

Semiparametric Methods for Gene-environment Case-control Studies

Raymond J. Carroll
Distinguished Professor
Department of Statistics
Texas A&M University

Abstract

We consider population-based case-control studies of gene and environment interactions using prospective logistic regression models. In a typical case-control study, neither the intercept of the logistic regression nor the population probability of disease can be identified. However, in many cases it is reasonable to assume that genotype and environment are independent in the population, possibly conditional on covariates to account for population stratification. In such a case, we show that the intercept and population probability of disease are identified. We develop a modern semiparametric likelihood approach for this problem, showing that it leads to much more efficient estimates of gene-environment interaction parameters and then gene main effect than the standard approach: decreases of standard errors for the former are often by factors of 50% and more. In addition, if the probability of disease is known in the population, we show efficiency gains for estimating gene-environment interactions, again in contrast to the standard approach. Multiple extensions are discussed, with applications to an important data set involving BRCA 1/2. The most important extensions are to the problems of missing genotype data (our example) and unphased haplotype data.

This is joint work with Nilanjan Chatterjee (National Cancer Institute).