Research Article

Statistics in Medicine

Received 13 October 2011, Accepted 30 May 2012 Published online in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.5530

Efficient analysis of case-control studies with sample weights

V. Landsman[a*†] and B. I. Graubard[b]

Analysis of population-based case–control studies with complex sampling designs is challenging because the sample selection probabilities (and, therefore, the sample weights) depend on the response variable and covariates. Commonly, the design-consistent (weighted) estimators of the parameters of the population regression model are obtained by solving (sample) weighted estimating equations. Weighted estimators, however, are known to be inefficient when the weights are highly variable as is typical for case–control designs. In this paper, we propose two alternative estimators that have higher efficiency and smaller finite sample bias compared with the weighted estimator. Both methods incorporate the information included in the sample weights by modeling the sample expectation of the weights conditional on design variables. We discuss benefits and limitations of each of the two proposed estimators emphasizing efficiency and robustness. We compare the finite sample properties of the two new estimators and traditionally used weighted estimators with the use of simulated data under various sampling scenarios. We apply the methods to the U.S. Kidney Cancer Case-Control Study to identify risk factors.

Keywords: informative sampling; modeling sample weights; sample distribution; weighted estimating equations

1. Introduction

The application of complex survey sample designs in population-based case-control studies have been discussed in the statistical literature [1–4]. In these designs, cases are selected from a population disease (case) registry, and controls are selected from the population at risk covered by the disease registry, usually using stratified or multistage cluster sampling. The sampling rates under these sample designs are informative, because they depend on the case–control status and other covariates that can be also related to the exposure of interest. For example, the U.S. Kidney Cancer Case-Control Study (KCS), considered in this paper, aimed to examine the association between hypertension and renal cell carcinoma in African American population. To attain sufficient sizes of African American cases and controls efficiently, African American cases were sampled with certainty while undersampling some age–sex–race strata of white cases, whereas the sample rates for the controls were higher for individuals from areas with greater percentages of African Americans [5]. Because hypertension is potentially associated with socioeconomic status and other factors related to geographical location, the informativeness of the sampling rates in the KCS needed to be accounted for to assure consistent estimation of the population associations of interest.

Various weighted estimation approaches where the observations are weighted in some fashion by the inverse of the sample rates, e.g., Horvitz-Thompson estimation, have been proposed for conducting logistic regression analyses with informatively sampled observations [1, 6–8]. However, the variances of the weighted estimators of logistic regression coefficients from these analyses can become quite large for sample designs with highly variable sample rates, such as those that can result from population-based case–control studies. Moreover, asymptotically consistent robust variance estimators

[a] *Center for Global Health Research, St. Michael's Hospital, Toronto, ON M5B1W8, Canada*
[b] *National Cancer Institute, Bethesda, MD 20852, U.S.A.*
**Correspondence to: V. Landsman, Center for Global Health Research, St. Michael's Hospital, 30 Bond Street, Toronto, ON M5B1W8, Canada.*
[†] *E-mail: landsmanv@smh.ca*

of estimated model parameters, such as Taylor linearization or delete-one jackknife [9], can seriously underestimate the true variances for moderate sample sizes (Section 6).

In this paper, we propose two new estimators for conducting logistic regression analyses for case–control studies: the sample pseudo likelihood (SPL) estimator and the semiparametric weighted (SPW) estimator. We adopt the underlying theory of these estimators from the survey sampling literature for cross-sectional samples to establish our estimators in the case–control setting [10]. Both estimators require fitting a regression model to the known sample weights versus the design variables and/or response variable and estimating the conditional expectation of the weights with respect to their distribution in the sample. For SPW estimation, we use the estimated conditional expectation of the weights to modify the original known sample weights. Estimators using estimated weights (as the SPW estimator) can be considerably more efficient (have smaller variances of estimated regression parameters) than the existing weighted estimators using known sample weights (e.g., [11]). Lumley *et al.* [12] in a recent paper have described insight into possible reasons for the improved efficiency of estimated weights with regard to calibrated weights. We empirically show that the variances of our proposed estimators can be estimated accurately using classic robust variance estimation methods applicable to complex survey data (Section 5) and that these estimated variances are considerably smaller compared with the variances of the other weighted estimators.

The SPL estimator generalizes the Breslow and Cain estimator [13] to general sampling designs where the selection mechanism is a function of the response, sample design variables, and their interaction, and the design variables may contain continuous as well as categorical variables. The SPW estimator can be viewed as an improved version of the design-based weighted estimator, with the improvement achieved by applying the appropriate modifications on the original sample weights prior to the estimation. SPW estimator is shown to have nearly the same efficiency and finite sample bias as the SPL estimator. However, the SPW estimator is shown to be robust to misspecification of the model fitted to the sample weights, which is an advantage over the SPL estimator in situations where the selection mechanism cannot be accurately modeled.

We organize the rest of the paper as follows. In the next section, we describe the data, the model, and the classical weighted estimator (Horvitz–Thompson-type estimator) commonly applied to informative samples. In Section 3, we derive the SPL estimator for the case–control data. In Section 4, we describe the SPW estimator and provide the theoretical basis for it. We address variance estimation for the proposed estimators in Section 5. In Section 6, we compare the proposed estimators along with the other weighted estimators described in the paper with the use of simulated data. We illustrate the estimation methods described in this paper with an analysis of risk factors for kidney cancer in the KCS data in Section 7. We conclude with a brief discussion in Section 8.

2. Model, data, and Horvitz–Thompson weighted estimator

Consider a finite population $\mathcal{U}$ of N individuals, consisting of N_1 cases and N_0 controls. Let Y denote a binary response variable in which $Y = 1$ for cases (diseased) and $Y = 0$ for controls (nondiseased). Let $\mathbf{z} = (\mathbf{x}', \mathbf{v}')'$ denote the vector of all measured covariates, where $\mathbf{x}$ are covariates associated with the outcome and $\mathbf{v}$ are design variables used in the process of sample selection. $\mathbf{x}$ and $\mathbf{v}$ may or may not have common variables.

Let $n = n_0 + n_1$ be the size of the observed sample S with n_1 cases and n_0 controls. The probability that individual i , $i = 1, 2, \ldots, n$, is included in the sample is denoted by π_i. In case–control studies, π_i generally depend on the outcome Y_i, design covariates $\mathbf{v}_i$, and the interactions $Y_i \times \mathbf{v}_i$, because different sampling rates (depending on $\mathbf{v}_i$) are applied for cases and controls. The (base) sample weight w_i is defined as a reciprocal of the sample inclusion probability π_i, that is, $\pi_i = 1/w_i$. The final sample weight is, in addition, adjusted for unit nonresponse and calibrated to population totals. Regression and poststratification calibration of the sample weights to the population totals are often used in survey research to reduce the variance and bias of the weighted estimators described in the rest of this section ([6]; [12, pp. 163–165]). We refer to the final sample weight as the sample weight of individual i and denote it by w_i. In some cases, with poststratification used, the final weights are constructed so that $\sum_{i \in S_l} w_i = N_l$, where S_l denotes the sample of cases for $l = 1$ and the sample of controls for $l = 0$ such that $S = S_1 \cup S_0$. Otherwise, we assume that, under appropriate conditions, $\sum_{i \in S_l} w_i$ converges to N_l as $N \to \infty$, $n_0 \to \infty$ and $n_1 \to \infty$. We assume throughout the paper that the observed data are the data for the sampled units, that is, for individual $i \in S$, $i = 1, 2, \ldots, n$, we observe $(y_i, \mathbf{z}_i', w_i)'$.

We also assume the availability of population proportions and means for the variables in $\mathbf{v}_i$, which can be used to calibrate the sample weights.

In the population, the outcomes Y_i conditional on covariates $\mathbf{x}_i$ have distribution $\Pr(Y_i = l|\mathbf{x}_i) = \mathcal{P}_l(\mathbf{x}_i;\boldsymbol{\beta})$, where $l = 0, 1$ and $\boldsymbol{\beta}$ is a vector of parameters. We refer to $\mathcal{P}_l(\mathbf{x}_i;\boldsymbol{\beta})$ as the *population distribution* of Y_i given $\mathbf{x}_i$. In this paper, we use the logistic regression model:

$$\text{logit}(\mathcal{P}_1(\mathbf{x}_i;\boldsymbol{\beta})) = \beta_0 + \mathbf{x}_i'\boldsymbol{\beta}_1, \tag{1}$$

where $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_1)$ and $\mathcal{P}_0(\mathbf{x}_i;\boldsymbol{\beta}) = 1 - \mathcal{P}_1(\mathbf{x}_i;\boldsymbol{\beta})$. The primary goal of this paper is in estimating the parameter $\boldsymbol{\beta}_1$ under a correctly specified population model (1). Estimation of $\boldsymbol{\beta}_1$ from the observed data requires taking into account the informativeness of the sample selection. A common way to account for informativeness of the sample selection is using the method of sample weighted estimating equations outlined in the following text.

If we fit model (1) to the entire finite population $\mathcal{U}$, we could estimate $\boldsymbol{\beta}$ consistently by solving a system of estimating equations with respect to $\mathbf{b}$:

$$\boldsymbol{G}(\mathbf{b}) = \sum_{t=1}^{N} \mathbf{x}_i(Y_i - \mathcal{P}_1(\mathbf{x}_i;\mathbf{b})) = \mathbf{0}, \tag{2}$$

where $\mathcal{P}_1(\mathbf{x}_i;\mathbf{b}) = \text{expit}(b_0 + \mathbf{x}_i'\mathbf{b}_1)$ and $\mathcal{P}_0(\mathbf{x}_i;\mathbf{b}) = 1 - \mathcal{P}_1(\mathbf{x}_i;\mathbf{b})$. Denote the solution of (2) by $\mathbf{B}$. $\mathbf{B}$ is often called a *census parameter*, and it is a consistent estimator of $\boldsymbol{\beta}$ under the true population model $\mathcal{P}_1$. Equation (2) are maximum-likelihood estimating equations based on all the values $(Y_i, \mathbf{x}_i) \in \mathcal{U}$ if $\mathcal{U}$ is a simple random sample from a superpopulation generated under model (1).

For any fixed value of $\mathbf{B}$, $\boldsymbol{G}(\mathbf{B})$ is a vector of finite population totals, and hence, it can be estimated from the sample S by

$$\hat{\boldsymbol{G}}_{\text{W}}(\mathbf{b}) = \sum_{i\in S} w_i\mathbf{x}_i(y_i - \mathcal{P}_1(\mathbf{x}_i;\mathbf{b})), \tag{3}$$

where y_i is the observed response for unit i. Denote the solution of the system of the equations $\hat{\boldsymbol{G}}_{\text{W}}(\mathbf{b}) = \mathbf{0}$ by $\hat{\boldsymbol{\beta}}_{\text{W}}$. We refer to $\hat{\boldsymbol{\beta}}_{\text{W}}$ as the weighted estimator (sometimes called the Horvitz–Thompson estimator). Binder [14] has established consistency and asymptotic normality of $\hat{\boldsymbol{\beta}}_{\text{W}}$ under suitable conditions.

We can refer the weighted estimation method described previously as a *design-based* method, because it utilizes the information related to the sampling design by directly including the final sample weights w_i in the estimating equations, and the estimating equations (3) are design based unbiased for (2). As mentioned previously, the major disadvantage of this approach is that it can provide very inefficient estimators, notably when there is large variation of the sample weights and $\mathbf{x}$ contains a continuous variable [7]. In the next two sections, we propose two alternative estimators that are later shown to have improved efficiency and smaller finite sample bias over the weighted estimator $\hat{\boldsymbol{\beta}}_{\text{W}}$.

3. Sample pseudo maximum likelihood estimator

The sample distribution approach is an alternative to a design-based (weighting) approach to account for informative sampling in estimation of population parameters [10]. This method determines the distribution of the observed data that results from a combination of the postulated population distribution and sample selection process, and fits this distribution to the data. In this section, we apply this method to the case–control setting and obtain an SPL estimator for parameter $\boldsymbol{\beta}$ in model (1).

Denote by $f_P(y_i|\mathbf{x}_i)$ the (conditional) *pdf* of the responses in the population (the population distribution). The observed data, however, arise from the sample distribution f_S, obtained by Bayes rule as

$$f_S(y_i|\mathbf{z}_i) = f_P(y_i|\mathbf{z}_i, i \in S) = \frac{\Pr(i \in S|y_i, \mathbf{v}_i) f_P(y_i|\mathbf{x}_i)}{\Pr(i \in S|\mathbf{z}_i)}, \tag{4}$$

where $\Pr(i \in S|\mathbf{z}_i) = \Pr(i \in S|Y_i = 1, \mathbf{v}_i)P(Y_i = 1|\mathbf{x}_i) + \Pr(i \in S|Y_i = 0, \mathbf{v}_i)P(Y_i = 0|\mathbf{x}_i)$ for binary Y. (We assume in Equation (4) and throughout the paper that $f_P(y_i|\mathbf{z}_i) = f_P(y_i|\mathbf{x}_i)$ and $\Pr(i \in S|y_i, \mathbf{z}_i) = \Pr(i \in S|y_i, \mathbf{v}_i)$ for each unit i.)

In a case–control study, the sample inclusion probabilities often depend on y values because the sampling mechanism is different for the cases and for the controls, implying that $\Pr(i \in S|y_i, \mathbf{v}_i) \neq \Pr(i \in S|\mathbf{v}_i)$. Then, it follows from (4) that the sample distribution is different from the population distribution, $f_P \neq f_S$. This is an example of informative sampling that should be accounted for in the inferential process.

Using the relationships between the moments of the population and sample distributions established in [15] and adding the unknown parameters to the notation, we write (4) equivalently as

$$f_S(y_i|\mathbf{z}_i;\boldsymbol{\gamma};\boldsymbol{\beta}) = \frac{E_S(w_i|\mathbf{z}_i;\boldsymbol{\gamma},\boldsymbol{\beta})\, f_P(y_i|\mathbf{x}_i;\boldsymbol{\beta})}{E_S(w_i|y_i,\mathbf{v}_i;\boldsymbol{\gamma})}, \tag{5}$$

where $E_S(w_i|\cdot)$ stands for the conditional expectation of the weights w_i with respect to the their distribution in the sample and $\boldsymbol{\gamma}$ is a vector of (unknown) regression coefficients of the regression model for sample weights on the design variables and the outcome. The main idea behind the sample distribution method is to recover the target parameters of the population distribution through (5) by fitting the distribution f_S to the observed data. Presenting the sample distribution in the form of (5) is convenient for practical implementation of the method, because the sample expectations $E_S(w_i|\cdot)$ are with respect to the distribution of the weights in the sample and can be estimated using the observed sample data.

By the 'independence result' established in Pfeffermann *et al.* [16], the sample observations $Y_i|\mathbf{z}_i$, $i \in S$ are approximately independent for commonly used sample designs (e.g., stratified sample designs) if the population observations $Y_i|\mathbf{z}_i$, $i \in \mathcal{U}$ are independent. For sample designs where the independence result holds, assuming that $\boldsymbol{\gamma}$ is known and fixed, $\boldsymbol{G}_S(\boldsymbol{\gamma};\boldsymbol{\beta}) = \sum_{i\in S} \frac{\partial \log f_S(y_i|\mathbf{z}_i;\boldsymbol{\gamma},\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$ define the sample log-likelihood score equations with respect to $\boldsymbol{\beta}$. If (1) is assumed to be a population distribution of Y_i given $\mathbf{x}_i$, it follows from (5) that these equations can be written equivalently as

$$\boldsymbol{G}_S(\boldsymbol{\gamma};\boldsymbol{\beta}) = \sum_{i\in S} \mathbf{x}_i(y_i - \mathcal{P}_1(\mathbf{x}_i;\boldsymbol{\beta})) + \sum_{i\in S} \frac{\partial \log E_S(w_i|\mathbf{z}_i;\boldsymbol{\gamma},\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}, \tag{6}$$

where $E_S(w_i|\mathbf{z}_i;\boldsymbol{\beta},\boldsymbol{\gamma})$ is differentiable with respect to $\boldsymbol{\beta}$ for any fixed $\boldsymbol{\gamma}$.

Fitting model (5) to the sample data is performed in two steps. At the first step, $\boldsymbol{\gamma}$ is estimated separately (and denoted by $\hat{\boldsymbol{\gamma}}$)) using individual-level data $w_i, y_i, \mathbf{v}_i$ observed for the sample units only, $i \in S$. Next, the sample expectations $E_S(w_i|\mathbf{z}_i)$ are expressed as functions of $\boldsymbol{\beta}$ and the estimated value of $\boldsymbol{\gamma}$ (see Appendix A for details). At the second step, the system of the equations $\boldsymbol{G}_S(\hat{\boldsymbol{\gamma}};\boldsymbol{\beta}) = \mathbf{0}$, obtained from (6) by substituting $\hat{\boldsymbol{\gamma}}$ in place of $\boldsymbol{\gamma}$, is solved iteratively with respect to $\boldsymbol{\beta}$.

It can be shown that if $\hat{\boldsymbol{\gamma}}$ is a $\sqrt{n}$-consistent estimator, the solution of the system (6) (with $\hat{\boldsymbol{\gamma}}$ substituted in place of $\boldsymbol{\gamma}$) is a consistent estimator for $\boldsymbol{\beta}$. We denote the solution of (6) by $\hat{\boldsymbol{\beta}}_{\mathrm{SPL}}$ and refer to it as an SPL estimator according to [17].

The Breslow and Cain estimator [13] is a special case of the SPL estimator (see Appendix B for the proof). The important feature of the SPL estimator is that it can be applied to case–control data collected by different complex sampling designs, where w_i can be functions of $y_i, \mathbf{v}_i$ and their cross products, and $\mathbf{v}_i$ may include also continuous components [see Section 6, sample plans A(E) and A(M)]. However, the application of the SPL estimator for general cluster designs is not straightforward. In the simulation study for this paper, we investigate a special case of cluster sampling (Section 6, sample plan B). In this sampling plan, the conditional responses $Y_i|\mathbf{z}_i$ are generated independently in the population, and hence the 'independence result' established in [16] still holds for the sample data despite the intraclass correlation (ICC) present in the covariates. In such situations, the SPL estimator is still valid, but the variance estimation should take the cluster sampling into account (Section 6.4).

4. Semiparametric weighted estimator

In this section, we present the SPW estimator, which can have better statistical properties than other weighted estimators considered in this paper. We obtain this estimator as a solution of a system of weighted estimating equations (3) but with modified rather than original sample weights. We obtain the modified sample weights by the following two-step procedure:

(1) Define the rescaled weight $\tilde{w}_i$ by

$$\tilde{w}_i = y_i w_i / M_1 + (1 - y_i) w_i / M_0,\ i \in S, \tag{7}$$

where $M_1 = \sum_{i \in S} y_i w_i / n_1$ is the mean of the weights within the cases and $M_0 = \sum_{i \in S}(1 - y_i)$ w_i / n_0 is the mean of the weights within the controls.

(2) Fit the regression model to the rescaled weights $\tilde{w}_i$ versus the covariates $\mathbf{z}_i$ and compute $\hat{E}_S(\tilde{w}_i|\mathbf{z}_i) = E_S(\tilde{w}_i|\mathbf{z}_i; \hat{\boldsymbol{\theta}})$, where $\hat{\boldsymbol{\theta}}$ denotes the estimated coefficients of the regression model relating $\tilde{w}_i$ to the covariates $\mathbf{z}_i$. Define the model-adjusted weights by $\tilde{w}_i^m = \tilde{w}_i / \hat{E}_S(\tilde{w}_i|\mathbf{z}_i)$.

After $\tilde{w}_i^m$ are estimated, the system of the weighted estimating equations

$$\hat{\boldsymbol{G}}_{\text{SPW}}(\mathbf{b}) = \sum_{i \in S} \tilde{w}_i^m \mathbf{x}_i (y_i - \mathcal{P}_1(\mathbf{x}_i; \mathbf{b})) = \mathbf{0} \tag{8}$$

is solved with respect to $\mathbf{b}$, and its solution, $\hat{\boldsymbol{\beta}}_{\text{SPW}}$, is the SPW estimator.

The first two steps of the algorithm define the transformations on the sample weights in order to reduce their variability. The first step is a rescaling transformation. Because $M_1 \approx N_1/n_1 \approx 1$ and $M_0 \approx N_0/n_0$ tend to be large numbers for case–control designs, it follows from (7) that the rescaled weights $\tilde{w}_i$ will have lower variation than the original weights w_i. The second step of the algorithm models the rescaled weights as a function of the design variables. It follows from the theory of the sample distribution [10, 15] that it is unnecessary to take into account the sampling effects attributed to the covariates (but not to the terms depending on the outcome) when estimating the parameters of the population distribution $\mathcal{P}_1(\mathbf{x}_i; \boldsymbol{\beta})$. By dividing the rescaled weights by the estimated sample expectations $\hat{E}_S(\tilde{w}_i|\mathbf{z}_i)$, we diminish their dependence on the sampling effects attributed to the covariates in $\mathbf{z}_i$ and often further reduce their variability. As a result, the efficiency of the estimators for the logistic regression coefficients of interest is increased.

Estimation of the model-adjusted weights $\tilde{w}_i$ requires modeling the sample expectation of the rescaled $\tilde{w}_i$ given the covariates $\mathbf{z}_i$. $\boldsymbol{\theta}$ is a vector of the coefficients in the regression model relating of $\tilde{w}_i$ on $\mathbf{z}_i$. In contrast, the use of the estimating equations (6) requires modeling and estimating the sample expectations of w_i given $\mathbf{z}_i$ as a function of the unknown population parameter $\boldsymbol{\beta}$. For this to be achieved, the sample expectations of w_i given y_i and $\mathbf{v}_i$ should be estimated first, and $\boldsymbol{\gamma}$ is the regression coefficient in a regression model relating of w_i on y_i and $\mathbf{v}_i$. In both cases, the only information required to estimate the expectations is the information on observed variables for the units in the sample.

The SPW estimator is a version of a semiparametric approach proposed in [15], which is extended in this paper to the case–control framework using the rescaled rather than original sample weights. The theoretical result presented in Section 4.1 establishes the consistency of the proposed estimator and its robustness to misspecification of the model fitted to the rescaled weights.

4.1. Theoretical properties of semiparametric weighted estimator

Let $\hat{\boldsymbol{\mu}}_l(\boldsymbol{\beta})$ define a vector of sample means that are consistent estimators of the means $\boldsymbol{E}_l\left[\dfrac{\mathbf{x}}{g(\mathbf{z})}\mathcal{P}_{1-l}(\mathbf{x}; \boldsymbol{\beta})\right]$, where $\boldsymbol{E}_l(\cdot)$ stands for the vector of expectations with respect to the conditional (population) distribution of $\mathbf{z}$ given $Y = l$ ($l = 0, 1$) and $g(\mathbf{z})$ is a strictly positive function from $\mathbb{R}_m$ to $\mathbb{R}_1$ (m is the length of $\mathbf{z}$). Consider the following system of estimating equations:

$$\hat{\boldsymbol{G}}(\boldsymbol{\beta}) = \lambda_1 \hat{\boldsymbol{\mu}}_1(\boldsymbol{\beta}) - \lambda_0 \hat{\boldsymbol{\mu}}_0(\boldsymbol{\beta}) = \mathbf{0}, \tag{9}$$

where λ_1 and λ_0 converge to some positive constants C_1 and C_0, respectively (numerically or in probability) as N, n_1 and n_0 go to infinity. Denote the solution of (9) by $\hat{\boldsymbol{\beta}}$. Then, the following result holds (we provide proof in Appendix C).

Result 1

If $N \to \infty$, $n_1 \to \infty$ and $n_0 \to \infty$ as described previously, then

(a) $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\boldsymbol{\beta}}_1{}')'$ converges in probability to $\tilde{\mathbf{b}} = (\tilde{b}_0, \tilde{\mathbf{b}}_1{}')'$, the solution of

$$C_1 \boldsymbol{E}_1\left[\frac{\mathbf{x}}{g(\mathbf{z})}\mathcal{P}_0(\mathbf{x}; \mathbf{b})\right] = C_0 \boldsymbol{E}_0\left[\frac{\mathbf{x}}{g(\mathbf{z})}\mathcal{P}_1(\mathbf{x}; \mathbf{b})\right], \tag{10}$$

and

(b) if the logistic model (1) is valid, $\tilde{\mathbf{b}}_1 = \boldsymbol{\beta}_1$ and $\tilde{b}_0 = \beta_0 + b_C$, where $b_C = \log \dfrac{C_1 Pr(Y = 0)}{C_0 Pr(Y = 1)}$.

Let $\hat{\boldsymbol{\beta}}_{\mathrm{SPW}} = (\hat{\beta}_{0,\mathrm{SPW}}, \hat{\boldsymbol{\beta}}'_{1,\mathrm{SPW}})'$. The consistency of the estimator $\hat{\boldsymbol{\beta}}_{1,\mathrm{SPW}}$ for $\boldsymbol{\beta}_1$ follows from Result 1. Indeed, by defining $\hat{\boldsymbol{\mu}}_l(\boldsymbol{\beta}) = \sum_{i \in S_l} w_i \mathbf{x}_i / g(\mathbf{z}_i) \mathcal{P}_{1-l}(\mathbf{x}_i; \boldsymbol{\beta}) / N_l$ $(l = 0, 1)$ and $g(\mathbf{z}_i) = E_S(\widetilde{w}_i | \mathbf{z}_i)$, the system of the weighted estimating equations (8) can be written in a form (9), that is,

$$\frac{n_1 N_1}{n \sum_{i \in S_1} w_i} \hat{\boldsymbol{\mu}}_1(\boldsymbol{\beta}) - \frac{n_0 N_0}{n \sum_{i \in S_0} w_i} \hat{\boldsymbol{\mu}}_0(\boldsymbol{\beta}) = \mathbf{0}.$$

Hence, Result 1 applies to $\hat{\boldsymbol{\beta}}_{1,\mathrm{SPW}}$ under suitable conditions on the rate of convergence of the ratios n_1/n and n_0/n to a positive constants. The estimator $\hat{\beta}_{0,\mathrm{SPW}}$ is not consistent for β_0. However, a consistent estimator for β_0 can be recovered in some cases with the use of the relationships in Result 1(b). The robustness of the estimator $\hat{\boldsymbol{\beta}}_{\mathrm{SPW}}$ to misspecification of the model assumed for $E_S(\widetilde{w}|\mathbf{z})$ follows from Result 1 because the solution to the asymptotic estimating equations is the population parameter $\boldsymbol{\beta}$ regardless of the form assumed for the function $g(\mathbf{z})$, which represents the modification applied to the weights in our setup. We anticipate, however, that model misspecification may have an impact on the efficiency of the estimator.

5. Variance estimation

In this section, we address variance estimation of the estimators proposed in the previous sections. In Section 6.4, we use the simulations to investigate empirically the finite sample properties of the variance estimators described in this subsection.

5.1. Variance of the sample pseudo likelihood estimator

Pfeffermann and Sverchkov [18, Section 12.4, Equation (12.22)] suggested using the inverse of the information matrix evaluated at $\hat{\boldsymbol{\beta}}_{\mathrm{SPL}}$ to estimate the covariance matrix of the SPL estimator. This variance estimator does not account for the fact that $\boldsymbol{\gamma}$ was estimated. Yuan and Jennrich [19, Equation (2.5)] gave a formula for the asymptotic variance of a general pseudo maximum likelihood estimator. This formula contains a correction term that takes into account the variability of $\hat{\boldsymbol{\gamma}}$. However, because the variance estimates obtained from the inverse of the information matrix have sufficient accuracy for the models we studied in our simulation studies (Section 6.4), we do not use the correction term in this paper.

The use of the inverse of the information matrix is inappropriate for the data obtained by cluster sampling designs because of ICC of the covariates from cluster sampling. Therefore, for cluster sampling designs, we recommend using robust or design-based variance estimation used in survey sampling but without sample weights [9] as it has good precision in the simulations (Section 6.4).

5.2. Variance of the semiparametric weighted estimator

We can estimate the variance of the SPW estimator described in the previous section with the use of the Taylor linearization method for variance estimation [9]. This method uses the delta method to approximate a nonlinear estimator by an estimator of a total. Standard analytical formulas are then used to estimate the standard error of this total depending on the sample design and the resulting variance estimates are consistent (see [20] for asymptotic results). We can use software packages designed for analysis of the binary data from surveys with complex sample designs to calculate the estimated covariance matrix of the weighted estimators using this method, for example, the R [21] package SURVEY [22].

It is important to emphasize that the design-based standard errors estimated in this way do not account for determination of the form of the expectations, $E_S(\tilde{w}|\mathbf{z}; \boldsymbol{\theta})$, imbedded in the denominator of the SPW estimator or estimation of the parameters $\boldsymbol{\theta}$ indexing them. A replication method such as delete-one jackknife, bootstrap or balanced half-sample repeated replication [9, 20] may be preferable as they can account for these additional sources of variation by re-estimating the weights at each step. We discuss the application of the Taylor linearization method and delete-one jackknife in Sections 6.4 and 7.

6. Simulation study

We used simulations to evaluate the methods described in Sections 3–5 and compared them with other weighted and unweighted (naive) estimators.

6.1. Data generation

We generated an artificial population of $N = 700,000$ units to approximate a population of a moderate size city in the USA. For this purpose, we obtained four covariates, high blood pressure (HBPR), age, gender, and race (denoted by x_1, x_2, x_3, x_4), from the KCS data (Section 7) and duplicated these as necessary to achieve the desired population size. x_5 is an artificial continuous variable generated from a normal distribution having different means and standard errors for the different levels of the variable x_2 such that the correlation between x_2 and x_5 is about 0.2. Further, we grouped the population data in $k = 3500$ clusters with $m = 200$ observations per cluster. We obtained the clusters by sorting the data preliminarily by the values of an artificially generated continuous covariate x_6, which is correlated with the variable x_2. The correlation between x_2 and x_6 induces ICC for the variable x_2 (for the clusters). In the current example, we set $\text{corr}(x_2, x_6) = 0.55$, such that ICC for x_2 approximately equals 0.3 (the resulting ICC for x_5 is negligible).

The population logistic regression model for $Y|\mathbf{x}$ is defined by (1), where $\mathbf{x}' = (x_1, x_2, x_3, x_4, x_5)$ and $\boldsymbol{\beta} = (-10.5, 1, 0.5, 1, 1, 1)$. The choice of the true values of the target parameter results in approximately 1000 cases in the population, approximating the number of the cases in the KCS. The Y-values were regenerated for each simulation, whereas the $\mathbf{x}$-values in the population are held fixed.

Sample plan A: Poisson sampling. We sample the subjects with the use of Poisson sampling with probabilities $\pi_i = 1/w_i$, where

$$w_i = \exp(\gamma_0 + \gamma_1 \cdot y + \gamma_2 \cdot y \cdot x_2 + \gamma_3 \cdot x_2 + \gamma_4 \cdot y \cdot x_3 + \gamma_5 \cdot x_3 + \gamma_6 \cdot y \cdot x_4 + \gamma_7 \cdot x_4 + u_i), \quad (11)$$

$u_i \sim U(0, 1/5)$ and $\boldsymbol{\gamma} = (\gamma_0, \ldots, \gamma_7)$. We chose two sets of values for the parameter $\boldsymbol{\gamma}$ that generate two different sets of weights with different amount of variation: moderate-to-high variation (hereafter, plan A(M)) and extremely high variation (hereafter, plan A(E)). In both cases, the generated weights fulfill two requirements: (1) the resulting simulated sample consists of almost all cases in the population and the same number of controls (approximately 1000 in each group on average); and (2) the selected controls match (approximately) in distribution to the cases on the variables x_2, x_3, x_4. Both versions of plan A can be viewed as an approximation of a stratified sample with two strata (cases and controls), because the number of sampled units from each stratum are held approximately fixed by the appropriate choice of the values of $\boldsymbol{\gamma}$. This sample design does not use the cluster structure of the population data.

Sample plan B: Cluster sampling. We sampled all cases with certainty. We obtain the controls by two-stage cluster sampling. At the first stage, we sample 35 clusters as a simple random sample from 3500 clusters in the population. At the second stage, we sample the units within the clusters with probabilities $1/w_i$, where w_i are defined by the equation $w_i = \exp(\gamma_0 + \gamma_1 \cdot x_{2i} + u_i)$, $u_i \sim U(0, 1/5)$, and γ_0, γ_1 are such that (1) approximately 30 controls are sampled from each cluster in average (yielding approximately 1000 sampled controls); and (2) the sampled controls are frequency matched to the cases on x_2.

6.2. Estimators of regression coefficients

We estimated the regression coefficients of the population model by the following five estimators:

(1) The unweighted (naive) estimator (U), $\hat{\boldsymbol{\beta}}_{\text{U}}$, defined as a solution of (3) with $w_i = 1$ for all $i \in S$;
(2) The SPL, $\hat{\boldsymbol{\beta}}_{\text{SPL}}$, defined in Section 3;
(3) The weighted (Horvitz–Thompson) estimator (W), $\hat{\boldsymbol{\beta}}_{\text{W}}$, defined in Section 2;
(4) The rescaled weighted estimator (RSW), $\hat{\boldsymbol{\beta}}_{\text{RSW}}$, defined as a solution of (3) with w_i replaced by the rescaled weights $\tilde{w}_i$ as defined in Section 4 [7];
(5) The SPW estimator, $\hat{\boldsymbol{\beta}}_{\text{SPW}}$, defined in Section 4.

We estimated the expectations $E_S(w|\mathbf{z}; \boldsymbol{\beta}; \hat{\boldsymbol{\gamma}})$ in (6) with the use of the algorithm outlined in Appendix A. First, we estimated $\hat{\boldsymbol{\gamma}}$ (steps 1(a)–(b)). For plan A, we set $\mathbf{v} = (x_2, x_3, x_4)$ and fitted the model defined by (11). For plan B, we fitted the linear regression model to $\log w$ with variables y, $y \times x_2$ and x_2. Next, we computed the expectations $E_S(w|\mathbf{z}; \boldsymbol{\beta}; \hat{\boldsymbol{\gamma}})$ (steps 1(c), 2, and 3). We estimate the expectations $E_S(\tilde{w}|\mathbf{z}; \boldsymbol{\theta})$ in the denominator of the model-adjusted weights $\tilde{w}_i^m$ by fitting the linear regression model to $\log \tilde{w}$ with covariates $\mathbf{z} = (x_1, \ldots, x_5)$ for plans A and B.

We computed the coefficient of variation (CV) of the original (w_i), rescaled ($\tilde{w}_i$), and model-adjusted weights ($\tilde{w}_i^m$) under three simulations plans. Even though $\tilde{w}_i$ have smaller variability than w_i in all plans, there is still a sizeable amount of variability in $\tilde{w}_i$. However, after adjusting the rescaled

weights by modeling, $\tilde{w}_i^m$ achieve an additional 70% and 50% reduction of variability under plans A and B, respectively.

We computed all the estimators, except SPL, with the use of the R [21] function svyglm in the package SURVEY [22]. For the SPL estimator, we used the R [21] function BBsolve in the package BB [23] to solve the estimating equations (6) (with $\hat{\boldsymbol{\gamma}}$ substituted in place of $\boldsymbol{\gamma}$). We repeated the simulations 1000 times for each of the sample plans. We computed the simulated bias for the regression coefficients as the average of the difference of the estimated coefficient minus the true coefficient. We computed the simulated standard deviation for a regression coefficient as the square root of the sample variance of the estimated coefficients. We computed the design-based standard error for a regression coefficient as the mean of the Taylor linearization estimated standard errors over the repeated simulations. We used the R [21] function svyglm in the package SURVEY [22] to compute the Taylor linearization standard errors.

6.3. Simulations results

Table I shows the biases and simulated standard deviations of the five estimators defined in Section 6.2 under plans A(M), A(E), and B, respectively. The results in Table I show that the SPW estimator has greater efficiency compared with the RSW estimator in all sample plans. Under plan A(E) with the largest sample weights variability, the simulated standard deviation of the SPW estimator is about 50% lower than for the Horvitz–Thompson weighted estimator for all components of $\boldsymbol{\beta}$. Moreover, the SPW estimator has smaller bias than all the other weighted estimators. In general, the SPW estimator is the

Table I. Simulated bias and standard deviation (in parentheses) under sample plans A(M), A(E), and B, over 1000 simulations.

Method	Plan	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$
U	A(M)	−11.544	−0.003	1.700	2.414	1.706	−0.004
		(0.259)	(0.135)	(0.121)	(0.147)	(0.113)	(0.058)
	A(E)	−12.524	−0.002	1.608	3.416	2.003	−0.004
		(0.294)	(0.158)	(0.121)	(0.212)	(0.141)	(0.049)
	B	−7.895	−0.001	1.089	−0.006	−0.004	−0.004
		(0.234)	(0.146)	(0.149)	(0.120)	(0.122)	(0.047)
SPL	A(M)	−0.005	−0.003	0.000	0.013	0.006	−0.004
		(0.259)	(0.135)	(0.122)	(0.147)	(0.113)	(0.058)
	A(E)	−0.025	−0.002	0.008	0.016	0.003	−0.004
		(0.294)	(0.158)	(0.122)	(0.213)	(0.141)	(0.049)
	B	0.006	−0.001	−0.001	−0.006	−0.004	−0.004
		(0.234)	(0.146)	(0.150)	(0.120)	(0.122)	(0.047)
W	A(M)	−0.050	−0.028	0.050	0.060	0.032	−0.039
		(0.378)	(0.186)	(0.198)	(0.217)	(0.172)	(0.105)
	A(E)	−0.253	−0.108	0.168	0.347	0.130	−0.123
		(0.716)	(0.420)	(0.325)	(0.542)	(0.381)	(0.153)
	B	0.108	−0.035	0.049	−0.033	−0.040	−0.055
		(0.436)	(0.251)	(0.225)	(0.219)	(0.214)	(0.107)
RSW	A(M)	−6.465	−0.018	0.039	0.041	0.020	−0.022
		(0.313)	(0.203)	(0.202)	(0.195)	(0.183)	(0.101)
	A(E)	−6.586	−0.038	0.110	0.143	0.065	−0.045
		(0.493)	(0.349)	(0.260)	(0.371)	(0.314)	(0.117)
	B	−6.441	−0.004	0.005	−0.010	−0.008	−0.007
		(0.212)	(0.163)	(0.155)	(0.130)	(0.132)	(0.053)
SPW	A(M)	−6.417	−0.002	0.005	0.018	0.007	−0.008
		(0.262)	(0.149)	(0.132)	(0.151)	(0.124)	(0.070)
	A(E)	−6.408	−0.014	0.025	0.039	0.014	−0.010
		(0.332)	(0.192)	(0.148)	(0.232)	(0.179)	(0.070)
	B	−6.445	−0.001	0.002	−0.006	−0.006	−0.004
		(0.210)	(0.150)	(0.151)	(0.122)	(0.125)	(0.049)

U, unweighted; SPL, sample pseudo likelihood; W, weighted; RSW, rescaled weighted; SPW, semiparametric weighted.

Table II. Simulated bias and standard deviation (in parentheses) of sample pseudo likelihood (SPL) and semiparametric weighted (SPW) estimators under sample plan A(E), over 1000 simulations.

Method	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$
SPL	−3.267 (0.330)	−0.002 (0.158)	0.006 (0.133)	3.416 (0.212)	−0.032 (0.159)	−0.004 (0.049)
SPW	−6.429 (0.377)	−0.017 (0.242)	0.036 (0.196)	0.049 (0.231)	0.017 (0.241)	−0.016 (0.087)

closest to the SPL estimator in efficiency and bias characteristics. As is anticipated, the SPL estimator has the best properties (consistency of the intercept, smallest bias, and highest efficiency) and is the best choice if the parametric model for mean and selection process can be accurately specified.

We also studied the SPL and SPW estimators under a misspecified selection model. For this purpose, we computed the SPL and SPW estimators from the data generated by plan A(E) but purposefully fitted the wrong selection model. Namely, we omitted the covariate x_3 and set $\mathbf{v} = (x_2, x_4)$ and $\mathbf{z} = (x_1, x_2, x_4, x_5)$ in estimation of the sample expectations used in model-adjusted weights. Table II presents the biases and simulated standard deviations of the SPL and SPW estimators under the misspecified model. We obtained the corresponding SPL and SPW estimates in Table I (plan A(E)) and Table II from exactly the same generated data sets. We can conclude from Table II that when the selection model is misspecified, the SPL estimator can be biased. In contrast, the SPW estimator is robust to this model misspecification, and the loss of efficiency under this misspecified model is negligible when compared with the results under the correct model.

6.4. Simulated versus estimated variances

We studied the performance of the variance estimators (of the weighted estimators and SPL estimator) described in Section 5.

Weighted estimators. We compared the design-based standard errors and the simulated standard deviations from the same sets of simulations for the weighted estimators (W, RSW, and SPW) under sample plans A and B. For purposes of variance estimation, we approximated the sampling design in plan A by a stratified sampling design with two strata (cases and controls). For variance estimation under plan B, we approximated the sampling design by stratified single stage cluster sample with replacement [6, pp. 19–20]. Table III presents the ratios of the simulated standard deviations divided by the design-based standard errors. These ratios are very close to 1 for the proposed estimator $\hat{\boldsymbol{\beta}}_{\mathrm{SPW}}$ under all three sample plans. However, significant differences exist between the two variance estimates (the ratios are much larger than 1) for the other weighted estimators. We observe particularly sizable differences under plan A(E), where the ratios between the two variance estimates are as large as 1.5 for all components of $\hat{\boldsymbol{\beta}}_{\mathrm{W}}$ and as large as 1.4 for all components of $\hat{\boldsymbol{\beta}}_{\mathrm{RSW}}$. Li *et al.* [8] reported a similar finding for $\hat{\boldsymbol{\beta}}_{\mathrm{W}}$. This finding shows that the design-based variances appear to perform poorly (for the conventionally used

Table III. Ratios of the simulated standard deviation to the design-based standard errors estimates (using the Taylor linearization method) under sample plans A(M), A(E), and B, over 1000 simulations (R_i, $i = 0, \ldots, 5$ is the ratio for the component i of $\hat{\boldsymbol{\beta}}$).

	Plan A(E)			Plan A(M)			Plan B		
	W	RSW	SPW	W	RSW	SPW	W	RSW	SPW
R_0	1.39	1.01	0.93	1.14	0.89	0.99	1.18	0.87	0.86
R_1	1.37	1.22	1.00	1.06	1.06	1.02	1.17	1.07	1.05
R_2	1.49	1.36	1.06	1.17	1.19	1.04	1.13	1.00	0.99
R_3	1.48	1.24	1.00	1.11	1.05	1.01	1.15	0.98	0.98
R_4	1.38	1.23	1.03	1.07	1.08	1.00	1.12	0.98	1.01
R_5	1.54	1.29	1.07	1.19	1.13	1.02	1.28	1.00	1.00

W, weighted; RSW, rescaled weighted; SPW, semiparametric weighted.

Table IV. Ratios of the simulated standard deviation to the estimated standard errors (using the inverse of the information matrix) under sample plans A(M), A(E), and B, over 1000 simulations (R_i, $i = 0, \ldots, 5$ is the ratio for the component i of $\hat{\boldsymbol{\beta}}$).

	Plan A(E)		Plan A(M)		Plan B	
	U	SPL	U	SPL	U	SPL
R_0	1.02	1.02	1.02	1.02	1.17	1.17
R_1	1.01	1.01	1.01	1.01	0.97	0.97
R_2	0.98	0.98	1.01	1.02	1.58	1.58
R_3	1.06	1.06	1.03	1.03	1.03	1.03
R_4	0.99	0.99	0.99	0.99	0.98	0.98
R_5	0.98	0.98	0.97	0.97	1.06	1.06

U, unweighted; SPL, sample pseudo likelihood.

weighted estimator) in finite samples with highly dispersed weights, although they are known to be consistent [14]; notice better results under plan A(M). Awareness of this problem is needed when analyzing real data.

Likelihood-based estimators. Table IV presents the ratios of the simulated standard deviations and the estimated standard errors (defined as a mean of the estimates of the standard errors computed using the inverse information matrix) from the same sets of simulations the unweighted (U) and SPL estimators under sample plans A and B. The two variance estimates are very close for both $\hat{\boldsymbol{\beta}}_{\mathrm{U}}$ and $\hat{\boldsymbol{\beta}}_{\mathrm{SPL}}$ under plans A(E) and A(M), indicating that the estimates obtained from the inverse information matrix capture the variability of the SPL estimator properly and that the additional variability caused from the estimation of $\boldsymbol{\gamma}$ in the SPL can be ignored for the models we studied. However, under plan B, the standard errors estimated using the information matrix seriously underestimate the simulated standard deviation for the intercept and the coefficient of x_2 (the ratios of the two estimates are 1.17 and 1.58, respectively). This occurs because of the high ICC in x_2. As mentioned at the end of Section 3, the sample observations $Y_i|\mathbf{z}_i$ are approximately independent, which validates the use of the SPL estimator in this case, but the variance estimator obtained using the inverse information matrix is inappropriate and should be replaced by a robust variance estimator. The robust standard errors estimates (we used Taylor linearization) are found to be very close to the simulated standard deviations for all the components of $\hat{\boldsymbol{\beta}}_{\mathrm{SPL}}$ (the ratios of the simulated standard deviations and robust standard errors estimates are 1.03, 1.00, 1.02, 1.02, 0.99, and 1.04 for the intercept and five coefficients, respectively).

7. Data example

The KCS is a population-based case–control study designed to investigate risk factors associated with kidney cancer in blacks and whites 20–79 years old [5]. We analyzed the data from the Detroit part of KCS. There are 1018 cases and 1038 controls for which we study the risk factor of high blood pressure on the kidney cancer incidence.

Kidney cancer cases were identified from the population-based cancer registry, the U.S. Surveillance, Epidemiology and End Results Program. In Detroit, all black cases were selected with certainty. White cases were either subsampled or selected with certainty depending on the period of accrual. Controls 20–64 years old were sampled from listings provided by Michigan's Department of Motor Vehicles, which also includes identification cards (for nondrivers). Controls aged 65–79 years were randomly sampled from the Medicaid and Medicare Beneficiaries database. All controls were (approximately) frequency matched to the cases on age, gender, and race.

The sampled cases and controls were assigned sample weights w_i adjusted for nonresponse and calibrated using poststratification to the population totals for age, race, and gender. Despite the calibration adjustment, which can decrease the variability of the weights and, as a result, increase the efficiency of weighted estimation, the calibrated weights in KCS data were still highly variable within the control group and also between the cases and the controls. The range of the weights in the sample was 1.4 to 39,830 with CV of 260%.

We fit the logistic regression model (1) to the data with the **x** vector consisting of the following variables ($I(\cdot)$ defines a dummy indicator variable): HBPR$=I$(High Blood Pressure); Age1$=I(20 \leqslant Age < 45)$; Age2$=I(45 \leqslant Age < 55)$; Age3$=I(55 \leqslant Age < 65)$; Age4$=I(65 \leqslant Age < 75)$; Gender$=I$(Male); Race$=I$(White); Smoke1$=I$(Never smoked); Smoke2$=I$(Occasional smoker); Smoke3$=I$(Former smoker); EDUC has four ordinal values (0, 1, 2, 3) corresponding to ($\leqslant$11 years HS, 12 years HS, 1–3 years college, and 3+ years college), and body mass index (BMI) has four ordinal values (0, 1, 2, 3) corresponding to (<25, 25–30, 30–35, and 35+ kg/m^2).

We estimated the coefficients of the model (1) with the use of five different estimators as defined in Section 6.2. For the SPL estimator, we estimated the expectations $E_S(w|\mathbf{z};\boldsymbol{\beta};\hat{\boldsymbol{\gamma}})$ with the use of the algorithm described in Appendix A. In step 1(a) of the algorithm, we fitted the linear regression model of $\log w$ on y, covariates Age1, Age2, Age3, Age 4, Gender, Race, and the cross products of these covariates with y. This model approximates the selection process of the controls that was based on age, gender, and race. For the SPW estimator, we estimated the expectations $E_S(\tilde{w}|\mathbf{z};\boldsymbol{\theta})$ by fitting the linear regression model of $\log \tilde{w}$ on the covariates **x**.

We obtained the standard errors of the three weighted estimators (W, RSW, and SPW) by a stratified delete-one jackknife [6, pp. 29–31] with $H = 21$ strata ($2 \times 2 \times 5$ strata of gender, race, and age within the controls and the stratum of the cases). We also obtained standard error estimates for W, RSW, and SPW with the use of the Taylor linearization method (not shown), and we found very close correspondence between the two variance estimates for all components of $\hat{\boldsymbol{\beta}}$. We obtained the standard errors for U and SPL with the use of the inverse information matrix evaluated at the estimated values of each estimate correspondently. Table V summarizes model estimates and their estimated standard errors.

Table V. The estimates and the estimated standard errors (in parentheses) of the logistic regression coefficients from the analysis of the U.S. Kidney Cancer Case-Control Study data using different estimation methods.

Variable	U	SPL	W	RSW	SPW
HBPR	0.823	0.823	0.869	0.923	0.889
	(0.103)	(0.103)	(0.123)	(0.128)	(0.117)
BMI	0.255	0.255	0.264	0.296	0.275
	(0.048)	(0.048)	(0.055)	(0.058)	(0.052)
EDUC	−0.177	−0.177	−0.149	−0.222	−0.180
	(0.050)	(0.050)	(0.056)	(0.058)	(0.052)
Never smoked	0.153	0.153	0.176	0.086	0.161
	(0.141)	(0.141)	(0.158)	(0.167)	(0.157)
Occasional smoker	0.244	0.244	0.499	0.285	0.320
	(0.248)	(0.248)	(0.271)	(0.274)	(0.254)
Former smoker	0.321	0.321	0.304	0.325	0.356
	(0.137)	(0.137)	(0.162)	(0.181)	(0.160)
Age: [20,45)	0.533	−1.589	−1.744	−1.762	−1.917
	(0.212)	(0.212)	(0.204)	(0.193)	(0.204)
Age: [45,55)	0.487	−0.611	−0.578	−0.674	−0.739
	(0.193)	(0.193)	(0.175)	(0.160)	(0.170)
Age: [55,65)	0.334	−0.054	−0.085	−0.106	−0.136
	(0.184)	(0.184)	(0.152)	(0.149)	(0.160)
Age: [65,75)	0.013	0.213	0.255	0.161	0.137
	(0.183)	(0.183)	(0.149)	(0.147)	(0.159)
Gender	0.022	0.618	0.522	0.485	0.613
	(0.096)	(0.096)	(0.082)	(0.081)	(0.077)
Race	0.843	0.053	0.083	0.233	0.116
	(0.104)	(0.104)	(0.098)	(0.095)	(0.089)
Intercept	−1.531	−7.454	−7.561	−0.349	−0.358
	(0.232)	(0.232)	(0.230)	(0.231)	(0.231)

U, unweighted; SPL, sample pseudo likelihood; W, weighted; RSW, rescaled weighted; SPW, semiparametric weighted; HBPR, high blood pressure; BMI, body mass index.

As is expected, the unweighted estimates of the regression coefficients of age, sex, and race differ from the estimates that account for the sample weights, because the sample inclusion probabilities depend on these variables and differ between cases and controls, that is, are informative. We obtain the unbiased estimates of the intercept with the use of weighted (W) or SPL estimation methods only. All other estimates are very similar among the different methods and confirm that HBPR and BMI are risk factors for kidney cancer. Although the standard errors for the SPW estimates are usually smaller than for the RSW estimates, the standard errors for all five methods are similar for most of the regression coefficient estimates. Notice, however, that on the basis of the simulation results, we might expect that the standard errors of W and RSW are underestimated.

8. Discussion

In this paper, we propose two estimators for logistic regression coefficients for analyses of population-based case–control studies with complex sample designs: an SPL estimator and an SPW estimator. Both estimators are derived from the survey sampling literature for cross-sectional designs and are new in the case–control framework. The estimators proposed provide an alternative to the conventionally used weighted (Horvitz–Thompson) estimator that can be very inefficient when applied to data with highly variable weights as often is the case with the case–control studies.

The two proposed estimators show significant improvement in efficiency and can be readily applied to the real data. The SPW estimator and its standard errors can be computed using existing software for complex survey designs. Therefore, its implementation is straightforward and requires only knowledge of the individual values of the weights and covariates for the cases and controls observed in the sample. We show the SPL estimator to have attractive finite sample properties (Section 6) but may suffer from two potential drawbacks. First, the validity of the SPL estimator is based on the 'independence' result (Section 3), and hence its implementation to cluster sample designs is not straightforward. Second, the SPL estimator can be sensitive to the misspecification of the selection model (Section 6.3), and hence it may not be appropriate for the analyses in which the selection model cannot be specified and estimated accurately. Also, unlike the SPW estimator, computing the SPL estimator requires additional programming to incorporate the selection model into the estimating equations.

In survey research, regression and poststratification calibration of the sample weights are used in weighted estimation [6, 12] to reduce the variance and bias (such as from deficient coverage of the sample frame) of the weighted estimators. Both types of calibration require knowledge of population proportions or means of design variables that are often available from population census data. In the KCS, the sample weights we used were calibrated by poststratifying to age, gender, and race categories. As we show, despite the calibration, the weights are still very variable and the weighted estimators using these weights are substantially less efficient than using the model adjusted weights proposed in this paper. For match case–control studies, the rescaled sample weights can be poststratified to the case distribution of the matching variables (e.g., age, gender, and race categories for the KCS) in order to maintain the matching in the weighted analyses [8]. These poststratified weights had modest efficiency gains over using only rescaled weights. The model adjusted weights that are used in the proposed SPW estimator are applicable to either unmatched or matched case–control studies, and our simulations show that the SPW estimator has much larger efficiency over the estimator using the rescaled sample weights.

To summarize, we recommend using the SPW estimator because our simulations show it to be nearly as efficient as the SPL estimator while being robust to misspecification of the selection model and applicable to multistage cluster designs. We recommend that the choice of the adjusted sample weights provided for analyses of a case–control study should be made without regard to any specific analysis but according to a general criteria of efficiency such as the CV of the weights. This will avoid choosing adjusted sample weights that provide the 'desired' results for a particular analysis. Further empirical research would be useful to investigate the finite sample properties of these estimators under other sample designs and types of sample weighting. Also, it would be useful to extend the SPL and SPW estimations to other types of studies such case–cohort designs and to case–control studies with multiple types of case series.

Appendix A. Estimation of $E_S(w_i|\mathbf{z}_i;\hat{\boldsymbol{\gamma}},\boldsymbol{\beta})$

We can estimate $E_S(w_i|\mathbf{z}_i;\hat{\boldsymbol{\gamma}},\boldsymbol{\beta})$ from the sample data by the following three-step procedure [18] adopted to the binary response setting.

(1)(a) Identify (using information available about sampling design) the regression model for $E_S(w_i|y_i,\mathbf{v}_i;\boldsymbol{\gamma})$ with $y_i,\mathbf{v}_i$ and their cross products as independent variables and $\boldsymbol{\gamma}$ as coefficients.
(b) Estimate $\hat{\boldsymbol{\gamma}}$ and compute $\hat{E}_S(w_i|y_i,\mathbf{v}_i)=E_S(w_i|y_i,\mathbf{v}_i;\hat{\boldsymbol{\gamma}})$.
(c) Denote $L_0(\hat{\boldsymbol{\gamma}})=\hat{E}_P(\pi_i|Y_i=0,\mathbf{v}_i)=1/\hat{E}_S(w_i|Y_i=0,\mathbf{v}_i)$ and $L_1(\hat{\boldsymbol{\gamma}})=\hat{E}_P(\pi_i|Y_i=1,\mathbf{v}_i)=1/\hat{E}_S(w_i|Y_i=1,\mathbf{v}_i)$.
(2) Obtain $E_P(\pi_i|\mathbf{z}_i;\hat{\boldsymbol{\gamma}},\boldsymbol{\beta})=L_0(\hat{\boldsymbol{\gamma}})\Pr(Y_i=0|\mathbf{x}_i;\boldsymbol{\beta})+L_1(\hat{\boldsymbol{\gamma}})\Pr(Y_i=1|\mathbf{x}_i;\boldsymbol{\beta})$ as a function of $\boldsymbol{\beta}$.
(3) Compute $\hat{E}_S(w_i|\mathbf{z}_i;\boldsymbol{\beta})=E_S(w_i|\mathbf{z}_i;\hat{\boldsymbol{\gamma}},\boldsymbol{\beta})=[E_P(\pi_i|\mathbf{z}_i;\hat{\boldsymbol{\gamma}},\boldsymbol{\beta})]^{-1}$.

Appendix B. Breslow and Cain estimator [13]

We show that Equation (2) in [13] is, in fact, another form of writing the sample distribution defined by (4).

Define: Y_i is a binary response [$Y_i=l$ denotes that unit i is a case ($l=1$) or a control ($l=0$)]; v_i is an ordinal variable taking J values ($j=1,\ldots,J$) denoting the population stratum from which unit i was sampled; $\mathbf{x}_i$ is a vector of covariates; n_{lj} is the size of the subsample drawn from the cell (l,j) ($l=0,1;\ j=1,2,\ldots,J$), $\sum_l\sum_j n_{lj}=n$, the total sample size. We also assume that $\Pr(i\in S|Y_i=l,v_i=j,\mathbf{x}_i)=\Pr(i\in S|Y_i=l,v_i=j)$ and $\Pr(Y_i=l|v_i,\mathbf{x}_i)=\Pr(Y_i=l|\mathbf{x}_i)$.

Using the notation and assumptions mentioned previously, we can write the quantity $R_{lji}(\mathbf{x})$ in the left-hand side of Equation (2) in [11] as

$$R_{lji}(\mathbf{x})=f_S(Y_i=l|v_i=j,\mathbf{x}_i)=\frac{\Pr(i\in S|Y_i=l,v_i=j)\Pr(Y_i=l|\mathbf{x}_i)}{\Pr(i\in S|v_i=j,\mathbf{x}_i)},\tag{B.1}$$

where the last equality follows from applying the Bayes rule.

Now, $\Pr(i\in S|Y_i=l,v_i=j)=\dfrac{(n_{lj}/n)\Pr(i\in S)}{Q_{lj}}$, where $Q_{lj}=\Pr(v_i=j,Y_i=l)$ and $n_{lj}/n=\Pr(v_i=j,Y_i=l|i\in S)$ (from the stratified sampling design). By multiplying the numerator and the denominator of (B.1) by the quantity $f_j(\mathbf{x}_i)=\Pr(\mathbf{x}_i|v_i=j)$, we obtain

$$f_S(Y_i=l|v_i=j,\mathbf{x}_i)=\frac{(n_{ij}/n)\Pr(Y_i=l|\mathbf{x}_i)f_j(\mathbf{x}_i)/Q_{lj}}{(n_{1j}/n)\Pr(Y_i=1|\mathbf{x}_i)f_j(\mathbf{x}_i)/Q_{1j}+(n_{0j}/n)\Pr(Y_i=0|\mathbf{x}_i)f_j(\mathbf{x}_i)/Q_{0j}}.$$

Because $\Pr(\mathbf{x}_i|Y_i=l,v_i=j)=\Pr(Y_i=l|\mathbf{x}_i)f_j(\mathbf{x}_i)/Q_{lj}$ $(l=0,1)$, we can rewrite the last equation equivalently as

$$f_S(Y_i=l|v_i=j,\mathbf{x}_i)=\frac{(n_{ij}/n)\Pr(\mathbf{x}_i|Y_i=l,v_i=j)}{(n_{1j}/n)\Pr(\mathbf{x}_i|Y_i=1,v_i=j)+(n_{0j}/n)\Pr(\mathbf{x}_i|Y_i=0,v_i=j)}.\tag{B.2}$$

The right-hand side of (B.2) equals the right-hand side of Equation (2) in [13].

Appendix C. Proof of result 1

(a) Equation (9) converges to Equation (10) by the weak law of large numbers.
(b) Equation (10) is equivalent to the following equation between vectors of integrals:

$$\frac{C_1}{\Pr(Y=1)}\int\frac{\mathbf{x}}{g(\mathbf{z})}\mathcal{P}_0(\mathbf{x};\mathbf{b})\Pr(Y=1|\mathbf{x};\boldsymbol{\beta})f(\mathbf{z})\mathrm{d}\mathbf{z}=\frac{C_0}{\Pr(Y=0)}\int\frac{\mathbf{x}}{g(\mathbf{z})}\mathcal{P}_1(\mathbf{x};\mathbf{b})\Pr(Y=0|\mathbf{x};\boldsymbol{\beta})f(\mathbf{z})\mathrm{d}\mathbf{z}.\tag{C.1}$$

Write $\mathbf{z}=(\mathbf{x}',\mathbf{v}_1')'$, where $\mathbf{x}=(1,\mathbf{x}_1')'$, $\mathbf{v}_1$ contains all the variables not included in $\mathbf{x}_1$; $\mathbf{b}=(b_0,\mathbf{b}_1')'$, $\boldsymbol{\beta}=(\beta_0,\boldsymbol{\beta}_1')'$, and substitute the corresponding logistic functions in (C.1). Then, equation between the first elements of the vectors in Equation (C.1) becomes the following:

$$\frac{C_1}{\Pr(Y=1)}\int\frac{\exp(\beta_0+\mathbf{x}_1'\boldsymbol{\beta}_1)}{g(\mathbf{z})(1+\exp(\mathbf{x}'\mathbf{b}))(1+\exp(\mathbf{x}'\boldsymbol{\beta}))}f(\mathbf{z})\mathrm{d}\mathbf{z}=\frac{C_0}{\Pr(Y=0)}\int\frac{\exp(b_0+\mathbf{x}_1'\mathbf{b}_1)}{g(\mathbf{z})(1+\exp(\mathbf{x}'\mathbf{b}))(1+\exp(\mathbf{x}'\boldsymbol{\beta}))}f(\mathbf{z})\mathrm{d}\mathbf{z}.\tag{C.2}$$

Given that $f(\mathbf{z})$ is strictly positive, it follows from (C.2) that $\mathbf{b}_1=\boldsymbol{\beta}_1$ and $b_0=\beta_0+b_C$, where $b_C=\log\dfrac{C_1Pr(Y=0)}{C_0Pr(Y=1)}$.

References

1. Scott AJ, Wild CJ. Fitting logistic models under case-control or choice based sampling. *JRSS, B* 1986; **48**:170–182.
2. Graubard BI, Fears TR, Gail MH. Effects of cluster sampling on epidemiologic analysis in population-based case-control studies. *Biometrics* 1989; **45**:1053–1071.
3. Fears TR, Gail MH. Analysis of a two-stage case-control study with cluster sampling of controls: application to nonmelanoma skin cancer. *Biometrics* 2000; **56**:190–198.
4. Kalton G, Piesse A. Survey research methods in evaluation and case-control studies. *Statistics in Medicine* 2007; **26**:1675–1687.
5. Colt JS, Schwartz K, Graubard BI, *et al.* Hypertension and risk of renal cell carcinoma among white and black Americans. *Epidemiology* 2011; **22**(6):797–804. [Epub ahead of print].
6. Korn EL, Graubard BI. *Analysis of Health Surveys*. Wiley: New York, NY, 1999.
7. Scott AJ, Wild CJ. Population-based case-control studies. In *Handbook of Statistics: Sample Surveys: Inference and Analysis*, Vol. 29B, Pfefferman D, Rao CR (eds). Elsevier: Amsterdam, 2009; 431–453.
8. Li Y, Graubard BI, DiGaetano R. Weighting methods for population-based case-control studies with complex sampling. *JRSS(C)* 2011; **60**:165–185.
9. Rust KF, Rao JNK. Variance estimation for complex surveys using replication techniques. *Statistical Methods in Medical Research* 1996; **5**:283–310.
10. Pfeffermann D, Sverchkov M. Inference under informative sampling. In *Handbook of Statistics: Sample Surveys: Inference and Analysis*, Vol. 29B, Pfefferman D, Rao CR (eds). Elsevier: Amsterdam, 2009; 455–487.
11. Robins JM, Rotnitzky A, Zhao L-P. Estimation of regression coefficients when some regressors are not always observed. *Journal of American Statistical Association* 1994; **89**:846–866.
12. Lumley T, Shaw PA, Dai JY. Connections between survey calibration estimators and semiparametric models for incomplete data. *International Statistical Review* 2011; **97**(2):200–220.
13. Breslow NE, Cain KC. Logistic regression for two-stage case-control data. *Biometrika* 1988; **75**:11–20.
14. Binder D. On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review* 1981; **51**:279–292.
15. Pfeffermann D, Sverchkov M. Parametric and semi-parametric estimation of regression models fitted to survey data. *Sankhya, Series B* 1999; **61**:166–186.
16. Pfeffermann D, Krieger A, Rinott Y. Parametric distributions of complex survey data under informative probability sampling. *Statistica Sinica* 1981; **8**:1087–1114.
17. Gong G, Samaniego F. Pseudo maximum likelihood estimation: theory and applications. *The Annals of Statistics* 1981; **9**:861–869.
18. Pfeffermann D, Sverchkov M. Fitting generalized linear models under informative sampling. In *Analysis of Survey Data*, Skinner C, Chambers R (eds). Wiley: New York, 2003; 175–195.
19. Yuan K-H, Jennrich R. Estimating equations with nuisance parameters: theory and applications. *Annals of the Institute of Statistical Mathematics* 2000; **52**:343–350.
20. Krewski D, Rao JNK. Inference from stratified samples: properties of linearization, jackknife and balanced repeated replication methods. *The Annals of Statistics* 1981; **9**:1010–1019.
21. R Development Core Team. R: *A Language and Environment for Statistical Computing. R Foundation for Statistical Computing*: Vienna, Austria, 2010. ISBN 3-900051-07-0, http://www.R-project.org [Accessed on July 12, 2012].
22. Lumley T. *Complex Surveys: A Guide to Analysis Using R*. Wiley: Hoboken NJ, 2010.
23. Varadhan R, Gilbert P. An R Package for solving a large system of nonlinear equations and for optimizing a high-dimensional nonlinear objective function. *Journal of Statistical Software* 2009; **32**:1–24. http://www.jstatsoft.org/v32/i04/ [Accessed on July 12, 2012].