UW Biostatistics Working Paper Series

12-30-2010

Oracle and Multiple Robustness Properties of Survey Calibration Estimator in Missing Response Problem

Kwun Chuen Gary Chan
University of Washington, kcgchan@u.washington.edu

Suggested Citation
Chan, Kwun Chuen Gary, "Oracle and Multiple Robustness Properties of Survey Calibration Estimator in Missing Response Problem" (December 2010). *UW Biostatistics Working Paper Series.* Working Paper 368.
http://biostats.bepress.com/uwbiostat/paper368

This working paper is hosted by The Berkeley Electronic Press (bepress) and may not be commercially reproduced without the permission of the copyright holder.
Copyright © 2011 by the authors

Oracle and multiple robustness properties of survey calibration estimator in missing response problem

Kwun Chuen Gary Chan

Department of Biostatistics and Department of Health Services,

University of Washington, Seattle, Washington 98195, U.S.A.

December 20, 2010

Author's Footnote:

Kwun Chuen Gary Chan is Assistant Professor, Department of Biostatistics and Department of Health Services, University of Washington, 1959 NE Pacific Street, Seattle WA 98195. (email: kcgchan@uw.edu).

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

Abstract

In the presence of missing response, reweighting the complete case subsample by the inverse of nonmissing probability is both intuitive and easy to implement. However, inverse probability weighting is not efficient in general and is not robust against misspecification of the missing probability model. Calibration was developed by survey statisticians for improving efficiency of inverse probability weighting estimators when population totals of auxiliary variables are known and when inclusion probability is known by design. In missing data problem we can calibrate auxiliary variables in the complete case subsample to the full sample. However, the inclusion probability is unknown in general and need to be estimated in missing data problems and it is unclear whether calibration is robust against misspecification of the missing probability model. It is also unclear how efficient calibration is for general missing data problem. This paper answers these two questions and presents two rather unexpected results. First, when the missing data probability is correctly specified and multiple working outcome regression models are posited, calibration enjoys an oracle property where the same semiparametric efficiency bound is attained as if the true outcome model is known in advance. Second, when the missing mechanism is misspecified, calibration can still be a consistent estimator when any one of the outcome regression model is correctly specified. This is a multiple robustness property more general than double robustness considered the missing data literature. We provide connections of a wide class of calibration estimator constructed based on generalized empirical likelihood to many existing estimators in biostatistics, econometrics and survey sampling and perform simulation studies to study the finite sample properties of calibration estimators.

Keywords: Generalized empirical likelihood, Model misspecification, Missing data, Robustness

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/uwbiostat/paper368

1. INTRODUCTION

Inverse probability weighting (IPW) was originally proposed by Horvitz and Thompson (1952) for reweighting a probability sample obtained from a complex survey design in order to represent an underlying study population. The estimator has also been widely used for missing data problems, where complete-case data is reweighted by the inverse of nonmissing probability. While IPW estimation is intuitive and easy to implement, the estimator is not efficient in general and is not robust against misspecification of missing probability.

Survey statisticians and biostatisticians have each developed methods to improve the IPW estimator. Calibration was proposed by Deville and Särndal (1992) in survey sampling literature to utilize information from auxiliary data. In missing data literature, Robins, Rotnitzky and Zhao (1994) considered a class of augmented inverse probability weighted (AIPW) estimating equation, which adds a mean zero augmentation term to the IPW estimating equation. The augmentation term utilizes information from fully observed variables in the full sample. While both AIPW and calibration estimators were proposed to improve efficiency of IPW estimator, little connection has been established in the literature until recently, see Qin and Zhang (2007) and Breslow et al. (2009).

While calibration estimators are well studied in the survey sampling literature, theoretical questions remain to be answered for its usage in missing data problem. Calibration is proposed when inclusion probability is known by design, but for missing data applications the nonmissing probability is usually not known but is being modeled. It is unclear whether calibration is robust against misspecification of the missing probability model. Also, it is unclear whether calibration can attain the semiparametric efficiency bound as for the AIPW estimators. This paper answers these two questions and presents two rather unexpected results. In section 2, we consider a missing response model and define calibration estimating equations to match moment conditions between complete-case subsample and the full sample. Calibration weighted is implemented using generalized empirical likelihood (Newey and Smith, 2004). Section 3 contains the main theoretical results of this paper. First, we

COBRA
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

will show that when the missing data probability is correctly specified and multiple working outcome regression models are posited, calibration enjoys an oracle property where the same semiparametric efficiency bound is attained as if the true outcome model is known in advance. Second, when the missing mechanism is misspecified, calibration can still be a consistent estimator when any one of the outcome regression model is correctly specified. Three important special cases of the generalized empirical likelihood calibration will be discussed in section 4 and is shown to be related to many existing estimators in the biostatistics, econometrics and survey sampling literature. Numerical examples, including simulation studies and an analysis of medical cost data from the Washington basic health plan will be presented in section 5. Discussions and several related extensions will be presented in section 6.

2. CALIBRATION ESTIMATORS

In this section we will consider a general framework for modifying inverse probability weights by calibration to include information from all observations using moment conditions. We consider the following missing response problem. Let Y be a random variable and X be a random vector. Suppose we observe $(y_1, x_1), \ldots, (y_n, x_n)$, $x_{n+1}, \ldots, x_N$, and the full data $(y_1, x_1), \ldots, (y_N, x_N)$ are *i.i.d.* from an unspecified distribution $F_0(y, x)$. Let R be a random variable correspond to the nonmissing indicator. The observed data can be represented as $(r_i, r_i y_i, x_i)$, where $r_i = 1$ for $i = 1, \ldots, n$ and $r_i = 0$ for $i = n+1, \ldots, N$. We are interested in estimating $\mu = E(Y)$, where Y is subject to missingness and auxiliary variables X are completely observed.

We consider the case under missing at random, *i.e.* $P(R = 1|Y, X) = P(R = 1|X) = \pi_0(X)$. Suppose $P(R = 1|X) = \pi(X; \beta_0)$, where β_0 is a finite dimensional parameter. A conventional choice of missing data model is a logistic regression model with linear predictors in X, though this is not necessary. Based on $(r_1, x_1), \ldots, (r_N, x_N)$, parameter β_0 can be estimated by solving a likelihood score equation $N^{-1} \sum_{i=1}^{N} s(x_i; \beta) = 0$ where $s(x; \beta) = [1 - \pi(x; \beta)]^{-1}[r_i - \pi(x; \beta)] \frac{\partial \pi}{\partial \beta}(x; \beta)$ and we denote $\hat{\beta}$ be the solution. When missing data

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/uwbiostat/paper368

mechanism is correctly modeled, the IPW estimator

$$\hat{\mu}_{IPW} = \frac{1}{N} \sum_{i=1}^{N} \frac{r_i}{\pi(x_i; \hat{\beta})} y_i \tag{1}$$

is a consistent estimator of μ. However, (1) is generally not fully efficient because information from $\tilde{x} = (x_{n+1}, \ldots, x_N)$ is not utilised except in the estimation of β_0 and such information may not be relevant to the estimation of μ. To improve efficiencies, we note that for $u(x) = (u_1(x), \ldots, u_q(x))$ such that $u_1,\ldots,u_q$ are linearly independent and $E_{F_0}(u^2(X))$ is finite, the two estimators $\tilde{u} = N^{-1} \sum_{i=1}^{N} r_i \pi^{-1}(x_i; \hat{\beta}) u(x_i)$ and $\bar{u} = N^{-1} \sum_{i=1}^{N} u(x_i)$ are both consistently estimating the same quantity, $E_{F_0}(u(X))$, while the latter is more efficient because information from all observations are utilized. Instead of using inverse probability weights in computing $\tilde{u}$ and in (1), we wish to find calibration weights $(p_1, \ldots, p_n)$ such that the following moment conditions are satisfied

$$\bar{u} = \sum_{i=1}^{n} p_i u(x_i) \tag{2}$$

The dimension of $u(\cdot)$ is assumed fixed and is much less than n. For $(p_1, \ldots, p_n)$ satisfying (2), the calibration weighted complete case estimate for $E_{F_0}(u(X))$ is more efficient than the IPW estimate $\tilde{u}$ because information from all observations is included. When Y and $u(X)$ are reasonably correlated, it is intuitive to expect that the calibration estimator $\hat{\mu}_{CAL} = \sum_{i=1}^{n} p_i y_i$ is possibly more efficient than the IPW estimator (1). The implied weights from moment restrictions (2) can be explicitly defined using generalized empirical likelihood (GEL) proposed by Newey and Smith (2004), a method originally proposed for efficient estimation of overidentified systems of estimating equations commonly encountered in econometrics applications. Calibration weights proposed by Deville and Särndal (1992) also satisfies (2) but the method to obtain the weights are different.

The construction of GEL calibration weights is as follows. Let $\rho(v)$ be a concave and twice differentiable function on $\mathcal{R}$ such that $\rho^{(1)} \neq 0$, where $\rho^{(j)}(v) = \partial^j \rho(v)/\partial v^j$ and $\rho^{(j)} = \rho^{(j)}(0)$. As suggested by Newey and Smith (2004), we can replace an arbitrary $\rho(v)$ by a normalised version $-\rho^{(2)}/(\rho^{(1)})^2 \rho([\rho^{(1)}/\rho^{(2)}]v)$ such that $\rho^{(1)} = \rho^{(2)} = -1$. This normalization will not

Hosted by The Berkeley Electronic Press

affect the results. The calibration weights are defined as

$$p_i = \frac{\pi^{-1}(x_i;\hat\beta)\rho^{(1)}(\hat\lambda^T(u(x_i)-\bar u))}{\sum_{j=1}^n \pi^{-1}(x_j;\hat\beta)\rho^{(1)}(\hat\lambda^T(u(x_j)-\bar u))} \tag{3}$$

where

$$\hat\lambda = \arg\max_\lambda \sum_{i=1}^n \pi^{-1}(x_i;\hat\beta)\rho(\lambda^T(u(x_i)-\bar u)) \tag{4}$$

We define a calibration (CAL) estimator to be $\hat\mu_{CAL} = \sum_{i=1}^n p_i y_i$. Moment restrictions (2) is satisfied as seen from the first order condition of (4).

In general, the calibration weights p_i are not guaranteed to be non-negative if λ is maximised globally in (4), except in the cases where $\rho^{(1)}(v) < 0\ \forall v \in \mathcal{R}$, such as $\rho(v) = -\exp(v)$. A way to produce non-negative weights for the whole GEL family, as suggested by Newey and Smith (2004), is to define $\hat\lambda$ to maximize the objective function in a restricted set $\Lambda = \{\lambda \in \mathcal{R}^q : \lambda^T(u_i(x_i)-\bar u) \in \mathcal{V}, i = 1,\dots,n\}$ where $\mathcal{V} \subset \mathcal{R}$ is an open interval containing zero. When we choose $\mathcal{V}$ to be a sufficiently small neighborhood around zero, p_i will be non-negative for all $i = 1,\dots,n$. When the missing data model is correctly specified, it follows from Newey and Smith (2004) that the restricted maximum exists with probability approaching 1 when n is large and is asymptotically equivalent to the unrestricted maximizer. The restricted maximization is implemented in the `gmm` package in R (Chaussé 2010). In econometrics, GEL estimators are usually solutions to saddlepoint problems and can be difficult to compute. However, the GEL calibration estimator is essentially a degenerate case of GEL with only auxiliary parameters λ appearing in (4) but not target parameters. In this case, $\hat\lambda$ is a solution to a convex maximization problem instead of a saddlepoint problem and can be computed easily.

3. ORACLE AND MULTIPLE ROBUST PROPERTIES

In this section we will examine statistical properties of calibration estimators in the context of missing data analysis, and show that the class of estimators enjoy an oracle property and a multiple robustness property. We consider model based calibration where the functions $u(x)$ in the moment condition (2) may depend on a finite dimensional parameter γ_0

COBRA
Collection of Biostatistics
Research Archive
http://biostats.bepress.com/uwbiostat/paper368

estimated by $\hat{\gamma}$. For instance, $u_1(X;\gamma_1),\ldots,u_q(X;\gamma_q)$ can be q non-nested working outcome regression models for $E(Y|X)$, and $\gamma_0 = (\gamma_1^T,\ldots,\gamma_q^T)^T$. We denote the sample mean $\bar{u}(\hat{\gamma}) = N^{-1}\sum_{i=1}^N u(x_i;\hat{\gamma})$ and the calibration weights $(p_1,\ldots,p_n)$ satisfy $\bar{u}(\hat{\gamma}) = \sum p_i u(x_i;\hat{\gamma})$, which are found by (3) and (4) with $u(x)$ and $\bar{u}$ replaced by $u(x;\hat{\gamma})$ and $\bar{u}(\hat{\gamma})$ respectively. Let $m(X) = c_0 + \sum_{j=1}^q c_j u_j(X;\gamma_0)$ where $c_0,\ldots,c_q$ minimizes

$$E((Y - c_0 - \sum_{j=1}^q c_j u_j(X;\gamma_0))^2). \tag{5}$$

That is, $m(X)$ is the best linear predictor of Y by $u(X)$. Suppose $\hat{\gamma}$ is a $\sqrt{N}$ consistent estimate of γ_0 and assume that the missing data model is correctly specified: $\pi_0(X) = \pi(X;\beta_0)$. We have the following lemma.

Lemma 1. *Under regularity conditions stated in the appendix,*

$$\hat{\mu}_{CAL} - \mu = \frac{1}{N}\sum_{i=1}^N \left[\frac{r_i}{\pi_0(x_i)}(y_i - \tilde{m}(x_i)) + (\tilde{m}(x_i) - \mu)\right] + o_p(N^{-1/2}) \tag{6}$$

where $\tilde{m}(X) = m(X) + A_2^T S^{-1}(1-\pi_0(X))^{-1}\frac{\partial\pi}{\partial\beta}(X;\beta_0)$, $A_2 = -E\left(\frac{\partial\pi}{\partial\beta}(X;\beta_0)\frac{1}{\pi(X;\beta_0)}(Y - m(X))\right)$ *and* $S = E\left(\pi_0^{-1}(X)(1-\pi_0(X))^{-1}\frac{\partial\pi}{\partial\beta}(X;\beta_0)\frac{\partial\pi}{\partial\beta}^T(X;\beta_0)\right)$

The above lemma holds for arbitrary sets of functions $u(\cdot)$ satisfying mild regularity conditions. The asymptotic expansion (6) depends on the choice of $u(X)$ implicitly through $m(X)$ and we may chose a particular $u(X)$ to minimize the asymptotic variance. Denote $m_0(X)$ be the true conditional expectation $E(Y|X)$. The optimality properties are stated in the following theorem.

Theorem 2. *(Semiparametric efficiency) Suppose regularity conditions in lemma 1 holds. In addition, if there exist $a_0,\ldots,a_q$ such that*

$$m_0(X) = a_0 + \sum_{i=1}^q a_j u_j(X;\gamma_0) \tag{7}$$

then the estimator $\hat{\mu}_{CAL}$ achieves the semiparametric efficiency bound as in Robins, Rotnitzky and Zhao (1994) and Hahn (1998).

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive
Hosted by The Berkeley Electronic Press

In Theorem 2, the constants $a_0, \dots, a_q$ are arbitrary and do not need to be estimated. Theorem 2 states that semiparametric efficiency is attained under a condition weaker than requiring the calibration function $u(X)$ to be identical to the true conditional expectation $m_0(X)$. An important implication of the theorem, an oracle property, is given as follows. Suppose $u_1(X;\gamma_1), \dots, u_q(X;\gamma_q)$ are q working models for $E(Y|X)$ and that one of them, without lot of generality say $u_1(X;\gamma_1)$, is the true conditional expectation. We have the following oracle property.

Corollary 3. *(Oracle Property) Under conditions in lemma 1 and suppose* $E(Y|X) = u_1(X;\gamma_1)$. *The estimator* $\hat{\mu}_{CAL,1}$ *where* $u = u_1$ *achieves the same semiparametric efficiency bound as the estimator* $\hat{\mu}_{CAL,2}$ *where* $u = (u_1, \dots, u_q)$.

While assuming multiple working regression models are similar to overfitting which should be avoided in usual statistical practice, we see following the oracle properties that the asymptotic efficiency of calibration estimators are not affected by multiple working models and attains the same semiparametric efficiency bound as if the true model is known in advance. In section 5, we show in simulation studies that multiple modeling lose a negligible amount of efficiency even in a practical sample size.

Next, we consider the validity of calibration estimators under misspecified missing data models. In this case, the estimator $\hat{\beta}$ will converge in probability to some constant vector β^*, but $\pi(X;\beta^*) \neq \pi_0(X)$. We denote $F_s(y,x) = P(Y \leq y, X \leq x | R = 1)$ be the biased sampling distribution of complete-case subsample. It follows from missing at random assumption that

$$dF_s(y,x) \propto \pi_0(x) dF_0(y,x).$$

The estimate $\hat{\lambda}$ will not converge in probability to 0, as in the case when missing data model is correctly specified, but will instead converge in probability to λ^* where

$$\lambda^* = \arg\max_{\lambda} E_s(\pi^{-1}(X;\beta^*)\rho(\lambda(u(X) - \mu_u)))$$

where $\mu_u = E(u(X))$, E_s denotes an expectation taken with respect to the distribution F_s.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/uwbiostat/paper368

We define a tilted distribution $F_t(y,x)$ such that

$$\begin{aligned} dF_t(y,x) &\propto \frac{1}{\pi(x;\beta^*)}\rho^{(1)}(\lambda^*(u(x)-\mu_u))dF_s(y,x) \\ &\propto \tilde{w}(x)dF_0(y,x) \end{aligned}$$

where $\tilde{w}(x) = k\pi_0(x)\pi^{-1}(x;\beta^*)\rho(\lambda^*(u(x)-\mu_u))$ for some constant k, and we denote E_t an expectation taken with respect to F_t. We have the following theorem.

Theorem 4. *(Robustness) When the missing data model is misspecified but condition (7) holds for the calibration function $u(X)$, calibration estimator $\hat{\mu}_{CAL}$ is still a consistent estimator for μ.*

The proof is as follows.

$$\begin{aligned} \hat{\mu}_{CAL} &= \sum_{i=1}^{n} p_i(y_i - m_0(x_i)) + \sum_{i=1}^{n} p_i m_0(x_i) \\ &= \sum_{i=1}^{n} p_i(y_i - m_0(x_i)) + \frac{1}{N}\sum_{i=1}^{N} m_0(x_i) \\ &= \sum_{i=1}^{n} \frac{\pi^{-1}(x_i;\hat{\beta})\rho(\hat{\lambda}(u(x_i)-\bar{u}))}{\sum_{j=1}^{n}\pi^{-1}(x_j;\hat{\beta})\rho(\hat{\lambda}(u(x_j)-\bar{u}))}(y_i - m_0(x_i)) + \frac{1}{N}\sum_{i=1}^{N} m_0(x_i) \\ &\xrightarrow{p} E_t((Y - m_0(X))) + E(m_0(X)) \\ &= E(\tilde{w}(X)(E(Y|X) - m_0(X))) + E(E(Y|X)) = 0 + \mu = \mu \end{aligned}$$

An immediate corollary is that when one of the q working models for $E(Y|X)$ is correctly specified, the calibration estimator is consistent even when the missing data model is misspecified. Therefore, calibration estimators enjoy the following multiple robust property: consistency holds when either the missing data model or any one of the working outcome regression models is correctly specified. Doubly robustness estimators (e.g. AIPW estimators) have been popular in missing data analysis because of its extra protection against misspecification of the missing data model. However, a working outcome regression model may be misspecified as well. Modified probability weighted estimators allow multiple non-nested working models to be assumed and is consistent when any one of the working models are

Hosted by The Berkeley Electronic Press

correctly specified. This provides an even better protection against model misspecification than the existing doubly robust estimators.

4. SPECIAL CASES AND RELATIONSHIP TO EXISTING ESTIMATORS

In this section, we consider several special cases of the GEL calibration estimator, and discuss their connections to existing estimators proposed in biostatistics, econometrics and survey sampling literature.

When ρ is a quadratic function, after normalization we have $\rho^{(1)}(v) = -v - 1$. From (4), $\hat{\lambda}$ has an explicit solution

$$\hat{\lambda} = -\left[\sum_{i=1}^{n} \pi^{-1}(x_i, \hat{\beta})(u(x_i) - \bar{u})^{\otimes 2}\right]^{-1} \left[\sum_{i=1}^{n} \pi^{-1}(x_i, \hat{\beta})(u(x_i) - \bar{u})\right]$$

where for a row vector a, $a^{\otimes 2} = aa^T$. The calibration weighted estimator is equivalent to

$$\hat{\mu}_{CAL,Q} = \frac{\sum_{i=1}^{N} r_i \pi^{-1}(x_i; \hat{\beta})[y_i - c_1^T u(x_i)]}{\sum_{i=1}^{N} r_i \pi^{-1}(x_i; \hat{\beta})} + c_1^T \frac{1}{N} \sum_{i=1}^{N} u(x_i) \tag{8}$$

where

$$c_1 = \sum_{i=1}^{n} \pi^{-1}(x_i, \hat{\beta}) \left[\sum_{i=1}^{n} \pi^{-1}(x_i, \hat{\beta})(u(x_i) - \bar{u})^{\otimes 2}\right]^{-1} [(u(x_i) - \bar{u})y_i]$$

This special case of GEL calibration estimator corresponds to the generalised regression estimator (Cassel, Särndal and Wretman 1976). Note that when the missingness model is correctly specified, the denominator $\sum_{i=1}^{N} r_i \pi^{-1}(x_i; \hat{\beta})$ on the left hand size of (8) is approximately N, so the estimator (8) is also similar to the augmented inverse probability weighted estimating equation proposed by Robins, Rotnitzky and Zhao (1994).

Empirical likelihood (EL) is another special case of GEL which is frequently studied in the literature (Owen 1988, Qin and Lawless 1994), which corresponds to $\rho(v) = \log(1 - v)$. In this case, $\hat{\lambda}$ is a solution of the system of equations

$$\sum_{i=1}^{n} \frac{\pi^{-1}(x_i; \hat{\beta})(u(x_i) - \bar{u})}{1 - \lambda^T(u(x_i) - \bar{u})} = 0$$

and

$$p_i = \frac{[\pi(x_i; \hat{\beta})(1 - \hat{\lambda}^T(u(x_i) - \bar{u}))]^{-1}}{\sum_{j=1}^{n}[\pi(x_j; \hat{\beta})(1 - \hat{\lambda}^T(u(x_j) - \bar{u}))]^{-1}}$$

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/uwbiostat/paper368

In this case, empirical likelihood has a pseudo nonparametric maximum likelihood interpretation, where p_i maximizes a weighted loglikelihood $\sum_{i=1}^n \pi^{-1}(x_i;\hat{\beta}) \log p_i$ subject to the moment condition (2). Moment matching using empirical likelihood for missing data have been discussed in the econometrics literature by Hellerstein and Imben (1999), and discussed recently in Qin and Zhang (2007) with an emphasis on causal inference applications. In survey sampling, the empirical likelihood based method has been proposed to calibrate design-based weights to auxiliary data by Chen and Sitter (1999), Wu and Sitter (2001), Chen, Sitter and Wu (2002), Kim (2009) among others.

Exponential tilting (ET) is also a special case of GEL where $\rho(v) = -\exp(v)$ (Kitamura and Stutzer 1997; Imbens, Spady and Johnson 1998). In this case, $\hat{\lambda}$ is a solution of the system of equations

$$\sum_{i=1}^n \pi^{-1}(x_i;\hat{\beta})(u(x_i) - \bar{u}) \exp(\lambda^T(u(x_i) - \bar{u})) = 0$$

and

$$p_i = \frac{\pi^{-1}(x_i;\hat{\beta}) \exp(\hat{\lambda}^T(u(x_i) - \bar{u}))}{\sum_{j=1}^n \pi^{-1}(x_j;\hat{\beta}) \exp(\hat{\lambda}^T(u(x_j) - \bar{u}))}$$

The estimator can also be formulated by maximizing a weighted entropy $\sum_{i=1}^n \pi^{-1}(x_i;\hat{\beta}) p_i \log p_i$ subject to the moment condition (2). This corresponds to raking estimators (Deming and Stephan 1940, Deville, Särndal and Sautory 1993) in the survey sampling literature, and An advantage of using the exponential tilting estimator is that the resulting weights p_i are always non-negative.

The class of GEL calibration estimators contain many more estimators than the three special cases mentioned above. For example, the family of power divergence statistics of Cressie and Read (1984) is a proper subclass of GEL, where for some scalar θ,

$$\rho(v) = -(1 + \theta v)^{(\theta+1)/\theta}/(\theta + 1)$$

The EL and ET estimators correspond to the limit as $\theta \to -1$ and $\theta \to 0$ respectively, and the quadratic estimator corresponds to $\theta = 1$. Several other cases have also been considered

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

in the literature, for example, $\theta = -\frac{1}{2}$ (Freeman-Tukey), $\theta = -2$ (calibration Neyman) and $\theta = \frac{2}{3}$ (Cressie-Read).

5. NUMERICAL STUDIES

5.1 Simulated data

In this section we present simulation studies and an analysis of Washington basic health plan data to study the finite sample performance of calibration estimators. The simulation studies followed a scenario in Kang and Schafer (2007) for estimation of a population mean. The scenario was designed so that the assumed outcome regression and missing data models are nearly correct under misspecification, but the AIPW estimator can be severely biased. Sample sizes for each simulated data set was 200 or 1000, and 1000 Monte Carlo datasets were generated. For each observation, a random vector $Z = (Z_1, Z_2, Z_3, Z_4)$ was generated from a standard multivariate normal distribution, and transformations $X_1 = \exp(Z_1/2), X_2 = Z_2/(1+\exp(Z_1)), X_3 = (Z_1 Z_3/25+0.6)^3$ and $X_4 = (Z_2+Z_4+20)^2$ were defined. The outcome of interest Y was generated from a normal distribution with mean $210 + 27.4Z_1 + 13.7Z_2 + 13.7Z_3 + 13.7Z_4$ and unit variance, and Y was observed with probability $\exp(\eta_0(Z))/(1 + \exp(\eta_0(Z)))$ where $\eta_0(Z) = -Z_1 + 0.5Z_2 - 0.25Z_3 - 0.1Z_4$. The correctly specified outcome and missing data models were regression models with Z as covariates, whereas we treated X to be the covariates instead of Z in misspecified models. Kang and Schafer (2007) showed that the missspecified models are nearly correctly specified. In each case we considered four possible combinations of correct and misspecified missing data and outcome regression models: (a) both correct; (b) correct missing data model and incorrect outcome regression; (c) incorrect missing data model but correct outcome regression and (d) both incorrect. For calibration estimators, we construct moment restrictions based on $u(Z) = (Z_1, Z_2, Z_3, Z_4)$ for correctly specified outcome model and to $u(X) = (X_1, X_2, X_3, X_4)$ for misspecified outcome model. We compared the performances of the inverse probability weighted estimator $\hat{\mu}_{IPW}$, the augmented inverse probability weighted estimator $\hat{\mu}_{AIPW}$, the calibration estimators $\hat{\mu}_{CAL,Q}, \hat{\mu}_{CAL,EL}, \hat{\mu}_{CAL,ET}$ corresponding to three special cases in the generalised empirical

http://biostats.bepress.com/uwbiostat/paper368

likelihood family: Quadratic (Q: $\rho(v) = -(v+1)^2/2$), empirical likelihood (EL: $\rho(v) = \ln(1-v)$) and exponential tilting (ET: $\rho(v) = -\exp(v)$). The results are shown in Table 1.

[Table 1 about here.]

Simulation results showed that both the AIPW estimator and the calibration estimators were more efficient than the IPW estimator. Both AIPW and calibration estimators had negligible bias when either the missing data model or the outcome regression model was correctly specified. When both models were correctly specified, AIPW and calibration estimators had very similar performances. When only one of the two models were correctly specified, the calibration estimators were more efficient than the AIPW estimator. When both models were misspecified, the AIPW estimator had a considerable bias and variability but the calibration estimators showed much better performance. Particular choices of calibration estimators within the GEL family did not affect their performance in general.

We next consider a case where the missing data mechanism was possibly misspecified and multiple working outcome regression models were assumed which contained the correctly specified model. Let $u_1 = (1, Z_1, Z_2, Z_3, Z_4)^T\hat{\gamma}_1$, $u_2 = (1, X_1, X_2, X_3, X_4)^T\hat{\gamma}_2$, $u_3 = (1, X_1, X_2, Z_3, Z_4)^T\hat{\gamma}_3$ and $u_4 = (1, Z_1, Z_2, X_3, X_4)^T\hat{\gamma}_4$, where $\hat{\gamma}_1, \hat{\gamma}_2, \hat{\gamma}_3$ and $\hat{\gamma}_4$ were least square estimates obtained from complete case data. We considered moment conditions from one to four working models: (a) one working model $u = u_1$, (b) two working models $u = (u_1, u_2)$, (c) three working models $u = (u_1, u_2, u_3)$ and (d) four working models $u = (u_1, u_2, u_3, u_4)$. Each of the four cases contained the correctly specified outcome regression model u_1. The simulation results are shown in Table 2. When multiple working outcome regression models were assumed that contained the correct model, calibration estimators were robust against misspecification of the missing data model, had negligible bias and negligible loss of efficiency compare to the estimator that calibrate only to the correct model known in advance.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

[Table 2 about here.]

Hosted by The Berkeley Electronic Press

5.2 Washington basic health plan data

We performed an analysis using the Washington basic health plan data. The dataset contained observations from 2687 households and collected a variety of health services variables. For the purpose of illustration, we chose outcome Y to be total household expenditure on outpatient visits, X_1 be the family size and X_2 be the total number of outpatient visits. The distribution of medical expenditure was highly skewed to the right with a lot of zeroes, and the mean household expenditure for outpatient visits was $\mu_y = 1948$. We drew a subsample following a model $logitP(R = 1|X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_1 I(X_1 \geq 3) + \beta_3 X_2$ and compared the performance of IPW and GEL calibration estimators as if Y were only observed in the subsamples. The resampling process was repeated $B = 1000$ times.

We evaluated the estimators by comparing two performance measures, relative bias (RB) and relative efficiency (RE), defined by

$$RB = \frac{1}{B} \sum_{b=1}^{B} \frac{\hat{\mu}_y - \mu_y}{\mu_y}$$

and

$$RE = \frac{MSE_{IPW}}{MSE}$$

where $\hat{\mu}_b$ is an estimator computed from the b^{th} sample, $MSE = B^{-1} \sum_{i=1}^{B} (\hat{\mu}_b - \mu_y)^2$ and MSE_{IPW} is the MSE of $\hat{\mu}_{IPW}$. The performance of estimators were evaluated under both a correctly specified missing data model and a misspecified working model $logitP(R = 1|X_1, X_2) = \delta_0 + \delta_1 X_1 + \delta_2 X_1 I(X_1 \geq 3)$. The misspecified model ignored the dependence between the missing mechanism and X_2. Under each scenario, we considered generalised empirical likelihood calibration estimators $\hat{\mu}_{CAL,Q,k}$, $\hat{\mu}_{CAL,EL,k}$ and $\hat{\mu}_{CAL,ET,k}$ where $i = 1, 2$ indicating two sets of working projection models. Under working assumption 1 ($k = 1$), we assumed a working linear model between Y and X_1 and calibrated to the corresponding linear projection. Under working assumption 2 ($k = 2$), we assumed two separate working linear models, one between Y and X_1 and the other between Y and X_2. Scatterplots showed that a working linear model between Y and X_1 may not be appropriate but a linear model

http://biostats.bepress.com/uwbiostat/paper368

between Y and X_2 was reasonable. Given this fact we expect calibration estimators under working assumption 2 shall perform better than the corresponding estimators for working assumption 1 both in terms of efficiency and bias reduction.

The results of the analyses are shown in table 3. In terms of bias, IPW estimators and calibration estimators showed considerable bias under working assumption 1 when the missing data mechanism is misspecified. However, the bias for calibration estimators was almost completely eliminated under working assumption 2 even when the missing data mechanism is misspecified. In terms of efficiency, calibration estimators under working model 2 demonstrated superior efficiency relative to the IPW estimator, both under correct and incorrect specification of missing data model. The results also suggested that particular choices of calibration estimators within the GEL family did not affect their performance in general.

[Table 3 about here.]

6. RELATED EXTENSIONS

In this article we study the statistical properties of GEL calibration calibration estimators in the context of missing data analysis. Calibration estimators allow multiple working outcome regression models to be assumed and enjoy an oracle property where the same semiparametric efficiency bound is attained as if the true outcome regression model is known in advance, when the missing data mechanism is correctly specified. The estimators also enjoy a multiple robustness property, where consistency holds when either the missing mechanism or any one of the working outcome regression models is correctly specified. Calibration estimators provide an even better protection against model misspecification than the existing doubly robust estimators. In this section we discuss several related extensions, including a different but related way to construct calibration estimators, calibration estimation of distribution functions and calibration estimating equations.

In previous sections, we focus on a class of calibration estimators satisfying moment conditions (2). There are many other calibration estimators that satisfy (2) and enjoy similar

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

statistical properties as the proposed class. A different but related calibration estimator can be constructed by noting that when the missingness model is correctly specified we have

$$E\left(\frac{R-\pi(X;\beta_0)}{\pi(X;\beta_0)}u(X)\right)=0$$

That is, $E(R\pi^{-1}(X;\beta_0)u(X)-\mu_u)=0$. We can define calibration weights as

$$p_i^*=\frac{1}{\pi(x_i;\hat\beta)}\rho^{(1)}\left(\hat\lambda_2^T\left(\pi^{-1}(x_i;\hat\beta)u(x_i)-\bar u\right)\right) \tag{9}$$

where

$$\hat\lambda_2=\arg\max_{\lambda}\sum_{i=1}^n\rho\left(\lambda^T\left(\pi^{-1}(x_i;\hat\beta)u(x_i)-\bar u\right)\right) \tag{10}$$

In this case, we assume that u contains a constant function. The moment condition $\bar u = \sum p_i^* u(x_i)$ is satisfied from the first order condition of (10). We can define a calibration estimator to be $\hat\mu_{CAL2}=\sum_{i=1}^n p_i^* y_i$. Suppose condition (7) holds,

$$\begin{aligned}\hat\mu3_{CAL2} &= \sum_{i=1}^n p_i^* y_i\\ &= \sum_{i=1}^n p_i^*(y_i-m_0(x_i))+\sum_{i=1}^n p_i^* m_0(x_i)\\ &= \sum_{i=1}^n p_i^*(y_i-m_0(x_i))+\frac{1}{N}\sum_{i=1}^N m_0(x_i)\end{aligned}$$

which converges in probability to μ by similar arguments as above. Therefore, the calibration estimator $\hat\mu_{CAL2}$ enjoys similar multiple robustness properties to those of the calibration estimator $\hat\mu_{CAL}$.

Although we focused on estimation of population mean, calibration is a general scheme that can be used in other estimation problems. For instance, if we are interested in estimating $F(y)=P(Y\le y)$, we can define a calibration estimator to be $\hat F_{CAL}(y)=\sum_{i=1}^n p_i I(y_i\le y)$, where p_i is found in (3) and (4). When $u_1,\dots u_q$ are q working models for $P(Y\le y|X)$ and contain the true model, by similar arguments as above we can show that $\hat F_{CAL}(y)$ converges in probability to $F(y)$ even when the missing data model is misspecified.

When we are interested in estimating a parameter θ_0 defined by an unbiased estimating function $g(y,x;\theta)$ such that $E(g(Y,X;\theta_0))=0$, we can define $\hat\theta_{CAL}$ to be the solution of

http://biostats.bepress.com/uwbiostat/paper368

a calibration estimating equation $g_{CAL}(\theta) = 0$ where $g_{CAL}(\theta) = \sum_{i=1}^{n} p_i g(y_i, x_i; \theta)$. Let $h_0(X) = E(g(Y, X; \theta_0)|X)$ and suppose there exist constants $a_0, \ldots, a_q$ such that $h_0(X) = a_0 + \sum_{j=1}^{q} a_j u_j(X)$, then

$$\begin{aligned} g_{CAL}(\theta) &= \sum_{i=1}^{n} p_i(g(y_i, x_i; \theta) - h_0(x_i)) + \sum_{i=1}^{n} p_i h_0(x_i) \\ &= \sum_{i=1}^{n} p_i^*(g(y_i, x_i; \theta) - h_0(x_i)) + \frac{1}{N}\sum_{i=1}^{N} h_0(x_i) \\ &\xrightarrow{p} E_t(g(Y, X; \theta) - h_0(X)) \end{aligned}$$

and $g_{CAL}(\theta_0) \xrightarrow{p} 0$. Let $Q(\theta) = E_t(g(Y, X; \theta) - h_0(X))^T E_t(g(Y, X; \theta) - h_0(X))$. Suppose the parameter space Θ is compact, $-Q(\theta)$ is uniquely maximised at θ_0, $Q(\theta)$ is continuous and $g_{CAL}(\theta)^T g_{CAL}(\theta) \to Q(\theta)$ uniformly in probability in a neighborhood of θ_0, then it follows from Newey and MacFadden (1994) that $\hat{\theta}_{CAL}$ is a consistent estimate of θ_0 even when the missing data model is misspecified.

Acknowledgement

The author would like to thank Professor Norman Breslow for helpful comments and suggestions that greatly improved the paper.

APPENDIX.

A.1 Proof of lemma 1

Suppose $E(Y^2) < \infty$, $u_1, \ldots, u_q$ are linearly independent, $\hat{\gamma}$ is a $\sqrt{N}$-consistent estimate of γ_0, π_0 is strictly between 0 and 1, $\rho(v)$ is twice continuously differentiable and uniformly bounded in a neighborhood of zero and, $u(\cdot;\gamma)$, $u^2(\cdot;\gamma)$, $\partial u(\cdot;\gamma)/\partial\gamma$, $\partial\pi(\cdot;\beta)/\partial\beta$ and $\partial\pi^2(\cdot;\beta)/\partial\beta\partial\beta^T$ are uniformly bounded by certain integrable functions in a neighborhood of $(\beta_0^T, \gamma_0^T)^T$. By standard asymptotic analysis it can be shown that $\hat{\lambda} \xrightarrow{p} 0$, $\hat{\beta} \xrightarrow{p} \beta_0$ and that $\hat{\lambda}$ and $\hat{\beta}$ are

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

$\sqrt{N}$-consistent. Moreover,

$$\begin{aligned}
\hat{\mu}_{CAL} - \mu &= \sum_{i=1}^{n} p_i(y_i - m(x_i)) + \sum_{i=1}^{n} p_i m(x_i) - \mu \\
&= \sum_{i=1}^{n} p_i(y_i - m(x_i)) + \frac{1}{N}\sum_{i=1}^{N}(m(x_i) - \mu) \\
&= \frac{1}{N}\sum_{i=1}^{N} r_i \left(\frac{\pi^{-1}(x_i;\hat{\beta})\rho^{(1)}(\hat{\lambda}(u(x_i;\hat{\gamma}) - \bar{u}(\hat{\gamma})))}{N^{-1}\sum_{i=1}^{N} r_j \pi^{-1}(x_j;\hat{\beta})\rho^{(1)}(\hat{\lambda}(u(x_j;\hat{\gamma}) - \bar{u}(\hat{\gamma})))} - \pi^{-1}(x_i;\beta_0) \right)(y_i - m(x_i)) \\
&\quad + \frac{1}{N}\sum_{i=1}^{N}\left[\frac{r_i}{\pi(x_i;\beta_0)}(y_i - m(x_i)) + (m(x_i) - \mu)\right] \\
&= A_1^T(\hat{\lambda} - 0) + A_2^T(\hat{\beta} - \beta_0) + A_3^T(\hat{\gamma} - \gamma_0) \\
&\quad + \frac{1}{N}\sum_{i=1}^{N}\left[\frac{r_i}{\pi_0(x_i)}(y_i - m(x_i)) + (m(x_i) - \mu)\right] + o_p(N^{-1/2})
\end{aligned}$$

where

$$A_1 = E[(u(X) - \mu_u)(Y - m(X))]$$

From the first order condition of (5), we have $E(Y - m(X)) = 0$ and $E(u(X)(Y - m(X))) = 0$, therefore $A_1 = 0$. Also,

$$A_2 = -E\left(\frac{\partial \pi}{\partial \beta}(X;\beta_0)\frac{1}{\pi(X;\beta_0)}(Y - m(X))\right)$$

and

$$A_3 = -E\left(\frac{\partial}{\partial \gamma}[\rho^{(1)}(\lambda^T u(X;\gamma))(Y - m(X))]\Big|_{\lambda=0,\gamma=\gamma_0}\right) = 0$$

Therefore,

$$\begin{aligned}
\hat{\mu}_{CAL} - \mu &= \frac{1}{N}\sum_{i=1}^{N}\left[\frac{r_i}{\pi_0(x_i)}(y_i - m(x_i)) + (m(x_i) - \mu)\right] + A_2^T(\hat{\beta} - \beta_0) + o_p(N^{-1/2}) \\
&= \frac{1}{N}\sum_{i=1}^{N}\left[\frac{r_i}{\pi_0(x_i)}(y_i - m(x_i)) + (m(x_i) - \mu)\right] \\
&\quad + A_2^T S^{-1}\frac{1}{N}\sum_{i=1}^{N}\frac{r_i - \pi_0(x_i)}{\pi_0(x_i)(1 - \pi_0(x_i))}\frac{\partial \pi}{\partial \beta}(x_i;\beta_0) + o_p(N^{-1/2}) \\
&= \frac{1}{N}\sum_{i=1}^{N}\left[\frac{r_i}{\pi_0(x_i)}(y_i - \tilde{m}(x_i)) + (\tilde{m}(x_i) - \mu)\right] + o_p(N^{-1/2})
\end{aligned}$$

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/uwbiostat/paper368

A.2 Proof of theorem 2

We start from the expression (6) in lemma 1. When (7) holds, then we have $m(X) = m_0(X)$ since $E(Y|X)$ minimizes (5). Furthermore, when condition (7) holds,

$$A_2 = -E\left(\frac{\partial \pi}{\partial \beta}(X;\beta_0)\frac{1}{\pi(X;\beta_0)}(E(Y|X) - m(X))\right) = 0$$

Under this special case, we have

$$\hat{\mu}_{CAL} - \mu = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{r_i}{\pi_0(x_i)}(y_i - m_0(x_i)) + (m_0(x_i) - \mu)\right] + o_p(N^{-1/2})$$

The influence function corresponds to the semiparametric efficiency bound.

A.3 Proof of corollary 3

Condition (7) is satisfied with $a_1 = 1$ and $a_0 = a_2 = \ldots = a_q = 0$. Applying theorem 2 it is straighforward to see that the two estimators attains the same semiparametric efficiency bound.

REFERENCES

Breslow, N., Lumley, T., Ballantyne, C., Chambless, L., and Kulich, M. (2009), "Improved Horvitz–Thompson Estimation of Model Parameters from Two-phase Stratified Samples: Applications in Epidemiology," *Statistics in Biosciences*, 1(1), 32–49.

Cassel, C., Särndal, C., and Wretman, J. (1976), "Some results on generalized difference estimation and generalized regression estimation for finite populations," *Biometrika*, 63(3), 615–620.

Chaussé, P. (2010), "Computing Generalized Method of Moments and Generalized Empirical Likelihood with R," *Journal of Statistical Software*, 34(11), 1–35.

Chen, J., and Sitter, R. (1999), "A pseudo empirical likelihood approach to the effective use of auxiliary information in complex surveys," *Statistica Sinica*, 9, 385–406.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

Chen, J., Sitter, R., and Wu, C. (2002), "Using empirical likelihood methods to obtain range restricted weights in regression estimators for surveys," *Biometrika*, 89(1), 230–237.

Cressie, N., and Read, T. (1984), "Multinomial goodness-of-fit tests," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 440–464.

Deming, W., and Stephan, F. (1940), "On a least squares adjustment of a sampled frequency table when the expected marginal totals are known," *The Annals of Mathematical Statistics*, 11(4), 427–444.

Deville, J. (1992), "Calibration estimators in survey sampling," *Journal of the American Statistical Association*, 87(418), 376–382.

Deville, J., Särndal, C., and Sautory, O. (1993), "Generalized raking procedures in survey sampling," *Journal of the American Statistical Association*, 88(423), 1013–1020.

Hahn, J. (1998), "On the role of the propensity score in efficient semiparametric estimation of average treatment effects," *Econometrica*, 66(2), 315–331.

Hansen, L., Heaton, J., and Yaron, A. (1996), "Finite-sample properties of some alternative GMM estimators," *Journal of Business & Economic Statistics*, 14(3), 262–280.

Hellerstein, J., and Imbens, G. (1999), "Imposing moment restrictions from auxiliary data by weighting," *Review of Economics and Statistics*, 81(1), 1–14.

Horvitz, D., and Thompson, D. (1952), "A generalization of sampling without replacement from a finite universe," *Journal of the American Statistical Association*, 47(260), 663–685.

Huang, C., Qin, J., and Follmann, D. (2008), "Empirical likelihood-based estimation of the treatment effect in a pretest-posttest study," *Journal of the American Statistical Association*, 103(483), 1270–1280.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/uwbiostat/paper368

Imbens, G., Spady, R., and Johnson, P. (1998), "Information theoretic approaches to inference in moment condition models," *Econometrica*, 66(2), 333–357.

Isaki, C., and Fuller, W. (1982), "Survey design under the regression superpopulation model," *Journal of the American Statistical Association*, 77(377), 89–96.

Kang, J., and Schafer, J. (2007), "Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data," *Statistical science*, 22(4), 523–539.

Kim, J. (2009), "Calibration estimation using empirical likelihood in survey sampling," *Statistica Sinica*, 19, 145–157.

Kitamura, Y., and Stutzer, M. (1997), "An information-theoretic alternative to generalized method of moments estimation," *Econometrica*, 65(4), 861–874.

Lindsay, B., and Qu, A. (2003), "Inference functions and quadratic score tests," *Statistical Science*, 18(3), 394–410.

Newey, W., and McFadden, D. (1994), "Large sample estimation and hypothesis testing," *Handbook of Econometrics*, 4, 2111–2245.

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

Owen, A. (1988), "Empirical likelihood ratio confidence intervals for a single functional," *Biometrika*, 75(2), 237–249.

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

Qin, J., and Zhang, B. (2007), "Empirical-likelihood-based inference in missing response problems and its application in observational studies," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(1), 101–122.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

Robins, J., Rotnitzky, A., and Zhao, L. (1994), "Estimation of Regression Coefficients When Some Regressors Are Not Always Observed.," *Journal of the American Statistical Association*, 89(427), 846–866.

Tan, Z. (2006), "A distributional approach for causal inference using propensity scores," *Journal of the American Statistical Association*, 101(476), 1619–1637.

Wu, C., and Sitter, R. (2001), "A Model-Calibration Approach to Using Complete Auxiliary Information from Survey Data.," *Journal of the American Statistical Association*, 96(453), 185–193.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/uwbiostat/paper368

List of Tables

1 Comparisons among the calibration estimators and IPW estimators under the Kang and Schafer scenario with four possible combinations of correct and misspecified missing data and outcome regression models, (a) both correct, (b) correct missing data model and incorrect outcome regression, (c) incorrect missing data model but correct outcome regression and (d) both incorrect. SSE represents the sampling standard deviation. 24

2 Performance of calibration estimators under correctly specified or misspecified missing data models and multiple working outcome regression models, (a) one working model, (b) two working models, (c) three working models and (d) four working models. SSE represents the sampling standard deviation. 25

3 Washington basic health plan data. Relative bias (RB) and relative efficiency (RE) of estimators under (a) correct specification of missing mechanism and (b) misspecification of missing mechanism. 26

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

Table 1: Comparisons among the calibration estimators and IPW estimators under the Kang and Schafer scenario with four possible combinations of correct and misspecified missing data and outcome regression models, (a) both correct, (b) correct missing data model and incorrect outcome regression, (c) incorrect missing data model but correct outcome regression and (d) both incorrect. SSE represents the sampling standard deviation.

n=200								
	(a)		(b)		(c)		(d)	
	Bias	SSE	Bias	SSE	Bias	SSE	Bias	SSE
$\hat{\mu}_{IPW}$	-0.74	12.62	-0.74	12.62	28.65	179.02	28.65	179.02
$\hat{\mu}_{AIPW}$	0.02	2.50	0.28	3.76	0.01	2.55	-8.01	40.30
$\hat{\mu}_{CAL,Q}$	0.02	2.50	0.50	3.11	0.02	2.50	-2.13	3.26
$\hat{\mu}_{CAL,EL}$	0.02	2.50	0.28	3.13	0.02	2.49	-2.73	3.98
$\hat{\mu}_{CAL,ET}$	0.02	2.50	0.38	3.09	0.02	2.50	-2.40	3.48

n=1000								
	(a)		(b)		(c)		(d)	
	Bias	SSE	Bias	SSE	Bias	SSE	Bias	SSE
$\hat{\mu}_{IPW}$	0.27	5.07	0.27	5.07	36.99	157.31	36.99	157.31
$\hat{\mu}_{AIPW}$	0.01	1.13	0.06	1.65	-0.01	1.25	-13.38	72.19
$\hat{\mu}_{CAL,Q}$	0.01	1.13	0.17	1.33	0.01	1.13	-2.94	1.45
$\hat{\mu}_{CAL,EL}$	0.01	1.13	0.10	1.35	0.01	1.13	-4.16	1.86
$\hat{\mu}_{CAL,ET}$	0.01	1.13	0.13	1.34	0.01	1.13	-3.45	1.86

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/uwbiostat/paper368

Table 2: Performance of calibration estimators under correctly specified or misspecified missing data models and multiple working outcome regression models, (a) one working model, (b) two working models, (c) three working models and (d) four working models. SSE represents the sampling standard deviation.

		n=200				n=1000			
		Correct		Misspecified		Correct		Misspecified	
		Bias	SSE	Bias	SSE	Bias	SSE	Bias	SSE
$\hat{\mu}_{CAL,Q}$	(a)	0.02	2.50	0.02	2.50	0.01	1.13	0.01	1.13
	(b)	0.02	2.50	0.02	2.50	0.01	1.13	0.01	1.13
	(c)	0.02	2.50	0.02	2.50	0.01	1.13	0.01	1.13
	(d)	0.02	2.50	0.02	2.50	0.01	1.13	0.01	1.13
$\hat{\mu}_{CAL,EL}$	(a)	0.02	2.50	0.02	2.50	0.01	1.13	0.01	1.13
	(b)	0.02	2.50	0.02	2.50	0.01	1.13	0.01	1.13
	(c)	0.02	2.50	0.03	2.49	0.01	1.13	0.02	1.13
	(d)	0.02	2.50	0.01	2.49	0.01	1.13	0.01	1.13
$\hat{\mu}_{CAL,ET}$	(a)	0.02	2.50	0.02	2.50	0.01	1.13	0.01	1.13
	(b)	0.02	2.50	0.02	2.50	0.01	1.13	0.01	1.13
	(c)	0.02	2.50	0.02	2.50	0.01	1.13	0.01	1.13
	(d)	0.02	2.50	0.02	2.50	0.01	1.13	0.01	1.13

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

Table 3: Washington basic health plan data. Relative bias (RB) and relative efficiency (RE) of estimators under (a) correct specification of missing mechanism and (b) misspecification of missing mechanism.

$(\beta_1, \beta_2, \beta_3, \beta_4)$	Measures	$\hat{\mu}_{IPW}$	$\hat{\mu}_{CAL,Q,1}$	$\hat{\mu}_{CAL,EL,1}$	$\hat{\mu}_{CAL,ET,1}$	$\hat{\mu}_{CAL,Q,2}$	$\hat{\mu}_{CAL,EL,2}$	$\hat{\mu}_{CAL,ET,2}$
				(a)				
(-0.2,0.1,-0.05,-0.01)	RB	-0.003	-0.003	-0.003	-0.003	<0.001	-0.002	-0.001
	RE	1.00	1.02	1.02	1.02	1.41	1.29	1.38
(-0.2,0.1,-0.05,0.05)	RB	-0.001	-0.001	-0.001	-0.001	<0.001	<0.001	<0.001
	RE	1.00	1.00	1.00	1.00	1.01	1.01	1.01
				(b)				
(-0.2,0.1,-0.05,-0.01)	RB	-0.091	-0.091	-0.091	-0.091	<0.001	-0.015	-0.009
	RE	1.00	1.00	1.00	1.00	11.83	8.09	10.61
(-0.2,0.1,-0.05,0.05)	RB	0.191	0.191	0.191	0.191	0.005	-0.093	-0.056
	RE	1.00	1.00	1.00	1.00	209.09	4.07	11.11

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/uwbiostat/paper368