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 methods that assume 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Asher, Liang, Carroll, Ma, and Chatterjee (2017, *Biometrika*, **104**, 801-812) and Wang, Asher, and Carroll (2018, unpublished), which exploit the assumption of gene-environment independence without placing any assumptions on the marginal distributions of the genetic or environmental variables. These methods are ideally suited for analyzing 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 by simulating and analyzing data from a case-control study of breast cancer.

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

1. Introduction

1.1. caseControlGE package

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.* (2018), 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.

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.* (2018). 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.

1.2. Background

Case-control studies are retrospective observational studies in which the sample consists of a group of healthy subjects and a group of diseased subjects. A crucial aspect of the case-control design is that the outcome, disease status, is known *before* sampling. The ability to deliberately oversample diseased subjects makes the case-control design cost effective, which is why it is widely popular in studies of gene-environment interactions.

Given the genetic and environmental covariates G and E, we assume the risk of disease D in the underlying population follows the model

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

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.

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 negligible bias (Stalder *et al.* 2017).

Wang *et al.* (2018) proposed an improvement to the method of Stalder *et al.* (2017) that increases the efficiency of the estimates with no additional assumptions. This development relies on the observation that the method of Stalder *et al.* removes dependence on the distribution of the genetic and environmental variables in two different fashions; by treating the genetic and environmental variables symmetrically Wang *et al.* generate two sets of parameter estimates that are combined to generate a more efficient estimate.

1.3. Implementation

The semiparametric method of Stalder *et al.* (2017) is implemented as the function `spmle` in **caseControlGE**, detailed in section 3. Estimating the semiparametric profile likelihood is a computationally intensive process, and significant effort was invested in speeding up calculations. Estimation functions, including the analytic gradient and hessian, are written in C++ and compiled using **Rcpp** (Eddelbuettel 2013), providing a tremendous speedup over native R code. Extensive benchmarking and code profiling was conducted, and estimation functions were written to apply matrix operations to contiguous blocks of memory whenever possible, reducing memory latency and allowing modern processors to exploit data level parallelism and perform the same operation on multiple data points simultaneously.

The estimated semiparametric likelihood is maximized using the quasi-Newton optimizer **ucminf** (Nielsen and Mortensen 2016) using starting values from logistic regression. `ucminf` is particularly well suited for this application because it allows us to precondition the optimization with the analytic hessian, and it evaluates the gradient after each call to the objective function. Calculating the gradient along with the likelihood adds negligible computational complexity, so we call a single C++ function to compute them both, then return them separately to `ucminf`. This leads `ucminf` to converge in roughly half the time of the next-fastest optimizers (several of the various R implementations of the BFGS algorithm tie for second place). The unmatched speed of `ucminf` means we are willing to tolerate its bugs, which include occasionally declaring convergence before actually converging. To address this, `spmle` checks the gradient at the reported optimum and restarts the optimization if necessary (with different starting values).

Computational complexity of the asymptotic covariance estimation, which contains a sum of the form $\sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} \partial \mathcal{L}_{ijk}(\Omega)/\partial\Omega$, was reduced from $O(n^3)$ to $O(n^2)$ by storing intermediate values in a three-dimensional array. This increases speed at the cost of memory usage, which climbs from $O(n)$ to $O(n^2)$, setting a practical limit on sample size in the low tens of thousands for average personal computers. This is sufficient to analyze all but the largest case-control studies; covariance estimates for larger studies should be computed using the bootstrap.

Asymptotic covariance estimates for the Symmetric Combination Estimator of Wang *et al.* converge slowly and unreliable in practice, often providing poor coverage. Wang *et al.* recommend a balanced bootstrap, with cases and controls resampled separately, to estimate covariance. **caseControlGE** offers users with multicore computers the option to speed up computation by using multiple processors. Parallelization is implemented using the R base package **parallel**, which is installed by default on all operating systems. Parallelization on computers running Linux or macOS is done by forking the active R session, saving time and memory. This option is unavailable in Windows, so parallelization is fractionally slower because a PSOCK cluster is created with a new instance of R running on each core.

2. Simulating case-control data with simulateCC

2.1. Data description

Wang *et al.* (2018) demonstrate the utility of their method by analyzing data from a case-control study of breast cancer. This case-control sample is taken from a large prospective cohort at the National Cancer Institute: the Prostate, Lung, Colorectal and Ovarian (PLCO) cancer screening trial (Canzian, Cox, Setiawan, Stram, Ziegler, Dossus, Beckmann, Blanché, Barricarte, Berg *et al.* 2010). The case-control study analyzed by Wang *et al.* consists of 658 cases and 753 controls sampled from a cohort of 64,440 non-Hispanic, white women aged 55 to 74, of whom 3.72% developed breast cancer (Pfeiffer, Park, Kreimer, Lacey Jr, Pee, Greenlee, Buys, Hollenbeck, Rosner, Gail *et al.* 2013). The data are available from the National Cancer Institute via a data transfer agreement, but cannot be distributed with the **caseControlGE** package. Fortunately, we can use the **caseControlGE** function `simulateCC` to generate a similar data set for analysis.

Each of the 1411 subjects in the PLCO sample was genotyped for 21 SNPs that have been previously associated with breast cancer based on large genome-wide association studies. These SNPs were weighted by their log-odds-ratio coefficients and summed to define a polygenic risk score (PRS). A standardized version of this PRS, with mean zero and standard deviation one, was used as the genetic risk factor G by Wang *et al.*. Early menarche is a known risk factor for breast cancer, and Wang *et al.* used a binary indicator of whether the subject underwent early menarche as E (age at menarche < 14). Several environmental variables were recorded as part of the PLCO study, including body mass index (BMI). In our simulation, we will consider BMI in addition to the variables modeled by Wang *et al.*.

2.2. Data simulation

To determine the appropriate distributions to use when simulating G and E, we examine the PLCO data. In doing so, it is important to keep in mind that the case control sample is not representative of the source population. Case-control studies deliberately oversample cases, so the distribution of G and E in the sample may be quite different from the distribution of G and E in the population (especially for variables that are strongly correlated with disease status). `simulateCC` works by simulating a population using user-specified parameters, then selecting `ncase` cases and `ncontrol` controls as the case-control sample. To accurately simulate the genetic and environmental variables from the PLCO study, we need to estimate their distributions *in the source population.*

The simplest and most common way to estimate population parameters is to calculate them using just the controls. Case-control designs are typically used to study relatively rare diseases, and the bias introduced by using the cases as a stand-in for the population is usually quite small. This *rare disease approximation* of population parameters is the best possible estimate when the true population disease rate π_1 is unknown or not well estimated. When π_1 is known, it is possible to calculate unbiased estimates by weighting the cases and controls by π_1 and $(1 - \pi_1)$, respectively. (This technique is employed to great effect by Stalder *et al.*, and is the reason that `spmle` requires the user to specify a value for `pi1`.)

Wang *et al.* report $\beta_G = 0.459$ with $p < 1e - 4$, but they standardized G to mean zero and standard deviation one *in the case-control sample.* G has a strong positive effect on disease risk, indicating that the distribution of $(G|D = 0)$ is meaningfully different from the distribution of $(G|D = 1)$. Specifically, $\mathbf{E}(G|D = 1) > \mathbf{E}(G|D = 0)$. With a population disease rate of 0.0372, this implies $\mathbf{E}_{\text{pop}}(G) \approx \mathbf{E}(G|D = 0) < 0$, where the subscript pop emphasizes that the expectation is in the source population. This causes no problem for Wang *et al.*, but it presents us with a dilemma: $G \nsim \mathrm{N}(0, 1)$. If we simulate $G \sim \mathrm{N}(0, 1)$ and use $\beta_G = 0.459$ as reported in Wang *et al.*, we

should try to approximate the d

To perform a realistic simulation we

Rather than trying

The entire population is used to calculate disease prevalence (reported when `control$trace > -1`).

yields the information necessary to simulate the genetic and environmental variables.

- G is a standardized polygenic risk score - simulate from N(0,1)

```
R> 1

[1] 1
```

2.3. Confirming the G-E independence assumption

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.

Canzian F, Cox DG, Setiawan VW, Stram DO, Ziegler RG, Dossus L, Beckmann L, Blanché H, Barricarte A, Berg CD, *et al.* (2010). "Comprehensive analysis of common genetic variation in 61 genes related to steroid hormone and insulin-like growth factor-I metabolism and breast cancer risk in the NCI breast and prostate cancer cohort consortium." *Human Molecular Genetics*, **19**(19), 3873–3884.

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

Eddelbuettel D (2013). *Seamless R and C++ Integration with Rcpp*. Springer, New York. doi:10.1007/978-1-4614-6868-4. ISBN 978-1-4614-6867-7.

Nielsen HB, Mortensen SB (2016). *ucminf: General-Purpose Unconstrained Non-Linear Optimization*. R package version 1.1-4, URL https://CRAN.R-project.org/package=ucminf.

Pfeiffer RM, Park Y, Kreimer AR, Lacey Jr JV, Pee D, Greenlee RT, Buys SS, Hollenbeck A, Rosner B, Gail MH, *et al.* (2013). "Risk prediction for breast, endometrial, and ovarian cancer in white women aged 50 y or older: derivation and validation from population-based cohort studies." *PLoS Medicine*, **10**, e1001492.

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 (2018). "Improved semiparametric analysis of polygenic gene-environment interactions in case-control studies." *To Appear*.

Affiliation:

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