Biostatistics 666
Problem Set 4
Due November 2, 2017

ApoE Haplotypes, LDL Cholesterol and Heart Disease.

Genetic variants in the ApoE gene have been associated with LDL cholesterol levels and susceptibility to heart disease.

a) **Suppose that genotypes for several markers in the ApoE gene were measured in a set of heart disease cases and controls. Sketch out a strategy for evaluating association between ApoE haplotypes and disease outcome.**

There are several possible strategies. One strategy we discussed would be to use maximum likelihood to calculate three sets of haplotype frequencies (in cases, in controls and in the combined sample) and three corresponding likelihoods (L_{cases}, $L_{controls}$, $L_{combined}$). Then, the statistic 2 ln ($\mathrm{L_{cases}}$ x $\mathrm{L_{controls}}$ / $\mathrm{L_{combined}}$) is approximately distributed as χ^2 with df equal to the number of haplotypes with nonzero frequencies.

b) **How could you ensure p-values calculated using the strategy you recommend above are accurate?**

One good strategy would be to shuffle case – control labels and repeat the analysis in each of the shuffled samples. We would then check how often χ^2 statistic in these shuffled samples exceeds that in the original data.

c) **Your colleagues estimated haplotype frequencies for the sample using Clark's algorithm for haplotype frequency estimation. Describe an advantage and a disadvantage of the algorithm.**

One advantage of Clark's algorithm is that it is relatively fast. As long as the number of haplotypes in a region is modest, the algorithm can often handle large numbers of markers – another attractive feature.

A major disadvantage of Clark's method is that, because it doesn't use haplotype frequency, it might resolve individuals with ambiguous haplotype configurations incorrectly.

d) **Propose an alternate and improved algorithm for haplotype frequency estimation. How does the algorithm improve upon Clark's method?**

Two possibilities would be to use an E-M algorithm (which would improve upon Clark's algorithm by also using frequency information) and to use an algorithm like Matthew Stephen's coalescent inspired algorithm (which would improve upon Clark's method both by using haplotype frequency information and by using haplotype similarities to resolve individuals with unique haplotypes [e.g. by favoring haplotypes unique haplotypes that are most similar to other previously seen haplotypes]).

e) **Suppose your colleagues have now measured ApoE genotypes in a population sample including 1,000 individuals for which measurements of LDL cholesterol levels are available. Sketch out a strategy for evaluating association between ApoE haplotypes and LDL cholesterol levels.**

We didn’t discuss this in class. When dealing with quantitative traits or wishing to incorporate covariates in the analysis, the most convenient strategy is to use regression based approaches. In these approaches, the trait of interest (say, disease status or quantitative trait levels) is usually the dependent variable, and predictors will include both relevant covariates (age, sex, other variables that are associated with the outcome of interest) and an haplotype dosage (which is the expected count of the haplotype being evaluated and is essentially the same fractional count of each haplotype in each individual used in the E-M algorithm).