Statistica Sinica **28** (2018), 2027-2048
doi:https://doi.org/10.5705/ss.202016.0291

EMPIRICAL LIKELIHOOD METHODS FOR COMPLEX SURVEYS WITH DATA MISSING-BY-DESIGN

Min Chen[1], Mary E. Thompson[2] and Changbao Wu[2]

[1]*Bank of Nova Scotia and* [2]*University of Waterloo*

Abstract: We consider nonrandomized pretest-posttest designs with complex survey data for observational studies. We show that two-sample pseudo empirical likelihood methods provide efficient inferences on the treatment effect, with a missing-by-design feature used for forming the two samples and the baseline information incorporated through suitable constraints. The proposed maximum pseudo empirical likelihood estimators of the treatment effect are consistent and pseudo empirical likelihood ratio confidence intervals are constructed through bootstrap calibration methods. The proposed methods require estimation of propensity scores which depend on the underlying missing-by-design mechanism. A simulation study was conducted to examine finite sample performances of the proposed methods under different scenarios of nonignorable and ignorable missing patterns. An application to the International Tobacco Control Policy Evaluation Project Four Country Surveys is also presented to demonstrate the use of the proposed methods for examining the mode effect in survey data collection.

Key words and phrases: Auxiliary information, complex survey, confidence interval, empirical likelihood, missing-by-design, pretest-posttest study, propensity scores, treatment and control.

1. Introduction

Two-sample problems are commonly encountered in many fields of scientific investigation, including classical designed experiments, modern case-control studies, and many observational studies in social and medical sciences. There exist standard statistical tools for handling problems with two independent samples, especially with parametric approaches where likelihood-based methods are readily available. The nonparametric empirical likelihood methods, first proposed by Owen (1988) for a single sample, have been extended to cover two independent samples. Wu and Yan (2012) contains detailed discussions and related references for two-sample empirical likelihood.

Pretest-posttest studies are a special type of two-sample problem. They constitute a very broad topic relevant to many subject areas; see, for instance,

Brogan and Kutner (1980) for a detailed discussion. Randomized pretest-posttest designs are often used in medical studies to examine a treatment effect where baseline (pretest) information is collected for all selected units before they are randomly assigned to treatment or control groups. Nonrandomized designs are common in observational studies to investigate the effect of a treatment or an intervention. There are two unique features for pretest-posttest studies: the baseline information collected for all units in the initial combined sample; and the missing-by-design structure for the two samples where each unit is assigned to either the treatment group or the control group but not both. How to best use the pretest information and how to incorporate the unique missing-by-design feature are the two major statistical research problems for pretest-posttest studies.

There have been several promising developments in empirical likelihood methods for pretest-posttest studies under randomized designs in recent literature. Huang, Qin and Follmann (2008) proposed an empirical likelihood method for estimating the treatment effect $\theta = \mu_1 - \mu_0$, where μ_1 and μ_0 are respectively the mean response under the treatment and the control. Their approach focused on estimating μ_1 and μ_0 separately while incorporating the pretest information through additional constraints for the maximum empirical likelihood estimators. Chen, Wu and Thompson (2015) proposed an imputation-based empirical likelihood approach to effectively combining the baseline information and the missing-by-design feature of pretest-posttest studies. Empirical likelihood ratio confidence intervals for θ can be constructed directly without involving μ_1 or μ_0. The authors showed that the empirical likelihood ratio test of the treatment effect θ is more powerful than existing alternative methods. Another important problem is to test the equality of the two distribution functions, H_0: $F_1(t) = F_0(t)$, where $F_1(t)$ and $F_0(t)$ are the distribution function of the response variable under the treatment and the control, respectively. This is equivalent to testing H_0: $S_1(t) = S_0(t)$, where $S_1(t)$ and $S_0(t)$ are the corresponding survival functions, which is often of primary interest in medical research. Chen, Wu and Thompson (2016) developed different versions of the Mann-Whitney test using empirical likelihood methods.

Pretest-posttest designs are also frequently used in observational studies with complex surveys. For example, a youth smoking intervention program may have the following design. First, a sample of grade six students is selected from a particular student population using a probability sampling method. Certain baseline information, such as gender, age, family background, and other social-economic indicators, is collected for all selected students. Then each selected student is

presented with the opportunity to join an intervention program. The treatment group consists of all participating students in the program and the control group includes those who choose not to participate. A key feature for such studies is that randomization is not used for assigning units to treatment or control. The main objective of the study is to examine the effectiveness of the intervention. In Section 5, we present an example from the ITC Four Country Survey on mode effect in data collection, where each respondent is given options to complete the same set of survey questionnaires through either a self-administered web survey or a telephone interview, but not both. The objective of the example is to demonstrate the use of the proposed methods for testing whether there is a non-negligible difference in the distribution of responses to a specific question between the two modes (web versus telephone) in data collection.

This paper develops empirical likelihood methods for complex surveys with nonrandomized pretest-posttest designs. We consider observational studies for which baseline information is gathered for all units in the initial survey sample but there is a self-selection for each unit on whether to be in the treatment group or the control group. Let $\mathbf{S}$ be the set of n units selected for the initial sample. Let $\pi_i = P(i \in \mathbf{S})$ be the first order inclusion probabilities for the survey design. Let $R_i = 1$ if unit i chooses to be in the treatment group and $R_i = 0$ if unit i is in the control group. We have $\mathbf{S} = \mathbf{S}_1 \cup \mathbf{S}_0$, where $\mathbf{S}_1 = \{i | i \in \mathbf{S} \text{ and } R_i = 1\}$ is the set of units in the treatment group and $\mathbf{S}_0 = \{i | i \in \mathbf{S} \text{ and } R_i = 0\}$ is the set of units in the control group. Let $n_1 = |\mathbf{S}_1|$ and $n_0 = |\mathbf{S}_0|$ be the sizes of the two groups, with $n = n_1 + n_0$. Let $\boldsymbol{x}_i$ be the value of the vector of auxiliary variables $\boldsymbol{x}$ for unit i; let y_{1i} be the value of the potential response variable y_1 if unit i is exposed to the treatment and y_{0i} be the value of the potential response variable y_0 if unit i is exposed to the control. The full observations on $\boldsymbol{x}$ and the missing-by-design feature of the responses can be represented by the following table:

i	1	2	$\cdots$	n_1	n_1+1	n_1+2	$\cdots$	n
$\boldsymbol{x}$	$\boldsymbol{x}_1$	$\boldsymbol{x}_2$	$\cdots$	$\boldsymbol{x}_{n_1}$	$\boldsymbol{x}_{n_1+1}$	$\boldsymbol{x}_{n_1+2}$	$\cdots$	$\boldsymbol{x}_n$
y_1	y_{11}	y_{12}	$\cdots$	y_{1n_1}	$*$	$*$	$\cdots$	$*$
y_0	$*$	$*$	$\cdots$	$*$	$y_{0(n_1+1)}$	$y_{0(n_1+2)}$	$\cdots$	y_{0n}

In the absence of randomization for the treatment assignments, the two samples $\mathbf{S}_1$ and $\mathbf{S}_0$ are not representative for the finite population under the original survey design. For instance, under the extreme scenario where all "male units" choose to be in the treatment group and all "female units" choose to belong

to the control group, the two samples would each represent a subpopulation of males or females. Let

$$r_i = P(R_i = 1 | i \in \mathbf{S}, y_{1i}, y_{0i}, \boldsymbol{x}_i) = P(i \in \mathbf{S}_1 | i \in \mathbf{S}, y_{1i}, y_{0i}, \boldsymbol{x}_i)\,.$$

It is a commonly acceptable assumption that $P(R_i = 1 | i \in \mathbf{S}, y_{1i}, y_{0i}, \boldsymbol{x}_i) = P(R_i = 1 | y_{1i}, y_{0i}, \boldsymbol{x}_i)$: the exposure of a unit to treatment or control is not confounded with the inclusion of the unit in the initial sample given all the characteristics of the unit to be measured by the survey. The missing-by-design mechanism (i.e., treatment assignment) is called ignorable (Rosenbaum and Rubin (1983)) if

$$a_i = P(R_i = 1 | y_{1i}, y_{0i}, \boldsymbol{x}_i) = P(R_i = 1 | \boldsymbol{x}_i)\,.$$

Our discussions in Sections 2 and 3 will adopt the assumption that the treatment assignment is ignorable. In the simulation studies presented in Section 4, we investigate practical scenarios where the missing-by-design mechanism is not ignorable.

2. The Propensity Score Adjusted Two-Sample Empirical Likelihood

Let $\boldsymbol{Y}_1 = (y_{11}, \ldots, y_{1N})'$, $\boldsymbol{Y}_0 = (y_{01}, \ldots, y_{0N})'$, $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$, where N is the population size. The parameter of interest is $\theta_N = \mu_1 - \mu_0$ where $\mu_1 = N^{-1}\sum_{i=1}^N y_{1i}$ and $\mu_0 = N^{-1}\sum_{i=1}^N y_{0i}$ are the population means of the potential responses under the treatment and the control. Let

$$\gamma_{1i} = P(i \in \mathbf{S}_1 | \boldsymbol{Y}_1, \boldsymbol{Y}_0, \boldsymbol{X}) \quad \text{and} \quad \gamma_{0i} = P(i \in \mathbf{S}_0 | \boldsymbol{Y}_1, \boldsymbol{Y}_0, \boldsymbol{X})$$

be, respectively, the inclusion probabilities for the treatment and control groups with the given finite population. We assume that

$$\begin{aligned} &P(i \in \mathbf{S}_1 | i \in \mathbf{S}, \boldsymbol{Y}_1, \boldsymbol{Y}_0, \boldsymbol{X}) \\ &= P(i \in \mathbf{S}_1 | i \in \mathbf{S}, y_{1i}, y_{0i}, \boldsymbol{x}_i) \\ &= P(R_i = 1 | y_{1i}, y_{0i}, \boldsymbol{x}_i)\,. \end{aligned}$$

Since $i \in \mathbf{S}_1$ implies $i \in \mathbf{S}$, we have

$$\begin{aligned} \gamma_{1i} &= P(i \in \mathbf{S}_1 | \boldsymbol{Y}_1, \boldsymbol{Y}_0, \boldsymbol{X}) \\ &= P(i \in \mathbf{S}_1, i \in \mathbf{S} | \boldsymbol{Y}_1, \boldsymbol{Y}_0, \boldsymbol{X}) \\ &= P(i \in \mathbf{S} | \boldsymbol{Y}_1, \boldsymbol{Y}_0, \boldsymbol{X})\, P(i \in \mathbf{S}_1 | i \in \mathbf{S}, \boldsymbol{Y}_1, \boldsymbol{Y}_0, \boldsymbol{X}) \\ &= \pi_i a_i\,, \end{aligned}$$

where a_i is the propensity score that unit i chooses the treatment. Similarly, we have $\gamma_{0i} = \pi_i(1 - a_i)$.

We first consider the hypothetical scenario where the propensity scores a_i are known for every unit in the population and not dependent on $\mathbf{S}$. The self-selection for treatment assignment is equivalent to an added last stage of Poisson sampling and γ_{1i} and γ_{0i} are design-based inclusion probabilities for the samples $\mathbf{S}_1$ and $\mathbf{S}_0$, respectively. Let $w_{1i} = 1/\gamma_{1i}$ and $w_{0i} = 1/\gamma_{0i}$ be the survey weights. Let

$$\tilde{w}_{1i}(\mathbf{S}_1) = \frac{w_{1i}}{\sum_{j\in\mathbf{S}_1} w_{1j}} \qquad \text{and} \qquad \tilde{w}_{0i}(\mathbf{S}_0) = \frac{w_{0i}}{\sum_{j\in\mathbf{S}_0} w_{0j}}$$

be the normalized survey weights. Following the formulation used in Wu and Rao (2006), the joint pseudo empirical (log) likelihood function is given by

$$\ell(\boldsymbol{p},\boldsymbol{q}) = \frac{1}{2}\sum_{i\in\mathbf{S}_1} \tilde{w}_{1i}(\mathbf{S}_1)\,\log(p_i) + \frac{1}{2}\sum_{i\in\mathbf{S}_0} \tilde{w}_{0i}(\mathbf{S}_0)\,\log(q_i)\,,$$

where $\boldsymbol{p} = (p_1,\ldots,p_{n_1})$ and $\boldsymbol{q} = (q_1,\ldots,q_{n_0})$ are nonparametric discrete probability measures over $\mathbf{S}_1$ and $\mathbf{S}_0$, respectively. The two factors $1/2$ appearing in $\ell(\boldsymbol{p},\boldsymbol{q})$ are for computational purposes and have no impact on inferences. The standard normalization constraints are given by

$$\sum_{i\in\mathbf{S}_1} p_i = 1\,, \qquad \sum_{i\in\mathbf{S}_0} q_i = 1\,. \tag{2.1}$$

The constraint induced by the parameter $\theta_{_N} = \mu_1 - \mu_0$ is given by

$$\sum_{i\in\mathbf{S}_1} p_i\, y_{1i} - \sum_{i\in\mathbf{S}_0} q_i\, y_{0i} = \theta\,. \tag{2.2}$$

Maximizing $\ell(\boldsymbol{p},\boldsymbol{q})$ under the normalization constraints (2.1) gives $\hat{p}_i = \tilde{w}_{1i}(\mathbf{S}_1)$ and $\hat{q}_i = \tilde{w}_{0i}(\mathbf{S}_0)$. Let $\hat{p}_i(\theta)$ and $\hat{q}_i(\theta)$ be the maximizer of $\ell(\boldsymbol{p},\boldsymbol{q})$ under both the normalization constraints (2.1) and the parameter constraint (2.2) for a fixed θ. Let

$$\ell\big(\hat{\boldsymbol{p}}(\theta),\hat{\boldsymbol{q}}(\theta)\big) = \frac{1}{2}\sum_{i\in\mathbf{S}_1} \tilde{w}_{1i}(\mathbf{S}_1)\,\log\big(\hat{p}_i(\theta)\big) + \frac{1}{2}\sum_{i\in\mathbf{S}_0} \tilde{w}_{0i}(\mathbf{S}_0)\,\log\big(\hat{q}_i(\theta)\big)\,.$$

The maximum empirical likelihood estimator of $\theta_{_N}$, which maximizes $\ell\big(\hat{\boldsymbol{p}}(\theta),$ $\hat{\boldsymbol{q}}(\theta)\big)$ with respect to θ, is given by

$$\hat{\theta}_{_N} = \sum_{i\in\mathbf{S}_1} \hat{p}_i\, y_{1i} - \sum_{i\in\mathbf{S}_0} \hat{q}_i\, y_{0i}\,. \tag{2.3}$$

It can be shown that $\hat{\theta}_{_N}$ is a design consistent estimator of the true parameter $\theta_{_N} = \mu_1 - \mu_0$. The empirical (log) likelihood ratio function for $\theta_{_N}$ is given by

$$r(\theta) = \ell\big(\hat{\boldsymbol{p}}(\theta),\hat{\boldsymbol{q}}(\theta)\big) - \ell\big(\hat{\boldsymbol{p}},\hat{\boldsymbol{q}}\big)\,, \tag{2.4}$$

where $\hat{\boldsymbol{p}} = (\hat{p}_1,\ldots,\hat{p}_{n_1})$ and $\hat{\boldsymbol{q}} = (\hat{q}_1,\ldots,\hat{q}_{n_0})$. There exists a simple and efficient

algorithm for computing $r(\theta)$ for a given θ; see Section 3 of Wu and Yan (2012) for details. It also follows from Theorem 3.1 in Wu and Yan (2012) that $-2r(\theta_N)$ is asymptotically distributed as a scaled χ^2 with one degree of freedom, with the scaling constant involving the design-based variance of $\hat{\theta}_N$.

One of the major advantages of the empirical likelihood approach is the power of combining additional available information for inferences. The baseline information on $\boldsymbol{x}$, collected for all units in the initial sample $\mathbf{S} = \mathbf{S}_1 \cup \mathbf{S}_0$, with the missing-by-design feature for response variables can be incorporated through the additional constraints

$$\sum_{i \in \mathbf{S}_1} p_i \, \boldsymbol{x}_i = \sum_{i \in \mathbf{S}_0} q_i \, \boldsymbol{x}_i \, . \tag{2.5}$$

Under this setting, the $\hat{\boldsymbol{p}}$ and $\hat{\boldsymbol{q}}$ maximizes $\ell(\boldsymbol{p}, \boldsymbol{q})$ subject to (2.1) and (2.5); the $\hat{\boldsymbol{p}}(\theta)$ and $\hat{\boldsymbol{q}}(\theta)$ maximizes $\ell(\boldsymbol{p}, \boldsymbol{q})$ subject to (2.1), (2.2), and (2.5); the maximum empirical likelihood estimator $\hat{\theta}_N$ is still given by (2.3), and the empirical likelihood ratio function $r(\theta)$ is given by (2.4).

We now discuss practical scenarios where the propensity scores a_i are unknown. If the missing-by-design mechanism is ignorable, we can estimate a_i using available data $\{(R_i, \boldsymbol{x}_i), i \in \mathbf{S}\}$ under a suitable model. A popular choice for $R|\boldsymbol{x}$ is the logistic regression model

$$\log\Big(\frac{a_i}{1 - a_i}\Big) = \eta_0 + \boldsymbol{x}_i' \boldsymbol{\eta}_1 \, , \tag{2.6}$$

where $\boldsymbol{\eta} = (\eta_0, \boldsymbol{\eta}_1')'$ is the vector of model parameters. This leads to $a_i = a(\boldsymbol{x}_i, \boldsymbol{\eta}) = \exp(\eta_0 + \boldsymbol{x}_i' \boldsymbol{\eta}_1)/\{1 + \exp(\eta_0 + \boldsymbol{x}_i' \boldsymbol{\eta}_1)\}$. Let $\hat{\boldsymbol{\eta}}$ be the maximum likelihood estimator of $\boldsymbol{\eta}$ under model (2.6). The estimated propensity scores are given by $\hat{a}_i = a(\boldsymbol{x}_i, \hat{\boldsymbol{\eta}})$.

When the missing-by-design mechanism is nonignorable, estimation of propensity scores becomes difficult and valid inferences often rely on strong assumptions or availability of additional information. In this paper we focus on estimating propensity scores based on model (2.6) under the assumption of ignorability, and examine the consequences of analysis through simulation studies to be reported in Section 4 on different scenarios of ignorable and nonignorable missing-by-design mechanisms.

Recall that $\gamma_{1i} = \pi_i a_i$ and $\gamma_{0i} = \pi_i(1 - a_i)$. For simplicity of notation without causing any confusion, we now let $\gamma_{1i} = \pi_i \hat{a}_i$ and $\gamma_{0i} = \pi_i(1 - \hat{a}_i)$. The maximum empirical likelihood estimator $\hat{\theta}_N$ and the empirical likelihood ratio function $r(\theta)$ can be computed in the same way as in (2.3) and (2.4), respectively. Under the current setting, however, asymptotic properties of $\hat{\theta}_N$ and the asymptotic distri-

bution of $r(\theta)$ need to be considered under the joint randomization framework with both the probability sampling design for the initial sample selection and the assumed logistic regression model (2.6) for the estimation of the propensity scores. Purely design-based inferences are no longer feasible.

It is possible to go through the traditional route $E(\hat{\theta}_N) = E_p E_\xi(\hat{\theta}_N|\mathbf{S}_1, \mathbf{S}_0)$ and $V(\hat{\theta}_N) = E_p V_\xi(\hat{\theta}_N|\mathbf{S}_1, \mathbf{S}_0) + V_p E_\xi(\hat{\theta}_N|\mathbf{S}_1, \mathbf{S}_0)$, where $E_p(\cdot)$, $E_\xi(\cdot)$, $V_p(\cdot)$, and $V_\xi(\cdot)$ denote, respectively, the expectation and the variance under the probability sampling design (p) and the assumed model (ξ) on the propensity scores. The consistency of the point estimator $\hat{\theta}_N$ under the joint randomization can be easily established since $\hat{\theta}_N$ is a smooth function of $\hat{\boldsymbol{\eta}}$ and we have $\hat{\boldsymbol{\eta}} = \boldsymbol{\eta} + O_p(n^{-1/2})$ under the assumed model (2.6).

For the empirical likelihood ratio statistic $r(\theta)$, a practically more convenient approach is to use a bootstrap calibration method. With complete observations, Wu and Rao (2010) developed bootstrap procedures for the one-sample pseudo empirical likelihood method for certain survey designs. Noting that $r(\theta_N) = h(\theta_N, \hat{\boldsymbol{\eta}})$ can be viewed as a smooth function of $\hat{\boldsymbol{\eta}}$, the current two-sample problem with the added complication from the model-based estimation of $\hat{\boldsymbol{\eta}}$ can be handled through an embedded procedure in the bootstrap method described below.

Consider single-stage unequal probability sampling with small sampling fractions. The conventional with-replacement bootstrap method provides valid inferences under the design-based framework for such cases. Let $\hat{\theta}_N$ be the maximum empirical likelihood estimator of θ_N computed with the estimated propensity scores $\hat{a}_i$. We assume that constraints (2.5) are included when computing $r(\theta)$ for the given θ. The asymptotic distribution of $r(\theta_N)$ can be approximated through the following bootstrap procedures.

1. Let $\mathbf{S}^*$ be the set of n units, including duplicated ones, selected from the initial sample $\mathbf{S}$ by simple random sampling with replacement. Let $\mathbf{S}_1^* = \{i|i \in \mathbf{S}^* \text{ and } R_i = 1\}$ and $\mathbf{S}_0^* = \{i|i \in \mathbf{S}^* \text{ and } R_i = 0\}$. Both $\mathbf{S}_1^*$ and $\mathbf{S}_0^*$ might include duplicated units from $\mathbf{S}$ and the total number of units in $\mathbf{S}_1^*$ (or $\mathbf{S}_0^*$) is not necessarily n_1 (or n_0).

2. Fit model (2.6) using data $\{(R_i, \boldsymbol{x}_i), i \in \mathbf{S}^*\}$; let $\hat{\boldsymbol{\eta}}^*$ be the estimate of $\boldsymbol{\eta}$. Compute the estimated propensity scores $\hat{a}_i^* = a(\boldsymbol{x}_i, \hat{\boldsymbol{\eta}}^*)$ for $i \in \mathbf{S}^*$. Compute the inclusion probabilities $\gamma_{1i}^* = \pi_i \hat{a}_i^*$ for $i \in \mathbf{S}_1^*$ and $\gamma_{0i}^* = \pi_i(1 - \hat{a}_i^*)$ for $i \in \mathbf{S}_0^*$. Let $w_{1i}^* = 1/\gamma_{1i}^*$ and $w_{0i}^* = 1/\gamma_{0i}^*$ and let

$$\tilde{w}^*_{1i}(\mathbf{S}^*_1) = \frac{w^*_{1i}}{\sum_{j\in\mathbf{S}^*_1} w^*_{1j}}\,, \quad i \in \mathbf{S}^*_1 \qquad \text{and} \qquad \tilde{w}^*_{0i}(\mathbf{S}^*_0) = \frac{w^*_{0i}}{\sum_{j\in\mathbf{S}^*_0} w^*_{0j}}\,, \quad i \in \mathbf{S}^*_0\,.$$

3. Compute the bootstrap version of $r(\theta)$ at $\theta = \hat{\theta}_{_N}$ using the bootstrap version of the empirical likelihood function

$$\ell(\boldsymbol{p}, \boldsymbol{q}) = \frac{1}{2}\sum_{i\in\mathbf{S}^*_1} \tilde{w}^*_{1i}(\mathbf{S}^*_1)\,\log(p_i) + \frac{1}{2}\sum_{i\in\mathbf{S}^*_0} \tilde{w}^*_{0i}(\mathbf{S}^*_0)\,\log(q_i)\,,$$

where $\boldsymbol{p}$ and $\boldsymbol{q}$ are, respectively, the discrete probability measures over $\mathbf{S}^*_1$ and $\mathbf{S}^*_0$, with the normalization constraints (2.1), the parameter constraint (2.2), and the constraint (2.5) for incorporating the baseline information replaced, respectively, by

$$\begin{aligned}
&\sum_{i\in\mathbf{S}^*_1} p_i = 1\,, \qquad \sum_{i\in\mathbf{S}^*_0} q_i = 1\,,\\
&\sum_{i\in\mathbf{S}^*_1} p_i\, y_{1i} - \sum_{i\in\mathbf{S}^*_0} q_i\, y_{0i} = \hat{\theta}_{_N}\,,\\
&\sum_{i\in\mathbf{S}^*_1} p_i\, \boldsymbol{x}_i \quad = \quad \sum_{i\in\mathbf{S}^*_0} q_i\, \boldsymbol{x}_i\,.
\end{aligned}$$

4. For a predetermined B, repeat Steps 1-3 independently to obtain B bootstrap copies of $r(\theta)$, all at $\theta = \hat{\theta}_{_N}$:

$$r^{[1]}(\theta)\,,\ r^{[2]}(\theta)\,,\ldots,\ r^{[B]}(\theta)\,.$$

The typical choice of the value for B is 1,000. Let $\alpha \in (0,1)$ and b_α be the lower 100αth quantile from the sequence given in Step 4. The bootstrap calibrated $(1-\alpha)$-level empirical likelihood ratio confidence interval for $\theta_{_N}$ is constructed as

$$\mathcal{C} = \left\{\theta | r(\theta) > b_\alpha\right\}. \tag{2.7}$$

Finding the lower and upper boundaries of the confidence interval $\mathcal{C}$ requires profiling. Detailed algorithms and computational code for finding empirical likelihood ratio confidence intervals can be found in Chen, Sitter and Wu (2002) and Wu (2004, 2005). Finite sample performances of the interval $\mathcal{C}$ were investigated through simulation studies to be presented in Section 4.

3. Two-Sample Empirical Likelihood with Post-Stratification by Propensity Scores

Pretest-posttest designs in the absence of randomization on treatment assignments have brought challenges for data analysis. There exists a rich literature on

analysis of observational data with such designs under an assumed model for the potential response variables Y_1 for treatment and Y_0 for control. The population level treatment effect is defined as $\theta = E(Y_1) - E(Y_0)$ where $E(\cdot)$ refers to the assumed model. Samples are obtained for the treatment ($R = 1$) and the control ($R = 0$). Without randomization of treatment assignments, the samples are not representative of their respective populations since $E(Y_1|R = 1) \neq E(Y_1)$ and $E(Y_0|R = 0) \neq E(Y_0)$. When data on covariates $\boldsymbol{X}$ are collected for all sampled units, conditional models for $Y_1|\boldsymbol{X}$ and $Y_0|\boldsymbol{X}$ can be used for inference. Estimation of treatment effect through conditional models often requires that the characteristics of $\boldsymbol{X}$ for the treatment group and the control group follow the same distributions, the covariates need to be balanced between the two groups. Balancing covariates through propensity scores has been studied by several authors, including Austin (2008, 2009), Imai and Ratkovic (2014) and Li, Morgan and Zaslavsky (2015). Rosenbaum and Rubin (1984) proposed using subclassification based on propensity scores, and Zanutto, Lu and Hornik (2005) provided an application of the subclassification method. The idea was also investigated by Lunceford and Davidian (2004) and Miratrix, Sekhon and Yu (2013) under slightly different terms such as "stratification" or "post-stratification". If we define the treatment effect as the finite population mean of y_1 minus the finite population mean of y_0 and estimate it using a design-based approach, we do not need to bring in the covariates or the models, and balancing covariates does not appear to be crucial. One nice thing about balancing covariates is that there is a kind of matching of treatment and control subjects on $\boldsymbol{x}$, and if under the model the mean functions of y_0 and y_1 differ by a constant, the estimation should be very efficient because of the matching.

It turns out that post-stratification by propensity scores provides an attractive alternative formulation for the joint pseudo empirical (log) likelihood function $\ell(\boldsymbol{p}, \boldsymbol{q})$. The key observation is that if the propensity scores $a_i = c$ are a constant for all units, the normalized survey weights $\tilde{w}_{1i}(\mathbf{S}_1) = w_{1i}/\sum_{j\in\mathbf{S}_1} w_{1j}$ and $\tilde{w}_{0i}(\mathbf{S}_0) = w_{0i}/\sum_{j\in\mathbf{S}_0} w_{0j}$, where $w_{1i} = 1/(\pi_i a_i)$ and $w_{0i} = 1/\{\pi_i(1-a_i)\}$, do not involve a_i and reduce to normalized weights based on the basic design weights $d_i = 1/\pi_i$ for the initial sample $\mathbf{S}$. Post-stratification by propensity scores breaks the initial sample into subsamples such that the propensity scores within each subsample have similar values. From the efficiency point of view, this approach may not do quite as well as adjustment by propensity scores, which is also what we observed from simulation studies. However, It does balance the covariates approximately within the post-strata, and it means that we have a bit of the

benefit of the conditional modeling even if that is not done.

Let $\hat{a}_i$, $i \in \mathbf{S}$ be the estimated propensity scores based on model (2.6) and order the n units in $\mathbf{S}$ according to the size of $\hat{a}_i$:

$$\hat{a}_{(1)} \ \le \ \hat{a}_{(2)} \ \le \ \cdots \ \le \ \hat{a}_{(n)} \, .$$

The initial sample $\mathbf{S}$ can be divided into K subsamples of equal or similar sample sizes based on suitable cut-offs of the estimated propensity scores. Let

$$\mathbf{S} = \mathbf{Q}_1 \ \cup \ \mathbf{Q}_2 \ \cup \ \cdots \cup \mathbf{Q}_{\kappa}$$

be the resulting post-stratification of $\mathbf{S}$. Within each post-stratified sample $\mathbf{Q}_k$, $k = 1, \ldots, K$, values of the propensity scores are similar. Let $n = \sum_{k=1}^{K} m_k$ be the corresponding breakdown of the sample sizes.

It is theoretically helpful (but not practically critical) to assume that there exists a conceptual stratification at the population level with stratum population size N_k, $k = 1, \ldots, K$ such that $N = \sum_{k=1}^{K} N_k$. Let $W_k = N_k/N$ be the stratum weights, which can be estimated by $\hat{W}_k = \hat{N}_k/\hat{N}$, where $\hat{N} = \sum_{i \in \mathbf{S}} d_i$ and $\hat{N}_k = \sum_{i \in \mathbf{Q}_k} d_i$. It follows that $\hat{N} = \sum_{k=1}^{K} \hat{N}_k$ and $\sum_{i=1}^{K} \hat{W}_k = 1$. Let $\mathbf{Q}_k = \mathbf{S}_{1k} \cup \mathbf{S}_{0k}$, where

$$\mathbf{S}_{1k} = \left\{ i | i \in \mathbf{Q}_k \ \text{ and } \ R_i = 1 \right\}, \qquad \mathbf{S}_{0k} = \left\{ i | i \in \mathbf{Q}_k \ \text{ and } \ R_i = 0 \right\}.$$

Let n_{1k} and n_{0k} be the sample size of $\mathbf{S}_{1k}$ and $\mathbf{S}_{0k}$, respectively. We have $m_k = n_{1k} + n_{0k}$, $k = 1, \ldots, K$.

One of the practical issues for post-stratification by propensity scores is the choice of K. A larger K would result in post-stratified samples with more uniform values of propensity scores within each subsample. However, a finer stratification is associated with smaller sample sizes m_k for $\mathbf{Q}_k$ and much smaller sample sizes n_{1k} for $\mathbf{S}_{1k}$ (or n_{0k} for $\mathbf{S}_{0k}$) for the first and the last subsamples. Rosenbaum and Rubin (1984) suggested using $K = 5$. Our simulation results also suggest that $K = 5$ is a reasonable choice for moderately large sample sizes.

We define the joint pseudo empirical (log) likelihood function for the post-stratified samples as

$$\begin{aligned} &\ell(\boldsymbol{p}_1, \ldots, \boldsymbol{p}_{\kappa}, \boldsymbol{q}_1, \ldots, \boldsymbol{q}_{\kappa}) \\ &= \sum_{k=1}^{K} \hat{W}_k \sum_{i \in \mathbf{S}_{1k}} \tilde{d}_i(\mathbf{S}_{1k}) \log(p_{ik}) + \sum_{k=1}^{K} \hat{W}_k \sum_{i \in \mathbf{S}_{0k}} \tilde{d}_i(\mathbf{S}_{0k}) \log(q_{ik}) \, , \end{aligned}$$

where $\boldsymbol{p}_k = (p_{1k}, \ldots, p_{n_{1k}k})$ and $\boldsymbol{q}_k = (q_{1k}, \ldots, q_{n_{0k}k})$ are the discrete probability measures over $\mathbf{S}_{1k}$ and $\mathbf{S}_{0k}$, respectively, $\tilde{d}_i(\mathbf{S}_{1k}) = d_i / \sum_{j \in \mathbf{S}_{1k}} d_j$, and $\tilde{d}_i(\mathbf{S}_{0k}) = d_i / \sum_{j \in \mathbf{S}_{0k}} d_j$, $k = 1, \ldots, K$. The estimated propensity scores are no longer used

explicitly in defining the joint empirical likelihood function other than in forming the stratification. The set of normalization constraints is given by

$$\sum_{i\in\mathbf{S}_{1k}} p_{ik} = 1\,, \quad \sum_{i\in\mathbf{S}_{0k}} q_{ik} = 1\,, \quad k = 1,\ldots,K\,. \tag{3.1}$$

The constraint associated with the parameter $\theta_{_N}$ is given by

$$\sum_{k=1}^{K} \hat{W}_k \sum_{i\in\mathbf{S}_{1k}} p_{ik} y_{1i} - \sum_{k=1}^{K} \hat{W}_k \sum_{i\in\mathbf{S}_{0k}} q_{ik} y_{0i} = \theta\,. \tag{3.2}$$

The baseline information on $\boldsymbol{x}$ can be incorporated through the constraints

$$\sum_{k=1}^{K} \hat{W}_k \sum_{i\in\mathbf{S}_{1k}} p_{ik} \boldsymbol{x}_i = \sum_{k=1}^{K} \hat{W}_k \sum_{i\in\mathbf{S}_{0k}} q_{ik} \boldsymbol{x}_i\,. \tag{3.3}$$

Let $\hat{\boldsymbol{p}}_k$ and $\hat{\boldsymbol{q}}_k$, $k = 1,\ldots,K$, be the maximizer of $\ell(\boldsymbol{p}_1,\ldots,\boldsymbol{p}_{_K},\boldsymbol{q}_1,\ldots,\boldsymbol{q}_{_K})$ under constraints (3.1) and (3.3); let $\hat{\boldsymbol{p}}_k(\theta)$ and $\hat{\boldsymbol{q}}_k(\theta)$, $k = 1,\ldots,K$, be the maximizer of $\ell(\boldsymbol{p}_1,\ldots,\boldsymbol{p}_{_K},\boldsymbol{q}_1,\ldots,\boldsymbol{q}_{_K})$ under constraints (3.1), (3.2), and (3.3) with a fixed θ. The maximum empirical likelihood estimator of $\theta_{_N}$ is given by

$$\hat{\theta}_{_N} = \sum_{k=1}^{K} \hat{W}_k \sum_{i\in\mathbf{S}_{1k}} \hat{p}_{ik} y_{1i} - \sum_{k=1}^{K} \hat{W}_k \sum_{i\in\mathbf{S}_{0k}} \hat{q}_{ik} y_{0i}\,.$$

The pseudo empirical (log) likelihood ratio statistic for θ is computed as

$$r(\theta) = \ell(\hat{\boldsymbol{p}}_1(\theta),\ldots,\hat{\boldsymbol{p}}_{_K}(\theta),\hat{\boldsymbol{q}}_1(\theta),\ldots,\hat{\boldsymbol{q}}_{_K}(\theta)) - \ell(\hat{\boldsymbol{p}}_1,\ldots,\hat{\boldsymbol{p}}_{_K},\hat{\boldsymbol{q}}_1,\ldots,\hat{\boldsymbol{q}}_{_K})\,.$$

The asymptotic distribution of $r(\theta_{_N})$ under the current setting, however, does not seem to have a tractable form due to multiple sources of variation: the estimated propensity scores $\hat{a}_i$ based on an assumed model; post-stratification by $\hat{a}_i$; the estimation of stratum population weights W_k; approximation of propensity scores within each stratum by a constant.

For single-stage unequal probability sampling designs for the initial sample $\mathbf{S}$, we propose to use the following bootstrap procedures to obtain an approximation to the sampling distribution of $r(\theta_{_N})$.

1. Select a sample $\mathbf{S}^*$ of size n from the initial sample $\mathbf{S}$ using simple random sampling with replacement ($\mathbf{S}^*$ typically contains duplicated units from $\mathbf{S}$).

2. Obtain the estimated propensity scores $\hat{a}_i$, $i \in \mathbf{S}^*$ using data $\{(R_i,\boldsymbol{x}_i), i \in \mathbf{S}^*\}$ and the model (2.6).

3. Break the bootstrap sample $\mathbf{S}^*$ into K subsamples $\mathbf{Q}_1^*,\ldots,\mathbf{Q}_{_K}^*$ based on the ordering of the $\hat{a}_i$, $i \in \mathbf{S}^*$.

4. Compute $\hat{W}_k^*$ using $\{d_i, i \in \mathbf{Q}_k^*\}$; Split $\mathbf{Q}_k^*$ into $\mathbf{S}_{1k}^*$ and $\mathbf{S}_{0k}^*$ based on whether $R_i = 1$ or $R_i = 0$; Calculate $\tilde{d}_i(\mathbf{S}_{1k}^*) = d_i / \sum_{j \in \mathbf{S}_{1k}^*} d_j$ for $i \in \mathbf{S}_{1k}^*$ and $\tilde{d}_i(\mathbf{S}_{0k}^*) = d_i / \sum_{j \in \mathbf{S}_{0k}^*} d_j$ for $i \in \mathbf{S}_{0k}^*$, $k = 1, \ldots, K$.

5. Compute $r^{[1]}(\theta)$ in the same way that $r(\theta)$ is computed, but use $\hat{W}_k^*$, $\mathbf{S}_{1k}^*$, $\mathbf{S}_{0k}^*$, $\tilde{d}_i(\mathbf{S}_{1k}^*)$, $\tilde{d}_i(\mathbf{S}_{0k}^*)$, and $\theta = \hat{\theta}_N$ in the formulation of $\ell(\boldsymbol{p}_1, \ldots, \boldsymbol{p}_K, \boldsymbol{q}_1, \ldots, \boldsymbol{q}_K)$, (3.1), (3.2) and (3.3).

6. Repeat Steps 1-5 independently $B = 1{,}000$ times to obtain $r^{[1]}(\theta)$, $r^{[2]}(\theta)$, $\ldots$, $r^{[B]}(\theta)$.

Let b_α be the lower 100αth quantile from the simulated sequence given in Step 6. The $(1-\alpha)$-level pseudo empirical likelihood ratio confidence interval for the treatment effect θ_N can be constructed as (2.7).

Under standard settings for stratified survey samples, the pseudo empirical likelihood ratio statistic follows a scaled chi-square distribution (Theorems 3 and 4, Wu and Rao (2006)). A bootstrap method was described in Wu and Rao (2010) and was shown to be valid if the sampling fractions within all strata are small. The proposed bootstrap method takes into account of the post-stratification through the estimated propensity scores, and is shown to perform well in the simulation studies.

A major computational task is the constrained maximization of $\ell(\boldsymbol{p}_1, \ldots, \boldsymbol{p}_K, \boldsymbol{q}_1, \ldots, \boldsymbol{q}_K)$ subject to (3.1), (3.2) and (3.3). A simple technique is to introduce the factor $1/2$, as in Section 2, and reformulate the problem as a single stratified sample, with a total of $2K$ strata and stratum weights $(\hat{W}_1/2, \ldots, \hat{W}_K/2, \hat{W}_1/2, \ldots, \hat{W}_K/2)$. The reformulated problem can then be handled straightforwardly by the algorithm of Wu (2004) on the empirical likelihood method for stratified survey samples. The R code presented in Wu (2005) can be modified to handle the current problem without any major difficulties. See Rao and Wu (2009) for an overview on the formulation as well as computational aspects of empirical likelihood methods for non-stratified and stratified complex surveys.

4. Simulation Studies

In this section we report results from a simulation study on performances of the maximum pseudo empirical likelihood estimators for the treatment effect and pseudo empirical likelihood ratio confidence intervals under different scenarios for the population and the missing-by-design mechanisms. The survey population consisted of $N = 20{,}000$ units, with five variables of interest:

$\{(y_{1i}, y_{0i}, x_{1i}, x_{2i}, x_{3i}), i = 1, \ldots, N\}$. The x variables were generated using $x_{1i} \sim$ Bernoulli(0.5) (gender), $x_{2i} \sim$ Uniform$(0, 1)$ (age), $x_{3i} \sim$ Exponential(1) (size measure of the unit). The two response variables were y_{1i} under treatment and y_{0i} under control, and both are associated with the x variables through a linear regression model

$$y_{ti} = \beta_{0t} + \beta_{1t}x_{1i} + \beta_{2t}x_{2i} + \beta_{3t}x_{3i} + \varepsilon_{ti}\,, \quad i = 1, \ldots, N\,, \quad t = 1, 0\,,$$

where the ε_{ti} were independently generated from $N(0, \sigma_t^2)$. The two auxiliary variables x_{1i} (gender) and x_{2i} (age) have different impacts on the response in terms of treatment or control and they played a bigger role than the variable x_{3i}, as is reflected by the choices $(\beta_{11}, \beta_{21}, \beta_{31}) = (1.0, 2.0, 0.3)$ and $(\beta_{10}, \beta_{20}, \beta_{30}) = (2.0, 1.0, 0.3)$. The two intercepts β_{01} and β_{00} were chosen such that $y_{1i} > 0$ and $y_{0i} > 0$ for all units. The residual variances σ_t^2, $t = 1, 0$ were used to control the correlation coefficients $\rho_t(y, \boldsymbol{x})$ between the response variable y_{ti} and the linear predictor $\beta_{0t} + \beta_{1t}x_{1i} + \beta_{2t}x_{2i} + \beta_{3t}x_{3i}$.

The values of the treatment assignment indicator R_i were generated for all units in the survey population using a logistic regression model similar to (2.6). We considered two scenarios.

(i) Ignorable missing-by-design: The propensity scores $a_i = P(R_i = 1|y_{1i}, y_{0i}, x_{1i}, x_{2i}, x_{3i}) = a(x_{1i}, x_{2i}, \boldsymbol{\eta})$ depend only on the two auxiliary variables gender and age.

(ii) Nonignorable missing-by-design: The values of $a_i = P(R_i = 1|y_{1i}, y_{0i}, x_{1i}, x_{2i}, x_{3i}) = a(y_{1i}, y_{0i}, \boldsymbol{\eta})$ depend on both potential outcome variables y_{1i} and y_{0i}.

The model parameters $\boldsymbol{\eta}$ were chosen to control the mean and the range of the propensity scores. We considered unbalanced treatment and control assignments with $\bar{R}_N = N^{-1}\sum_{i=1}^N R_i \doteq 0.65$, on average 65% of units would have chosen the treatment group if selected by the initial sample. The survey population along with the treatment assignments was held fixed for repeated simulation runs. This assured that treatment assignments were not confounded with selection of units for simulated samples. The parameter of interest was the treatment effect $\theta_N = N^{-1}\sum_{i=1}^N y_{1i} - N^{-1}\sum_{i=1}^N y_{0i}$.

We considered single-stage unequal probability sampling designs for the initial survey sample $\mathbf{S}$, selected by the randomized systematic PPS sampling method of Goodman and Kish (1950) and Hartley and Rao (1962), with inclusion probabilities $\pi_i = P(i \in \mathbf{S}) \propto x_{3i}$. In the simulation we added a constant

c to all x_{3i} to avoid very small size measures, which led to $\max \pi_i / \min \pi_i \approx 12$ for the PPS samples. The simulated sample data can be represented by $\{(R_i, y_{1i}, y_{0i}, x_{1i}, x_{2i}, \pi_i), i \in \mathbf{S}\}$. For all the calculations, however, the responses only involved $\{y_{1i} | i \in \mathbf{S} \text{ and } R_i = 1\}$ and $\{y_{0i} | i \in \mathbf{S} \text{ and } R_i = 0\}$. We considered sample sizes $n = 200$, 400 and 600, corresponding to sampling fractions 1%, 2%, and 3%, which were viewed as small. For $n = 200$, we encountered computational issues for the bootstrap calibration methods with the post-stratification approach. The subsample sizes were $m_k = 40$ with $K = 5$. Under the unbalanced treatment assignments with $\bar{R}_N \doteq 0.65$, some bootstrap subsamples corresponding to the low or high propensity scores could have no units belonging to the treatment (or control) group. The issues disappeared completely for $n = 600$. The results reported below are based on $n = 400$.

For each simulated sample, we first fit the logistic regression model (2.6) to obtain estimated propensity scores $\hat{a}_i$, $i \in \mathbf{S}$ using data $\{(R_i, x_{1i}, x_{2i}), i \in \mathbf{S}\}$. This is the correct model for Scenario (i) but an incorrect model for Scenario (ii). Results under Scenario (ii) would shed light on the consequences of analyzing nonignorable missing-by-design pretest-posttest studies based on the ignorable treatment assignment assumption. We considered six maximum pseudo empirical likelihood estimators of θ_N using six different approaches.

*A*1. The naive two-sample pseudo empirical likelihood method, which is equivalent to setting the propensity scores as a constant for the method presented in Section 2 and using $\tilde{w}_{1i}(\mathbf{S}_1) = d_i / \sum_{j \in \mathbf{S}_1} d_j$ and $\tilde{w}_{0i}(\mathbf{S}_0) = d_i / \sum_{j \in \mathbf{S}_0} d_j$ where $d_i = 1/\pi_i$. The constraints (2.5) on auxiliary variables x_{1i} and x_{2i} were not used.

*A*2. The naive two-sample pseudo empirical likelihood method with constraints (2.5) on auxiliary variables x_{1i} and x_{2i}.

*A*3. The propensity score adjusted pseudo empirical likelihood method presented in Section 2 without constraints (2.5).

*A*4. The propensity score adjusted pseudo empirical likelihood method presented in Section 2 with constraints (2.5).

*A*5. The pseudo empirical likelihood method under post-stratification by the estimated propensity scores, presented in Section 3 with the choice $K = 5$, without constraints (3.3).

*A*6. The pseudo empirical likelihood method under post-stratification by the

estimated propensity scores, presented in Section 3 with the choice $K = 5$, with constraints (3.3).

Performance of a point estimator $\hat{\theta}_N$ was evaluated by the simulated relative bias (in percentage, RB%) and mean squared error (MSE) computed as

$$\text{RB\%} = 100 \times \frac{1}{M}\sum_{m=1}^{M}\frac{(\hat{\theta}_N^{(m)} - \theta_N)}{|\theta_N|}, \quad \text{MSE} = \frac{1}{M}\sum_{m=1}^{M}\left(\hat{\theta}_N^{(m)} - \theta_N\right)^2,$$

where $\hat{\theta}_N^{(m)}$ is the estimator $\hat{\theta}_N$ computed from the mth simulated sample and $M = 2{,}000$ is the total number of simulation runs. Our simulations were programmed in R; the simulation codes are available from the authors upon request.

The range of the propensity scores, denoted by $(a_{\min}, a_{\max})$, has a major impact on the performances of different methods. We considered the cases $(a_{\min}, a_{\max}) = (0.56, 0.73)$ and $(a_{\min}, a_{\max}) = (0.20, 0.95)$. The first of them represents situations where the propensity scores are less variable and hence the missing-by-design mechanism leans toward missing completely at random. The second represents the other end of the spectrum. For all scenarios investigated in the simulation the average propensity score is 0.65.

The simulated relative bias (RB%) and the mean squared error (MSE) of point estimators of θ_N for different settings are presented in Table 1. We have several observations from the simulation results with $n = 400$.

1. When the estimation uses the correct model for the propensity scores and $(a_{\min}, a_{\max}) = (0.56, 0.73)$, the methods $A3$ and $A4$ of Section 2 and $A5$ and $A6$ of Section 3 all show excellent results. The relative biases are all within 4% except for $(a_{\min}, a_{\max}) = (0.20, 0.95)$ and $\rho_t(y, \boldsymbol{x}) = 0.30$ where the RBs are around 6%.

2. When the estimation uses an incorrect model for the propensity scores, the methods $A3 - A6$ provide acceptable results only for one particular setting: $(a_{\min}, a_{\max}) = (0.56, 0.73)$ and $\rho_t(y, \boldsymbol{x}) = 0.80$. Here propensity scores are relatively uniform and the correlation between y and $\boldsymbol{x}$ is strong.

3. The naive method $A1$ does not provide any valid results. The naive estimator from method $A2$ with calibration constraints on auxiliary variables x_{1i} and x_{2i} has excellent performance, very similar to $A3$ when $(a_{\min}, a_{\max}) = (0.56, 0.73)$, but better than $A3$ when $(a_{\min}, a_{\max}) = (0.20, 0.95)$.

4. All methods under highly variable propensity scores ($(a_{\min}, a_{\max}) = (0.20, 0.95)$) coupled with misspecification of the propensity score model do not

Table 1. Relative bias (in %) and mean squared error of point estimators of θ_N (I : $(a_{\min}, a_{\max}) = (0.56, 0.73)$, II : $(a_{\min}, a_{\max}) = (0.20, 0.95)$; (i): $P(R = 1|y, \boldsymbol{x}) = a(x_1, x_2)$, (ii): $P(R = 1|y, \boldsymbol{x}) = a(y_1, y_0)$).

		$\rho_t(y, \boldsymbol{x})$		$A1$	$A2$	$A3$	$A4$	$A5$	$A6$
I	(i)	0.80	RB%	14.3	−0.4	−0.8	−0.8	0.3	−0.9
			MSE	0.040	0.010	0.010	0.010	0.011	0.010
		0.30	RB%	11.6	−3.2	−3.8	−3.8	−2.7	−3.9
			MSE	0.158	0.135	0.136	0.136	0.139	0.139
	(ii)	0.80	RB%	11.4	3.2	2.7	2.7	3.6	2.7
			MSE	0.033	0.011	0.011	0.011	0.012	0.011
		0.30	RB%	22.6	19.7	19.4	19.4	19.6	19.4
			MSE	–	–	–	–	–	–
II	(i)	0.80	RB%	80.6	1.0	−2.7	−1.7	2.3	−2.1
			MSE	0.667	0.014	0.020	0.015	0.016	0.015
		0.30	RB%	77.6	−1.3	−6.3	−5.5	−2.4	−6.5
			MSE	0.753	0.203	0.225	0.217	0.215	0.220
	(ii)	0.80	RB%	53.8	18.5	16.9	17.0	18.9	17.0
			MSE	–	–	–	–	–	–
		0.30	RB%	104.2	91.1	90.4	90.4	91.7	90.5
			MSE	–	–	–	–	–	–

produce acceptable results. All estimators are seriously biased for those cases and the related MSEs are not reported in Table 1.

With $n = 600$, all scenarios with small relative biases for $n = 400$ continue to have small biases with decreased mean squared errors. Where the relative biases are unacceptably large for $n = 400$ and the corresponding methods are invalid, the relative biases remain at the same magnitude, where the mean squared errors are no longer relevant.

That the naive estimator from method $A2$ with calibration constraints on auxiliary variables performs well is not a surprise as calibration has been shown to be a useful tool to adjust for nonresponse; see, for instance, Chang and Kott (2008). Chen, Wu and Thompson (2015) observe that the calibration approach used in Huang, Qin and Follmann (2008) performs very well for the empirical likelihood-based estimation of the treatment effect in a randomized pretest-posttest study. The use of calibration constraints on $\boldsymbol{x}$ in methods $A4$ and $A6$ provides mild improvement over $A3$ and $A5$ for some cases, but none are dramatically different.

Performances of pseudo empirical likelihood ratio confidence intervals are assessed through average length (AL) and coverage probability (CP) of the interval

Table 2. Average length and coverage probability of 95% confidence intervals for θ_N (I : $(a_{\min}, a_{\max}) = (0.56, 0.73)$, II : $(a_{\min}, a_{\max}) = (0.20, 0.95)$; (i): $P(R = 1|y, \boldsymbol{x}) = a(x_1, x_2)$, (ii): $P(R = 1|y, \boldsymbol{x}) = a(y_1, y_0)$).

		$\rho_t(y, \boldsymbol{x})$		$A1$	$A2$	$A3$	$A4$	$A5$	$A6$
I	(i)	0.80	AL	0.556	0.383	0.384	0.384	0.429	0.396
			CP	83.6	94.9	95.0	94.9	96.1	94.9
		0.30	AL	1.52	1.46	1.47	1.47	1.51	1.51
			CP	94.6	95.7	95.7	95.8	95.7	95.4
	(ii)	0.80	AL	0.562	0.386	0.387	0.387	0.444	0.398
			CP	88.6	94.0	94.2	94.3	95.6	94.6
		0.30	AL	1.52	1.46	1.47	1.47	1.52	1.51
			CP	91.1	91.7	91.9	91.5	92.3	92.6
II	(i)	0.80	AL	0.530	0.457	0.539	0.467	0.517	0.486
			CP	0.0	94.5	94.8	93.3	95.3	94.6
		0.30	AL	1.52	1.78	1.83	1.81	1.86	1.85
			CP	49.5	93.9	93.8	93.7	95.2	94.4
	(ii)	0.80	AL	–	–	–	–	–	–
			CP	3.5	55.3	62.3	61.5	60.2	63.7
		0.30	AL	–	–	–	–	–	–
			CP	21.4	29.0	30.9	31.1	31.4	33.1

computed as

$$\text{AL} = \frac{1}{M}\sum_{m=1}^{M}\left(\hat{\theta}_{U}^{(m)} - \hat{\theta}_{L}^{(m)}\right), \quad \text{CP} = \frac{1}{M}\sum_{m=1}^{M} I\big(\hat{\theta}_{L}^{(m)} \le \theta_{N} < \hat{\theta}_{U}^{(m)}\big),$$

where $\big(\hat{\theta}_{L}^{(m)}, \hat{\theta}_{U}^{(m)}\big)$ is the confidence interval computed from the mth simulated sample and $I(\cdot)$ is the indicator function. Simulation results are reported in Table 2. The performances of the six confidence intervals follow those of the point estimators, if the point estimators perform well, the corresponding confidence intervals also perform well. All methods fail when $(a_{\min}, a_{\max}) = (0.20, 0.95)$ and the model for propensity scores is misspecified.

5. The ITC Four Country Survey

The International Tobacco Control (ITC) Policy Evaluation Project was created in 2002 to measure the effectiveness of national-level tobacco control policies in selected countries which signed and ratified the Framework Convention on Tobacco Control (FCTC). The ITC project is a prospective cohort study and first started in Canada, USA, Australia, and the UK. The first wave ITC Four Country Survey used a stratified sampling design and conducted telephone interviews of over 2000 adult smokers in each of the four countries. The initial group of

respondents was followed in subsequent waves and a new cross-sectional replenishment sample was added at each wave to make up for the reduced size of the longitudinal sample due to attrition. The ITC survey questionnaires cover the domains which are relevant to the implementation of FCTC: demographic variables, smoking behaviour, warning labels, advertising and promotion, light/mild brand descriptors, taxation and purchase behaviour, stop-smoking medications and alternative nicotine products, cessation and quitting behaviour as well as key psychosocial measures. Thompson et al. (2006) contains further details on the ITC Four Country Survey.

The ITC Four Country Survey conducted a pilot study at Wave 7 to evaluate whether an online version of the survey would be a viable option for further waves. The study was to determine the amount of cost savings that could be achieved if some of the cohort participants completed the survey online, and to determine whether some people could be retained who might otherwise be lost. After the pilot study, it was decided that the web survey option would be offered to all respondents starting from Wave 8. The Wave 8 Recontact Survey employed a mixed mode for data collection, combining telephone interviews with self-administered web surveys. Each recontact respondent of Wave 8 received either an email invitation (if an email address was available at Wave 7) or a mailed letter invitation to respond online. Among the 5135 recontact respondents who were invited, 2006 (39%) participated through the web survey. Chen (2014) provides detailed descriptions of the survey data set.

One of the research problems for mixed mode surveys is to examine whether there is a mode effect in data collection. Web surveys are self-administered with the questionnaires visually presented to the respondent while telephone surveys require the reading of questions and possible answers in a particular order by the interviewer. In addition, answers to certain questions may require a respondent's recall of activities in the past, and answering those questions online in a self-controlled manner might differ from speaking to an interviewer.

We applied the methods developed in Sections 2 and 3 on two-sample empirical likelihood for nonrandomized pretest-posttest studies to the ITC Four Country survey to investigate a mode effect in data collection. The data set we used was the Canadian sample of smokers present in both Wave 7 and Wave 8 who provided answers to a question on number of cigarettes smoked per day (y, CPD). The data set contained $n = 900$ respondents, with $n_1 = 398$ participating through the web survey. Let μ_1 be the population mean response of y in Wave 8 if everyone were to answer the question through the web survey and μ_0 be the

Table 3. Point estimates and 95% confidence intervals for the mode effect: The ITC Canada survey (W8, all values multiplied by -1).

$A1$	$A2$	$A3$	$A4$	$A5$	$A6$
2.38	1.98	1.96	1.97	2.09	2.04
$(0.9, 3.8)$	$(0.5, 3.5)$	$(0.5, 3.4)$	$(0.5, 3.4)$	$(0.6, 3.6)$	$(0.5, 3.6)$

population mean response of y if the survey were done entirely through telephone interviews. The parameter of interest is $\theta = \mu_1 - \mu_0$.

The ITC survey contains a long list of questions over several areas, and it is reasonable to assume that the propensity scores of choosing the web survey depend on certain basic demographic variables but not on some of the key variables to be measured. Chen (2014) contains detailed discussions on building a model for the propensity scores with the ITC Four Country survey data set. One interesting finding is that the variable on invitation method ("Email" versus "Regular Mail") is highly correlated to the treatment assignment. Chen (2014) also presents balance checks among several covariates between the treatment group and the control group. To simplify the application and for the purpose of illustration, we fit the logistic regression model (2.6) involving only gender (x_1) and age (x_2). It has been found that both variables are highly significant. The fitted propensity scores ranged from 0.21 to 0.59. The Wave 7-8 longitudinal weights were used and the survey design is treated as if it is single stage unequal probability sampling. The six different approaches $A1 - A6$ described in Section 4 are used to compute the point estimate $\hat{\theta}$ and the 95% confidence interval $(\hat{\theta}_L, \hat{\theta}_U)$ associated with each method. The results are presented in Table 3.

There are two major findings. First, the calibration method $A2$ and the propensity score adjusted methods $A3$ and $A4$ give very similar results, while $A5$ and $A6$, based on post-stratification of the propensity scores, provide slightly different results. This is consistent with observations from the simulation results reported in Section 4. Second, all methods show that there is a mode effect for measuring CPD, with none of the 95% confidence intervals containing zero. On average, the value of CPD reported through telephone interviews tends to be two cigarettes higher than the one obtained by the web survey.

6. Concluding Remarks

We demonstrate in this paper that the pseudo empirical likelihood approach can be used for nonrandomized pretest-posttest studies with observational survey data. Our proposed approaches provide design-consistent estimators as well

as confidence intervals for the treatment effect, and allow the use of auxiliary information through additional constraints.

The missing-by-design mechanism and the distribution of the propensity scores are among factors that affect the effectiveness of these methods. There might be scenarios, such as in the ITC Four Country Survey, where the propensity scores of selecting the treatment most likely depend on basic characteristics of the respondents rather than the variables to be measured, which lends itself to the reasonable assumption of ignorable missing-by-design. Such scenarios can be efficiently handled by the proposed approaches but testing for ignorable treatment assignment has been shown in applications to be a challenging task. With nonignorable missingness-by-design, strong assumptions about the model structure for the propensity scores or additional information are typically required to obtain estimates for the propensity scores. Chen and Kim (2014) proposed a two-phase sampling method to conduct a test for ignorable missingness using additional information collected from the second phase sample. Wang, Shao and Kim (2014) discussed an instrumental variable approach for identification and estimation with nonignorable nonresponse. Shao and Wang (2016) studied semiparametric inverse propensity weighting methods for nonignorable missing data assuming that information on an instrumental variable is available. Whether these approaches can be applied to the missing-by-design pretest-posttest problems requires further investigation. Furthermore, rigorous asymptotic developments for the second proposed approach based on post-stratified samples have not been pursued in this paper and remain as a challenging problem.

Acknowledgment

This research was supported by grants from the Natural Sciences and Engineering Research Council (NSERC) of Canada. The authors also thank the International Tobacco Control Policy Evaluation Project (The ITC Project) for allowing the use of its Four Country Survey data for Canada. Waves 7 and 8 of the ITC Four Country Survey in Canada were funded by the Canadian Institutes for Health Research (79551), the Ontario Institute for Cancer Research (Senior Investigator Award) and the National Cancer Institute, US (RO1 CA100362, P50 CA111236 and P01 CA138389).

References

Austin, P. C. (2008). A critical appraisal of propensity score matching in the medical literature

from 1996–2003. *Statistics in Medicine* **27**, 2037–2049.

Austin, P. C. (2009). Balance diagnostics for comparing the distribution of baseline co-variates between treatment groups in propensity-score matched samples. *Statistics in Medicine* **28**, 3083–3107.

Brogan, D. R. and Kutner, M. H. (1980). Comparative analyses of pretest-posttest research designs. *The American Statistician* **34**, 229–232.

Chang, T. and Kott, P. S. (2008). Using calibration weighting to adjust for nonresponse under a plausible model. *Biometrika* **95**, 555–571.

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, M. (2014). Empirical likelihood methods for pretest-posttest studies, unpublished PhD dissertation, University of Waterloo.

Chen, M., Wu, C. and Thompson, M. E. (2015). An imputation based empirical likelihood approach to pretest-posttest studies. *The Canadian Journal of Statistics* **43**, 378–402.

Chen, M., Wu, C. and Thompson, M. E. (2016). Mann-Whitney test with empirical likelihood methods for pretest-posttest studies. *Journal of Nonparametric Statistics* **28**, 360–374.

Chen, S. and Kim, J. K. (2014). Two-phase sampling experiment for propensity score estimation in self-selected samples. *The Annals of Applied Statistics* **8**, 1492–1515.

Goodman, R. and Kish, L. (1950). Controlled selection - a technique in probability sampling. *Journal of the American Statistical Association* **45**, 350–372.

Hartley, H. O. and Rao, J. N. K. (1962). Sampling with unequal probabilities and without replacement. *Annals of Mathematical Statistics* **33**, 350–374.

Huang, C. Y., Qin, J. and Follmann, D. A. (2008). Empirical likelihood-based estimation of the treatment effect in a pretest-posttest study. *Journal of the American Statistical Association* **103**, 1270–1280.

Imai, K. and Ratkovic (2014). Covariate balancing propensity score. *Journal of the Royal Statistical Society, Ser. B (Statistical Methodology)* **76**, 243–263.

Li, F., Morgan, K. L. and Zaslavsky, A. M. (2015). Balancing covariates via propensity score weighting. arXiv: 1404.1785v2 [stat.ME] 27 Jan 2015.

Lunceford, J. K. and Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Statistics in Medicine* **23**, 2937–2960.

Miratrix, L., Sekhon, J. S. and Yu, B. (2013). Adjusting treatment effect estimates by post-stratification in randomized experiments. *Journal of the Royal Statistical Society, Ser. B (Statistical Methodology)* **75**, 369–396.

Owen, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* **75**, 237–249.

Rao, J. N. K. and Wu, C. (2009). Empirical likelihood methods. In *Handbook of Statistics 29B Sample Surveys: Inference and Analysis* (Edited by D. Pfeffermann and C. R. Rao), 189–207.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55.

Rosenbaum, P. R. and Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association* **79**, 516–524.

Shao, J. and Wang, L. (2016). Semiparametric inverse propensity weighting for nonignorable missing data. *Biometrika* **103**, 175–187.

Thompson, M. E., Fong, G. T., Hammond, D. et al. (2006). Methods of the international tobacco control (ITC) four country survey. *Tobacco Control* **15** (suppl III), iii12–18.

Wang, S., Shao, J. and Kim, J. K. (2014). An instrumental variable approach for identification and estimation with nonignorable nonresponse. *Statistica Sinica* **24**, 1097–1116.

Wu, C. (2004). Some algorithmic aspects of the empirical likelihood method in survey sampling. *Statistica Sinica* **14**, 1057–1067.

Wu, C. (2005). Algorithms and R codes for the pseudo empirical likelihood methods in survey sampling. *Survey Methodology* **31**, 239–243.

Wu, C. and Rao, J. N. K. (2006). Pseudo-empirical likelihood ratio confidence intervals for complex surveys. *The Canadian Journal of Statistics* **34**, 359–375.

Wu, C. and Rao, J. N. K. (2010). Bootstrap procedures for the pseudo empirical likelihood method in sample surveys. *Statistics and Probability Letters* **80**, 1472–1478.

Wu, C. and Yan, Y. (2012). Empirical likelihood inference for two-sample problems. *Statistics and Its Interface* **5**, 345–354.

Zanutto, E., Lu, B. and Hornik, R. (2005). Using propensity score subclassification for multiple treatment doses to evaluate a national antidrug media campaign. *Journal of Educational and Behavioral Statistics* **30**, 59–73.

Bank of Nova Scotia, 4 King Street West, Toronto, ON, Canada M5H 1B6.

E-mail: minchen12@gmail.com

Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, ON, N2L 3G1, Canada.

E-mail: methomps@uwaterloo.ca

Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, ON, N2L 3G1, Canada.

E-mail: cbwu@uwaterloo.ca

(Received June, 2016; accepted July, 2017)