A simple multiple robust estimator for missing response problem

K. C. G. Chan

Received 00 Month 2012; Accepted 00 Month 2012

Han and Wang (2013) proposed a multiple robust estimator for a missing response problem that is a more robust than doubly robust estimators proposed in the literature. Their formulation is based on empirical likelihood, which solves an implicit Lagrangian equation and often encounters computational problems such as multiple roots or non-convergence. An alternative multiple robust estimator is proposed which is computed by least squares and can be implemented easily in practice. We show this multiple robust estimator is locally semiparametric efficient. Copyright © 2012 John Wiley & Sons, Ltd.

Keywords: Least square; Model averaging; Semiparametric efficiency

1. Introduction and background

Estimation with missing data had been well studied in the statistical literature but remains to be an active research area due to its complex nature and practical applicability. When a response variable is subject to missingness and certain covariates are fully observable, two different approaches can be employed to estimate the mean of the response variable under a missing at random assumption. One approach is to weight the observed response using the inverse of their estimated selection probability based on a propensity score model (Horvitz and Thompson, 1952; Rosenbaum and Rubin, 1983). Another approach is to fit a regression model for the conditional mean of the response given covariates, followed by averaging fitted values from all subjects. Both approaches have merits but fail to be consistent when their underlying model is misspecified. Doubly robust estimators remain consistent when one of the two models are correct and have gained considerable popularity in the recent statistical literature. Doubly robust estimators can be constructed in several different ways: augmented inverse probability weighting (Robins and Rotnitzky, 1994; Cao et al., 2009), regression estimation with propensity-based covariates (Scharfstein et al., 1999; Little and An, 2004; Bang and Robins, 2005), empirical likelihood (Qin and Zhang, 2007; Tan, 2006, 2010) and survey sampling methods (Cassel et al., 1976; Kott, 1994).

Doubly robust estimators are constructed based on one propensity score model and one outcome regression model. As noted in Kang and Schafer (2007), the doubly robust estimators can perform poorly when both models are misspecified.

aDepartment of Biostatistics, University of Washington, Seattle WA 98195, USA
Email: kcgchan@u.washington.edu
More robust procedures are needed because no model is perfect in practice. A recent paper of Han and Wang (2013) showed that the empirical likelihood approach of Qin and Zhang (2007) can be extended to include multiple working propensity score models and multiple outcome regression models. They show that the estimator enjoys a multiple robustness property that consistency is guaranteed when any one of the multiple propensity score or outcome regression models are correctly specified. While the extension from double robustness to multiple robustness is a major leap, the empirical likelihood estimator with many model-based constraints often suffer from computational problems. As noted in Han and Wang (2013), multiple roots are commonly encountered and can lead to inconsistent estimation. Empirical likelihood with misspecified models also has a slower than parametric asymptotic convergence rate (Schennach, 2007). To maintain the multiple robustness property that consistency is guaranteed when any one of the multiple propensity score or outcome regression models is correctly specified, we propose an alternative estimator from a different route: regression estimation with propensity-based covariates. The estimator can be simply computed by alleviating some implementation concerns of empirical likelihood, we propose a multiple robustness estimator based on empirical likelihood weights (Chen et al., 2002; Newey and Smith, 2004). Empirical likelihood with misspecified models also has a slower than parametric asymptotic convergence rate (Schennach, 2007). To maintain the multiple robustness properties while alleviating some implementation concerns of empirical likelihood, we propose a multiple robustness estimator based on the working models in $\mathcal{P}$ and $\mathcal{A}$. For instance, $\tilde{\alpha}_j$ can be a maximum likelihood estimate based on $\pi_j(x; \alpha_j)$ and $\tilde{\gamma}_k$ can be a least square estimate when $a_k(x; \gamma_k)$ is linear in $\gamma_k$.

The empirical likelihood estimator of Han and Wang (2013) is constructed by matching population moments of model-based fitted values from the complete-case subsample to the full sample. Empirical likelihood weights $\{w, i = 1, \ldots, m\}$ satisfy the constraints

$$\sum_{i=1}^{m} w_i = 1, \quad \sum_{i=1}^{m} w_i \pi_j(x; \tilde{\alpha}_j) = \hat{\theta}_j, j = 1, \ldots, J$$

$$\sum_{i=1}^{m} w_i a(x; \tilde{\gamma}_k) = \hat{\gamma}_k, k = 1, \ldots, K$$

where $\hat{\theta}_j = n^{-1} \sum_{i=1}^{n} \pi_j(x; \tilde{\alpha}_j)$ and $\hat{\gamma}_k = n^{-1} \sum_{i=1}^{n} a_k(x; \tilde{\gamma}_k)$. Let $\hat{w}$ be the weights that satisfy (1), which can be founded by first solving an implicit normal equation derived from a Langrange multiplier method. The empirical likelihood estimator is $\hat{\mu}_{EL} = \sum_{i=1}^{m} \hat{w}_i y_i$ which was shown to enjoy a multiple robustness property. When $j = 1$, the estimator reduces to the estimator of Qin and Zhang (2007). As mentioned in Section 1, the empirical likelihood with multiple constraints often suffer from implementation problems. We propose an alternative estimator from a different route.

2. Multiple robust estimation

Suppose $Y$ is a response variable which is subject to missingness. $X$ is a vector of covariates which is fully observable. $R$ is a non-missing indicator where $R = 1$ if $Y$ is observed and $R = 0$ if $Y$ is missing. We assume a missing at random condition $P(R = 1 | Y, X) = P(R = 1 | X)$. The observed data $(R, R, Y, X_i), i = 1, \ldots, n$ are independent and identically distributed. Without loss of generality, we assume that $R_i = 1$ for $i = 1, \ldots, m$ and $R_i = 0$ for $i = m + 1, \ldots, n$.

Since the functional form of the true propensity score model $\pi_0(x) = \text{pr}(R = 1 | X = x)$ and the true outcome conditional expectation function $a_0(x) = E(Y | X = x)$ are unknown in general, we assume multiple working propensity score models $\mathcal{P} = \{\pi_j(x; \alpha_j) : j = 1, \ldots, J\}$ and multiple working outcome regression models $\mathcal{A} = \{a_k(x; \gamma_k) : k = 1, \ldots, K\}$. Denote $\tilde{\alpha}_j, j = 1, \ldots, J$ and $\tilde{\gamma}_k, k = 1, \ldots, K$ the estimators based on the working models in $\mathcal{P}$ and $\mathcal{A}$. For instance, $\tilde{\alpha}_j$ can be a maximum likelihood estimate based on $\pi_j(x; \alpha_j)$ and $\tilde{\gamma}_k$ can be a least square estimate when $a_k(x; \gamma_k)$ is linear in $\gamma_k$.
Let \( u(X; \hat{\alpha}, \hat{\gamma}) = (1, (\pi_1(X; \hat{\alpha}_1))^{-1}, \ldots, (\pi_J(X; \hat{\alpha}_J))^{-1}, a_1(X; \hat{\gamma}_1), \ldots, a_K(X; \hat{\gamma}_K)) \). Using the complete case subsample, we regress \( Y \) on \( u(X; \hat{\alpha}, \hat{\gamma}) \) by least square to obtain \( \hat{\beta} \). That is, when \( u(X; \hat{\alpha}, \hat{\gamma}) = (u_1(X; \hat{\alpha}, \hat{\gamma})^T, \ldots, u_m(X; \hat{\alpha}, \hat{\gamma})^T)^T \) and \( Y = (Y_1, \ldots, Y_m)^T \).

\[
\hat{\beta} = [u(X; \hat{\alpha}, \hat{\gamma})^T u(X; \hat{\alpha}, \hat{\gamma})]^{-1} u(X; \hat{\alpha}, \hat{\gamma})^T Y
\]

The proposed least square estimator of \( \mu = E(Y) \) is

\[
\hat{\mu}_{LS} = \frac{1}{n} \sum_{i=1}^{n} u(X_i; \hat{\alpha}, \hat{\gamma}) \hat{\beta}.
\]

The least square estimator is motivated from doubly robust estimators constructed from regression estimators with propensity-based covariates (Scharfstein et al., 1999; Bang and Robins, 2005). However, their estimator only includes one outcome regression model and one propensity score model. Furthermore, even when \( J = 1 \) and \( K = 1 \), our estimator is different from the doubly robust estimator of Scharfstein et al. (1999) and Bang and Robins (2005). Suppose \( a_i(x; \gamma) = \Psi(s(\gamma; x)) \), where \( \Psi \) is a known link function and \( s(x; \gamma) \) is a known regression function. The estimator of Scharfstein et al. (1999) first fit an extended model \( \tilde{a}_i(x; \gamma) = \Psi(s(\gamma; x) + \phi / \pi_1(x; \hat{\alpha})) \), where \( \phi \) is an additional scalar parameter, and their estimator is an average of fitted values of this extended model. In contrast, our method first directly fit the model \( a_i(x; \gamma) \), followed by fitting another least square regression of \( Y \) against \( a_i(X; \hat{\gamma}) \) and \( 1/\pi_1(X; \hat{\alpha}) \), and to average the predictions obtained from the latter least square fit. Our method allows the multiple outcome models \( a_1, \ldots, a_J \) to have different link functions. As mentioned in Basu et al. (2011), health care cost data usually exhibit skewness and linearity and fitting outcome models with different link functions are often desired.

To show that \( \hat{\mu}_{LS} \) is multiply robust, we first consider the case where one of the models in \( \mathcal{P} \), say \( \pi_1(x; \alpha_1) \), is the true propensity score \( \pi_0(x) \). The maximum likelihood estimator \( \hat{\alpha}_1 \) is a consistent estimate of \( \alpha_1 \). The least square estimate \( \hat{\beta} \) solves the score equation

\[
\sum_{i=1}^{n} u_i(X_i)(Y_i - u(X_i) \hat{\beta}) = 0.
\]

In particular,

\[
\sum_{i=1}^{n} \frac{R_i}{\pi_1(X_i; \hat{\alpha}_1)}(Y_i - u(X_i; \hat{\alpha}, \hat{\gamma}) \hat{\beta}) = 0.
\]

Therefore,

\[
\hat{\mu}_{LS} = \frac{1}{n} \sum_{i=1}^{n} \frac{R_i}{\pi_1(X_i; \hat{\alpha})} (Y_i - u(X_i; \hat{\alpha}_1) \hat{\beta}) + \frac{1}{n} \sum_{i=1}^{n} u(X_i; \hat{\alpha}, \hat{\gamma}) \hat{\beta}
\]

\[
= \frac{1}{n} \sum_{i=1}^{n} \frac{R_i}{\pi_1(X_i; \hat{\alpha})} Y_i - \frac{1}{n} \sum_{i=1}^{n} \frac{R_i}{\pi_1(X_i; \hat{\alpha})} u(X_i; \hat{\alpha}, \hat{\gamma}) \hat{\beta}
\]

When \( \pi_1 \) is correctly specified, the second term converges to \( 0 \) in probability and the first term converges to \( E(Y) \). Therefore, \( \hat{\mu}_{LS} \) is consistent when a model in \( \mathcal{P} \) is the correct propensity score model.

Suppose instead one of the working regression model, say \( a_K(x; \gamma_K) \), is correctly specified and \( \hat{\gamma}_K \) converges to \( \gamma_K \) in probability under the data generating mechanism. The model on \( E(Y|X) \) can be written as

\[
E(Y|X) = u(X) \beta_0
\]
where \( \beta_0 = (0, \ldots, 0, 1) \). Under (4), \( \hat{\beta} \) converges to \( \beta_0 \) in probability and \( \hat{\mu}_{LS} \) converges to \( E(\alpha_k(X; \gamma_k)) = E(Y) \) in probability. Therefore, \( \hat{\mu}_{LS} \) is consistent when a model in \( A \) is the correct outcome regression model.

In addition to the multiple robustness properties, the proposed estimator is locally semiparametric efficient, according to the following theorem.

**Theorem 1**

When \( \mathcal{P} \) contains a correctly specified model \( \pi_0(x) \) and \( A \) contains a correctly specified model \( m_0(x) \), then \( \hat{\mu}_{LS} \) attains the semiparametric efficiency bound.

The proof is given as follows. Suppose \( \hat{\alpha} = (\hat{\alpha}_1^T, \ldots, \hat{\alpha}_J^T)^T \) and \( \hat{\gamma} = (\hat{\gamma}_1^T, \ldots, \hat{\gamma}_K^T)^T \). Without loss of generality assume \( \pi_0(x) = \pi_1(x; \alpha_1) \) and \( a_0(x) = a_k(x; \gamma_k) \). Under the assumed data generating mechanism, \( (\hat{\alpha}, \hat{\beta}, \hat{\gamma}) \) converges to constant vectors \( (\alpha_0, \beta_0, \gamma_0) \) in probability (White, 1982). Suppose \( (\hat{\alpha} - \alpha_0), (\hat{\beta} - \beta_0) \) and \( (\hat{\gamma} - \gamma_0) \) are all \( O_p(n^{-1/2}) \), which is typically true under standard regularity conditions. Let

\[
\phi(R_i, Y_i, X_i; \alpha, \beta, \gamma) = \frac{R_i}{\pi_1(X_i; \alpha)} \{Y_i - u(X_i; \alpha, \gamma)\beta\} - \{u(X_i; \alpha, \gamma)\beta - \mu\}.
\]

From (3) and Taylor series expansion,

\[
\hat{\mu}_{LS} - \mu = \frac{1}{n} \sum_{i=1}^{n} \phi(R_i, Y_i, X_i; \hat{\alpha}, \hat{\beta}, \hat{\gamma})
\]

\[
= \frac{1}{n} \sum_{i=1}^{n} \phi(R_i, Y_i, X_i; \alpha_0, \beta_0, \gamma_0)
\]

\[
+ E \left( \frac{\partial \phi(R_i, Y_i, X_i; \alpha_0, \beta_0, \gamma_0)}{\partial \alpha} \right) (\hat{\alpha} - \alpha_0)
\]

\[
+ E \left( \frac{\partial \phi(R_i, Y_i, X_i; \alpha_0, \beta_0, \gamma_0)}{\partial \beta} \right) (\hat{\beta} - \beta_0)
\]

\[
+ E \left( \frac{\partial \phi(R_i, Y_i, X_i; \alpha_0, \beta_0, \gamma_0)}{\partial \gamma} \right) (\hat{\gamma} - \gamma_0) + o_p(n^{-1/2})
\]

Note that

\[
E \left( \frac{\partial \phi(R_i, Y_i, X_i; \alpha_0, \beta_0, \gamma_0)}{\partial \alpha} \right)
\]

\[
= - E \left( \frac{R_i}{\pi_0(X)} \frac{\partial \pi_0(X)}{\partial \alpha} \{Y_i - a_0(X_i)\} \right) - E \left( \frac{R_i - \pi_0(X)}{\pi_0(X)} \frac{\partial u(X; \alpha_0, \gamma_0)\beta_0}{\partial \alpha} \right) = 0
\]

The first term is 0 since \( E(Y - a_0(X)|X) = 0 \) and the second term is 0 since \( E(R - \pi_0(X)|X) = 0 \). Furthermore,

\[
E \left( \frac{\partial \phi(R_i, Y_i, X_i; \alpha_0, \beta_0, \gamma_0)}{\partial \beta} \right) = -E \left( \frac{R_i - \pi_0(X)}{\pi_0(X)} u(X; \alpha_0, \gamma_0) \right) = 0
\]

and

\[
E \left( \frac{\partial \phi(R_i, Y_i, X_i; \alpha_0, \beta_0, \gamma_0)}{\partial \gamma} \right) = E \left( \frac{R_i - \pi_0(X)}{\pi_0(X)} \frac{\partial u(X; \alpha_0, \gamma_0)\beta_0}{\partial \gamma} \right) = 0.
\]

Since

\[
\phi(R_i, Y_i, X_i; \alpha_0, \beta_0, \gamma_0) = \frac{R_i}{\pi_0(X)} \{Y_i - a_0(X_i)\} - \{a_0(X_i) - \mu\}.
\]
we have
\[ \hat{\mu}_{LS} - \mu = \frac{1}{n} \sum_{i=1}^{n} R_i \{ Y_i - a_0(X_i) \} - \{ a_0(X_i) - \mu \} + o_p(n^{-1/2}) \]
which attains the semiparametric efficiency bound as in Robins and Rotnitzky (1994) and Hahn (1998).

The semiparametric efficiency result looks counterintuitive at first, because there is a common belief that overfitting leads to decrement of estimation efficiency. This is indeed true that the variability of \( \hat{\beta} \) increases with more redundant models being added to \( u(X) \). However, our main interest is not the estimation of \( \hat{\beta}_0 \) but \( \mu \), and interestingly, the asymptotic behavior of \( \hat{\beta}_{LS} \) is unaffected by estimated \( \hat{\beta} \) because a condition of Randles (1982), namely (5), is satisfied. Therefore, Theorem 1 does not contradict with existing theories for least square.

3. Numerical studies

We conducted simulation studies based on a scenario considered in Han and Wang (2013). The sample size was 300 or 1000, and 5000 replications were being generated for each situation. A covariate \( X \) was being generated from a uniform \((-2.5, 2.5)\) distribution. Conditioning on \( X \), \( Y \) followed a normal distribution with mean \( a(x) \) and variance \( 4x^2 + 2 \), and \( R \) followed a Bernoulli distribution with probability of success \( \pi(x) \). We considered two propensity score models, \( \pi(x) = \{1 + \exp(0.8 + 0.5x - 0.3x^2)\}^{-1} \) or \( 1 - \exp[-\exp(0.5 + 0.5x - 0.3\exp(x))] \) and two outcome regression models \( a(x) = 1 + 2x + 3x^2 \) or \( 1 + 2x + 3\exp(x) \). There were four combinations of \( \{ \pi(x), a(x) \} \). We considered two working propensity models \( \pi_1(x; \alpha_1) = \{1 + \exp(\alpha_{11} + \alpha_{12}x - \alpha_{13}x^2)\}^{-1} \) and \( \pi_2(x; \alpha_2) = 1 - \exp[-\exp(\alpha_{21} + \alpha_{22}x - \alpha_{23}\exp(x))] \), as well as two outcome regression models \( a_1(x; \gamma_1) = \gamma_{11} + \gamma_{12}x + \gamma_{13}x^2 \) and \( a_2(x; \gamma_2) = \gamma_{21} + \gamma_{22}x + \gamma_{23}\exp(x) \). Following the notation of Han and Wang (2013), we use a four-digit subscript to indicate the working models that entered the computation of \( \hat{\mu}_{LS} \). For example, \( \hat{\mu}_{1001} \) denotes the proposed estimator based on a working propensity score model \( \pi_1(x) \) and a working outcome regression model \( a_1(x) \); and \( \hat{\mu}_{1110} \) denotes the proposed estimator based on two working propensity score models \( \pi_1(x), \pi_2(x) \) and one working outcome model \( a_2(x) \).

The results for \( n = 300 \) and \( n = 1000 \) are shown in Table 1 and 2 respectively. Since \( \hat{\mu}_{LS} \) is multiple robust, the estimators \( \hat{\mu}_{1011}, \hat{\mu}_{0111}, \hat{\mu}_{1101}, \hat{\mu}_{1111} \) and \( \hat{\mu}_{1111} \) are consistent under all four combinations of \( \{ \pi(x), a(x) \} \). Simulation results agreed with this theoretical finding and these estimators had negligible bias under all models. The estimators \( \hat{\mu}_{1010}, \hat{\mu}_{0101}, \hat{\mu}_{1001} \) and \( \hat{\mu}_{0110} \) are doubly robust because only one propensity score model and one outcome regression model were being assumed. We observed that the estimation bias was noticeable when both models were misspecified. It follows that the multiple robust estimators are more robust than the doubly robust estimators. By comparing the root mean squared error (RMSE) of the estimators, we noticed that the multiple robust estimator is efficient. Adding more models did not lead to any noticeable increase in RMSE, consistent with what the theoretical results suggested. In conclusion, the simulation results suggested that multiple robust estimators provide extra protection against model misspecification while estimation efficiency is not compromised.

References


Table 1. Simulation results for the proposed estimator with \( n = 300 \). Values have been multiplied by 100. RMSE denotes root mean squared error.

| \( \pi(x) = \{1 + \exp(0.8 + 0.5x - 0.3x^2)\}^{-1} \) | \( a(x) = 1 + 2x + 3x^2 \) | \( \hat{\mu} \) | Bias | RMSE | \( \pi(x) = 1 - \exp[-\exp(0.5 + 0.5x - 0.3\exp(x))] \) | \( a(x) = 1 + 2x + 3\exp(x) \) | \( \hat{\mu} \) | Bias | RMSE |
|---|---|---|---|---|---|---|---|---|---|---|---|
| \( \mu_{1010} \) | 0 | 49 | 0 | 75 | 0 | 47 | -21 | 76 |
| \( \mu_{1011} \) | 12 | 52 | 0 | 74 | 1 | 50 | 1 | 73 |
| \( \mu_{1001} \) | 0 | 50 | 0 | 74 | 20 | 54 | 1 | 74 |
| \( \mu_{0110} \) | 0 | 49 | -9 | 80 | 0 | 48 | 1 | 74 |
| \( \mu_{0111} \) | 0 | 49 | 0 | 74 | 0 | 48 | 1 | 73 |
| \( \mu_{1101} \) | 0 | 49 | 0 | 74 | 1 | 48 | 1 | 73 |
| \( \mu_{1110} \) | 0 | 49 | 0 | 74 | 0 | 48 | 4 | 75 |
| \( \mu_{1111} \) | 0 | 49 | 0 | 74 | -4 | 50 | 1 | 74 |

Table 2. Simulation results for the proposed estimator with \( n = 1000 \). Values have been multiplied by 100. RMSE denotes root mean squared error.

| \( \pi(x) = \{1 + \exp(0.8 + 0.5x - 0.3x^2)\}^{-1} \) | \( a(x) = 1 + 2x + 3x^2 \) | \( \hat{\mu} \) | Bias | RMSE | \( \pi(x) = 1 - \exp[-\exp(0.5 + 0.5x - 0.3\exp(x))] \) | \( a(x) = 1 + 2x + 3\exp(x) \) | \( \hat{\mu} \) | Bias | RMSE |
|---|---|---|---|---|---|---|---|---|---|---|---|
| \( \mu_{1010} \) | -1 | 25 | 0 | 40 | 0 | 25 | -21 | 46 |
| \( \mu_{1011} \) | 12 | 30 | 0 | 40 | 1 | 27 | 0 | 42 |
| \( \mu_{1001} \) | 0 | 27 | 0 | 40 | 19 | 33 | 0 | 42 |
| \( \mu_{0110} \) | -1 | 25 | -9 | 41 | 0 | 26 | 0 | 42 |
| \( \mu_{0111} \) | -1 | 25 | 0 | 40 | 0 | 26 | 0 | 42 |
| \( \mu_{1101} \) | 0 | 25 | 0 | 40 | 0 | 26 | 1 | 42 |
| \( \mu_{1110} \) | 0 | 25 | 0 | 40 | -1 | 27 | 0 | 42 |
| \( \mu_{1111} \) | 0 | 25 | 0 | 40 | 0 | 26 | 0 | 42 |

Cao, W, Tsiatis, AA & Davidian, M (2009), 'Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data.' *Biometrika*, 96(3), pp. 723–734.


Han, P & Wang, L (2013), 'Estimation with missing data: beyond double robustness.' *Biometrika*, in press.
Simple multiple robust estimator


