Semiparametric Analysis of Polygenic Gene-Environment Interactions in Case-Control Studies with caseControlGE

Alex Asher
Texas A&M University

Abstract

Gene-environment interactions can be efficiently estimated in case-control data by exploiting the assumption of gene-environment independence in the source population, but until recently such techniques required parametric modelling of the genetic variables. The **caseControlGE** package implements the methods of Stalder, Asher, Liang, Carroll, Ma, and Chatterjee (2017, *Biometrika*, **104**, 801-812) and Wang, Asher, and Carroll (2018, unpublished), which exploit the assumption of gene-environment independence without placing any assumptions on the marginal distributions of the genetic and environmental variables. These methods are ideally suited for analysis of complex polygenic data for which parametric distributional models are not feasible. In addition to the two estimators, the package also supplies a function to simulate case-control data and several helper functions for use on model objects. Use of this package is illustrated using simulated data from a case-control study of breast cancer.

Keywords: case-control study; gene-environment interaction; genetic epidemiology; retrospective method; semiparametric analysis; pseudolikelihood; polygenic analysis.

1. Introduction

The **caseControlGE** package (Asher 2018) contains tools for the analysis of case-control data using R (R Core Team 2018). It implements the methods of Stalder *et al.* (2017) and Wang *et al.*, both of which fall under the class of semiparametric retrospective profile likelihood estimators. These methods are the first available to exploit the assumption of gene-environment independence while treating the genetic component nonparametrically. As such, they are well suited to replace logistic regression as the preferred method in situations where parametric distributional models are not feasible, such as in the analysis of complex polygenic data.

An important aspect of case-control studies is that the covariates are sampled conditional on the response, disease status. Given the genetic and environmental covariates G and E, we assume the risk of disease D in the underlying population follows the model

$$\text{pr}(D = 1 \mid G, X) = H\{\alpha_0 + m(G, X, \boldsymbol{\beta})\},$$

where $H(x) = \{1 + \exp(-x)\}^{-1}$ is the logistic distribution function and $m(G, X, \boldsymbol{\beta})$ is a function that describes the joint effect of G and X and is known up to the unspecified parameters of interest $\boldsymbol{\beta}$.

Given the retrospective nature of case-control sampling, it is surprising that standard prospective logistic regression can be used to obtain unbiased estimates of $\boldsymbol{\beta}$ (Prentice and Pyke 1979). Logistic regression requires no assumptions about the joint distribution of G and E, but it suffers from low power when estimating $G*E$ interaction effects. To gain efficiency, Chatterjee and Carroll (2005) exploited the assumption of gene-environment independence in the source population to maximize the retrospective likelihood while profiling out the distribution of E. Their method is available as the function `snp.logistic` in the *Bioconductor* package **CGEN** (Bhattacharjee, Chatterjee, Han, Song, and Wheeler 2012).

The method of Chatterjee and Carroll, and subsequent methods utilizing the same retrospective profile likelihood framework, require a parametric model for the distribution of G given E. This becomes difficult as the number and complexity of genetic variables in the model grows. Capitalizing on advances in high-throughput genomics, genome-wide association studies have identified scores of SNPs associated with complex diseases such as cancers and diabetes. Modern case-control studies of gene-environment interactions need efficient methodology that allows for a flexible and arbitrarily complex genetic component, such as multiple correlated SNPs and/or continuous polygenic risk scores (PRSs).

The method of Stalder *et al.* (2017) extends the retrospective profile likelihood framework of Chatterjee and Carroll, dispensing with the need to model G parametrically. When the population disease rate π_1 is known, the retrospective profile loglikelihood can be estimated (up to an additive constant) using just the case-control sample and without modeling the distribution of G. When π_1 is unknown but the disease is rare, estimates can be obtained using the *rare disease approximation* that $\pi_1 \approx 0$, which typically introduces negligable bias (Stalder *et al.* 2017).

The semiparametric method of Stalder *et al.* (2017) is implemented

-
-

These methods extend the retrospective profile likelihood framework developed by Chatterjee and Carroll (2005), which addresses the retrospective sampling scheme of case-control studies by which gains efficiency over logistic regression by exploiting the assumption of gene-environment independence in the population. The method of Chatterjee and Carroll

caseControlGE contains three main functions: `simulateCC`, `spmle`, and `spmleCombo`, as well as several helper functions. Section 2 of this paper introduces `simulateCC` in the context of simulating case-control data analogous to the data analyzed in Wang *et al.*. Section 3 introduces `spmle` as a tool to analyze the simulated data, and section 4 introduces `spmleCombo` to conduct a more efficient analysis of the simulated data.

2. Simulating case-control data with simulateCC

Wang *et al.* demonstrate the utility of their method

```
R> 1
```

```
[1] 1
```

3. Analyzing case-control data with spmle

```
R> 1
```

```
[1] 1
```

4. Analyzing case-control data with spmleCombo

References

Asher A (2018). *Semiparametric Gene-Environment Interactions in Case-Control Studies*. R package version 0.2.

Bhattacharjee S, Chatterjee N, Han S, Song M, Wheeler W (2012). *CGEN: An R package for analysis of case-control studies in genetic epidemiology*. R package version 3.6.2.

Chatterjee N, Carroll RJ (2005). "Semiparametric maximum likelihood estimation in case-control studies of gene-environment interactions." *Biometrika*, **92**, 399–418.

Prentice RL, Pyke R (1979). "Logistic disease incidence models and case-control studies." *Biometrika*, **66**, 403–411.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Stalder O, Asher A, Liang L, Carroll RJ, Ma Y, Chatterjee N (2017). "Semiparametric analysis of complex polygenic gene-environment interactions in case-control studies." *Biometrika*, **104**(4), 801–812.

Wang T, Asher A, Carroll RJ (????). "Improved semiparametric analysis of polygenic gene-environment interactions in case-control studies." *Unpublished*.

Affiliation:

Alex Asher
Texas A&M University
Department of Statistics
E-mail: alexasher@stat.tamu.edu