The Canadian Journal of Statistics
Vol. 49, No. 2, 2021, Pages 344–361
La revue canadienne de statistique

Empirical and conditional likelihoods for two-phase studies

Menglu CHE[1], Jerald F. LAWLESS[1*], and Peisong HAN[2]

[1]*Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, Canada*
[2]*Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, MI, U.S.A.*

Key words and phrases: Conditional likelihood; empirical likelihood; missing-at-random; response-dependent sampling; two-phase studies.

MSC 2010: Primary 62J99; secondary 62F12.

Abstract: Two-phase, response-dependent sampling is often used in regression settings that involve expensive covariate measurements. Conditional maximum likelihood (CML) is an attractive approach in many cases as it avoids modelling of the covariate distribution, unlike full maximum likelihood. Scott & Wild (2011) introduced an augmented CML approach which is semiparametric efficient in certain settings with a discrete response variable. We consider general regression models and show the Scott–Wild estimator of covariate effects has the same asymptotic efficiency as two empirical likelihood estimators, and that these estimators dominate the CML estimator. We compare the efficiencies of various estimators in simulation studies and illustrate the methodology in a two-phase genetics study. *The Canadian Journal of Statistics* 49: 344–361; 2021 © 2020 Statistical Society of Canada

Résumé: Un échantillonnage en deux phases dépendant de la réponse est couramment utilisé dans les contextes de régression dont les covariables nécessitent des mesures coûteuses. Le maximum de vraisemblance conditionnelle (MVC) constitue une approche intéressante pour de nombreux cas puisqu'il évite de modéliser la distribution des covariables, contrairement au maximum de vraisemblance complète. Scott et Wild (2011) ont proposé une approche de MVC augmentée qui est semi-paramétriquement efficace dans certains contextes avec une variable réponse discrète. Les auteurs considèrent les modèles de régression en général et montrent que l'estimateur de Scott-Wild (SW) pour l'effet des covariables offre la même efficacité asymptotique que deux estimateurs issus de la vraisemblance empirique, et que ces estimateurs dominent celui du MVC. Les auteurs comparent l'efficacité de différents estimateurs dans des études de simulation et illustrent la méthodologie dans le cadre d'une étude génétique à deux phases. *La revue canadienne de statistique* 49: 344–361; 2021 © 2020 Société statistique du Canada

1. INTRODUCTION

Response-dependent samples arise in a variety of contexts; see for example, Qin (2017), where numerous examples and associated methodology are discussed. We consider two-phase studies with response-dependent samples that are designed to provide efficient estimation of covariate effects while limiting the cost of obtaining expensive covariate measurements (e.g., Song, Zhou & Kosorok, 2009). Phase 1 data consisting of responses Y_i and covariates Z_i are available for a cohort of individuals $i = 1, \ldots, N$ and in phase 2, values for expensive covariates X_i are

Additional Supporting Information may be found in the online version of this article at the publisher's website.

*Author to whom correspondence may be addressed.
E-mail: jlawless@uwaterloo.ca

© 2020 Statistical Society of Canada / Société statistique du Canada

obtained for a subset of n individuals. By appropriately basing phase 2 sample selection on the observed phase 1 responses and covariates, estimation efficiency can be increased. Such designs include generalized case–control and case–cohort designs used with rare outcomes (e.g., Scott & Wild, 2001; Borgan & Samuelsen, 2014); other designs with response-dependent sampling are frequently used in genetic studies (e.g., Breslow et al., 2009; Barnett, Lee & Lin, 2013; Lin, Zeng & Tang, 2013) and many other areas.

We assume that triplets $(Y_i, X_i, Z_i), i = 1, \ldots, N$ are iid and have joint probability density or mass function $f(y|x, z)g(x, z)$. Our interest is in estimation of $\boldsymbol{\beta}$ in a regression model $f(y|x, z; \boldsymbol{\beta})$; the distribution $g(x, z)$ of the covariates is not of direct interest and we may wish to avoid modelling it (Scott & Wild, 2011). Methods that avoid consideration of $g(x, z)$ include Horwitz–Thompson or inverse-probability-weighted (IPW) estimation and conditional maximum likelihood (CML), e.g., see Lawless, Kalbfleisch & Wild (1999). Both approaches use only the data for phase 2 individuals but require specification of phase 2 selection probabilities. We denote $R_i = \mathrm{I}$ (individual i is selected for phase 2), $i = 1, \ldots, N$ and let $\pi(y, z; \boldsymbol{\alpha}) = P(R_i = 1|Y_i = y, Z_i = z)$. The phase 2 sample is typically selected either by variable probability sampling (VPS), where the R_i $(i = 1, \ldots, n)$ are independent Bernoulli random variables with selection probabilities $\pi(y_i, z_i)$ or by basic stratified sampling (BSS), in which case the N individuals are stratified on the basis of their phase 1 variables Y, Z, and a simple random sample is drawn from each stratum for phase 2. In the former case the true value $\boldsymbol{\alpha}_0$ of $\boldsymbol{\alpha}$ is specified by design, but using an estimate $\widehat{\boldsymbol{\alpha}}$ instead can increase efficiency (Robins, Rotnitzky & Zhao, 1994; Lawless, Kalbfleisch & Wild, 1999; Scott & Wild, 2011). For simplicity we focus on VPS in the discussion, but the results also apply to BSS with minor modifications (Scott & Wild, 2011).

CML, described in the next section, has been widely used when Y is categorical (Scott & Wild, 1986; Breslow & Cain, 1988; Keogh & Cox, 2014; Qin, 2017) but also with continuous responses (Huang & Lin, 2007; Li et al., 2011; Barnett, Lee & Lin, 2013) and with censored failure time outcomes (Shen et al., 2015). Semiparametric maximum likelihood (ML) methods are also used; they do not require a selection model $\pi(y, z; \boldsymbol{\alpha})$ but unlike CML, the covariate distribution must be estimated (Zhang & Rockette, 2005, 2007; Zhao, Lawless & McLeish, 2009; Song, Zhou & Kosorok, 2009; Zeng & Lin, 2014; Tao, Zeng & Lin, 2017). Full maximum likelihood is problematic computationally when Z is continuous or high-dimensional, and when X is high-dimensional, and CML is especially attractive in such cases. In addition, CML can be used when some $\pi(y, z)$ may be zero (Huang & Lin, 2007; Barnett, Lee & Lin, 2013), whereas IPW estimation cannot. CML estimates are not semiparametric efficient in general, but Scott & Wild (2011) showed how to augment CML so that full efficiency is achieved in certain situations. This occurs only for special discrete response models, and they did not show that their Scott–Wild (SW) estimator dominates the CML estimator in general situations. We prove here that this is indeed the case.

We make four new contributions in this article. First, we combine CML and empirical likelihood (EL) (Owen, 2001; Qin, 2017) to improve on the efficiency of CML. Second, we prove that the EL and SW estimators of $\boldsymbol{\beta}$ have the same asymptotic variance and that in general, both dominate CML. This is of practical as well as theoretical interest because the SW estimators are computationally easier to obtain than the EL estimators. Third, we show that using a value of $\boldsymbol{\alpha}$ that is known by design is asymptotically equivalent to estimating it for the EL estimators; this reduces computation somewhat. Finally, we provide numerical comparisons of CML, SW and EL estimators with both discrete and continuous response, and in the latter case compare them with semiparametric maximum likelihood. Section 2 outlines the CML, SW and EL procedures, and derives and compares asymptotic variances. Section 3 provides numerical comparisons of estimators in finite samples, and Section 4 discusses an illustration in a genetic testing context. Section 5 has concluding remarks. Details associated with derivations in Section 2 are given in the Appendix. The Supplementary Material contains two simulation studies that

complement those in the article, and a few additional technical details which are mentioned in the article.

2. ESTIMATORS AND ASYMPTOTIC VARIANCES

2.1. The CML Estimator

We will briefly review the CML estimator and set some notation. We assume that the phase 2 sample is selected using VPS. The CML estimator for $\boldsymbol{\beta}$ is based on the distribution of responses for the phase 2 individuals, conditional on covariate values and the fact that they were selected for phase 2. The conditional likelihood (CL) is then

$$L_c(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \prod_{R_i=1} P(Y = y_i | X = x_i, Z = z_i, R_i = 1) = \prod_{R_i=1} f_c(y_i | x_i, z_i; \boldsymbol{\beta}, \boldsymbol{\alpha}),$$

where, when Y is continuous,

$$f_c(y|x, z; \boldsymbol{\beta}, \boldsymbol{\alpha}) = \frac{f(y|x, z; \boldsymbol{\beta})\pi(y, z; \boldsymbol{\alpha})}{\int f(y|x, z; \boldsymbol{\beta})\pi(y, z; \boldsymbol{\alpha})dy}$$

and the conditional log-likelihood is

$$l_c(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{i=1}^{N} r_i \log f_{ci}(\boldsymbol{\beta}, \boldsymbol{\alpha}), \tag{1}$$

where $f_{ci}(\boldsymbol{\beta}, \boldsymbol{\alpha}) := f_c(y_i|x_i, z_i; \boldsymbol{\beta}, \boldsymbol{\alpha})$. With $\boldsymbol{\alpha} = \boldsymbol{\alpha}_0$ known by design, we can solve $\partial l_c(\boldsymbol{\beta}, \boldsymbol{\alpha}_0)/\partial \boldsymbol{\beta} = \sum_{i=1}^{N} \boldsymbol{S}_{1,i}(\boldsymbol{\beta}, \boldsymbol{\alpha}_0) = \mathbf{0}$, where $\boldsymbol{S}_{1,i}(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \boldsymbol{S}_1(y_i, x_i, z_i, r_i; \boldsymbol{\beta}, \boldsymbol{\alpha}) = r_i \partial \log f_{ci}(\boldsymbol{\beta}, \boldsymbol{\alpha})/\partial \boldsymbol{\beta}$; we term the resulting estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{CML0}}$. The asymptotic covariance matrix of $\sqrt{N}(\widehat{\boldsymbol{\beta}}_{\mathrm{CML0}} - \boldsymbol{\beta}_0)$ is then

$$\begin{aligned}\boldsymbol{\Sigma}_0 &= E\left\{-\frac{\partial \boldsymbol{S}_1(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0)}{\partial \boldsymbol{\beta}^T}\right\}^{-1} \operatorname{cov}\left\{\boldsymbol{S}_1(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0)\right\} E\left\{-\frac{\partial \boldsymbol{S}_1(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0)}{\partial \boldsymbol{\beta}^T}\right\}^{-T} \\ &:= \mathcal{I}_{11}^{-1} \mathcal{C}_{11} \mathcal{I}_{11}^{-1}.\end{aligned}$$

Expectations here and later are taken with respect to the true distribution of (Y, Z, X, R), given by $f(y|x, z; \boldsymbol{\beta}_0)$, $g(x, z)$ and $\pi(y, z; \boldsymbol{\alpha}_0)$. Since the components of (1) are log-likelihoods, it is easy to see that $\mathcal{I}_{11} = \mathcal{C}_{11}$, so $\boldsymbol{\Sigma}_0 = \mathcal{C}_{11}^{-1}$.

Efficiency can be increased by estimating $\boldsymbol{\alpha}$ even when it is known by design (Lawless, Kalbfleisch & Wild, 1999); this can be done by using the likelihood function for $\boldsymbol{\alpha}$ based on the observed values of $R_1, \ldots, R_N$. Let $\boldsymbol{\phi} = (\boldsymbol{\beta}^T, \boldsymbol{\alpha}^T)^T$ with true value $\boldsymbol{\phi}_0$; then the CML estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{CML}}$ and ML estimator $\widehat{\boldsymbol{\alpha}}_{\mathrm{ML}}$ solve the estimating equation

$$\sum_{i=1}^{N} \boldsymbol{S}_{\mathrm{C},i}(\boldsymbol{\phi}) = \sum_{i=1}^{N} \boldsymbol{S}_{\mathrm{C}}(y_i, x_i, z_i, r_i; \boldsymbol{\phi}) = \sum_{i=1}^{N} \begin{pmatrix} \boldsymbol{S}_{1,i}(\boldsymbol{\beta}, \boldsymbol{\alpha}) \\ \boldsymbol{S}_{\pi,i}(\boldsymbol{\alpha}) \end{pmatrix} = \sum_{i=1}^{N} \begin{pmatrix} \frac{r_i \partial \log f_{ci}}{\partial \boldsymbol{\beta}} \\ \frac{r_i - \pi_i}{\pi_i(1-\pi_i)} \frac{\partial \pi_i}{\partial \boldsymbol{\alpha}} \end{pmatrix} = \mathbf{0}, \tag{2}$$

where $\pi_i = \pi(y_i, z_i; \boldsymbol{\alpha})$, $f_{ci} = f_{ci}(\boldsymbol{\beta}, \boldsymbol{\alpha})$ and $\boldsymbol{S}_{\pi,i}(\boldsymbol{\alpha}) = \boldsymbol{S}_{\pi}(y_i, z_i, r_i; \boldsymbol{\alpha}) = (r_i - \pi_i)/\{\pi_i(1 - \pi_i)\}\partial \pi_i/\partial \boldsymbol{\alpha}$. Then $\sqrt{N}\{(\widehat{\boldsymbol{\beta}}_{\mathrm{CML}}, \widehat{\boldsymbol{\alpha}}_{\mathrm{ML}})\text{-}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0)\}$ is asymptotically normal with covariance matrix $\mathcal{I}_{\mathrm{CML}}^{-1} \mathcal{C}_{\mathrm{CML}} \mathcal{I}_{\mathrm{CML}}^{-T}$, where

$$\mathcal{C}_{\mathrm{CML}} = \begin{bmatrix} \mathcal{C}_{11} & \mathcal{C}_{1\pi} \\ \mathcal{C}_{1\pi}^T & \mathcal{C}_{\pi\pi} \end{bmatrix} = \begin{bmatrix} E\{\boldsymbol{S}_1(\boldsymbol{\phi}_0)\boldsymbol{S}_1^T(\boldsymbol{\phi}_0)\} & E\{\boldsymbol{S}_1(\boldsymbol{\phi}_0)\boldsymbol{S}_{\pi}^T(\boldsymbol{\phi}_0)\} \\ E\{\boldsymbol{S}_{\pi}(\boldsymbol{\phi}_0)\boldsymbol{S}_1^T(\boldsymbol{\phi}_0)\} & E\{\boldsymbol{S}_{\pi}(\boldsymbol{\phi}_0)\boldsymbol{S}_{\pi}^T(\boldsymbol{\phi}_0)\} \end{bmatrix},$$

and

$$\mathcal{I}_{\mathrm{CML}} = E\left\{-\frac{\partial \boldsymbol{S}_{\mathrm{C}}(\boldsymbol{\phi}_0)}{\partial \boldsymbol{\phi}^T}\right\} = \begin{bmatrix} \mathcal{I}_{11} & \mathcal{I}_{1\pi} \\ \mathbf{0} & \mathcal{I}_{\pi\pi} \end{bmatrix}.$$

Since

$$\mathcal{I}_{\mathrm{CML}}^{-1} = \begin{bmatrix} \mathcal{I}_{11}^{-1} & -\mathcal{I}_{11}^{-1}\mathcal{I}_{1\pi}\mathcal{I}_{\pi\pi}^{-1} \\ \mathbf{0} & \mathcal{I}_{\pi\pi}^{-1} \end{bmatrix},$$

and $\mathcal{I}_{1\pi} = E\{-\partial \boldsymbol{S}_1(\boldsymbol{\phi}_0)/\partial \boldsymbol{\alpha}^T\} = E\{\boldsymbol{S}_1(\boldsymbol{\phi}_0)\boldsymbol{S}_\pi^T(\boldsymbol{\phi}_0)\} = \mathcal{C}_{1\pi}$, $\mathcal{I}_{11} = \mathcal{C}_{11}$, and $\mathcal{I}_{\pi\pi} = \mathcal{C}_{\pi\pi}$ (details are in Appendix A1), we find that the asymptotic covariance matrix of $\sqrt{N}(\widehat{\boldsymbol{\beta}}_{\mathrm{CML}} - \boldsymbol{\beta}_0)$ is

$$\begin{aligned} \boldsymbol{\Sigma}_1 &= \mathcal{I}_{11}^{-1} - \mathcal{I}_{11}^{-1}\mathcal{I}_{1\pi}\mathcal{I}_{\pi\pi}^{-1}\mathcal{I}_{1\pi}^T\mathcal{I}_{11}^{-1} \\ &= \mathcal{C}_{11}^{-1} - \mathcal{C}_{11}^{-1}\mathcal{C}_{1\pi}\mathcal{C}_{\pi\pi}^{-1}\mathcal{C}_{1\pi}^T\mathcal{C}_{11}^{-1} \le \mathcal{C}_{11}^{-1} = \boldsymbol{\Sigma}_{\mathbf{0}}, \end{aligned} \tag{3}$$

consistent with the well-known result that using the true value of $\boldsymbol{\alpha}$ does not reduce the asymptotic variance of $\widehat{\boldsymbol{\beta}}$. Here and subsequently $A \le B$ for positive definite symmetric matrices A and B means that $B - A$ is positive semi-definite.

2.2. The SW Estimator

Noting that $l_c(\boldsymbol{\beta}, \boldsymbol{\alpha})$ also carries information about $\boldsymbol{\alpha}$, Scott & Wild (2011) proposed an estimator that uses the score function $\boldsymbol{S}_2(y, x, z, r; \boldsymbol{\beta}, \boldsymbol{\alpha}) = r\partial \log f_c(y|x, z; \boldsymbol{\beta}, \boldsymbol{\alpha})/\partial \boldsymbol{\alpha}$. This was based on noticing that when Y is binary and when $\pi(y, z; \boldsymbol{\alpha})$ depends only on a finite partition of Z, the semiparametric efficient maximum likelihood (ML) estimator developed in Scott & Wild (1997) satisfies the estimating equation

$$\begin{aligned} \sum_{i=1}^N \boldsymbol{S}_{\mathrm{E},i}(\boldsymbol{\phi}) = \sum_{i=1}^N \boldsymbol{S}_{\mathrm{E}}(y_i, x_i, z_i, r_i; \boldsymbol{\phi}) &= \sum_{i=1}^N \begin{pmatrix} \boldsymbol{S}_{1,i}(\boldsymbol{\beta}, \boldsymbol{\alpha}) \\ \boldsymbol{S}_{\pi,i}(\boldsymbol{\alpha}) - \boldsymbol{S}_{2,i}(\boldsymbol{\beta}, \boldsymbol{\alpha}) \end{pmatrix} \\ &= \mathbf{0}. \end{aligned} \tag{4}$$

However, this estimator is not fully efficient in general regression settings, for example with continuous Y, and Scott & Wild (2011) did not establish that it dominates the CML estimator based on (2) in general settings; we now establish this.

The asymptotic covariance matrix of the solution to the estimating equation (4), which we denote as $\widehat{\boldsymbol{\phi}}_{\mathrm{SW}}$, is given by $\mathcal{I}_{\mathrm{SW}}^{-T}\mathcal{C}_{\mathrm{SW}}\mathcal{I}_{\mathrm{SW}}^{-1}$, where

$$\begin{aligned} \mathcal{I}_{\mathrm{SW}} &= \begin{bmatrix} -E\left\{\partial \boldsymbol{S}_1(\boldsymbol{\phi}_0)/\partial \boldsymbol{\beta}^T\right\} & -E\left\{\partial \boldsymbol{S}_1(\boldsymbol{\phi}_0)/\partial \boldsymbol{\alpha}^T\right\} \\ E\left\{\partial \boldsymbol{S}_2(\boldsymbol{\phi}_0)/\partial \boldsymbol{\beta}^T\right\} & -E\left\{\partial(\boldsymbol{S}_\pi(\boldsymbol{\phi}_0) - \boldsymbol{S}_2(\boldsymbol{\phi}_0))/\partial \boldsymbol{\alpha}^T\right\} \end{bmatrix} \\ &:= \begin{bmatrix} \mathcal{I}_{11} & \mathcal{I}_{1\pi} \\ -\mathcal{I}_{21} & \mathcal{I}_{\pi\pi} - \mathcal{I}_{2\pi} \end{bmatrix}, \\ \mathcal{C}_{\mathrm{SW}} &= \begin{bmatrix} \mathcal{C}_{11} & \mathcal{C}_{1\pi} - \mathcal{C}_{12} \\ \mathcal{C}_{1\pi}^T - \mathcal{C}_{12}^T & \mathcal{C}_{\pi\pi} - \mathcal{C}_{2\pi}^T - \mathcal{C}_{2\pi} + \mathcal{C}_{22} \end{bmatrix}, \end{aligned}$$

with $\mathcal{C}_{ij} = E\{\boldsymbol{S}_i(\boldsymbol{\phi}_0)\boldsymbol{S}_j^T(\boldsymbol{\phi}_0)\}$ for $i, j \in \{1, 2, \pi\}$.

In the Appendix we show that $\mathcal{I}_{1\pi} = \mathcal{C}_{1\pi} = \mathcal{C}_{12}$ and $\mathcal{I}_{2\pi} = \mathcal{C}_{2\pi} = \mathcal{C}_{22}$. Also, $\mathcal{C}_{22}$ is symmetric, so $\mathcal{C}_{2\pi} = \mathcal{C}_{22}^T = \mathcal{C}_{2\pi}^T$. Thus the asymptotic covariance matrix of $\sqrt{N}(\widehat{\boldsymbol{\phi}}_{\mathrm{SW}} - \boldsymbol{\phi}_0)$ is

$$\begin{aligned}\mathcal{I}_{\mathrm{SW}}^{-T}\mathcal{C}_{\mathrm{SW}}\mathcal{I}_{\mathrm{SW}}^{-1} &= \begin{bmatrix} \mathcal{C}_{11} & \mathcal{C}_{12} \\ -\mathcal{C}_{12}^T & \mathcal{C}_{\pi\pi} - \mathcal{C}_{22} \end{bmatrix}^{-1} \begin{bmatrix} \mathcal{C}_{11} & \mathbf{0} \\ \mathbf{0} & \mathcal{C}_{\pi\pi} - \mathcal{C}_{22} \end{bmatrix} \begin{bmatrix} \mathcal{C}_{11} & -\mathcal{C}_{12} \\ \mathcal{C}_{12}^T & \mathcal{C}_{\pi\pi} - \mathcal{C}_{22} \end{bmatrix}^{-1} \\ &= \begin{bmatrix} \{\mathcal{C}_{11} + \mathcal{C}_{12}(\mathcal{C}_{\pi\pi} - \mathcal{C}_{22})^{-1}\mathcal{C}_{12}^T\}^{-1} & \mathbf{0} \\ \mathbf{0} & \{\mathcal{C}_{12}^T\mathcal{C}_{11}^{-1}\mathcal{C}_{12} + \mathcal{C}_{\pi\pi} - \mathcal{C}_{22}\}^{-1} \end{bmatrix}\end{aligned}$$

and the asymptotic covariance matrix of $\sqrt{N}(\widehat{\boldsymbol{\beta}}_{\mathrm{SW}} - \boldsymbol{\beta}_0)$ is

$$\begin{aligned}\boldsymbol{\Sigma}_2 &= \{\mathcal{C}_{11} + \mathcal{C}_{12}(\mathcal{C}_{\pi\pi} - \mathcal{C}_{22})^{-1}\mathcal{C}_{12}^T\}^{-1} \\ &= \{\mathcal{C}_{11} + \mathcal{C}_{1\pi}(\mathcal{C}_{\pi\pi} - \mathcal{C}_{22})^{-1}\mathcal{C}_{1\pi}^T\}^{-1} \\ &= \mathcal{C}_{11}^{-1} - \mathcal{C}_{11}^{-1}\mathcal{C}_{1\pi}(\mathcal{C}_{\pi\pi} - \mathcal{C}_{22} + \mathcal{C}_{1\pi}^T\mathcal{C}_{11}^{-1}\mathcal{C}_{1\pi})^{-1}\mathcal{C}_{1\pi}^T\mathcal{C}_{11}^{-1}.\end{aligned}$$

Interestingly, $\widehat{\boldsymbol{\beta}}_{\mathrm{SW}}$ and $\widehat{\boldsymbol{\alpha}}_{\mathrm{SW}}$ are asymptotically uncorrelated. Since $\sqrt{N}(\widehat{\boldsymbol{\beta}}_{\mathrm{CML}} - \boldsymbol{\beta}_0)$ has asymptotic covariance matrix $\mathcal{C}_{11}^{-1} - \mathcal{C}_{11}^{-1}\mathcal{C}_{1\pi}\mathcal{C}_{\pi\pi}^{-1}\mathcal{C}_{1\pi}^T\mathcal{C}_{11}^{-1}$ from (3), whether SW will improve upon CML depends on $\mathcal{C}_R := \mathcal{C}_{22} - \mathcal{C}_{1\pi}^T\mathcal{C}_{11}^{-1}\mathcal{C}_{1\pi} = \mathcal{C}_{22} - \mathcal{C}_{12}^T\mathcal{C}_{11}^{-1}\mathcal{C}_{12}$. This is the asymptotic variance of $\boldsymbol{S}_2$ given $\boldsymbol{S}_1$ and so it is positive definite; SW is therefore guaranteed to improve upon CML. That is, $\boldsymbol{\Sigma}_2 \leq \boldsymbol{\Sigma}_1$.

2.3. The EL Estimator

The SW approach uses three estimating functions $\boldsymbol{S}_1(\boldsymbol{\beta}, \boldsymbol{\alpha})$, $\boldsymbol{S}_2(\boldsymbol{\beta}, \boldsymbol{\alpha})$ and $\boldsymbol{S}_\pi(\boldsymbol{\alpha})$, and it is possible that some information is lost by using only the difference of $\boldsymbol{S}_\pi$ and $\boldsymbol{S}_2$. EL allows us to exploit more estimating functions than the number of parameters (Qin & Lawless, 1994; Owen, 2001; Qin, 2017). Qin, Zhang & Leung (2009), Xie & Zhang (2017) and Che, Han & Lawless (2020) have used this to formulate EL approaches in other missing data settings. For two-phase, response-dependent samples, Zhou et al. (2011) give an example of applying EL but their method still needs to model the covariate distribution; we avoid this. We note that another approach to combining estimating functions is generalized method of moments, or GMM (Hansen, 1982; Newey & Smith, 2004). It is known that GMM and EL estimators based on a specific set of estimating functions have the same asymptotic distributions, but EL has some higher order asymptotic advantages (Newey & Smith, 2004).

Let the dimensions of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ be p and q, respectively and let $\mathbf{U}(\boldsymbol{\phi}) = (\boldsymbol{S}_1(\boldsymbol{\beta}, \boldsymbol{\alpha})^T, \boldsymbol{S}_2(\boldsymbol{\beta}, \boldsymbol{\alpha})^T, \boldsymbol{S}_\pi(\boldsymbol{\alpha})^T)^T$. Then, EL involves maximizing the empirical likelihood $\prod_{i=1}^N p_i$ with respect to $\mathbf{p} = (p_1, \ldots, p_N)$, $\boldsymbol{\beta}, \boldsymbol{\alpha}$ subject to $p_i \geq 0$, $\sum_{i=1}^N p_i = 1$ and the $p + 2q$ constraints

$$\sum_{i=1}^N p_i \mathbf{U}_i(\boldsymbol{\phi}) = \sum_{i=1}^N p_i \begin{pmatrix} r_i \frac{\partial \log f_{ci}}{\partial \boldsymbol{\beta}} \\ r_i \frac{\partial \log f_{ci}}{\partial \boldsymbol{\alpha}} \\ \left(\frac{r_i}{\pi_i} - \frac{1-r_i}{1-\pi_i}\right) \frac{\partial \pi_i}{\partial \boldsymbol{\alpha}} \end{pmatrix} = \mathbf{0}. \tag{5}$$

This produces an EL estimator of $\boldsymbol{\phi}$, which we denote as $\widehat{\boldsymbol{\phi}}_{\mathrm{EL}}$. By Qin & Lawless (1994), it is asymptotically normal, with covariance matrix

$$(\mathcal{J}^T\mathcal{C}^{-1}\mathcal{J})^{-1} = \left[E\left\{\frac{\partial \mathbf{U}(\boldsymbol{\phi}_0)}{\partial \boldsymbol{\phi}^T}\right\}^T E\left\{\mathbf{U}(\boldsymbol{\phi}_0)\mathbf{U}^T(\boldsymbol{\phi}_0)\right\}^{-1} E\left\{\frac{\partial \mathbf{U}(\boldsymbol{\phi}_0)}{\partial \boldsymbol{\phi}^T}\right\}\right]^{-1},$$

and in the Appendix we show

$$\mathcal{C} = E\{\mathbf{U}(\boldsymbol{\phi}_0)\mathbf{U}^T(\boldsymbol{\phi}_0)\} = \begin{bmatrix} \mathcal{C}_{11} & \mathcal{C}_{12} & \mathcal{C}_{12} \\ \mathcal{C}_{12}^T & \mathcal{C}_{22} & \mathcal{C}_{22} \\ \mathcal{C}_{12}^T & \mathcal{C}_{22}^T & \mathcal{C}_{\pi\pi} \end{bmatrix},$$

$$\mathcal{J} = -E\left\{\frac{\partial \mathbf{U}(\boldsymbol{\phi}_0)}{\partial \boldsymbol{\phi}^T}\right\} = \begin{bmatrix} \mathcal{C}_{11} & \mathcal{C}_{12} \\ \mathcal{C}_{12}^T & \mathcal{C}_{22} \\ \mathbf{0} & \mathcal{C}_{\pi\pi} \end{bmatrix}.$$

We now show that $\widehat{\boldsymbol{\beta}}_{\mathrm{SW}}$ and $\widehat{\boldsymbol{\beta}}_{\mathrm{EL}}$ have the same asymptotic covariance matrix. We write the inverse of the asymptotic covariance matrix of $\sqrt{N}(\widehat{\boldsymbol{\phi}}_{\mathrm{EL}} - \boldsymbol{\phi}_0)$ in block form as

$$\begin{bmatrix} V^{11} & V^{12} \\ V^{21} & V^{22} \end{bmatrix},$$

and note that the asymptotic covariance matrix of $\sqrt{N}(\widehat{\boldsymbol{\beta}}_{\mathrm{EL}} - \boldsymbol{\beta}_0)$ can then be expressed as

$$\boldsymbol{\Sigma}_3 := \{V^{11} - V^{12}(V^{22})^{-1}V^{21}\}^{-1}.$$

In the Appendix we show that

$$V^{11} = \mathcal{C}_{11} + \mathcal{C}_{12}(\mathcal{C}_{\pi\pi} - \mathcal{C}_{22})^{-1}\mathcal{C}_{12}^T$$

and $V^{12} = \mathbf{0}$. Therefore,

$$\begin{aligned} \boldsymbol{\Sigma}_3 &= \{V^{11} - V^{12}(V^{22})^{-1}V^{21}\}^{-1} \\ &= \{\mathcal{C}_{11} + \mathcal{C}_{12}(\mathcal{C}_{\pi\pi} - \mathcal{C}_{22})^{-1}\mathcal{C}_{12}^T\}^{-1} = \boldsymbol{\Sigma}_2. \end{aligned}$$

Thus $\widehat{\boldsymbol{\beta}}_{\mathrm{SW}}$ and $\widehat{\boldsymbol{\beta}}_{\mathrm{EL}}$ have the same asymptotic covariance matrix. This is of practical as well as theoretical importance, since the SW estimators are computationally easier to obtain than the EL estimators.

We remark that since phase 2 sampling is by design, the true value of $\boldsymbol{\alpha}_0$ is known. Qin, Zhang & Leung (2009) found with estimating functions used in the augmented inverse probability weighted (AIPW) estimator (Robins et al., 1994) that efficiency could in some cases be improved further by using the known $\boldsymbol{\alpha}_0$, unlike the situation with CML. Xie & Zhang (2017) found a similar effect in another missing data setting. We call this estimator $\hat{\boldsymbol{\beta}}_{\mathrm{EL0}}$ in the setting here. Another alternative estimator uses the maximum likelihood estimate $\widehat{\boldsymbol{\alpha}}_{\mathrm{ML}}$ in (5); Qin, Zhang & Leung (2009) called this a pseudo EL estimator. We denote the resulting estimator as $\widehat{\boldsymbol{\beta}}_{\mathrm{PEL}}$. In our setting these two estimators respectively involve replacing $\boldsymbol{\phi}$ in (5) with $(\boldsymbol{\beta}^T, \boldsymbol{\alpha}_0^T)^T$ to give $\tilde{\mathbf{U}}(\boldsymbol{\beta}, \boldsymbol{\alpha}_0)$, and with $(\boldsymbol{\beta}^T, \widehat{\boldsymbol{\alpha}}_{\mathrm{ML}}^T)^T$ to give $\tilde{\mathbf{U}}(\boldsymbol{\beta}, \widehat{\boldsymbol{\alpha}}_{\mathrm{ML}})$. The asymptotic covariance matrices for $\hat{\boldsymbol{\beta}}_{\mathrm{EL0}}$ and $\hat{\boldsymbol{\beta}}_{\mathrm{PEL}}$, denoted as $\boldsymbol{\Sigma}_4$ and $\boldsymbol{\Sigma}_5$ respectively, can both be shown to equal $(\widetilde{\mathcal{I}}^T\mathcal{C}^{-1}\widetilde{\mathcal{I}})^{-1}$, where $\mathcal{C}$ is given above and where

$$\widetilde{\mathcal{I}} = -E\left\{\frac{\partial \tilde{\mathbf{U}}(\boldsymbol{\beta}, \boldsymbol{\alpha}_0)}{\partial \boldsymbol{\beta}^T}\right\} = \begin{bmatrix} \mathcal{C}_{11} \\ \mathcal{C}_{21} \\ \mathbf{0} \end{bmatrix}.$$

Similarly to the derivation above, we find $\boldsymbol{\Sigma}_4^{-1} = \boldsymbol{\Sigma}_5^{-1} = V^{11}$ and therefore $\boldsymbol{\Sigma}_4 = \boldsymbol{\Sigma}_5 = \boldsymbol{\Sigma}_3 = \boldsymbol{\Sigma}_2$. Thus, in our setting all the EL estimators of $\boldsymbol{\beta}$ have the same asymptotic variance. Another option

is to use known $\boldsymbol{\alpha}_0$ only in $\boldsymbol{S}_2(\boldsymbol{\beta}, \boldsymbol{\alpha})$ and to estimate $\boldsymbol{\alpha}$; this again gives the same asymptotic variance.

The asymptotic variances for all estimators of $\boldsymbol{\beta}$ depend on the phase 2 selection model $\pi(y, z; \boldsymbol{\alpha})$ that is used in the system of estimating functions. It has been shown that efficiency can be improved by using a "highly stratified" model that includes the actual sampling probability model but incorporates a finer stratification of (Y, Z) than was actually used for sampling in the study, (e.g., Lawless, Kalbfleisch & Wild, 1999; Scott & Wild, 2011). For example, if phase 2 sampling depended only on Y, we can do better by using a working model that involves both Y and Z than with one that involves only Y. This is implicit in the asymptotic variances given here, and numerical studies in the next section illustrate the efficiency gains in finite samples.

3. SIMULATION STUDIES

The implementation of the EL estimators is based on Lagrange multiplier methods and Newton–Raphson iterations. We write the Lagrangian function of the constrained optimization in Section 2.3 as

$$\mathcal{L} = \sum_{i=1}^{N} \log p_i + \sum_{i=1}^{N} \lambda^T \mathbf{U}_i(\boldsymbol{\phi}) + \mu(\sum_{i=1}^{N} p_i - 1).$$

The Lagrange multiplier λ can be found by solving $\sum_{i=1}^{N}[\mathbf{U}_i(\boldsymbol{\phi})/\{1 + \lambda^T \mathbf{U}_i(\boldsymbol{\phi})\}] = \mathbf{0}$, which is equivalent to minimizing the function

$$l(\boldsymbol{\phi}, \lambda) = -\sum_{i=1}^{N} \log\{1 + \lambda^T \mathbf{U}_i(\boldsymbol{\phi})\}$$

with respect to λ. The algorithm we use is as follows.

We start from an initial estimate of $\boldsymbol{\phi}$, say the CML estimate, which we denote as $\boldsymbol{\phi}^0$. In the kth iteration ($k = 1, 2, \ldots$) we have steps

Step k.1. Given $\boldsymbol{\phi}^{k-1}$, we use the optimization function "fmincon" in MATLAB to solve for the values λ^k which minimize $l(\boldsymbol{\phi}^{k-1}, \lambda)$.

Step k.2. We compute the first-order derivative vector (Jacobian) of $l(\boldsymbol{\phi}, \lambda^k)$ with respect to $\boldsymbol{\phi}$ and evaluate it at $\boldsymbol{\phi}^{k-1}$ as

$$\frac{\partial l(\boldsymbol{\phi}^{k-1}, \lambda^k)}{\partial \boldsymbol{\phi}} = -\sum_{i=1}^{N} \frac{\mathbf{U}_{\phi,i}(\boldsymbol{\phi}^{k-1})^T \lambda^k}{1 + \lambda^{k^T}(\boldsymbol{\phi}^{k-1})\mathbf{U}_i(\boldsymbol{\phi}^{k-1})} := J(\boldsymbol{\phi}^{k-1}, \lambda^k),$$

where the entries in matrices $U_{\phi,i}(\boldsymbol{\phi}) = \partial U_i(\boldsymbol{\phi})/\partial \boldsymbol{\phi}^T$ are computed by numerical differentiation.

The second-order derivative (Hessian) matrix is approximated by

$$\frac{\partial^2 l(\boldsymbol{\phi}^{k-1}, \lambda^k)}{\partial \boldsymbol{\phi} \partial \boldsymbol{\phi}^T} \approx -\frac{\partial l(\boldsymbol{\phi}^{k-1}, \lambda^k)}{\partial \boldsymbol{\phi}} \frac{\partial l(\boldsymbol{\phi}^{k-1}, \lambda^k)}{\partial \boldsymbol{\phi}^T} := H(\boldsymbol{\phi}^{k-1}, \lambda^k).$$

Step k.3. We update the estimate of $\boldsymbol{\phi}$ with

$$\boldsymbol{\phi}^k = \boldsymbol{\phi}^{k-1} - s\{H(\boldsymbol{\phi}^{k-1}, \lambda^k)\}^{-1} J(\boldsymbol{\phi}^{k-1}, \lambda^k),$$

where s is an adaptive step length which guarantees $l(\boldsymbol{\phi}^k, \lambda^k) \geq l(\boldsymbol{\phi}^{k-1}, \lambda^k)$, and can be found via, for example, bi-section search.

We repeat the above steps until convergence. This so-called saddle-point algorithm is widely recommended in the EL literature (e.g., Owen, 2001). In our settings it converges acceptably fast.

We describe numerical comparisons of the CML, SW and EL estimators for two scenarios here. The first is similar to one in Scott & Wild (2011) and involves a binary response Y with two covariates X and Z. Study 2 involves a continuous response. Two additional related scenarios are considered in the Supplementary Material, as described below.

3.1. Simulation Study 1

This study involves a binary covariate X and continuous covariate Z, which are correlated. We consider a phase 1 sample of 10,000 subjects with data generated as follows. A continuous standard normal covariate Z_i is first generated and then a Bernoulli covariate X_i is generated with probability $P(X_i = 1) = 0.2\mathrm{I}(Z_i < 0) + 0.5\mathrm{I}(Z_i \geq 0)$. We then generate the response Y_i using a logistic regression model; with $\mathrm{expit}(u)$ denoting $e^u/(1 + e^u)$, it is

$$P(Y = 1|x, z) = \mathrm{expit}(\beta_c + \beta_x x + \beta_z z), \tag{6}$$

with $\boldsymbol{\beta}_0 = (-2.8, 0.5, 1)$. This gives marginal probabilities for $Y = 0$ and $Y = 1$ approximately equal to 0.9 and 0.1. We use a Bernoulli VPS selection scheme for phase 2, according to the sampling model

$$P(R = 1|y, z) = \pi(y, z; \boldsymbol{\alpha}) = \mathrm{expit}(\alpha_c + \alpha_y y), \tag{7}$$

where we chose $\boldsymbol{\alpha}_0 = (-4.1, 2.5)$. This gives marginal selection probabilities for $Y = 0$ and $Y = 1$ of approximately 0.0163 and 0.168. We can use these known selection probabilities in our estimating functions; we refer to the CML estimator in this case as CML0. Estimation efficiency can be increased by estimating selection probabilities using the model (7), and we denote estimates obtained using this model with the suffix "est" in Table 1. It is possible, however, to further increase efficiency by using a more highly stratified selection model that conditions on observed z values, similar to calibration or post-stratification in sampling contexts. We consider two such models, referred to with the suffixes "sat1" and "sat2" in Table 1. For sat1 we use a binary covariate $v = \mathrm{I}(z > 0.5)$ and the model

$$P(R = 1|y, v) = \pi_{sat1}(y, v; \boldsymbol{\alpha}) = \mathrm{expit}(\alpha_c + \alpha_y y + \alpha_v v + \alpha_{yv} yv). \tag{8}$$

The sat2 model uses the continuous covariate z in a more highly stratified logistic regression model for phase 2 selection, namely

$$P(R = 1|y, z) = \pi_{sat2}(y, z; \boldsymbol{\alpha}) = \mathrm{expit}(\alpha_c + \alpha_y y + \alpha_z z + \alpha_{yz} yz). \tag{9}$$

Note that working models (8) and (9) both include the true phase 2 sampling model (7) as special cases.

We also considered pseudo empirical likelihood (PEL) estimators, where the $\boldsymbol{\alpha}$ parameters in models (7), (8) or (9) are first estimated by maximum likelihood from $\boldsymbol{S}_\pi(\boldsymbol{\alpha}) = 0$ and then fixed in the estimating function $\mathbf{U}(\boldsymbol{\phi}) = \mathbf{U}(\boldsymbol{\beta}, \widehat{\boldsymbol{\alpha}}_{\mathrm{ML}})$. This EL procedure is slightly easier to implement since the estimating function $\boldsymbol{S}_\pi(\widehat{\boldsymbol{\alpha}}_{\mathrm{ML}})$ equals zero.

We mention that in this example the estimating functions $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$ are not linearly independent when the π_{sat1} model is used. Then $\dim(\boldsymbol{\beta}) = 3$ and $\dim(\boldsymbol{\alpha}) = 4$ so the dimension of $(\boldsymbol{S}_1^T, \boldsymbol{S}_2^T)^T$ is 7. However, in Section S.3 in the Supplementary Material we show that the actual rank of these 7 estimating equations is 4. Therefore, we use here only the first element of $\boldsymbol{S}_2$ for the EL estimator in this case. This phenomenon is an example of the well-known fact that $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are not identifiable from the conditional likelihood $l_c(\boldsymbol{\beta}, \boldsymbol{\alpha})$ alone in this setting.

TABLE 1: Simulation results for Study 1.

	Mean (Empirical SE)[Estimated SE]		
Method	β_c ($\beta_{c0} = -2.8$)	$\beta_z(\beta_{z0} = 0.5)$	$\beta_x(\beta_{x0} = 1)$
CML0	−2.814 (0.168)[0.165]	0.519 (0.258)[0.257]	0.997 (0.244)[0.252]
CML-est	−2.802 (0.117)[0.123]	0.519 (0.258)[0.257]	0.997 (0.244)[0.252]
CML-sat1	−2.803 (0.116)[0.121]	0.509 (0.204)[0.200]	0.997 (0.245)[0.251]
CML-sat2	−2.803 (0.113)[0.119]	0.514 (0.123)[0.123]	0.997 (0.244)[0.251]
EL-est	−2.802 (0.117)[0.123]	0.519 (0.258)[0.257]	0.997 (0.244)[0.252]
EL-sat1	−2.803 (0.115)[0.121]	0.502 (0.133)[0.133]	0.997 (0.244)[0.250]
EL-sat2	−2.803 (0.113)[0.119]	0.515 (0.123)[0.122]	0.997 (0.245)[0.251]
SW-est	−2.802 (0.117)[0.123]	0.519 (0.258)[0.257]	0.997 (0.244)[0.252]
EL-sat1	−2.803 (0.115)[0.121]	0.509 (0.139)[0.129]	0.997 (0.245)[0.250]
EL-sat2	−2.803 (0.113)[0.119]	0.514 (0.123)[0.122]	0.997 (0.244)[0.251]

In Table 1, we compare the performance of CML, SW and EL estimators based on 500 simulations, using each of the three π models (7)–(9) as well as the known selection probabilities, denoted for CML as CML0. The EL0 and PEL estimator with each π model are asymptotically equivalent to the corresponding EL estimator so are omitted; their finite sample performances are close to those of the EL estimators. We show empirical standard deviations and average standard errors (SEs) for each estimator; SEs are obtained by estimating asymptotic covariance matrices with sample covariance matrices evaluated at estimates of $\boldsymbol{\phi}$. These are labelled as empirical and estimated SE respectively in the table and they are seen to be close in value. In this case, CML performs about as well as the EL and SW methods. A substantial efficiency gain for estimation of β_z, the coefficient for the covariate that is known for all individuals, occurs when the stratified selection model (8) is used instead of (7) for the EL and SW estimators. A big increase in efficiency for CML and small further increases in efficiency for EL and SW result from using the more highly stratified model (9).

In the Supplementary Material, we present a simulation study based on the same data generating model as here, in which BSS instead of VPS is used for phase 2 selection. It shows results very similar to the ones in Table 1.

3.2. Simulation Study 2

In Study 2, we simulate a normal linear regression model, $f(y|x, z; \beta) = N(\beta_c + \beta_x x + \beta_z z, \sigma^2)$, with stratified phase 2 sampling based on Y. We generate X from a standard normal distribution; an auxiliary random variable W is generated by $W|X \sim \text{Exp}(0.2\text{I}(X > 0) + 0.5\text{I}(X \leq 0))$, and then Z is a categorization of W based on break-points at 0.5 and 1.5, which are approximately the 1/3 and 2/3 quantiles of W, and coded as 0, 1, 2 in the first, second and third tertile, respectively. We take $\beta_0 = (0, 0.5, 0.5)$ and $\sigma = 1$. We consider a phase 2 sampling plan often used in genetic epidemiological studies, in which the lower and upper tails of the Y distribution are over-sampled. In phase 1, (y, z) are observed for every individual and in phase 2, we use VPS with $P(R = 1|y, z) = \pi(y; \alpha) = \sum_{k=1}^{K} \alpha_k \text{I}(y \in S_k)$, where the values α_k are specified probabilities and $S_k = (c_{k-1}, c_k]$ is the kth stratum in a partition of the sample space for Y. We set $K = 3$, and $S_1 = (-\infty, -0.05], S_2 = (-0.05, 0.9], S_3 = (0.9, \infty)$; the values −0.05 and 0.9 are approximately the first and third quartiles of Y. We consider frequently used extreme response

TABLE 2: Simulation results for Study 2.

Method	Mean (Empirical SE)[Estimated SE]			
	β_c ($\beta_{c0} = -2.8$)	$\beta_z(\beta_{z0} = 0.5)$	$\beta_x(\beta_{x0} = 1)$	$\sigma(\sigma_0 = 1)$
CML0	0.007 (0.118)[0.117]	0.496 (0.087)[0.090]	0.505 (0.075)[0.077]	0.991 (0.049)[0.050]
CML-est	0.006 (0.106)[0.109]	0.497 (0.087)[0.090]	0.506 (0.075)[0.076]	0.992 (0.049)[0.050]
CML-sat	0.003 (0.092)[0.099]	0.500 (0.071)[0.078]	0.507 (0.075)[0.076]	0.093 (0.048)[0.050]
EL-est	0.006 (0.104)[0.104]	0.497 (0.087)[0.090]	0.505 (0.075)[0.076]	0.992 (0.049)[0.050]
EL-sat	0.004 (0.088)[0.084]	0.500 (0.068)[0.067]	0.506 (0.076)[0.075]	0.992 (0.049)[0.049]
SW-est	0.006 (0.103)[0.104]	0.497 (0.087)[0.090]	0.506 (0.075)[0.076]	0.992 (0.049)[0.050]
SW-sat	0.002 (0.085)[0.084]	0.501 (0.067)[0.067]	0.507 (0.074)[0.075]	0.993 (0.048)[0.049]
SML-ZLM	0.002 (0.075)	0.497 (0.058)	0.505 (0.064)	1.004 (0.034)
SML-L	0.003 (0.079)	0.500 (0.064)	0.505 (0.067)	1.004 (0.039)

sampling and set $\alpha_0 = (\alpha_1, \alpha_2, \alpha_3) = (0.2, 0, 0.3)$, so about 17% of a phase 1 sample is selected for phase 2.

In this scenario, the conditional density function for Y is

$$f_c(y|x, z; \boldsymbol{\beta}, \boldsymbol{\alpha}) = \frac{\exp\{-(y - \beta_c - \beta_x x - \beta_z z)^2/(2\sigma^2)\} \sum_{k=1}^{3} \alpha_k \mathrm{I}(y \in S_k)}{\int \exp\{-(y - \beta_c - \beta_x x - \beta_z z)^2/(2\sigma^2)\} \sum_{k=1}^{3} \alpha_k \mathrm{I}(y \in S_k) dy}$$
$$= \frac{\exp\{-(y - \beta_c - \beta_x x - \beta_z z)^2/(2\sigma^2)\} \sum_{k=1}^{3} \alpha_k \mathrm{I}(y \in S_k)}{\sum_{k=1}^{3} \alpha_k \{F((c_k - \beta_c - \beta_x x - \beta_z z)/\sigma) - F((c_{k-1} - \beta_c - \beta_x x - \beta_z z)/\sigma)\}},$$

where F is the standard normal distribution function. The parameter $\boldsymbol{\alpha} = (\alpha_1, 0, \alpha_3)$, though known, can be estimated through estimating equation

$$\left(\frac{\sum_{y_i \in S_1} R_i}{\sum_i \mathrm{I}(y_i \in S_1)} - \widehat{\alpha}_1, \frac{\sum_{y_i \in S_3} R_i}{\sum_i \mathrm{I}(y_i \in S_3)} - \widehat{\alpha}_3 \right) = \mathbf{0}.$$

We also consider a more highly stratified model $\pi(y, z)$; we define $\tilde{\boldsymbol{\alpha}} = (\alpha_{11}, \alpha_{12}, \alpha_{13}, \alpha_{31}, \alpha_{32}, \alpha_{33})$, where $\alpha_{ij} = P(R = 1 | Y \in S_i, Z = j + 1)$. We compare the performances of CML, EL, and SW estimators in 500 simulations with phase 1 sample size $N = 1{,}000$ in Table 2. Two semiparametric maximum likelihood (SML) estimators are also included, one based on the likelihood in Zhao et al. (2009), referred to as SML-ZLM, and one based on the estimator in Lawless (1997), referred to as SML-L. The SML-L estimator uses only the phase 1 stratum information for Y, Z whereas the SML-ZLM estimator uses the exact values for Y, Z. The results for EL and SW estimators are similar to those for study 1; CML is slightly less efficient than these estimators. However, we see that the SW and EL estimators are less efficient than SML; SEs for SW and EL regression coefficients are 12–18% larger than those for SML-ZLM.

In the Supplementary Material we present a second simulation study with a continuous response, but where X and Z are both continuous. Comparisons of CML, SW and EL are similar to those in Table 2, but with SW and EL showing slightly less improvement over CML.

4. ILLUSTRATION

We consider an application to genetic association testing based on data from Genetic Analysis Workshop 17 (Almasy et al., 2011). These data use real genotype sequences from 697 individuals, obtained from the 1000 Genomes Project. The workshop organizers simulated data on three continuous traits (outcomes) by using models based on the genotype information on each individual along with their sex, age and smoking status. There has been much interest in the last decade on testing for associations between rare genetic variants and specific traits or outcomes (e.g., Barnett et al., 2013; Derkach, Lawless, & Sun, 2014). We consider the problem of testing for an association between the quantitative trait called Q1 and rare variants in gene FLT1 on chromosome 13 (Yilmaz & Bull, 2011). For the purposes of illustration, we consider the 321 individuals of Chinese or Japanese origin; this reduces population heterogeneity and avoids stratification issues. There are 18 single nucleotide polymorphisms (SNPs) with minor allele frequency less than 0.05 in this sub-population, and from these we derive a rare variant score for each individual as the total number of minor alleles across the 18 SNPs. These scores ranged from 0 to 3 across the 321 individuals.

As in Yilmaz & Bull (2011) we mimic a study in which the cost of genotyping all individuals is prohibitive and suppose that a two-phase design is used, in which the trait Q1 (Y in our notation) and covariates Z represented by age (in years, and standardized), sex (male =1, female = 0) and smoking status (yes = 1, no = 0) are known for all 321 individuals. A phase 2 sample is then selected and from it the rare variant score X is obtained.

We stratify the phase 1 sample by the quartiles of Q1 (Y). This gives three strata $S_1 = (-\infty, -0.6213)$, $S_2 = (-0.6213, 0.7826)$ and $S_3 = (0.7826, \infty)$, where -0.6213 and 0.7826 are the first and third quartiles of Q1 respectively. We consider a phase 2 VPS design which samples only from the first and third strata, with $P(R = 1|Y \in S_1) = P(R = 1|Y \in S_3) = 0.625$, giving approximately 100 individuals for phase 2. The true model for phase 2 selection is therefore $P(R = 1|y) = \pi_{est}(y; \boldsymbol{\alpha}) = \alpha_1 \mathrm{I}(y < -0.6213) + \alpha_3 \mathrm{I}(y \geq 0.7826)$, with $\alpha_1 = \alpha_3 = 0.625$. We also consider a more highly stratified model $P(R = 1|y, z) = \pi_{sat}(y, z; \tilde{\boldsymbol{\alpha}})$ where the π model is stratified also by smoking status and age. More specifically, we consider four strata determined by the sets $T_1 = \mathrm{I}(\text{smoke} = 1, \text{age} \geq -0.087)$, $T_2 = \mathrm{I}(\text{smoke} = 0, \text{age} \geq -0.087)$, $T_3 = \mathrm{I}(\text{smoke} = 1, \text{age} < -0.087)$, $T_4 = \mathrm{I}(\text{smoke} = 0, \text{age} < -0.087)$, where -0.087 is the median of (adjusted) age. Thus $\tilde{\boldsymbol{\alpha}} = (\alpha_{11}, \alpha_{12}, \alpha_{13}, \alpha_{14}, \alpha_{31}, \alpha_{32}, \alpha_{33}, \alpha_{34})$ with $\alpha_{ij} = P(R = 1|(Y, Z) \in S_i \times T_j)$. The estimates of regression coefficients in the normal regression model $N(\beta_c + \beta_x x + \boldsymbol{\beta}_z' z)$ for Y are shown in Table 3. For comparison we also show the estimates based on the full data for all 321 individuals, which is available. We see that the most efficient CML, SW and EL estimates based on X data

TABLE 3: Regression coefficient estimates for the GAW 17 two-phase data.

	Mean [Estimated SE]				
Method	Intercept	# of rare variants	Sex	Age	Smoke
Full data	−0.2489 [0.1701]	0.3164 [0.0857]	0.0344 [0.1011]	0.2805 [0.0505]	0.6878 [0.1176]
CML0	−0.2722 [0.2376]	0.2126 [0.0912]	0.1070 [0.1401]	0.3269 [0.0740]	0.7272 [0.1912]
CML-est	−0.2837 [0.2348]	0.2126 [0.0912]	0.1070 [0.1401]	0.3275 [0.0735]	0.7291 [0.1912]
CML-sat	−0.2703 [0.2310]	0.2128 [0.0912]	0.1086 [0.1374]	0.3239 [0.0700]	0.6655 [0.1667]
SW-est	−0.2889 [0.2348]	0.2126 [0.0912]	0.1070 [0.1401]	0.3278 [0.0735]	0.7300 [0.1913]
SW-sat	−0.2635 [0.2260]	0.2143 [0.0910]	0.1091 [0.1349]	0.3259 [0.0693]	0.6683 [0.1609]

for only about 30% of individuals have SEs that are only moderately larger than those for the estimates from the full data. As in the numerical studies, the highly stratified selection models give more precise estimates, but the reductions in SEs are not large in this setting.

5. CONCLUDING REMARKS

Our results show that for response-dependent two-phase studies the Scott & Wild (2011) estimator of $\boldsymbol{\beta}$ is optimal in the class of estimators based on the estimating functions $\boldsymbol{S}_1(\boldsymbol{\beta}, \boldsymbol{\alpha})$, $\boldsymbol{S}_2(\boldsymbol{\beta}, \boldsymbol{\alpha})$ and $\boldsymbol{S}_\pi(\boldsymbol{\alpha})$ for general regression models. The SW estimator, which is computationally straightforward to obtain, has the same asymptotic efficiency as more computationally demanding EL estimators. We note that EL estimates of α appear slightly more efficient than SW estimates, but this is of no practical importance for two-phase studies since $\boldsymbol{\alpha}_0$ is known by design.

As expected, the finite sample performances of SW and EL in simulation studies were very similar. We have demonstrated that using more highly stratified models for $\pi(y, z)$ in estimating functions can greatly increase efficiency. We also found that CML is very efficient when a sufficiently highly stratified model for π is used, especially with a binary response variable. SML gave significant, though not huge, gains in efficiency over SW and EL in Study 3, where the response is continuous. When feasible computationally, SML is thus advantageous. Moreover, it can handle situations where phase 2 sampling is based on residuals from a fitted phase 1 model for Y given Z (e.g., Derkach, Lawless & Sun, 2015; Tao et al., 2017); CML and the corresponding SW and EL estimators cannot do this. However, in spite of recent advances in SML that use kernel- or sieve-based methods for estimation of covariate distributions (Zeng & Lin, 2014; Tao et al., 2017), CML, SW and EL estimations remain attractive, especially with discrete responses and more generally, when covariate distributions are complicated.

Finally, we note that recent advances have also been made in the optimal design of two-phase studies when SML is used (Tao, Zeng & Lin, 2020). The designs are presumably highly efficient when CML, SW or EL estimation are used, but it would be useful to explore this.

ACKNOWLEDGEMENTS

We thank the Genetic Analysis Workshop (GAW) for permission to use the data in Section 4, and Yildiz Yilmaz for helpful comments on the data. The GAW is supported by National Institutes of Health (NIH) grant R01 GM031575 and preparation of the simulated exome data set for GAW 17 was supported in part by NIH grant R01 MH059490. The research of Dr. Lawless was supported by the Natural Sciences and Engineering Research Council of Canada under Discovery Grant RGPIN 8597.

BIBLIOGRAPHY

Almasy, L. A., Dyer, T. D., Peralta, J. M., Kent, J. W., Charlesworth, J. C., Curran, J. E., & Blangero, J. (2011). Genetic analysis workshop 17 mini-exome simulation. *BMC Proceedings*, S2.

Barnett, I. J., Lee, S., & Lin, X. (2013). Detecting rare variant effects using extreme phenotype sampling in sequencing association studies. *Genetic Epidemiology*, 37, 142–151.

Borgan, O. & Samuelsen, S. O. (2014). Nested case–control and case–cohort studies. In Klein, J. P., Van Houwelingen, H. C., Ibrahim, J. G., & Scheike, T. H. (Eds.) *Handbook of Survival Analysis*, CRC Press, Boca Raton, 343–367.

Breslow, N. E. & Cain, K. C. (1988). Logistic regression for two-stage case–control data. *Biometrika*, 75, 11–20.

Breslow, N. E., Lumley, T., Ballantyne, C. M., Chambless, L. E., & Kulich, M. (2009). Using the whole cohort in the analysis of case–cohort data. *American Journal of Epidemiology*, 169, 1398–1405.

Derkach, A., Lawless, J. F., & Sun, L. (2014). Pooled association tests for rare variants: A review and some new results. *Statistical Science*, 29, 302–321.

Derkach, A., Lawless, J. F., & Sun, L. (2015). Score tests for association under response-dependent sampling designs for expensive covariates. *Biometrika*, 102, 988–994.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica*, 50, 1029–1054.

Huang, B. & Lin, D. (2007). Efficient association mapping of quantitative trait loci with selective genotyping. *American Journal of Human Genetics*, 80, 567–576.

Keogh, R. H. & Cox, D. R. (2014). *Case–Control Studies*. Cambridge University Press, Cambridge, UK.

Lawless, J. F., Kalbfleisch, J. D., & Wild, C. J. (1999). Semiparametric methods for response-selective and missing data problems in regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61, 413–438.

Lawless, J. F. (1997). Likelihood and pseudo likelihood estimation based on response-biased observation. In Basawa, I. V., Godambe, V. P., & Taylor, R. L. (Eds.) *Selected Proceedings of the Symposium on Estimating Functions*, IMS, Hayward CA, 43–55.

Li, D., Lewinger, J. P., Gauderman, W. J., Murcray, C. E., & Conti, D. (2011). Using extreme phenotype sampling to identify the rare causal variants of quantitative traits in association studies. *Genetic Epidemiology*, 35, 790–799.

Lin, D. -Y., Zeng, D., & Tang, Z. Z. (2013). Quantitative trait analysis in sequencing studies under trait-dependent sampling. *Proceedings of the National Academy of Sciences of the United States of America*, 110, 12247–12252.

Newey, W. K. & Smith, R. J. (2004). Higher order properties of GMM and generalized empirical likelihood estimators. *Econometrica*, 72, 219–255.

Owen, A. B. (2001). *Empirical Likelihood*. Chapman and Hall/CRC, Boca Raton.

Qin, J. (2017). *Biased Sampling, Over-identified Parameter Problems and Beyond*. Springer Nature, Singapore.

Qin, J. & Lawless, J. (1994). Empirical likelihood and general estimating equations. *The Annals of Statistics*, 22, 300–325.

Qin, J., Zhang, B., & Leung, D. H. (2009). Empirical likelihood in missing data problems. *Journal of the American Statistical Association*, 104, 1492–1503.

Robins, J. R., Rotnitzky, A., & Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89, 846–866.

Scott, A. J. & Wild, C. J. (1986). Fitting logistic models under case–control or choice based sampling. *Journal of the Royal Statistical Society: Series B (Methodological)*, 48, 170–182.

Scott, A. J. & Wild, C. J. (1997). Fitting regression models to case–control data by maximum likelihood. *Biometrika*, 84, 57–71.

Scott, A. J. & Wild, C. J. (2001). Maximum likelihood for generalised case–control studies. *Journal of Statistical Planning and Inference*, 96, 3–27.

Scott, A. J. & Wild, C. J. (2011). Fitting regression models with response-biased samples. *Canadian Journal of Statistics*, 39, 519–536.

Shen, Y., Cai, T., Chen, Y., Yang, Y., & Chen, J. (2015). Retrospective likelihood-based methods for analyzing case–cohort genetic association studies. *Biometrics*, 71, 960–968.

Song, R., Zhou, H., & Kosorok, M. R. (2009). A note on semiparametric efficient inference for two-stage outcome-dependent sampling with a continuous outcome. *Biometrika*, 96, 221–228.

Tao, R., Zeng, D., & Lin, D. Y. (2017). Efficient semiparametric inference under two-phase sampling, with applications to genetic association studies. *Journal of the American Statistical Association*, 112, 1468–1476.

Tao, R., Zeng, D., & Lin, D. Y. (2020). Optimal design of two-phase studies. *Journal of the American Statistical Association*, 115, 1946–1959.

Xie, Y. & Zhang, B. (2017). Empirical likelihood in nonignorable covariate-missing data problems. *The International Journal of Biostatistics*, 13, 10.1515/ijb-2016-0053.

Yilmaz, Y. & Bull, S. B. (2011). Are quantitative trait-dependent sampling designs cost effective for analysis of rare and common variants?. *BMC Proceedings*, 5, S111.

Zeng, D. & Lin, D. -Y. (2014). Efficient estimation of semiparametric transformation models for two-phase cohort studies. *Journal of the American Statistical Association*, 109, 371–383.

Zhang, Z. & Rockette, H. E. (2005). On maximum likelihood estimation in parametric regression with missing covariates. *Journal of Statistical Planning and Inference*, 134, 206–223.

Zhang, Z. & Rockette, H. E. (2007). An EM algorithm for regression analysis with incomplete covariate information. *Journal of Statistical Computation and Simulation*, 77, 163–173.

Zhao, Y., Lawless, J. F., & McLeish, D. L. (2009). Likelihood methods for parametric regression models with missing covariates or outcomes. *Biometrical Journal*, 51, 1–14.

Zhou, H., Song, R., Wu, Y., & Qin, J. (2011). Statistical inference for a two-stage outcome-dependent sampling design with a continuous outcome. *Biometrics*, 67, 194–202.

APPENDIX

A1. Derivations for relationships between the matrices $\mathcal{C}$ and $\mathcal{J}$ in Section 2.3

We will show that

$$\mathcal{C} = E\mathbf{U}(\boldsymbol{\phi}_0)\mathbf{U}(\boldsymbol{\phi}_0)^T = \begin{bmatrix} \mathcal{C}_{11} & \mathcal{C}_{12} & \mathcal{C}_{1\pi} \\ \mathcal{C}_{12}^T & \mathcal{C}_{22} & \mathcal{C}_{2\pi} \\ \mathcal{C}_{1\pi}^T & \mathcal{C}_{2\pi}^T & \mathcal{C}_{\pi\pi} \end{bmatrix} = \begin{bmatrix} \mathcal{C}_{11} & \mathcal{C}_{12} & \mathcal{C}_{12} \\ \mathcal{C}_{12}^T & \mathcal{C}_{22} & \mathcal{C}_{22} \\ \mathcal{C}_{12}^T & \mathcal{C}_{22} & \mathcal{C}_{\pi\pi} \end{bmatrix} \tag{A1}$$

and

$$\begin{aligned} \mathcal{J} = -E\left\{\frac{\partial \mathbf{U}(\boldsymbol{\phi}_0)}{\partial \boldsymbol{\phi}^T}\right\} &= \begin{bmatrix} -E\left(R\frac{\partial^2 \log f_c}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}\right) & -E\left(R\frac{\partial^2 \log f_c}{\partial\boldsymbol{\beta}\partial\boldsymbol{\alpha}^T}\right) \\ -E\left(R\frac{\partial^2 \log f_c}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\beta}^T}\right) & -E\left(R\frac{\partial^2 \log f_c}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}^T}\right) \\ \mathbf{0} & -E\left\{\frac{\partial}{\partial\boldsymbol{\alpha}^T}\left(\frac{R}{\pi} - \frac{1-R}{1-\pi}\right)\frac{\partial\pi}{\partial\boldsymbol{\alpha}}\right\} \end{bmatrix} \\ &= \begin{bmatrix} \mathcal{I}_{11} & \mathcal{I}_{12} \\ \mathcal{I}_{12}^T & \mathcal{I}_{22} \\ \mathbf{0} & \mathcal{I}_{\pi\pi} \end{bmatrix} = \begin{bmatrix} \mathcal{C}_{11} & \mathcal{C}_{12} \\ \mathcal{C}_{12}^T & \mathcal{C}_{22} \\ \mathbf{0} & \mathcal{C}_{\pi\pi} \end{bmatrix}. \end{aligned} \tag{A2}$$

For a function h of y, x, z and r where r is binary, we use $\int h d\mu(y, r)$ to denote the integral with respect to y and summed over $r = 1$ and $r = 0$. Then we can compute

$$\begin{aligned} \mathcal{I}_{11} &= -E\int \frac{\partial \boldsymbol{S}_1}{\partial\boldsymbol{\beta}^T} f(y|x, z; \boldsymbol{\beta}) d\mu(y, r) \\ &= -E\int r\frac{\partial^2 \log f_c(y|x, z; \boldsymbol{\beta}, \boldsymbol{\alpha})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} f(y|x, z; \boldsymbol{\beta})\pi(y, z; \boldsymbol{\alpha}) d\mu(y, r) \\ &= -E\int r\frac{\partial^2 \log f_c(y|x, z; \boldsymbol{\beta}, \boldsymbol{\alpha})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} f_c(y|x, z; \boldsymbol{\beta}, \boldsymbol{\alpha})\pi^*(x, z; \boldsymbol{\beta}, \boldsymbol{\alpha}) d\mu(y, r) \\ &= -E\int r\frac{\partial^2 \log f_c(y|x, z; \boldsymbol{\beta}, \boldsymbol{\alpha})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} f_c(y|x, z; \boldsymbol{\beta}, \boldsymbol{\alpha}) d\mu(y, r)\pi^*(x, z; \boldsymbol{\beta}, \boldsymbol{\alpha}) \\ &:= E\{\mathcal{I}_{c\beta\beta}(x, z)\pi^*(x, z; \boldsymbol{\beta}, \boldsymbol{\alpha})\}, \end{aligned} \tag{A3}$$

where the expectation is taken with respect to the marginal distribution $g(x, z)$; $\pi^*(x, z; \boldsymbol{\beta}, \boldsymbol{\alpha}) = f(y|x, z; \boldsymbol{\beta})\pi(y, z; \boldsymbol{\alpha})/f_c(y|x, z; \boldsymbol{\beta}, \boldsymbol{\alpha}) = \int f(y|x, z; \boldsymbol{\beta})\pi(y, z; \boldsymbol{\alpha})dy$, and

$$\begin{aligned}
\mathcal{I}_{c\beta\beta}(x, z) &= -\int r\frac{\partial^2 \log f_c(y|x, z; \boldsymbol{\beta}, \boldsymbol{\alpha})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} f_c(y|x, z; \boldsymbol{\beta}, \boldsymbol{\alpha})d\mu(y, r) \\
&= -\int r\left\{-\frac{1}{f_c^2}\left(\frac{\partial f_c}{\partial\boldsymbol{\beta}}\right)^{\otimes 2} + \frac{1}{f_c}\frac{\partial^2 f_c}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}\right\} f_c(y|x, z; \boldsymbol{\beta}, \boldsymbol{\alpha})d\mu(y, r) \\
&= \int r\left(\frac{\partial \log f_c}{\partial\boldsymbol{\beta}}\right)^{\otimes 2} f_c(y|x, z; \boldsymbol{\beta}, \boldsymbol{\alpha})d\mu(y, r) \\
&\quad + \frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} E\int r f_c(y|x, z; \boldsymbol{\beta}, \boldsymbol{\alpha})d\mu(y, r) \\
&= \int r\left(\frac{\partial \log f_c}{\partial\boldsymbol{\beta}}\right)^{\otimes 2} f_c(y|x, z; \boldsymbol{\beta}, \boldsymbol{\alpha})d\mu(y, r).
\end{aligned}$$

Also,

$$\begin{aligned}
\mathcal{C}_{11} &= E\int r\left(\frac{\partial \log f_c}{\partial\boldsymbol{\beta}}\right)^{\otimes 2} f(y|x, z; \boldsymbol{\beta})\pi(y, z; \boldsymbol{\alpha})d\mu(y, r) \\
&= E\int r\left(\frac{\partial \log f_c}{\partial\boldsymbol{\beta}}\right)^{\otimes 2} f_c(y|x, z; \boldsymbol{\beta}, \boldsymbol{\alpha})d\mu(y, r)\pi^*(x, z; \boldsymbol{\beta}, \boldsymbol{\alpha}) \\
&= E\{\mathcal{I}_{c\beta\beta}(x, z)\pi^*(x, z; \boldsymbol{\beta}, \boldsymbol{\alpha})\}. \qquad \text{(A4)}
\end{aligned}$$

So by (A3) and (A4), we have $\mathcal{I}_{11} = \mathcal{C}_{11}$.

Similarly, we compute

$$\begin{aligned}
\mathcal{I}_{1\pi} = \mathcal{I}_{12} &= -E\int \frac{\partial \boldsymbol{S}_1}{\partial\boldsymbol{\alpha}^T} f(y|x, z; \boldsymbol{\beta})\pi(y, z; \boldsymbol{\alpha})d\mu(y, r) \\
&= -E\int r\frac{\partial^2 \log f_c(y|x, z; \boldsymbol{\beta}, \boldsymbol{\alpha})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\alpha}^T} f_c(y|x, z; \boldsymbol{\beta}, \boldsymbol{\alpha})d\mu(y, r)\pi^*(x, z; \boldsymbol{\beta}, \boldsymbol{\alpha}) \\
&:= E\{\mathcal{I}_{c\beta\alpha}(x, z)\pi^*(x, z; \boldsymbol{\beta}, \boldsymbol{\alpha})\}, \qquad \text{(A5)}
\end{aligned}$$

where

$$\begin{aligned}
\mathcal{I}_{c\beta\alpha}(x, z) &= -\int r\frac{\partial^2 \log f_c(y|x, z; \boldsymbol{\beta}, \boldsymbol{\alpha})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\alpha}^T} f_c(y|x, z; \boldsymbol{\beta}, \boldsymbol{\alpha})d\mu(y, r) \\
&= -\int r\left\{-\frac{1}{f_c^2}\frac{\partial f_c}{\partial\boldsymbol{\beta}}\frac{\partial f_c}{\partial\boldsymbol{\alpha}^T} + \frac{1}{f_c}\frac{\partial^2 f_c}{\partial\boldsymbol{\beta}\partial\boldsymbol{\alpha}^T}\right\} f_c(y|x, z; \boldsymbol{\beta}, \boldsymbol{\alpha})d\mu(y, r) \\
&= \int r\frac{\partial \log f_c}{\partial\boldsymbol{\beta}}\frac{\partial \log f_c}{\partial\boldsymbol{\alpha}^T} f_c(y|x, z; \boldsymbol{\beta}, \boldsymbol{\alpha})d\mu(y, r) \\
&\quad + \frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\alpha}^T}\int r f_c(y|x, z; \boldsymbol{\beta}, \boldsymbol{\alpha})d\mu(y, r)
\end{aligned}$$

$$= \int r \frac{\partial \log f_c}{\partial \boldsymbol{\beta}} \frac{\partial \log f_c}{\partial \boldsymbol{\alpha}^T} f_c(y|x,z;\boldsymbol{\beta},\boldsymbol{\alpha}) d\mu(y,r).$$

$$\begin{aligned} C_{12} &= E \int r \frac{\partial \log f_c}{\partial \boldsymbol{\beta}} \frac{\partial \log f_c}{\partial \boldsymbol{\alpha}^T} f_c(y|x,z;\boldsymbol{\beta},\boldsymbol{\alpha}) d\mu(y,r) \pi^*(x,z;\boldsymbol{\beta},\boldsymbol{\alpha}) \\ &= E\{\mathcal{I}_{c\beta\alpha}(x,z)\pi^*(x,z;\boldsymbol{\beta},\boldsymbol{\alpha})\}. \end{aligned} \tag{A6}$$

At the same time, if we differentiate both sides of $E(\boldsymbol{S}_1|x,z) = \mathbf{0}$ with respect to $\boldsymbol{\alpha}$, we have

$$\begin{aligned} \mathbf{0} &= \frac{\partial}{\partial \boldsymbol{\alpha}^T} \int r \frac{\partial \log f_c(y|x,z,\boldsymbol{\beta},\boldsymbol{\alpha})}{\partial \boldsymbol{\beta}} f(y|x,z;\boldsymbol{\beta}) \pi(y,z;\boldsymbol{\alpha}) d\mu(y,r) \\ &= \int r \frac{\partial^2 \log f_c(y|x,z,\boldsymbol{\beta},\boldsymbol{\alpha})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\alpha}^T} f(y|x,z;\boldsymbol{\beta}) \pi(y,z;\boldsymbol{\alpha}) d\mu(y,r) \\ &\quad + \int r \frac{\partial \log f_c(y|x,z,\boldsymbol{\beta},\boldsymbol{\alpha})}{\partial \boldsymbol{\beta}} f(y|x,z;\boldsymbol{\beta}) \frac{\partial \pi(y,z;\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}^T} d\mu(y,r). \end{aligned}$$

Taking the expectation with respect to the marginal distribution $g(x,z)$, we get

$$\mathbf{0} = -\mathcal{I}_{1\pi} + E \int r \frac{\partial \log f_c(y|x,z,\boldsymbol{\beta},\boldsymbol{\alpha})}{\partial \boldsymbol{\beta}} \frac{\partial \pi(y,z;\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}^T} f(y|x,z;\boldsymbol{\beta}) d\mu(y,r). \tag{A7}$$

Note that the first term above is equal to $-\mathcal{I}_{12}$. Since $r^2 = r$ and $r(1-r) = 0$, we can rewrite the second term as

$$\begin{aligned} & E \int r \frac{\partial \log f_c(y|x,z,\boldsymbol{\beta},\boldsymbol{\alpha})}{\partial \boldsymbol{\beta}} \left\{ \frac{r}{\pi(y,z;\boldsymbol{\alpha})} - \frac{1-r}{1-\pi(y,z;\boldsymbol{\alpha})} \right\} \frac{\partial \pi(y,z;\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}^T} \\ & \cdot f(y|x,z;\boldsymbol{\beta}) \pi(y,z,\boldsymbol{\alpha}) d\mu(y,r) \\ &= E \int \boldsymbol{S}_1 \boldsymbol{S}_\pi^T f(y|x,z;\boldsymbol{\beta}) \pi(y,z,\boldsymbol{\alpha}) d\mu(y,r) \\ &= E(\boldsymbol{S}_1 \boldsymbol{S}_\pi^T) = \mathcal{C}_{1\pi}. \end{aligned} \tag{A8}$$

Thus by (A5), (A6), (A7), (A8), we get

$$\mathcal{I}_{1\pi} = \mathcal{I}_{12} = \mathcal{C}_{1\pi} = \mathcal{C}_{12}. \tag{A9}$$

Similarly,

$$\begin{aligned} \mathcal{I}_{22} &= -E \int \frac{\partial \boldsymbol{S}_2}{\partial \boldsymbol{\alpha}^T} f(y|x,z;\boldsymbol{\beta}) \pi(y,z;\boldsymbol{\alpha}) d\mu(y,r) \\ &= -E \int r \frac{\partial^2 \log f_c(y|x,z;\boldsymbol{\beta},\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\alpha}^T} f_c(y|x,z;\boldsymbol{\beta},\boldsymbol{\alpha}) d\mu(y,r) \pi^*(x,z;\boldsymbol{\beta},\boldsymbol{\alpha}) \\ &:= E\{\mathcal{I}_{c\alpha\alpha}(x,z)\pi^*(x,z;\boldsymbol{\beta},\boldsymbol{\alpha})\} \end{aligned} \tag{A10}$$

and

$$\begin{aligned} \mathcal{I}_{c\alpha\alpha}(x,z) &= -\int r \frac{\partial^2 \log f_c(y|x,z;\boldsymbol{\beta},\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\alpha}^T} f_c(y|x,z;\boldsymbol{\beta},\boldsymbol{\alpha}) d\mu(y,r) \\ &= -\int r \left\{ -\frac{1}{f_c^2} \left(\frac{\partial f_c}{\partial \boldsymbol{\alpha}} \right)^{\otimes 2} + \frac{1}{f_c} \frac{\partial^2 f_c}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\alpha}^T} \right\} f_c(y|x,z;\boldsymbol{\beta},\boldsymbol{\alpha}) d\mu(y,r) \end{aligned}$$

$$= \int r \left(\frac{\partial \log f_c}{\partial \boldsymbol{\alpha}} \right)^{\otimes 2} f_c(y|x,z;\boldsymbol{\beta},\boldsymbol{\alpha}) d\mu(y,r) + \frac{\partial^2}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\alpha}^T} \int r f_c(y|x,z;\boldsymbol{\beta},\boldsymbol{\alpha}) d\mu(y,r)$$

$$= E \int r \left(\frac{\partial \log f_c}{\partial \boldsymbol{\alpha}} \right)^{\otimes 2} f_c(y|x,z;\boldsymbol{\beta},\boldsymbol{\alpha}) d\mu(y,r).$$

We also have

$$\begin{aligned} \mathcal{C}_{22} &= E \int r \left(\frac{\partial \log f_c}{\partial \boldsymbol{\alpha}} \right)^{\otimes 2} f(y|x,z;\boldsymbol{\beta}) \pi(y,z;\boldsymbol{\alpha}) d\mu(y,r) \\ &= E \int r \left(\frac{\partial \log f_c}{\partial \boldsymbol{\alpha}} \right)^{\otimes 2} f_c(y|x,z;\boldsymbol{\beta},\boldsymbol{\alpha}) d\mu(y,r) \pi^*(x,z;\boldsymbol{\beta},\boldsymbol{\alpha}) \\ &= E\{\mathcal{I}_{c\alpha\alpha}(x,z)\pi^*(x,z;\boldsymbol{\beta},\boldsymbol{\alpha})\}. \end{aligned} \tag{A11}$$

Differentiate both sides of $E(\boldsymbol{S}_2|x,z) = \mathbf{0}$ with respect to $\boldsymbol{\alpha}$, and we have

$$\begin{aligned} \mathbf{0} &= \frac{\partial}{\partial \boldsymbol{\alpha}^T} \int r \frac{\partial \log f_c(y|x,z,\boldsymbol{\beta},\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} f(y|x,z;\boldsymbol{\beta}) \pi(y,z;\boldsymbol{\alpha}) d\mu(y,r) \\ &= \int r \frac{\partial^2 \log f_c(y|x,z,\boldsymbol{\beta},\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\alpha}^T} f(y|x,z;\boldsymbol{\beta}) \pi(y,z;\boldsymbol{\alpha}) d\mu(y,r) \\ &\quad + \int r \frac{\partial \log f_c(y|x,z,\boldsymbol{\beta},\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} f(y|x,z;\boldsymbol{\beta}) \frac{\partial \pi(y,z;\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}^T} d\mu(y,r) \end{aligned}$$

and by taking expectation with respect to $g(x,z)$,

$$\begin{aligned} \mathbf{0} &= -\mathcal{I}_{22} + E \int r \frac{\partial \log f_c(y|x,z,\boldsymbol{\beta},\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} \frac{\partial \pi(y,z;\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}^T} f(y|x,z;\boldsymbol{\beta}) d\mu(y,r) \\ &= -\mathcal{I}_{22} + E(\boldsymbol{S}_2 \boldsymbol{S}_\pi^T) = -\mathcal{I}_{21} + \mathcal{C}_{2\pi}. \end{aligned} \tag{A12}$$

Thus by (A10), (A11), (A12), we see that

$$\mathcal{I}_{22} = \mathcal{C}_{2\pi} = \mathcal{C}_{22}. \tag{A13}$$

Lastly, by maximum likelihood for logistic regression we have $-E\left\{\frac{\partial \boldsymbol{S}_\pi(\boldsymbol{\phi}_0)}{\partial \boldsymbol{\alpha}^T}\right\} = E(\boldsymbol{S}_\pi \boldsymbol{S}_\pi^T)$, so

$$\mathcal{I}_{\pi\pi} = \mathcal{C}_{\pi\pi}. \tag{A14}$$

A2. Computation of $\mathrm{ACov}(\widehat{\boldsymbol{\beta}}_{\mathrm{EL}})$

We denote

$$\mathcal{C}^{-1} = D = \begin{bmatrix} D^{11} & D^{12} & D^{13} \\ D^{21} & D^{22} & D^{23} \\ D^{31} & D^{32} & D^{33} \end{bmatrix}, \tag{A15}$$

and then by $\mathcal{C}D = I$, we have $\mathcal{C}_{11}D^{11} + \mathcal{C}_{12}D^{21} + \mathcal{C}_{12}D^{31} = I$, $\mathcal{C}_{11}D^{12} + \mathcal{C}_{12}D^{22} + \mathcal{C}_{12}D^{32} = \mathbf{0}$, $\mathcal{C}_{11}D^{13} + \mathcal{C}_{12}D^{23} + \mathcal{C}_{12}D^{33} = \mathbf{0}$, and therefore

$$\begin{aligned} V^{11} &= \begin{bmatrix} \mathcal{C}_{11} & \mathcal{C}_{12} & \mathbf{0} \end{bmatrix} \begin{bmatrix} D^{11} & D^{12} & D^{13} \\ D^{21} & D^{22} & D^{23} \\ D^{31} & D^{32} & D^{33} \end{bmatrix} \begin{bmatrix} \mathcal{C}_{11} \\ \mathcal{C}_{12}^T \\ \mathbf{0} \end{bmatrix} \\ &= \begin{bmatrix} I - \mathcal{C}_{12}D^{31} & -\mathcal{C}_{12}D^{32}, & -\mathcal{C}_{12}D^{33} \end{bmatrix} \begin{bmatrix} \mathcal{C}_{11} \\ \mathcal{C}_{12}^T \\ \mathbf{0} \end{bmatrix} \\ &= \mathcal{C}_{11} - \mathcal{C}_{12}(D^{31}\mathcal{C}_{11} + D^{32}\mathcal{C}_{12}^T) \\ &= \mathcal{C}_{11} + \mathcal{C}_{12}D^{33}\mathcal{C}_{12}^T. \end{aligned}$$

We can further compute D^{33} as:

$$D^{33} = \left(\mathcal{C}_{\pi\pi} - \begin{bmatrix} \mathcal{C}_{12}^T & \mathcal{C}_{22} \end{bmatrix} \begin{bmatrix} \mathcal{C}_{11} & \mathcal{C}_{12} \\ \mathcal{C}_{12}^T & \mathcal{C}_{22} \end{bmatrix}^{-1} \begin{bmatrix} \mathcal{C}_{12} \\ \mathcal{C}_{22} \end{bmatrix} \right)^{-1}.$$

Note that $\begin{bmatrix} \mathcal{C}_{11} & \mathcal{C}_{12} \\ \mathcal{C}_{12}^T & \mathcal{C}_{22} \end{bmatrix}^{-1} \begin{bmatrix} \mathcal{C}_{11} & \mathcal{C}_{12} \\ \mathcal{C}_{12}^T & \mathcal{C}_{22} \end{bmatrix} = I$, so

$$\begin{bmatrix} \mathcal{C}_{12}^T & \mathcal{C}_{22} \end{bmatrix} \begin{bmatrix} \mathcal{C}_{11} & \mathcal{C}_{12} \\ \mathcal{C}_{12}^T & \mathcal{C}_{22} \end{bmatrix}^{-1} \begin{bmatrix} \mathcal{C}_{12} \\ \mathcal{C}_{22} \end{bmatrix} = \mathcal{C}_{22},$$

and therefore

$$D^{33} = (\mathcal{C}_{\pi\pi} - \mathcal{C}_{22})^{-1}. \tag{A16}$$

Thus, finally we obtain

$$V^{11} = \mathcal{C}_{11} + \mathcal{C}_{12}(\mathcal{C}_{\pi\pi} - \mathcal{C}_{22})^{-1}\mathcal{C}_{12}^T.$$

Similarly, V^{12} is the $(1,2)$ block of $\mathcal{I}^T\mathcal{C}^{-1}\mathcal{I}$, so

$$\begin{aligned} V^{12} &= \begin{bmatrix} \mathcal{C}_{11} & \mathcal{C}_{12} & \mathbf{0} \end{bmatrix} \begin{bmatrix} D^{11} & D^{12} & D^{13} \\ D^{21} & D^{22} & D^{23} \\ D^{31} & D^{32} & D^{33} \end{bmatrix} \begin{bmatrix} \mathcal{C}_{12} \\ \mathcal{C}_{22} \\ \mathcal{C}_{\pi\pi} \end{bmatrix} \\ &= \begin{bmatrix} \mathcal{C}_{11} & \mathcal{C}_{12} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ I \end{bmatrix} = \mathbf{0}. \end{aligned}$$

Received 09 January 2020
Accepted 16 April 2020