Statistica Sinica Preprint No: SS-2016-0320.R1	
Title	Propensity Score Matching Analysis for Causal Effects with MNAR Covariates
Manuscript ID	SS-2016-0320.R1
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202016.0320
Complete List of Authors	Bo Lu and Robert Ashmead
Corresponding Author	Bo Lu
E-mail	lu.232@osu.edu

Statistica Sinica

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0320

Statistica Sinica

Propensity Score Matching Analysis for Causal Effects with MNAR Covariates

Bo Lu[1] and Robert Ashmead[2]

[1] *The Ohio State University,* [2] *U.S. Census Bureau*

Abstract: In observational studies, propensity score methods are popular for estimating causal effects. With completely observed data, this approach is valid under several assumptions; however, in practice data are often missing which can have a substantial impact on the estimation. Current remedies to deal with missing covariates in propensity score methods generally fall into two categories. Some authors propose to account for the missing data patterns in propensity score estimation. Others propose to first impute the missing data, then utilize conventional propensity score adjustment methods. Both approaches assume that the data are missing at random (MAR), and there is little discussion regarding the impact on treatment effect estimation if covariates are missing not at random (MNAR). In this paper, we first examine the implication of the MAR assumption under the potential outcome framework. We then propose a sensitivity analysis method for assessing the impact of a MNAR covariate on treatment effect estimation with a matching estimator, with varying magnitudes of unmeasured confounding effect due to the missing covariate. Our method takes full advantage of the information contained in the partially missing covariate by matching

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0320

on the observed portion and identifying a bounding distribution for the missing portion. It can be interpreted similarly as Rosenbaum's sensitivity analysis, and the results are robust since we make few parametric assumptions. We illustrate the application of the method using the 2012 Ohio Medicaid Assessment Survey (OMAS) to investigate the effect of health insurance on health outcomes, where an important covariate, household income, is partially missing.

Key words and phrases: Propensity score; Matching; Not Missing At Random; Sensitivity Analysis.

1. Introduction

Modern causal inference methodology is built on the potential outcome framework, which assumes the existence of a pair of outcomes for each subject, one under treatment and the other under control. In observational data we can never observe both potential outcomes for each subject at the same time. Therefore causal inference can be thought of as a missing data problem where we try to infer about the missing potential outcomes and estimate the causal effect. The assumption that the unobserved potential outcomes are missing at random (MAR) given observed confounders provides the basis for estimation of the treatment effect. The missing data mechanism of the confounders plays a critical role in estimating the causal effect. While some authors have focused research on understanding the impact of completely unobserved confounders (Rosenbaum (1987)), an often

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0320

overlooked detail in causal inference is partially missing confounders, which seems inevitable in observational data.

It is quite challenging to handle missingness in both the potential outcomes and covariates as the missing covariate not only has implications for the treatment assignment, but it may affect the outcome as well. Without appropriate adjustment of the partially missing confounder, the treatment effect estimate may be biased. In observational studies, propensity score-based methods are widely used to estimate the average causal effect, due to the lack of randomization. The propensity score is commonly used to stratify the data, to construct weights, or to create matched pairs. Propensity score-based methods are shown to yield unbiased treatment effect estimates under certain assumptions, which include treatment assignment ignorability and stable unit treatment values (Rubin (1978), Rosenbaum and Rubin (1983), and Lunceford and Davidian (2004)). These assumptions are usually not verifiable with observed data. For example, the ignorability assumption implies that there is no unmeasured confounding given the observed set of covariates.

There are two general strategies for dealing with missing covariate data in the literature: a pattern-mixture model approach using a generalized propensity score and a multiple imputation approach. Rosenbaum and

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0320

Rubin (1984) proposed the generalized propensity score, defined as the probability of treatment conditional on a set of observed covariates and a missing data pattern indicator

$$\mathbf{e}^* = P(T = 1|\mathbf{X}_{obs}, \mathbf{M}), \tag{1.1}$$

where T is the treatment indicator, $\mathbf{X}_{obs}$ are the non-missing values of the complete data $\mathbf{X}$, and $\mathbf{M}$ is a vector that indicates which variables in the complete data $\mathbf{X}$ are missing. Let $\mathbf{X}^*$ denote the non-missing values of $\mathbf{X}$ with indicators ($*$'s) in place of any missing data. Rosenbaum and Rubin proved that $\mathbf{X}^* \perp\!\!\!\perp T|\mathbf{e}^*$, meaning that $\mathbf{e}^*$ balances the observed covariates along with the missing data patterns. This does not imply that $\mathbf{X} \perp\!\!\!\perp T|\mathbf{e}^*$, meaning that the unobserved values of $\mathbf{X}$ are not necessarily balanced conditional on $\mathbf{e}^*$, only the missingness patterns. Rosenbaum and Rubin recommended using separate models to estimate the generalized propensity score for each missingness pattern in large studies, and treating the missing values as additional categories for discrete variables in a single model when there are a large number of missing data patterns. D'Agostino and Rubin (2000) extended the idea by using a general location model to jointly model $(Z, \mathbf{X}, \mathbf{M})$, where model parameters are estimated using EM and ECM algorithms.

Another popular alternative to handling incomplete data is imputation.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0320

Using Rubin's methodology for multiple imputation (Rubin (2009)), Crowe et al. (2010) examined the properties of various multiple imputation strategies in a simulation for a binary outcome using a stratification estimator. They found that multiple imputation methods outperformed the complete case analysis, and the best performance came when including the outcome in the imputation model along with the treatment and covariates. Hill (2004) considered two modified imputation strategies for missing covariate data which either combined treatment effect estimates across multiply imputed datasets or averaged the propensity score estimates across the imputations, then estimated a single treatment effect. Through simulation, the former procedure was found to be slightly less biased as well as having a smaller Monte Carlo variance than the latter. Qu and Lipkovich (2009) proposed a multiple imputation missingness pattern (MIMP) approach for dealing with missing covariate data in a propensity score analysis. Simulation results showed that the MIMP approach performed better than a multiple imputation without indicators for the missing data pattern in the presence of a MNAR mechanism. However, it showed no improvements over multiple imputation under MCAR or MAR mechanisms. Interestingly, Rubin and Rosenbaum's pattern mixture model (Rosenbaum and Rubin (1984)), which included no imputation, performed as well or better than MIMP

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0320

in the MNAR case. Mitra and Reiter (2011) proposed another method combining D'Agostino and Rubin's general location model with multiple imputation.

A key assumption for most existing methods is that covariates are missing at random, but covariates not missing at random have seldom been discussed in the literature. Moreover, most imputation approaches try to take advantage of as much observed information as possible, i.e., the observed outcome is used in the imputation. This practice ignores the subtle but critical point that the observed outcome is a function of the potential outcomes, which cannot be fully observed in nature. Our paper discusses the MAR and MNAR assumptions under the potential outcome framework, and proposes a sensitivity analysis method for assessing potential changes of causal effect with a MNAR covariate. In Section 2, we elaborate MAR and MNAR assumptions under the potential outcome framework. In Section 3, we propose a sensitivity analysis approach to assessing the impact due to a MNAR covariate. In Section 4, we discuss the large sample approximation for the proposed method to speed up computations. In Section 5, we apply our method to a dataset which suffered from the missingness of an important covariate. Lastly, in Section 6, we provide some guidelines on using our method in practice, and discuss the limitations.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0320

2. Ignorability: Missingness and Treatment Assignment

In the causal inference problem with missing covariates, the ignorability assumption needs to be carefully examined. It involves two different but related components: ignorability of missing covariates and ignorability of the treatment assignment.

2.1 Notation

To facilitate the discussion, we introduce the following notation:

T: Treatment indicator ($T = 1$ for treated and $T = 0$ for control)

Y^1: Potential outcome under treatment

Y^0: Potential outcome under control

Y: Observed outcome, $Y = T \times Y^1 + (1 - T) \times Y^0$

X: Fully observed covariates

Z: A partially observed covariate

R: Response indicator for Z ($R = 1$ for observed and $R = 0$ for missing)

The observed data vector includes (T, Y, X, Z^{obs}, R), and the missing data vector includes $((1 - T)Y^1, TY^0, Z^{mis})$. We only focus on the case with a partially missing covariate in this paper. The fully missing covariate

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0320

case is equivalent to the the unmeasured confounder problem in causal inference, and it is usually handled with a sensitivity analysis approach such as Rosenbaum (1987) that varies the assumption about treatment assignment ignorability.

2.2 Ignorability Assumptions

There are two important ignorability assumptions for causal inference with partially missing covariates, one for treatment assignment and the other for the missingness mechanism. Following Rosenbaum and Rubin (1983), in general, with confounders X and Z the strongly ignorable treatment assignment assumption implies:

$$(Y^1, Y^0) \perp\!\!\!\perp T | X, Z$$

When Z is fully observed, this is the classic causal inference problem. When Z is only partially observed, strictly speaking, the strongly ignorable treatment assignment assumption does not hold because it is necessary to condition on a variable with unobserved values to achieve conditional independence between the treatment and potential outcomes. To proceed, we need to make an additional assumption on the missingness mechanism.

If the missingness of Z depends only on the observed X, an ignorable missing or MAR case, we can obtain an unbiased estimate of the population average treatment effect (PATE), $\Delta_{PATE} = E[Y^1 - Y^0]$, using the observed

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0320

data, as shown in the following proposition:

Proposition 1. *Assume the strongly ignorable treatment assignment defined above and that the missingness of the covariate Z is ignorable depending only on X, $Z \perp\!\!\!\perp R|X$. Then the population average treatment effect can be unbiasedly estimated based on the observed data.*

Proof. Since the missingness of Z depends only on X, we note that Z is missing completely at random (MCAR) given X. Therefore, we have

$$(Y^1, Z, T) \perp\!\!\!\perp R|X.$$

This implies that $Y^1 \perp\!\!\!\perp R|X, Z, T$, the marginal mean of $Y^1|X, Z, T$ is the same regardless of R:

$$E(Y^1|X, Z, T) = E(Y^1|X, Z, T, R = 1).$$

Under the strong ignorability of treatment assignment, we have $E(Y^1|X, Z, T) = E(Y^1|X, Z)$. The fact that Z is MCAR also implies $(Y^1, Z) \perp\!\!\!\perp R|X$. Then we know $Y^1 \perp\!\!\!\perp R|X, Z$, which leads to

$$E(Y^1|X, Z) = E(Y^1|X, Z, R = 1). \tag{2.1}$$

So, we have $E(Y^1|X, Z, T, R = 1) = E(Y^1|X, Z, R = 1)$.

Similarly, we can show $E(Y^0|X, Z, T, R = 1) = E(Y^0|X, Z, R = 1)$.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0320

Then,

$$\Delta_{PATE} = E(Y^1 - Y^0) = E[E(Y^1 - Y^0|X, Z)] = E[E(Y^1 - Y^0|X, Z, R = 1)]$$

$$= E[E(Y^1|X, Z, R = 1) - E(Y^0|X, Z, R = 1)]$$

$$= E[E(Y^1|X, Z, T = 1, R = 1) - E(Y^0|X, Z, T = 0, R = 1)]$$

$$= E[E(Y|X, Z, T = 1, R = 1) - E(Y|X, Z, T = 0, R = 1)]$$

where the third equality follows from (2.1). To obtain the overall population average effect, the outer expectation is taken over the joint distribution of (X, Z). Even though this joint distribution is not directly observed due to missing data, it can be recovered since Z is MCAR given X. Noting that $P(X, Z) = P(Z|X)P(X)$, one can capture the conditional distribution of Z given X by either assuming a parametric distributional relationship between X and Z or imputing Z based on X. □

For imputation methods dealing with missing covariate data, many authors (Qu and Lipkovich (2009), Crowe et al. (2010)) propose to include the observed outcome Y in an effort to take advantage of all available information. Even if the outcome is not related to the missingness mechanism, by its inclusion in the confounder set we relate Z to Y. Therefore, Y may provide information about the distribution of Z whether it is related to its missingness or not. This is a plausible approach but its implication has not

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0320

been fully explored. Under the potential outcome framework, the observed Y is an intermediate variable, as a function of T, Y^1, and Y^0.

Assuming that the missingness depends on Y has complex implications. It essentially assumes that the missingness depends on Y^1 for the treatment group and Y^0 for the control group. Simply saying that missingness depends on observed Y tends to downplay the importance of T. If the treatment has any effect, we need to adjust both T and Y to capture the missing information correctly. A small simulation study illustrates that imputing Z solely using observed Y generates biased results even if the missingness can be characterized as a function of Y.

In this simple example, we simulated $Y^0 \sim \mathcal{N}(50, 10)$ and used a constant effect model, $Y^1 = Y^0 + 10$. We considered one covariate Z set to be $Z = Y^0 + \epsilon$, where $\epsilon \sim \mathcal{N}(0, 1)$. We assumed Z partially missing, and its missingness depended on Y, in the sense that missing occurs when $Y > 55$. The treatment group was randomly assigned to units with a 50% probability independently of everything else. Using the observed data, we imputed for the missing Zs in two ways. Using the fully observed data, we fit a linear model using only Y or using both T and Y to impute for Z. The simulation was repeated 5000 times with a sample size of 100. Table 1 presents the results of percentage bias of Z's mean after imputation for

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0320

Table 1: Bias associated with imputation based on observed outcome

	$T = 0$	$T = 1$
True $\bar{Z}$	50	50
% Bias of $\bar{Z}$, using only Y	-3.6	4.9
% Bias of $\bar{Z}$, using T and Y	0.009	0.018

treated and control groups, with missing data rates of 30% on average.

The estimates of the covariate mean are biased in both groups when only Y is used in imputation. This is because the distribution of $Z|Y = y$ differs from the distribution of $Z|Y = y, T = 1$ or the distribution of $Z|Y = y, T = 0$ (due to the non-zero treatment effect). Note that $(Z|Y = y, T = 1) = (Z|Y^1 = y)$ and $(Z|Y = y, T = 0) = (Z|Y^0 = y)$. This implies that when Y is used in the imputation model, we should either impute Z separately for the treated and control units or include the treatment indicator T in the model as well.

In reality, things are likely much more complicated than this. In the next section, we propose an approach based on sensitivity analysis to gauge the impact of MNAR covariate on treatment effect estimation.

3. Covariates Missing Not At Random

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0320

If the missingness of the covariate depends on unobserved values, we should treat the missing as MNAR. For example, the missingness may depend on the covariate's own values, or another unobserved variable. To investigate the impact of MNAR covariates on the causal effect estimation, we propose a sensitivity analysis strategy for a matching estimator to handle the observed component and the missing component separately. We use propensity score matching to balance the observed data and identify bounds on the treatment effect p-value due to the missing covariate given a hypothetical parameter that represents the assumed association between the variable with missing data and the treatment assignment.

We assume a relationship between the treatment assignment and covariates based on a logit model, which allows us to mathematically identify the bounds for the p-value of the test of no treatment effect for various magnitude of the sensitivity parameter. Then we can reach qualitative conclusions on how likely the causal findings will change due to a partially observed confounder that is MNAR. The larger the hypothetical parameter value required to change the qualitative conclusion, the less likely the causal estimation is affected by the MNAR covariate. This is similar to the idea of Rosenbaum's sensitivity analysis for unmeasured confounding (Rosenbaum (1987)), but the setup is more general. It handles a missing covariate

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0320

with arbitrary missing rate. With 100% missingness, this is a unmeasured confounding problem and our formula is equivalent to Rosenbaum's sensitivity analysis. With 0% missingness, all confounders are observed and our method is simplified to the conventional propensity score matching approach. When the missing data fraction is low or the causal relationship is very strong based on the observed data, even an MNAR confounder may not change the study conclusion and our approach provides a way to gauge how likely this is the case. This is very useful for practitioners since it adds substantial robustness to the study findings, eliminating questions related to different missing data mechanism assumptions in the analysis.

3.1 Identifying Bounds for the Causal Effect under MNAR

Following the notation outlined in section 2.1, we assume that the treatment assignment T is strongly ignorable given $(\mathbf{X}, Z)$. If Z is fully observed, adjusting for the propensity score including both $\mathbf{X}$ and Z would suffice to remove the confounding bias. The propensity score can have a very general form:

$$e = P(T = 1|\mathbf{X}, Z) = g^{-1}[f(\mathbf{X}) + \gamma Z] \tag{3.1}$$

where $g()$ denotes the link function and $f()$ is any function with respect to covariate vector $\mathbf{X}$. We can also assume that

$$Z \perp\!\!\!\perp \mathbf{X}. \tag{3.2}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0320

In practice this assumption has little impact because we could replace Z with $Z' = Z - E(Z|\mathbf{X})$, which, while not necessarily independent of $\mathbf{X}$, is uncorrelated with $\mathbf{X}$ (Rosenbaum (1987)).

When Z is only partially observed, we need to further assume a certain functional form of the relationship between Z and T in order to identify the sensitivity. We pick the logit model as inspired by Rosenbaum's sensitivity analysis and write the propensity score as

$$e = P(T = 1|\mathbf{X}, Z) = logit^{-1}[f(\mathbf{X}) + \gamma Z] = \frac{exp[f(\mathbf{X}) + \gamma Z]}{1 + exp[f(\mathbf{X}) + \gamma Z]}. \quad (3.3)$$

We assume that (3.3) is the true relationship between $T, \mathbf{X}$, and Z going forward. In practice, e cannot be estimated directly due to the missingness of Z, so we instead work with the generalized propensity score which includes the response indicator and is written as

$$e^* = P(T = 1|\mathbf{X}, Z_{obs}, R). \quad (3.4)$$

The generalized propensity score e^* is not known in practice either, but it can be estimated. We discuss how to estimate e^* in Section 3.2.

Assuming the true generalized propensity score e^* is known, it will balance the observed covariates and the missing data patterns when Z is not observed (Rosenbaum and Rubin (1984)). Assume we create S matched pairs exactly on e^* and R with one treated and one untreated subjects in

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0320

each pair. Specifically, we first match on R to ensure a perfect balance on missingness status, and then match on e^* to ensure balance on $\mathbf{X}$ and Z_{obs}. Therefore, among the matched pairs with completely observed data ($R = 1$), both $\mathbf{X}$ and Z are balanced to remove the confounding. Among the matched pairs with missing data ($R = 0$), only $\mathbf{X}$ values are balanced. Since the exact values of Z are not observed, the treatment odds in each pair may differ by $exp(\gamma(z - z'))$, where z and z' are the unobserved values of Z for the matched pair.

Under the potential outcome framework, let (Y_{s1}^1, Y_{s1}^0) denote the pair of potential outcomes for first unit in the s^{th} pair. Then $\{(Y_{s1}^1, Y_{s1}^0), (Y_{s2}^1, Y_{s2}^0)\}$ are the two pairs of potential outcomes in this matched pair. Assuming an additive treatment effect τ, we can write the potential outcomes as

$$Y_{s1}^1 = Y_{s1}^0 + (T_{s1})\tau \quad \text{and} \quad Y_{s2}^1 = Y_{s2}^0 + (T_{s2})\tau.$$

The observed outcomes are then

$$Y_{s1} = T_{s1}Y_{s1}^1 + (1 - T_{s1})Y_{s1}^0 \quad \text{and} \quad Y_{s2} = T_{s2}Y_{s2}^1 + (1 - T_{s2})Y_{s2}^0.$$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0320

If D_s is the difference in observed responses, then

$$\begin{aligned} D_s &= Y_{s1} - Y_{s2} \\ &= T_{s1}(Y_{s1}^0 + T_{s1}\tau) + (1 - T_{s1})Y_{s1}^0 - [T_{s2}(Y_{s2}^0 + T_{s2}\tau) + (1 - T_{s1})Y_{s2}^0] \\ &= Y_{s1}^0 T_{s1} + \tau T_{s1} + (1 - T_s 1)Y_{s1}^0 - [Y_{s2}^0 T_{s2} + \tau T_{s2} + Y_{s2}^0(1 - T_{s1})] \\ &= \tau(T_{s1} - T_{s2}) + Y_{s1}^0(T_{s1} + 1 - T_{s1}) - Y_{s2}^0(T_{s2} + 1 - T_{s2}) \\ &= (Y_{s1}^0 - Y_{s2}^0) + (T_{s1} - T_{s2})\tau. \end{aligned}$$

Let $V_s = T_{s1} - T_{s2}$. Then the treated-minus-control difference in responses in s^{th} pair is $D_s V_s$, where

$$\begin{aligned} D_s V_s &= \left[(Y_{s1}^0 - Y_{s2}^0) + (T_{s1} - T_{s2})\tau\right](T_{s1} - T_{s2}) \\ &= (T_{s1} - T_{s2})(Y_{s1}^0 - Y_{s2}^0) + (T_{s1} - T_{s2})^2\tau \\ &= (T_{s1} - T_{s2})(Y_{s1}^0 - Y_{s2}^0) + \tau. \end{aligned}$$

If each unit in a matched pair has the same probability of treatment assignment e, then the distribution of $(T_{s1} - T_{s2})(Y_{s1}^0 - Y_{s2}^0)$ is centered at zero, which implies that the mean or median of the S treated-minus-control differences yields an estimate of τ.

Without further parametric assumptions, the Wilcoxon signed-rank statistic is used; it can be represented as

$$T = t(\mathbf{D}, \mathbf{V}) = \sum_{s=1}^{S} I(V_s D_s > 0) q(D_s),$$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0320

where $q(D_s)$ is the rank of $|D_s|$. The Wilcoxon signed-rank statistic is not the only statistic that can be used for this method. For example, the treated minus control differences could instead be used. However, to compute the null distribution for such a statistic it is necessary to compute every possible treatment assignment permutation for the set of matched pairs. The signed-rank statistic is more accessible because a closed-form solution for the approximate distribution is available.

Consider the assumptions:

(B.1) Treatment assignment is strongly ignorable given $(\mathbf{X}, Z)$;

(B.2) Treatment assignment is specified by the logit model (3.3) ;

(B.3) S treated/control pairs are matched on R and the true generalized propensity score e^*.

Let $w_s = z_{s1} - z_{s2}$, the difference in values of Z for a matched pair, and $\mathbf{w}$ be the vector of the S observed w_s values. Let B be the set containing the 2^S possible treatment assignments of the matched pairs. Define n_{obs} to be the number of pairs for which Z is not missing, $n_{obs} \leq S$.

Proposition 2. *Under assumptions (B.1) through (B.3) and under the*

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0320

null hypothesis of zero treatment effect, for each k

$$pr\{t(\mathbf{D},\mathbf{V}) \geq k|\mathbf{D}=\mathbf{d},\mathbf{X},\mathbf{Z},\mathbf{R}=\mathbf{r}\} = pr\{t(\mathbf{d},\mathbf{V}) \geq k|\mathbf{W}=\mathbf{w},\mathbf{R}=\mathbf{r}\} \tag{3.5}$$

$$= \sum_{\mathbf{v}\in B} I\{t(\mathbf{d},\mathbf{v}) \geq k\} \left(\frac{1}{2}\right)^{n_{obs}} \prod_{s\in\{r_s=0\}} \frac{exp(\frac{1}{2}\gamma v_s w_s)}{exp(\frac{1}{2}\gamma w_s) + exp(\frac{-1}{2}\gamma w_s)} \tag{3.6}$$

$$= h_k(\mathbf{d},\mathbf{w}).$$

The proof is included in the Appendix.

To identify the explicit form of the bounds on the p-value, we make one additional assumption on Z. We assume that $Z \in [0,1]$ so that $w_s \in [-1,1]$ and the log odds of treatment differ by at most γ for a matched pair with missing Z values. This assumption is minor because, as long as Z is bounded, which is usually the case in data, we can always rescale Z to be within the range of 0 and 1.

Since $\mathbf{W}$ is not completely observed, we can calculate a bound for (3.6) using the "least favorable" value of $\mathbf{w}$ and some estimate of γ. "Least favorable" refers to selecting $\mathbf{w}$ to create a conservative bound. Rosenbaum (1987) shows that the least favorable $\mathbf{w}$ is correlated with d. Then setting w_s equal to 1 or -1 (whichever is in the same direction as d_s) for all unobserved values makes (3.6) larger and thus more conservative.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0320

Some intuition about the least favorable value is as follows. Assume that the treatment effect τ is positive and that $\gamma > 0$ (Z is positively associated with treatment). If the treatment difference $d > 0$, then $w = 1$ would account for some of the difference. This is because $w = 1$ implies that the first observation in the pair was more likely to get the treatment, which in turn had a positive effect on the outcome, giving the first observation higher probability of having a larger outcome.

The most extreme w is positively associated with D, so it can explain away the observed association as much as possible. Let $\tilde{w}(d)_s = sgn(d_s)$ when $R_s = 0$ and $\tilde{w}(d)_s = w_s$ when $R_s = 1$. Also take $\tilde{w}^*(d)_s = -sgn(d_s)$ when $R_s = 0$ and $\tilde{w}^*(d)_s = w_s$ when $R_s = 1$. Then we can identify the bounds for $P(T(\mathbf{D}, \mathbf{V}) \geq k|\mathbf{D} = \mathbf{d}, \mathbf{W} = \mathbf{w})$ with the following.

Proposition 3. *Assume conditions (B.1) through (B.3), the null hypothesis of zero treatment effect, and that $\gamma \geq 0$ hold. Also assume that the partially observed covariate $Z \in [0, 1]$. Then for each possible* $\mathbf{w}$

$$h_k\{\mathbf{d}, \tilde{\mathbf{w}}^*(\mathbf{d})\} \leq P(T(\mathbf{D}, \mathbf{V}) \geq k|\mathbf{D} = \mathbf{d}, \mathbf{W} = \mathbf{w}) \leq h_k\{\mathbf{d}, \tilde{\mathbf{w}}(\mathbf{d})\} \quad (3.7)$$

A proof can be found in the Appendix for Proposition 3; it follows from the idea of Rosenbaum (1987).

3.2 Estimating The Generalized Propensity Score

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0320

We do not observe the generalized propensity score for each unit, and therefore must estimate it. In our scenario we only have two missing data patterns: missing data for Z and no missing data. Consider estimating the generalized propensity scores using logistic regression under the model

$$log\frac{pr(T=1|\mathbf{X},Z_{obs},R)}{pr(T=0|\mathbf{X},Z_{obs},R)} = \beta'\mathbf{X} + \eta I(R=1)Z_{obs} + \alpha I(R=0). \quad (3.8)$$

Ideally, this model should balance the observed covariates and missing data pattern. Specifically, the model building for propensity score can be an iterative process that should follow the general guidelines in Rubin (2007). Another option would be to estimate the generalized propensity score separately for each missing data pattern, but we prefer to borrow strength in case of a small sample. In an extension of this method with multiple missing data patterns, it might be necessary to collapse patterns with small sizes using techniques as suggested in Qu and Lipkovich (2009).

4. Large Sample Approximation

The rank-based test statistics are not easy to compute. Especially, in the presence of missing data, we have to vary the magnitude of γ to assess the impact over a wide range of values. This implies that we need to calculate a series of p-values. In this section, we provide a large sample approximation to the test statistic based on the normal distribution, which is easy to compute.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0320

4.1 Normal Approximation

We derive the mean and variance of the test statistic $t(\mathbf{D}, \mathbf{V})$ under the null hypothesis of no treatment effect when $\mathbf{W} = \tilde{\mathbf{w}}(\mathbf{d})$, which provides the upper bound of the p-value for our test. This computation is conditional on the ranks that have missing data. Under the null hypothesis $\tau = 0$, our test statistic is $T = t(\mathbf{D}, \mathbf{V}) = \sum I(V_s D_s > 0) q(D_s)$. Reorder the matches by their ranked differences $q(\mathbf{D})$, and let the subscript i, $i = 1, \cdots, S$ denote the ordered pairs. In the simplest case when there are no ties in the ranks, $q(D_i) = i$. Then define $P_i = I(V_i D_i > 0) q(D_i), i = 1, \ldots, S$. We can write $T = \sum P_i$.

When $R_i = 1$, $V_i = 1$ with probability $1/2$ and $V_i = -1$ with probability $1/2$. This implies that $P(V_i D_i > 0) = 1/2$ for any D_i. Thus P_i takes on the value $q(D_i)$ with probability $1/2$ and 0 with probability $1/2$.

When $R_i = 0$, P_i takes on the value $q(D_i)$ with probability

$$P(P_i = q(D_i) | R_i = 0) = \frac{exp(\frac{1}{2}\gamma)}{exp(\frac{1}{2}\gamma) + exp(\frac{-1}{2}\gamma)}. \tag{4.1}$$

which follows from considering two cases. When $D_i > 0$, $w = \tilde{w}(d) = 1$. Thus

$$P(V_i = 1 | R_i = 0, W = 1) = \frac{exp(\frac{1}{2}\gamma(1))}{exp(\frac{1}{2}\gamma(1)) + exp(\frac{-1}{2}\gamma(1))}.$$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0320

When $D_i < 0$, $w = \tilde{w}(d) = -1$. We know

$$P(V_i = 1|R_i = 0, W = -1) = \frac{exp(\frac{1}{2}\gamma(-1))}{exp(\frac{1}{2}\gamma(-1)) + exp(\frac{-1}{2}\gamma(-1))},$$

then

$$\begin{aligned} P(V_i = -1|R_i = 0, W = -1) &= 1 - \frac{exp(\frac{1}{2}\gamma(-1))}{exp(\frac{1}{2}\gamma(-1)) + exp(\frac{-1}{2}\gamma(-1))} \\ &= \frac{exp(\frac{1}{2}\gamma(1))}{exp(\frac{1}{2}\gamma(1)) + exp(\frac{-1}{2}\gamma(1))}. \end{aligned}$$

$I(V_i D_i > 0) = 1$ holds when $D_i > 0$ and $V_i = 1$ or when $D_i < 0$ and $V_i = -1$. So we have the distribution of P_i as $P_i = q(D_i)$ with probability at (4.1), and that $P_i = 0$ otherwise.

We can generalize our probabilities as

$$f(r_i) = \frac{exp(\frac{1}{2}r_i\gamma)}{exp(\frac{1}{2}r_i\gamma) + exp(\frac{-1}{2}r_i\gamma)},$$

which equals 1/2 when $r_i = 0$. Then we have

$$E\left[\sum_{i=1}^{S} P_i\right] = \sum_{i=1}^{S} E[P_i] = \sum_{i=1}^{S} f(r_i)q(D_i) = \mu. \tag{4.2}$$

As the P_i's are mutually independent, we have

$$var\left(\sum_{i=1}^{S} P_i\right) = \sum_{i=1}^{S} var(P_i).$$

Then

$$\begin{aligned} var(P_i) &= E(P_i^2) - (E(P_i))^2 = q(D_i)^2(f(r_i)) - (q(D_i)f(r_i))^2 \\ &= q(D_i)^2(f(r_i)) - q(D_i)^2 f(r_i)^2 = q(D_i)^2(f(r_i))(1 - f(r_i)) \end{aligned}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0320

which implies that

$$\sum var(P_i) = \sum_{i=1}^{S} q(D_i)^2 (f(r_i))(1 - f(r_i)) = \sigma^2. \tag{4.3}$$

These results hold generally regardless of missing data rates. When $\gamma = 0$ (when there is no missing data), the mean and variance expressions (4.2) and (4.3) simplify to the traditional mean and variance expressions of the normal approximation to a Wilcoxon signed-rank test. When all pairs have missing data for Z (100% missing), they simplify to Rosenbaum's sensitivity analysis expressions when Z is an unobserved confounder (Rosenbaum (1987)). For any missing data rate in between, it is a mixture of r_i taking values from 0 and 1's.

When using the Wilcoxon signed-rank test we need a modification to the test statistic because of ties between matched pair outcome differences ($D_s = Y_{s1} - Y_{s2} = 0$). We adopt the method described in Section 3.1 of Hollander and Wolfe (1999) to deal with this circumstance. If there are zero values in matched pair outcome differences D_s, discard the non-zero values and redefine the number of matched pairs S appropriately.

4.2 Calculating the P-Value

For a given value of the sensitivity parameter, γ, we use the expressions (4.2) and (4.3) as the null mean and variance of our test statistic $t(\mathbf{D}, \mathbf{V})$ and calculate the p-value of our observed data. We take the upper bound

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0320

of the p-value to be conservative. Let γ_s be the assumed sensitivity value for γ, then the normal approximation of our test statistic is given by

$$Z^*_{\gamma_s} = \frac{t(\mathbf{D}, \mathbf{V}) - \mu_{\gamma_s}}{\sqrt{\sigma^2_{\gamma_s}}}, \tag{4.4}$$

where μ_{γ_s} and $\sigma^2_{\gamma_s}$ are defined by (4.2) and (4.3) for a given value γ_s.

If we are testing the one-sided alternative of a positive treatment effect, $H_0 : \tau = 0$ vs. $H_1 : \tau > 0$, then the p-value for a given γ_s is calculated as $p = P(Z > Z^*_{\gamma_s})$, where $Z \sim N(0, 1)$. If we are testing the two-sided alternative $H_0 : \tau = 0$ vs. $H_1 : \tau \neq 0$, the calculation is slightly more complicated. Let Z^*_0 be the value of Z^* when $\gamma_s = 0$, the unadjusted value. To ensure that the p-value is larger than the unadjusted value, if $Z^*_0 > 0$, then the p-value is given by

$$p(\gamma_s) = min[2 \times min\{0.5, P(Z > Z^*_{\gamma_s})), 2 \times min(0.5, P(Z > Z^*_{-\gamma_s})\}]. \tag{4.5}$$

If $Z^*_0 < 0$ then

$$p(\gamma_s) = min[2 \times min\{0.5, P(Z < Z^*_{\gamma_s})), 2 \times min(0.5, P(Z < Z^*_{-\gamma_s})\}]. \tag{4.6}$$

5. Data Example

While a majority of adults ages 19 through 64 years in Ohio had employer-sponsored health insurance (54.4%) in 2012, approximately one out of every six adults ages 19 through 64 years were uninsured (17.3%)

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0320

Tumin et al. (2013). Using the 2012 Ohio Medicaid Assessment Survey (OMAS), a dual-frame telephone health survey representative of Ohio's non-institutionalized adults, we are interested in estimating the effect of health insurance coverage on the self-rated health of adults 19-64 years of age with incomes between 138% and 400% of the FPL (Federal Poverty Level). This is a relevant segment of the population in reference to the healthcare insurance exchange of the 2010 Patient Protection and Affordable Care Act (ACA).

In the OMAS, there are a series of questions asking about the respondents' income, like most surveys, some people chose not to provide an answer to such questions. In the 2012 OMAS, self-reported annual household income is missing for 18% of the data, while no other variable has a missing rate more than 5%. OMAS has provided a complete dataset with missing values imputed by either hot-deck imputation or regression imputation. For our analysis we first created a study population dataset consisting of adults with reported or imputed incomes between 138% and 400% FPL and reported or imputed ages between 19 and 64 years.

Household income is a key factor in determining health insurance status, could be related to health outcomes, and is likely to be missing not at random (Gelman and Hill (2006)). To illustrate our proposed method-

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0320

ology, we treat income as the only missing covariate in the analysis. After discussions with content experts, we included 13 covariates available to us in the survey in the analysis as predictors in the propensity score model. These variables were race-ethnicity, age, gender, working status, education, disability status, income, county type, children in the household, marital status, smoking status, drink alcohol, and mental health distress. The outcome of interest was self-rated health status based on a five-point Likert scale (1=Excellent, 5=Poor).

In total there were 3,920 persons in our analytical dataset, 483 without health insurance and 3,437 with health insurance. We regarded those without health insurance as "treated" and calculated the treatment effect as the difference between those not having insurance and those do have. We estimated the "average treatment effect on the treated" (ATT), $ATT = E[Y^1 - Y^0|T = 1]$. This can be interpreted as what would have happened to those uninsured if they had been insured. The missing data rate for income was 17% for those without health insurance and 18% for those with health insurance. We computed the propensity score, matched pairs, and the treatment effect for different missing data assumptions.

We used R (R Development Core Team (2008)) for data analysis, and the *Matchby* function in the package *Matching* (Sekhon (2011)) with a

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0320

caliper of 0.25 standard deviations of the propensity score in order to make treatment-control matches. First, under MCAR, a complete case analysis that excluded those units with missing income resulted in 384 matched pairs with a mean difference of self-rated health status of 0.03, which implies having health insurance tends to improve health status by 0.03 points on average. The p-value for the one-sided Wilcoxan signed-rank test was 0.3262. If the cases without missing data are representative of the entire target population, there is no evidence of an effect of insurance status on self-reported health status.

In comparison, we also included two other MAR-based methods in the data analysis. First, we used the OMAS imputed values for income to estimate the propensity score and create matched pairs. Using the imputed values we created 464 matched pairs with a mean difference of 0.20 and a Wilcoxan signed-rank test p-value of 0.0017. This implies that if the MAR assumption and the parametric model used in the OMAS's imputation hold, there is significant evidence of a treatment effect. Additionally, we utilized the method of Qu and Lipkovich (2009) for an additional comparison. This method used a weighting estimator (instead of a matching estimator) of the ATT, so the results might not be as comparable for that reason. We used 5 multiple imputations of the missing values of income and 500 bootstraps

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0320

of the data to estimate the within-imputation variance. The analysis found a highly significant estimated mean difference of self-rated health status of 0.83 points ($p < 0.0001$). These results are summarized in Table 2.

Analyses based on MCAR and MAR show different results. This implies that the missing data mechanism plays a role in how to interpret the evidence from the data. It is sensible to explore further when the missing data differ from the observed data in some unknown way. As pointed out by many researchers, the missingness of income is more likely to be MNAR (Gelman and Hill (2006)). So we applied our sensitivity method to this data to gain more insight on what the inference on the treatment effect might be if income is MNAR. Including units with missing income and estimating the generalized propensity score with a missing variable indicator resulted in 459 matched pairs and a mean difference was 0.13 points. The p-value for the one-sided Wilcoxon signed-rank test was 0.046. This result gives more evidence towards an effect of insurance status on self-reported health status, but we have not taken into account that the missing data may be MNAR. On this same set of matched pairs, we applied our proposed sensitivity method to investigate how the conclusions might change if income is MNAR. The p-values for this method along with the γ values are presented in Table 2 (labeled as "18% missing"). It shows that the

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0320

p-value jumps over the threshold of 0.05 for a small deviation from MAR ($\gamma > 0$). The sensitivity p-value increased to above 0.1 for $\gamma > 0.05$, which means a potential 5% difference in odds of having insurance, due to missing income, may account for the observed health status discrepancy. The sensitivity p-values increased quickly as γ went up. This implies the causal effect estimated based on MAR assumption is sensitive to a small influence from the unknown income if the missing data is MNAR. From a practical perspective, this is plausible since income is such an important determinant for health outcomes. Therefore, we need to be very careful interpreting the findings from the observed data.

To gauge the impact due to a different missing data rate, we used some of the imputed values in the OMAS dataset at random so that 9% of the income values would be missing (half of the original missing rate). We then repeated the propensity score estimation and matching using the generalized propensity score and our method. As shown in Table 2, with a smaller missing data rate, the observed association is slightly more robust to hidden bias. It remains significant for a small influence due to the MNAR covariate with $\gamma = 0.05$. Overall, it still suggests that the significant findings under MAR are sensitive to non-ignorable missing income data. From a statistical perspective, this is because the matched pairs with missing income are as-

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0320

Table 2: Comparison of OMAS analysis results

Method	P-Value
Complete Case Analysis	0.3262
Generalized PS	0.0459
OMAS Imputed Values	0.0017
Qu and Lipkovich	< 0.0001
MNAR Sensitivity, 18% missing, $\gamma = 0$	0.0459
MNAR Sensitivity, 18% missing, $\gamma = 0.05$	0.1008
MNAR Sensitivity, 18% missing, $\gamma = 0.1$	0.19238
MNAR Sensitivity, 18% missing, $\gamma = 0.2$	0.4782
MNAR Sensitivity, 18% missing, $\gamma = 0.3$	0.7763
MNAR Sensitivity, 9% missing, $\gamma = 0$	0.0174
MNAR Sensitivity, 9% missing, $\gamma = 0.05$	0.0447
MNAR Sensitivity, 9% missing, $\gamma = 0.1$	0.0987
MNAR Sensitivity, 9% missing, $\gamma = 0.2$	0.3187
MNAR Sensitivity, 9% missing, $\gamma = 0.3$	0.6350

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0320

sociated with some large outcome differences, which contribute large rank scores in the test. The uncertainty of these income values could substantially weaken the significance of the test under the worst case scenario.

6. Discussion

The implementation of our proposed method is simple, similar to how Rosenbaum's sensitivity analysis was implemented. First we make inference assuming MAR (setting the sensitivity parameter to zero). If the hypothesis test is significant, we then increase γ to the point where the test is no longer significant. This tells us how sensitive the causal conclusion is to assumptions about the missing data. Practically, it remains a question how to effectively choose the value of γ to assess the sensitivity. Since the missing part of the covariate is not ignorable, it is hard to guess the true value of γ. However, in some occasions, the researchers may have a better idea of γ. For example, if there are external data on the covariate and treatment assignment from a similar population, we can estimate their association and use it as a reasonable value for γ. Or we may consult the content expert to solicit a range of reasonable values for γ.

There are two limitations of the method. First, our proposed method can be quite conservative due to the fact that we find the bounding distribution using the "least favorable" approach. The method does not assume

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0320

any parametric assumption for the non-ignorable missing data, only a logit treatment assignment model. We follow Rosenbaum's sensitivity analysis framework to assess the impact of a MNAR covariate on treatment effect estimation. The goal is to evaluate the robustness of the results based on observed data if there is a non-ignorable missing covariate. This approach is designed to be conservative, in the sense that, if the observed finding remains significant under a moderate to large influence from the MNAR covariate, the result is trustworthy. The conservativeness of the method depends on several factors, including the strength of evidence contained in the observed data, the belief of the impact of the MNAR covariate and the missing data rate. These factors vary from study to study. Similar to what is shown in Rosenbaum (2002), some studies are sensitive to small influence from missing covariates, but other studies are more robust. In general, with large γ and a high missingness rate, the method could be very conservative. Usually there is no good way to handle this unless the researcher is willing to make more parametric assumptions. Another alternative is to seek help from instrumental variable (IV) approach if a good IV exists. Yang et al. (2014) discussed the development of IV based strategy for estimating causal effect with potential MNAR covariates. Second, the current development of the method only works for a single missing covariate. This certainly limits

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0320

its practical utility. More research is needed regarding how to generalize it to the situation with multiple missing covariates. Like the OMAS example, when there are other covariates that are rarely missing, we could impute them first before applying our method. In general, with multiple missing covariates, it is difficult to find the bounding distribution of the test statistic and it tends to be very conservative if the impact of those covariates are cumulative. Alternatively, we could assume some parametric structure among missing covariates and observed variables, indexed by several sensitivity parameters. We then can impute the missing data for various values of the sensitivity parameters and proceed our analysis. This is similar to the idea of the imputation based sensitivity analysis for complex observational studies (Lu et al. (2012)).

Acknowledgements

This work was partially supported by grant 1R01 HS024263-01 from the Agency of Healthcare Research and Quality of the U.S. Department of Health and Human Services. We would also like to thank the referee and an associate editor for their insightful comments, which lead to substantial improvement of the paper.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0320

References

Crowe, B. J., Lipkovich, I., and Wang, O. (2010). Comparison of Several Imputation Methods for Missing Baseline Data in Propensity Scores Analysis of Binary Outcome. *Pharmaceutical Statistics* **9**, 269–279.

D'Agostino Jr R. B. and Rubin D. B. (2000). Estimating and Using Propensity Scores with Partially Missing Data. *J. Amer. Statist. Assoc.* **95**, 749-759.

Gelman, A. and Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

Hill, J. (2004). Reducing bias in treatment effect estimation in observational studies suffering from missing data. *Columbia University Institute for Social and Economic Research and Policy (ISERP)* working paper, 04–01.

Hollander, M. and Wolfe, D. A. (1999). *Nonparametric Statistical Methods*. John Wiley & Sons.

Lu, B., Qian, Z., Cunningham, A., and Li., C. (2012). Estimating the Effect of Premarital Cohabitation on Timing of Marital Disruption: Using Propensity Score Matching in Event History Analysis. *Sociological Methods and Research* **41**, 440-466.

Lunceford, J. K. and Davidian M. (2004). Stratification and Weighting via the Propensity Score in Estimation of Causal Treatment Effects: a Comparative Study. *Statistics in Medicine* **23**, 2937-2960.

Mitra, R. and Reiter, J. P. (2011). Estimating Propensity Scores with Missing Covariate Data

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0320

Using General Location Mixture Models. *Statistics in Medicine* **30**, 627-641.

Qu, Y. and Lipkovich, I. (2009). Propensity Score Estimation with Missing Values Using a Multiple Imputation Missingness Pattern (MIMP) Approach. *Statistics in Medicine* **28**, 1402-1414.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org`. ISBN 3-900051-07-0.

Rosenbaum P. R. (1984). Conditional Permutation Tests and the Propensity Score in Observational Studies. *J. Amer. Statist. Assoc.* **79**, 565-574.

Rosenbaum, P. R. (1987). Sensitivity Analysis for Certain Permutation Inferences in Matched Observational Studies. *Biometrika* **74**, 13–26.

Rosenbaum, P. R. (2002). *Observational Studies*. Springer.

Rosenbaum, P. R. and Rubin D. B. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika* **70**, 41–55.

Rosenbaum, P. R. and Rubin D. B. (1984). Reducing Bias in Observational Studies Using Subclassification on the Propensity Score. *J. Amer. Statist. Assoc.* **79**, 516-524.

Rubin, D. B. (1978). Bayesian Inference for Causal Effects: The Role of Randomization. *Annals of Statistics* **6**, 34–58.

Rubin, D. B. (2007). The Design Versus the Analysis of Observational Studies for Causal Effects:

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0320

Parallels with the Design of Randomized Trials. *Statistics in Medicine* **26**, 20–36.

Rubin D. B. (2009). *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons.

Sekhon, J. S. (2011). Multivariate and Propensity Score Matching Software with Automated Balance Optimization: The Matching Package for R. *Journal of Statistical Software*, **42**, 1–52. URL `http://www.jstatsoft.org/v42/i07/.`.

Tumin, R., Ashmead, R. and Sahr, T. (2013). *Types of Insurance Coverage Among Ohios Non-Senior Adult and Child Populations in 2012*. The Ohio Colleges of Medicine Government Resource Center, Technical Report.

Yang, F., Lorch, S., and Small, D. (2014). The Estimation of Causal Effects Using Instrumental Variables with Nonignorable Missing Covariates: Application to Effect of Type of Delivery NICU on Premature Infants. *The Annals of Applied Statistics* **8**, 48-73.

College of Public Health, The Ohio State University, Columbus, OH 43210

E-mail: lu.232@osu.edu

Center for Statistical Research and Methodology, U.S. Census Bureau, Suitland, MD

E-mail: robert.douglas.ashmead@census.gov

Appendix

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0320

A Details of proof of Proposition 2

Proof. Under the null hypothesis of no treatment effect, $D_s = Y_{s1}^0 - Y_{s2}^0$. By the assumption that the treatment assignment is strongly ignorable given $(\mathbf{X}, Z)$, our generalized propensity scores, and our matching procedure,

$$(Y_{s1}^1, Y_{s1}^0) \perp\!\!\!\perp (Y_{s2}^1, Y_{s2}^0) \perp\!\!\!\perp V_s | e_s^*, R_s = 1, \text{and}$$

$$(Y_{s1}^1, Y_{s1}^0) \perp\!\!\!\perp (Y_{s2}^1, Y_{s2}^0) \perp\!\!\!\perp V_s | e_s^*, Z_{s1}, Z_{s2}, R_s = 0.$$

These along with assumption of the null hypothesis imply that

$$D_s \perp\!\!\!\perp V_s | e_s^*, Z_{s1}, Z_{s2}.$$

$$\begin{aligned}
&P(V_s = 1 | R_s, e_s^*, Z_{s1}, Z_{s2}) \\
&= \frac{P(T_{s1} = 1, T_{s2} = 0 | R_s, e_s^*, Z_{s1}, Z_{s2})}{P(T_{s1} + T_{s2} = 1 | R_s, e_s^*, Z_{s1}, Z_{s2})} \\
&= \frac{P(T_{s1} = 1 | R_s, e_s^*, Z_{s1}) P(T_{s2} = 0 | R_s, e_s^*, Z_{s2})}{P(T_{s1} = 1 | R_s, e_s^*, Z_{s1}) P(T_{s2} = 0 | R_s, e_s^*, Z_{s2}) + P(T_{s1} = 0 | R_s, e_s^*, Z_{s1}) P(T_{s2} = 1 | R_s, e_s^*, Z_{s2})} \\
&= \frac{e_{s1}(1 - e_{s2})}{e_{s1}(1 - e_{s2}) + (1 - e_{s1}) e_{s2}}
\end{aligned}$$

If $R = 1$, (3.8) implies that $e_{s1}^* = e_{s1}$ and $e_{s2}^* = e_{s2}$. Also, based on matching $e_{s1}^* = e_{s2}^* = e_s^*$, so then

$$P(V_s = 1 | R_s = 1, e_s^*, Z_{s1}, Z_{s2}) = \frac{e_{s1}(1 - e_{s2})}{e_{s1}(1 - e_{s2}) + (1 - e_{s1}) e_{s2}} = \frac{1}{2}$$

If $R = 0$, because of matching $e_{s1}^* = e_{s2}^* = e_s^*$. However, (3.8) implies that

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0320

$$e_{s1} = logit^{-1}\left(logit(e_s^*) + \gamma Z_{s1}\right) = \frac{exp\left(logit(e_s^*) + \gamma U_{s1}\right)}{1 + exp\left(logit(e_s^*) + \gamma Z_{s1}\right)}.$$

Then we can write $P(V_s = 1|R_s = 0, e_s^*, Z_{s1}, Z_{s2})$ as

$$\begin{aligned}
&P(V_s = 1|R_s = 0, e_s^*, Z_{s1}, Z_{s2}) \\
&= \frac{e_{s1}(1-e_{s2})}{e_{s1}(1-e_{s2}) + (1-e_{s1})e_{s2}} \\
&= \frac{\frac{exp\left(logit(e_s^*)+\gamma Z_{s1}\right)}{1+exp\left(logit(e_s^*)+\gamma Z_{s1}\right)}\left(\frac{1}{1+exp\left(logit(e_s^*)+\gamma Z_{s2}\right)}\right)}{\frac{exp\left(logit(e_s^*)+\gamma Z_{s1}\right)}{1+exp\left(logit(e_s^*)+\gamma Z_{s1}\right)}\left(\frac{1}{1+exp\left(logit(e_s^*)+\gamma Z_{s2}\right)}\right) + \left(\frac{1}{1+exp\left(logit(e_s^*)+\gamma Z_{s1}\right)}\right)\frac{exp\left(logit(e_s^*)+\gamma Z_{s}\right)}{1+exp\left(logit(e_s^*)+\gamma Z_{s2}\right)}} \\
&= \frac{exp\left(logit(e_s^*) + \gamma Z_{s1}\right)}{exp\left(logit(e_s^*) + \gamma Z_{s1}\right) + exp\left(logit(e_s^*) + \gamma Z_{s2}\right)} \\
&= \frac{exp(\gamma Z_{s1})}{exp(\gamma Z_{s1}) + exp(\gamma Z_{s2})}.
\end{aligned}$$

In summary we have that

$$P(V_s = 1|e_s^*, R_s = 1) = 1/2, \text{and} \tag{A.1}$$

$$\begin{aligned}
P(V_s = 1|e_s^*, R_s = 0, Z_{s1} = z_{s1}, Z_{s2} = z_{s2}) &= \frac{exp(\gamma z_{s1})}{exp(\gamma z_{s1}) + exp(\gamma z_{s2})} \times \frac{exp(-\frac{1}{2}\gamma(z_{s1} - z_{s2}))}{exp(-\frac{1}{2}\gamma(z_{s1} - z_{s2}))} \\
&= \frac{exp(\frac{1}{2}\gamma(z_{s1} - z_{s2}))}{exp(\frac{1}{2}\gamma(z_{s1} - z_{s2})) + exp(-\frac{1}{2}\gamma(z_{s1} - z_{s2}))}
\end{aligned} \tag{A.2}$$

Since (A.1) and (A.2) only depend on (e_s^*, z_{s1}, z_{s2}) through $w_s = z_{s1} - z_{s2}$, we can ignore e_s^* in (3.5). From Theorem 1 in Rosenbaum (1984), we note that D is fixed and the only random

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0320

variable is the treatment assignment indicator. Then (3.6) follows from (A.1) and (A.2) because

$$pr\{t(d,V) \geq k|\mathbf{W}=\mathbf{w},\mathbf{R}=\mathbf{r}\} = \sum_{v\in B} I\{t(d,v)\geq k\}|V=v\}pr(V=v|\mathbf{W}=\mathbf{w},\mathbf{R}=\mathbf{r}).$$

□

B Details of proof of Proposition 3

Proof. Since $Z \in [0,1]$, $w_s \in [-1,1]\forall s$. This implies that for all s, $\{\tilde{w}(d)_s - w_s\}d_s \geq 0$ and $\{w_s - \tilde{w}^*(d)_s\}d_s \geq 0$. It follows that for each possible $\mathbf{d}$,

$$\tilde{\mathbf{w}}^*(\mathbf{d}) \lesssim_d \mathbf{w} \lesssim_d \tilde{\mathbf{w}}(\mathbf{d}) \tag{B.1}$$

where $\lesssim_d$ is defined for any S-dimensional vector $\mathbf{d}$ such that $a \lesssim_d b$ if $\{(b_s - a_s)d_s \geq 0\,; s = 1,\ldots,S\}$. Now $t(\mathbf{D},\mathbf{V})$ is isotonic with respect to $\lesssim_d$, meaning that $(\mathbf{V},\mathbf{V}' \in B; \mathbf{V} \lesssim_d \mathbf{V}') \rightarrow \{t(\mathbf{d},\mathbf{V}) \leq t(\mathbf{d},\mathbf{V}')\}$. Then the result follows directly from Lemma 1 from Rosenbaum (1987) and equation (B.1). □