SEMIPARAMETRIC BAYESIAN ANALYSIS OF GENE-ENVIRONMENT INTERACTIONS

Iryna Lobach
New York University
New York, NY
iryna.lobach@nyumc.org

Abstract

A key component to prevention and control of complex diseases, such as cancer, diabetes, hypertension, is to analyze the genetic and environmental factors that lead to the development of these complex diseases. We propose a Bayesian approach for analysis of gene-environment interactions that efficiently models information available in the observed data and a priori biomedical knowledge.

1 Introduction

The analysis of gene-environment interactions is complicated by the fact that many variables of interest to biomedical researchers, such as dietary intake and cigarette smoking are very difficult to measure on individuals. For example, in large epidemiologic studies of an impact of diet on development of a disease, nutrient intake is commonly measured using the food frequency questionnaire (FFQ). It is well known that the FFQ as a measure of long-term diet is subject both to biases and random errors [1]. The measurement error causes bias in estimates of gene-environment interactions [2] thus masking the features of the data and hence leads to loss of power. The loss of power prevents researchers from detecting important relationships among variables [3]. Massive measurement error is well known [4] to result in skewed sampling distribution of parameter estimates and the skewness is more pronounced for small sample sizes. Hence conventional estimation and inference based on Normality assumption are not precise. The Bayesian approach offers an advantage by specifying a priori distribution that can shrink the parameter estimates towards the prior thus bringing the sampling distribution of parameter estimates closer to Normal [5].

Conventionally, case-control data are analyzed using prospective logistic regression ignoring the fact that under this design subjects are sampled into the study conditionally on their disease status. [2] and [5] developed an efficient semiparametric pseudo-likelihood approach for analysis of case-control studies. The form of this pseudo-likelihood function offers several advantages. One is that it allows to incorporate information about the probability of disease, what cannot be done in the conventional analysis. Further, the formulation of the pseudo-likelihood function does not require specification of the distribution of environmental variables measured exactly. These variables include age, ethnicity, bmi and other demographic and clinical measurements. Thus gains in efficiency can be achieved by not having to model a distribution of a multivariate vector of these measurements. Validity of the Bayesian analysis needs to be examined when the proposed likelihood function is not a proper likelihood.

2 Semiparametric Pseudo-Likelihood

Let D be the categorical indicator of disease status and let $D = 0$ denote the disease-free (control) subjects and $D = 1$ - the diseased (case) subjects. Suppose there are I genetic markers are spanning the genomic region of interest and define G be the observed genetic markers. Denote (X, Z) denote all of the environmental (non-genetic) covariates of interest with X denoting the factors susceptible to measurement error. Given the environmental covariates X and Z and genetic data G, the risk of the disease in the underlying population is given by the polytomous logistic regression model.

$$\mathrm{pr}(D = 1|G, X, Z) = \frac{\exp\{\beta_0 + m(G, X, Z, \beta)\}}{1 + \exp\{\beta_0 + m(G, X, Z, \beta)\}}. \tag{1}$$

Here $m()$ is a known function parameterizing the joint risk of the disease from G, X and Z in terms of the odds-ratio parameters β. For the i-th marker, denote the two alleles by M_i and m_i, with frequencies P_{M_i} and P_{m_i}. Define the dummy variables that model genetic effect in the following form.

$$A_i \;=\; \begin{cases} 1 & \text{if } G_i = M_iM_i \\ 0 & \text{if } G_i = M_im_i \\ -1 & \text{if } G_i = m_im_i \end{cases}, \quad B_i \;=\; \begin{cases} -P_{m_i}^2 & \text{if } G_i = M_iM_i \\ P_{M_i}P_{m_i} & \text{if } G_i = M_im_i \\ -P_{M_i}^2 & \text{if } G_i = m_im_i \end{cases}. \tag{2}$$

The following specification of the risk function models both additive and dominance effects of genotype, as well as the multiplicative gene-environment interaction.

$$\begin{aligned} m_k(G, X, Z; \beta) = \quad & X\beta_{kX} + Z\beta_{kZ} + \sum_{i=1}^{I} A_i\beta_{kAi} + \sum_{i=1}^{I} XA_i\beta_{kAXi} + \sum_{i=1}^{I} ZA_i\beta_{kAZi} \\ & + \sum_{i=1}^{I} B_i\beta_{kDi} + \sum_{i=1}^{I} XB_i\beta_{kDXi} + \sum_{i=1}^{I} ZB_i\beta_{kDZi}. \end{aligned} \tag{3}$$

The regression coefficients β_{kA_i} and β_{kD_i} model risk due to the additive and dominance effect, respectively [5]. The remaining terms capture the multiplicative gene-environmental interaction. Form (3) of the risk function offers an advantage, namely that the linkage disequilibrium (the genetic term used to describe dependence between genetic markers that are located closely) is captured in the regression coefficients [5]. The model (1)-(3) cannot be used directly for analysis since the covariate X is measured with error. Let W denote the error-prone version of X. We assume a parametric model for the measurement error process of the form $f_{\mathrm{mem}}(w|D, G, X, Z; \xi)$.

Let n_0 and n_1 be the number of control and case subjects, respectively. In addition, let us denote $\pi_k = \mathrm{pr}(D = k), k = 0, 1$. Consider a sampling scenario where each subject from the underlying population is selected into the case-control study using a Bernoulli sampling scheme, where the selection probability for a subject given his/her disease status $D = 1$ is proportional to $\mu = n_1/\pi_1$. Let $R = 1$ denote the indicator of whether a subject is selected in the sample. Let us denote $\kappa = \beta_0 + \log(n_1/n_0) - \log(\pi_1/\pi_0)$. In addition, let $\Omega = (\widetilde{\beta}_0^{\mathrm{T}}, \beta^{\mathrm{T}}, \Theta^{\mathrm{T}}, \widetilde{\kappa}^{\mathrm{T}})^{\mathrm{T}}$, $\mathcal{B} = (\Omega^{\mathrm{T}}, \eta^{\mathrm{T}})^{\mathrm{T}}$ and $\upsilon = (\eta^{\mathrm{T}}, \xi^{\mathrm{T}})^{\mathrm{T}}$.

Define

$$S(k,g,x,z;\Omega) \;=\; \frac{\exp\left[1_{(k=1)}(k)\left\{\kappa + m(g,x,z;\beta)\right\}\right]}{1+\exp\left\{\beta_0 + m_j(g,x,z;\beta)\right\}}\mathrm{pr}(g;\theta).$$

Motivated by [5], we employ the following pseudo-likelihood function in place of the likelihood function. Note that by design the data are collected retrospectively (case-control sampling design), but the pseudo-likelihood is describing the data as if they were coming from a random sample.

$$\begin{aligned} L_{Pseudo}(k,g,w,z;\Omega,\eta,\xi) &\equiv \mathrm{pr}(D=k,G=g,W=w|Z=z,R=1) \\ &= \frac{\int S(k,g,x,z;\Omega)f_{\mathrm{mem}}(w|k,g,x,z;\xi)f_X(x|z;\eta)dx}{\sum_{k^*}\sum_{g^*}\int S(k^*,g^*,x,z;\Omega)f_X(x|z;\eta)dx}. \quad (4)\end{aligned}$$

Note that this pseudo-likelihood function does not assume any distribution on the environmental variables measured exactly, Z, thus creating a semiparametric feature of the model.

3 Bayesian Analysis

Validity of the Bayesian analysis based on (4) needs to be examined when the proposed likelihood function is not a proper likelihood. [6] proposed a numeric technique that can be used to validate our Bayesian approach under this pseudo-likelihood function and exploit it to draw inference about parameters based on the posterior distribution. Due to the complexity of the pseudo-likelihood function, the posterior distribution of the parameters is not in explicit form, therefore Markov Chain Monte Carlo (MCMC) algorithms are required to sample from this posterior distribution to make necessary inference. The joint posterior distribution for the MCMC calculations can be written in the following form.

$$\begin{aligned} &L_{Pseudo}(k,g,w,z;\Omega,\eta,\xi)\times|\Sigma_{\mathcal{B}}|^{-1/2}\exp\left\{-\frac{1}{2}(\mathcal{B}-\mu_{\mathcal{B}})^{\mathrm{T}}\Sigma_{\mathcal{B}}^{-1}(\mathcal{B}-\mu_{\mathcal{B}})\right\} \\ &\times\eta_2^{-1/2}\exp\left\{-(\mu_x-\eta_1)^2/(2\eta_2)\right\}\times(\sigma_x^2)^{-A-1}\exp\left\{-B/\sigma_x^2\right\}\prod_{i=1}^{I}\mathcal{I}_{(0,1)}(\theta_i).\end{aligned}$$

4 Simulation Experiments

To illustrate performance of the proposed method we performed a simulation study. The genetic information was simulated according to the Hardy-Weinberg Equilibrium for two marker loci $P_{M_i}=0.25,\ i=1,2$. The environmental covariate (X) is binary and measured with error with misclassification probabilities with misclassification probabilities being 0.20 for exposed and 0.25 for non-exposed subjects. The results are based on 500 replicates of 1000 cases and 1000 controls. Simulation results presented in Table 1 illustrate that the Naive approach that ingores existance of measurement error results in biased parameter estimates, while the proposed approach eliminates bias and results in parameter estimates that are less variable.

Table 1: Bias and Root Mean Squared Errors (RMSE) of the Naive approach that ingores existance of the measurement error and the proposed method.

		Naive Approach		Proposed Method	
Parameter	True Value	Bias	RMSE	Bias	RMSE
β_0	-5.000	-0.071	0.037	-0.023	0.013
β_X	1.099	0.083	0.035	0.003	0.027
β_{A1}	0.693	-0.057	0.017	-0.009	0.013
β_{A2}	0.000	-0.073	0.025	0.003	0.015
β_{AX1}	0.693	0.175	0.071	0.006	0.021
β_{AX2}	0.693	0.118	0.051	0.005	0.023

5 Discussion

We proposed a semiparametric Bayesian approach for the analysis of gene-environment interactions to address a difficult but common situation when the environmental exposure is measured with error. The proposed approach was successfully applied to the Colorectal Adenoma Analysis.

References

[1] Subar A.F., Kipnis V., Troiano R.P., et al. (2003). Using intake biomarkers to evaluate the extent of dietary misreporting in a large sample of adults: The Observing Protein and Energy Nutrition (OPEN) study. *American Journal of Epidemiology*. Vol. **158** pp. 1-13.

[2] Lobach I., Carroll R. J., Spinka C., Gail M. H., and Chatterjee N. (2008). Haplotype-based regression analysis of case-control studies with unphased genotypes and measurement errors in environmental exposures. *Biometrics*. Vol **64** pp. 673-684.

[3] Carroll R. J., Ruppert D., Stefanski L. A. and Crainiceanu C.M. (2006). Measurement Error in Nonlinear Models, Second Edition. *Chapman & Hall CRC Press.*

[4] Schafer D. W. and Purdy K. G. (1996). Likelihood analysis for errors-in-variables regression with replicate measurements. *Biometrika*. Vol. **83**, pp. 813-824.

[5] Lobach I., Fan R. and Carroll R.J. (2010) Genotype-Based Association Mapping of Complex Diseases: Gene-Environment Interactions with Multiple Genetic Markers and Measurement Errors in Environmental Exposures. *Genetic Epidemiology*

[6] Monahan J. F. and Boos D. D. (1992). Proper likelihood for Bayesian analysis. *Biometrika*. Vol. **79(2)**, pp. 271-8.