Statistica Sinica Preprint No: SS-2017-0308.R1	
Title	Empirical Likelihood Estimation Using Auxiliary Summary Information with Different Covariate Distributions
Manuscript ID	SS-2017-0308.R1
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202017.0308
Complete List of Authors	Peisong Han and Jerald F. Lawless
Corresponding Author	Peisong Han
E-mail	peisonghan@uwaterloo.ca
Notice: Accepted version subject to English editing.	

Statistica Sinica

Empirical Likelihood Estimation Using Auxiliary Summary Information with Different Covariate Distributions

Peisong Han AND Jerald F. Lawless

Abstract: The potential use of auxiliary summary information to improve estimation efficiency for a current study has attracted significant interest. Most existing methods assume the data distribution is the same for the current study and for the population that generates the auxiliary information, but recent work has relaxed this assumption by allowing heterogeneity between the two covariate distributions. We consider an empirical likelihood approach which guarantees that use of the auxiliary information increases efficiency when the variability associated with this information is sufficiently small. We also investigate the effects of this variability on efficiency. Implementation through a nested optimization procedure and a Newton-Raphson-type algorithm is described. Simulation results that demonstrate efficiency gains and confirm large sample approximations are provided.

Key words and phrases: Auxiliary information; Combining data; Empirical likelihood; Estimation efficiency; Information uncertainty; Summary information.

1. Introduction

In many settings a main statistical objective is to fit models for a response variable conditional on certain covariates. With the increasing availability of large databases there is much interest in the possibility of enhancing modeling, prediction and inference for a current study by using auxiliary information. Such methodology has been used for many years in survey sampling, where summary population-level data, say from a census, are used to provide calibration factors so as to increase the efficiency of estimation based on survey data (e.g., Deville and Särndal 1992; Chen and Qin 1993; Chaudhuri et al. 2008; Chen and Kim 2014). Similar problems have also been considered in economics; for example, Imbens and Lancaster (1994) considered a longitudinal employment study involving covariates, with auxiliary data provided by longitudinal unemployment rates. More recently, studies in medicine and public health have received attention (e.g., Qin et al. 2015; Chatterjee et al. 2016; Huang et al. 2016). The auxiliary data typically have much less detailed covariate information than the study data, and are often in summary or aggregate form. For example,

Huang et al. (2016) consider detailed models for the time to some event such as the recurrence of cancer in a group of treated patients, along with auxiliary data summarizing recurrence rates by a specific time, available from a population cancer registry. Another example is the use of large cohorts or populations as a basis for two-phase studies (e.g., Breslow et al. 2009; Lumley et al. 2011), where a subset of the cohort is selected for the measurement of detailed information on certain covariates. See also Kim and Rao (2012) in the survey sampling context.

Several methods for using auxiliary data have been proposed, including weight calibration (e.g., Lumley et al. 2011), generalized regression (e.g., Chen and Chen 2000; Lawless and Kalbfleisch 2011), constrained maximum likelihood (Handcock et al. 2000; Chatterjee et al. 2016), generalized method of moments (e.g., Imbens and Lancaster 1994) and empirical likelihood (e.g., Chen and Qin 1993; Qin 2000; Chaudhuri et al. 2008; Chen and Kim 2014; Qin et al. 2015; Huang et al. 2016). Most of these methods make the assumptions that (a) the conditional distribution of the response variable given covariates of interest is the same in the populations that provide the study data and auxiliary data, and (b) the covariate distributions in the two populations are also the same. These assumptions are reasonable when the current study is based on a sample of individuals from the

population providing the auxiliary data. However, there is much recent interest in using big data bases that are external to the current study (e.g. Qin et al. 2015; Chatterjee et al. 2016; Huang et al. 2016). In many such contexts assumption (a) may be plausible but assumption (b) is more likely to be violated; see, for example, Keiding and Louis (2016) who note that conditional features or distributions are more likely to be "transportable" from one population to another than are marginal distributions. In this case methods developed assuming both (a) and (b) hold can lead to biased estimation and misleading conclusions. Our objective in this paper is to provide an empirical likelihood-based method for the case where assumption (b) does not hold.

To utilise auxiliary summary information when covariate distributions are different, we require the availability of a supplementary sample from the auxiliary data population, for which measurements on covariates of interest are collected. This supplementary sample can be of small size. It could be either independent of, or a subset of, the original units on which the auxiliary information is based. It is sometimes relatively easy to obtain such a supplementary sample, for example when the covariates of interest represent demographic characteristics in a large data base with accessible micro-data on individuals. However, there will often be significant incre-

mental costs to obtaining a supplementary sample and so it is important to weigh this against potential efficiency gains.

We study here an empirical likelihood-based method which treats the covariate distributions as nuisance parameters and leaves them completely unspecified. The only model we specify is for the conditional distribution of the response given covariates of interest. The approach was proposed by Han and Lawless (2016) in a discussion of Chatterjee et al. (2016), but not developed or studied there. When the variability associated with the auxiliary summary information is negligible compared to that in the study data, we show the proposed estimators are more efficient than the maximum likelihood estimator based on the study data alone. When the variability for the auxiliary summary information is non-negligible, we show explicitly how it affects the efficiency of the proposed estimators. We discuss how to implement the proposed method and provide numerical results on efficiency for binary logistic and normal linear regression. Some comments on assumptions and related issues are made in the final section, including the potential uses of large data sets.

2. Setup and review of some existing methods

The setting we consider is as follows. Let $(Y_i, X_i^{\mathrm{T}}, Z_i^{\mathrm{T}})^{\mathrm{T}}$, $i = 1, \ldots, n$, denote the random sample collected in the current study, where Y is the response and X and Z are vectors of covariates. Our interest is in $f(Y|X, Z)$, the distribution of Y given X and Z. We consider a family of models $f(Y \mid X, Z; \beta)$ that is parametrized by parameter β and assume that $f(Y|X, Z) = f(Y \mid X, Z; \beta_0)$ for some β_0. In addition to the study data, some auxiliary summary information is available in the form of an estimate $\hat{\theta}$ and its variance estimate, based on a known set of estimating functions $h(Y, X; \theta)$ applied to the auxiliary data set. The summary data reflects measurements on Y and X but not on Z. This is common; the current study is typically tailored to particular scientific questions and has numerous relevant covariates measured, whereas the auxiliary data summarize only a few features. Examples can be found in Imbens and Lancaster (1994), Chaudhuri et al. (2008), Qin et al. (2015) and Huang et al. (2016). It is assumed that the populations represented by the study data and by the auxiliary data share the same conditional distribution $f(Y|X, Z)$. The goal is to make inference about β_0 using both the study data and the auxiliary summary information, in the hope this improves efficiency over inference based solely on the study data.

Most existing methods assume that the two populations share the same covariate distribution $f(X, Z)$. To describe the situation, assume for now that there is no variability or uncertainty associated with the estimate $\hat{\theta}$. In other words, the auxiliary data summary consists of the vector θ^* that satisfies $E\{h(Y, X; \theta^*)\} = 0$, where the expectation $E(\cdot)$ is taken under the joint distribution $f(Y \mid X, Z)f(X, Z)$. Let $s(Y, X, Z; \beta) = \partial \log f(Y \mid X, Z; \beta)/\partial\beta$ be the score function of model $f(Y \mid X, Z; \beta)$. Estimation based on the current study data alone solves the estimating equation $\sum_{i=1}^{n} s(Y_i, X_i, Z_i; \beta) = 0$. The simplest way to utilise the auxiliary information is to treat $\{s(Y, X, Z; \beta)^{\mathrm{T}}, h(Y, X; \theta^*)^{\mathrm{T}}\}^{\mathrm{T}}$ as a set of estimating functions and apply the generalized method of moments (Hansen 1982) or empirical likelihood (Qin and Lawless 1994; Owen 2001) to the current study data. However, this approach does not yield a fully efficient estimator because it does not make full use of the fact that $f(Y \mid X, Z; \beta)$ is a likelihood function (Imbens and Lancaster 1994).

Since $E\{h(Y, X; \theta^*)\} = E[E\{h(Y, X; \theta^*) \mid X, Z\}]$, it follows that $E\{u(X, Z; \beta_0, \theta^*)\} = 0$ where

$$u(X, Z; \beta, \theta^*) = \int h(Y, X; \theta^*) f(Y \mid X, Z; \beta) dY.$$

This moment equality provides a constraint on $f(X, Z)$, and an application of semi-empirical likelihood (Qin 2000; Chatterjee et al. 2016) leads to an

estimator of β_0 defined through

$$\max_{\beta,p_1,\ldots,p_n} \prod_{i=1}^{n} f(Y_i \mid X_i, Z_i; \beta) p_i \qquad \text{subject to}$$
$$p_i \geq 0, \quad \sum_{i=1}^{n} p_i = 1, \quad \sum_{i=1}^{n} p_i u(X_i, Z_i; \beta, \theta^*) = 0,$$

where the p_i denote an empirical distribution for (X, Z) supported on the study data. This estimator has been shown to be more efficient than $\hat{\beta}_{\text{MLE}}$, the maximum likelihood estimator based on current study data alone (Qin 2000; Chatterjee et al. 2016). Asymptotically equivalent estimators can be derived by treating $\{s(Y, X, Z; \beta)^{\text{T}}, u(X, Z; \beta, \theta^*)^{\text{T}}\}^{\text{T}}$ as a set of estimating functions and then straightforwardly applying the generalized method of moments or empirical likelihood (Imbens and Lancaster 1994; Han and Lawless 2016).

As noted, the assumption of the same $f(X, Z)$ is often implausible (e.g., Keiding and Louis 2016) and when it is violated, the aforementioned estimators, except for $\hat{\beta}_{\text{MLE}}$, are biased. To relax this assumption, denote $f(X, Z)$ and $f^*(X, Z)$ as the covariate distributions for the study data and the auxiliary data populations, respectively. The auxiliary summary information θ^* then satisfies $E^*\{h(Y, X; \theta^*)\} = 0$, where the expectation $E^*(\cdot)$ is taken under the joint distribution $f(Y \mid X, Z) f^*(X, Z)$. Similar to before, it is seen that $E^*\{u(X, Z; \beta_0, \theta^*)\} = 0$. However, with the study data alone

the auxiliary estimating function $u(X, Z; \beta, \theta^*)$ cannot be used because $E^*(\cdot)$ is taken under the auxiliary data covariate distribution $f^*(X, Z)$. To use the auxiliary information, Chatterjee et al. (2016) assumed that a small random sample $(X_j^{*\mathrm{T}}, Z_j^{*\mathrm{T}})^{\mathrm{T}}$, $j = 1, \ldots, n^*$, is available from the auxiliary data population, referred to here as the supplementary sample. They proposed a constrained maximum likelihood estimator by maximizing $\prod_{i=1}^{n} f(Y_i \mid X_i, Z_i; \beta)$ under the constraint $n^{*-1} \sum_{j=1}^{n^*} u(X_j^*, Z_j^*; \beta, \theta^*) = 0$, but this estimator can be less efficient than $\hat{\beta}_{\mathrm{MLE}}$, especially when n^*/n is not large. Han and Lawless (2016) observed that an empirical likelihood approach could be applied. We develop this idea in the following sections.

3. The proposed empirical likelihood-based method

3.1 The proposed estimators

With the auxiliary summary information and the supplementary sample $(X_j^{*\mathrm{T}}, Z_j^{*\mathrm{T}})^{\mathrm{T}}$, $j = 1, \ldots, n^*$, we can construct estimators that are guaranteed to be more efficient than $\hat{\beta}_{\mathrm{MLE}}$. Han and Lawless (2016) noted that an approach of Qin (2000), also considered by Chen et al. (2003), could be applied. This involves $p_j^*, j = 1, \ldots, n^*$, an empirical distribution for the

supplementary sample, and defines an estimator $\hat{\beta}_{\mathrm{EL1}}$ through

$$\max_{\beta,p_1^*,\ldots,p_{n^*}^*} \prod_{i=1}^{n} f(Y_i \mid X_i, Z_i; \beta) \prod_{j=1}^{n^*} p_j^* \qquad \text{subject to}$$

$$p_j^* \geq 0, \quad \sum_{j=1}^{n^*} p_j^* = 1, \quad \sum_{j=1}^{n^*} p_j^* u(X_j^*, Z_j^*; \beta, \theta^*) = 0. \tag{3.1}$$

For convenience write $f(\beta) = f(Y \mid X, Z; \beta)$, $s(\beta) = \partial \log f(\beta)/\partial\beta$ and $u^*(\beta) = u(X^*, Z^*; \beta, \theta^*)$. In the Appendix we show that $\hat{\beta}_{\mathrm{EL1}}$ is the component of $(\hat{\beta}_{\mathrm{EL1}}^{\mathrm{T}}, \hat{\lambda}^{\mathrm{T}})^{\mathrm{T}}$ that satisfies

$$\sum_{i=1}^{n} s_i(\hat{\beta}_{\mathrm{EL1}}) + \sum_{j=1}^{n^*} \frac{\partial u_j^*(\hat{\beta}_{\mathrm{EL1}})/\partial\beta^{\mathrm{T}}}{1 - \hat{\lambda}^{\mathrm{T}} u_j^*(\hat{\beta}_{\mathrm{EL1}})} \hat{\lambda} = 0, \tag{3.2}$$

$$\sum_{j=1}^{n^*} \frac{u_j^*(\hat{\beta}_{\mathrm{EL1}})}{1 - \hat{\lambda}^{\mathrm{T}} u_j^*(\hat{\beta}_{\mathrm{EL1}})} = 0, \tag{3.3}$$

where λ is a vector of Lagrange multipliers and $\hat{p}_j^* = 1/[n^*\{1 - \hat{\lambda}^{\mathrm{T}} u_j^*(\hat{\beta}_{\mathrm{EL1}})\}]$ with

$$1 - \hat{\lambda}^{\mathrm{T}} u_j^*(\hat{\beta}_{\mathrm{EL1}}) > 1/n^*, \qquad j = 1, \ldots, n^*. \tag{3.4}$$

Based on Z-estimator theory (e.g., van der Vaart 1998), it is easy to see that $(\hat{\beta}_{\mathrm{EL1}}^{\mathrm{T}}, \hat{\lambda}^{\mathrm{T}})^{\mathrm{T}} \xrightarrow{p} (\beta_0^{\mathrm{T}}, 0^{\mathrm{T}})^{\mathrm{T}}$, and thus $\hat{\beta}_{\mathrm{EL1}}$ is a consistent estimator of β_0. To introduce the asymptotic distribution of $\hat{\beta}_{\mathrm{EL1}}$, write $S = E\{s(\beta_0)s(\beta_0)^{\mathrm{T}}\}$, $G^* = E^*\{\partial u^*(\beta_0)/\partial\beta\}$, $\Omega^* = E^*\{u^*(\beta_0)u^*(\beta_0)^{\mathrm{T}}\}$ and $\kappa = \lim_{n\to\infty} n^*/n$. In the Appendix we show that

$$\sqrt{n}(\hat{\beta}_{\mathrm{EL1}} - \beta_0) \xrightarrow{d} N\left(0, (S + \kappa G^{*\mathrm{T}} \Omega^{*-1} G^*)^{-1}\right). \tag{3.5}$$

It is clear that $G^{*\mathrm{T}}\Omega^{*-1}G^*$ is positive-definite, and thus the above asymptotic variance $V_{\mathrm{EL}} \equiv (S + \kappa G^{*\mathrm{T}}\Omega^{*-1}G^*)^{-1}$ is always smaller than $V_{\mathrm{MLE}} \equiv S^{-1}$, the asymptotic variance of $\hat{\beta}_{\mathrm{MLE}}$. Therefore, $\hat{\beta}_{\mathrm{EL1}}$ is guaranteed to be more efficient than $\hat{\beta}_{\mathrm{MLE}}$, and the efficiency improvement gets larger as κ increases.

The formulation (3.1) following Qin (2000) uses a parametric likelihood multiplied by a nonparametric likelihood. A full empirical likelihood formulation as in Qin and Lawless (1994) can be given by letting p_i, $i = 1, \ldots, n$, denote an empirical distribution supported on the study data sample. Then we define an estimator $\hat{\beta}_{\mathrm{EL2}}$ through

$$\max_{\beta, p_i\text{'s}, p_j^*\text{'s}} \prod_{i=1}^{n} p_i \prod_{j=1}^{n^*} p_j^* \qquad \text{subject to}$$

$$p_i \geq 0, \quad \sum_{i=1}^{n} p_i = 1, \quad \sum_{i=1}^{n} p_i s(Y_i, X_i, Z_i; \beta) = 0,$$

$$p_j^* \geq 0, \quad \sum_{j=1}^{n^*} p_j^* = 1, \quad \sum_{j=1}^{n^*} p_j^* u(X_j^*, Z_j^*; \beta, \theta^*) = 0.$$

Arguments similar to those in the Appendix leading to (3.2)-(3.3) show that

$\hat{\beta}_{\mathrm{EL2}}$ is the component of $(\hat{\beta}_{\mathrm{EL2}}^{\mathrm{T}}, \hat{\lambda}^{\mathrm{T}}, \hat{\rho}^{\mathrm{T}})^{\mathrm{T}}$ that satisfies

$$\sum_{i=1}^{n} \frac{\partial s_i(\hat{\beta}_{\mathrm{EL2}})/\partial\beta}{1-\hat{\lambda}^{\mathrm{T}} s_i(\hat{\beta}_{\mathrm{EL2}})}\hat{\lambda} + \sum_{j=1}^{n^*} \frac{\partial u_j^*(\hat{\beta}_{\mathrm{EL2}})/\partial\beta^{\mathrm{T}}}{1-\hat{\rho}^{\mathrm{T}} u_j^*(\hat{\beta}_{\mathrm{EL2}})}\hat{\rho} = 0, \tag{3.6}$$

$$\sum_{i=1}^{n} \frac{s_i(\hat{\beta}_{\mathrm{EL2}})}{1-\hat{\lambda}^{\mathrm{T}} s_i(\hat{\beta}_{\mathrm{EL2}})} = 0, \tag{3.7}$$

$$\sum_{j=1}^{n^*} \frac{u_j^*(\hat{\beta}_{\mathrm{EL2}})}{1-\hat{\rho}^{\mathrm{T}} u_j^*(\hat{\beta}_{\mathrm{EL2}})} = 0, \tag{3.8}$$

and $\hat{p}_i = 1/[n\{1-\hat{\lambda}^{\mathrm{T}} s_i(\hat{\beta}_{\mathrm{EL2}})\}]$ and $\hat{p}_j^* = 1/[n^*\{1-\hat{\rho}^{\mathrm{T}} u_j^*(\hat{\beta}_{\mathrm{EL2}})\}]$ with

$$1-\hat{\lambda}^{\mathrm{T}} s_i(\hat{\beta}_{\mathrm{EL2}}) > 1/n, \quad i=1,\ldots,n; \qquad 1-\hat{\rho}^{\mathrm{T}} u_j^*(\hat{\beta}_{\mathrm{EL2}}) > 1/n^*, \quad j=1,\ldots,n^*. \tag{3.9}$$

Based on Z-estimator theory again, it is easy to see that $(\hat{\beta}_{\mathrm{EL2}}^{\mathrm{T}}, \hat{\lambda}^{\mathrm{T}}, \hat{\rho}^{\mathrm{T}})^{\mathrm{T}} \xrightarrow{p} (\beta_0^{\mathrm{T}}, 0^{\mathrm{T}}, 0^{\mathrm{T}})^{\mathrm{T}}$, showing the consistency of $\hat{\beta}_{\mathrm{EL2}}$. In the Appendix, we show that $\sqrt{n}(\hat{\beta}_{\mathrm{EL2}} - \beta_0)$ has the same asymptotic distribution as in (3.5). In other words, $\hat{\beta}_{\mathrm{EL2}}$ is asymptotically equivalent to $\hat{\beta}_{\mathrm{EL1}}$, and thus is guaranteed to be more efficient than $\hat{\beta}_{\mathrm{MLE}}$.

3.2 Numerical implementation

Reliable procedures for obtaining empirical or constrained maximum likelihood estimates are sometimes elusive (e.g. Chaudhuri et al. 2008). A simple way to compute $\hat{\beta}_{\mathrm{EL1}}$ and $\hat{\beta}_{\mathrm{EL2}}$ seems to be to solve (3.2)-(3.3) for $(\beta^{\mathrm{T}}, \lambda^{\mathrm{T}})^{\mathrm{T}}$ and (3.6)-(3.8) for $(\beta^{\mathrm{T}}, \lambda^{\mathrm{T}}, \rho^{\mathrm{T}})^{\mathrm{T}}$, respectively. However, this way

is not recommended due to its unstable behavior: equations (3.3) and (3.7)-(3.8), viewed as equations for λ and $(\lambda^{\mathrm{T}}, \rho^{\mathrm{T}})^{\mathrm{T}}$ for a fixed β, typically have many roots (Han and Wang 2013), yet the ones we need are $\hat{\lambda}$ and $(\hat{\lambda}^{\mathrm{T}}, \hat{\rho}^{\mathrm{T}})^{\mathrm{T}}$ that satisfy (3.4) and (3.9), respectively. Directly solving those equations can lead to an unwanted root.

A more reliable implementation is to follow recommendations from the empirical likelihood literature by considering the saddle-point representation of $\hat{\beta}_{\mathrm{EL1}}$ and $\hat{\beta}_{\mathrm{EL2}}$ (e.g., Owen 2001; Imbens 2002; Kitamura 2007), based on which we outline a Newton-Raphson-type algorithm that can be shown to have good performance. The following discussion will focus on $\hat{\beta}_{\mathrm{EL1}}$ for simplicity. From the derivation of (3.2)-(3.3) in the Appendix, we have that, for a fixed β, the solution $\hat{p}_j^*(\beta)$ to (3.1) is given by $\hat{p}_j^*(\beta) = 1/[n^*\{1 - \hat{\lambda}(\beta)^{\mathrm{T}} u_j^*(\beta)\}]$, where $\hat{\lambda}(\beta)$ solves $\sum_{j=1}^{n^*} u_j^*(\beta)/\{1 - \lambda^{\mathrm{T}} u_j^*(\beta)\} = 0$. It is then easy to see that $\hat{\lambda}(\beta)$ actually minimizes

$$L^*(\lambda, \beta) \equiv -\sum_{j=1}^{n^*} \log\{1 - \lambda^{\mathrm{T}} u_j^*(\beta)\}, \tag{3.10}$$

and $\sum_{j=1}^{n^*} \log \hat{p}_j^*(\beta) = L^*\{\hat{\lambda}(\beta), \beta\} - n^* \log n^*$. Therefore, $\hat{\beta}_{\mathrm{EL1}}$ defined in (3.1) can be equivalently defined as

$$\hat{\beta}_{\mathrm{EL1}} = \arg\max_{\beta} \left\{ \sum_{i=1}^{n} \log f_i(\beta) + \min_{\lambda} L^*(\lambda, \beta) \right\} \equiv \arg\max_{\beta} M(\beta).$$

This is the so-called saddle-point representation due to the nested optimiza-

tions.

An implementation based on the Newton-Raphson algorithm requires the Jacobian $M_\beta(\beta) = \partial M(\beta)/\partial\beta$ and the Hessian $M_{\beta\beta}(\beta) = \partial^2 M(\beta)/\partial\beta\partial\beta^{\mathrm{T}}$ of $M(\beta)$, the expressions of which, together with some simplifications, are given in the Appendix. Both $M_\beta(\beta)$ and $M_{\beta\beta}(\beta)$ involve $\hat{\lambda}(\beta)$, whose value at the current β in each iteration can be calculated by minimizing $L^*(\lambda, \beta)$ in (3.10) with respect to λ. This minimization requires the Jacobian $L^*_\lambda(\lambda, \beta)$ and Hessian $L^*_{\lambda\lambda}(\lambda, \beta)$:

$$L^*_\lambda(\lambda, \beta) = \sum_{j=1}^{n^*} \frac{u^*_j(\beta)}{1 - \lambda^{\mathrm{T}} u^*_j(\beta)}, \qquad L^*_{\lambda\lambda}(\lambda, \beta) = \sum_{j=1}^{n^*} \frac{u^*_j(\beta) u^*_j(\beta)^{\mathrm{T}}}{\{1 - \lambda^{\mathrm{T}} u^*_j(\beta)\}^2}.$$

The implementation consists of two loops: the outer loop updates β using $M_\beta(\beta)$ and $M_{\beta\beta}(\beta)$, where the needed $\hat{\lambda}(\beta)$ at the current β is calculated by the inner loop. The following is an algorithm describing how this works.

Outer loop:

Step 0: Set $l = 0$ (iteration count), $\sigma = 5$ (maximum number of attempts within each iteration to find a steplength making $M(\beta)$ increase), and $\epsilon = 10^{-4}$ (algorithm convergence criterion). Let $\hat{\beta}^{(0)} = \hat{\beta}_{\mathrm{MLE}}$ (initial value of β) and $M^{(0)} = M(\hat{\beta}^{(0)})$ (the calculation here invokes the inner loop).

Step 1: Calculate $\Delta^{(l)} = M_{\beta\beta}(\hat{\beta}^{(l)})^{-1}M_{\beta}(\hat{\beta}^{(l)})$ (direction for updating $\hat{\beta}^{(l)}$, the calculation here invokes the inner loop). Set $\tau = 1$ (the initial steplength taken along the direction $\Delta^{(l)}$) and $t = 0$ (number of attempts to find a steplength making $M(\beta)$ increase).

Step 2: Calculate $\hat{\beta}^{\text{temp}} = \hat{\beta}^{(l)} - \tau\Delta^{(l)}$ and $M^{\text{temp}} = M(\hat{\beta}^{\text{temp}})$ (the calculation here invokes the inner loop). If $M^{\text{temp}} > M^{(l)}$ or $t = \sigma$, then go to Step 3; otherwise let $t = t + 1$ and $\tau = \tau/2$ and repeat Step 2.

Step 3: Let $\hat{\beta}^{(l+1)} = \hat{\beta}^{\text{temp}}$ and $M^{(l+1)} = M^{\text{temp}}$. If $\parallel \hat{\beta}^{(l+1)} - \hat{\beta}^{(l)} \parallel_1 < \epsilon$, let $\hat{\beta}_{\text{EL1}} = \hat{\beta}^{(l+1)}$ and stop the algorithm; otherwise set $l = l + 1$ and go back to Step 1.

Inner loop:

Step 0: Set $l = 0$ (iteration count) and $\epsilon = 10^{-4}$ (algorithm convergence criterion). Let $\hat{\lambda}^{(0)} = 0$ (initial value of λ) and $L^{*(0)} = 0$.

Step 1: Calculate $\Delta^{(l)} = L^{*}_{\lambda\lambda}(\hat{\lambda}^{(l)}, \beta)^{-1}L^{*}_{\lambda}(\hat{\lambda}^{(l)}, \beta)$ (direction for updating $\hat{\lambda}^{(l)}$). Set $\tau = 1$ (the initial steplength taken along the direction $\Delta^{(l)}$).

Step 2: Calculate $\hat{\lambda}^{\text{temp}} = \hat{\lambda}^{(l)} - \tau\Delta^{(l)}$. If $\hat{\lambda}^{\text{temp}}$ satisfies $1 - (\hat{\lambda}^{\text{temp}})^{\text{T}}u^{*}_{j}(\beta) > 1/n^{*}$ for $j = 1, \ldots, n^{*}$ and $L^{*\text{temp}} \equiv L^{*}(\hat{\lambda}^{\text{temp}}, \beta) < L^{*(l)}$, then go to Step 3; otherwise let $\tau = \tau/2$ and repeat Step 2.

Step 3: Let $\hat{\lambda}^{(l+1)} = \hat{\lambda}^{\text{temp}}$ and $L^{*(l+1)} = L^{*\text{temp}}$. If $\| \hat{\lambda}^{(l+1)} - \hat{\lambda}^{(l)} \|_1 < \epsilon$, let $\hat{\lambda} = \hat{\lambda}^{(l+1)}$ and stop the algorithm; otherwise set $l = l+1$ and go back to Step 1.

In the outer loop, Step 2 sequentially tries steplengths 1, 2^{-1}, ..., 2^{-5} along the direction for updating β, and will take the first one that makes $M(\beta)$ increase. When such an increase does not occur for any of those steplengths, we still update β by taking the steplength to be 2^{-5}. This is because the current β might be a local maximizer instead of the global one, and continuing updating β could take the iterations out of this region. In the inner loop, with β fixed, Step 2 sequentially tries steplengths 1, 2^{-1}, ..., along the direction for updating λ, and will take the first one that satisfies $1 - \lambda^{\mathrm{T}} u_j^*(\beta) > 1/n^*$ for $j = 1, \ldots, n^*$ and makes $L^*(\lambda, \beta)$ decrease. Such a steplength always exists because $\hat{\lambda}^{(0)} = 0$ and $L^*(\lambda, \beta)$ is a strictly convex function of λ. The inner loop usually always converges (Chen et al. 2002; Han 2014). The initial value for the outer loop, $\hat{\beta}^{(0)} = \hat{\beta}_{\text{MLE}}$, is a consistent estimator of β_0, and the initial value for the inner loop, $\hat{\lambda}^{(0)} = 0$, is the probability limit of $\hat{\lambda}$. Therefore, the convergence of the above algorithm is usually fast.

For $\hat{\beta}_{\text{EL2}}$, we have the following saddle-point representation:

$$\hat{\beta}_{\mathrm{EL2}} = \arg\max_{\beta} \left\{ \min_{\lambda} \left[-\sum_{i=1}^{n} \log\{1 - \lambda^{\mathrm{T}} s_i(\beta)\} \right] + \min_{\rho} \left[-\sum_{j=1}^{n^*} \log\{1 - \rho^{\mathrm{T}} u_j^*(\beta)\} \right] \right\}.$$

The determination of $\hat{\beta}_{\mathrm{EL2}}$ is similar to before and details are omitted here.

3.3 Uncertainty of the auxiliary summary information

If the auxiliary data set is not sufficiently large, the variability associated with the auxiliary summary information may be non-negligible and affect the properties of $\hat{\beta}_{\mathrm{EL1}}$ and $\hat{\beta}_{\mathrm{EL2}}$. To study this effect, let N^* denote the sample size for the auxiliary data set from which the auxiliary estimate $\hat{\theta}$ is derived based on solving $\sum_{k=1}^{N^*} h(Y_k, X_k; \theta) = 0$, and let V_{θ^*} be the asymptotic variance of $\sqrt{N^*}(\hat{\theta} - \theta^*)$, where now θ^* is the unknown probability limit of $\hat{\theta}$. The auxiliary summary information now includes $\hat{\theta}$ and $\hat{V}_{\hat{\theta}}$, where $\hat{V}_{\hat{\theta}}$ is an estimate of V_{θ^*}.

The estimation procedures are as before but with $\hat{\theta}$ replacing θ^*. It turns out that in this case the asymptotic distribution of $\hat{\beta}_{\mathrm{EL1}}$ depends on whether the supplementary sample is independent of or a subset of the auxiliary data set. In the former case, as shown in the Appendix, $\sqrt{n}(\hat{\beta}_{\mathrm{EL1}}^{\mathrm{T}} - \beta_0^{\mathrm{T}}, \hat{\lambda}^{\mathrm{T}})^{\mathrm{T}}$ has

an asymptotic Normal distribution with mean zero and variance

$$\begin{pmatrix} -S, & \kappa G^{*\mathrm{T}} \\ \kappa G^{*}, & \kappa \Omega^{*} \end{pmatrix}^{-1} \begin{pmatrix} S, & 0 \\ 0, & \kappa(\Omega^{*} + \kappa^{*} Q^{*} V_{\theta^{*}} Q^{*\mathrm{T}}) \end{pmatrix} \begin{pmatrix} -S, & \kappa G^{*\mathrm{T}} \\ \kappa G^{*}, & \kappa \Omega^{*} \end{pmatrix}^{-1}, \tag{3.11}$$

where $\kappa^{*} = \lim_{n\to\infty} n^{*}/N^{*}$ and $Q^{*} = E^{*}\{\partial u^{*}(\beta_0, \theta^{*})/\partial\theta\}$. An explicit but messy expression for the asymptotic variance of $\sqrt{n}(\hat{\beta}_{\mathrm{EL1}} - \beta_0)$ may then be derived, but is not necessary for implementation purposes since we can calculate (3.11) as a whole and then extract the corresponding sub-matrix for $\hat{\beta}_{\mathrm{EL1}}$.

From the proof of (3.5) in the Appendix, the asymptotic variance of $\sqrt{n}(\hat{\beta}_{\mathrm{EL1}}^{\mathrm{T}} - \beta_0^{\mathrm{T}}, \hat{\lambda}^{\mathrm{T}})^{\mathrm{T}}$ when θ^{*} is used instead of $\hat{\theta}$ is

$$\begin{pmatrix} -S, & \kappa G^{*\mathrm{T}} \\ \kappa G^{*}, & \kappa \Omega^{*} \end{pmatrix}^{-1} \begin{pmatrix} S, & 0 \\ 0, & \kappa \Omega^{*} \end{pmatrix} \begin{pmatrix} -S, & \kappa G^{*\mathrm{T}} \\ \kappa G^{*}, & \kappa \Omega^{*} \end{pmatrix}^{-1}. \tag{3.12}$$

Since $Q^{*} V_{\theta^{*}} Q^{*\mathrm{T}}$ is positive-definite, a comparison between (3.11) and (3.12) reveals that the variability of $\hat{\theta}$ always increases the asymptotic variance of $\hat{\beta}_{\mathrm{EL1}}$. Therefore, confidence intervals for $\hat{\beta}_{\mathrm{EL1}}$ ignoring this uncertainty will have coverage rate smaller than the nominal level.

When the supplementary sample is a subset of the auxiliary data set, it is shown in the Appendix that $\sqrt{n}(\hat{\beta}_{\mathrm{EL1}}^{\mathrm{T}} - \beta_0^{\mathrm{T}}, \hat{\lambda}^{\mathrm{T}})^{\mathrm{T}}$ has an asymptotic

Normal distribution with mean zero and variance

$$\begin{pmatrix} -S, & \kappa G^{*\mathrm{T}} \\ \kappa G^*, & \kappa \Omega^* \end{pmatrix}^{-1} \begin{pmatrix} S, & 0 \\ 0, & \kappa\{(1-2\kappa^*)\Omega^* + \kappa^* Q^* V_{\theta^*} Q^{*\mathrm{T}}\} \end{pmatrix} \begin{pmatrix} -S, & \kappa G^{*\mathrm{T}} \\ \kappa G^*, & \kappa \Omega^* \end{pmatrix}^{-1}. \tag{3.13}$$

A comparison between (3.13) and (3.12) leads to a surprising observation: the variability of $\hat{\theta}$ increases the asymptotic variance of $\hat{\beta}_{\mathrm{EL1}}$ when $Q^* V_{\theta^*} Q^{*\mathrm{T}} > 2\Omega^*$ and reduces the asymptotic variance when $Q^* V_{\theta^*} Q^{*\mathrm{T}} < 2\Omega^*$. Some calculation shows that $Q^* V_{\theta^*} Q^{*\mathrm{T}} = \Omega^* + E^*[\mathrm{Var}\{h(\theta^*) \mid X, Z\}]$. Therefore, the uncertainty of $\hat{\theta}$ will reduce the asymptotic variance of $\hat{\beta}_{\mathrm{EL1}}$ when $E^*[\mathrm{Var}\{h(\theta^*) \mid X, Z\}] < \Omega^*$. In other words, confidence intervals for $\hat{\beta}_{\mathrm{EL1}}$ ignoring this uncertainty will have coverage rate smaller than the nominal level when $E^*[\mathrm{Var}\{h(\theta^*) \mid X, Z\}] > \Omega^*$ and larger than the nominal level when $E^*[\mathrm{Var}\{h(\theta^*) \mid X, Z\}] < \Omega^*$.

In general, the asymptotic variance of the proposed estimators is affected by κ and κ^*. Because this effect is tangled up with quantities depending on the data distribution, a quantitative assessment of it is difficult. For example, consider the case where the supplementary sample is independent of the auxiliary data set. The asymptotic variance is determined by (3.11). When $\kappa = 0$, the asymptotic variance becomes that of $\hat{\beta}_{\mathrm{MLE}}$ based on the current study data alone, and thus the auxiliary information

is no longer useful. When $\kappa^* = 0$, (3.11) reduces to (3.12), the asymptotic variance with θ^* known. The case where $\kappa = \infty$ or $\kappa^* = \infty$ is not practically meaningful because n^* is typically small as the micro data from the auxiliary data set is not available. When n^*/n and n^*/N^* vary, the asymptotic variance varies between that of $\hat{\beta}_{\mathrm{MLE}}$ and that using a known θ^*, but a quantification is difficult because the quantities in (3.11) depend on the data distribution.

In the special case that the summary information $\hat{\theta}$ is an estimate of $\theta^* = E^*(Y)$ for the auxiliary study population, calculated as the sample average of the auxiliary data $\{Y_k : \ k = 1, \ldots, N^*\}$, we have $h(Y; \theta) = Y - \theta$. Some calculation shows that $G^* = E^*\{Y s(\beta_0)^{\mathrm{T}}\}$, $\Omega^* = \mathrm{Var}^*\{E(Y \mid X, Z)\}$, $Q^* = -1$ and $V_{\theta^*} = \mathrm{Var}^*(Y)$ in this case. Therefore, when the supplementary sample is a subset of the auxiliary data set, the uncertainty of $\hat{\theta}$ will reduce the asymptotic variance when $E^*\{\mathrm{Var}(Y \mid X, Z)\} < \mathrm{Var}^*\{E(Y \mid X, Z)\}$.

All of the above conclusions also apply to $\hat{\beta}_{\mathrm{EL2}}$ because it has the same asymptotic expansion as $\hat{\beta}_{\mathrm{EL1}}$.

4. Some simulation and analytical results

4.1 Logistic regression

We first examine efficiency for logistic regression, which was considered by Qin et al. (2015) and Chatterjee et al. (2016) in case-control settings. Two covariates X and Z are assumed to jointly follow a bivariate normal distribution with marginal means zero and marginal variances one. The correlation coefficient is taken for illustration as $\rho = 0.5$ for the current study population and $\rho^* = 0.1$ for the auxiliary data population. Given X and Z, Y follows a Bernoulli distribution with $\text{logit}\{P(Y = 1 \mid X, Z)\} = \beta_{0c} + \beta_{0X}X + \beta_{0Z}Z + \beta_{0XZ}XZ$, where $\text{logit}(\pi) = \log\{\pi/(1-\pi)\}$ and $\beta_0^{\mathrm{T}} = (\beta_{0c}, \beta_{0X}, \beta_{0Z}, \beta_{0XZ}) = (0.5, -0.5, -0.5, 0.5)$. For the auxiliary data set, we assume the model $\text{logit}\{P(Y = 1 \mid X)\} = \theta_c + \theta_X X$ was fitted for $f^*(Y \mid X)$ using the maximum likelihood score function $h(Y, X; \theta) = (1, X)^{\mathrm{T}}\{Y - \text{expit}(\theta_c + \theta_X X)\}$, where $\text{expit}(\gamma) = e^\gamma/(1 + e^\gamma)$. Notice that this is a misspecified model for $f^*(Y \mid X)$. The value of θ^* may be calculated numerically as the solution to $E^*\{h(Y, X; \theta)\} = 0$. We then have $u(X, Z; \beta, \theta^*) = P(Y = 1 \mid X, Z; \beta)h(Y = 1, X; \theta^*) + P(Y = 0 \mid X, Z; \beta)h(Y = 0, X; \theta^*)$.

Table 1 contains simulation results based on 1000 replications. Scenar-

ios 1-3 correspond to (i) no uncertainty in the auxiliary summary information, (ii) uncertainty in the auxiliary summary information (that is, $\hat{\theta}$ replaces θ^*) and the supplementary sample is independent of the auxiliary data set, and (iii) uncertainty in the auxiliary summary information and the supplementary sample is a subset of the auxiliary data set, respectively. For all scenarios, we take $n = 300$ for the study sample and $n^* = 100$ for the supplementary sample; for scenarios 2 and 3, we take $N^* = 500$ for the auxiliary data set used to calculate $\hat{\theta}$ and $\hat{V}_{\hat{\theta}}$. In all scenarios, $\hat{\beta}_{\mathrm{EL1}}$ and $\hat{\beta}_{\mathrm{EL2}}$ have almost identical performance, and both have considerably smaller empirical standard error than $\hat{\beta}_{\mathrm{MLE}}$ for the components β_c and β_X corresponding to the regressors that were included in the auxiliary data model. The reduction is much less in scenarios 2 and 3 where variability in $\hat{\theta}$ is non-negligible. We note that $N^* = 500$ is small for most auxiliary data bases and in practice large data sets will give gains close to those in scenario 1. We also note that efficiency gains are small for β_Z and β_{XZ}, even in scenario 1. These results agree qualitatively with results of Qin et al. (2015) and Chatterjee et al. (2016) in case-control settings, though the former paper assumes the covariate distributions are the same. The coverage probabilities of the 95% confidence intervals constructed using the asymptotic distributions, with no uncertainty in scenario 1 and with ad-

justment for uncertainty in scenarios 2 and 3, are very close to the nominal level.

4.2 Normal linear regression

To gain more insight on how efficiency improvement might be affected by different factors we next consider linear regression, where mathematical calculation is feasible. Let X and Z be generated as before but with ρ and ρ^* unspecified. We assume that the model $N(\theta_c + \theta_X X, 1)$ was fitted for the auxiliary data, and the model $N(\beta_c + \beta_X X + \beta_Z Z + \beta_{XZ} XZ, 1)$ holds for $f(Y \mid X, Z)$. Here the variances of the normal distributions are assumed known to simplify the calculations. The auxiliary data model leads to $h(Y, X; \theta) = (1, X)^{\mathrm{T}}(Y - \theta_c - \theta_X X)$ and then $u(X, Z; \beta, \theta) = (1, X)^{\mathrm{T}}(\beta_c + \beta_X X + \beta_Z Z + \beta_{XZ} XZ - \theta_c - \theta_X X)$. Solving $E^*\{h(Y, X; \theta^*)\} = 0$ gives $\theta_c^* = \beta_{0c} + \beta_{0XZ}\rho^*$ and $\theta_X^* = \beta_{0X} + \beta_{0Z}\rho^*$. Thus the auxiliary data provide information only on a two-dimensional function of β_{0c}, β_{0X}, β_{0Z} and β_{0XZ}. Some straightforward calculations show that

$$S = \begin{pmatrix} 1 & 0 & 0 & \rho \\ 0 & 1 & \rho & 0 \\ 0 & \rho & 1 & 0 \\ \rho & 0 & 0 & 1 + 2\rho^2 \end{pmatrix}, \qquad G^* = \begin{pmatrix} 1 & 0 & 0 & \rho^* \\ 0 & 1 & \rho^* & 0 \end{pmatrix},$$

$$\text{and} \quad \Omega^* = \begin{pmatrix} \beta_{0Z}^2(1-\rho^{*2}) + \beta_{0XZ}^2(1+\rho^{*2}) & 2\beta_{0Z}\beta_{0XZ}(1-\rho^{*2}) \\ 2\beta_{0Z}\beta_{0XZ}(1-\rho^{*2}) & \beta_{0Z}^2(1-\rho^{*2}) + \beta_{0XZ}^2(3+7\rho^{*2}) \end{pmatrix}. \tag{4.1}$$

From (3.5), it is then seen that efficiency improvement becomes more dramatic when $|\beta_{0Z}|$ and $|\beta_{0XZ}|$ are small. For example, taking $\rho = 0.5$, $\rho^* = 0.1$ and $\kappa = 1/3$, the square root of the ratio of the diagonal elements of V_{EL} to those of V_{MLE} is $(0.68, 0.81, 0.97, 0.97)$ when $\beta_0^{\text{T}} = (0.5, -0.5, -0.5, 0.5)$ and $(0.40, 0.52, 0.93, 0.95)$ when $\beta_0^{\text{T}} = (0.5, -0.5, -0.2, 0.2)$. This observation makes intuitive sense because weak association between Y and (Z, XZ) means that the fitted auxiliary data model $N(\theta_c + \theta_X X, 1)$ is close to the model of interest $N(\beta_c + \beta_X X + \beta_Z Z + \beta_{XZ} XZ, 1)$, and thus should lead to better efficiency improvement. This observation is also confirmed by simulation results for the models here, which are omitted due to their similarity to those based on logistic regression. Imbens and Lancaster (1994) found similar behaviour for probit binary response models using generalized method of moments estimators when the covariate distributions are the same.

For the above linear regression case, we are also able to more closely examine the effect of the uncertainty in $\hat{\theta}$ for scenario 3 where, as shown

by our theoretical results, the definiteness of $E^*[\mathrm{Var}\{h(\theta^*) \mid X, Z\}] - \Omega^*$ determines whether using $\hat{\theta}$ increases or reduces the asymptotic variance of $\hat{\beta}_{\mathrm{EL1}}$ and $\hat{\beta}_{\mathrm{EL2}}$ compared to that using θ^*. For this case, Ω^* is given in (4.1), and simple calculation shows that $E^*[\mathrm{Var}\{h(\theta^*) \mid X, Z\}]$ is the identity matrix. Taking $\beta_0^{\mathrm{T}} = (1, 1, 1, 1)$ as an example, it is easy to see that $E^*[\mathrm{Var}\{h(\theta^*) \mid X, Z\}] - \Omega^*$ is never positive-definite and is negative-definite when $|\rho^*| > 0.27$. In other words, using $\hat{\theta}$ instead of θ^* will reduce the asymptotic variance of $\hat{\beta}_{\mathrm{EL1}}$ and $\hat{\beta}_{\mathrm{EL2}}$ when $|\rho^*| > 0.27$. Figure 1 plots, as a function of ρ^*, the ratio of the asymptotic standard deviation of $\sqrt{n}(\hat{\beta}_{\mathrm{EL1}} - \beta_0)$ using $\hat{\theta}$ versus that using θ^*, taking $\rho = 0.5$, $\kappa = 1/3$ and $\kappa^* = 1/5$. Clearly, when $|\rho^*| > 0.27$, the uncertainty of $\hat{\theta}$ reduces the asymptotic variance. When $|\rho^*| < 0.27$, this uncertainty may reduce the asymptotic variance for some regression coefficients but increase it for others. The impact, however, is very small as can be seen from the scale of the y-axis. Other aspects of efficiency can be examined using (4.1), such as the effects of ρ and ρ^*.

5. Concluding remarks

The fact that a supplementary sample of $(X^{\mathrm{T}}, Z^{\mathrm{T}})^{\mathrm{T}}$ from the auxiliary data population is needed when $f(X, Z) \neq f^*(X, Z)$ limits the use of auxiliary

summary information, and the methodology here, to cases where this can be obtained. This is feasible in settings where individual-level data can be produced from the auxiliary data base. If, however, micro data are not available or if covariates Z are not included in such data, then a randomly selected supplementary sample of individuals from the auxiliary data population is needed to measure the covariates, and the cost of doing this should be weighted against potential efficiency gains or, in some cases, the cost of expanding the current study. Awareness is growing of the need for a careful consideration of covariate distributions, and for methodology to deal with situations where they differ across populations. In addition to the analysis of a current study "borrowing strength" from external summary data, this also applies to the comparison or integration of results from different studies. It is essential that information concerning different populations be compared critically in order for auxiliary data's usefulness to be assessed more fully.

The two proposed estimators here are asymptotically equivalent and performed similarly in simulations, though more studies are needed to compare their finite-sample behavior in more detail. On computational grounds we would recommend the use of $\hat{\beta}_{\mathrm{EL1}}$ since its implementation involves only one Lagrange multiplier $\hat{\lambda}$, whereas the implementation of $\hat{\beta}_{\mathrm{EL2}}$ involves

both $\hat{\lambda}$ and $\hat{\rho}$. A larger set of Lagrange multipliers may affect the performance of the optimization procedures.

Let $\hat{g}(\beta) = \{n^{-1}\sum_{i=1}^{n} s_i(\beta)^{\mathrm{T}}, (n^*)^{-1}\sum_{j=1}^{n^*} u_j^*(\beta)^{\mathrm{T}}\}^{\mathrm{T}}$. An alternative to the proposed empirical likelihood estimators is the generalized method of moments estimator minimizing $\hat{g}(\beta)^{\mathrm{T}}\hat{C}(\beta)^{-1}\hat{g}(\beta)$, where

$$\hat{C}(\beta) = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n} s_i(\beta)s_i(\beta)^{\mathrm{T}} & 0 \\ 0 & \frac{n}{n^*}\frac{1}{n^*}\sum_{j=1}^{n^*} u_j^*(\beta)u_j^*(\beta)^{\mathrm{T}} \end{pmatrix}$$

is the sample version of $C(\beta_0) = \mathrm{diag}(S, \kappa^{-1}\Omega^*)$, the asymptotic variance of $\sqrt{n}\hat{g}(\beta_0)$. Standard results on generalized method of moments (Hansen 1982) show that this estimator is asymptotically equivalent to the ones we have proposed. The well established comparisons between generalized method of moments and empirical likelihood apply here (e.g., Imbens 2002; Newey and Smith 2004). Imbens and Lancaster (1994) considered auxiliary information and the generalized method of moments when $f(X, Z) = f^*(X, Z)$.

The parametric model $f(Y \mid X, Z; \beta)$ assumed in the current study can be checked using relevant goodness-of-fit tests. The assumption that the study population and the auxiliary data population have the same distribution $f(Y \mid X, Z)$ is more difficult to check in the setting considered here, where only summary information plus a supplementary sample on

$(X^{\mathrm{T}}, Z^{\mathrm{T}})^{\mathrm{T}}$ is available for the latter population. One check is to compare $\hat{\beta}_{\mathrm{MLE}}$ with $\hat{\beta}_{\mathrm{EL1}}$ or $\hat{\beta}_{\mathrm{EL2}}$, with a significant lack of agreement suggesting departures from this assumption (e.g., Imbens and Lancaster 1994; Chatterjee et al. 2016). Another check is to evaluate the average of $u(X, Z; \hat{\beta}_{\mathrm{MLE}}, \hat{\theta})$ over the supplementary sample, a significant difference from 0 indicating violation of this assumption. We note, however, that such checks cannot detect certain types of differences in the distributions for Y given X and Z (e.g., Newey 1985) and that a supplementary sample of $(Y, X^{\mathrm{T}}, Z^{\mathrm{T}})^{\mathrm{T}}$ or background information is needed to remedy this. These issues will be examined in more detail elsewhere.

Acknowledgements We wish to thank the Editor, an Associate Editor and two reviewers for their helpful comments. This research was partially supported by grants from the Natural Sciences and Engineering Research Council of Canada (NSERC) to the two authors.

Appendix

Derivation of (3.2) and (3.3)

The Lagrangian corresponding to the constrained optimization problem (3.1) is

$$\mathcal{L} = \sum_{i=1}^{n} \log f_i(\beta) + \sum_{j=1}^{n^*} \log p_j^* + n^* \lambda^{\mathrm{T}} \sum_{j=1}^{n^*} p_j^* u_j^*(\beta) - \mu(\sum_{j=1}^{n^*} p_j^* - 1),$$

where λ and μ are the Lagrange multipliers. At the solution $\hat{\beta}_{\mathrm{EL1}}$ and $\hat{p}_j^*$ we must have $\partial\mathcal{L}/\partial p_j^* = 0$ and $\partial\mathcal{L}/\partial\beta = 0$. Multiplying both sides of $\partial\mathcal{L}/\partial p_j^* = 1/p_j^* + n^*\lambda^{\mathrm{T}} u_j^*(\beta) - \mu$ by p_j^* and summing over j, the constraints in (3.1) lead to $\hat{\mu} = n^*$, which, combined with $\partial\mathcal{L}/\partial p_j^* = 0$ yields $\hat{p}_j^* = 1/[n^*\{1 - \hat{\lambda}^{\mathrm{T}} u_j^*(\hat{\beta}_{\mathrm{EL1}})\}]$. Then $\partial\mathcal{L}/\partial\beta = 0$ gives (3.2) and the constraint $\sum_{j=1}^{n^*} \hat{p}_j^* u_j^*(\hat{\beta}_{\mathrm{EL1}}) = 0$ gives (3.3).

Proof of (3.5)

Applying mean-value theorem to (3.2)-(3.3) around $(\beta_0^{\mathrm{T}}, 0^{\mathrm{T}})^{\mathrm{T}}$ leads to

$$0 = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n} s_i(\beta_0) \\ \frac{\sqrt{n^*}}{n}\frac{1}{\sqrt{n^*}}\sum_{j=1}^{n^*} u_j^*(\beta_0) \end{pmatrix} + \begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n} \frac{\partial s_i(\bar{\beta})}{\partial\beta}, & \frac{n^*}{n}\frac{1}{n^*}\sum_{j=1}^{n^*} \frac{\partial u_j^*(\hat{\beta}_{\mathrm{EL1}})/\partial\beta^{\mathrm{T}}}{1-\hat{\lambda}^{\mathrm{T}} u_j^*(\hat{\beta}_{\mathrm{EL1}})} \\ \frac{n^*}{n}\frac{1}{n^*}\sum_{j=1}^{n^*} \frac{\partial u_j^*(\bar{\beta})/\partial\beta}{1-\bar{\lambda}^{\mathrm{T}} u_j^*(\hat{\beta}_{\mathrm{EL1}})}, & \frac{n^*}{n}\frac{1}{n^*}\sum_{j=1}^{n^*} \frac{u_j^*(\bar{\beta}) u_j^*(\hat{\beta}_{\mathrm{EL1}})^{\mathrm{T}}}{\{1-\bar{\lambda}^{\mathrm{T}} u_j^*(\hat{\beta}_{\mathrm{EL1}})\}^2} \end{pmatrix} \begin{pmatrix} \hat{\beta}_{\mathrm{EL1}} - \beta_0 \\ \hat{\lambda} \end{pmatrix}$$

where $\bar{\beta}$ is some value between $\hat{\beta}_{\mathrm{EL1}}$ and β_0 and $\bar{\lambda}$ is some value between $\hat{\lambda}$ and 0. Then we have

$$\sqrt{n}\begin{pmatrix} \hat{\beta}_{\mathrm{EL1}} - \beta_0 \\ \hat{\lambda} \end{pmatrix} = -\begin{pmatrix} -S, & \kappa G^{*\mathrm{T}} \\ \kappa G^*, & \kappa\Omega^* \end{pmatrix}^{-1} \begin{pmatrix} \frac{1}{\sqrt{n}}\sum_{i=1}^{n} s_i(\beta_0) \\ \frac{\sqrt{n^*}}{\sqrt{n}}\frac{1}{\sqrt{n^*}}\sum_{j=1}^{n^*} u_j^*(\beta_0) \end{pmatrix} + o_p(1). \tag{5.1}$$

From Central Limit Theorem, $\sqrt{n}(\hat{\beta}_{\mathrm{EL1}}^{\mathrm{T}} - \beta_0^{\mathrm{T}}, \hat{\lambda}^{\mathrm{T}})^{\mathrm{T}}$ has an asymptotic Normal distribution with mean 0 and variance

$$\begin{aligned}
&\begin{pmatrix} -S, & \kappa G^{*\mathrm{T}} \\ \kappa G^*, & \kappa\Omega^* \end{pmatrix}^{-1} \begin{pmatrix} S, & 0 \\ 0, & \kappa\Omega^* \end{pmatrix} \begin{pmatrix} -S, & \kappa G^{*\mathrm{T}} \\ \kappa G^*, & \kappa\Omega^* \end{pmatrix}^{-1} \\
= &\begin{pmatrix} (S + \kappa G^{*\mathrm{T}}\Omega^{*-1}G^*)^{-1}, & 0 \\ 0, & (\kappa\Omega^* + \kappa^2 G^* S^{-1} G^{*\mathrm{T}})^{-1} \end{pmatrix},
\end{aligned}$$

which shows (3.5).

Derivation of the asymptotic distribution of $\sqrt{n}(\hat{\beta}_{\mathbf{EL2}} - \beta_0)$

Applying mean-value theorem to (3.6)-(3.8) around $(\beta_0^{\mathrm{T}}, 0^{\mathrm{T}}, 0^{\mathrm{T}})^{\mathrm{T}}$ leads to

$$\begin{aligned}
0 = &\begin{pmatrix} \frac{1}{n}\sum_{i=1}^n s_i(\beta_0) \\ \frac{\sqrt{n^*}}{n}\frac{1}{\sqrt{n^*}}\sum_{j=1}^{n^*} u_j^*(\beta_0) \\ 0 \end{pmatrix} + \\
&\begin{pmatrix} \frac{1}{n}\sum_{i=1}^n \frac{\partial s_i(\bar{\beta})/\partial\beta}{1-\bar{\lambda}^{\mathrm{T}} s_i(\hat{\beta}_{\mathrm{EL2}})}, & \frac{1}{n}\sum_{i=1}^n \frac{s_i(\bar{\beta}) s_i(\hat{\beta}_{\mathrm{EL2}})^{\mathrm{T}}}{\{1-\bar{\lambda}^{\mathrm{T}} s_i(\hat{\beta}_{\mathrm{EL2}})\}^2}, & 0 \\ \frac{n^*}{n}\frac{1}{n^*}\sum_{j=1}^{n^*} \frac{\partial u_j^*(\bar{\beta})/\partial\beta}{1-\bar{\rho}^{\mathrm{T}} u_j^*(\hat{\beta}_{\mathrm{EL2}})}, & 0, & \frac{n^*}{n}\frac{1}{n^*}\sum_{j=1}^{n^*} \frac{u_j^*(\bar{\beta}) u_j^*(\hat{\beta}_{\mathrm{EL2}})^{\mathrm{T}}}{\{1-\bar{\rho}^{\mathrm{T}} u_j^*(\hat{\beta}_{\mathrm{EL2}})\}^2} \\ 0, & \frac{1}{n}\sum_{i=1}^n \frac{\partial s_i(\hat{\beta}_{\mathrm{EL2}})/\partial\beta}{1-\hat{\lambda}^{\mathrm{T}} s_i(\hat{\beta}_{\mathrm{EL2}})}, & \frac{n^*}{n}\frac{1}{n^*}\sum_{j=1}^{n^*} \frac{\partial u_j^*(\hat{\beta}_{\mathrm{EL2}})/\partial\beta^{\mathrm{T}}}{1-\hat{\rho}^{\mathrm{T}} u_j^*(\hat{\beta}_{\mathrm{EL2}})} \end{pmatrix} \begin{pmatrix} \hat{\beta}_{\mathrm{EL2}} - \beta_0 \\ \hat{\lambda} \\ \hat{\rho} \end{pmatrix} \\
= &\begin{pmatrix} \frac{1}{\sqrt{n}}\sum_{i=1}^n s_i(\beta_0) \\ \frac{\sqrt{n^*}}{\sqrt{n}}\frac{1}{\sqrt{n^*}}\sum_{j=1}^{n^*} u_j^*(\beta_0) \\ 0 \end{pmatrix} + \begin{pmatrix} -S, & S, & 0 \\ \kappa G^*, & 0, & \kappa\Omega^* \\ 0, & -S, & \kappa G^{*\mathrm{T}} \end{pmatrix} \sqrt{n} \begin{pmatrix} \hat{\beta}_{\mathrm{EL2}} - \beta_0 \\ \hat{\lambda} \\ \hat{\rho} \end{pmatrix} + o_p(1),
\end{aligned}$$

where $\bar{\beta}$ is some value between $\hat{\beta}_{\mathrm{EL2}}$ and β_0, $\bar{\lambda}$ is some value between $\hat{\lambda}$ and 0, and $\bar{\rho}$ is some value between $\hat{\rho}$ and 0. Then we have $S\hat{\lambda} = \kappa G^{*\mathrm{T}}\hat{\rho} + o_p(1)$, and thus the above equality becomes

$$0 = \begin{pmatrix} \frac{1}{\sqrt{n}}\sum_{i=1}^{n} s_i(\beta_0) \\ \frac{\sqrt{n^*}}{\sqrt{n}}\frac{1}{\sqrt{n^*}}\sum_{j=1}^{n^*} u_j^*(\beta_0) \end{pmatrix} + \begin{pmatrix} -S, & \kappa G^{*\mathrm{T}} \\ \kappa G^*, & \kappa\Omega^* \end{pmatrix} \sqrt{n} \begin{pmatrix} \hat{\beta}_{\mathrm{EL2}} - \beta_0 \\ \hat{\rho} \end{pmatrix} + o_p(1),$$

which has exactly the same structure as (5.1). Therefore, $\sqrt{n}(\hat{\beta}_{\mathrm{EL2}} - \beta_0)$ and $\sqrt{n}(\hat{\beta}_{\mathrm{EL1}} - \beta_0)$ have the same asymptotic distribution.

Expressions for $M_\beta(\beta)$ and $M_{\beta\beta}(\beta)$

Bearing in mind the implicit dependence of $\hat{\lambda}(\beta)$ on β, some routine but tedious calculation leads to

$$M_\beta(\beta) = \sum_{i=1}^{n} s_i(\beta) + \sum_{j=1}^{n^*} \frac{u_{\beta j}^*(\beta)^{\mathrm{T}}}{1 - \hat{\lambda}(\beta)^{\mathrm{T}} u_j^*(\beta)} \hat{\lambda}(\beta),$$

$$\begin{aligned} M_{\beta\beta}(\beta) \;=\; & \sum_{i=1}^{n} s_{\beta i}(\beta) + \sum_{j=1}^{n^*} \frac{u_{\beta j}^*(\beta)^{\mathrm{T}} \hat{\lambda}(\beta)\hat{\lambda}(\beta)^{\mathrm{T}} u_{\beta j}^*(\beta)}{\{1 - \hat{\lambda}(\beta)^{\mathrm{T}} u_j^*(\beta)\}^2} + \sum_{j=1}^{n^*} \frac{\sum_{k=1}^{m} u_{\beta\beta j}^{*[k]}(\beta)\hat{\lambda}^{[k]}(\beta)}{1 - \hat{\lambda}(\beta)^{\mathrm{T}} u_j^*(\beta)} \\ & - \left(\sum_{j=1}^{n^*} \frac{u_{\beta j}^*(\beta)^{\mathrm{T}} \hat{\lambda}(\beta) u_j^*(\beta)^{\mathrm{T}}}{\{1 - \hat{\lambda}(\beta)^{\mathrm{T}} u_j^*(\beta)\}^2} + \sum_{j=1}^{n^*} \frac{u_{\beta j}^*(\beta)^{\mathrm{T}}}{1 - \hat{\lambda}(\beta)^{\mathrm{T}} u_j^*(\beta)} \right) \left(\sum_{j=1}^{n^*} \frac{u_j^*(\beta) u_j^*(\beta)^{\mathrm{T}}}{\{1 - \hat{\lambda}(\beta)^{\mathrm{T}} u_j^*(\beta)\}^2} \right)^{-1} \\ & \times \left(\sum_{j=1}^{n^*} \frac{u_j^*(\beta)\hat{\lambda}(\beta)^{\mathrm{T}} u_{\beta j}^*(\beta)}{\{1 - \hat{\lambda}(\beta)^{\mathrm{T}} u_j^*(\beta)\}^2} + \sum_{j=1}^{n^*} \frac{u_{\beta j}^*(\beta)}{1 - \hat{\lambda}(\beta)^{\mathrm{T}} u_j^*(\beta)} \right), \end{aligned}$$

where $u^{*[k]}(\beta)$ and $\hat{\lambda}^{[k]}(\beta)$ are the k-th component of $u^*(\beta)$ and $\hat{\lambda}(\beta)$, respectively.

When $\hat{\lambda}(\beta)$ is close to 0, as is the case for our implementation in Section 3.2 because $\hat{\lambda} \xrightarrow{p} 0$ and $\hat{\lambda}^{(0)} = 0$, we have the approximation

$$M_{\beta\beta}(\beta) \approx \sum_{i=1}^{n} s_{\beta i}(\beta) - \left(\sum_{j=1}^{n^*} \frac{u^*_{\beta j}(\beta)^{\mathrm{T}}}{1 - \hat{\lambda}(\beta)^{\mathrm{T}} u^*_j(\beta)} \right) \left(\sum_{j=1}^{n^*} \frac{u^*_j(\beta) u^*_j(\beta)^{\mathrm{T}}}{\{1 - \hat{\lambda}(\beta)^{\mathrm{T}} u^*_j(\beta)\}^2} \right)^{-1} \left(\sum_{j=1}^{n^*} \frac{u^*_{\beta j}(\beta)}{1 - \hat{\lambda}(\beta)^{\mathrm{T}} u^*_j(\beta)} \right),$$

which is used in our implementation.

Proof of (3.11)

Similar to the derivation of (5.1), we have

$$\sqrt{n} \begin{pmatrix} \hat{\beta}_{\mathrm{EL1}} - \beta_0 \\ \hat{\lambda} \end{pmatrix} = - \begin{pmatrix} -S, & \kappa G^{*\mathrm{T}} \\ \kappa G^*, & \kappa \Omega^* \end{pmatrix}^{-1} \begin{pmatrix} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} s_i(\beta_0) \\ \frac{\sqrt{N^*}}{\sqrt{n}} \frac{1}{\sqrt{N^*}} \sum_{j=1}^{n^*} u^*_j(\beta_0, \hat{\theta}) \end{pmatrix} + o_p(1). \tag{5.2}$$

Then the mean-value theorem and law of large numbers lead to

$$\frac{1}{\sqrt{N^*}} \sum_{j=1}^{n^*} u^*_j(\beta_0, \hat{\theta}) = \frac{1}{\sqrt{N^*}} \sum_{j=1}^{n^*} u^*_j(\beta_0, \theta^*) + \kappa^* Q^* \sqrt{N^*}(\hat{\theta} - \theta^*) + o_p(1), \tag{5.3}$$

These facts lead to (3.11).

Proof of (3.13)

Let $\{(Y^*_k, X^{*\mathrm{T}}_k)^{\mathrm{T}} : k = 1, \ldots, N^*\}$ denote the auxiliary data. From (5.3) we have

$$\frac{1}{\sqrt{N^*}} \sum_{j=1}^{n^*} u^*_j(\beta_0, \hat{\theta}) = \frac{1}{\sqrt{N^*}} \sum_{k=1}^{N^*} \left\{ R_k u^*_k(\beta_0, \theta^*) - \kappa^* Q^* H^{*-1} h(Y^*_k, X^*_k; \theta^*) \right\} + o_p(1), \tag{5.4}$$

where R_k is the indicator of whether the k-th subject in the sample $\{k : k = 1, \ldots, N^*\}$ is also in the supplementary sample $\{j : j = 1, \ldots, n^*\}$, and $H^* = E^*\{\partial h(Y, X; \theta^*)/\partial\theta\}$. Without loss of generality, we assume that the underlying mechanism that generates R_k is sampling n^* subjects from a finite population of N^* subjects without replacement. Let $W^* = Ru^*(\beta_0, \theta^*) - \kappa^* Q^* H^{*-1} h(Y^*, X^*; \theta^*)$. It is easy to verify that W^* has mean zero and that the covariance of $W^*_{k_1}$ and $W^*_{k_2}$ is zero when $k_1 \neq k_2$. In addition, using the fact that $H^* = Q^*$, some calculation shows that the variance of W^* is equal to $(\kappa^* - 2\kappa^{*2})\Omega^* + \kappa^{*2} Q^* V_{\theta^*} Q^{*\mathrm{T}}$. Therefore, from (5.2) and (5.4), (3.13) follows from the Central Limit Theorem for dependent random variables (e.g. Billingsley 1995).

References

Billingsley, P. (1995). *Probability and Measure 3rd Edition.* Wiley-Interscience.

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

Chatterjee, N., Chen, Y. H., Maas, P., and Carroll, R. J. (2016). Con-

strained maximum likelihood estimation for model calibration using summary-level information from external big data sources. *Journal of the American Statistical Association*, 111:107–117.

Chaudhuri, S., Handcock, M. S., and Rendall, M. S. (2008). Generalized linear models incorporating population level information: An empirical-likelihood-based approach. *Journal of the Royal Statistical Society Series B*, 70:311–328.

Chen, J. and Qin, J. (1993). Empirical likelihood estimation for finite populations and the effective usage of auxiliary information. *Biometrika*, 80:107–116.

Chen, J., Sitter, R. R., and Wu, C. (2002). Using empirical likelihood methods to obtain range restricted weights in regression estimators for surveys. *Biometrika*, 89:230–237.

Chen, S. and Kim, J. K. (2014). Population empirical likelihood for non-parametric inference in survey sampling. *Statistica Sinica*, 24:335–355.

Chen, S. X., Leung, D. H. Y., and Qin, J. (2003). Information recovery in a study with surrogate endpoints. *Journal of the American Statistical Association*, 98:1052–1062.

REFERENCES

Chen, Y. H. and Chen, H. (2000). A unified approach to regression analysis under double sampling design. *Journal of the Royal Statistical Society, Series B*, 62:449–460.

Deville, J. and Särndal, C. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87:376–382.

Han, P. (2014). Multiply robust estimation in regression analysis with missing data. *Journal of the American Statistical Association*, 109:1159–1173.

Han, P. and Lawless, J. F. (2016). Discussion of "constrained maximum likelihood estimation for model calibration using summary-level information from external big data sources". *Journal of the American Statistical Association*, 111:118–121.

Han, P. and Wang, L. (2013). Estimation with missing data: beyond double robustness. *Biometrika*, 100:417–430.

Handcock, M. S., Huovilainen, S. M., and Rendall, M. S. (2000). Combining registration-system and survey data to estimate birth probabilities. *Demography*, 37:187–192.

REFERENCES

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

Huang, C.-Y., Qin, J., and Tsai, H.-T. (2016). Efficient estimation of the cox model with auxiliary subgroup survival information. *Journal of the American Statistical Association*, 111:787–799.

Imbens, G. W. (2002). Generalized method of moments and empirical likelihood. *Journal of Business and Economic Statistics*, 20:493–506.

Imbens, G. W. and Lancaster, T. (1994). Combining micro and macro data in microeconometric models. *Review of Economic Studies*, 61:655–680.

Keiding, N. and Louis, T. A. (2016). Perils and potentials of self-selected entry to epidemiological studies and surveys. *Journal of the Royal Statistical Society Series A*, 179:319–376.

Kim, J. K. and Rao, J. N. K. (2012). Combining data from two independent surveys: a model-assisted approach. *Biometrika*, 99:85–100.

Kitamura, Y. (2007). *Advances in Economics and Econometrics: Theory and Applications, Ninth World Congress, Vol. 3.*, chapter Empirical Likelihood Methods in Econometrics: Theory and Practice, pages 174–237. Cambridge University Press.

Lawless, J. F. and Kalbfleisch, J. D. (2011). Discussion of "connections between survey calibration estimators and semiparametric models for incomplete data". *International Statistical Review*, 79:225–228.

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

Lumley, T., Shaw, P. A., and Dai, J. Y. (2011). Connections between survey calibration estimators and semiparametric models for incomplete data. *International Statistical Review*, 79:200–220.

Newey, W. K. (1985). Generalized method of moments specification testing. *Journal of Econometrics*, 29:229–256.

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

Owen, A. (2001). *Empirical Likelihood*. Chapman & Hall/CRC Press, New York.

Qin, J. (2000). Combining parametric and empirical likelihoods. *Biometrika*, 87:484–490.

REFERENCES

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

Qin, J., Zhang, H., Li, P., Albanes, D., and Yu, K. (2015). Using covariate-specific disease prevalence information to increase the power of case-control studies. *Biometrika*, 102:169–180.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press.

Department of Statistics and Actuarial Science

University of Waterloo

Waterloo, ON

Canada N2L 3G1

E-mail: peisonghan@uwaterloo.ca

Department of Statistics and Actuarial Science

University of Waterloo

Waterloo, ON

Canada N2L 3G1

E-mail: jlawless@uwaterloo.ca

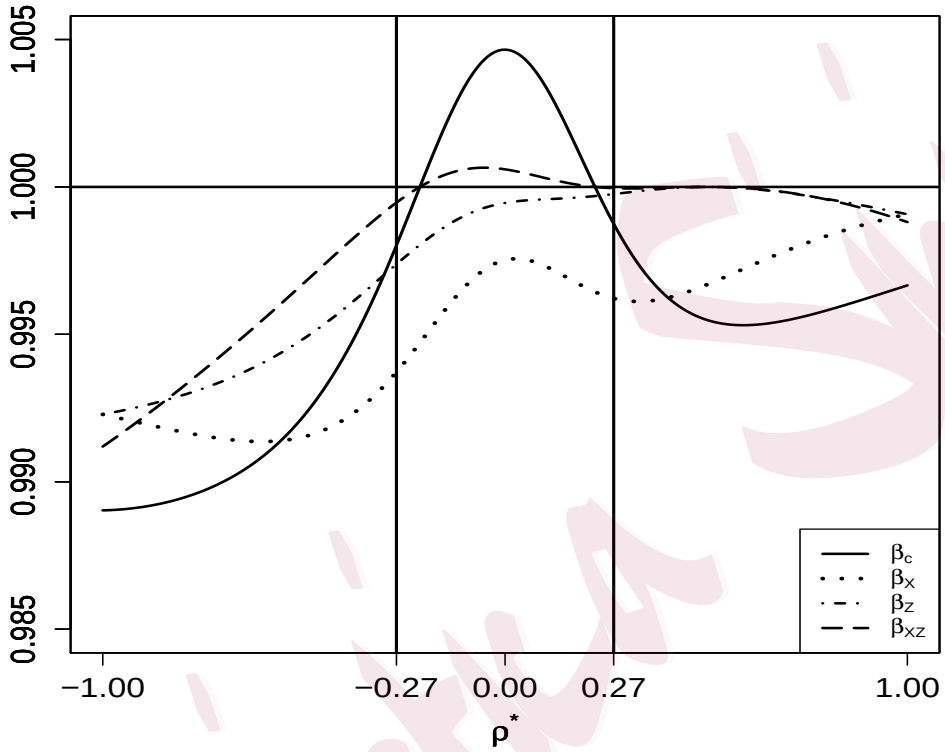

Figure 1: Plot of the ratio of the asymptotic standard deviation of $\sqrt{n}(\hat{\beta}_{\mathrm{EL1}} - \beta_0)$ using $\hat{\theta}$ versus that using θ^* when the supplementary sample is a subset of the auxiliary data set, taking $\beta_0^{\mathrm{T}} = (1, 1, 1, 1)$, $\rho = 0.5$, $\kappa = 1/3$ and $\kappa^* = 1/5$.

Table 1: Simulation results for logistic regression models based on 1000 replications. All numbers other than the percentages have been multiplied by 1000. Scenarios 1-3 correspond to no uncertainty in the auxiliary summary information, uncertainty in the auxiliary summary information and the supplementary sample is independent of the auxiliary data set, and uncertainty in the auxiliary summary information and the supplementary sample is a subset of the auxiliary data set, respectively. For all scenarios, $n = 300$ and $n^* = 100$. For scenario 2 and 3, $N^* = 500$.

	Scenario 1				Scenario 2				Scenario 3			
	β_c	β_X	β_Z	β_{XZ}	β_c	β_X	β_Z	β_{XZ}	β_c	β_X	β_Z	β_{XZ}
current study sample only												
bias	11	-14	-25	30	7	-9	-24	32	8	-12	-23	31
SE-EMP	139	168	167	173	144	169	179	166	139	166	181	168
empirical likelihood 1												
bias	-13	-9	-24	35	13	-13	-22	30	9	-12	-23	31
SE-EMP	66	86	159	163	106	127	172	169	101	120	172	167
SE-NAIVE	65	87	159	160	65	86	158	159	65	86	159	159
CP-95%	93.7	95.3	95.4	95.5	76.1	80.3	94.1	93.6	79.3	83.1	92.8	93.7
SE-EST	–	–	–	–	105	124	162	164	101	118	161	163
CP-95%-ADJ	–	–	–	–	94.6	95.1	94.3	94.3	94.6	95.7	93.3	94.2
empirical likelihood 2												
bias	-13	-9	-25	35	13	-12	-23	30	9	-13	-23	32
SE-EMP	67	86	159	164	106	127	172	169	101	120	172	167
SE-NAIVE	65	87	159	160	65	86	159	159	65	86	159	159
CP-95%	93.8	95.5	95.4	95.5	75.9	80.7	93.9	93.7	79.4	83.2	92.5	93.3
SE-EST	–	–	–	–	105	124	162	164	101	118	161	163
CP-95%-ADJ	–	–	–	–	94.5	94.7	94.1	94.2	94.6	95.6	93.1	93.9

SE-EMP: empirical standard error; SE-NAIVE: mean of estimated standard error based on asymptotic variance without accounting for the uncertainty in $\hat{\theta}$; CP-95%: coverage probability of the 95% confidence interval based on asymptotic distribution without accounting for the uncertainty in $\hat{\theta}$; SE-EST: mean of estimated standard error based on asymptotic variance adjusting for the uncertainty in $\hat{\theta}$; CP-95%-ADJ: coverage probability of the 95% confidence interval based on asymptotic distribution adjusting for the uncertainty in $\hat{\theta}$