Semiparametric Analysis of Case-Control Studies, With Applications to Gene-Environment Interactions

Raymond J. Carroll

Department of Statistics

3143 TAMU, Texas A&M University, College Station TX 77843-3143

E-Mail: carroll@stat.tamu.edu

Abstract: Consider a standard retrospective case-control study involving interactions, with covariates (G, X) and logits of the form $\beta_0 + m(X, G, \beta_1)$ for an arbitrary known function $m(\cdot)$. If the function $m(\cdot)$ is known, and if the distribution of the covariates (X, G) is modeled nonparametrically, then only β_1 is identifiable and Prentice and Pyke (1979, Biometrika) showed that the semiparametric efficient estimator for β_1 is obtained via ordinary logistic regression, ignoring the case-control sampling scheme. The same result applies for partially linear models, with logits $m(G, \beta_1) + \theta(X)$, where $\theta(\cdot)$ is modeled nonparametrically. Indeed, if the distribution of (X, G) is nonparametric, inference that ignores the case-control sampling scheme and pretends that the observed data are from a prospective random sample is asymptotically efficient and asymptotically correct.

In genetic epidemiology, however, it is often reasonable to make assumptions about the distribution of the gene, G, given the environment X, while still treating the distribution of X nonparametrically. For example, suppose that G is binary (you have a mutation or you do not) while X is multivariate. There is a considerable methodological and applied literature that assumes that genetic status is independent of environment in the population. If you make this assumption, and no other, and then compute the semiparametric MLE for retrospective case-control data in the parametric model with logits $\beta_0 + m(X, G, \beta_1)$, the resulting efficient estimator is very different from ordinary logistic regression. In practical cases, we have seen decreases in standard errors for interactions that are a fraction of 2 or more. Put less precisely, the assumption that G and X are independent, without modeling either component, is effectively the same as having 4 times more data for certain parameters. Similar statements apply if the logistic model has nonparametric components, e.g., varying coefficient interaction models.

I will describe a very general theory when things are known about the distribution of G given X, and apply it to an example of testing whether

oral contraceptive use among those carrying the BRCA1/2 gene mutation is protective against cervical cancer.