Retrospective Sampling and Gene-Environment Interaction Studies

Raymond J. Carroll
Department of Statistics, Nutrition and Toxicology
Texas A&M University
College Station TX 77843-3143 USA
http://www.stat.tamu.edu/~carroll
http://statbio.stat.tamu.edu

Abstract

Risks of complex diseases, such as cancers, are determined by both genetic and environmental factors. Advances in human genome research have thus led to epidemiologic investigations not only of the effects of genes alone, but also of their effects in combination with environmental exposures. The case-control study design, which has been widely used in classical questionnaire based epidemiologic studies, is now being increasingly used to study the role of genes and gene-environment interactions in the etiology of complex diseases.

The traditional approach for analysis of case-control studies is prospective logistic regression. Here the basis of inference is formed by the likelihood of the disease (D) outcome data conditional on covariate information (X) ignoring the fact that under the case-control sampling design data are observed on X conditional on D. It is known that such a prospective approach is actually equivalent to the retrospective maximum likelihood analysis that properly accounts for the case-control sampling design, provided that the distribution of the covariates are treated completely nonparametrically. It is also known that even in the presence of covariate missing data or/and measurement error, the prospective and retrospective maximum-likelihood methods for analyzing case-control studies are equivalent as long as the underlying model for the covariate distribution is nonparametric.

In studies of genetic epidemiology, it often may be reasonable to assume certain parametric or semi-parametric models for the covariate distribution in the underlying source population. For example, if G represents one of the three possible genotypes a subject can have at a particular bi-allelic locus, the population frequencies of the three genotypes could be specified in terms of the allele frequency of one of the alleles under the Hardy-Weinberg Equilibrium (HWE) assumption. Another assumption that is commonly invoked in practice is that genetic susceptibility and environmental exposures are independently distributed in the population. The prospective logistic regression analysis, being the semiparametric maximum likelihood solution for the problem that allows an arbitrary covariate distribution, clearly remains a valid option for analyzing case-control studies in such setting. However, retrospective methods that can exploit these various covariate distributional assumptions can be more efficient, sometimes dramatically so, especially for understanding the interaction between gene and the environment. Decreases in standard errors of 50% are common in this area.

This course is meant to provide an overview of the emerging area of gene-environment interaction studies, when parametric or semi-parametric models for the covariate distribution in the underlying source population are employed.

Outline

- Retrospective case-control studies
- Prospective analysis of retrospective studies: Cornfield (1956), Andersen (1970) and Prentice and Pyke (1979).
    - Consistency, efficiency and inference
- Gene-environment independence
    - Binary cases, rare disease assumption, case-only method
- Semiparametric retrospective likelihood methods
    - Derivation
    - Inference
    - Gains in efficiency over prospective methods
    - Rare disease simplifications
    - Application to the Israeli Ovarian Cancer Study
- Missing genotype data
- Haplotype analysis of retrospective case-control gene-environment data, a complex missing data setting
- Robustness against assumptions about the environmental variable
    - Modeling: allowing the haplotype distribution to depend on covariates
    - Model averaging in genotype studies
- Improving the prospective analysis of interactions via a Tukey-type score test