Some Old and New Tests of Disease Association with Multiple SNPs in Linkage Disequilibrium

Wei Pan[1]

[1]Division of Biostatistics, School of Public Health

University of Minnesota

Dec 3, 2008

Outline

- Introduction: problem
- Review: some existing methods
- New methods: SumSq tests
  Some theory; numerical results...
- Discussion

Introduction

- Single Nucleotide Polymorphisms (SNP)
  Example:
  DNA seq 1 – AAGC**C**TA
  DNA seq 2 – AAGC**T**TA
  two alleles, C and T; 3 genotypes: CC, TT, CT;
  SNP: a minor allele freq (MAF) $\geq$ 5% (or 1% or ...).
- Problem: Genome-wide *association* studies (GWAS)
  Goal: to detect assoc b/w a phenotype (e.g. disease status) and genetic variants (e.g. SNPs);
  Ultimate goal: to detect *causal* genetic variants.
- As of 11/24/08, the Catalog of Published Genome-Wide Association Studies "includes 202 publications and 435 SNPs" that are associated with some phenotypes, such as prostate cancer, diabetes, bipolar disorder...

- Most common study design: case-control;
  n in thousands;
  hundreds of thousands SNPs (e.g. 500K Affy arrays);
  $OR: \sim 1.5$.

- Data:

Obs	Y	SNP1	...	SNP2	...	(SNP0)	...	SNPk
1	1	CT	...	AG	...	CG	...	AC
2	1	TT	...	AG	...	GG	...	AA
3	1	CT	...	AA	...	CG	...	CC
......								
1001	0	CT	...	AG	...	CC	...	AC
1002	0	TT	...	GG	...	CC	...	AC
1003	0	CC	...	GG	...	CC	...	CC
......								

- A binary response: $Y = 0$ or 1;
  each SNP j has up to 3 possible values; coded as $X_j = 0$, 1 or 2, though other codings are possible.
- The causal SNP0 may not be observed.
- Linkage disequilibrium (LD): SNP0 and its nearby SNPs are

correlated (and form an LD block).
$\Longrightarrow$ If SNP0 is causal, then its nearby SNPs are associated with Y!

- Statistical question: any SNP associated with Y? univariate or multivariate?

- Here we only consider $k > 1$ SNPs inside an LD block.

Existing methods

- Single-locus (or SNP-by-SNP or univariate) analysis:
  - Model: $Y \sim SNP_j$

    $$\text{Logit Pr}(Y_i = 1) = \beta_{M,0j} + X_{ij}\beta_{M,j}, \qquad (1)$$

  - $H_{0,j}$: $\beta_{M,j} = 0$ for each $j = 1, ..., k$
    $\Longrightarrow p_j$.
  - Combining: $p = \min(p_1, p_2, ..., p_k)$
    Need to do multiple test adjustment!
    Time-consuming with permutation, or conservative with Bonferroni method.
  - Model (1): as a 2×3 table; Cochran-Armitage trend test.

- Multivariate (or global or joint) analysis:
  - Model: $Y \sim SNP_1 + ... + SNP_k$

$$\text{Logit Pr}(Y_i = 1) = \beta_0 + \sum_{j=1}^{k} X_{ij}\beta_j, \tag{2}$$

  - H_0: $\beta_1 = ... = \beta_k = 0$
  - Use the score, Wald or LR test:
    $T_W = \hat{\beta}' V^{-1} \hat{\beta}$, $T_S = U' V_U^{-1} U \sim \chi^2_k$ under H_0;
    $V = Cov(\hat{\beta})$, $V_U = Cov(U)$;
    Possibly large $DF = k$!
  - Hotelling's T^2 test: similar to the above global test.

- Weight score test (WST) (Wang and Elston, 2007, AJHG):
  - High cost of multiple test adjustment or a large DF!
  - WST: 1) apply a Fourier transform on X;
    2) test on no assoc b/w each component and Y;
    3) form a weighted sum of the score stat's in 2).
  - worked well in their numerical examples.

- Sum test
  - *Working* assumption: $\beta_1 = ... = \beta_k \equiv \beta_c$.
    in general, *incorrect*!
  - Model:

$$\text{Logit Pr}(Y_i = 1) = \beta_{0,c} + \sum_{j=1}^{k} X_{ij}\beta_c = \beta_{0,c} + X_{i,c}\beta_c, \tag{3}$$

  - $H_{0,c}$: $\beta_c = 0$
  - Apply the score, Wald or LR test:
    $T_W = \hat{\beta}_c^2 / V_c \sim \chi_1^2$ under $H_{0,c}$.
  - Feature: DF=1; no multiple test!
  - Correct test size:
    $H_0 \Longrightarrow H_{0,c}$!
  - Power: simulation results; $n = 500 + 500$

Corr	OR	Sum	WST	L-G	T^2	U-P	Go-P
CS	1.0	.051	.053	.047	.049	.046	.047
	1.2	.098	.096	.059	.062	.072	.084
	1.4	.235	.226	.089	.093	.153	.206
	1.6	.395	.399	.145	.153	.239	.366
	1.8	.578	.578	.255	.262	.379	.530
	2.0	.711	.713	.357	.366	.480	.670
AR-1	1.0	.055	.048	.053	.054	.037	.049
	1.2	.132	.115	.078	.080	.107	.131
	1.4	.350	.315	.192	.194	.289	.354
	1.6	.599	.549	.361	.370	.504	.583
	1.8	.798	.743	.549	.560	.704	.796
	2.0	.895	.868	.726	.727	.845	.907

Corr	OR	Sum	WST	L-G	T^2	U-P	Go-P
Rand	1.0	.044	.043	.048	.051	.050	.048
	1.2	.134	.130	.078	.079	.087	.121
	1.4	.320	.318	.148	.153	.200	.290
	1.6	.546	.550	.243	.246	.360	.523
	1.8	.753	.748	.383	.391	.537	.729
	2.0	.863	.864	.530	.540	.688	.848

HapMap data for gene CHI3L2; #SNP=16:

n	OR	Sum	WST	L-G	T^2	U-P	Go-P
200	1.0	.050	.041	.094	.036	.053	.052
200	1.2	.181	.160	.142	.058	.169	.182
200	1.4	.521	.480	.292	.173	.483	.516
200	1.6	.803	.774	.521	.375	.764	.818
500	1.0	.051	.043	.074	.032	.054	.057
500	1.2	.387	.356	.188	.113	.333	.381
500	1.4	.886	.867	.606	.483	.886	.899
500	1.6	.994	.992	.927	.879	.997	.995

- What is β_c?
  Some average of $\beta_1, ..., \beta_k$? why?
- For linear models,

$$\hat{\beta}_c = (X_c'X_c)^{-1}1'(X'X)\hat{\beta},$$

$$(X_c'X_c)^{-1}1'(X'X)1 = 1,$$

- Why better? with collinearity,

$$Cov(\hat{\beta}) = \sigma^2(X'X)^{-1},$$

$$Var(\hat{\beta}_c) = \sigma^2(X_c'X_c)^{-1}.$$

- A limitation: $\hat{\beta}_c$ depends on the signs of $\hat{\beta}_j$'s!
  Codings of X_j's (vs $2 - X_j$'s) matter!
  A heuristic: flip the codings of X_j's to minimize # of negative pairwise correlations, but enough?
  Same with the WST.

HapMap CEU data for gene IL21R; #SNP=27:

n	OR	Sum	WST	L-G	T^2	U-P	Go-P
200	1.0	.046	.050	.098	.063	.057	.052
200	1.2	.078	.078	.107	.078	.087	.087
200	1.4	.204	.215	.200	.148	.256	.265
200	1.6	.351	.366	.344	.275	.500	.474
500	1.0	.050	.049	.054	.031	.055	.047
500	1.2	.165	.174	.142	.107	.183	.204
500	1.4	.432	.444	.408	.333	.652	.600
500	1.6	.607	.611	.717	.667	.908	.831

- Chapman and Whittaker (2008, *Genetic Epi*):
  1) The Sum test may not be good;
  2) The U-P and a test by Goeman et al (2006, JRSS-B) work best.

- Goeman's test:
  - Set-up: "large k, small n" as for microarray data;
  - Main idea:
    Prior for $\beta = (\beta_1, ..., \beta_k)'$: $E(\beta) = 0$, $Cov(\beta) = \tau^2 I$.
    Now test H_{0,τ^2}: $\tau^2 = 0$.
  - For logistic regression:
    $T_{Go} = \frac{1}{2}(U'U - \text{Trace}(I_f))$, where $U = X'(Y - \bar{Y}) = U_M$, and $I_f = Cov(U) = \bar{Y}(1 - \bar{Y})(X - \bar{X})'(X - \bar{X})$.
  - Null distribution unknown; use simulation or permutation.

- Why does Goeman's test work here ("large n, small k")?

New methods

- How to fix the problem?

$$\hat{\beta}_c = (X_c' X_c)^{-1} 1' (X'X) \hat{\beta} = \frac{(\sum_{i=1}^m X_{i1}^2, ..., \sum_{i=1}^m X_{ik}^2)\hat{\beta}_M}{\sum_{i=1}^m \left(\sum_{j=1}^k X_{ij}\right)^2}.$$

- Use squared $\hat{\beta}_{M,j}$'s:

$$SumSqB = \hat{\beta}_M' \hat{\beta}_M = \sum_{j=1}^k \hat{\beta}_{M,j}^2,$$

$$SumSqBw = \hat{\beta}_M' \mathrm{Diag}(V_M)^{-1} \hat{\beta}_M = \sum_{j=1}^k \hat{\beta}_{M,j}^2 / v_{M,j},$$

- Null distributions for $Q = \hat{\beta}_M' W^{-1} \hat{\beta}_M$:
  1) $W = I$ and $W = \mathrm{Diag}(V_M)$ in the above;
  2) $Q \sim \sum_{j=1}^k c_j \chi_1^2$, where c_j's are the eigen values of $V_M W^{-1}$;

3) Zhang (2005, JASA): approximate by $a\chi^2_d + b$ with

$$a = \frac{\sum_{j=1}^k c_j^3}{\sum_{j=1}^k c_j^2}, \quad b = \sum_{j=1}^k c_j - \frac{\left(\sum_{j=1}^k c_j^2\right)^2}{\sum_{j=1}^k c_j^3}, \quad d = \frac{\left(\sum_{j=1}^k c_j^2\right)^3}{\left(\sum_{j=1}^k c_j^3\right)^2}.$$

4) $Pr(SumSqB > s|H_0) \approx Pr\left(\chi^2_d > (s-b)/a\right)$.

- Analogs of the score test:

$$U_{M,j} = \sum_{i=1}^m X_{ij}(Y_i - \bar{Y}) = X'_{.j}(Y - \bar{Y}),$$

$$SumSqU = U'_M U_M = (Y - \bar{Y})' X X' (Y - \bar{Y}),$$

$$SumSqUw = U'_M \mathrm{Diag}(I_f)^{-1} U_M,$$

where $I_f = Cov(U_M) = \bar{Y}(1-\bar{Y})(X - \bar{X})'(X - \bar{X})$.

- Null distributions: approximated as before.

Simulation with CS; #SNP=10; $n = 500 + 500$:

					SumSq				
OR	Sum	L-G	U-P	Go-P	Bw	B	Uw	U	EMP
1.0	.051	.047	.046	.047	.044	.046	.044	.043	.045
1.2	.098	.059	.072	.084	.076	.076	.077	.080	.080
1.4	.235	.089	.153	.206	.198	.199	.199	.193	.199
1.6	.395	.145	.239	.366	.357	.363	.358	.356	.360
1.8	.578	.255	.379	.530	.518	.506	.518	.519	.520
2.0	.711	.357	.480	.670	.661	.657	.661	.662	.666

Simulation with AR-1; #SNP=10; $n = 500 + 500$:

					SumSq				
OR	Sum	L-G	U-P	Go-P	Bw	B	Uw	U	EMP
1.0	.055	.053	.037	.049	.047	.047	.048	.048	.049
1.2	.132	.078	.107	.131	.123	.123	.124	.125	.127
1.4	.350	.192	.289	.354	.354	.353	.354	.352	.357
1.6	.599	.361	.504	.583	.584	.583	.585	.577	.589
1.8	.798	.549	.704	.796	.782	.779	.783	.785	.785
2.0	.895	.726	.845	.907	.897	.891	.896	.901	.898

Simulation with corr randomly b/w 0.2–0.7; #SNP=10; $n = 500 + 500$:

					SumSq				
OR	Sum	L-G	U-P	Go-P	Bw	B	Uw	U	EMP
1.0	.044	.048	.050	.048	.044	.046	.044	.046	.046
1.2	.134	.078	.087	.121	.116	.113	.116	.114	.117
1.4	.320	.148	.200	.290	.279	.280	.281	.284	.281
1.6	.546	.243	.360	.523	.505	.510	.505	.500	.506
1.8	.753	.383	.537	.729	.716	.717	.718	.721	.720
2.0	.863	.530	.688	.848	.837	.835	.837	.836	.840

HapMap data for gene CHI3L2; #SNP=16:

					SumSq				
OR	Sum	L-G	U-P	Go-P	Bw	B	Uw	U	EMP
					$(n = 200)$				
1.0	.050	.094	.053	.052	.051	.049	.052	.053	.055
1.2	.181	.142	.169	.182	.177	.181	.177	.179	.180
1.4	.521	.292	.483	.516	.512	.513	.512	.513	.518
1.6	.803	.521	.764	.818	.814	.816	.813	.811	.818
					$(n = 500)$				
1.0	.051	.074	.054	.057	.056	.056	.056	.054	.057
1.2	.387	.188	.333	.381	.370	.376	.370	.370	.371
1.4	.886	.606	.886	.899	.901	.901	.901	.896	.901
1.6	.994	.927	.997	.995	.995	.997	.995	.994	.995

HapMap CEU data for gene IL21R; #SNP=27:

					SumSq				
OR	Sum	L-G	U-P	Go-P	Bw	B	Uw	U	EMP
					$(n = 200)$				
1.0	.046	.098	.057	.052	.046	.047	.047	.047	.048
1.2	.078	.107	.087	.087	.078	.078	.079	.084	.082
1.4	.204	.200	.256	.265	.260	.264	.265	.261	.267
1.6	.351	.344	.500	.474	.451	.457	.457	.464	.470
					$(n = 500)$				
1.0	.050	.054	.055	.047	.042	.045	.044	.042	.045
1.2	.165	.142	.183	.204	.207	.202	.208	.202	.211
1.4	.432	.408	.652	.600	.587	.582	.589	.594	.594
1.6	.607	.717	.908	.831	.833	.836	.836	.828	.839

- $SumSqB \approx SumSqU$ and $SumSqBw \approx SumSqUw$
  $\hat{\beta}_M = I_{M,d}^{-1} U_M + O_p(m^{-1})$.

- $SumSqB \neq SumSqB_w$ except $diag(V_M) \approx v\mathbf{1}$.

- Connection b/w SumSq and Goeman tests:

$$
\begin{aligned}
T_{Go} &= \frac{1}{2}(Y-\bar{Y})'XX'(Y-\bar{Y}) - \\
&\quad \frac{1}{2}\bar{Y}(1-\bar{Y})\text{Trace}((X-\bar{X})'(X-\bar{X})),
\end{aligned}
$$

  Conditional on Y the second term is fixed (i.e. non-random) and can be dropped:

$$T_{Go} = \frac{1}{2}U'U + c_0 = \frac{1}{2}U_M' U_M + c_0 \propto SumSqU.$$

  And $\hat{\tau}^2 = \sum_{j=1}^{k} \hat{\beta}_{M,j}^2 / k \propto SumSqB$.

- Why do they work?
  How could they beat “optimal” score, Wald and LR tests???

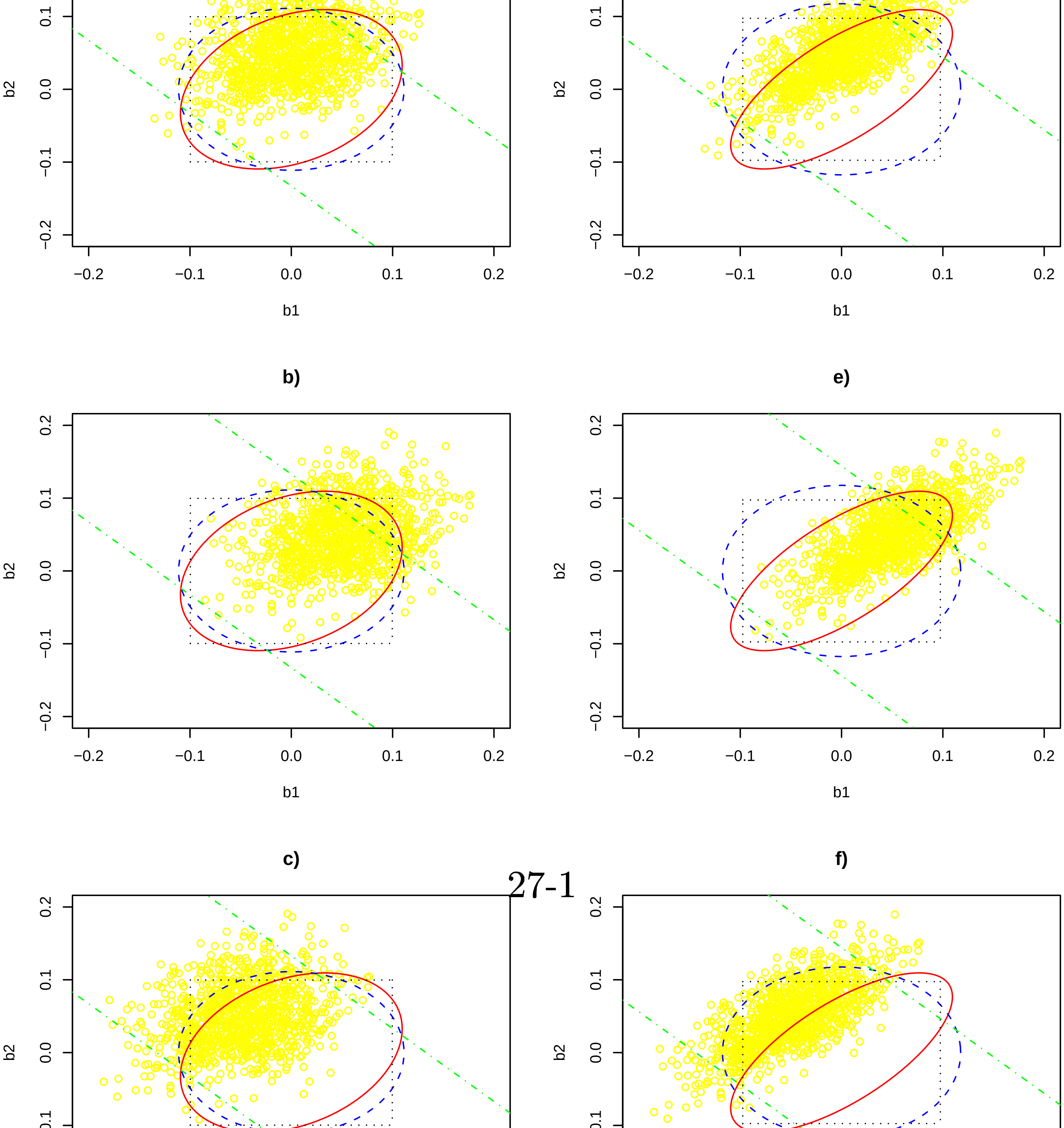
- Cox and Hinkley, *Theoretical Statistics*, 1974:
  - Optimality of the score, Wald and LR tests:
    locally most powerful, but only for ...;
    o/w, no uniformly most power (unbiased) (UMPU) test!
  - If we knew β, then
    $T_{MP} = \beta' U$, **but** ...
  - Try $\max_b b'U$ s.t. $Var(b'U) = b' I_f b = 1$?
- We estimate T_{MP} by
  $T_{EMP} = \hat{\beta}'_M U_M$.
- $T_{EMP} \approx SumSqUw = U'_M \text{Diag}(I_f)^{-1} U_M$ because

$$\hat{\beta}_M = I_{M,d}^{-1} U_M + O_p(m^{-1}). \tag{4}$$

- How about estimating β by $\hat{\beta}$?
  $T_{EMP,J} = \hat{\beta}' U \approx U' I_f^{-1} U$, which is ...

- Any intuitive explanation for using $diag(V)$ or I, not V?
- Is $\hat{V}$ problematic?
- Consider a simple situation:
  1) $k = 2$, $\beta = (\beta_1, \beta_2)'$;
  2) $\hat{\beta} \sim N(\beta, V)$;
  3) V **known**: $Var(\hat{\beta}_1) = Var(\hat{\beta}_2) = 1/500$, $corr(\hat{\beta}_1, \hat{\beta}_2) = \rho$;
  4) Test H_0: $\beta = 0$
- Compare 4 tests:
  1) Wald: $T_W = \hat{\beta}' V^{-1} \hat{\beta}$;
  2) SumSqB: $SumSqB = \hat{\beta}'\hat{\beta}$;
  3) univariate test: $Max = \max(|\hat{\beta}_1|, |\hat{\beta}_2|)$;
  4) Sum test: $Sum = \hat{\beta}_1 + \hat{\beta}_2$.
- Obtain their rejection regions: $R_T(c) = \{\beta : |T(\beta)| > c\}$ for test stat $T = T(\beta)$.
  numerically solve $\int_{R_T(c)} f_0(\beta) d\beta = \alpha$, thanks to Fang Han!

- Fig:



b2
b1
−0.2
−0.1
0.0
0.1
0.2
b)
e)
c)
f)

Empirical powers with $\alpha = 0.05$:

Set-up	ρ	β	Wald	SumSqB	Max	Sum
a	0.3	$(0, .05)'$	0.164	0.143	0.158	0.121
b	0.3	$(.05, .05)'$	0.226	0.258	0.242	0.312
c	0.3	$(-.05, .05)'$	0.373	0.239	0.274	0.059
d	0.7	$(0, .05)'$	0.263	0.102	0.158	0.133
e	0.7	$(.05, .05)'$	0.180	0.224	0.222	0.296
f	0.7	$(-.05, .05)'$	0.725	0.171	0.292	0.082

Discussion

- No UMPU test!
- A practical question: which one to use?
- Tried with real data (GAW16) and found that the univariate test, global/joint score (or Wald or LR) test, the sum test and SumSqU (or SumSqUw) could each have highest power, depending on chromosome regions.
- Use all of the above, then combine?
various combination methods; no uniform winner!
- Extended to haplotype analyses?
- Multiple unlinked loci and their interactions (*epistasis*)?
Use biological knowledge, e.g. gene networks (Pan 2008, *Hum Genet*).

- Main results applicable to other GLMs or regressions in general!
  Why do we always use the score/Wald/LR test in regression?
  They are **not** UMPU (though they are UMPI).
  Ignore correlations, as in the SumSq tests?
  Reduce # parameters, as in the sum test? Tukey's 1-DF test!

Acknowledgement: This research was supported by NIH.

You can download our papers from
http://www.biostat.umn.edu/rrs.php

Thank you!