Journal of Statistical Planning and Inference 156 (2015) 31–47

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Semiparametric nonlinear regression for detecting gene and environment interactions

Shujie Ma [a,*], Shizhong Xu [b]

[a] Department of Statistics, University of California at Riverside, Riverside, CA 92521, United States
[b] Center for Plant Cell Biology, University of California at Riverside, Riverside, CA 92521, United States

ARTICLE INFO

Article history:
Received 10 February 2014
Received in revised form 10 August 2014
Accepted 10 August 2014
Available online 3 September 2014

Keywords:
G×E interactions
Nonlinearity
Semiparametric models
B-splines
Profile estimation
Score test

ABSTRACT

It is commonly accepted that gene and environment (G×E) interactions play a pivotal role in determining the risk of human diseases. In conventional parametric models such as linear models and generalized linear models which are applied frequently to study statistical interactions, effects of covariates are decomposed into main effects and interaction effects (products of two components). Such decomposition, however, may not reflect the true interaction effect of gene and environment. In this paper, we propose a semiparametric regression approach to capture possible nonlinear G×E interactions. A profile quasi-log-likelihood estimation method is applied with asymptotic consistency and normality established for the profile estimators. Moreover, we develop Rao-score-type test procedures based on the profile estimation for regression parameters and nonparametric coefficient functions, respectively. Our models and methods are illustrated by both simulation studies and analysis of a dataset application.

1. Introduction

It is commonly accepted that most human diseases result from a complex interaction between genetic and environmental factors, such as obesity (Hebebrand and Hinney, 2009), psychiatric disorders (Tsuang et al., 2004; Caspi and Moffitt, 2006), heart disease (Talmud, 2007), diabetes (Grarup and Andersen, 2007) and cancer (Song et al., 2011). By learning how genetic and environmental factors jointly influence the risk of developing a human disease, it would help scientists to develop new methods for prevention and treatment of illnesses. Despite the enthusiasm for investigation of gene–environment (G×E) interactions, published works on studying these interactions via statistical modeling are limited. In the literature, linear models as well as generalized linear models such as logistic and log-linear models are used frequently to study statistical interactions. In the conventional parametric models, the effects of covariates are decomposed into main effects and interaction effects (products of two components). Such decomposition, however, may not reflect the true nonlinear interaction between gene and environment. As a result, mis-specification in parametric models could lead to a large estimation bias. To overcome this limitation, different non- and semi-parametric modeling methods have been recently applied to study G×E interactions. For example, Chatterjee and Carroll (2005) and Chen et al. (2012) studied semiparametric maximum likelihood estimates of logistic regression parameters in case-control studies. Maity et al. (2009) developed a score test for parametric main effects of genetic factors in a semiparametric model with Tukey's form of interaction, and Wei et al. (2011) derived a generalized likelihood ratio test for nonparametric effects. Lobach et al. (2011) and Ahn et al. (2013)

* Corresponding author.
E-mail addresses: shujie.ma@ucr.edu (S. Ma), shizhong.xu@ucr.edu (S. Xu).

http://dx.doi.org/10.1016/j.jspi.2014.08.005
0378-3758/

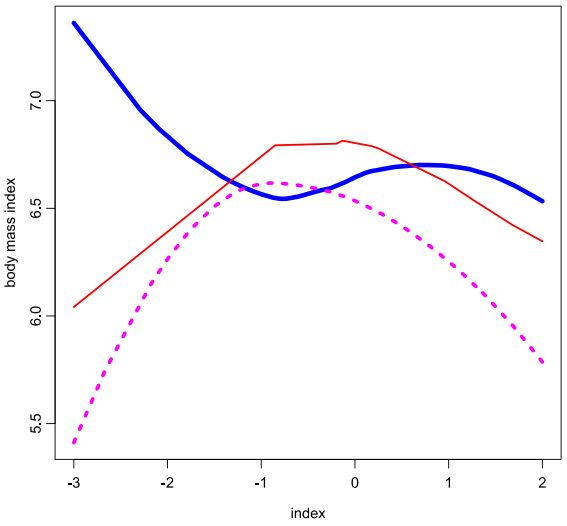

Fig. 1. Plots of the estimated BMI against the index value $U = \widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{X}$ for the three groups with genotype *aa* (thick line), *Aa* (thin line) and *AA* (dashed line) of SNP ss66155100.

proposed semiparametric Bayesian analysis of G×E interaction. Ma et al. (2011) applied a varying-coefficient model for G×E interaction.

In this paper, we apply a generalized partially linear single-index coefficient model (GPLSiCM) to study nonlinear G×E interactions. Let $\mathbf{Z} = (Z_1, Z_2, \ldots, Z_p)^{\mathrm{T}} = (1, Z_2, \ldots, Z_p)^{\mathrm{T}}$ be the p-dimensional vector, where $(Z_2, \ldots, Z_p)^{\mathrm{T}}$ are the genetic factors. A parametric model for studying genetic effects on phenotype is given as $E(Y|\mathbf{Z}) = \mu(\mathbf{Z}) = g^{-1}(\eta)$ with $\eta = \sum_{\ell=1}^{p} \beta_\ell Z_\ell$, where Y is the response variable which can be continuous or discrete such as binary variables or counts, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ is the regression coefficient vector, and g is a known monotone link function. By considering effects of genetic factors interacting with environmental factors, we may employ the parametric model with both main and interaction effects given as

$$\eta = \beta_1 + \sum_{k=1}^{d_1} \beta_{k1} X_k + \sum_{\ell=2}^{p} \beta_\ell Z_\ell + \sum_{\ell=2}^{p} \sum_{k=1}^{d} \beta_{k\ell} X_k Z_\ell, \tag{1}$$

where $(\beta_1, \beta_{k1}, \beta_\ell, \beta_{k\ell})$ are the regression coefficients and $\mathbf{X} = (X_1, \ldots, X_{d_1})^{\mathrm{T}}$ is the d_1-dimensional vector of the environmental factors. By simple calculation, (1) can be written as

$$\eta = \sum_{\ell=1}^{p} \left(\beta_\ell + \sum_{k=1}^{d_1} \beta_{k\ell} X_k \right) Z_\ell.$$

Thus the coefficient for the ℓth genetic factor, which is $\beta_\ell + \sum_{k=1}^{d_1} \beta_{k\ell} X_k$, is indeed a linear function of $(X_1, \ldots, X_{d_1})$.

The linearity assumption can be easily violated due to the underlying nonlinear mechanism of the relationship between the response and explanatory variables. For example, research on causes of obesity has brought tremendous attention due to its high prevalence and the associated medical and psychosocial risks. It is known that obesity is related to not only genetic factors but also some environmental factors such as sleeping hours (Knutson, 2012) and physical activity (Wareham et al., 2005). Thus, people may wonder how genetic and environmental factors together influence people's weight. Using data from the Framingham Heart Study (Dawber et al., 1951), we let X_1 = sleeping hours per day, X_2 = hours of light activity, and X_3 = hours of moderate activity, be the environmental factors. The body mass index (BMI) is used as the response variable. After deleting missing data, 299 subjects remain in our study. To illustrate possible G×E interactions, we divide people in the study into three groups based on the three genotypes of SNP ss66155100. For each group, a linear model $E(Y|\mathbf{X}) = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}$ can be fitted. However, in order to study possible nonlinear relationship between Y and $\mathbf{X}$, we let $E(Y|\mathbf{X})$ be a nonlinear function of $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}$, such that $E(Y|\mathbf{X}) = m\left(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}\right)$, where m is an unknown but smooth function, and $U = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}$ is the index. This model is the semiparametric single-index model, estimation of which has been substantially studied, see Carroll et al. (1997) for the backfitting method, Liang et al. (2010) for the profile method and Xia and Härdle (2006) for the minimum average variance estimation (MAVE) method. Fig. 1 shows the plots of the estimates of $m(\cdot)$ against the index value by using the np package in R for the three groups with genotypes *aa* (thick line), *Aa* (thin line) and *AA* (dashed line). Here *A* is the minor allele. We can clearly observe different nonlinear change patterns of the estimated BMI mean functions among the three groups. For example, for people with genotype *aa*, their BMI shows a decreasing pattern as index value increases. However, for people