Published in final edited form as:
J Am Stat Assoc. 2013 ; 108(502): . doi:10.1080/01621459.2013.793121.

Unified Analysis of Secondary Traits in Case-Control Association Studies

Arpita Ghosh[1], **Fred A. Wright**[2], and **Fei Zou**[2]
[1]Public Health Foundation of India, New Delhi, India
[2]Department of Biostatistics, University of North Carolina at Chapel Hill, North Carolina, USA

Abstract

It has been repeatedly shown that in case-control association studies, analysis of a secondary trait which ignores the original sampling scheme can produce highly biased risk estimates. Although a number of approaches have been proposed to properly analyze secondary traits, most approaches fail to reproduce the marginal logistic model assumed for the original case-control trait and/or do not allow for interaction between secondary trait and genotype marker on primary disease risk. In addition, the flexible handling of covariates remains challenging. We present a general retrospective likelihood framework to perform association testing for both binary and continuous secondary traits which respects marginal models and incorporates the interaction term. We provide a computational algorithm, based on a reparameterized approximate profile likelihood, for obtaining the maximum likelihood (ML) estimate and its standard error for the genetic effect on secondary trait, in presence of covariates. For completeness we also present an alternative pseudo-likelihood method for handling covariates. We describe extensive simulations to evaluate the performance of the ML estimator in comparison with the pseudo-likelihood and other competing methods.

2 Introduction

The retrospective case-control design is one of the most important tools for epidemiology, and for rare diseases/traits may offer tremendous savings in time and expense compared to a prospective design. Even so, case-control designs remain costly, and efficiency is further maximized by gathering additional clinical phenotypes/traits for the sampled individuals. We refer to the dichotomous case/control variable as the *primary* trait (or phenotype), and other traits, which may be discrete or continuous, as secondary. Methods for analyzing such data have received considerable recent attention due to the availability of genome-wide association (GWA) datasets, which often follow a case-control design and include numerous secondary traits, which may be correlated with the primary trait (Frayling et al., 2007; Weedon et al., 2008; Spitz et al., 2008). In such studies the risk variables of main interest are genotypes of single nucleotide polymorphisms (SNPs), with possible additional covariate effects. The approaches described here apply generally to secondary analysis of case-control data, but the notation and examples are applicable to genetic association studies.

It is widely understood that, for case-control designs, a (prospective) analysis which ignores the sampling scheme yields consistent estimates of the odds ratio for disease risk (e.g., using logistic regression models (Prentice and Pyke, 1979)). For the analysis of secondary traits within a case-control design, a number of approaches have been described (Jiang et al., 2006; Lin and Zeng, 2009; He et al., 2012; Wang and Shete, 2011, 2012). The existing methods can be broadly divided as follows: a) the naïve method of analyzing the combined sample of cases and controls, without accounting for the case-control ascertainment; b) performing analysis within case and control groups; c) combining estimates from the case and control groups either by meta-analysis with inverse variances as weights or by using case/control status as a covariate; d) correcting bias using weighting schemes originally developed for sample surveys; and d) explicitly accounting for the case/control sampling scheme via a retrospective likelihood.

The naïve method can lead to severely biased estimates of the risk effects for the secondary traits, except under restrictive conditions (Nagelkerke et al., 1995; Jiang et al., 2006; Lin and Zeng, 2009; Monsees et al., 2009). Simple methods which account for the biased sampling include adjusting for the disease status in a regression model or restricting the analysis to cases or controls only. For rare diseases, the controls-only analysis is approximately unbiased, but may be highly inefficient. If the association between the primary and the secondary traits does not depend on the SNP genotype, then the cases also give a valid risk effect estimate and can be combined with the controls-only estimate via a inverse-variance meta-analytic procedure to provide a weighted estimate with improved efficiency. Li et al. (2010) describe an adaptively weighted method for rare diseases with binary secondary trait and genotype, that further combines the controls-only and the weighted estimates. Recently, they proposed another adaptive procedure to analyze secondary phenotypes for data from a case-control study of a primary disease that is not rare (Li and Gail, 2012).

The survey approach uses weights inversely proportional to the sampling fractions (Jiang et al., 2006; Scott and Wild, 2002). Richardson et al. (2007) and Monsees et al. (2009) use this standard survey-weighted approach, applied to the analysis of binary secondary traits, termed the inverse-probability-of-sampling-weighted (IPW) regression by Monsees et al. (2009). Technically this approach requires knowledge of the case-control sampling fractions, but the disease prevalence is often available externally and may be used as in Wang and Shete (2011).

In general, most of the approaches described above result in inconsistent risk estimates, although the bias may be low under some specific assumptions. The efficiency of the procedures also varies widely (Jiang et al., 2006). An alternative approach to account for the sampling mechanism is to use the retrospective likelihood, explicitly conditioning on the sampling scheme (Jiang et al., 2006; Scott and Wild, 1997b, 2001, 1991; Lee et al., 1998; Lin and Zeng, 2009; He et al., 2012). The retrospective likelihood models the joint distribution for the primary and secondary traits as a function of the genotype and other covariates. There are a number of attractive features to this approach. If the model is correct, then it will provide large-sample optimality for the maximum likelihood (ML) estimates. Provided the model is sufficiently rich, it enables partitioning of the correlations between primary and secondary phenotypes into the portion due to the risk variable (genotype), as well as a residual portion that may be due to other genes or environment. In addition, standard semi-parametric maximum likelihood (SPML) approaches are available to handle covariates in a flexible manner, with minimal assumptions for the potentially complex interplay of covariates with other data. For this reason the retrospective approaches are collectively termed "SPML" in the taxonomy of Jiang et al. (2006).

Although the motivation for the retrospective likelihood is clear, specific implementations vary. One approach is to factor the joint distribution into the marginal for the secondary trait and the conditional for the primary trait given secondary (SPML2 in Jiang et al., 2006). For example, Lin and Zeng (2009) handle both binary and continuous secondary traits by modeling the disease status given the secondary trait as a logistic regression. Wang and Shete (2011) use this joint model and apply a method of moments approach to produce bias-corrected odds ratio estimates for binary secondary traits using prevalence estimates for the primary and secondary traits from the literature. Recently they have shown that their method is robust even when there is an interactive effect of the SNP and secondary phenotype on the primary disease risk (Wang and Shete, 2012). He et al. (2012) have proposed a Gaussian copula-based approach that models the joint distribution in terms of the marginals for the primary and secondary phenotypes (SPML3 in Jiang et al., 2006) and uses the multivarite normal distribution to build in correlation between the phenotypes. Their method can handle multiple correlated secondary phenotypes.

Despite all of these efforts, a number of deficiencies remain. Most retrospective likelihood approaches do not incorporate interaction between the genetic variant and the secondary trait on primary disease risk. Other than the Gaussian copula approach (He et al., 2012), retrospective methods do not generally preserve the marginal logistic model for the disease trait, creating a contradiction between the primary trait risk estimates obtained from the marginal model and from the joint primary-secondary analysis. Additionally, the nonparametric handling of continuous covariates remains challenging; when there are multiple covariates, including one or more continuous ones, direct maximization of the likelihood is infeasible.

We propose an approach that specifies the joint distribution of primary and secondary traits in terms of the marginals for the two traits, with terms to govern their association (SPML3 in Jiang et al., 2006). Our framework enables association testing for both binary and continuous secondary traits, while respecting the desired logistic model for the primary trait and standard marginal models for the secondary trait. We demonstrate how this approach can incorporate covariates, and easily allow for interaction between the genetic variant and the secondary trait on primary disease risk. To handle the computational complexity introduced by the covariates, we reparameterize the profile likelihood, and derive a closed form expression which provides ML estimates of risk effects. For completeness, we briey describe a pseudo-likelihood approach that bypasses the need to involve the potentially high-dimensional covariate distribution. We perform extensive simulations to evaluate the performance of our profile likelihood method and compare it with the pseudo-likelihood and other competing methods.

The remainder of the paper is organized as follows. In Section 3 we lay down the details for our proposed profile likelihood-based method. In particular, in Subsection 3.1 we describe our joint model for the primary and secondary traits and in the following subsection we provide details of the estimation procedure. In Section 4 we demonstrate the performance of the proposed method as compared to other competing methods via a real data example and simulations. Section 5 concludes with some comments on related work and future directions. In Section 7 we expand on the theoretical details.

3 Methods

Let D denote the disease status or the primary trait (0=control, 1=case), Y the secondary trait, G the genotype at a biallelic locus, and $\mathbf{Z}$ the vector of covariates. Under additive model, G represents the number of minor alleles (0, 1, or 2) at the locus; under dominant (recessive) model, G denotes whether the individual carries at least one minor allele (two minor alleles). We randomly sample n_0 and n_1 individuals from the controls ($D = 0$) and the cases ($D = 1$) in the population respectively, and observe their Y-, G-, and $\mathbf{Z}$-values. The log-likelihood for our case-control sampled data (d_u, y_u, g_u, $\mathbf{z}_u$), $u = 1, 2, \dots, n$, takes the retrospective form $\sum_{u=1}^{n} \log P(y_u, g_u, \mathbf{z}_u | d_u)$ where

$$\log P(y_u, g_u, \mathbf{z}_u | d_u) = \log P(d_u, y_u | g_u, \mathbf{z}_u) + \log P(g_u, \mathbf{z}_u) - \log P(d_u).$$

We model *P*(*d*, *y*|*g*, **z**) parametrically, the regression *P*(*y*|*g*, **z**) being of primary interest. Note that *P*(*g*, **z**), the joint distribution for *G* and **Z** in the population, cannot be ignored, since

$$P(d)=\sum_{y,g,\mathbf{z}} P(d,y,|g,\mathbf{z})P(g,\mathbf{z}).$$

We deal with *P*(*g*, **z**) nonparametrically, thereby taking a semi-parametric approach to modeling the likelihood. In the following subsections we describe the joint model for the bivariate response (*D*, *Y*) and the estimation procedure.

3.1 Joint modeling of the primary and secondary traits

In modeling the joint distribution for the traits we try to preserve the marginal logistic model typically assumed for the original case-control trait *D*. We specify the bivariate distribution, *P*(*D*, *Y*|*g*, **z**), by parametrically modeling the marginals for the primary and secondary traits and also building a parametric model for their association given the genotype and the covariates. The natural choice for the marginal distribution for disease status is logistic, and that for the secondary trait is logistic or normal depending on whether the trait is binary or continuous respectively. For binary *Y*, the Palmgren model (Palmgren, 1989), the Bahadur model (Bahadur, 1959), and models based on copula theory (Meester and MacKay, 1994) have been used previously to specify joint distributions for bivariate binary responses. But for continuous *Y*, there is no standard bivariate distribution that yields logistic and normal marginals for *D* and *Y* respectively.

3.1.1 Binary secondary trait—The bivariate logistic model, considered by Palmgren (1989), has been used previously to model the joint distribution for correlated binary data (Jiang et al., 2006; Lee et al., 1998) and is conceptually very simple. It is based on the fact that the joint distribution of two binary variables can be specified in terms of their marginal probabilities and their odds ratio. Thus, for a randomly sampled individual in the population we specify the joint distribution of *D* and *Y* given *g* and **z** as

$$logit P(D=1|g,\mathbf{z})=\alpha_1+\beta_1 g+\gamma_{\mathbf{1}}'\mathbf{z},\ \ logit P(Y=1|g,\mathbf{z})=\alpha_2+\beta_2 g+\gamma_{\mathbf{2}}'\mathbf{z},$$

and $\log\, OR(D,Y|g,\mathbf{z})=\log\frac{P(D=1,\,Y=1|g,\mathbf{z})P(D=0,\,Y=0|g,\mathbf{z})}{P(D=1,\,Y=0|g,\mathbf{z})P(D=0,\,Y=1|g,\mathbf{z})}=\alpha_3+\beta_3 g$.

We are interested in inferring about β_2, the risk effect for the secondary trait.

3.1.2 Continuous secondary trait—For continuous *Y*, we consider joint models for (*D*, *Y* |*g*, **z**) such that the marginal distributions for *D* and *Y* given *g* and **z** correspond to the standard logistic and linear regressions respectively, that is,

$$logit P(D=1|g,\mathbf{z})=\alpha_1+\beta_1 g+\gamma_{\mathbf{1}}'\mathbf{z}\ \text{and}\ Y|g,\mathbf{z}\sim N(\alpha_2+\beta_2 g+\gamma_{\mathbf{2}}'\mathbf{z},\sigma_2^2).$$

To come up with a joint model that complies with the above marginals, we start with the following bivariate normal distribution,

$$\left(\begin{array}{c} V \\ Y \end{array}\right)\Bigg| g,\mathbf{z}\sim N\left(\left(\begin{array}{c}\mu_1\\ \mu_2\end{array}\right),\left(\begin{array}{cc}1 & \rho\sigma_2\\ \rho\sigma_2 & \sigma_2^2\end{array}\right)\right),$$

where $\mu_1=\alpha_1+\beta_1 g+\gamma_{\mathbf{1}}'\mathbf{z}$, $\mu_2=\alpha_2+\beta_2 g+\gamma_{\mathbf{2}}'\mathbf{z}$, and $\log\frac{1+\rho}{1-\rho}=\alpha_3+\beta_3 g$. We then introduce another latent variable to produce the logistic marginal. To transform the normal variate *V*, we define $U=\mu_1+\log\frac{\Phi(V-\mu_1)}{1-\Phi(V-\mu_1)}$. It follows that the density of *U* given *g* and **z** is logistic with location parameter μ_1 and scale parameter 1,

$$p_{U|G,\boldsymbol{Z}}(u|g,\mathbf{z})=\frac{\exp(-(u-\mu_1))}{(1+\exp(-(u-\mu_1)))^2}, -\infty<u<\infty.$$

We then threshold *U* at 0 to derive *D*. That is, *D*, defined as

$$D=\begin{cases}1, & U\geq 0\\ 0, & U<0\end{cases},$$

follows

$$logit P(D=1|g,\mathbf{z})=\alpha_1+\beta_1 g+\gamma_{\mathbf{1}}'\mathbf{z}.$$

As in the binary case, β_2 is our parameter of interest.

3.2 Estimation of β_2

3.2.1 ML estimate in absence of covariates—Let us first consider the situation without any covariates. The retrospective log-likelihood is

$$l=\sum_{u=1}^{n}\log P(y_u,g_u|d_u)=\sum_{u=1}^{n}\log P(d_u,y_u|g_u)+\sum_{u=1}^{n}\log P(g_u)-\sum_{i=0}^{1}n_i\log P(D=i). \quad (1)$$

We assume Hardy-Weinberg equilibrium and parameterize the genotype distribution in terms of the minor allele frequency, δ. We use θ to denote the set of parameters describing the joint distribution for *D* and *Y* given *g*. We write θ as $(\theta_1, \theta_2)'$, where $\theta_1 = (\alpha_1, \beta_1, \gamma_{\mathbf{1}})'$ parameterizes the disease model. Then the log-likelihood in (1) can be written as

$$l=\sum_{u=1}^{n}\log P(d_u,y_u|g_u;\theta)+\sum_{u=1}^{n}\log P(g_u;\delta)-\sum_{i=0}^{1}n_i\log\sum_{k=0}^{1}P(D=i|G=k;\theta_1)P(G=k;\delta). \quad (2)$$

We maximize (2) with respect to $\eta = (\theta, \delta)'$ to derive $\hat{\eta}$, the ML estimate for η. We can use the standard Newton-Raphson method to obtain the ML estimate iteratively, or other optimization tools such as quasi-Newton methods, Nelder-Mead simplex algorithm for derivative-free maximization, simulated annealing. Specifically, using Newton-Raphson method, we update the parameter in the $(k + 1)^{th}$ iteration by

$$\hat{\eta}^{(k+1)}=\hat{\eta}^{(k)}-\left\{\frac{\partial^2}{\partial\eta\partial\eta'}l(\hat{\eta}^{(k)})\right\}^{-1}\frac{\partial}{\partial\eta}l(\hat{\eta}^{(k)}).$$

The inverse of the observed information matrix, $\left\{-\frac{\partial^2}{\partial\eta\partial\eta'}l(\hat{\eta})\right\}^{-1}$, gives standard error estimate for $\hat{\eta}$.

3.2.2 ML estimate via profiling in presence of covariates—Let us now consider the situation where the joint model for the disease and the secondary trait involves covariates. The retrospective log-likelihood is

$$l=\sum_{u=1}^{n}\log P(d_u,y_u|g_u,\mathbf{z}_u;\theta)+\sum_{u=1}^{n}\log P(g_u,\mathbf{z}_u)-\sum_{i=0}^{1}n_i\log\sum_{g,\mathbf{z}}P(D=i|g,\mathbf{z};\theta_1)P(g,\mathbf{z}). \quad (3)$$

We can reasonably assume that *G* and **Z** are independently distributed in the underlying population, but must somehow account for the distribution of **Z**. Since the covariate structure is usually far too complicated to model parametrically (considering that it may be correlated with the primary trait), we want to make no assumptions about the form of the covariate distribution.. For a single binary covariate parameterized by ψ, we can easily derive the ML estimate by maximizing the retrospective likelihood with respect to (η, ψ). However, this approach is infeasible for a continuous covariate as the ML estimate involves maximization with respect to a high-dimensional nuisance parameter. In the following section we describe a computational technique to derive a closed-form expression, which may be thought of as an approximate profile likelihood up to one nuisance parameter, that upon maximization provides the ML estimate for the parameters of interest.

Let {$\mathbf{z_1}, \mathbf{z_2}, \ldots, \mathbf{z_L}$} represent unique values of **Z** in the case-control sample. We parameterize the distribution function for **Z** in terms of the probability masses {$\psi_1, \psi_2, \ldots, \psi_L$} that we assign to {$\mathbf{z_1}, \mathbf{z_2}, \ldots, \mathbf{z_L}$}. The retrospective log-likelihood in (3) can now be written as

$$l=\sum_{u=1}^{n}\log P(d_u,y_u|g_u,\mathbf{z}_u;\theta)+\sum_{u=1}^{n}\log P(g_u;\delta)+\sum_{u=1}^{n}\log P(\mathbf{z}_u;\boldsymbol{\psi})-\sum_{i=0}^{1}n_i\log\sum_{k=0}^{1}\sum_{l=1}^{L}P(D=i|G=k,\mathbf{z}_l;\theta_1)P(G=k;\delta)\psi_l. \quad (4)$$

When ψ is high-dimensional, maximizing the log-likelihood with respect to (η, ψ) can be a daunting task. We take the approximate profile likelihood approach described by Chatterjee and Carroll (2005) to handle the nuisance parameter ψ (hereafter dropping the “approximate” qualifier). To obtain the overall ML estimate for η, we maximize the profile log-likelihood, *l*(η) = *sup*$_\psi$*l*(η, ψ) with respect to η. Following Scott and Wild (1997a, 2001) and Chatterjee and Carroll (2005), we show in the Appendix (Section 7.1) that the profile log-likelihood *l*(η) can be equivalently expressed as

$$l^*(\eta,\kappa)=\sum_{u=1}^{n}\log P^*_{d_u,y_u,g_u}(\mathbf{z}_u;\eta,\kappa)$$

where

$$P^*_{d_u,y_u,g_u}(\mathbf{z}_u;\eta,\kappa)=\frac{P(d_u,y_u|g_u,\mathbf{z}_u;\theta)P(g_u;\delta)\kappa^{d_u}}{\sum_{i=0}^{1}P(D=i|\mathbf{z}_u;\theta_1)\kappa^i}.$$

The parameter κ represents the ratio of the sampling fractions,

$$\kappa=\frac{n_1/P(D=1)}{n_0/P(D=0)},$$

and satisfies the equation $\frac{\partial l^*(\eta,\kappa)}{\partial\kappa}=0$. We thus have effectively reduced the number of parameters from (p + L − 1) to (p + 1), with p used to denote the dimension of η. We provide a closed form expression, l*(φ), that on maximization with respect to φ = (η, κ)′ gives $\hat{\eta}$, thereby bypassing the need to numerically maximize the likelihood with respect to a high-dimensional nuisance parameter.

Although l*(φ) is not a true log-likelihood the relevant asymptotic theory for $\hat{\varphi}$ can be obtained by working with $S^*(\boldsymbol{\phi})=\frac{\partial}{\partial\boldsymbol{\phi}}l^*(\boldsymbol{\phi})$, referred to as the "pseudo" score-equations by Scott and Wild (2001). Under the assumption that n goes to infinity with n_1/n and n_2/n remaining fixed, we show in the Appendix (Section 7.2) that $\sqrt{n}(\hat{\phi}-\phi)$ converges in distribution to a normal random vector with mean zero and covariance matrix

$$n\,Cov(\hat{\phi})=n(J^*(\boldsymbol{\phi})^{-1}-J^*(\boldsymbol{\phi})^{-1}\Gamma(\boldsymbol{\phi})J^*(\boldsymbol{\phi})^{-1}),$$

where

$$J^*(\boldsymbol{\phi})=-E\left(\frac{\partial}{\partial\boldsymbol{\phi}'}S^*(\boldsymbol{\phi})\right)$$

and

$$\Gamma(\boldsymbol{\phi})=\left(\frac{1}{n_0}+\frac{1}{n_1}\right)E_{\mathbf{z}}\left(f(\mathbf{z})\frac{\partial}{\partial\boldsymbol{\phi}}P_1^*(\mathbf{z})\right)E_{\mathbf{z}}(f(\mathbf{z})\frac{\partial}{\partial\boldsymbol{\phi}'}P_1^*(\mathbf{z}))$$

with $P_1^*(\mathbf{z})=\frac{P(D=1|\mathbf{z})\kappa}{\sum_i\kappa^iP(D=i|\mathbf{z})}$ and $f(\mathbf{z})=(\frac{n_1}{\kappa}+n_0)\sum_i\kappa^iP(D=i|\mathbf{z})$. In practice, we can use

$$J^*(\hat{\phi})^{-1}=-\left.\frac{\partial^2}{\partial\boldsymbol{\phi}\partial\boldsymbol{\phi}'}l^*(\boldsymbol{\phi})\right|_{\hat{\phi}}^{-1},$$

the inverse of the observed information matrix based on l*, to estimate $Cov(\hat{\varphi})$. Since Γ(φ) ≥ 0, $J^*(\hat{\varphi})^{-1}$ provides a conservative estimate for $Cov(\hat{\varphi})$. Note that if the disease prevalence is known or well-estimated, we fix κ at its true value and work with l*(η) to obtain $\hat{\eta}$ and its standard error.

3.2.3 Pseudo-likelihood estimate—Motivated by the fact that β_2, the parameter of interest, appears only in the first term of the expression for the log-likelihood in (3), one might attempt to handle the covariates via a previously described pseudo-likelihood approach (Gong and Samaniego, 1981). For the sake of completeness, we lay down the framework for this approach to estimate β_2 and its standard error, and evaluate its performance via simulations. The retrospective log-likelihood in (4) can be viewed as

$$l=l_1(\theta_1,\theta_2)+l_2(\theta_1,\delta,\boldsymbol{\psi}), \quad (5)$$

where

$$l_1(\theta_1,\theta_2)=\sum_{u=1}^{n}\log P(d_u,y_u|g_u,\mathbf{z}_u;\theta_1,\theta_2), \quad (6)$$

and

$$l_2(\theta_1,\delta,\boldsymbol{\psi}=\sum_{u=1}^{n}\log P(g_u;\delta)+\sum_{u=1}^{n}\log P(\mathbf{z}_u;\boldsymbol{\psi})-\sum_{i=0}^{1}n_i\log\sum_{k=0}^{1}\sum_{l=1}^{L}P(D=i|G=k,\mathbf{z}_l;\theta_1)P(G=k;\delta)\psi_l.$$

As β_2 is of primary interest, we treat θ_1 as nuisance parameter. Gong and Samaniego proposed first obtaining an estimate for θ_1. When disease prevalence is known, we derive an estimate for θ_1, say $\tilde{\theta}_1$, from $l_2(\theta_1, \delta, \psi)$ (in general different from the ML estimate $\hat{\theta}_1$ derived from l). We then plug this estimate into $l_1(\theta_1, \theta_2)$ to obtain the pseudo log-likelihood

$$\tilde{\mathrm{l}}_1(\theta_2)=\sum_{u=1}^{n}\log P(d_u,y_u|g_u\mathbf{z}_u;\tilde{\theta}_1,\theta_2). \quad (7)$$

The pseudo-likelihood estimate, $\tilde{\theta}_2$, is obtained by maximizing $\tilde{l}_1$ with respect to θ_2. We discuss the asymptotic properties of the pseudo-likelihood estimate and provide formulae for its variance estimation in the Appendix (Section 7.3).

4 Results

We present a data example and simulation results to demonstrate the performance of the profile likelihood-based **ML** estimate in comparison with the **pseudo**-likelihood estimate and other competing methods: the **naïve** method ignoring the sampling mechanism, the **cases**-only and **controls**-only methods, the **weighted** method combining the cases-only and controls-only estimators via inverse-variance, the **adaptively weighted** method, proposed by Li et al. (2010), combining the controls-only and weighted estimators, the **adjusted** method including the disease status as a covariate in the regression for the secondary trait, and finally the **survey** approach using weights inversely proportional to the sampling fractions. All analyses were performed in `R v.2.15.0`. In the table and figures we use ‘Wtd’, ‘Awtd’, and ‘Adj’ to denote the weighted, the adaptively weighted, and the adjusted estimates respectively.

4.1 Data example

We reanalyze the data on colorectal cancer, smoking, and N-acetyltransferase 2 (*NAT2*) presented in Li et al. (2010). The authors compare different analysis methods for a binary secondary trait using case-control data for colorectal adenoma described in Moslehi et al. (2006). Moslehi et al. explore how variants in *NAT1* and *NAT2* genes affect the smoking-colorectal cancer relationship using cases with colorectal adenoma and controls selected from the screening arm of the Prostate, Lung, Colorectal and Ovarian (PLCO) cancer screening trial. Of 42,037 participants who provided a blood sample, 4,834 were excluded. 772 cases were randomly selected from 1234 participants who had at least one advanced colorectal adenoma detected at baseline screen and 777 gender and age-matched controls were sampled from 26,651 participants with a negative baseline sigmoidoscopy screening. Li et al. use this case-control data for colorectal adenoma to analyze the effect of *NAT2* gene on smoking (secondary trait). In Table 1 we display the data analyzed in Li et al. (2010) and in Table 2 we illustrate our re-analysis of the data.

There are no covariates here, so the ML estimate based on the Palmgren model is easily derivable from the retrospective likelihood, without requiring additional profiling or pseudo-likelihood. We used a prevalence of 0.04 (=1234/(1234+26651)) based on Moslehi et al. (2006). Under known prevalence, the ML based on the Palmgren model and the survey approach are the same. Li et al., in their paper, treat the robust controls-only estimate as the gold standard and show that the adaptively weighted estimate is similar to the controls-only estimate. The authors point out that the non-zero interaction between *NAT2* and smoking on colorectal adenoma risk results in high bias for the commonly used methods, apparent in the table for the naïve, the cases-only, the weighted, and the adjusted estimates. In contrast, our results show that the ML/survey estimate is similar to the adaptively weighted and controls-only estimates, with the ML/survey estimate enjoying smaller standard error.

4.2 Simulations

We compare performances of the different methods for both binary and continuous secondary traits, for biologically plausible scenarios, by generating data from the corresponding bivariate models (see Methods). We consider a biallelic SNP with an additive mode of inheritance and a minor allele frequency of 0.25. For demonstration purposes we include two covariates, one following a Bernoulli distribution with 0.45 success probability and the other a standard normal variate. For our simulations we consider the SNP genotype and the covariates as independent in the population. We fix the disease prevalence at 5% and the mean or prevalence of the secondary trait at 20%. Rather than specifying α_1 and α_2 directly, it is more interpretable to solve for them for specified values of the prevalences and other parameters. We fix β_1 at 0.25 (odds ratio = 1.28). We consider $\beta_2 = 0$ to reect the null and compare Type I error for different methods. To simulate data under alternative hypotheses and examine power, we fix β_2 at either −0.25 or 0.25 for the binary secondary trait; for the continuous setup we use −0.15 and 0.15. We assign the values $(-0.94, -0.37)'$ and $(-1.82, 0.30)'$, each generated independently from the standard normal, to the parameters γ_1 and γ_2, respectively. We fix α_3 at 1.0 and examine β_3 ranging from −2.0 to 2.0. We use $\sigma_2 = 1.0$ for the continuous scenario. For fixed prevalences, α_1 and α_2 are calculated according to the following equations:

$$P(D=1)=\int_{z_2}\sum_{g,z_1}\frac{exp(\alpha_1+\beta_1 g+\gamma_1'\mathbf{z})}{(1+exp(\alpha_1+\beta_1 g+\gamma_1'\mathbf{z}))}\begin{pmatrix}2\\g\end{pmatrix}\delta^g(1-\delta)^{(2-g)}0.45^{z_1}(1-0.45)^{(1-z_1)}\phi(z_2)dz_2,$$

$$P(Y=1)=\int_{z_2}\sum_{g,z_1}\frac{exp(\alpha_2+\beta_2 g+\gamma_2'\mathbf{z})}{(1+exp(\alpha_2+\beta_2 g+\gamma_2'\mathbf{z}))}\begin{pmatrix}2\\g\end{pmatrix}\delta^g(1-\delta)^{(2-g)}0.45^{z_1}(1-0.45)^{(1-z_1)}\phi(z_2)dz_2,$$

and

$$E(Y)=E_{G,\mathbf{Z}}E(Y|g,\mathbf{z})=\alpha_2+\beta_2\delta.$$

For each combination of parameter values, we ran 10,000 simulations, each consisting of 1500 cases and 1500 controls. For each simulated dataset, we derive the ML, pseudo-likelihood, naïve, cases-only, controls-only, weighted, adaptively weighted, adjusted and survey-weighted estimates of the risk effect for the secondary trait and the corresponding standard error estimates. We used the “nmk” function within the “dfoptim” package in R to implement the robust Nelder-Mead algorithm for derivative-free optimization of the likelihood and the “hessian” function from the “numDeriv” package to derive Hessian matrices at parameter estimates. We judge the different estimators based on the amount of bias they incur in estimating the risk effect and their mean squared error (MSE). Using the standard error estimates we construct 95% Wald-type confidence intervals for the risk effect and study their coverage probabilities. Also, we examine Type I error and power for corresponding Wald tests conducted at an intended 5% significance level.

Figures 1 and 2 present the key simulation results for the binary and continuous secondary traits, respectively. The upper row plots the bias, MSE, and coverage probability estimates for the different estimators. In the lower row the estimated Type I error and power are displayed for the corresponding tests under the null and the alternative scenarios. In Figures 1 and 2 we demonstrate the performance of the likelihood-based methods in comparison with only three other methods: the controls-only, adaptively weighted, and survey approaches. The remaining methods (naïve, cases-only, weighted and adjusted), are hugely biased, exhibit extremely large MSEs, and suffer from poor coverage (Figure 4 in Appendix).

We observe the same performance patterns for binary (Figure 1) and continuous (Figure 2) secondary traits. The leftmost plot in the upper row displays the bias incurred by the different estimators. The controls-only and the adaptively weighted methods are considerably biased for non-zero values of β_3. The likelihood-based and survey-weighted estimators are essentially unbiased. The corresponding MSEs for the estimators are shown in the middle plot of the upper row. Both the ML and pseudo-likelihood estimators, as well as the survey-weighted estimator, have smaller mean squared errors than the controls-only and the adaptively weighted estimators for most β_3 values examined. In particular, the two likelihood-based estimators have practically the same MSEs and perform slightly better than the survey-weighted estimator, especially for the continuous trait. The rightmost plot in the upper row presents the estimated coverage probabilities. The coverage for the controls-only and the adaptively weighted methods drops considerably for large β_3. For the confidence intervals corresponding to the likelihood-based and survey methods, the coverage is maintained at the nominal level.

Only the likelihood and the survey approaches maintain the correct Type I error throughout. The power for the Wald tests corresponding to the likelihood and survey methods ranges between 62% to 75% for the binary setup, and varies roughly from 71% to 95% for the continuous setup, whereas the controls-only and the adaptive methods have power as low as 6% (21%) to as high as 100% (95%) for the continuous (binary) trait. But their apparently superior power for large values of β_3 is largely illusory, as they suffer from substantially inflated Type I error for those β_3 values. The key message from the simulations is that the relative advantage of the likelihood and the survey methods over the controls-only and the adaptively weighted methods is preserved across the simulation setups. Furthermore, the likelihood methods can have substantial power advantage over the survey method, evident in the plots for the continuous secondary trait. The mean squared error estimates for the continuous trait suggest that the likelihood estimators can be almost one and a half times more efficient than the survey-weighted estimator. The performances of the ML and pseudo-likelihood estimators are comparable.

4.2.1 Evaluation of robustness—We illustrate the robustness of the proposed ML method by generating data for the binary secondary trait under an alternative model. Earlier, we had simulated data from the Palmgren model. Here we generate data from the following bivariate model, proposed by Lin and Zeng (2009),

$$logit P(Y=1|g,\mathbf{z})=\alpha_2+\beta_2 g+\boldsymbol{\gamma}_{\mathbf{2}}'\mathbf{z} \;\; logit P(D=1|g,\mathbf{z},y)=\alpha_1+\beta_1 g+\gamma_{\mathbf{1}}'\mathbf{z}+\tau y.$$

We note that the disease model above does not preserve the standard logistic marginal distribution for primary disease risk and does not allow for interaction between the genetic variant and the secondary trait on disease risk. The simulation setup is as before. We vary the correlation between the primary and secondary phenotypes, denoted by τ in the disease model, between −0.7 and 0.7. We note that estimates from all the methods are practically unbiased and that Type I error is maintained; the adaptively weighted method is slightly conservative.

5 Discussion

The recent attention to the analysis of secondary traits in case/control genome scans has spurred a variety of methodological efforts. However, it is not clear that the genetics literature has utilized the full range of longstanding methods, including survey-based weighted estimators (Scott and Wild, 2002; Binder, 1983), the Palmgren model (Palmgren, 1989) and the work of Lee et al. (1998). The paper by Jiang et al. (2006) provided a relatively complete taxonomy of various approaches to the problem, but key choices in the modeling of primary and secondary effects had remained open. The more recent work of Lin and Zeng (2009); Li et al. (2010); Li and Gail (2012); He et al. (2012); Wang and Shete (2011, 2012) have provided specific solutions, relevant for the genomic scan context, but with constraints such that it is not clear that a comprehensive approach has been available.

Our likelihood approach is designed to encompass the data types typically encountered for genome scans. The proposed framework can handle continuous secondary traits, as well as the binary traits which have received more attention. In addition, our models allow for interaction between the genetic variant and the secondary trait on primary disease risk. We do not require the rare-disease assumption, i.e., our method can be used to to analyze secondary phenotypes for data from case-control studies of both rare and common primary diseases. We have developed our approach in the context of genome scans, but it is relevant to any standard case/control design. However, our joint model may be particularly useful in a genetic context, for which the parameters are interpretable quantities elucidating genetic effects on risk for both traits, as well as additional trait-trait correlation unexplained by the SNP.

Although demonstrated for a single SNP in this paper, our method can, in principle, be applied to genome-wide association data. However, it will be computationally challenging, as will also be any other retrospective method handling covariates nonparametrically. Derivative-based numerical optimization techniques, as well as C routines for numerical optimization, are likely to provide a faster solution. To reduce computational burden of a genome scan, we can employ a two-stage approach whereby we first screen SNPs using one of the faster but reasonably accurate methods, such as the survey approach, to identify SNPs that might be potentially significant and then follow-up on the interesting SNPs using the ML method.

Our formulation of the joint model for the two traits arises from viewing the pair as a bivariate response, and differs from the usual models based on conditional factorization of the joint distribution. In particular, specifying a logistic model for the conditional distribution of disease status given the secondary trait distorts the natural logistic marginal distribution for primary disease risk, resulting in incompatible models for the primary trait when performing primary vs. secondary analysis. Our parameterization of the joint distribution respects the conventional logistic choice for the marginal distribution of the disease trait, whether the secondary trait is binary or continuous. For binary secondary traits, this property is achieved using a well-known bivariate (Palmgren) logistic model. For continuous secondary traits we achieve the intended marginal model via a two-stage latent variable approach. It turns out that our model for the continuous trait is a special case of the Gaussian copula model (He et al., 2012) for a single normally distributed secondary phenotype.

We emphasize that we provide algorithms that allow us to deal with covariates nonparametrically. In our approximate profile likelihood approach, we reparameterize the profile likelihood to provide a closed form expression eliminating the need for explicit maximization over high-dimensional nuisance parameters, thus moving a step forward from theoretically discussing profile likelihood approaches to handling covariates nonparametrically. We present the pseudo-likelihood approach as another attractive way of dealing with the covariate distribution. We note that both the likelihood approaches yield almost identical results. The closed form expression for the reparameterized profile likelihood can be treated just like a likelihood to obtain the ML estimate and its standard error. The pseudo-likelihood approach, on the other hand, is intuitively very appealing and easily provides estimates; but variance estimation adds to the computational complexity. Yet another option would be to use Bayesian techniques. As one reviewer pointed out, we can use a Dirichlet prior on the probability masses $\psi = \{\psi_1, \psi_2, \ldots, \psi_L\}$ for **Z** and then integrate out ψ as in Zhang and Liu (2007). However, in absence of prior knowledge on covariate distribution, the likelihood and Bayesian approaches will provide very similar results.

We have assumed the genetic and environmental exposures to be independently distributed in the underlying population. We emphasize that the independence assumption applies to the general population, so that after selection by case-control status, they may be dependent (due to their common dependence on D). Thus the assumption is not as restrictive as it may seem. Moreover, for specific genetic applications, the assumption may be reasonable. For random environmental exposures, for instance, it may be entirely reasonable to assume that the genotype does not cause the exposure. The reciprocal assumption is also very standard, i.e. the exposure does not ”cause” genotype, which is fixed from conception. However, if gene-environment independence assumption seems biologically implausible, we can easily relax this assumption. Earlier, we used **Z** to denote covariates and G the SNP genotype. Instead, **Z** will now contain all genetic and environmental factors together, including gene-environment interactions, and we model $P(\mathbf{z})$ nonparametrically as before. On the other hand, if it is reasonable to assume independence of genetic and environmental exposures in the population, our framework allows us to exploit this and still leave the distribution of the environmental exposures to be nonparametric.

Our setup lends itself to performing corrections for secondary trait risk estimates which are subject to significance bias, commonly known as the “winner’s curse”. Such bias correction is necessary if the SNPs have been initially selected based on association with the primary trait, and approaches based on likelihood models (Ghosh et al., 2008) can be extended to encompass our joint likelihood.

Several proposed methods for secondary analysis require knowledge of the disease prevalence. The survey approach uses the ratio of sampling fractions, a function of disease prevalence, for appropriate weighting of cases and the controls. The method of moments approach proposed by Wang and Shete (2011) involves knowing the prevalence for the secondary trait in addition to the disease prevalence. In principle, likelihood-based methods do not require specification of the disease prevalence. However, when disease prevalence is assumed unknown, the estimation of intercept parameter α_1 can lead to numerically unstable solutions, as has been noted by others (Lin and Zeng, 2009; Chatterjee and Carroll, 2005). In practice, a sensitivity analysis is warranted to determine the degree to which inference is affected by the assumed prevalence, and robust approaches to estimate the prevalence in the context of our models is an important area for further exploration.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors appreciate the insightful and constructive comments made by the two anonymous reviewers. The authors thank Dr. Ravi Varadhan for helpful suggestions regarding numerical optimization in R. Part of Dr. Ghosh's research was supported by the Intramural Research Program of the Division of Cancer Epidemiology and Genetics of the National Cancer Institute.

References

Bahadur, R. A representation of the joint distribution of responses to n dichotomous items. In: Solomon, H., editor. In Studies in Item Analysis and Prediction. Stanford, California: Stanford University Press; 1959. Stanford Mathematical Studies in the Social Sciences VI.

Binder D. On the variances of asymptotically normal estimators from complex surveys. International Statistical Review/Revue Internationale de Statistique. 1983:279–292.

Chatterjee N, Carroll R. Semiparametric maximum likelihood estimation exploiting gene-environment independence in case-control studies. Biometrika. 2005; 92:399–418.

Frayling T, Timpson N, Weedon M, Zeggini E, Freathy R, Lindgren C, Perry J, Elliott K, Lango H, Rayner N, et al. A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. Science. 2007; 316:889–894. [PubMed: 17434869]

Ghosh A, Zou F, Wright F. Estimating odds ratios in genome scans: an approximate conditional likelihood approach. The American Journal of Human Genetics. 2008; 82:1064–1074.

Gong G, Samaniego F. Pseudo maximum likelihood estimation: theory and applications. The Annals of Statistics. 1981:861–869.

He J, Li H, Edmondson A, Rader D, Li M. A gaussian copula approach for the analysis of secondary phenotypes in case–control genetic association studies. Biostatistics. 2012; 13:497–508. [PubMed: 21933777]

Jiang Y, Scott AJ, Wild CJ. Secondary analysis of case-control data. Stat Med. 2006; 25:1323–1339. [PubMed: 16220494]

Lee A, McMURCHY L, Scott A. Re-using data from case-control studies. Statistics in medicine. 1998; 16:1377–1389. [PubMed: 9232759]

Li H, Gail M. Efficient adaptively weighted analysis of secondary phenotypes in case-control genome-wide association studies. Human Heredity. 2012; 73:159–173. [PubMed: 22710642]

Li H, Gail M, Berndt S, Chatterjee N. Using cases to strengthen inference on the association between single nucleotide polymorphisms and a secondary phenotype in genome-wide association studies. Genetic epidemiology. 2010; 34:427–433. [PubMed: 20583284]

Lin D, Zeng D. Proper analysis of secondary phenotype data in case-control association studies. Genetic epidemiology. 2009; 33:256–265. [PubMed: 19051285]

Meester S, MacKay J. A parametric model for cluster correlated categorical data. Biometrics. 1994; 50:954–963. [PubMed: 7787008]

Monsees G, Tamimi R, Kraft P. Genome-wide association scans for secondary traits using case-control samples. Genetic epidemiology. 2009; 33:717–728. [PubMed: 19365863]

Moslehi R, Chatterjee N, Church T, Chen J, Yeager M, Weissfeld J, Hein D, Hayes R. Cigarette smoking, N-acetyltransferase genes and the risk of advanced colorectal adenoma. Pharmacogenomics. 2006; 7:819–829. [PubMed: 16981843]

Nagelkerke N, Moses S, Plummer F, Brunham R, Fish D. Logistic regression in case-control studies: The effect of using independent as dependent variables. Statistics in medicine. 1995; 14:769–775. [PubMed: 7644857]

Palmgren, J. Regression models for bivariate binary responses; UW Biostatistics Working Paper Series; 1989. p. 101

Prentice R, Pyke R. Logistic disease incidence models and case-control studies. Biometrika. 1979; 66:403–411.

Richardson DB, Rzehak P, Klenk J, Weiland SK. Analyses of case-control data for additional outcomes. Epidemiology. 2007; 18:441–445. [PubMed: 17473707]

Scott A, Wild C. Fitting logistic regression models in stratified case-control studies. Biometrics. 1991:497–510.

Scott A, Wild C. Fitting regression models to case-control data by maximum likelihood. Biometrika. 1997a; 84:57–71.

Scott A, Wild C. Maximum likelihood for generalised case-control studies. Journal of Statistical Planning and Inference. 2001; 96:3–27.

Scott A, Wild C. On the robustness of weighted methods for fitting models to case-control data. Journal of the Royal Statistical Society. Series B, Statistical Methodology. 2002:207–219.

Scott AJ, Wild CJ. Fitting regression models to case-control data by maximum likelihood. Biometrika. 1997b; 84:57–71.

Spitz M, Amos C, Dong Q, Lin J, Wu X. The chrna5-a3 region on chromosome 15q24-25.1 is a risk factor both for nicotine dependence and for lung cancer. Journal of the National Cancer Institute. 2008; 100:1552–1556. [PubMed: 18957677]

Wang J, Shete S. Estimation of odds ratios of genetic variants for the secondary phenotypes associated with primary diseases. Genetic epidemiology. 2011; 35:190–200. [PubMed: 21308766]

Wang J, Shete S. Analysis of secondary phenotype involving the interactive effect of the secondary phenotype and genetic variants on the primary disease. Annals of Human Genetics. 2012; 76:484–499. [PubMed: 22881407]

Weedon M, Lango H, Lindgren C, Wallace C, Evans D, Mangino M, Freathy R, Perry J, Stevens S, Hall A, et al. Genome-wide association analysis identifies 20 loci that influence adult height. Nature Genetics. 2008; 40:575–583. [PubMed: 18391952]

Zhang Y, Liu J. Bayesian inference of epistatic interactions in case-control studies. Nature genetics. 2007; 39:1167–1173. [PubMed: 17721534]

Appendix

7.1 The profile log-likelihood

Let n_{+++l} denote the number of observations in the sample with $\mathbf{Z} = \mathbf{z}_l$, $l = 1, \ldots, L$. The retrospective log-likelihood in (4) can then be expressed as

$$l=\sum_{u=1}^{n}\log P(d_u, y_u|g_u, \mathbf{z}_u;\theta)+\sum_{u=1}^{n}\log P(g_u;\delta)+\sum_{l=1}^{L}n_{+++l}\log\psi_l-\sum_{i=0}^{1}n_i\log\sum_{l=1}^{L}P(D=i|\mathbf{z}_l;\theta_1)\psi_l. \quad (8)$$

Using a Lagrange multiplier λ to account for the constraint $\sum_{l=1}^{L}\psi_l=1$ and then maximizing the retrospective log-likelihood with respect to ψ_l, we obtain

$$\frac{n_{+++l}}{\psi_l}-\sum_{i}n_i\frac{P(D=i|\mathbf{z}_l;\theta_1)}{P(D=i)}-\lambda=0. \quad (9)$$

Multiplying (9) by ψ_l and summing over l gives us $\lambda = 0$ which on substitution results in

$$\psi_l=\frac{n_{+++l}}{\sum_i \mu_i P(D=i|\mathbf{z}_l;\theta_1)}, \quad (10)$$

where $\mu_i=\frac{n_i}{P(D=i)}$. We note that $\frac{n_0}{\mu_0}+\frac{n_1}{\mu_1}=1$. Substituting ψ_l in the expression for μ_i we have

$$n_i=\sum_l n_{+++l}\frac{P(D=i|\mathbf{z}_l;\theta_1)\mu_i}{\sum_i P(D=i|\mathbf{z}_l;\theta_1)\mu_i}, i=0,1. \quad (11)$$

We substitute (10) in (8) to obtain the following expression that is equivalent to the profile log-likelihood up to two additional nuisance parameters, μ_i, i = 0, 1,

$$\begin{aligned} & l^*(\eta,\mu(\eta)) \\ & =\sum_{u=1}^{n}\log P(d_u,y_u|g_u,\mathbf{z}_u;\theta)+\sum_{u=1}^{n}\log P(g_u;\delta) \\ & -\sum_{l=1}^{L} n_{+++l}\log\sum_i \mu_i P(D \\ & =i|\mathbf{z}_l;\theta_1) \\ & +\sum_{i=0}^{1} n_i\log\mu_i \\ & =\sum_{u=1}^{n}\log P^*_{d_u,y_u,g_u}(\mathbf{z}_u;\eta,\mu). \end{aligned}$$

where μ_i's satisfy (11) and $P^*_{d_u,y_u,g_u}(\mathbf{z}_u;\eta,\mu)$ is defined as

$$P^*_{d_u,y_u,g_u}(\mathbf{z}_u;\eta,\mu)=\frac{P(d_u,y_u|g_u,\mathbf{z}_u;\theta)P(g_u;\delta)\mu_u}{\sum_{i=0}^{1} P(D=i|\mathbf{z}_u;\theta_1)\mu_i}.$$

We note that

$$\frac{\partial l^*(\eta,\mu)}{\partial\mu_i}=\frac{n_i-\sum_l n_{+++l}\frac{P(D=i|\mathbf{z}_l;\theta_1)\mu_i}{\sum_i P(D=i|\mathbf{z}_l;\theta_1)\mu_i}}{\mu_i}. \quad (12)$$

Comparing (12) with (11), we realize that (11) can be replaced by $\frac{\partial l*(\eta,\mu)}{\partial\mu_i}=\mathbf{0}, i=0,1$. Thus, we can treat l*(**η, μ**) as a log-likelihood, although it is not an actual log-likelihood, and maximize it with respect to (**η, μ**) to obtain $\hat{\eta}$. We further note that l*(**η, μ**) can be expressed in terms of η and $\kappa=\frac{\mu_1}{\mu_0}$ as follows:

$$l^*(\eta,\mu)=\sum_{u=1}^{n}\log P^*_{d_u,y_u,g_u}(\mathbf{z}_u;\eta,\kappa)=l^*(\eta,\kappa), \quad (13)$$

with

$$P^*_{d_u,y_u,g_u}(\mathbf{z}_u;\eta,\kappa)=\frac{P(d_u,y_u|g_u,\mathbf{z}_u;\theta)P(g_u;\delta)\kappa^{d_u}}{\sum_{i=0}^{1}P(D=i|\mathbf{z}_u;\theta)\kappa^i}.$$

7.2 Asymptotic theory for ML estimate

We first show that *E*(*S**(φ)) = 0. The proof is presented in the following subsection. We then assume that *n* goes to infinity with *n*$_1$/*n* and *n*$_2$/*n* remaining fixed. Under this assumption it can be shown, by expanding *S**($\hat{\varphi}$) about the true value φ and then applying standard procedures, that $\sqrt{n}(\hat{\phi}-\boldsymbol{\phi})$ converges in distribution to a normal random variable with mean zero and covariance matrix

$$nCov(\hat{\phi})=nJ^*(\boldsymbol{\phi})^{-1}Cov(S^*(\boldsymbol{\phi}))J^*(\boldsymbol{\phi})^{-1}.$$

We then show in Subsection 7.2.2 that *Cov*(*S**(φ)) = *J**(φ) − Γ(φ). This leads to

$$Cov(\hat{\phi})=J^*(\boldsymbol{\phi})^{-1}-J^*(\boldsymbol{\phi})^{-1}\Gamma(\boldsymbol{\phi})J^*(\boldsymbol{\phi})^{-1}.$$

7.2.1 *E*(*S**(φ)) = 0

$$\begin{aligned}
E(S^*(\boldsymbol{\phi}))=&\sum_u E_{d_u}\left(\frac{\partial}{\partial\boldsymbol{\phi}}\log P^*_{d_u,y_u,g_u}(\mathbf{z}_u;\boldsymbol{\phi})\right), \text{where} E_{d_u}(.)\\
&=E(.|D\\
&=d_u)=\sum_i n_i E_i\left(\frac{\partial\log P^*_{iyg}(\mathbf{z};\boldsymbol{\phi})}{\partial\boldsymbol{\phi}}\right)\\
&=\sum_i n_i\sum_{y,g,\mathbf{z}}\frac{\partial\log P^*_{iyg}(\mathbf{z};\boldsymbol{\phi})}{\partial\boldsymbol{\phi}}P(y,g,\mathbf{z}|D\\
&=i)=\sum_i n_i\sum_{y,g,\mathbf{z}}\frac{1}{P^*_{iyg}(\mathbf{z})}\frac{\partial P^*_{iyg}(\mathbf{z};\boldsymbol{\phi})}{\partial\boldsymbol{\phi}}\frac{P(D=i,y|g,\mathbf{z})P(g)}{P(D=i)}P(\mathbf{z})\\
&=\sum_i\sum_{y,g,\mathbf{z}}\frac{\partial P^*_{iyg}(\mathbf{z};\boldsymbol{\phi})}{\partial\boldsymbol{\phi}}f(\mathbf{z})P(\mathbf{z}), \text{where} f(\mathbf{z})\\
&=(\frac{n1}{\kappa}+n_0)\sum_i\kappa^i P(D\\
&=i|\mathbf{z})\\
&=\sum_{\mathbf{z}}\sum_{i,y,g}\frac{\partial P^*_{iyg}(\mathbf{z};\boldsymbol{\phi})}{\partial\boldsymbol{\phi}}f(\mathbf{z})P(\mathbf{z})\\
&=0, \text{since}\sum_{i,y,g}P^*_{iyg}(\mathbf{z};\boldsymbol{\phi})=1.
\end{aligned}$$

7.2.2 *Cov*(*S**(φ)) = *J**(φ) − Γ(φ)

$$\begin{aligned}
Cov(S^*(\boldsymbol{\phi}))=&\sum_u E_{d_u}(\frac{\partial}{\partial\boldsymbol{\phi}}\log P^*_{d_u,y_u,g_u}(\mathbf{z}_u;\boldsymbol{\phi}\frac{\partial}{\partial\boldsymbol{\phi}'}\log P^*_{d_u,y_u,g_u}(\mathbf{z}_u;\boldsymbol{\phi})\\
&-\sum_u E_{d_u}\left(\frac{\partial}{\partial\boldsymbol{\phi}}\log P^*_{d_u,y_u,g_u}(\mathbf{z}_u;\boldsymbol{\phi})\right)E_{d_u}\left(\frac{\partial}{\partial\boldsymbol{\phi}'}\log P^*_{d_u,y_u,g_u}(\mathbf{z}_u;\boldsymbol{\phi})\right)\\
=&\sum_i n_iE_i\left(\frac{\partial}{\partial\boldsymbol{\phi}}\log P^*_{iyg}(\mathbf{z};\boldsymbol{\phi})\frac{\partial}{\partial\boldsymbol{\phi}'}\log P^*_{iyg}(\mathbf{z};\boldsymbol{\phi})\right)\\
&-\sum_i n_iE_i\left(\frac{\partial}{\partial\boldsymbol{\phi}}\log P^*_{iyg}(\mathbf{z};\boldsymbol{\phi})\right)E_i\left(\frac{\partial}{\partial\boldsymbol{\phi}'}\log P^*_{iyg}(\mathbf{z};\boldsymbol{\phi})\right).
\end{aligned}$$

Using the same argument as before it can be shown that

$$\sum_i n_iE_i\left(\frac{1}{P^*_{iyg}(\mathbf{z})}\frac{\partial^2}{\partial\boldsymbol{\phi}\partial\boldsymbol{\phi}'}P^*_{iyg}(\mathbf{z})\right)=0$$

which implies that

$$\sum_i n_iE_i\left(\frac{\partial^2}{\partial\boldsymbol{\phi}\partial\boldsymbol{\phi}'}\log P^*_{iyg}(\mathbf{z})\right)=-\sum_i n_iE_i\left(\frac{\partial}{\partial\boldsymbol{\phi}}\log P^*_{iyg}(\mathbf{z})\frac{\partial}{\partial\boldsymbol{\phi}'}\log P^*_{iyg}(\mathbf{z})\right).$$

We have also shown before that

$$E_i\left(\frac{\partial}{\partial\boldsymbol{\phi}}\log P^*_{iyg}(\mathbf{z};\boldsymbol{\phi})\right)=\frac{1}{n_i}\sum_{y,g,\mathbf{z}}\frac{\partial}{\partial\boldsymbol{\phi}}P^*_{iyg}(\mathbf{z};\boldsymbol{\phi})f(\mathbf{z})P(\mathbf{z})=\frac{1}{n_i}E\left(f(\mathbf{z})\frac{\partial}{\partial\boldsymbol{\phi}}P^*_i(\mathbf{z};\boldsymbol{\phi})\right),$$

where $P^*_i(\mathbf{z})=\frac{P(D=i|\mathbf{z})\kappa^i}{\sum_i\kappa^iP(D=i|\mathbf{z})}$. Then

$$
\begin{aligned}
&\sum_i n_i E_i\left(\frac{\partial}{\partial \boldsymbol{\phi}}\log P^*_{iyg}(\mathbf{z};\boldsymbol{\phi})\right)E_i\left(\frac{\partial}{\partial \boldsymbol{\phi}'}\log P^*_{iyg}(\mathbf{z};\boldsymbol{\phi})\right)\\
&=\sum_i E(f(\mathbf{z})\frac{\partial}{\partial \boldsymbol{\phi}}P^*_i(\mathbf{z};\boldsymbol{\phi})\frac{1}{n_i}E\left(f(\mathbf{z})\frac{\partial}{\partial \boldsymbol{\phi}'}P^*_i(\mathbf{z};\boldsymbol{\phi})\right)\\
&=\frac{1}{n_0}E\left(f(\mathbf{z})\frac{\partial}{\partial \boldsymbol{\phi}}P^*_0(\mathbf{z};\boldsymbol{\phi})\right)E\left(f(\mathbf{z})\frac{\partial}{\partial \boldsymbol{\phi}'}P^*_0(\mathbf{z};\boldsymbol{\phi})\right)\\
&+\frac{1}{n_1}E\left(f(\mathbf{z})\frac{\partial}{\partial \boldsymbol{\phi}}P^*_1(\mathbf{z};\boldsymbol{\phi})\right)E\left(f(\mathbf{z})\frac{\partial}{\partial \boldsymbol{\phi}'}P^*_1(\mathbf{z};\boldsymbol{\phi})\right)\\
&=\frac{1}{n_0}E\,(\\
&-f(\mathbf{z})\frac{\partial}{\partial \boldsymbol{\phi}}P^*_1(\mathbf{z};\boldsymbol{\phi})E\,(\\
&\quad -f(\mathbf{z})\frac{\partial}{\partial \boldsymbol{\phi}'}P^*_1(\mathbf{z};\boldsymbol{\phi})\Big)\\
&+\frac{1}{n_1}E\left(f(\mathbf{z})\frac{\partial}{\partial \boldsymbol{\phi}}P^*_1(\mathbf{z};\boldsymbol{\phi})\right)E\left(f(\mathbf{z})\frac{\partial}{\partial \boldsymbol{\phi}'}P^*_1(\mathbf{z};\boldsymbol{\phi})\right)\\
&=\left(\frac{1}{n_0}+\frac{1}{n_1}\right)E\left(f(\mathbf{z})\frac{\partial}{\partial \boldsymbol{\phi}}P^*_1(\mathbf{z};\boldsymbol{\phi})\right)E\left(f(\mathbf{z})\frac{\partial}{\partial \boldsymbol{\phi}'}P^*_1(\mathbf{z};\boldsymbol{\phi})\right).
\end{aligned}
$$

$\therefore$

$$
\begin{aligned}
Cov(S^*(\boldsymbol{\phi}))=&-\sum_i n_i E_i\left(\frac{\partial^2}{\partial \boldsymbol{\phi}\partial \boldsymbol{\phi}'}\log P^*_{iyg}(\mathbf{z};\boldsymbol{\phi})\right)\\
&-\left(\frac{1}{n_0}+\frac{1}{n_1}\right)E\left(f(\mathbf{z})\frac{\partial}{\partial \boldsymbol{\phi}}P^*_1(\mathbf{z};\boldsymbol{\phi})\right)E\left(f(\mathbf{z})\frac{\partial}{\partial \boldsymbol{\phi}'}P^*_1(\mathbf{z};\boldsymbol{\phi})\right)\\
=&-E(\frac{\partial}{\partial \boldsymbol{\phi}'}S^*(\boldsymbol{\phi}))\\
&-\Gamma(\boldsymbol{\phi})\\
=&J^*(\boldsymbol{\phi})-\Gamma(\boldsymbol{\phi}).
\end{aligned}
$$

7.3 Asymptotic theory for pseudo-likelihood estimate

Gong and Samaniego (1981), under some regularity conditions, showed that $\tilde{\theta}_2$ is consistent when $\tilde{\theta}_1$ is consistent. Also, suppose that

$$
\sqrt{n}\begin{pmatrix}\tilde{\theta}_1-\theta_1^0\\ \frac{\partial l_1(\theta_1^0,\theta_2^0)/\partial\theta_2}{n}\end{pmatrix}\to_d N\left(\mathbf{0},\begin{pmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{12}' & I_{22}\end{pmatrix}\right),
$$

where $(\theta_1^0,\theta_2^0)'$ is the true value of $(\theta_1, \theta_2)'$ and

$$
I_{22}=\lim_{n\to\infty}\left[E\left\{\frac{-\partial^2 l_1(\theta_1^0,\theta_2^0)/\partial\theta_2^2}{n};\theta_1^0,\theta_2^0\right\}\right].
$$

Then, under some regularity conditions, discussed in Gong and Samaniego,

$$\sqrt{n}\left(\tilde{\theta}_2-\theta_2^0\right)\rightarrow_d N(\mathbf{0},\Sigma_{22}),$$

with

$$\Sigma_{22}=I_{22}^{-1}(I_{22}-2I_{12}^{'}\Sigma_{12}+I_{12}^{'}\Sigma_{11}I_{12})I_{22}^{-1}=I_{22}^{-1}I_{22}^{*}I_{22}^{-1}, \quad (14)$$

where

$$I_{12}=\lim_{n\to\infty}\left[E\left\{\frac{-\partial^2 l_1(\theta_1^0,\theta_2^0)/\partial\theta_1\partial\theta_2}{n};\theta_1^0,\theta_2^0\right\}\right].$$

I_{22} is equal to the variance-covariance matrix of the score function for θ_2 evaluated at (θ_1^0,θ_2^0) and I_{22}^* is the limiting variance-covariance matrix of $\sqrt{n}\partial\tilde{l}_1(\theta_2)/\partial\theta_2$.

7.3.1 Variance Estimation

We use

$$\hat{I}_{22}=-\frac{1}{n}\frac{\partial^2 l_1(\theta_1,\theta_2)}{\partial\theta_2^2}\bigg|_{(\tilde{\theta}_1,\tilde{\theta}_2)}=-\frac{1}{n}\frac{\partial^2\tilde{l}_1(\theta_2)}{\partial\theta_2^2}\bigg|_{\tilde{\theta}_2}$$

and

$$\hat{I}_{12}=-\frac{1}{n}\frac{\partial^2 l_1(\theta_1,\theta_2)}{\partial\theta_1\partial\theta_2}\bigg|_{(\tilde{\theta}_1,\tilde{\theta}_2)}.$$

We note that

$$\Sigma_{12}=cov(\tilde{\theta}_1-\theta_1^0,\frac{\partial}{\partial\theta_2}l_1(\theta_1^0,\theta_2^0))$$

If $\tilde{\theta_1}$ is the ML estimate obtained by maximizing $g(\theta_1)$, by Taylor expansion, $\tilde{\theta}_1-\theta_1^0$ is asymptotically equivalent to

$$Var(\tilde{\theta}_1;\theta_1^0)\frac{\partial}{\partial\theta_1}g(\theta_1^0)$$

which leads to

$$\Sigma_{12}=Var(\tilde{\theta}_1;\theta_1^0)cov(\frac{\partial}{\partial\theta_1}g(\theta_1^0),\frac{\partial}{\partial\theta_2}l_1(\theta_1^0,\theta_2^0)).$$

We can easily estimate the asymptotic variance of $\tilde{\theta_1}$ by

$$\widehat{\mathrm{Var}}(\tilde{\theta}_1)=\left(-\frac{\partial^2}{\partial{\theta_1}^2}g(\tilde{\theta}_1)\right)^{-1}.$$

Now we have only to estimate $cov(\frac{\partial}{\partial\theta_1}g(\theta_1^0),\frac{\partial}{\partial\theta_2}l_1(\theta_1^0,\theta_2^0))$. If we can write

$$\frac{\partial}{\partial\theta_1}g(\theta_1^0)=\sum_u\frac{\partial}{\partial\theta_1}g_u(\theta_1^0)$$

and

$$\frac{\partial}{\partial\theta_2}l_1(\theta_1^0,\theta_2^0)=\sum_u\frac{\partial}{\partial\theta_2}l_{1u}(\theta_1^0,\theta_2^0),$$

then we can use the following to estimate the covariance,

$$\widehat{Cov}(\frac{\partial}{\partial\theta_1}g(\theta_1^0),\frac{\partial}{\partial\theta_2}l_1(\theta_1^0,\theta_2^0))=\sum_u\frac{\partial}{\partial\theta_1}g_u(\tilde{\theta}_1),\frac{\partial}{\partial\theta_2}l_{1u}(\tilde{\theta}_1,\tilde{\theta}_2)'=\sum_u\frac{\partial}{\partial\theta_1}g_u(\tilde{\theta}_1)\frac{\partial}{\partial\theta_2}\tilde{\mathrm{l}}_{1u}(\tilde{\theta}_2)'$$

Thus,

$$\hat{\Sigma}_{12}=\left(-\frac{\partial^2}{\partial{\theta_1}^2}g(\tilde{\theta}_1)\right)^{-1}\sum_u\frac{\partial}{\partial\theta_1}g_u(\tilde{\theta}_1)\frac{\partial}{\partial\theta_2}\tilde{\mathrm{l}}_{1u}(\tilde{\theta}_2)'.$$

Also,

$$\hat{\Sigma}_{11}=n\left(-\frac{\partial^2}{\partial{\theta_1}^2}g(\tilde{\theta}_1)\right)^{-1}$$

Plugging in all the corresponding estimates in (14) gives

$$\hat{\Sigma}_{22}=\hat{I}_{22}^{-1}\left(\hat{I}_{22}-2\hat{I}_{12}^{'}\hat{\Sigma}_{12}+\hat{I}_{12}^{'}\hat{\Sigma}_{11}\hat{I}_{22}\right)\hat{I}_{22}^{-1}. \quad (15)$$

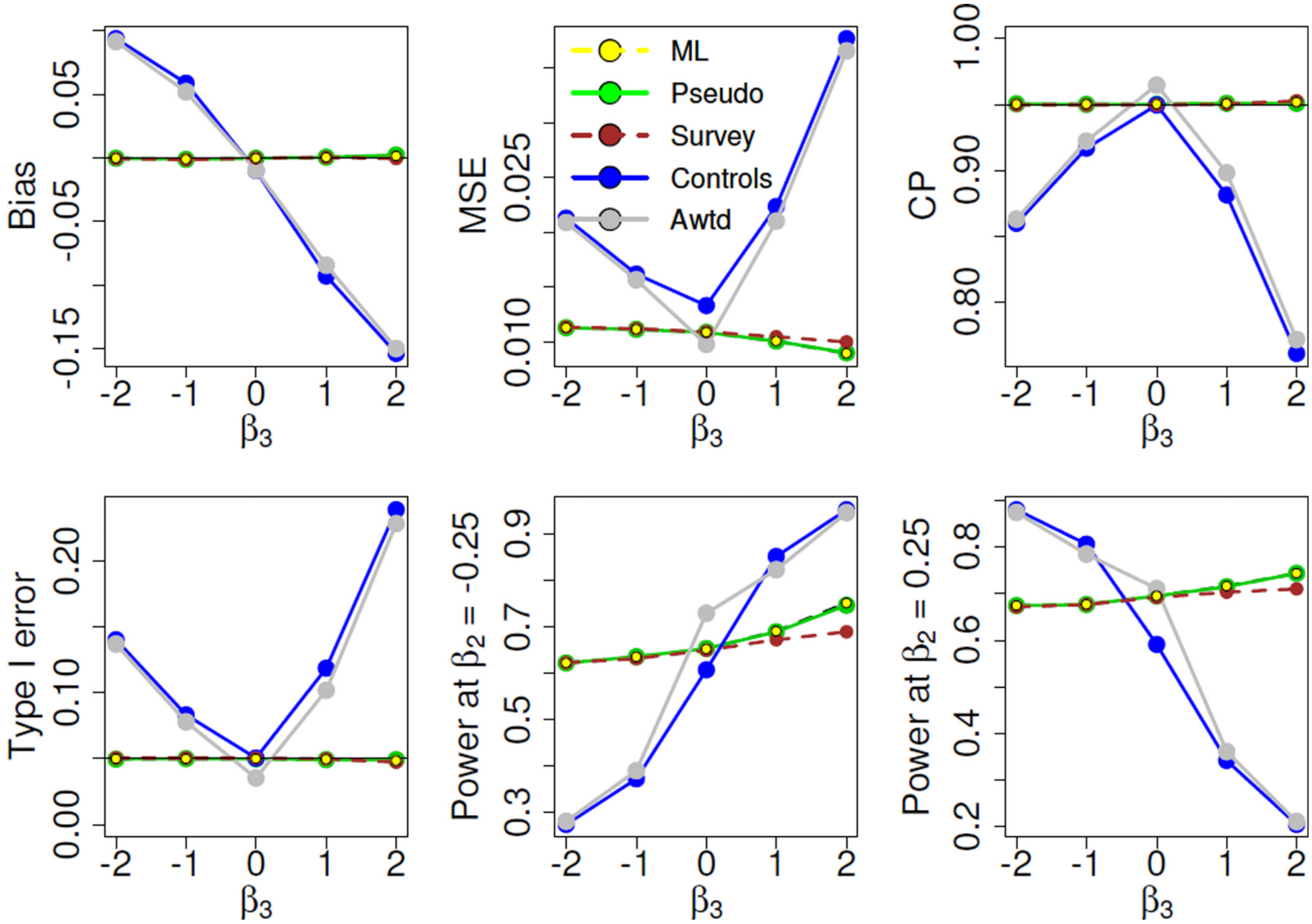

Figure 1.

Binary secondary trait

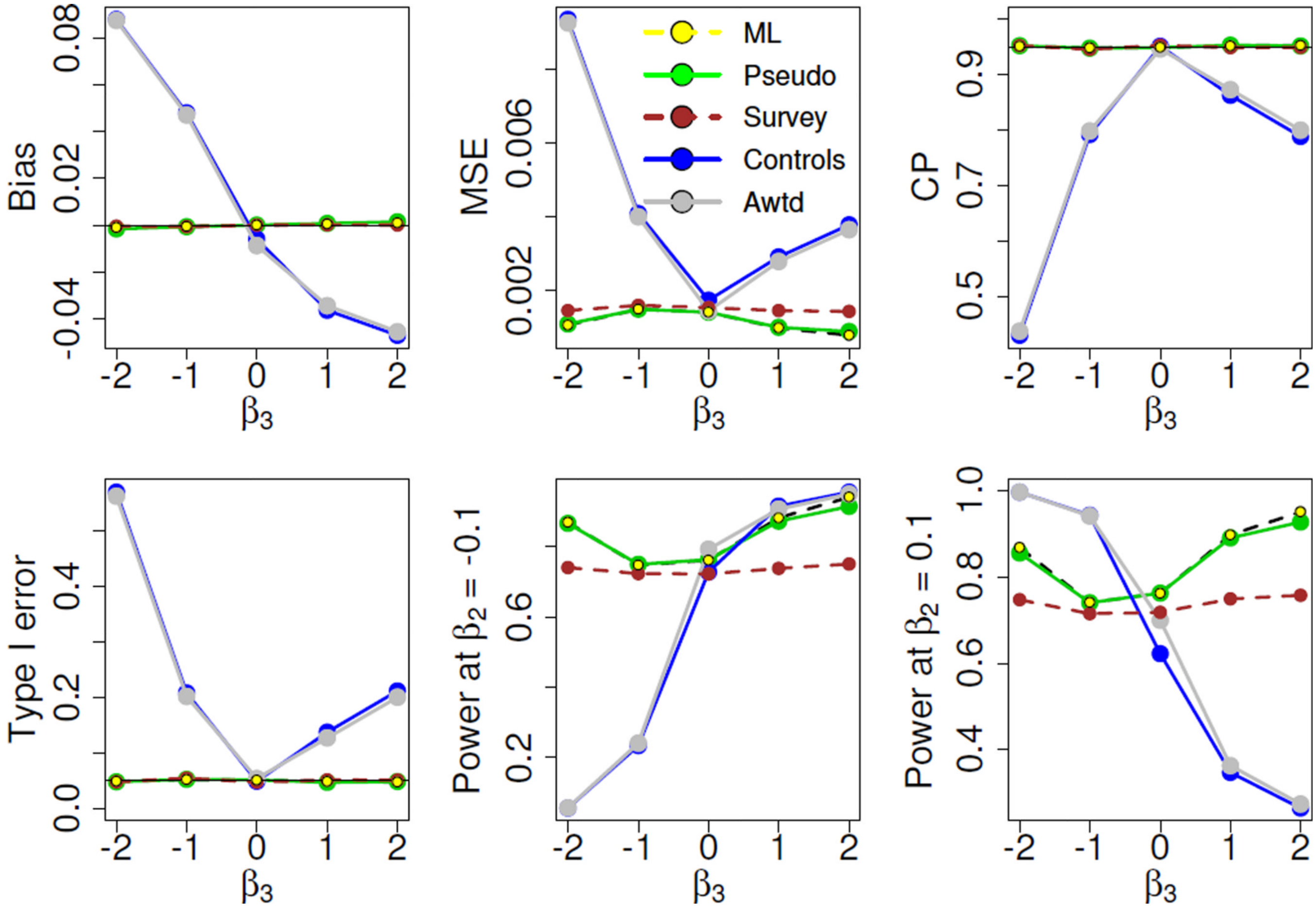

Figure 2.
Continuous secondary trait

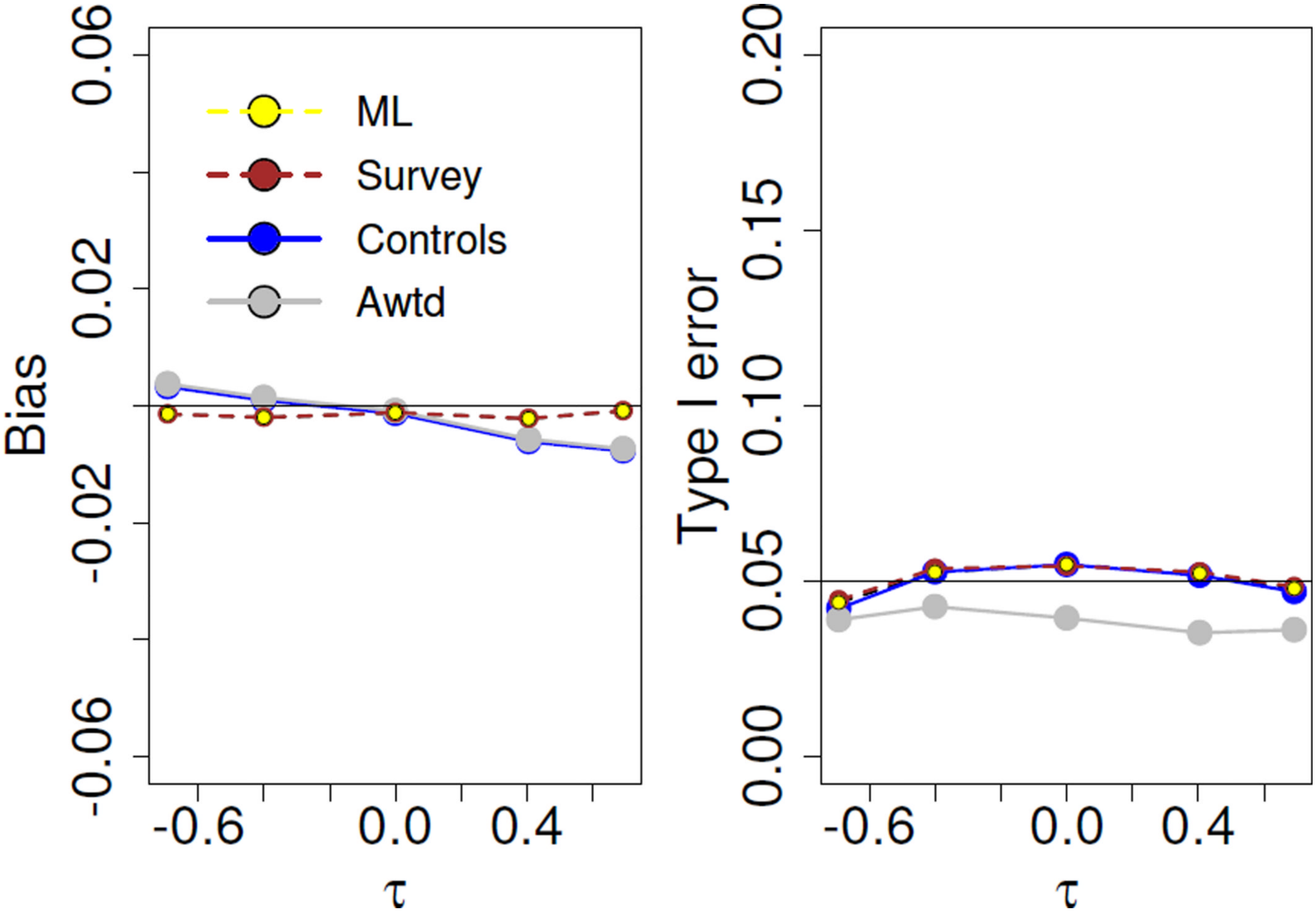

Figure 3.
Performance under alternative model

Table 1

Data for studying effect of *NAT2* on smoking from a case-control study for colorectal adenoma

Cases of colorectal adenoma				**Matched controls**			
	Smoking status		**Total**		**Smoking status**		**Total**
	Never/Former	**Current**			**Never/Former**	**Current**	
NAT2=0	199	380	579	*NAT2*=0	255	317	572
NAT2=1	18	13	31	*NAT2*=1	10	23	33
Total	217	393	610	Total	265	340	605

Table 2

Estimates of effect of *NAT2* on smoking from a case-control study for colorectal adenoma

Method	**log(OR)**	**SE**
Naïve	−0.18	0.26
Cases	−0.97	0.37
Controls	0.62	0.39
Adj	−0.17	0.26
Wtd	−0.21	0.27
Awtd	0.57	0.40
ML/Survey	0.54	0.37