Missing Data under the Matched-Pair Design: A Practical Guide *

Kentaro Fukumoto †

July 17, 2015

*Paper prepared for delivery at the 32nd Annual Summer Meeting of Society for Political Methodology, Rochester, July 23-25, 2015. Earlier versions, whose title was "Blocking Reduces, if not Removes, Attrition Bias," were presented at the Workshop on Directional Statistics, the Institute of Statistical Mathematics, Tokyo, July 5, 2012, the Symposium on Incomplete Data Analysis and Causal Inference, Osaka University, September 22-23, 2013 and the 1st Asian Political Methodology Meeting, Tokyo Institute of Technology, January 6-7, 2014, the Annual Meeting of the Midwest Political Science Association, Chicago, April 16-19, 2015, and workshops at Dartmouth, UCSD, and Washington University in St. Louis. I appreciate comments from Raymond Duch, Jeff Gill, Donald Green, Yusaku Horiuchi, Kosuke Imai, Luke Keele, Gary King, Dean Lacy, Ryan T. Moore, Brendan Nyhan, Keith Schnakenberg, and Susumu Shikano. The usual disclaimer applies.

†Professor, Department of Political Science, Faculty of Law, Gakushuin University, 1-5-1 Mejiro, Toshima, Tokyo 171-8588 Japan. Tel: +81-3-3986-0221 ext. 4913. Fax: +81-3-5992-1006. Email: Kentaro.Fukumoto@gakushuin.ac.jp. URL: http://www-cc.gakushuin.ac.jp/~e982440/index_e.htm.

Abstract

The matched-pair design in experiments and matching in observational studies are powerful tools that enable robust causal inference. What if outcomes of some units are missing? First, deleting units with missing will lead to bias. Second, some scholars recommend deleting the other unit in the pair, while others argue that it also results in biased estimates if missing is not independent of potential outcomes. Third, imputation is problematic unless missing is at random. By using the potential outcome framework, this study formalizes estimation error and bias of each method, and establishes that, when the design and implementation of research is sufficiently good, blockwise deletion is better than others. This argument is supported by application of these methods to experimental data of the Mexican universal health insurance program (King et al., 2007) and non-experimental matched data of training programs (Dehejia and Wahba, 1999; Lalonde, 1986). This paper also shows that the single imputation method is equivalent to blockwise deletion.

INTRODUCTION

The matched-pair design in experiments (Imai, King, and Nall, 2009*a*; Moore, 2012; Paluck and Green, 2009) and matching in observational studies (Ho et al., 2007) are powerful tools to enable pretreatment balance and, therefore, robust causal inference of average treatment effects (hereafter, ATEs) on an outcome. In the matched-pair design, experimenters make pairs (or blocks) of two units, each of which shares the exact or simiar values of some matched-on variables, and randomize treatment assignment within each pair ("block what you can and randomize what you cannot" (Box, Hunter, and Hunter, 1978, 103). This paper uses the words "pair" and "block" exchangeably). Or, in the case of observational studies, analysts match pairs of treated and controlled units so that their covariates are as similar as possible within each pair and discard non-matched units. In either case, the difference-in-means of outcomes between the treated and controlled groups estimates the ATE without bias.

Unfortunately, however, outcome values of units are sometimes missing. This problem arises more frequently than generally acknowledged. For instance, in the case of randomized controlled trial, subjects may die, be tired of responding to follow-up survey (Enos, 2014), or be interrupted politically (King et al., 2007) or elections are uncontested (Panagopoulos and Green, 2008). When it comes to observational study, examples are that legislators retire, or duration such as political career is censored (Fukumoto and Masuyama, 2015).

The first, but naive, approach to missing data is the unitwise deletion estimator, that is, to delete missing observations alone and apply the difference-in-means estimator to all the remaining units to estmate the ATE. An application example is Panagopoulos and Green (2008). It is, however, well known that the unitwise deletion estimator can be biased unless missing is (completely) at random (e.g. Honaker, King, and Blackwell, 2012).

The second, textbook tool to address attrition is the blockwise deletion estimator, which deletes missing units as well as the other units in the same pairs (namely, blocks) and calculate difference-in-means by using units in the remaining *blocks* only (Donner and Klar, 2000, 40, Imai, King, and Nall, 2009*a*, 44, Moore, 2010). Application examples are Enos (2014) and Family Heart Study Group (1994). It is true that blockwise deletion protects *balance of matched-on variables* which would result without missing outcomes, though it is not sufficient to guarantee *unbiased estimation of the ATE*. Rather, as Gerber and Green (2012) warn and this paper also formalizes, the blockwise deletion estimator can "lead[s] to bias when attrition is a function of potential outcomes" (p. 243). Though King et al. (2007) argue that the blockwise deletion estimator is unbiased as long as units are missing "for a reason related to one or more of the variables we matched on" (p. 490), this paper shows that this statement does not necessarily hold in the presence of heterogeneous treatment effects. In addition, while Imai, King, and Nall (2009*a*) claim that the blockwise deletion estimator "retain[s] the benefits of randomization ... regardless of the missing data mechanism" (p. 44), this paper demonstrates that the missing data mechanism matters for how much biased the blockwise deletion estimator is.

The third, general fix for unobserved outcome is the imputation estimators. There are two variants. On the one hand, the single imputation estimator imputes one value for every missing outcome by OLS and derives the difference-in-means using both observed and imputed outcomes (Giesbrecht and Gumpertz, 2004, 102, Hinkelmann and Kempthorne, 1994, 265-271,Yates, 1933=1970. Details will be explained later). In fact, this paper shows that, in the matched-pair design, the single imputation estimator is equivalent to the blockwise deletion estmator. On the other hand, the multiple imputation estimator imputes several values for each missing outcome, calculates the differences-in-means for each imputed datasets, and uses their average as the estimate of the ATE. There are some methods to impute missing outcomes, while this paper focuses on Amelia II (Honaker, King, and Blackwell, 2012). Here is a caveat, though. Multiple imputation assumes "that the data are missing at random ... that the pattern of missingness only depends on the observed data ... not the *unobserved* data" (Honaker, King, and Blackwell, 2012, 4, emphasis added). This assumption is violated when attrition is a function of *potential* outcomes. Therefore, imputation of missing values does not necessarily promise solution to attrition bias.

To sum, all the three estimators scholars have uses – unitwise deletion, blockwise deletion, and multiple imputation – can be biased in the presence of missing outcome.[1] Which estimator should we use, practically? Previous works do not examine, much less compare, the degree to which these estimators suffer from attrition bias. By employing the potential outcome framework, this paper formalizes estimation error and bias of the unitwise deletion and blockwise deletion estimators in various conditions. This study also applies the three estimators to experimental data of the Mexican universal health insurance program (King et al., 2007) and observational data of training programs (Dehejia and Wahba, 1999; Lalonde, 1986). In a nutshell, the message of this paper is that the blockwise deletion estimator is better than others as long as the design achieves balance of covariates sufficiently and treatment effect is not so heterogeneous. Moreover, though King et al. (2007, 490) argue that "whether we delete or impute" does not matter for *balance of covariates*, this paper shows that it does matter for *estimation of the ATE.*

The plan of the paper is as follows. The next section formalizes estimators, lays out the potential outcome framework, and characterises estimation error and bias of estimators under some scenarios. The following section demonstrates two applications. The final section concludes.

FORMAL ANALYSIS

The matched-pair design of either experiments or observational studies is formalized as follows. Suppose that there are $\bar{b}$ blocks and each block is composed of two units. Let $Y_{b,u}$ denote the outcome of unit $u \in \{1, 2\}$ in block $b \in \{1, 2, \ldots, \bar{b}\}$ (which is called unit (b, u)). Moreover, let $R_{b,u}$ denote the response of unit (b, u). $R_{b,u}$ is equal to 1 if $Y_{b,u}$ is observed, and 0 if $Y_{b,u}$ is missing. Finally, let $T_{b,u}$ denote the treatment status of unit (b, u). $T_{b,u}$ is

[1] If a block is composed of more than two units and has at least one treated unit and one controlled unit observed, inverse probability weighting is also avialable (Gerber and Green, 2012, 222).

equal to 1 if unit (b, u) is assigned and recieves treatment, and 0 otherwise.[2] For every block b, it holds that either $T_{b,1} = 1, T_{b,2} = 0$ or $T_{b,1} = 0, T_{b,2} = 1$. At this moment, we have not yet assumed that treatment is randomly assigned. [3]

Estimators

Let $W_{b,u}$ denote genetic weight dummy variable of unit (b, u). Weighted mean of outcome is calculated as:

$$E(W \circ Y) \equiv \frac{\sum_{b=1}^{\bar{b}} \sum_{u=1}^{2} W_{b,u} Y_{b,u}}{\sum_{b=1}^{\bar{b}} \sum_{u=1}^{2} W_{b,u}}.$$

where, for genetic variable $V_{b,u}$, its vector is defined as $V \equiv (V_{1,1}, V_{1,2}, V_{2,1}, V_{2,2}, \ldots, V_{\bar{b},1}, V_{\bar{b},2})$ and we assume that $\sum_{b=1}^{\bar{b}} \sum_{u=1}^{2} W_{b,u} \neq 0$. A genetic estimator of the ATE (which we will define formally in the next subsection) is defined as difference in *weighted* means:

$$\hat{\tau} \equiv E(W^T \circ Y) - E(W^C \circ Y).$$

It is easy to see that $W_{b,u}^T$ and $W_{b,u}^C$ indicate whether unit (b, u) is counted for calculating the average of outcomes for treated and controlled groups, respectively. Below, different estimators have different weights.

If no unit is missing, the full sample estimator is available;

$$\hat{\tau}_F \equiv E(W_F^T \circ Y) - E(W_F^C \circ Y)$$
$$W_{F,b,u}^T \equiv T_{b,u},$$
$$W_{F,b,u}^C \equiv 1 - T_{b,u}.$$

[2]This paper assumes that all units comply treatment assignment.

[3]If we assume that treatment is randomly assigned, this setup is random allocation of treatment in each block and is not complete randomization (Moore, 2012), where $T_{b,u} \perp\!\!\!\perp T_{b',u'}$ unless $b = b'$ and $u = u'$, and it is possible that $T_{b,1} = 1, T_{b,2} = 1$ or $T_{b,1} = 0, T_{b,2} = 0$.

The unitwise deletion estimator is defined as

$$\hat{\tau}_U \equiv E(W_U^T \circ Y) - E(W_U^C \circ Y)$$
$$W_{U,b,u}^T \equiv R_{b,u} T_{b,u},$$
$$W_{U,b,u}^C \equiv R_{b,u}(1 - T_{b,u}).$$

The blockwise deletion estimator is defined as

$$\hat{\tau}_B \equiv E(W_B^T \circ Y) - E(W_B^C \circ Y)$$
$$W_{B,b,u}^T \equiv R_{b,u} R_{b,-u} T_{b,u},$$
$$W_{B,b,u}^C \equiv R_{b,u} R_{b,-u}(1 - T_{b,u}).$$

The multiple imputation estimator is defined as

$$\hat{\tau}_{MI} \equiv \frac{1}{5} \sum_{i=1}^{5} [E(W_F^T \circ \hat{Y}^{(i)}) - E(W_F^C \circ \hat{Y}^{(i)})],$$

where $\hat{Y}^{(i)}$ is the i-th imputed data of Y by Amelia II. Note that the weights are the same as those of the full sample estimator.

The single imputation estimator is defined as

$$\hat{\tau}_{SI} \equiv E(W_F^T \circ \hat{Y}^{(S)}) - E(W_F^C \circ \hat{Y}^{(S)}),$$

where $\hat{Y}^{(S)}$ is the imputed data of Y by the classic method of Hinkelmann and Kempthorne (1994, 67-72, 265-271). Briefly, it is introduced here (for details, see Appendix). Rearrange the order of (b, u) so that $Y_{b,u}$ of the first L units, $Y_{R=1}$, are observed and $Y_{b,u}$ of the next M units, $Y_{R=0}$, are missing ($L + M = 2\bar{b}$). Let X the $2\bar{b} \times (\bar{b} + 1)$ design matrix such that, for

$k \leq \bar{b}$,

$$X_{(b,u),k} = \begin{cases} 1 & \text{if} \quad b = k \\ 0 & \text{otherwise} \end{cases}$$

and

$$X_{(b,u),\bar{b}+1} = T_{b,u},$$

where $X_{(b,u),k}$ denotes the element of X in the row of unit u of block b and the k-th column. Let matrices U, V and W such that

$$(I_{2\bar{b}} - X(X'X)^{-1}X') = \begin{pmatrix} U_{L\times L} & V_{L\times M} \\ V'_{M\times L} & W_{M\times M} \end{pmatrix},$$

where I_K is the $K \times K$ identity matrix. Then, in the spirit of OLS, the missing values are imputed by

$$\hat{Y}_{R=0} = -W^{-1}V'Y_{R=1}$$

and the imputed outcomes are $\hat{Y}^{(S)} = (Y_{R=1}, \hat{Y}_{R=0})$.

To my knowledge, this paper establishes the following theorem for the first time (all proofs are in Appendix).

Theorem 1 *Equivalence between the single imputation estimator and the blockwise deletion estimator.*

$$\hat{\tau}_{SI} = \hat{\tau}_B$$

Thus, we do not have to examine the single imputation estimator, if we do the blockwise deletion estimator. This theorem also gives another rationale of the blockwise deletion estimator. Moreover, if two units of the same block become missing at the same time ($R_{b,1} = R_{b,2} = 0$), W is not invertible and, thus, the single imputation estimator does not work but the blockwise deletion estimator does. Therefore, practitioners only have to employ the blockwise deletion estimator instead of the single imputation estimator.

Needless to say, if no unit has missing outcome, all estimators are reduced to the full sample estimator because, for all b and u, $R_{b,u} = 1$ and $\hat{Y}^{(i)} = \hat{Y}^{(S)} = Y$.

Potential Outcome and Response

The conventional stable unit treatment value assumption (SUTVA) is made. Let $Y_{b,u}(1)$ and $Y_{b,u}(0)$ denote potential outcomes of unit (b, u) in the case of treatment and control received, respectively. The realized (but not necessarily observed) outcome can be expressed as $Y_{b,u} = T_{b,u}Y_{b,u}(1) + (1 - T_{b,u})Y_{b,u}(0)$. The estimand, the ATE, can be defined in the same formula as estimators;

$$\bar{\tau} \equiv E(W_P^T \circ Y(1)) - E(W_P^C \circ Y(0))$$

$$W_{P,b,u}^T \equiv 1,$$

$$W_{P,b,u}^C \equiv 1.$$

This paper decomposes potential outcomes into six parts;

$$Y_{b,u}(t) = \quad \alpha + (\bar{\tau} + \Delta_b^B \tau + \Delta_{b,u}^W \tau)t + \Delta_b^B \epsilon + \Delta_{b,u}^W \epsilon$$

provided that

$$\sum_{b=1}^{\bar{b}} \Delta_b^B \tau = \sum_{u=1}^{2} \Delta_{b,u}^W \tau = \sum_{b=1}^{\bar{b}} \Delta_b^B \epsilon = \sum_{u=1}^{2} \Delta_{b,u}^W \epsilon = 0,$$

where $t \in \{0, 1\}$ (for details, see Appendix). This decomposition is unique. This paper refers to α as the average of control potential outcomes, $\Delta_b^B \epsilon$ as between-block error, $\Delta_{b,u}^W \epsilon$ as within-block error, $\Delta_b^B \tau$ as between-block heterogeneity of treatment effects, and $\Delta_{b,u}^W \tau$ as within-block heterogeneity of treatment effects.

In addition to standard potential outcomes, this paper formalizes missing in the same framework. Let $R_{b,u}(1)$ and $R_{b,u}(0)$ denote potential *responses* of unit (b, u) in the case of treatment and control received, respectively.[4] $R_{b,u}(1)$ is equal to 1 if $Y_{b,u} = Y_{b,u}(1)$ would be observed, and 0 if $Y_{b,u} = Y_{b,u}(1)$ would be missing, in the case of *treatment* received. $R_{b,u}(0)$ is equal to 1 if $Y_{b,u} = Y_{b,u}(0)$ would be observed, and 0 if $Y_{b,u} = Y_{b,u}(0)$ would be missing, in the case of *control* received. $R_{b,u}(1)$ and $R_{b,u}(0)$ are fixed as $Y_{b,u}(1)$ and $Y_{b,u}(0)$ are ($T_{b,u}$ are stochastic). The realized reponse can be rewritten as $R_{b,u} = T_{b,u}R_{b,u}(1) + (1 - T_{b,u})R_{b,u}(0)$.

Below, this paper examines and compares estimation error (in the next subsection) and bias (in the following subsection) of the unitwise deletion, blockwise deletion, and, as a reference, full sample estimators.[5]

Estimation Error

Denote the estimation error of a genetic estimator $\hat{\tau}$ by $\Delta\hat{\tau} \equiv \hat{\tau}_U - \bar{\tau}$. We have the following proposition.

Proposition 1 (Estimation Error)

$$
\begin{aligned}
\Delta\hat{\tau}_U =& E(W_U^T \circ (\Delta^B\epsilon + \Delta^W\epsilon + \Delta^B\tau + \Delta^W\tau)) - E(W_U^C \circ (\Delta^B\epsilon + \Delta^W\epsilon)) \\
\Delta\hat{\tau}_B =& E(W_B^T \circ (2\Delta^W\epsilon + \Delta^B\tau + \Delta^W\tau)) \\
\Delta\hat{\tau}_F =& E(W_F^T \circ (2\Delta^W\epsilon + \Delta^W\tau)).
\end{aligned}
$$

To put another way, $\hat{\tau}_B$ is free of $\Delta^B\epsilon$, while $\hat{\tau}_F$ is free of $\Delta^B\epsilon$ and $\Delta^B\tau$. This paper emphasizes that, unlike bias, Proposition 1 holds in any one trial without any assumption on treatment assignment (such as randomness). A caveat is that the order of the estimation error size among the three estimators, $|\Delta\hat{\tau}|$, depends.

When $\Delta_{b,u}^W\epsilon = \Delta_{b,u}^W\tau = 0$, this paper calls the matched-pair design "perfect." The reason is that the realized outcome of one unit is the counter-factual of the other unit in the same

[4] For this notation, see, for instance, Frangakis and Rubin (1999), Gerber and Green (2012, ch. 7), and Morton and Williams (2010, 182-192).

[5] This subsection does not consider the multiple imputation estimator because its property depends on covariates Amelia II employs.

pair:

$$Y_{b,1}(0) = Y_{b,2}(0) = \alpha + \Delta_b^B \epsilon$$

$$Y_{b,1}(1) = Y_{b,2}(1) = \alpha + \Delta_b^B \epsilon + \bar{\tau} + \Delta_b^B \tau.$$

Corollary 1 is easily derived by substituting $\Delta_{b,u}^W \epsilon = \Delta_{b,u}^W \tau = 0$ in Proposition 1.

Corollary 1 (Estimation Error under Perfect Matched-Pair Design) *When* $\Delta_{b,u}^W \epsilon = \Delta_{b,u}^W \tau = 0$,

$$\Delta\hat{\tau}_U = E(W_U^T \circ (\Delta^B \epsilon + \Delta^B \tau)) - E(W_U^C \circ \Delta^B \epsilon)$$

$$\Delta\hat{\tau}_B = E(W_B^T \circ \Delta^B \tau)$$

$$\Delta\hat{\tau}_F = 0.$$

When the matched-pair design is perfect, usually due to covariate balance, the estimation errors are free of within-block factors, $\Delta^W \equiv \{\Delta^W \epsilon, \Delta^W \tau\}$. Note that the full sample estimator is free of estimation error. Moreover, even if the matched-pair design is not perfect but *nearly* so, namely, if $\Delta_{b,u}^W \epsilon$ and $\Delta_{b,u}^W \tau$ are *sufficiently small* (compared to $\Delta_{b,u}^B \epsilon$ and $\Delta_{b,u}^B \tau$), the estimation errors are *almost* free of within-block factors and the estimation error size of $\hat{\tau}_F$ *is more likely to* be smaller than those of $\hat{\tau}_B$ and $\hat{\tau}_U$. By contrast, if a matched-pair design does not succeed in covariates balance and is far from perfect, it would not be helpful anyway even without missing values ($\Delta\hat{\tau}_F$ in Proposition 1).

In addition to $\Delta_{b,u}^W \tau = 0$, if $\Delta_{b,u}^B \tau = 0$, the treatment effects are homogeneous in the sense that $Y_{b,u}(1) - Y_{b,u}(0) = \bar{\tau}$ for all (b, u). Corollary 2 is easily derived by substituting $\Delta_{b,u}^B \tau = 0$ in Corollary 1.

Corollary 2 *Estimation Error under Perfect Matched-Pair Design with Homogeneous Treat-*

ment Effects. When $\Delta^{W}_{b,u}\epsilon = \Delta^{W}_{b,u}\tau = \Delta^{B}_{b,u}\tau = 0$,

$$\begin{aligned}\Delta\hat{\tau}_U =& E(W^T_U \circ \Delta^B\epsilon) - E(W^C_U \circ \Delta^B\epsilon)\\ \Delta\hat{\tau}_B =& 0\\ \Delta\hat{\tau}_F =& 0.\end{aligned}$$

In this case, not only the full sample estimator but also the blockwise deletion estimator is free of estimation error. Moreover, even if the treatment effects are not homogeneous but *nearly* so, namely, if $\Delta^{B}_{b,u}\tau$ is *sufficiently small* (relative to $\Delta^{B}_{b,u}\epsilon$), the estimation errors are *almost* free of between-block heterogeneity of treatment effects, and it is *more likely* that $|\hat{\tau}_F| < |\hat{\tau}_U|$ and $|\hat{\tau}_B| < |\hat{\tau}_U|$.

Note that perfect matched-pair design is a matter of design, while homogeneous treatment effects is a matter of implementation. On the one hand, researchers can approach perfect matched-pair design by improving covariate balance between units within block. On the other hand, analysts may achieve homogeneous treatment effects by assigning treatment in as similar a way among units and blocks as possible. However, the latter has already been implied by SUTVA and the remaining between-block heterogeneity of treatment effects is a matter of nature and out of control of scholars. By trying to make the matched-pair design as perfect as possible, designers manage to sort heterogeneity of treatment effects into not within-block heterogeneity $\Delta^W\tau$ but between-block heterogeneity $\Delta^B\tau$. Therefore, unless total heterogeneity of treatment effects decreases, between-block heterogeneity is unavoidable even under a perfect matched-pair design. In this sense, between-block heterogeneity is more severe than covariate balance.

Bias

General Case. Now, this paper assumes ignorability of treatment assignment

$$(Y_{b,1}(0), Y_{b,1}(1), Y_{b,2}(0), Y_{b,2}(1)) \perp\!\!\!\perp (T_{b,1}, T_{b,2})$$

and equal probability of treatment assignment

$$\Pr(T_{b,1} = 1, T_{b,2} = 0) = \Pr(T_{b,1} = 0, T_{b,2} = 1) = \frac{1}{2}.$$

Denote bias of genetic estimator $\hat{\tau}$ by $\bar{\Delta}\hat{\tau} \equiv \mathbb{E}(\Delta\hat{\tau})$ where expectation operator $\mathbb{E}$ is taken over random allocation of T: $\mathbb{E}(f(T)) \equiv \sum_{T\in\{(0,1),(1,0)\}^{\bar{b}}} f(T)\Pr(T)$ for a genetic function $f(\cdot)$ of T.[6] We assume neither perfect matched-pair design nor homogeneous treatment effects.

Proposition 2 (Bias)

$$\begin{aligned}
\bar{\Delta}\hat{\tau}_U =& \mathbb{E}[E(W_U^T \circ (\Delta^B\epsilon + \Delta^W\epsilon + \Delta^B\tau + \Delta^W\tau)) - \\
& E(W_U^C \circ (\Delta^B\epsilon + \Delta^W\epsilon))] \\
\bar{\Delta}\hat{\tau}_B =& \mathbb{E}[E(W_B^T \circ (2\Delta^W\epsilon + \Delta^B\tau + \Delta^W\tau))] \\
\bar{\Delta}\hat{\tau}_F =& 0.
\end{aligned}$$

The bias of the full sample estimator is free of within-block factors and zero unlike its estimation error (Proposition 1). Thus, when we examine biases in special conditions below, we do not mention $\bar{\Delta}\hat{\tau}_F$ any more. Let $\Delta y \in \{\Delta^B\epsilon, \Delta^W\epsilon, \Delta^B\tau, \Delta^W\tau\}$, $W_U \in \{W_U^T, W_U^C\}$, and $W_B \in \{W_B^T, W_B^C\}$. Rough interpretation of $\mathbb{E}[E(W_U \circ \Delta y)]$ (or $\mathbb{E}[E(W_B \circ \Delta y)]$) is scaled covariance between $R_{b,u}$ (or $R_{b,u}R_{b,-u}$) and $\Delta y_{b,u}$. Or, when $\mathbb{E}[E(W \circ \Delta y)] = 0$, we may say that W and Δy are "uncorrelated."

Proposition 2 does not confirm that the blockwise deletion estimator "retain[s] the benefits of randomization ... regardless of the missing data mechanism" (Imai, King, and Nall, 2009*a*, 44); rather, the bias of the blockiwise deletion estimator ($\bar{\Delta}\hat{\tau}_B$) depends on relationship between the missing data mechanism $(R(0), R(1))$ and the potential outcome Δy, $\mathbb{E}[E(W_B \circ \Delta y)]$. In addition, according to King et al. (2007, 490), "if we lose a

[6] This paper does not assume that treatment assignment is independent between pairs: $(T_{b,1}, T_{b,2}) \perp\!\!\!\perp (T_{b',1}, T_{b',2})$ for $b \neq b'$.

cluster [i.e. unit] for a reason related to one or more of the variables we matched on [i.e. $\mathbb{E}[E(W \circ \Delta^W \epsilon)] = \mathbb{E}[E(W \circ \Delta^W \tau)] = 0$], ... then no bias would be induced for the remaining clusters [i.e. $\Delta\hat{\tau}_B = 0$]." Nonetheless, according to Proposition 2, if $\mathbb{E}[E(W_B^T \circ \Delta^B \tau)] \neq 0$, it can be that $\Delta\hat{\tau}_B \neq 0$. Thus, existence of $\Delta^B \tau$ poses difficulty to $\hat{\tau}_B$.

Special Cases. When $R_{b,u}(1) = R_{b,u}(0) = R_{b,u}$, we say the missing data mechanism achieves "within-unit balance of potential responses." Good examples are (single or double) blind test or subliminal stimuli. Since a subject is blind to treatment status (T), whether the subject responds or not (R) will not depend on whether treatment is assigned or not (T). In this situation, the bias of the blockiwise deletion estimator is free of within-block factors ($\mathbb{E}[E(W_B \circ \Delta^W)] = 0$), even if the missing data mechanism is a function of within-block factors in the sense that $\mathbb{E}[E(W_U \circ \Delta^W)] \neq 0$.[7]

Proposition 3 (Bias Given Within-Unit Balance of Potential Responses) *When* $R_{b,u}(1) = R_{b,u}(0)$,

$$\begin{aligned}\bar{\Delta}\hat{\tau}_U =& \mathbb{E}[E(W_U^T \circ (\Delta^B \epsilon + \Delta^W \epsilon + \Delta^B \tau + \Delta^W \tau)) - \\ & E(W_U^C \circ (\Delta^B \epsilon + \Delta^W \epsilon))] \\ \bar{\Delta}\hat{\tau}_B =& \mathbb{E}[E(W_B^T \circ \Delta^B \tau)].\end{aligned}$$

When $R_{b,u}(t) = R_{b,-u}(t)$, we say the missing data mechanism achieves "within-block balance of potential responses." For instance, if matched-on variables (e.g. DNA) completely explain the missing data mechanism, either two units or no unit in a pair (e.g. twin) should respond if treatment (or control) is assigned, because two units in a pair share the same (or similar) value of matched-on variables. Within-block factors (which varies across units in a block) cannot be responsible for the missing data mechanism (which is the same across units in a block). In this situation, not only the bias of the blockwise deletion estimator but also that of the unitwise deletion estimator are free of within-block factors ($\mathbb{E}[E(W \circ \Delta^W)] = 0$).

[7] In the terminolgy of Gerber and Green (2012, 224-228), in this situation, there are only always-reporters ($R_{b,u}(1) = R_{b,u}(0) = 1$) and never-reporters ($R_{b,u}(1) = R_{b,u}(0) = 0$). The unitwise deletion estimator provides an unbiased estimate of the local ATE for always-reporters.

Proposition 4 (Bias Given Within-Block Balance of Potential Responses) *When* $R_{b,u}(t) = R_{b,-u}(t)$,

$$\begin{aligned}
\bar{\Delta}\hat{\tau}_U =& \mathbb{E}[E(W_U^T \circ (\Delta^B \epsilon + \Delta^B \tau))] - \mathbb{E}[E(W_U^C \circ \Delta^B \epsilon)] \\
\bar{\Delta}\hat{\tau}_B =& \mathbb{E}[E(W_B^T \circ \Delta^B \tau)].
\end{aligned}$$

Dunning (2011, 15) denounces blocking, arguing that, for the blockwise deletion estimator to be unbiased, "we have to assume that all units with the same values of the blocked covariate respond similarly to treatment assignment," that is, within-block balance of potential responses. However, even in this condition, the blockwise deletion estimator is still biased, unless $\mathbb{E}[E(W_B^T \circ \Delta^B \tau)] = 0$, according to Proposition 4. Or, even if within-block balance of potential responses does not hold, but if within-unit balance of potential responses (Proposition 3) or perfect matched-pair design (Corollary 3, see below) is achieved, it follows anyway that $\bar{\Delta}\hat{\tau}_B = \mathbb{E}[E(W_B^T \circ \Delta^B \tau)]$.

Under the perfect matched-pair design, bias is simply expectation of estimation error and Δ^B is not averaged out. Corollary 3 is easily derived by substituting $\Delta_{b,u}^W \epsilon = \Delta_{b,u}^W \tau = 0$ in Proposition 2.

Corollary 3 (Bias under Perfect Matched-Pair Design) *When* $\Delta_{b,u}^W \epsilon = \Delta_{b,u}^W \tau = 0$,

$$\begin{aligned}
\bar{\Delta}\hat{\tau}_U =& \mathbb{E}[E(W_U^T \circ (\Delta^B \epsilon + \Delta^B \tau)] - \mathbb{E}[E(W_U^C \circ \Delta^B \epsilon)] \\
\bar{\Delta}\hat{\tau}_B =& \mathbb{E}[E(W_B^T \circ \Delta^B \tau)].
\end{aligned}$$

Note that Proposition 4 and Corollary 3 lead to the same equations under different conditions. Both are concerned with the design of how to match two units into pairs. If researchers succeed in balancing all pretreatment variables which decide potential outcomes ($Y(1)$ and $Y(0)$, Proposition 4) or potential responses ($R(1)$ and $R(0)$, Corollary 3) across units in each pair by design, $\mathbb{E}[E(W \circ \Delta^W)] = 0$ results. By contrast, Proposition 3 is concerned with not design but implementation of how treatment status (does not) affect(s) potential responses. If scholars can prevent subjects from knowing treatment status, Proposition 3 is applicable. In addition, unlike Proposition 4 where within-block balance of potential responses across units ($u = 1, 2$) in each block (b) in each treatment status (t) is achieved, Proposition 3 holds when within-unit balance of potential responses across treatment status ($t = 0, 1$) in each unit ((b, u)) is achieved.

Ignorability. Finally, this paper introduces ignorability of unitwise response,[8]

$$(Y_{b,u}(0), Y_{b,u}(1)) \perp\!\!\!\perp R_{b,u}$$

which is equivalent to, given ignorability of treatment assignment,

$$F(Y(0)|R(0) = 1) = F(Y(0)|R(0) = 0) = F(Y(0))$$
$$F(Y(1)|R(1) = 1) = F(Y(1)|R(1) = 0) = F(Y(1)),$$

where $F(\cdot)$ is the cumulative distribution function. In this condition, $\mathbb{E}[E(W_U \circ \Delta y)] = 0$. Similary, ignorability of blockwise response is defines as

$$(Y_{b,u}(0), Y_{b,u}(1)) \perp\!\!\!\perp R_{b,u}R_{b,-u}$$

which is equivalent to, given ignorability of treatment assignment,

$$F(Y_{b,u}(0)|R_{b,u}(0)R_{b,-u}(1) = 1) = F(Y_{b,u}(0)|R_{b,u}(0)R_{b,-u}(1) = 0) = F(Y_{b,u}(0))$$
$$F(Y_{b,u}(1)|R_{b,u}(1)R_{b,-u}(0) = 1) = F(Y_{b,u}(1)|R_{b,u}(1)R_{b,-u}(0) = 0) = F(Y_{b,u}(1)),$$

where $F(\cdot)$ is the cumulative distribution function. In this condition, $\mathbb{E}[E(W_B \circ \Delta y)] = 0$.

Proposition 5 *Unbiasedness Given Ignorability of Response.*

[8] Gerber and Green (2012, 229) call this "missing independent of potential outcomes" (MIPO).

(1) When $(Y_{b,u}(0), Y_{b,u}(1)) \perp\!\!\!\perp R_{b,u}$,

$$\bar{\Delta}\hat{\tau}_U = 0.$$

(2) When $(Y_{b,u}(0), Y_{b,u}(1)) \perp\!\!\!\perp R_{b,u}R_{b,-u}$,

$$\bar{\Delta}\hat{\tau}_B = 0.$$

Ignorability of unitwise (or blockwise) response is a sufficient, but not a necessary, condition for unbiasedness of the unitwise (or blockwise) deletion estimator.[9]

Unfortunately, however, it is unusual that ignorability of response holds. For example, treated subjects might not feel like showing up in the follow up survey if their outcomes are not so good in spite of treatment ($R_{b,u}(1)$ is more likely to be 0 as $Y(1)$ becomes smaller, and $\mathbb{E}[E(W^T \circ \Delta y)]$ is more likely to be larger), while controlled subjects might think they don't have to continue experiments if their outcomes are good ($R_{b,u}(0)$ is more likely to be 0 as $Y(0)$ becomes larger, and $\mathbb{E}[E(W^C \circ \Delta\epsilon)]$ is more likely to be smaller). Thus, estimates of the ATE are expcted be larger than the true ATE ($\bar{\Delta}\hat{\tau}_U > 0$ and $\bar{\Delta}\hat{\tau}_B > 0$).

In Appendix, other special cases where $\mathbb{E}[E(W \circ \Delta y)] = 0$ for some W and Δy are introduced.

APPLICATION

The most important message of the previous section is that $\Delta\hat{\tau}_B$ is free from $\Delta^B\epsilon$, though $\Delta\hat{\tau}_U$ is not. The previous section does not characterize properties of $\Delta\hat{\tau}_{MI}$ analytically because they depend on behavior of matched-on variables. This section applies these three estimators to real experimental and non-experimental datasets, and compares their performance.[10]

[9] Frangakis and Rubin (1999) considers "latent ignorability" where response is ignorable given compliance status and monotonicity, and derives an unbiased estimator.

[10] The previous section does not examine efficiency of estimators. It is difficult to derive their variance formally because the denominator in $\frac{\sum_{b=1}^{\bar{b}}\sum_{u=1}^{2} W_{b,u}Y_{b,u}}{\sum_{b=1}^{\bar{b}}\sum_{u=1}^{2} W_{b,u}}$ varies across trials. Instead, this section considers variance of estimators by simulation.

Though these datasets have no missing value, this paper makes missing values by simulation, where potential responses ($R(1)$ and $R(0)$) are functions of $\Delta^B\epsilon$ only ($\mathbb{E}[E(W_B \circ \Delta^B\epsilon)] \neq 0$ and $\mathbb{E}[E(W_U \circ \Delta^B\epsilon)] \neq 0$), neither $\Delta^W\epsilon$ nor $\Delta^B\tau$ nor $\Delta^W\tau$ ($\mathbb{E}[E(W_B \circ (\Delta^W\epsilon + \Delta^B\tau + \Delta^W\tau))] = \mathbb{E}[E(W_U \circ (\Delta^W\epsilon + \Delta^B\tau + \Delta^W\tau))] = 0$, because the matched-pair design is perfect and treatment effects are homogeneous). It is not formally clear how $\bar{\Delta}\hat{\tau}_{MI}$ behaves, though missing-at-random assumption (or ignorability of response or MIPO) is violated.[11]

Experiment

Data. This subsection considers one of the largest randomized field experiments, evaluation of the universal health insurance program, *Segro Popular de Salud* (hereafter, SPS), in Mexico (Imai, 2008; Imai, King, and Nall, 2009*a*; King et al., 2007, 2009).

A pair of similar areas, which are called "health clusters," were matched. Treatment (T) was randomly assigned to either of two clusters in each pair. In a treated cluster, people were encouraged to sign up SPS, while they were not in a controlled cluster. A resident of each sampled household in each cluster was interviewed in the baseline survey (at the time of treatment assignment in August 2005) and the follow-up survey (10 months after the treatment assignment). The unit of observation is household (the dataset contains 32,515 respondents), though the unit of randomization is cluster (50 pairs of clusters). Thus, this paper analyzes cluster-level data, as Imai (2008) does.[12]

Among hundreds of outcomes, previous studies have examined most intensively "catastrophic health expenditure," i.e. out-of-pocket health-care expenditures greater than 30 percent of disposal income. Following Imai (2008, 4868), the outcome this paper studies (Y) is the proportion of households suffering catastrophic health expenditure in every cluster.

Indeed, the investigators utilized the matched-pair design exactly because they prepared

[11] In the previous section, $R(0)$ and $R(1)$ are fixed, while T is random. In this section, $R(0)$ and $R(1)$ are random, while T is fixed ($Y(0)$ and $Y(1)$ are fixed in both sections). Therefore, admittedly, both sections are not strictly comparable. However, Monte Carlo simulation of this section will give us practiacl sense of how each estimator behaves in real setups.

[12] The data this paper uses is Imai, King, and Nall (2009*b*). This paper aggregates the original individual-level data into cluster-level data (Imai, 2008, 4868).

for missing values. Before they observed outcomes, King et al. (2007, 489, emphasis original) wrote,

> [T]he PROGRESA evaluation ... had some loss of observations ... given this previous experience, we must expect to lose health clusters, and so we need a design that allows some clusters to be lost, under at least some circumstances, without also losing the advantages of randomization. Thus, we turn to what is known as a *randomized cluster matched pair design.*

Eventually, no cluster has missing values. Thus, this paper turns some actually observed outcome values into missing ones by way of the following procedure.

Method. According to the framework of this paper, below, clusters and their pairs are referred to as "units" and "blocks," respectively ($\bar{b} = 50$). First, the fixed effects of block b, $\Delta_b^B \equiv \alpha + \Delta_b^B \epsilon + \Delta_b^B \tau$, is estimated by $\hat{\Delta}_b^B = (y_{b,1} + y_{b,2} - \hat{\tau}_F)/2$ (note that the full sample estimator is applicable because no cluster is missing in fact).[13] Then, the weight for randomly assigning missing to unit (b, u) is calculated as $\omega_{b,u} \equiv \Phi(\rho[t_{b,u}\hat{\Delta}_b^B - (1 - t_{b,u})\hat{\Delta}_b^B])$, where $\Phi(\cdot)$ is the cumulative density function of the standard normal distribution. In the case of $\rho > 0$, the larger $\hat{\Delta}_b^B$, the more (or less) likely $R_{b,u}(1)$ (or $R_{b,u}(0)$) is to be 0. In the case of $\rho < 0$, the opposite holds. Thus, ρ represents direction and strength of the relationship between within-unit-unbalanced potential responses and between-block factor.[14] Denote the proportion of missing units by m. The values of ρ and m are specified shortly.

Fix the values of ρ and m. For one trial, $m\bar{b}$ units are chosen randomly from treated and controlled groups, respectively, following the above weight $\omega_{b,u}$, and their values of $Y_{b,u}$ are turned into missing. Then, $\hat{\tau}_B$ and $\hat{\tau}_U$ are estimated. When $\hat{\tau}_{MI}$ is estimated, missing values are imputed by Amelia II using 17 pretreatment variables and making 5 imputed datasets.[15]

[13] The next version of the paper will use $\Delta_b^B \equiv \Delta_b^B \epsilon + \Delta_b^B \tau$, which is estimated by $\hat{\Delta}_b^B \equiv ((y_{b,1} + y_{b,2} - \hat{\tau}_F)/2) - \hat{\alpha}$, where $\hat{\alpha} \equiv \bar{b}^{-1} \sum_{b=1}^{\bar{b}} ((y_{b,1} + y_{b,2} - \hat{\tau}_F)/2)$.

[14] If we make within-unit-*balanced* potential response by, say, $\omega_{b,u} \equiv \Phi(\rho\hat{\Delta}_b^B)$, both $\hat{\tau}_U$ and $\hat{\tau}_B$ are unbiased according to Proposition 3 and $\mathbb{E}[E(W_U \circ (\Delta^W \epsilon + \Delta^W \tau))] = 0$ (though $\mathbb{E}[E(W \circ (\Delta^B \epsilon + \Delta^B \tau))] \neq 0$. Why this is the case is a future agenda). Since this is not interesting, this paper does not take this strategy.

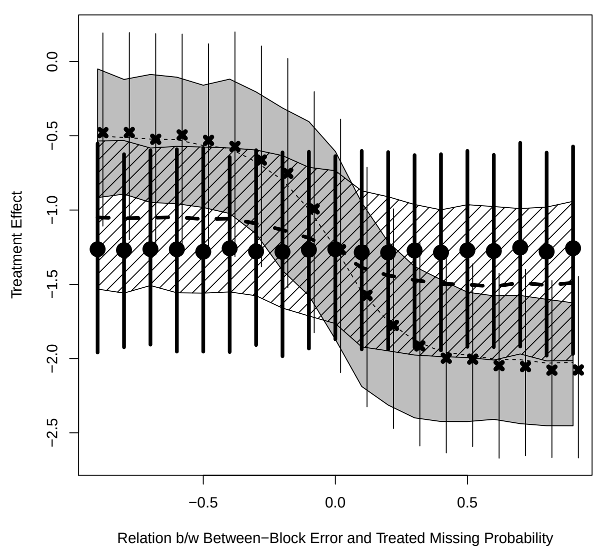
[15] For details about 17 matched-on pretreatment variables, see Appendix.

In addition, in order to examine what if blocking is not successful, the failing blockwise deletion estimator, $\hat{\tau}_{NB}$, is also applied. Namely, this paper randomly makes pairs of units irrespetive of the original block, regard the new pairs as blocks and applies the blockwise deletion estimator.

This trial is repeated 1,000 times for one set of values of ρ and m. For each estimator, the average and the 2.5-97.5 percentile range of 1,000 estimates are recorded.

Results. Figure 1 shows the case where ρ varies from -0.9 to 0.9 by 0.1 and m is constantly equal to 0.1.[16] The horizontal axis represents ρ, while the vertical axis indicates $\hat{\tau}$. The baseline is the estimate by the full sample estimator, $\hat{\tau}_F = -1.27$ (Imai, 2008, 4868), because no unit is missing actually and $\hat{\tau}_F$ is available.[17]

Figure 1: Estimates by Four Estimators of ATE ($\hat{\tau}$, vertical axis) against Relationship (ρ, horizontal axis) between $\hat{\Delta}_b^B$ and $R_{b,u}(t)$. Missing proportion is $m = 0.1$. Experimental Data.



The vertical thick lines and big points correspond to the 2.5-97.5 percentile ranges and

[16]This study uses R Development Core Team (2012) for conducting simulation, analyzing data, and drawing figures.

[17]Imai (2008, 4868) reports $\hat{\tau}_F = 1.27$, which seems to be simply a typo. The true ATE, $\bar{\tau}$, is obviously unknown.

averages of 1,000 estimates by the blockwise deletion estimator, $\hat{\tau}_B$, for each value of ρ. It is easy to see that the points are almost equal to $\hat{\tau}_F$, because the missing data mechanism is a function of Δ^B but not Δ^W, and $\Delta\hat{\tau}_B$ is free of $\Delta^B\epsilon$. [18]

The gray area and the thin dotted line represent the 2.5-97.5 percentile ranges and averages of estimates by the unitwise deletion estimator, $\hat{\tau}_U$. The larger the size of $|\rho|$, the more $\hat{\tau}_U$ deviates from $\hat{\tau}_F$. The size and direction of the deviation depends on $\mathbb{E}[E(W_U^T \circ \Delta^B\epsilon)] = -\mathbb{E}[E(W_U^C \circ \Delta^B\epsilon)]$, which is managed by ρ. Furthermore, the gray area does not contain $\hat{\tau}_F$ for relatively larger values of ρ; thus, the null hypothesis of $\bar{\tau} = \hat{\tau}_F$ would be rejected.

The shaded area and the thick dotted line show the 2.5-97.5 percentile ranges and averages of estimates by the multiple imputation estimator, $\hat{\tau}_{MI}$. The thick dotted line deviates from $\hat{\tau}_F$ for most of values of ρ, though to lesser degree than $\hat{\tau}_U$. Blockwise deletion makes most of blocking, while multiple imputation usually cannot use it because the number of blocks is too large (half of the number of units). When the matched-pair design is (almost) perfect, pairs can explain more variation in dependent variables than covariates multiple imputation utilizes can do and, therefore, the blockwise deletion estimator outperforms the multiple imputation estimator. Recall also that assumption of missing at random is violated. Fortunately, the shaded area always includes $\hat{\tau}_F$.

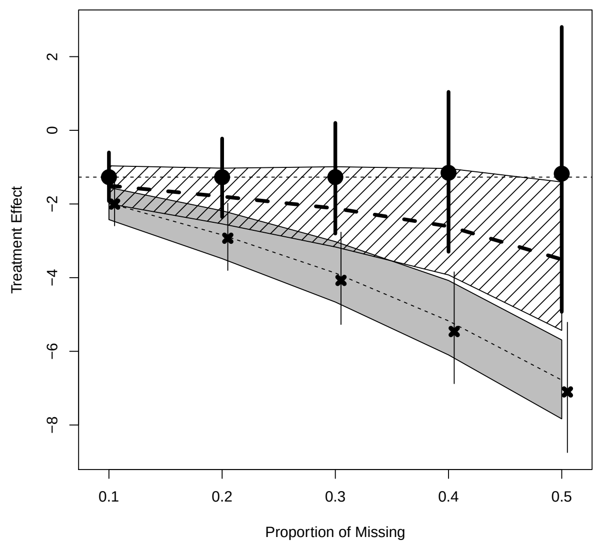
Finally, the vertical thin lines and small points correspond to the 2.5-97.5 percentile ranges and averages of estimates by the failing blockwise deletion estimator, $\hat{\tau}_{NB}$. Clearly, it deviates from $\hat{\tau}_F$ almost as much as $\hat{\tau}_U$. Therefore, when the matched-pair design is far from perfect, the blockwise deletion estimator does not work well.

Two general remarks are in order. First, when attrition is not a function of potential outcomes ($\rho = 0$), the averages of all four estimators convege to $\hat{\tau}_F$. Second, a downside of the blockwise deletion estimator is its larger 2.5-97.5 percentile range relative to those of other estimators, naturally because $\hat{\tau}_B$ utilizes smaller number of units than $\hat{\tau}_U$ and $\hat{\tau}_{MI}$.

[18] $\hat{\Delta}^B$ includes $\Delta^B\epsilon$ and $\Delta^B\tau$. One cannot, however, identify $\Delta^B\epsilon$ and $\Delta^B\tau$ separately.

Figure 2 shows the case where the missing proportion m varies from 0.1 to 0.5 by 0.1 and ρ is constantly equal to 0.5. The horizontal axis represents m, while the vertical axis indicates $\hat{\tau}$. The horizontal thin dotted line indicates $\hat{\tau}_F$. The estimates of all the four estimators are displayed in the same way as in Figure 1. Implication of Figure 1 becomes more severe, as more units are missing (m becomes larger).

Figure 2: Estimates by Four Estimators ($\hat{\tau}$, vertical axis) against Missing Proportion (m, horizontal axis). $\rho = 0.5$. Experimental Data.



Observatioal Study

Data. This subsection switches to one of the most famous non-experimental data, which Dehejia and Wahba (1999) and Lalonde (1986) analyze. Their goal is to estimate the ATE of a labor training program (T) in the National Supported Work (NSW) Demonstration on post-intervention income levels in 1978 (Y). To that end, Lalonde (1986) constructs a dataset by combining the experimental treated units from the NSW dataset with non-experimental controlled units from the Current Population Survey (CPS). Since the NSW and CPS units are not comparable, Dehejia and Wahba (1999) apply propensity score matching to the subset of the Lalonde's (1986) original dataset (only those whose earnings in 1974 are available) so that observed preintervention variables are balanced between the treated and controlled groups.

This paper applies Sekhon's (2011) multivariate and propensity score matching algorithm, `Matching` package in a statistical environment `R`, to Dehejia and Wahba's (1999) dataset.[19] Among 185 treated units and 15,992 controlled units, $\bar{b} = 223$ matched pairs are coupled. In the framework of this paper, a pair corresponds to a block.

Method. The simulation procedure is basically the same as that of the previous Mexican application. Exceptions are as follows. The fixed effects of block b, $\Delta_b^B \equiv \Delta_b^B \epsilon + \Delta_b^B \tau + \bar{\tau}$, is estimated by $\hat{\Delta}_b^B = (y_{b,1} + y_{b,2} - \hat{\alpha})/2$, where $\hat{\alpha} \equiv \bar{b}^{-1} \sum_{b=1}^{\bar{b}} ((y_{b,1} + y_{b,2} - \hat{\tau}_F)/2)$. [20] When $\hat{\tau}_{MI}$ is estimated, missing values are imputed by Amelia II using the same nine pretreatment variables as `Matching` package uses.[21]

Results. Figure 3 shows the case where ρ ranges between -0.0009 to 0.0009 by 0.0001 and m is constantly equal to 0.1. The horizontal axis represents ρ, while the vertical axis indicates $\hat{\tau}$. The estimates of all the four estimators are displayed in the same way as in Figure 1. If no unit is missing, the estimate by the full sample estimator, $\hat{\tau}_F = 1913.2$, is available, which is indicated by the horizontal thin dotted line.[22] Implication of the results is the same as that of the Mexican application. Noticeable difference is that $\hat{\tau}_{MI}$ is worse than that in the previous subsection and is now as bad as $\hat{\tau}_U$. It calls attention given that matching algorithm (`Matching`, which affects performance of $\hat{\tau}_B$) and multiple imputation algorithm (Amelia II, which affects performance of $\hat{\tau}_{MI}$) take advantage of the same nine pretreatment variables.

Figure 4 shows the case where the missing proportion m varies from 0.1 to 0.5 by 0.1 and ρ is constantly equal to 0.0005. The horizontal axis represents m, while the vertical axis indicates $\hat{\tau}$. The estimates of all the four estimators are displayed in the same way

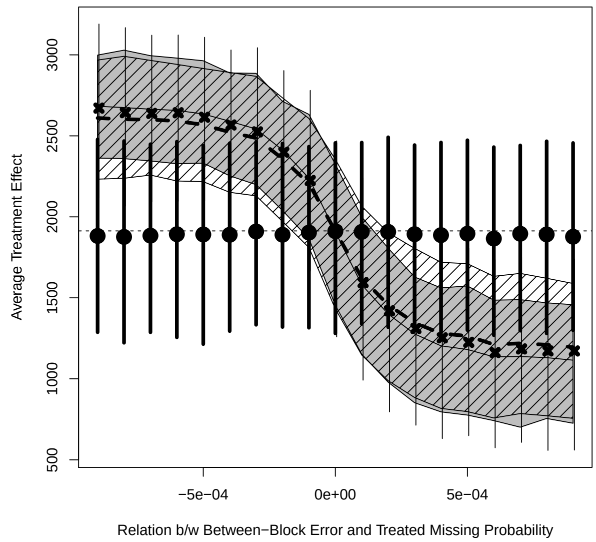
[19]For details about data and matching procedure, see Appendix.

[20]The next version of the paper will use $\Delta_b^B \equiv \Delta_b^B \epsilon + \Delta_b^B \tau$, which is estimated by $\hat{\Delta}_b^B = ((y_{b,1} + y_{b,2} - \hat{\tau}_F)/2) - \hat{\alpha}$.

[21]As for the nine matched-on variables, see Appendix.

[22]The true ATE, $\bar{\tau}$, is obviously unknown.

Figure 3: Estimates by Four Estimators of ATE ($\hat{\tau}$, vertical axis) against Relationship (ρ, horizontal axis) between $\hat{\Delta}_b^B$ and $R_{b,u}(t)$. Missing proportion is $m = 0.1$. Observational Study Data.

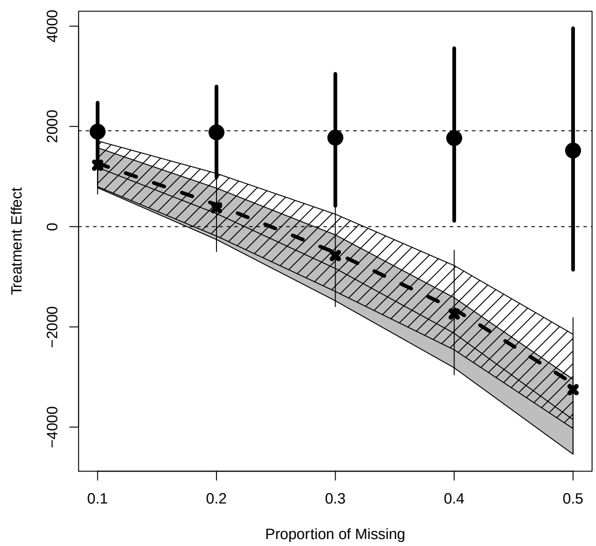


as in Figure 2. The upper and lower horizontal thin dotted lines indicate $\hat{\tau}_F$ and $\hat{\tau} = 0$, respectively. Again, implication of the results is the same as that of the previous Mexican application. Implication of Figure 3 becomes more severe, as more units are missing (m becomes larger). In particular, when $m \geq 0.4$, the 2.5-97.5 percentile ranges of $\hat{\tau}_U$, $\hat{\tau}_{MI}$, and $\hat{\tau}_{NB}$ are below zero; thus, the alternative hypothesis $\hat{\tau} < 0$ would be supported by these estimators, even if $\hat{\tau}_F > 0$.

CONCLUSION

When the missing data mechanism is not independent of potential outcomes in experiments with the matched-pair design or observational study with matching, all of the blockwise deletion, unitwise deletion, single imputation and multiple imputation estimators can be biased. Balance of pretreatment variables does not guarantee unbiased estimation of the ATE in the presence of attrition. In practice, however, analysts should choose at least one

Figure 4: Estimates by Four Estimators ($\hat{\tau}$, vertical axis) against Missing Proportion (m, horizontal axis). $\rho = 0.0005$. Observational Study Data.



estimator. A practical advice of this paper is simple: use the blockwise deletion estimator. The reasons are summarized as follows.

1. The blockwise deletion estimator is equivalent to the single imputation estimator (Theorem 1).

2. The blockwise deletion estimator cancels out between-block error and, thus, *reduces* a source of estimation error (Proposition 1), while the unitwise deletion estimator does not.

3. The blockwise deletion estimator is more likely to have smaller estimation error than the unitwise deletion estimator, as a matched-pair design succeeds in achieving balance of more pretreatment variables which decide potential outcome, and treatment effects are less heteregeneous (Corollary 2).

4. The blockwise deletion estimator is more likely to have smaller *bias* than the unitwise deletion estimator, as a matched-pair design succeeds in achieving balance of more

pretreatment variables which decide potential *response*, and treatment effects are less heteregeneous (Proposition 4).

5. The blockwise deletion estimator is more likely to have smaller bias than the unitwise deletion estimator, as a researcher can implement treatment assignment with subjects blind to it more successfully, and treatment effects are less heteregeneous (Proposition 3).

6. The blockwise deletion estimator takes advantage of information about which unit is paired with which unit, while the multiple imputation estimator cannot. It is difficult to find such covariates that are as informative as blocks and, thus, enable as good imputation as the matched-pair design. (as shown in Application)

At the same time, researchers should not forget a caveat that even a perfect matched-pair design *cannot remove* attrition bias due to between-block heterogeneity of treatment effects, which is almost unavoidable (Corollary 3). The main contribution of this paper is to clarify the findings above by formalizing the relationship between components of potential outcomes and potential reponses.

The relationship between blocking and attrition bias has attracted less attention than it should and is complicated enough to need thorough consideration. This work aims to contribute to this issue, though there remain many research agendas. For instance, this study focuses on blocks composed of a pair of units, while, in many studies, blocks are composed of more than two units. In this case, inverse probability weighting (Gerber and Green, 2012, 222) is available, where weight variable W is not a dummy but a real number which is not smaller than one. For another example, units sometimes do not comply with treatment assignment, even if their outcomes are observed. A typical solution is instrumental variable estimation (Angrist, Imbens, and Rubin, 1996). It is not clear, however, whether or in what condition the blockwise deletion estimator is better than inverse probability weighting or instrumental variable estimation. Studying these topics is future agenda.

References

Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin. 1996. “Indentification of Causal Effects Using Instrumental Variables.” *Journal of the American Statistical Association* 91(434): 444–455.

Box, George E. P., William G. Hunter, and J. Stuart Hunter. 1978. *Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building*. New York: Wiley.

Dehejia, Rajeev, and Sadek Wahba. 1999. “Causal Effects in Non-Experimental Studies: Reevaluating the Evaluation of Training Programs.” *Journal of the American Statistical Association* 94(448): 1053–1062.

Donner, Allan, and Neil Klar. 2000. *Design and Analysis of Cluster Randomization Trials in Health Research*. London: Arnold.

Dunning, Thad. 2011. “Does Blocking Reduce Attrition Bias?” *Experimental Political Scientist: Newsletter of the APSA Experimental Section* 2(1): 12–16.

Enos, Ryan D. 2014. “Causal effect of intergroup contact on exclusionary attitudes.” *Proceedings of the National Academy of Sciences* 111(10): 3699–3704.

Family Heart Study Group. 1994. “Randomised Controlled Trial Evaluating Cardiovascular Screening and Intervention in General Practice: Principal Results of British Family Heart Study.” *British Medical Journal* 308(6924): 313–20.

Frangakis, Constantine E., and Donald B. Rubin. 1999. “Addressing Complications of Intention-to-Treat Analysis in the Combined Presence of All-or-None Treatment-Noncompliance and Subsequent Missing Outcomes.” *Biometrika* 86(2): 365–379.

Fukumoto, Kentaro, and Mikitaka Masuyama. 2015. “Measuring Judicial Independence

Reconsidered: Survival Analysis, Matching, and Average Treatment Effects." *Japanese Journal of Political Science* 16(1): 33–51.

Gerber, Alan S., and Donald P. Green. 2012. *Field Experiments: Design, Analysis, and Interpretation.* New York: W. W. Norton & Company.

Giesbrecht, Francis G., and Marcia L. Gumpertz. 2004. *Planning, Construction, and Statistical Analysis of Comparative Experiments.* New York: Wiley.

Hinkelmann, Klaus, and Oscar Kempthorne. 1994. *Design and Analysis of Experiments. Volume I: Introduction to Experimental Design.* New York: John Wiley and Sons, Inc.

Ho, Daniel, Kosuke Imai, Gary King, and Elizabeth Stuart. 2007. "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference." *Political Analysis* 15: 199–236.

Honaker, James, Gary King, and Matthew Blackwell. 2012. "AMELIA II: A Program for Missing Data.".

Imai, Kosuke. 2008. "Variance Identification and Efficiency Analysis in Randomized Experiments under the Matched-Pair Design." *Statistics in Medicine* 27(24): 4857–4873.

Imai, Kosuke, Gary King, and Clayton Nall. 2009*a*. "The Essential Role of Pair Matching in Cluster-Randomized Experiments, with Application to the Mexican Universal Health Insurance Evaluation (with discussions and rejoinder)." *Statistical Science* 24(1): 29–53.

Imai, Kosuke, Gary King, and Clayton Nall. 2009*b*. "Replication data for: The Essential Role of Pair Matching in Cluster-Randomized Experiments, with Application to the Mexican Universal Health Insurance Evaluation.".

King, Gary, Emmanuela Gakidou, Kosuke Imai, Jason Lakin, Ryan T. Moore, Clayton Nall, Nirmala Ravishankar, Manett Vargas, Martha Mara Tllez-Rojo, Juan Eugenio Hernndez vila, Mauricio Hernndez vila, and Hctor Hernndez Llamas. 2009. "Public Policy for the

Poor? A Randomised Assessment of the Mexican Universal Health Insurance Programme." *Lancet* 373: 1447–1454.

King, Gary, Emmanuela Gakidou, Nirmala Ravishankar, Ryan T. Moore, Jason Lakin, Manett Vargas, Martha Mara Tllez-Rojo, Juan Eugenio Hernndez vila, Mauricio Hernndez vila, and Hctor Hernndez Llamas. 2007. "A "Politically Robust" Experimental Design for Public Policy Evaluation, with Application to the Mexican Universal Health Insurance Program." *Journal of Policy Analysis and Management* 26: 479–506.

Lalonde, Robert. 1986. "Evaluating the Econometric Evaluations of Training Programs." *American Economic Review* 76: 604–620.

Moore, Ryan T. 2010. "Blocking Political Science Experiments: Why, How, and Then What?" *Experimental Political Scientist: Newsletter of the APSA Experimental Section* 1(1): 3–5.

Moore, Ryan T. 2012. "Multivariate Continuous Blocking to Improve Political Science Experiments." *Political Analysis* 20(4): 460–479.

Morton, Rebecca B., and Kenneth C. Williams. 2010. *Experimental Political Science and the Study of Causality: From Nature to the Lab.* Cambridge: Cambridge University Press.

Paluck, Elizabeth Levy, and Donald P. Green. 2009. "Deference, Dissent, and Dispute Resolution: An Experimental Intervention Using Mass Media to Change Norms and Behavior in Rwanda." *American Political Science Review* 103(4): 622–644.

Panagopoulos, Costas, and Donald P Green. 2008. "Field experiments testing the impact of radio advertisements on electoral competition." *American Journal of Political Science* 52(1): 156–168.

R Development Core Team. 2012. "R: A Language and Environment for Statistical Computing.".

Sekhon, Jasjeet S. 2011. "Multivariate and Propensity Score Matching Software with Automated Balance Optimization: The Matching package for R." *Journal of Statistical Software* 42(7): 1–52.

Yates, Frank. 1933=1970. *The Analysis of Replicated Experiments When the Field Results Are Incomplete*. London: Griffin pp. 41–56.

APPENDIX

Single Imputation

Definition. The single imputation method explained by Hinkelmann and Kempthorne (1994, 67-72, 265-271) is as follows. Let $(b, T) = (b, 1)$ and $(b, C) = (b, 2)$ if $T_{b,1} = 1$ and $(b, T) = (b, 2)$ and $(b, C) = (b, 1)$ if $T_{b,2} = 1$. Note that $R_{b,T} = R_{b,1}T_{b,1} + R_{b,2}T_{b,2}$ and $R_{b,C} = R_{b,1}(1 - T_{b,1}) + R_{b,2}(1 - T_{b,2})$. Suppose that two units of the same block never become missing at the same time and rearrange block number b so that

$$\{b|R_{b,T} = 1, R_{b,C} = 1\} \equiv \mathcal{B}^L = \{1, \ldots, L\}$$
$$\{b|R_{b,T} = 1, R_{b,C} = 0\} \equiv \mathcal{B}^T = \{L + 1, \ldots, L + M^T\}$$
$$\{b|R_{b,T} = 0, R_{b,C} = 1\} \equiv \mathcal{B}^C = \{L + M^T + 1, \ldots, L + M^T + M^C\}$$
$$\{b|R_{b,T}(1) = 0, R_{b,C}(0) = 0\} = \emptyset$$
$$\therefore L + M^T + M^C = \bar{b} = N/2.$$

Let Y^* such that

$$\begin{aligned} Y^*_{b,u} &= Y_{b,u} \quad \text{if} \quad R_{b,u} = 1 \\ Y^*_{b,u} &= 0 \qquad \text{if} \quad R_{b,u} = 0. \end{aligned}$$

Make the following model:

$$\begin{aligned} Y^* &= X\beta + ZY_{R=0} + \epsilon \\ \beta &= (\beta_1, \beta_2, \ldots, \beta_{\bar{b}}, \bar{\tau}) \\ \beta_b &= \alpha + \Delta^B_b \epsilon \\ \epsilon_{b,u} &= \Delta^W_{b,u}\epsilon + (\Delta^B_b \tau + \Delta^W_{b,u}\tau)T_{b,u} \\ Y_{R=0} &= (Y_{L+1,C}, Y_{L+2,C}, \ldots, Y_{L+M^T,C}, Y_{L+M^T+1,T}, Y_{L+M^T+2,T}, \ldots, Y_{L+M^T+M^C,T}) \end{aligned}$$

where X is the design matrix

$$X = \begin{array}{c} \\ (b \in \mathcal{B}^L, T) \\ (b \in \mathcal{B}^L, C) \\ (b \in \mathcal{B}^T, T) \\ (b \in \mathcal{B}^C, C) \\ (b \in \mathcal{B}^T, C) \\ (b \in \mathcal{B}^C, T) \end{array} \begin{array}{c} \begin{array}{cccc} I(b=1) & \cdots & I(b=\bar{b}) & T_{b,u} \end{array} \\ \begin{pmatrix} I_L & 0_{L\times M^T} & 0_{L\times M^C} & 1_{L\times 1} \\ I_L & 0_{L\times M^T} & 0_{L\times M^C} & 0_{L\times 1} \\ 0_{M^T\times L} & I_{M^T} & 0_{M^T\times M^C} & 1_{M^T\times 1} \\ 0_{M^C\times L} & 0_{M^C\times M^T} & I_{M^C} & 0_{M^C\times 1} \\ 0_{M^T\times L} & I_{M^T} & 0_{M^T\times M^C} & 0_{M^T\times 1} \\ 0_{M^C\times L} & 0_{M^C\times M^T} & I_{M^C} & 1_{M^C\times 1} \end{pmatrix} \end{array},$$

and Z is the missing indicator matrix

$$Z = \begin{array}{c} \\ (b \in \mathcal{B}^L, T) \\ (b \in \mathcal{B}^L, C) \\ (b \in \mathcal{B}^T, T) \\ (b \in \mathcal{B}^C, C) \\ (b \in \mathcal{B}^T, C) \\ (b \in \mathcal{B}^C, T) \end{array} \begin{array}{c} \begin{array}{cc} I(b=L+1) & I(b=\bar{b}) \end{array} \\ \begin{pmatrix} 0_{L\times M^T} & 0_{L\times M^C} \\ 0_{L\times M^T} & 0_{L\times M^C} \\ 0_{M^T\times M^T} & 0_{M^T\times M^C} \\ 0_{M^C\times M^T} & 0_{M^C\times M^C} \\ -I_{M^T} & 0_{M^T\times M^C} \\ 0_{M^C\times M^T} & -I_{M^C} \end{pmatrix} \end{array},$$

where I_K is the $K \times K$ identity matrix, $0_{K_1\times K_2}$ is the $K_1 \times K_2$ matrix whose elements are all zero, and $1_{K_1\times K_2}$ is the $K_1 \times K_2$ matrix whose elements are all one. Let matrices U, V and W such that

$$(I_N - X(X'X)^{-1}X') = \begin{pmatrix} U_{(2L+M^T+M^C)\times(2L+M^T+M^C)} & V_{(2L+M^T+M^C)\times(M^T+M^C)} \\ V'_{(M^T+M^C)\times(2L+M^T+M^C)} & W_{(M^T+M^C)\times(M^T+M^C)} \end{pmatrix}.$$

By OLS (namely, minimizing $\sum \epsilon^2$), it follows

$$\hat{Y}_{R=0} = -W^{-1}V'Y_{R=1},$$

where $Y_{R=1} = (Y_{1,T}, \ldots, Y_{2L+M^T+M^C,C})$.[23] Denote the imputed outcome by

$$\begin{aligned}\hat{Y}^{(S)} &\equiv (Y_{R=1}, \hat{Y}_{R=0}) \\ &= Y^* - Z\hat{Y}_{R=0} \\ &= X\beta + \epsilon.\end{aligned}$$

Then, β, including $\bar{\tau}$, is estimated by regressing $\hat{Y}^{(S)}$ on X;

$$\hat{\beta} = (X'X)^{-1}X'\hat{Y}^{(S)}.$$

Or, in order to estimate $\bar{\tau}$ alone, we only have to apply the full sample estimator $\hat{\tau}_F$ to the imputed outcome $\hat{Y}^{(S)}$;

$$\hat{\tau}_{SI} = E(W_F^T \circ \hat{Y}^{(S)}) - E(W_F^C \circ \hat{Y}^{(S)}).$$

[23] If two units of the same block become missing at the same time ($\{b | R_{b,T}(1) = 0, R_{b,C}(0) = 0\} \neq \emptyset$), W is not invertible.

Note that, by OLS, one implicitly assumes that, for missing units,

$$\begin{aligned}\hat{\epsilon}_{b,u} &= \Delta^W_{b,u}\epsilon + (\Delta^B_b\tau + \Delta^W_{b,u}\tau)T_{b,u} \\ &= 0.\end{aligned}$$

This assumption is true if (but not only if)

$$\Delta^W_{b,u}\epsilon = \Delta^B_b\tau = \Delta^W_{b,u}\tau = 0,$$

when the blockwise estimator is unbiased.

Proof of Theorem 1. The variance covariance matrix of the design matrix is

$$
X'X = \begin{array}{c} \\ b_1 \\ b_2 \\ \vdots \\ b_{\bar{b}} \\ T_{b,u} \end{array}
\begin{array}{c}
\begin{array}{ccccc} b_1 & b_2 & \cdots & b_{\bar{b}} & T_{b,u} \end{array} \\
\begin{pmatrix}
2 & 0 & \cdots & 0 & 1 \\
0 & 2 & \cdots & 0 & 1 \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & \cdots & 2 & 1 \\
1 & 1 & \cdots & 1 & \bar{b}
\end{pmatrix}
\end{array}.
$$

Its inverse is

$$
(X'X)^{-1} = \begin{array}{c} \\ b_1 \\ b_2 \\ \vdots \\ b_{\bar{b}} \\ T_{b,u} \end{array}
\begin{array}{c}
\begin{array}{ccccc} b_1 & b_2 & \cdots & b_{\bar{b}} & T_{b,u} \end{array} \\
\begin{pmatrix}
2^{-1}+N^{-1} & N^{-1} & \cdots & N^{-1} & -2N^{-1} \\
N^{-1} & 2^{-1}+N^{-1} & \cdots & N^{-1} & -2N^{-1} \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
N^{-1} & N^{-1} & \cdots & 2^{-1}+N^{-1} & -2N^{-1} \\
-2N^{-1} & -2N^{-1} & \cdots & -2N^{-1} & 4N^{-1}
\end{pmatrix}
\end{array}.
$$

The projection matrix is

$$
X(X'X)^{-1}X' = \begin{array}{c} \\ (\mathcal{B}^L,T) \\ (\mathcal{B}^L,C) \\ (\mathcal{B}^T,T) \\ (\mathcal{B}^C,C) \\ (\mathcal{B}^T,C) \\ (\mathcal{B}^C,T) \end{array}
\begin{array}{c}
\begin{array}{cccccc} (\mathcal{B}^L,T) & \mathcal{B}^L,C) & (\mathcal{B}^T,T) & (\mathcal{B}^C,C) & (\mathcal{B}^T,C) & (\mathcal{B}^C,T) \end{array} \\
\begin{pmatrix}
Ss_{L\times L} & Sd_{L\times L} & Ds_{L\times M^T} & Dd_{L\times M^C} & Dd_{L\times M^T} & Ds_{L\times M^C} \\
Sd_{L\times L} & Ss_{L\times L} & Dd_{L\times M^T} & Ds_{L\times M^C} & Ds_{L\times M^T} & Dd_{L\times M^C} \\
Ds_{M^T\times L} & Dd_{M^T\times L} & Ss_{M^T\times M^T} & Dd_{M^T\times M^C} & Sd_{M^T\times M^T} & Ds_{M^T\times M^C} \\
Dd_{M^C\times L} & Ds_{M^C\times L} & Dd_{M^C\times M^T} & Ss_{M^C\times M^C} & Ds_{M^C\times M^T} & Sd_{M^C\times M^C} \\
Dd_{M^T\times L} & Ds_{M^T\times L} & Sd_{M^T\times M^T} & Ds_{M^T\times M^C} & Ss_{M^T\times M^T} & Dd_{M^T\times M^C} \\
Ds_{M^C\times L} & Dd_{M^C\times L} & Ds_{M^C\times M^T} & Sd_{M^C\times M^C} & Dd_{M^C\times M^T} & Ss_{M^C\times M^C}
\end{pmatrix}
\end{array},
$$

where the $K \times K$ design covariance matrix of the $\underline{\text{S}}$ame block group and $\underline{\text{s}}$ame assignment is

$$Ss_{K\times K} = \begin{pmatrix} 2^{-1}+N^{-1} & N^{-1} & \cdots & N^{-1} \\ N^{-1} & 2^{-1}+N^{-1} & \cdots & N^{-1} \\ \vdots & \vdots & \ddots & \vdots \\ N^{-1} & N^{-1} & \cdots & 2^{-1}+N^{-1} \end{pmatrix},$$

the $K \times K$ design covariance matrix of the $\underline{\text{S}}$ame block group and $\underline{\text{d}}$ifferent assignment is

$$Sd_{K\times K} = \begin{pmatrix} 2^{-1}-N^{-1} & -N^{-1} & \cdots & -N^{-1} \\ -N^{-1} & 2^{-1}-N^{-1} & \cdots & -N^{-1} \\ \vdots & \vdots & \ddots & \vdots \\ -N^{-1} & -N^{-1} & \cdots & 2^{-1}-N^{-1} \end{pmatrix},$$

the $K_1 \times K_2$ design covariance matrix of $\underline{\text{D}}$ifferent block groups and $\underline{\text{s}}$ame assignment is

$$Ds_{K_1\times K_2} = \begin{pmatrix} N^{-1} & N^{-1} & \cdots & N^{-1} \\ N^{-1} & N^{-1} & \cdots & N^{-1} \\ \vdots & \vdots & \ddots & \vdots \\ N^{-1} & N^{-1} & \cdots & N^{-1} \end{pmatrix},$$

and the $K_1 \times K_2$ design covariance matrix of $\underline{\text{D}}$ifferent block groups and $\underline{\text{d}}$ifferent assignment is

$$Dd_{K_1\times K_2} = \begin{pmatrix} -N^{-1} & -N^{-1} & \cdots & -N^{-1} \\ -N^{-1} & -N^{-1} & \cdots & -N^{-1} \\ \vdots & \vdots & \ddots & \vdots \\ -N^{-1} & -N^{-1} & \cdots & -N^{-1} \end{pmatrix}.$$

It follows

$$V = \begin{array}{c} \\ (b \in \mathcal{B}^L, T) \\ (b \in \mathcal{B}^L, C) \\ (b \in \mathcal{B}^T, T) \\ (b \in \mathcal{B}^C, C) \end{array} \begin{array}{c} \begin{array}{cc} (\mathcal{B}^T, C) & (\mathcal{B}^C, T) \end{array} \\ \begin{pmatrix} -Dd_{L \times M^T} & -Ds_{L \times M^C} \\ -Ds_{L \times M^T} & -Dd_{L \times M^C} \\ -Sd_{M^T \times M^T} & -Ds_{M^T \times M^C} \\ -Ds_{M^C \times M^T} & -Sd_{M^C \times M^C} \end{pmatrix} \end{array}$$

$$W = \begin{array}{c} \\ (b \in \mathcal{B}^T, C) \\ (b \in \mathcal{B}^C, T) \end{array} \begin{array}{c} \begin{array}{cc} (\mathcal{B}^T, C) & (\mathcal{B}^C, T) \end{array} \\ \begin{pmatrix} I_{M^T \times M^T} - Ss_{M^T \times M^T} & -Dd_{M^T \times M^C} \\ -Dd_{M^C \times M^T} & I_{M^C \times M^C} - Ss_{M^C \times M^C} \end{pmatrix} \end{array}$$

$$W^{-1} = \begin{array}{c} \\ (b \in \mathcal{B}^T, C) \\ (b \in \mathcal{B}^C, T) \end{array} \begin{array}{c} \begin{array}{cc} (\mathcal{B}^T, C) & (\mathcal{B}^C, T) \end{array} \\ \begin{pmatrix} S_{M^T \times M^T} & D_{M^T \times M^C} \\ D_{M^C \times M^T} & S_{M^C \times M^C} \end{pmatrix} \end{array},$$

where

$$S_{K \times K} = \begin{pmatrix} 2 + 2L^{-1} & 2L^{-1} & \cdots & 2L^{-1} \\ 2L^{-1} & 2 + 2L^{-1} & \cdots & 2L^{-1} \\ \vdots & \vdots & \ddots & \vdots \\ 2L^{-1} & 2L^{-1} & \cdots & 2 + 2L^{-1} \end{pmatrix}$$

$$D_{K_1 \times K_2} = \begin{pmatrix} -2L^{-1} & -2L^{-1} & \cdots & -2L^{-1} \\ -2L^{-1} & -2L^{-1} & \cdots & -2L^{-1} \\ \vdots & \vdots & \ddots & \vdots \\ -2L^{-1} & -2L^{-1} & \cdots & -2L^{-1} \end{pmatrix}.$$

We have

$$V' Y_{R=1} = \begin{array}{c} (\mathcal{B}^T, T) \\ (\mathcal{B}^C, C) \end{array} \begin{pmatrix} \frac{1}{N}\Big(\sum_{b \in \mathcal{B}^L \cup \mathcal{B}^T} Y_{b,T} - \sum_{b \in \mathcal{B}^L \cup \mathcal{B}^C} Y_{b,C} \Big) - \frac{1}{2} Y_{\mathcal{B}^T, T} \\ -\frac{1}{N}\Big(\sum_{b \in \mathcal{B}^L \cup \mathcal{B}^T} Y_{b,T} - \sum_{b \in \mathcal{B}^L \cup \mathcal{B}^C} Y_{b,C} \Big) - \frac{1}{2} Y_{\mathcal{B}^C, C} \end{pmatrix}.$$

Since

$$-W^{-1}V'Y_{R=1} = \hat{Y}_{R=0}$$
$$\hat{\tau}_B = \frac{1}{L}\Big(\sum_{b\in\mathcal{B}^L} Y_{b,T} - \sum_{b\in\mathcal{B}^L} Y_{b,C}\Big),$$

the imputed outcomes are

$$\begin{aligned}
\hat{Y}_{b\in\mathcal{B}^T,C} &= Y_{b,T} - \Big[\Big(\frac{2}{N} + \frac{2(M^T+M^C)}{NL}\Big)\Big(\sum_{b\in\mathcal{B}^L\cup\mathcal{B}^T} Y_{b,T} - \sum_{b\in\mathcal{B}^L\cup\mathcal{B}^C} Y_{b,C}\Big) - \frac{1}{L}(\sum_{b\in\mathcal{B}^T} Y_{b,T} - \sum_{b\in\mathcal{B}^C} Y_{b,C})\Big] \\
&= Y_{b,T} - \hat{\tau}_B \\
\hat{Y}_{b\in\mathcal{B}^C,T} &= Y_{b,C} + \Big[\Big(\frac{2}{N} + \frac{2(M^T+M^C)}{NL}\Big)\Big(\sum_{b\in\mathcal{B}^L\cup\mathcal{B}^T} Y_{b,T} - \sum_{b\in\mathcal{B}^L\cup\mathcal{B}^C} Y_{b,C}\Big) - \frac{1}{L}(\sum_{b\in\mathcal{B}^T} Y_{b,T} - \sum_{b\in\mathcal{B}^C} Y_{b,C})\Big] \\
&= Y_{b,C} + \hat{\tau}_B.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\hat{\tau}_{SI} &= E(W_F^T \circ \hat{Y}^{(S)}) - E(W_F^C \circ \hat{Y}^{(S)}) \\
&= \bar{b}^{-1}\sum_{b=1}^{\bar{b}}(\hat{Y}_{b,T} - \hat{Y}_{b,C}) \\
&= \frac{1}{L}\sum_{b\in\mathcal{B}^L}(Y_{b,T} - Y_{b,C}) + \frac{1}{M^T}\sum_{b\in\mathcal{B}^T}(Y_{b,T} - \hat{Y}_{b,C}) + \frac{1}{M^C}\sum_{b\in\mathcal{B}^C}(\hat{Y}_{b,T} - Y_{b,C}) \\
&= \hat{\tau}_B + \frac{1}{M^T}\sum_{b\in\mathcal{B}^T}\hat{\tau}_B + \frac{1}{M^C}\sum_{b\in\mathcal{B}^C}\hat{\tau}_B \\
&= \hat{\tau}_B. \square
\end{aligned}$$

Unique Decomposition of Potential Outcome

The ATE is calculated as

$$\begin{aligned}
\bar{\tau} &\equiv E(W_P^T \circ Y) - E(W_P^C \circ Y) \\
&= \frac{\sum_{b=1}^2\sum_{u=1}^2 W_{P,b,u}^T Y_{b,u}(1)}{\sum_{b=1}^2\sum_{u=1}^2 W_{P,b,u}^T} - \frac{\sum_{b=1}^2\sum_{u=1}^2 W_{P,b,u}^C Y_{b,u}(0)}{\sum_{b=1}^2\sum_{u=1}^2 W_{P,b,u}^C}.
\end{aligned}$$

The block treatment effect of block b is defined as

$$\bar{\tau}_b \equiv \frac{\sum_{u=1}^{2} W_{P,b,u}^T Y_{b,u}(1)}{\sum_{u=1}^{2} W_{P,b,u}^T} - \frac{\sum_{u=1}^{2} W_{P,b,u}^C Y_{b,u}(0)}{\sum_{u=1}^{2} W_{P,b,u}^C}.$$

The unit treatment effect of unit (b, u) is defined as

$$\tau_{b,u} \equiv \frac{W_{P,b,u}^T Y_{b,u}(1)}{W_{P,b,u}^T} - \frac{W_{P,b,u}^C Y_{b,u}(0)}{W_{P,b,u}^C}.$$

Between-block heterogeneity of treatment effect of block b is derived by

$$\Delta_b^B \tau = \bar{\tau}_b - \bar{\tau}.$$

Within-block heterogeneity of treatment effect of unit (b, u) is derived by

$$\Delta_{b,u}^W \tau = \tau_{b,u} - \bar{\tau}_b.$$

The average of control potential outcomes is derived by

$$\begin{aligned} \alpha &\equiv E(W_P^C \circ Y) \\ &= \frac{\sum_{b=1}^{2} \sum_{u=1}^{2} W_{P,b,u}^C Y_{b,u}(0)}{\sum_{b=1}^{2} \sum_{u=1}^{2} W_{P,b,u}^C}. \end{aligned}$$

The average of control potential outcomes of block b is defined as

$$\alpha_b \equiv \frac{\sum_{u=1}^{2} W_{P,b,u}^C Y_{b,u}(0)}{\sum_{u=1}^{2} W_{P,b,u}^C}.$$

Between-block error of block b is derived by

$$\Delta_b^B \epsilon = \alpha_b - \alpha.$$

Within-block error of unit (b, u) is derived by

$$\Delta^W_{b,u}\epsilon = Y_{b,u}(0) - \alpha_b.$$

Sketch of Proof

Proposition 1. Before we prove Proposition 1, we prove the following two lemmas.

Lemma 1

$$E(W \circ (\Delta y + \Delta y')) = E(W \circ \Delta y) + E(W \circ \Delta y')$$

Proof

$$\begin{aligned} E(W \circ (\Delta y + \Delta y')) &= \frac{\sum_{b=1}^{\bar{b}} \sum_{u=1}^{2} W_{b,u}(\Delta y + \Delta y')}{\sum_{b=1}^{\bar{b}} \sum_{u=1}^{2} W_{b,u}} \\ &= \frac{\left(\sum_{b=1}^{\bar{b}} \sum_{u=1}^{2} W_{b,u}\Delta y\right) + \left(\sum_{b=1}^{\bar{b}} \sum_{u=1}^{2} W_{b,u}\Delta y'\right)}{\sum_{b=1}^{\bar{b}} \sum_{u=1}^{2} W_{b,u}} \\ &= E(W \circ \Delta y) + E(W \circ \Delta y') \end{aligned}$$

□

Lemma 2

$$\begin{aligned} \Delta\hat{\tau} = & E(W^T \circ (\Delta^B\epsilon + \Delta^W\epsilon + \Delta^B\tau + \Delta^W\tau)) - \\ & E(W^C \circ (\Delta^B\epsilon + \Delta^W\epsilon)). \end{aligned}$$

Proof *Note that* $T^2_{b,u} = T_{b,u}$, $(1 - T_{b,u})^2 = 1 - T_{b,u}$ *and* $T_{b,u}(1 - T_{b,u}) = 0$. *It follows that* $W^T_{b,u}T_{b,u} = W^T_{b,u}, W^C_{b,u}(1 - T_{b,u}) = W^C_{b,u}$, *and* $W^T_{b,u}(1 - T_{b,u}) = W^C_{b,u}T_{b,u} = 0$. *Thus, by Lemma*

1,

$$
\begin{aligned}
E(W^T \circ Y) &= E(W^T \circ (T(1)Y(1) + (1 - T(1))Y(0))) \\
&= E(W^T \circ Y(1)) \\
E(W^C \circ Y) &= E(W^T \circ (T(1)Y(1) + (1 - T(1))Y(0))) \\
&= E(W^C \circ Y(0))
\end{aligned}
$$

By Lemma 1,

$$
\begin{aligned}
\hat{\tau} - \bar{\tau} =& E(W^T \circ Y) - E(W^C \circ Y) - \bar{\tau} \\
=& E(W^T \circ Y(1)) - E(W^C \circ Y(0)) - \bar{\tau} \\
=& E(W^T \circ (\alpha + \Delta^B \epsilon + \Delta^W \epsilon + \bar{\tau} + \Delta^B \tau + \Delta^W \tau)) \\
& - E(W^C \circ (\alpha + \Delta^B \epsilon + \Delta^W \epsilon)) - \bar{\tau} \\
=& E(W^T \circ \alpha) + E(W^T \circ \bar{\tau}) + E(W^T \circ (\Delta^B \epsilon + \Delta^W \epsilon + \Delta^B \tau + \Delta^W \tau)) \\
& - E(W^C \circ \alpha) - E(W^C \circ (\Delta^B \epsilon + \Delta^W \epsilon)) - \bar{\tau} \\
=& \alpha + \bar{\tau} + E(W^T \circ (\Delta^B \epsilon + \Delta^W \epsilon + \Delta^B \tau + \Delta^W \tau)) \\
& - \alpha - E(W^C \circ (\Delta^B \epsilon + \Delta^W \epsilon)) - \bar{\tau} \\
=& E(W^T \circ (\Delta^B \epsilon + \Delta^W \epsilon + \Delta^B \tau + \Delta^W \tau)) \\
& - E(W^C \circ (\Delta^B \epsilon + \Delta^W \epsilon)).
\end{aligned}
$$

□

Note that

$$\begin{aligned}\sum_{u=1}^{2} W_{B,b,u}^{T}\Delta_{b,u}^{W}\epsilon &= \sum_{u=1}^{2} R_{b,u}(1)R_{b,-u}(0)T_{b,u}\Delta_{b,u}^{W}\epsilon \\ &= \sum_{-u=1}^{2} R_{b,-u}(1)R_{b,u}(0)T_{b,-u}\Delta_{b,-u}^{W}\epsilon \\ &= \sum_{-u=1}^{2} R_{b,u}(0)R_{b,-u}(1)(1-T_{b,u})(-\Delta_{b,u}^{B}\epsilon) \\ &= -\sum_{u=2}^{1} R_{b,u}(0)R_{b,-u}(1)(1-T_{b,u})\Delta_{b,u}^{B}\epsilon \\ &= -\sum_{u=1}^{2} R_{b,u}(0)R_{b,-u}(1)(1-T_{b,u})\Delta_{b,u}^{B}\epsilon \\ &= -\sum_{u=1}^{2} W_{B,b,u}^{C}\Delta_{b,u}^{W}\epsilon. \qquad (1)\end{aligned}$$

By replacing $\Delta_{b,u}^{W}\epsilon$ and $-\Delta_{b,u}^{W}\epsilon$ by 1 in Equatiion (1), we can derive

$$\sum_{u=1}^{2} W_{B,b,u}^{T} = \sum_{u=1}^{2} W_{B,b,u}^{C}.$$

Thus,

$$E(W_B^T \circ \Delta^W \epsilon) = -E(W_B^C \circ \Delta^W \epsilon).$$

By replacing $\Delta_{b,u}^{W}\epsilon$ and $-\Delta_{b,u}^{W}\epsilon$ by $\Delta_{b}^{B}\epsilon$ in Equatiion (1), we can derive

$$E(W_B^T \circ \Delta^B \epsilon) = E(W_B^C \circ \Delta^B \epsilon).$$

By replacing $R_{b,u}(t)$ by 1 in Equatiion (1), we can derive

$$E(W_F^T \circ \Delta^W \epsilon) = -E(W_F^C \circ \Delta^W \epsilon)$$

$$E(W_F^T \circ \Delta^B \epsilon) = E(W_F^C \circ \Delta^B \epsilon).$$

□

Proposition 2.

$$\begin{aligned}\mathbb{E}[E(W_F^T \circ \Delta^W)] =& \mathbb{E}\Big[\frac{\sum_{b=1}^{\bar{b}} \sum_{u=1}^{2} T_{b,u}\Delta_{b,u}^W}{\sum_{b=1}^{\bar{b}} \sum_{u=1}^{2} T_{b,u}}\Big] \\ =& \mathbb{E}\Big[\frac{\sum_{b=1}^{\bar{b}} \sum_{u=1}^{2} T_{b,u}\Delta_{b,u}^W}{\sum_{b=1}^{\bar{b}} 1}\Big] \\ =& \frac{\sum_{b=1}^{\bar{b}} \frac{\Delta_{b,1}^W+\Delta_{b,2}^W}{2}}{\bar{b}} \\ =& 0.\end{aligned}$$

□

Proposition 3. When $R_{b,u}(1) = R_{b,u}(0) = R_{b,u}$,

$$\begin{aligned}\mathbb{E}[E(W_B^T \circ \Delta^W)] =& \mathbb{E}\Big[\frac{\sum_{b=1}^{\bar{b}} \sum_{u=1}^{2} R_{b,u}(1)R_{b,-u}(0)T_{b,u}\Delta_{b,u}^W}{\sum_{b=1}^{\bar{b}} \sum_{u=1}^{2} R_{b,u}(1)R_{b,-u}(0)T_{b,u}}\Big] \\ =& \mathbb{E}\Big[\frac{\sum_{b=1}^{\bar{b}} R_{b,1}R_{b,2} \sum_{u=1}^{2} T_{b,u}\Delta_{b,u}^W \epsilon}{\sum_{b=1}^{\bar{b}} R_{b,1}R_{b,2} \sum_{u=1}^{2} T_{b,u}}\Big] \\ =& \mathbb{E}\Big[\frac{\sum_{b=1}^{\bar{b}} R_{b,1}R_{b,2} \sum_{u=1}^{2} T_{b,u}\Delta_{b,u}^W \epsilon}{\sum_{b=1}^{\bar{b}} R_{b,1}R_{b,2}}\Big] \\ =& \frac{\sum_{b=1}^{\bar{b}} R_{b,1}R_{b,2} \frac{\Delta_{b,1}^W+\Delta_{b,2}^W}{2}}{\sum_{b=1}^{\bar{b}} R_{b,1}R_{b,2}} \\ =& 0.\end{aligned}$$

□

Special Cases

Ready-made Blockwise Deletion. Suppose that, whenever one unit in a pair has missing outcome, the other unit in the same pair also has missing outcome. For instance, when a treated twin does not respond, the other controled twin living together may not, either. In this case, obviously, the unitwise deletion estimator is equivalent to the blockwise deletion estimator.

Proposition 6 (Ready-made Blockwise Deletion) *When* $R_{b,u}(t) = R_{b,-u}(1-t)$*,*

$$\hat{\tau}_U = \hat{\tau}_B$$

Proof

$$\begin{aligned} W_{B,b,u}^T =& R_{b,u} R_{b,-u} T_{b,u} \\ =& R_{b,u}(1) R_{b,-u}(0) T_{b,u} \\ =& (R_{b,u}(1))^2 T_{b,u} \\ =& R_{b,u}(1) T_{b,u} \\ =& W_{U,b,u}^T \end{aligned}$$

Similarly, $W_{B,b,u}^C = W_{U,b,u}^C$. □

Inversed Correlation between $R(t)$ and $\Delta^B \tau$.

Proposition 7 (Inversed Correlation between $R(t)$ and $\Delta^B \tau$) *Suppose that, in the case of* $R_{b,u}(t) = 0$*, there exists* $-b$ *such that* $\Delta_b^B \tau = -\Delta_{-b}^B \tau$ *and* $R_{b,-u}(t) = R_{-b,u}(1-t) = R_{-b,-u}(1-t) = 0$*. Then,*

$$\Delta \hat{\tau}_B = E(W_B^T \circ (2\Delta^W \epsilon + \Delta^W \tau))$$

Proof *Denote* $\mathcal{B} = \{b | R_{b,u}(t) = 0\}$.

$$
\begin{aligned}
\sum_{b=1}^{\bar{b}} \sum_{u=1}^{2} W_{B,b,u}^T \Delta_{b,u}^B \tau =& \sum_{b \notin \mathcal{B}} \sum_{u=1}^{2} R_{b,u}(1) R_{b,-u}(0) T_{b,u} \Delta_b^B \tau + \\
& \sum_{b \in \mathcal{B}} \sum_{u=1}^{2} R_{b,u}(1) R_{b,-u}(0) T_{b,u} \Delta_b^B \tau \\
=& \sum_{b \notin \mathcal{B}} \sum_{u=1}^{2} T_{b,u} \Delta_b^B \tau \\
=& \sum_{b \notin \mathcal{B}} \Delta_b^B \tau \\
=& \sum_{b \notin \mathcal{B}} \Delta_b^B \tau + \sum_{b \in \mathcal{B}} \Delta_b^B \tau \\
=& \sum_{b=1}^{\bar{b}} \Delta_b^B \tau \\
=& 0 \\
\therefore E(W_B^T \circ \Delta^B \tau) =& 0
\end{aligned}
$$

□

Therefore, $\Delta \hat{\tau}_B$ is free of $\Delta^B \tau$. This setup is helpful, though it seems to be rare, unfortunately.

Application

Experiment. For multiple imputation, this paper chooses 17 pretreatment variables as follows. Supposing missing occurred only after treatment assignment, not before it, this paper does not use variables which are obtained after treatment assignment (i.e. which are included in the follow up survey, not the baseline survey, and whose variable names end with `.T2`) except the outcome (`cata1.30.T2`). Moreover, since the imputation procedure cannot deal with many variables relative to the number of units (100), this paper focuses on expenditure related variables as well as face sheet variables and exclude health-care expenditure variables (except cata1.30 in the baseline survey, which corresponds to the outcome (cata1.30.T2 in the follow up survey)) due to high correlation among them. Nominal variables (cluster ID and its counterpart ID) and a constant variable (BMI) are also discarded. Finally, four variables (sp.op, hhage (age of head of household), hheduc (education of head of household), and interact) are dropped because they are highly correlated with other variables. After this procedure, 17 variables remain. Their names are assets, age, sex, educ, hhsex, hhsize, in.sp, treatment, urbrur, indigenous, Oportunidades, margw, dependents, media, P06D01, P12D01, and cata1.30. They overlap the variables King et al. (2007, 492, f.n. 9) used for matching pairs.

Observational Study. The dataset file, cps1re74.dta, was downloaded from http://economics.mit.edu/faculty/angrist/data1/mhe/dehejia (access on December 23, 2013). Readers might be interested in websites of Dehejia and Wahba (1999).[24] The nine matched-on variables which this paper utilizes for matching and multiple imputation are age, its square, years of schooling, four indicator variables for blacks, Hispanics, martial status, and high school diploma, real earnings in 1974, and real earnings in 1975. Following demonstration in Matching package, the next version of the paper will also use three square terms of years of schooling, real earnings in 1974, and real earnings in 1975.[25] It will also add two unemployment dummies in 1974 and 1975. Since randomization occurred between March 1975 and July 1977 (Dehejia and Wahba, 1999, 1054), real earnings in 1974 and real earnings in 1975 are pretreatment variables and real earnings in 1978 is a post-treatment outcome.

[24] http://users.nber.org/~rdehejia/nswdata.html and http://users.nber.org/~rdehejia/nswdata2.html.

[25] http://sekhon.berkeley.edu/matching/DehejiaWahba.Rout. As for Matching package and its companion papers, see http://sekhon.berkeley.edu/matching/.