Journal of the Korean Data & Information Science Society
2013, **24**(3), 659–667

http://dx.doi.org/10.7465/jkdi.2013.24.3.659
한국데이터정보과학회지

Usage of auxiliary variable and neural network in doubly robust estimation[†]

Hyeonah Park[1] · Wonjun Park[2]

[12]Department of Statistics, Seoul National University

Received 3 April 2013, revised 7 May 2013, accepted 13 May 2013

Abstract

If the regression model or the propensity model is correct, the unbiasedness of the estimator using doubly robust imputation can be guaranteed. Using a neural network instead of a logistic regression model for the propensity model, the estimators using doubly robust imputation are approximately unbiased even though both assumed models fail. We also propose a doubly robust estimator of ratio form using population information of an auxiliary variable. We prove some properties of proposed theory by restricted simulations.

Keywords: Doubly robust estimator, doubly robust imputation, neural network, response probability.

1. Introduction

Missing data frequently occurs in sample survey and one cause is nonresponse. There are two kinds of nonresponses, that is, unit nonresponse and item nonresponse, and imputation is commonly used for compensating for item nonresponses in sample survey. Ratio imputation and regression imputation which are model-based imputations depend on the assumption of the regression model for missing data, where target variables with nonresponse and auxiliary variables without nonresponse are employed. If the assumption of the regression model is wrong, ratio and regression imputations cannot have good properties, for example, estimators using ratio and regression imputations under the incorrect model are not unbiased. Various imputation procedures containing model-based imputations are introduced by Kalton and Kaspryzk (1986), Groves (2002) and Little and Rubin (2002).

When the regression model or the propensity model is correct, estimators using doubly robust imputation has good property such as unbiasedness. In other words, estimators using doubly robust imputation are asymptotically unbiased whether the regression model is correct or not. The properties of the estimators are considered under the assumption that the propensity model is correct. Carpenter and Kenward (2006), Kim and Park (2006), Qin

[†] This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (NRF-2012R1A1A3003761).

[1] Corresponding author: Postdoc, Department of Statistics, Seoul National University, Seoul 151-742, Korea. E-mail: parkha03@yahoo.co.kr

[2] Master candidate, Department of Statistics, Seoul National University, Seoul 151-742, Korea.

et al. (2008) and Cao *et al.* (2009) introduced doubly robust imputation, whose shapes are explained by response values of a target variable, values of auxiliary variables without nonresponse and the estimated propensity function. For the estimation of the propensity function, one can refer to Rosenbaum (1987), Ekholm and Laaksonen (1991), Robins *et al.* (1994), Iannacchione (2003) and Kim and Park (2006). Specially, doubly robust imputation using population information of an auxiliary variable was used in Park *et al.* (2011).

In this paper, neural network is used for the estimation of the propensity function. Estimators using doubly robust imputation by neural network are asymptotically unbiased though both the regression model and the propensity model are incorrect. We suggest a doubly robust estimator of ratio form using population information of an auxiliary variable and a propensity function. In Section 2, estimators using existed ratio imputation and various doubly robust imputation methods to estimate the population mean are summarized and the estimation of the propensity function using the logistic model and neural network is introduced. We also propose an estimator having doubly robust property in Section 2. In Section 3, the result of computational studies under various settings that prove the doubly robust property of the proposed estimators using estimated propensity functions by neural network is provided.

2. Doubly robust estimators and neural network

Let the finite population be $\{(y_i, x_i), i = 1, ..., N\}$, where N is the population size, y_i is the value of the target variable of unit i and x_i is the value of the auxiliary variable. Let the parameter of interest in sample survey be the population mean $\mu_y = N^{-1}\sum_{i=1}^{N} y_i$. We assume that the population information of the auxiliary variable x_i is known. The estimator of the population mean under a probability sampling scheme is $\bar{y}_n = N^{-1}\sum_{i=1}^{n} \pi_i^{-1} y_i$ based on the sample size n, where $\pi_i = P(i \in S)$ is the inclusion probability for element i and S is the set of indices in the sample. The estimator containing inclusion probability is called Horvitz and Thompson (1952) estimator and has the unbiasedness property under the finite population such that

$$E(\bar{y}_n) = \mu_y. \tag{2.1}$$

We define the response indicator variable of y_i as

$$R_i = \begin{cases} 1, & \text{if unit } i \text{ responds} \\ 0, & \text{otherwise} \end{cases}$$

for $i \in S$. Let $\phi_i = P(R_i = 1 | i \in S)$ be the propensity function of sample unit i under the probability sampling scheme. We assume that R_i is ignorable or missing at random(MAR) such that ϕ_i depends on the auxiliary variable x_i but not on the target variable y_i (Little and Rubin, 2002). If we consider the ratio imputation in Rao and Sitter (1995) and Rao (1996), the imputed value for missing data is

$$y_i^I = x_i \left(\sum_{i=1}^{n} \pi_i^{-1} R_i x_i \right)^{-1} \sum_{i=1}^{n} \pi_i^{-1} R_i y_i$$

and the estimator of μ_y based on the ratio imputation can be written by

$$\bar{y}_{RI} = N^{-1}\left[\sum_{i=1}^{n}\pi_i^{-1}R_i y_i + \sum_{i=1}^{n}\pi_i^{-1}(1-R_i)y_i^I\right]. \tag{2.2}$$

If we assume a regression model

$$E(y_i) = x_i r, \quad i = 1, ..., N, \tag{2.3}$$

then the expectation of the estimator (2.2) is the population mean μ_y. If the assumption of the regression model (2.3) is wrong, we cannot guarantee the unbiasedness of the imputed estimator (2.2). In other words, when the regression model is satisfied, we can consider the usage of ratio imputation.

The estimators using doubly robust imputation are unbiased even though the regression model fails. The imputed estimator of Kim and Park (2006) using propensity function for the population mean is

$$\bar{y}_{Id} = N^{-1}\left[\sum_{i=1}^{n}\pi_i^{-1}R_i y_i + \sum_{i=1}^{n}\pi_i^{-1}(1-R_i)y_i^{Id}\right], \tag{2.4}$$

where the imputed value for missing data is

$$y_i^{Id} = x_i\left[\sum_{i=1}^{n}\pi_i^{-1}R_i x_i(\phi_i^{-1}-1)\right]^{-1}\sum_{i=1}^{n}\pi_i^{-1}R_i y_i(\phi_i^{-1}-1).$$

Note that for the regression model or the response model

$$E(\bar{y}_{Id}) = \mu_y + o(n^{-1/2})$$

under some assumptions (Kim and Park, 2006). The doubly robust imputation y_i^{Id} using auxiliary variable and propensity function for missing data decreases the dependence of regression model because the unbiasedness of the estimator using doubly robust imputation can be guaranteed as long as the regression model or the propensity model is correct.

Park *et al.* (2011) proposed a doubly robust imputation using population information of an auxiliary variable to increase the efficiency of an imputed estimator. The imputed value for missing data is

$$y_i^{ARI} = x_i\left[\sum_{i=1}^{n}\pi_i^{-1}R_i x_i(\phi_i^{-1}-1)\right]^{-1}\sum_{i=1}^{n}\pi_i^{-1}R_i y_i\left(\frac{\phi_i^{-1}\sum_{i=1}^{N}x_i}{\sum_{i=1}^{n}\pi_i^{-1}x_i}-1\right)$$

and the estimator using y_i^{ARI} is

$$\bar{y}_{ARI} = N^{-1}\left[\sum_{i=1}^{n}\pi_i^{-1}R_i y_i + \sum_{i=1}^{n}\pi_i^{-1}(1-R_i)y_i^{ARI}\right]. \tag{2.5}$$

The approximative unbiasedness of the imputed estimator (2.5) for population mean was proved in Theorem 1 of Park *et al.* (2011). For the comparison of the efficiency, observe that (2.4) has larger variance than (2.5) when

$$\rho > [2CV(\bar{y}_n)]^{-1} CV(\bar{x}_n),$$

where CV refers to the coefficient of variation, $\bar{x}_n = N^{-1} \sum_{i=1}^{n} \pi_i^{-1} x_i$ and ρ is the correlation coefficient between y_i and x_i (Park *et al.*, 2011).

When population information of an auxiliary variable is known, we propose a simple doubly robust estimator of ratio form using a propensity function. The adjusted ratio estimator using the propensity function is

$$\bar{y}_{Rp} = (\bar{x}_r^{-1} \bar{y}_r) \mu_x, \tag{2.6}$$

where $\bar{x}_r = \sum_{i=1}^{n} \pi_i^{-1} \phi_i^{-1} R_i x_i$, $\bar{y}_r = \sum_{i=1}^{n} \pi_i^{-1} \phi_i^{-1} R_i y_i$ and $\mu_x = N^{-1} \sum_{i=1}^{N} x_i$. By assumptions (9),(10) and (11) in Kim and Park (2006), Taylor expansion of $\bar{x}_r^{-1} \bar{y}_r$ is

$$\bar{x}_r^{-1} \bar{y}_r - r = \left(\sum_{i=1}^{N} x_i \right)^{-1} \left[(\bar{y}_r - \sum_{i=1}^{N} y_i) - r(\bar{x}_r - \sum_{i=1}^{N} x_i) \right] + o_p(n^{1/2}),$$

where $r = (\sum_{i=1}^{N} x_i)^{-1} \sum_{i=1}^{N} y_i$. We can prove that under the regression model (2.3) or the response model,

$$E(\bar{y}_{Rp}) = \mu_y + o(n^{-1/2})$$

and under the propensity model,

$$V(\bar{y}_{Rp}) = V_D \left[N^{-1} \sum_{i=1}^{n} \pi_i^{-1} (y_i - r x_i) \right] + E_D \left[N^{-2} \sum_{i=1}^{n} \pi_i^{-2} (\phi_i^{-1} - 1)(y_i - r x_i)^2 \right] + o(n^{-1}),$$

where E_D and V_D are expectation and variance under sample design.

The propensity function used in this doubly robust imputation is also contained in weighting adjustment for a complement of nonresponse. In practice, because the propensity function is unknown, we must estimate it. We consider a parametric method and a nonparametric method to estimate the propensity function. We assume a logistic model for parametric method, for example, and estimate the coefficients. The logistic model with a single auxiliary variable is defined by

$$\phi_i = (1 + \exp(-\alpha_0 - \alpha_1 x_i))^{-1}. \tag{2.7}$$

We use the maximum likelihood method to estimate the logistic model and compute the values of α_0 and α_1 iteratively using the Newton-Raphson method.

A nonparametric method for estimating the propensity function is to use the method of neural network among data mining methods which is a learning algorithm studied from the structure of biological neural networks and a complex nonlinear method between input variables and output variables. A typical neural network is made up of input layers, hidden layers and output layer, called multi-layer perceptron and linked with activation function and output function. Backpropagation algorithm is also used to decide the weights of neural networks. When the neural network has one input node, one output node and single hidden layer, the model of neural network is

$$\begin{aligned} z_m &= \sigma(\alpha_{0m} + \alpha_m x), \ m = 1, ..., M \\ t &= \beta_0 + \sum_{i=1}^{M} \beta_i z_i \\ f(x) &= g(t), \end{aligned} \tag{2.8}$$

where $\sigma(\cdot)$ is the activation function, $g(\cdot)$ is the output function and $f(x)$ is the value of output node for input node x. Various contents for neural network were referred in Hastie *et al.* (2009) and Izenman (2008). Another data mining technique except neural network was introduced by Cho and Park (2012).

Secondly, we divide the set of indices of the population $U = \{1, 2, \ldots, N\}$ into some nonresponse cells $U = \bigcup_{g=1}^{G} U_g$. A nonresponse cell consists of sample units which are believed to have approximately equal propensity functions. When the set of indices of the sample $S = \{1, 2, \ldots, n\}$ is partitioned into G exhaustive and exclusive nonresponse cells $S = \bigcup_{g=1}^{G} S_g$, an estimated propensity function is

$$\hat{\phi}_g = \frac{\sum_{i \in S_g} \pi_i^{-1} R_i}{\sum_{i \in S_g} \pi_i^{-1}}. \tag{2.9}$$

Little (1986) proposed the method of dividing cells using an estimated logistic model to decide nonresponse cells. We suggest a method of dividing nonresponse cells using a neural network, where the estimated propensity function (2.9) is used in gth nonresponse cell.

3. Computational study

In this section, we provide the result of a limited computational study performed to test and support our theory. In the computational study, $B = 10,000$ samples of size $n = 100, 300$ are generated by a linear model

$$y_i = 3.9x_i + \sqrt{x_i}\epsilon_i \tag{3.1}$$

and by a nonlinear model

$$y_i = (1.8x_i - 1)^2 + \sqrt{x_i}\,\epsilon_i, \tag{3.2}$$

where $x_i \sim \text{Uniform}(0.1, 2.1)$, $\epsilon_i \sim N(0, 1)$ for $i = 1, \ldots, n$, and x_i and ϵ_i are independent (Park *et al.*, 2011). Note that the population mean of target variable y_i is 4.29 for the linear model, and 2.04 for the nonlinear model. The method of selecting samples is simple random sampling and the inclusion probability is $\pi_i = N^{-1}n$. For the propensity function, we use the logistic model of (2.7), where $x_i \sim \text{Uniform}(0.1, 2.1)$ and the value of (α_0, α_1) is assumed to be (-1.0, 3.2) and (-3.3, 4.0) such that the overall response rate becomes 0.83 and 0.63, respectively. We also use the nonlinear response model

$$\phi_i = \beta_0 + \beta_1(1.1 - x_i)^2, \tag{3.3}$$

where the value of (β_0, β_1) is assumed to be (1.0, -0.5) and (0.8, -0.5) such that the overall response rate becomes 0.83 and 0.63, respectively.

For computational setting, we use R Package. The propensity function is estimated by the logistic model and the neural network model. We use weight decay for the penalty of parameters in the neural network whose value is 0.01. The number of nodes in the single hidden layer that make up the neural network is five. The set of indices of the sample is partitioned into eight exhaustive and exclusive nonresponse cells by the estimated logistic

model and the estimated neural network. The equation of (2.9) in each cell is used to estimate the propensity function.

From each simulation (x_i, ϕ_i, R_i, y_i), $i = 1, \ldots, n$, we computed various imputed estimators for the population mean of y_i :

Table 3.1 Doubly robust estimators for the population mean

Estimator	Explanation
$\bar{y}_{Ie1}$	the estimator (2.4) using the propensity function estimated by the logistic model
$\bar{y}_{Ie2}$	the estimator (2.4) using the propensity function estimated by the neural network
$\bar{y}_{Ie3}$	the estimator (2.4) using (2.9) after groups are divided using the propensity function estimated by the logistic model
$\bar{y}_{Ie4}$	the estimator (2.4) using (2.9) after groups are divided using the propensity function estimated by the neural network
$\bar{y}_{Ae1}$	the estimator (2.5) using the propensity function estimated by the logistic model
$\bar{y}_{Ae2}$	the estimator (2.5) using the propensity function estimated by the neural network
$\bar{y}_{Ae3}$	the estimator (2.5) using (2.9) after groups are divided using the propensity function estimated by the logistic model
$\bar{y}_{Ae4}$	the estimator (2.5) using (2.9) after groups are divided using the propensity function estimated by the neural network
$\bar{y}_{Rp1}$	the estimator (2.6) using the propensity function estimated by the logistic model
$\bar{y}_{Rp2}$	the estimator (2.6) using the propensity function estimated by the neural network
$\bar{y}_{Rp3}$	the estimator (2.6) using (2.9) after groups are divided using the propensity function estimated by the logistic model
$\bar{y}_{Rp4}$	the estimator (2.6) using (2.9) after groups are divided using the propensity function estimated by the neural network

In Tables 3.2-3.5, there are computed means, variances and standardized MSEs of the point estimators based on 10,000 samples, where the standardized MSE is the relative MSE compared with that of the complete sample estimator $\bar{y}_n = n^{-1} \sum_{i=1}^{n} y_i$. Simulated values for (x_i, y_i) in Table 3.2 are generated by the linear model (3.1) and simulated values for (ϕ_i, R_i) are generated by the logistic model.

Simulated values for (x_i, ϕ_i, R_i, y_i) in Table 3.3 are generated by the linear model (3.1) and the nonlinear response model (3.3). In Tables 3.2-3.3, the point estimators are unbiased and the estimators $\bar{y}_{Ae1}$, $\bar{y}_{Ae2}$, $\bar{y}_{Ae3}$, $\bar{y}_{Ae4}$, $\bar{y}_{Rp1}$, $\bar{y}_{Rp2}$, $\bar{y}_{Rp3}$, $\bar{y}_{Rp4}$ are more efficient than $\bar{y}_{RI}$, $\bar{y}_{Ie1}$, $\bar{y}_{Ie2}$, $\bar{y}_{Ie3}$, $\bar{y}_{Ie4}$ regardless of sample size and response rate.

Means, variances and standardized MSEs in Table 3.4 are calculated using simulated values of (x_i, ϕ_i, R_i, y_i) which are generated by the nonlinear model (3.2) and the logistic model. Under the nonlinear model, the imputed estimator $\bar{y}_{RI}$ is biased though the estimator shows good properties under the linear model. When response rate is low, bias and MSE of $\bar{y}_{RI}$ become large. It is observed that point estimators except $\bar{y}_{RI}$ are approximately unbiased for large sample size. The estimators $\bar{y}_{Ae1}$, $\bar{y}_{Ae2}$, $\bar{y}_{Ae3}$, $\bar{y}_{Ae4}$, $\bar{y}_{Rp1}$, $\bar{y}_{Rp2}$, $\bar{y}_{Rp3}$, and $\bar{y}_{Rp4}$ are also more efficient than the estimators $\bar{y}_{RI}$, $\bar{y}_{Ie1}$, $\bar{y}_{Ie2}$, $\bar{y}_{Ie3}$, and $\bar{y}_{Ie4}$ regardless of sample size and response rate, which is caused by the fact that the population information of the auxiliary variable is used. The estimators using (2.9) for estimating the propensity function have a little bias.

Simulated values for (x_i, y_i) in Table 3.5 are generated by the nonlinear model (3.2) and simulated values for (ϕ_i, R_i) are generated by the nonlinear response model (3.3). Under the nonlinear model and the nonlinear response model, the estimator $\bar{y}_{RI}$ and the estimators using the estimated propensity function by logistic model $\bar{y}_{Ie1}$, $\bar{y}_{Ae1}$, and $\bar{y}_{Rp1}$ are biased. The estimators $\bar{y}_{Ie2}$, $\bar{y}_{Ie4}$, $\bar{y}_{Ae2}$, $\bar{y}_{Ae4}$, $\bar{y}_{Rp2}$, and $\bar{y}_{Rp4}$ are unbiased when sample size is large and response rate is high.

Table 3.2 Mean, variance, standardized mse for linear model and logistic model

Sample size		n=100			n=300		
(α_0, α_1)	estimator	Mean	Variance	MSE	Mean	Variance	MSE
	$\bar{y}_n$	4.29	0.062	1.000	4.29	0.020	1.000
	$\bar{y}_{RI}$	4.29	0.063	1.016	4.29	0.021	1.050
	$\bar{y}_{Ie1}$	4.29	0.064	1.032	4.29	0.021	1.050
	$\bar{y}_{Ie2}$	4.29	0.064	1.032	4.29	0.021	1.050
	$\bar{y}_{Ie3}$	4.29	0.064	1.032	4.29	0.021	1.050
	$\bar{y}_{Ie4}$	4.29	0.064	1.032	4.29	0.021	1.050
$(-1.0, 3.2)$	$\bar{y}_{Ae1}$	4.29	0.013	0.210	4.29	0.004	0.200
	$\bar{y}_{Ae2}$	4.29	0.013	0.210	4.29	0.004	0.200
	$\bar{y}_{Ae3}$	4.29	0.013	0.210	4.29	0.004	0.200
	$\bar{y}_{Ae4}$	4.29	0.013	0.210	4.29	0.004	0.200
	$\bar{y}_{Rp1}$	4.29	0.013	0.210	4.29	0.004	0.200
	$\bar{y}_{Rp2}$	4.29	0.013	0.210	4.29	0.004	0.200
	$\bar{y}_{Rp3}$	4.29	0.013	0.210	4.29	0.004	0.200
	$\bar{y}_{Rp4}$	4.29	0.013	0.210	4.29	0.004	0.200
	$\bar{y}_n$	4.29	0.061	1.000	4.29	0.020	1.000
	$\bar{y}_{RI}$	4.29	0.064	1.049	4.29	0.021	1.050
	$\bar{y}_{Ie1}$	4.29	0.070	1.148	4.29	0.023	1.150
	$\bar{y}_{Ie2}$	4.29	0.069	1.131	4.29	0.023	1.150
	$\bar{y}_{Ie3}$	4.29	0.071	1.164	4.29	0.023	1.150
	$\bar{y}_{Ie4}$	4.29	0.071	1.164	4.29	0.023	1.150
$(-3.3, 4.0)$	$\bar{y}_{Ae1}$	4.29	0.021	0.344	4.29	0.007	0.350
	$\bar{y}_{Ae2}$	4.29	0.020	0.328	4.29	0.007	0.350
	$\bar{y}_{Ae3}$	4.29	0.023	0.377	4.29	0.007	0.350
	$\bar{y}_{Ae4}$	4.29	0.023	0.377	4.29	0.007	0.350
	$\bar{y}_{Rp1}$	4.29	0.022	0.361	4.29	0.007	0.350
	$\bar{y}_{Rp2}$	4.29	0.019	0.311	4.29	0.006	0.300
	$\bar{y}_{Rp3}$	4.29	0.022	0.361	4.29	0.007	0.350
	$\bar{y}_{Rp4}$	4.29	0.021	0.344	4.29	0.007	0.350

Table 3.3 Mean, variance, standardized mse for linear model and nonlinear response model

Sample size		n=100			n=300		
(β_0, β_1)	estimator	Mean	Variance	MSE	Mean	Variance	MSE
	$\bar{y}_n$	4.29	0.062	1.000	4.29	0.021	1.000
	$\bar{y}_{RI}$	4.29	0.064	1.032	4.29	0.022	1.048
	$\bar{y}_{Ie1}$	4.29	0.064	1.032	4.29	0.022	1.048
	$\bar{y}_{Ie2}$	4.29	0.065	1.048	4.29	0.022	1.048
	$\bar{y}_{Ie3}$	4.29	0.065	1.048	4.29	0.022	1.048
	$\bar{y}_{Ie4}$	4.29	0.065	1.048	4.29	0.022	1.048
$(1.0, -0.5)$	$\bar{y}_{Ae1}$	4.29	0.013	0.210	4.29	0.004	0.190
	$\bar{y}_{Ae2}$	4.29	0.014	0.226	4.29	0.005	0.238
	$\bar{y}_{Ae3}$	4.29	0.014	0.226	4.29	0.005	0.238
	$\bar{y}_{Ae4}$	4.29	0.016	0.258	4.29	0.005	0.238
	$\bar{y}_{Rp1}$	4.29	0.013	0.210	4.29	0.004	0.190
	$\bar{y}_{Rp2}$	4.29	0.014	0.226	4.29	0.005	0.238
	$\bar{y}_{Rp3}$	4.29	0.014	0.226	4.29	0.005	0.238
	$\bar{y}_{Rp4}$	4.29	0.014	0.226	4.29	0.005	0.238
	$\bar{y}_n$	4.29	0.062	1.000	4.29	0.021	1.000
	$\bar{y}_{RI}$	4.29	0.069	1.113	4.29	0.023	1.095
	$\bar{y}_{Ie1}$	4.29	0.069	1.113	4.29	0.023	1.095
	$\bar{y}_{Ie2}$	4.29	0.070	1.129	4.29	0.023	1.095
	$\bar{y}_{Ie3}$	4.29	0.071	1.145	4.29	0.023	1.095
	$\bar{y}_{Ie4}$	4.29	0.072	1.161	4.29	0.023	1.095
$(0.8, -0.5)$	$\bar{y}_{Ae1}$	4.29	0.018	0.290	4.29	0.006	0.286
	$\bar{y}_{Ae2}$	4.29	0.019	0.306	4.29	0.006	0.286
	$\bar{y}_{Ae3}$	4.29	0.020	0.323	4.29	0.006	0.286
	$\bar{y}_{Ae4}$	4.29	0.021	0.339	4.29	0.006	0.286
	$\bar{y}_{Rp1}$	4.29	0.018	0.290	4.29	0.006	0.286
	$\bar{y}_{Rp2}$	4.29	0.019	0.306	4.29	0.006	0.286
	$\bar{y}_{Rp3}$	4.29	0.020	0.323	4.29	0.006	0.286
	$\bar{y}_{Rp4}$	4.29	0.021	0.339	4.29	0.006	0.286

Table 3.4 Mean, variance, standardized mse for nonlinear model and logistic model

Sample size		n=100			n=300		
(α_0, α_1)	estimator	Mean	Variance	MSE	Mean	Variance	MSE
$(-1.0, 3.2)$	$\bar{y}_n$	2.04	0.062	1.000	2.04	0.020	1.000
	$\bar{y}_{RI}$	2.14	0.065	1.210	2.14	0.021	1.550
	$\bar{y}_{Ie1}$	2.04	0.064	1.032	2.04	0.021	1.050
	$\bar{y}_{Ie2}$	2.04	0.064	1.032	2.04	0.021	1.050
	$\bar{y}_{Ie3}$	2.06	0.064	1.039	2.06	0.021	1.070
	$\bar{y}_{Ie4}$	2.06	0.064	1.039	2.06	0.021	1.070
	$\bar{y}_{Ae1}$	2.04	0.032	0.516	2.04	0.011	0.550
	$\bar{y}_{Ae2}$	2.04	0.031	0.500	2.04	0.010	0.500
	$\bar{y}_{Ae3}$	2.05	0.032	0.518	2.06	0.011	0.570
	$\bar{y}_{Ae4}$	2.05	0.032	0.518	2.06	0.011	0.570
	$\bar{y}_{Rp1}$	2.04	0.032	0.516	2.04	0.011	0.550
	$\bar{y}_{Rp2}$	2.04	0.031	0.500	2.04	0.010	0.500
	$\bar{y}_{Rp3}$	2.05	0.032	0.518	2.05	0.010	0.505
	$\bar{y}_{Rp4}$	2.05	0.032	0.518	2.05	0.010	0.505
$(-3.3, 4.0)$	$\bar{y}_n$	2.04	0.062	1.000	2.04	0.021	1.000
	$\bar{y}_{RI}$	2.35	0.068	2.647	2.35	0.023	5.671
	$\bar{y}_{Ie1}$	2.04	0.075	1.210	2.04	0.025	1.190
	$\bar{y}_{Ie2}$	2.04	0.072	1.161	2.04	0.024	1.143
	$\bar{y}_{Ie3}$	2.05	0.076	1.227	2.04	0.025	1.190
	$\bar{y}_{Ie4}$	2.05	0.076	1.227	2.04	0.025	1.190
	$\bar{y}_{Ae1}$	2.03	0.043	0.695	2.04	0.014	0.667
	$\bar{y}_{Ae2}$	2.03	0.039	0.631	2.03	0.013	0.624
	$\bar{y}_{Ae3}$	2.05	0.042	0.679	2.04	0.014	0.667
	$\bar{y}_{Ae4}$	2.05	0.042	0.679	2.04	0.014	0.667
	$\bar{y}_{Rp1}$	2.03	0.043	0.695	2.04	0.014	0.667
	$\bar{y}_{Rp2}$	2.04	0.040	0.645	2.04	0.013	0.619
	$\bar{y}_{Rp3}$	2.05	0.045	0.727	2.03	0.014	0.671
	$\bar{y}_{Rp4}$	2.05	0.045	0.727	2.03	0.014	0.671

Table 3.5 Mean, variance, standardized mse for nonlinear model and nonlinear response model

Sample size		n=100			n=300		
(β_0, β_1)	estimator	Mean	Variance	MSE	Mean	Variance	MSE
$(1.0, -0.5)$	$\bar{y}_n$	2.04	0.062	1.000	2.04	0.021	1.000
	$\bar{y}_{RI}$	1.87	0.061	1.450	1.87	0.021	2.376
	$\bar{y}_{Ie1}$	1.87	0.064	1.498	1.87	0.021	2.376
	$\bar{y}_{Ie2}$	2.03	0.066	1.066	2.04	0.022	1.048
	$\bar{y}_{Ie3}$	2.02	0.066	1.071	2.03	0.022	1.052
	$\bar{y}_{Ie4}$	2.03	0.068	1.098	2.04	0.022	1.048
	$\bar{y}_{Ae1}$	1.87	0.034	1.015	1.87	0.011	1.900
	$\bar{y}_{Ae2}$	2.03	0.033	0.534	2.04	0.011	0.524
	$\bar{y}_{Ae3}$	2.02	0.034	0.555	2.02	0.011	0.543
	$\bar{y}_{Ae4}$	2.03	0.034	0.550	2.04	0.011	0.524
	$\bar{y}_{Rp1}$	1.87	0.034	1.015	1.87	0.011	1.900
	$\bar{y}_{Rp2}$	2.02	0.034	0.555	2.04	0.011	0.524
	$\bar{y}_{Rp3}$	2.02	0.034	0.555	2.02	0.011	0.543
	$\bar{y}_{Rp4}$	2.02	0.037	0.603	2.04	0.011	0.524
$(0.8, -0.5)$	$\bar{y}_n$	2.04	0.062	1.000	2.04	0.021	1.000
	$\bar{y}_{RI}$	1.81	0.070	1.982	1.81	0.023	3.614
	$\bar{y}_{Ie1}$	1.82	0.070	1.910	1.81	0.023	3.614
	$\bar{y}_{Ie2}$	2.01	0.072	1.176	2.03	0.024	1.148
	$\bar{y}_{Ie3}$	2.02	0.073	1.184	2.02	0.024	1.162
	$\bar{y}_{Ie4}$	2.02	0.078	1.265	2.03	0.024	1.148
	$\bar{y}_{Ae1}$	1.81	0.041	1.515	1.81	0.014	3.186
	$\bar{y}_{Ae2}$	2.01	0.040	0.660	2.03	0.013	0.624
	$\bar{y}_{Ae3}$	2.01	0.041	0.676	2.02	0.013	0.638
	$\bar{y}_{Ae4}$	2.01	0.045	0.740	2.03	0.013	0.624
	$\bar{y}_{Rp1}$	1.81	0.041	1.515	1.81	0.014	3.186
	$\bar{y}_{Rp2}$	2.00	0.040	0.671	2.02	0.013	0.638
	$\bar{y}_{Rp3}$	2.01	0.041	0.676	2.02	0.013	0.638
	$\bar{y}_{Rp4}$	2.01	0.047	0.773	2.03	0.014	0.671

4. Concluding remarks

If the regression model is incorrect and the propensity model is correct, the estimator using regression imputation is biased but the estimators using doubly robust imputation

are unbiased. Doubly robust estimators using estimated propensity function by neural network are asymptotically unbiased but doubly robust estimators using estimated propensity function by logistic model are biased as long as the regression model is incorrect and nonlinear response model for propensity model is used. Simulation results were suggested to test our theory under some restricted population model and propensity model in this thesis. An auxiliary variable with population information is only used for doubly robust estimators in this paper. Doubly robust estimators using various auxiliary variables and data mining techniques for estimating propensity function can be researched in this future.

References

Carpenter, J. R. and Kenward, M. G. (2006). A comparison of multiple imputation and doubly robust estimation for analyses with missing data. *Journal of the Royal Statistical Society A*, **169**, 571-584.

Cao, W., Tsiatis, A. A. and Davidian, M. (2009). Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika*, **96**, 723-734.

Cho, K. H. and Park, H. C. (2012). A study on decision tree creation using marginally conditional variables. *Journal of the Korean Data & Information Science Society*, **23**, 299-307.

Ekholm, A. and Laaksonen, S. (1991). Weighting via response modeling in the finnish household budget survey. *Journal of Official statistics*, **7**, 325-337.

Groves, R., Dillman, D., Eltinge, J. and Little, R. J. A. (2002). *Survey nonresponse*, Wiley, New York.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The element of statistical learning: Data mining, inference, and prediction*, 2nd edition, Springer, New York.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Society*, **47**, 663-685.

Iannacchione, V. G. (2003). Sequential weight adjustment for location and cooperation propensity for the 1995 national survey of family growth. *Journal of Official statistics*, **19**, 31-43.

Izenman, A. J. (2008). *Modern multivariate statistical techniques: Regression, classification, and manifold learning*, Springer, New York.

Kalton, G. and Kasprzyk, D. (1986). The treatment of missing survey data. *Survey Methodology*, **12**, 1-16.

Kim, J. K. and Park, H. (2006). Imputation using response probability. *The Canadian Journal of Statistics*, **34**, 171-182.

Little, R. J. A. (1986). Survey nonresponse adjustments for estimates of means. *International Statistical Review*, **54**, 139-157.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical analysis with missing data*, Wiley, New York.

Park, H., Jeon, J. and Na, S. (2011). Doubly robust imputation using auxiliary information. *Communications of the Korean Statistical society*, **18**, 47-55.

Qin,J., Shao, J. and Zhang, B. (2008). Efficient and doubly robust imputation for covariate-dependent missing responses. *Journal of the American Statistical Association*, **103**, 797-810.

Rao, J. N. K. and Sitter, R. R. (1995). Variance estimation under two-phase sampling with application to imputation for missing data. *Biometrika*, **82**, 453-460.

Rao, J. N. K. (1996). On variance estimation with imputed survey data. *Journal of the American Statistical Association*, **91**, 499-506.

Robins, J. M, Rotnitzky, A. and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, **89**, 846-866.

Rosenbaum, P. R. (1987). Model-based direct adjustment. *Journal of the American Statistical Association*, **82**, 387-394.