse and Alcoholism (NIAAA).

- Binary phenotype (Affected with ALDX1 or “pure unaffected”).

Most significant SNPs in COGA Data Set

(p-value $<$ 5.0e-5 by at least 1 test)

					p-value		
chr	pos.	n_{ca}	n_{co}	p	$W_{\chi^2_{corr}}$	W_{QLS}	M_{QLS}
16	59.8	383	137	0.85	1.0e-4	1.2e-4	$\underline{\text{2.5e-7}}$
6	153.7	397	144	0.85	1.6e-2	3.3e-1	$\underline{\text{9.2e-6}}$
3	158.2	350	142	0.82	1.1e-2	7.2e-3	$\underline{\text{2.3e-5}}$
7	123.7	419	159	0.84	1.8e-3	$\underline{\text{2.7e-5}}$	3.7e-2
18	104.7	266	111	0.62	3.3e-1	8.3e-1	$\underline{\text{4.1e-5}}$
18	95.8	394	143	0.67	8.7e-3	6.3e-3	$\underline{\text{4.6e-5}}$
1	188.1	477	183	0.92	3.1e-2	3.7e-2	$\underline{\text{4.9e-5}}$

Markers are tsc1750530, tsc1288916, tsc0175005, tsc0043946, tsc0046696, tsc0054146, and tsc0275539, respectively.

- SNP on chromosome 16 is genome-wide significant (p-value .008) after Bonferroni correction for 10,081 SNPs and 3 tests per SNP: 2.5e-7 $\times 10,081 \times 3 =$ 7.6e-3

- The most significant results are generally obtained with M_{QLS}.

Summary

- The QL framework provides a way to extend standard testing and estimation methods to the situation when sampled individuals are related.
- First moments under null and alternative and 2nd moments under null are required for our testing methods.
- Methods are very fast, even in complex inbred pedigrees.
- We currently use QL methods for a number of genetic problems when samples contain related individuals:
  - allele frequency estimation
  - Hardy-Weinberg testing
  - case-control association testing
    - allelic tests (← discussed today)
    - genotypic tests (analogue of Armitage trend test)
- For case-control association testing, use of modified QLS greatly improves power

Extensions

- Analyze quantitative traits
- Incorporate covariates
- Extension to haplotype analysis when complete haplotype information is not available

Acknowledgments

Collaborators:

Former post-doc: Catherine Bourgain (INSERM, Paris)

Former Ph.D. student: Timothy Thornton (U.C. Berkeley)

We thank COGA and GAW for permission to use the GAW 14 COGA data. Support from NIH grants HG001645 and DK55889 and from an Institut National de la Sante et de la Recherche Medicale postdoctoral fellowship is gratefully acknowledged.