

Case-Control Studies, Semiparametrics, Genetics and Interactions

Raymond J. Carroll
Department of Statistics
3143 TAMU
Texas A&M University
College Station TX 77843-3143
carroll@stat.tamu.edu
<http://www.stat.tamu.edu/~carroll>

Abstract:

Consider simple case-control sampling for logistic regression models in which predictors are called G and X , and in which strata are called S . In certain circumstances, it is reasonable to posit a model for the distribution of G given (X,S) *in the population*. For example, in genetics, a reasonable assumption is that the distribution of the gene G is independent of the environment X given the strata S of different populations. We show that assumptions such as stated above carry Fisher information about parameters in the original logistic model, especially about main effects for G and interactions between G and X . Our approach is via a semiparametric profiling approach based on nonparametric maximum likelihood, plus an old trick that is well-known in the survey sampling literature. Mean squared error efficiency improvements easily reach 400% for interactions. The resulting semiparametric profile likelihood is explicit, easily maximized, and handles missing data in G trivially: the latter point is of major interest in genetics where G might represent unphased haplotypes or missing genotypes. A detailed example involving missing genetic information, breast cancer and the BRCA1/BRCA2 mutation is discussed to illustrate the methods and its subtleties.

Co-Author:

This work is based on a series of papers with Nilanjan Chatterjee.