Mann-Whitney Test with Adjustments to Pre-treatment Variables for Missing Values and Observational Study[1]

Song Xi Chen[1], Jing Qin[2] and Cheng Yong Tang[3]

[1] Guanghua School of Management and Center for Statistical Science, Peking University and
Department of Statistics, Iowa State University

[2] National Institute of Allergy and Infectious Diseases, National Institute of Health
Email: jingqin@niaid.nih.gov

[3] Department of Statistics and Applied Probability, National University of Singapore
Email: statc@nus.edu.sg

SUMMARY

The conventional Wilcoxon/Mann-Whitney test can be invalid for comparing treatment effects in the presence of missing values or in observational studies. This is because the missingness of the outcomes or the participation in the treatments may depend on certain pre-treatment variables. We propose an approach to adjust the Mann-Whitney test by correcting the potential bias via consistently estimating the conditional distributions of the outcomes given the pre-treatment variables. We also propose semiparametric extensions of the adjusted Mann-Whitney test which leads to dimension reduction for high dimensional covariate. A novel bootstrap procedure is devised to approximate the null distribution of the test statistics for practical implementations. Results from simulation studies and an economic observational study data analysis are presented to demonstrate the performance of the proposed approach.

Key Words: Dimension reduction; Kernel smoothing; Mann-Whitney statistic; Missing outcomes; Observational studies; Selection bias.

1. Introduction

In statistical, epidemiological and economic literature, the average treatment effect is a widely employed measure to evaluate the impact of a treatment. There has been a recent surge in econometric and epidemiological studies focusing on estimating and comparing treatment effects under various scenarios; see for example Hahn (1998); Korn and Baumrind (1998); Hirano et al. (2003) and Imbens (2004). If the outcome distributions are symmetric, the difference between the average effects is a good measure (Imbens, 2004) for comparison.

[1] Address for Correspondence: Song X Chen, Guanghua School of Management and Center for Statistical Science, Peking University, Beijing 100871, China.
Email: csx@gsm.pku.edu.edu.cn

If, in particular, the outcomes are normally distributed with equal variances, the t-test is preferred for comparing univariate outcomes. However, if the observed outcome distributions are quite away from the normal distributions, nonparametric Wilcoxon/Mann-Whitney tests may be the choice.

Missing values are commonly encountered in survey sampling, medical, social and economic studies; see Rubin (1976) and Little and Rubin (2002) for comprehensive discussions. In particular, the outcome variables can be missing, which may be influenced by a set of covariates. A popular but misguided method is to use only the observed portion of the data. This method might cause the t-test and the Wilcoxon/Mann-Whitney tests to be invalid if the missingness is contributed by certain covariate (pre-treatment variables). A similar issue occurs in observational studies (Rosenbaum, 2002) where the choice of treatment or control on an individual is not purely random and depends on certain pre-treatment variables.

To obtain a valid test for comparing the treatment outcome distributions, one has to adjust for the effect of pre-treatment variables on the missing propensity or on the allocation for treatment and control. Rosenbaum and Rubin (1983) proposed a propensity score matching method which assigns each individual a propensity score calculated from the baseline covariates. By grouping the scores of individuals into intervals, individuals with similar scores are compared. A drawback of this method is the lack of a general guidance on how the groups should be formed. Inverse probability weighting (Korn and Baumrind, 1998; Imbens, 2004) based on the approach of Horvitz and Thompson (1952) is a method that weighs each individual by his or her propensity of treatment or control; see for example Hirano et al. (2003) and Tsiatis (2006). Another popular approach is based on the so-called regression method (Matloff, 1981) by assuming a conditional model for the response variable given the observed covariate, which is efficient provided the underlying model assumption is correct. Nonparametric methods have been also introduced to bring in robustness for the regression approach. Kuk (1993) proposed a marginal distribution estimation by averaging the nonparametrically estimated conditional distribution in finite population sampling problem. Cheng (1994) proposed using the kernel method to estimate the regression function first, followed by averaging the estimated regression function to estimate the mean of the marginal response variable. Cheng's approach is shown by Hahn (1998) to be semiparametric efficient. Generalizations of Cheng (1994)'s method to misclassified binary responses were studied by Chu and Chen (1995). For discrete baseline covariates, Cheung (2005) studied a test for distributional equivalence based on a version of the Wilcoxon statistic. Similar statistic was discussed by Bilker and Wang (1996) for truncated semiparametric models. In the latter two formulations, the baseline covariate information was not fully employed since items associated missing values were not utilized. This can lead to a loss of efficiency since the response variable is correlated with the covariate in general.

In this paper, we propose an adjusted Mann-Whitney test to compare outcome distributions between treatment and control that can accommodate both missing values and observational studies. The adjustment is carried out by a nonparametric kernel estimation to the conditional distributions of the outcomes given the pre-treatment variables. This leads to estimators of the marginal outcome distributions which then produces a Mann-Whitney type statistic. Semiparametric adjustments are also proposed which give rise to a general working model based smoothed Mann-Whitney statistic that reduces the impacts of high dimensional covariate. We show that both approaches are model robust and are able to utilize the data information in the common pre-treatment baseline covariates. The efficiency gains of both proposed adjustments are quantified by reductions in the variances of the test statistics. How to approximate the null distribution of the adjusted Mann-Whitney statistic is a challenge in the conditional setting we face. We propose a novel bootstrap approach which respects the underlying conditional distributions of the outcomes given the pre-treatment covariates while maintaining the null hypothesis.

This paper is organized as follows. The adjusted Mann-Whitney statistic is proposed in Section 2, whose asymptotic distribution is evaluated in Section 3. Semiparametric extensions of the adjusted Mann-Whitney test are discussed in Section 4. Section 5 outlines and justifies the bootstrap resampling approach in approximating the critical values of the adjusted test. Results from simulation experiments are reported in Section 6. An empirical study on a dataset from an economic observational study is presented in Section 7. All the technical details are relegated to the Appendix.

2. **A Covariate Adjusted Mann-Whitney Statistic**

We first introduce the proposed adjusted Mann-Whitney statistic when outcome variables can be missing. Later in this section, we will illustrate how to extend it for observational studies. In a randomized clinical trial, patients are randomly assigned to a treatment arm and or a placebo. For each patient, one can observe a d-variate baseline covariate X. This gives rise to $X_{11}, ..., X_{1n_1}$ for one group and $X_{21}, ..., X_{2n_2}$ for the other. Due to the randomization, X_{1i} $(i = 1, 2, ..., n_1)$ and X_{2j} $(j = 1, 2, ..., n_2)$ have the same marginal distribution F_x. After starting the trial, patients are followed for a period of time, the outcome variables $Y_{11}, ..., Y_{1n_1}$ and $Y_{21}, ..., Y_{2n_2}$ are respectively observed for the two groups. Let F_m be the marginal distribution function of Y_{mi} for $m = 1, 2$, that is assumed to be continuous. We are interested in testing

$$H_0 : F_1 = F_2.$$

If there is no missing value for the outcome variable Y, one may directly compare the Y_{1i}'s $(i = 1, 2, ..., n_1)$ and Y_{2j}'s $(j = 1, 2, ..., n_2)$ distributions to evaluate the treatment effect. Both the t-test and the Wilcoxon/Mann-Whitney tests are popular methods for this purpose. However, often in practice some Ys are missing during the follow-up.

Let $(X_{m1}, S_{m1}, Y_{m1}), ..., (X_{mn_m}, S_{mn_m}, Y_{mn_m})$, for $m = 1$ and 2, be two independent random samples, where the d-variate baseline covariate X_{mi} is always observed, and S_{mi} is a retention indicator such that $S_{mi} = 1$ if Y_{mi} is observed, and 0 otherwise. We assume completely ignorable missing at random (MAR), a notion introduced in Rubin (1976), such that

$$P(S_{mi} = 1|X_{mi}, Y_{mi}) = P(S_{mi} = 1|X_{mi}) = \pi_m(X_{mi}) \tag{1}$$

where $\pi_m(x)$ is the missing (selection) propensity function in the m-th sample. The MAR implies that the conditional distribution of Y_{mi} given X_{mi} and S_{mi} is the same as that of Y_{mi} given X_{mi}, which is denoted by $F_m(y|x)$. If one makes inference based only on the so-called complete data (those with $S_{mi} = 1$), biased results may occur (Breslow, 2003) since the distribution of the complete data may have been distorted away from the truth of the underlying population, which is the case if π_m is not a constant function of the covariate X_i. To avoid this distortion, we propose an approach via estimating $F_m(y|x)$ to filter out the potential bias caused by X_i.

We propose using the kernel method to estimate the conditional distribution function $F_m(y|x)$ based on the m-th sample $\{(X_{mi}, Y_{mi}, S_{mi})\}_{i=1}^{n_m}$. Specifically, let K be a d-dimensional kernel function which is a symmetric (in each dimension) probability density function with finite second moment σ_K^2 in R^d and $K_{h_m}(t) = h_m^{-d}K(t/h_m)$ where h_m is a smoothing bandwidth. The kernel estimator of $F_m(y|x)$ is

$$\hat{F}_m(y|x) = n_m^{-1} \sum_{i=1}^{n_m} \frac{I(Y_{mi} \leq y)K_{h_m}(X_{mi} - x)S_{mi}}{\hat{\eta}_m(x)}, \tag{2}$$

where $\hat{\eta}_m(x) = \frac{1}{n_m} \sum_{i=1}^{n_m} S_{mi}K_{h_m}(X_{mi} - x)$ is a kernel estimator of $\eta_m(x) = \pi_m(x)f_x(x)$ and f_x is the common density of the covariate X_{mi}. As $F_m(y) = \int F_m(y|x)dF_x(x)$, $F_m(y)$ can be estimated by

$$\hat{F}_m(y) = \int \hat{F}_m(y|x)dF_{nx}(x) = \frac{1}{nn_m} \sum_{j=1}^{n} \sum_{i=1}^{n_m} \frac{I(Y_{mi} \leq y)K_{h_m}(X_{mi} - X_j)S_{mi}}{\hat{\eta}_m(X_j)} \tag{3}$$

where F_{nx} is the empirical distribution function based on the pooled covariates $\{X_i\}_{i=1}^n = (X_{11},...X_{1n_1},X_{21},....,X_{2n_2})$ with $n = n_1 + n_2$. The adjusted Mann-Whitney statistic is

$$\begin{aligned} W_n &= \int \hat{F}_1(y)d\hat{F}_2(y) \\ &= \frac{1}{n^2 n_1 n_2}\sum_{l=1}^{n}\sum_{k=1}^{n_2}\sum_{j=1}^{n}\sum_{i=1}^{n_1}\frac{I(Y_{1i} \le Y_{2k})K_{h_1}(X_{1i}-X_j)K_{h_2}(X_{2k}-X_l)S_{1i}S_{2k}}{\hat{\eta}_1(X_j)\hat{\eta}_2(X_l)}. \end{aligned} \quad (4)$$

To reduce the bias of the kernel estimator, one can also adapt the cross-validated estimator so that in obtaining $\hat{\eta}_m(X_j)$, the X_j is not used.

The test statistic W_n in (4) can be readily modified to compare treatment effects in observational studies. Our discussion three paragraphs earlier on potential bias induced by the pre-treatment variables and the need for correction in the context of missing values remains intact. We can understand an observational study as follows. Let (Y_1, Y_0, S, X) be the treatment outcome, control outcome, treatment indicator and baseline covariate, where Y_1 is observed if $S = 1$ but Y_0 is missing, whereas Y_0 is observed if $S = 0$ but Y_1 is missing. Clearly, Y_1 and Y_0 are correlated since they come from the same individual. The basic assumption in casual inference is that $P(S = 1|x, y_1, y_0) = P(S = 1|x)$ i.e. the propensity score only depends on the observable baseline covariate, which is similar to the notion of MAR in (1). Moreover, the conditional densities of Y_1 and Y_0 given covariate X and treatment assignment S satisfy

$$f_1(y_1|x, S) = f_1(y_1|x), \;\; f_0(y_0|x, S) = f_0(y_0|x).$$

namely given covariate X, the treatment or control outcomes does not depend on the choice of treatment or control. Let $F_1(y)$ and $F_2(y)$ be the marginal distributions of Y_1 and Y_0 respectively. Since Y_1 and Y_0 are not available for each individual simultaneously, it is impossible to estimate the joint distribution of (Y_1, Y_0). We get around the problem by adopting our early strategies used in formulating the W_n statistic. Specifically, we treat Y_1's as missing for those individuals who had made the choice of controls ($S = 0$), and similarly Y_0's are regarded as missing for those who had made the choice of "treatment". And the common baseline covariate X is available for each individual. Then, all we need to do is to change the missing indicator S_{mi} to be the indicator for a treatment, and F_1 and F_2 represent the marginal distributions of the two outcome variables.

3. **Properties of the Adjusted Test Statistic**

To analyze the adjusted Mann-Whitney statistic W_n, we apply the projection method (Hoeffding, 1948; Serfling, 1980) to approximate W_n. Let $n = n_1+n_2$ and $\theta = \int F_1(y)dF_2(y)$. Clearly $\theta = 1/2$ under H_0 (no treatment effect). The following conditions are assumed:

C1: The conditional distribution functions $F_m(y|x)$ have continuous second order derivatives with respect to x and y for all (x, y) in their support $\mathcal{S}_{x,y} \subset R^{p+1}$; the density function $f_m(x)$ of the covariate X and the propensity functions $\pi_m(x)$ have continuous second order derivatives for all x in its support $\mathcal{S}_x \subset R^p$, and are both bounded away from zero.

C2: As $\min\{n_1, n_2\} \to \infty$, $n/n_1 \to \rho_1$ and $n/n_1 \to \rho_2$.

C3: The kernel function K is a symmetric probability density function in R^p such that $\int u^2 K(u)du < \infty$ and $\int K^2(u)du < \infty$; the smoothing bandwidth h_m satisfies that $h_m \to 0$, $n_m h_m^d \to \infty$ and $\sqrt{n_m} h_m^2 \to 0$ as $n \to \infty$.

We note that the latter part of Condition C3 prescribes undersmoothing in the kernel estimation for the purpose of bias reduction. It rules out situations where $d \geq 4$, namely X having four or more covariates. For $d \geq 4$, we advocate a semiparametric adjustment that we will propose in Section 4. Let $\bar{F}(y) = 1 - F(y)$ be the survival function and

$$\xi_1(X) = \int \bar{F}_2(y)dF_1(y|X) \text{ and } \xi_2(X) = \int F_1(y)dF_2(y|X)$$

be, respectively, the conditional expectations of $\bar{F}_2(Y_{1i})$ and $F_1(Y_{2k})$. Furthermore, define the conditional variances of $\bar{F}_2(Y_{1i})$ and $F_1(Y_{2k})$

$$v_1^2(X) = \int \bar{F}_2^2(y)dF_1(y|X) - \xi_1^2(X) \text{ and } v_2^2(x) = \int F_1^2(y)dF_2(y|X) - \xi_2^2(X). \tag{5}$$

Let $O_{mi} = (X_{mi}, Y_{mi}, S_{mi})$, $i = 1, \ldots, n_m$, the following lemma provides an approximation to $W_n - \theta$ by projecting it onto the space of all $\{O_{mi}\}_{i=1}^{n_m}$ for $m = 1$ and 2.

Lemma 1. Under Conditions C1-C3, as $\min\{n_1, n_2\} \to \infty$,

$$\begin{aligned} W_n - \theta = {} & n_1^{-1} \sum_{i=1}^{n_1} \left[\frac{S_{1i}}{\pi_1(X_{1i})} \left\{ \bar{F}_2(Y_{1i}) - \xi_1(X_{1i}) \right\} \right] + n_2^{-1} \sum_{k=1}^{n_2} \left[\frac{S_{2k}}{\pi_2(X_{2k})} \left\{ F_1(Y_{2k}) - \xi_2(X_{2k}) \right\} \right] \\ & + n^{-1} \sum_{j=1}^{n} \{\xi_1(X_j) + \xi_2(X_j) - 2\theta\} + o_p(n^{-1/2}). \end{aligned} \tag{6}$$

It can be checked that the first three terms on the right of (6) are mutually un-correlated. The asymptotic normality of W_n is now readily available by applying the central limit theorem and the Slutsky's theorem.

Theorem 1. Under Conditions C1-C3, as $\min\{n_1, n_2\} \to \infty$, $\sqrt{n}(W_n - \theta) \xrightarrow{d} N\{0, v^2(\theta)\}$ where

$$v^2(\theta) = E\left[\rho_1 \pi_1^{-1}(X) v_1^2(X) + \rho_2 \pi_2^{-1}(X) v_2^2(X) + \{\xi_1(X) + \xi_2(X) - 2\theta\}^2 \right]. \tag{7}$$

Remark 1. Despite the covariates can be multivariate such that $d \geq 1$, the kernel smoothing leaves no first order impacts on the asymptotic distribution of W_n. This is due to the averaging in W_n with respect to the pre-treatment covariates as well as the under-smoothing by requiring $\sqrt{n}h_m^2 \to 0$.

Remark 2. Let us consider the classical Mann-Whitney test, in the absence of missing values, defined by

$$W_{0n} = (n_1 n_2)^{-1} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} I(Y_{1i} < Y_{2j}). \tag{8}$$

By carrying out similar projection to that in Lemma 1, we have

$$\begin{aligned} W_{0n} - \theta = {} & n_1^{-1} \sum_{i=1}^{n_1} \{\bar{F}_2(Y_{1i}) - \xi_1(X_{1i})\} + n_2^{-1} \sum_{j=1}^{n_2} \{F_1(Y_{2j}) - \xi_2(X_{2j})\} \\ & + n_1^{-1} \sum_{i=1}^{n_1} \{\xi_1(X_{1i}) - \theta\} + n_2^{-2} \sum_{j=1}^{n_2} \{\xi_2(X_{2j}) - \theta\} + o_p(n^{-1/2}). \end{aligned} \tag{9}$$

Hence the asymptotic variance of W_{0n} is

$$\lim_{n\to\infty} n\text{var}(W_{0n}) = E\left[\rho_1 v_1^2(X) + \rho_2 v_2^2(X) + \rho_1\{\xi_1(X) - \theta\}^2 + \rho_2\{\xi_2(X) - \theta\}^2\right]. \tag{10}$$

As $\rho_1^{-1} + \rho_2^{-1} = 1$, $(\rho_1 - 1)(\rho_2 - 1) = 1$. Thus, in the absence of missing data,

$$\lim_{n\to\infty} n\{\text{var}(W_{0n}) - \text{var}(W_n)\} = E\left[\sqrt{\rho_1 - 1}\{\xi_1(X) - \theta\} - \sqrt{\rho_2 - 1}\{\xi_2(X) - \theta\}\right]^2 \geq 0.$$

This implies that W_n has smaller limiting variance than the classical Mann-Whitney statistic when all observations are complete. This also illustrates the benefit by incorporating data information from the covariates. If X is not informative in the conditional expectations of $\bar{F}_2(Y_1)$ and $F_1(Y_2)$ so that $\xi_1(X) = \xi_2(X) = \theta$, the limiting variances of W_n and W_{0n} are identical. When $\xi_1(X)$ and $\xi_2(X)$ are not constant, W_n can improve on W_{0n}, which demonstrates an advantage of the proposed approach. Our discussion above and elsewhere in the paper is footed on a fact that is if two test statistics are both asymptotically normal with the same asymptotic mean, the test based on the statistic with smaller asymptotic variance is more powerful asymptotically.

Remark 3. We can also compare W_n with

$$Q_n = n_1^{-1} n_2^{-1} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \frac{I(Y_{1i} \leq Y_{2j}) S_{1i} S_{2j}}{\hat{\pi}_1(X_{1i}) \hat{\pi}_2(X_{2j})} \tag{11}$$

which adjusts the Mann-Whitney statistic via the kernel estimated propensity functions $\hat{\pi}_m(x) = \sum_{j=1}^{n_m} K_{h_m}(X_{mi} - x)S_i / \sum_{j=1}^{n_m} K_{h_m}(X_{mi} - x)$. By conducting a similar analysis as that in Remark 2, we can show that

$$\lim_{n\to\infty} \{\mathrm{var}(Q_n) - \mathrm{var}(W_n)\} \geq 0.$$

This confirms again the benefit of incorporating common covariates in W_n. The covariate adjusted Mann-Whitney test by parametrically estimating $\pi_m(x)$ will be discussed in the next section.

An α-level adjusted Mann-Whitney test rejects $H_0 : F_1 = F_2$ if $|\sqrt{n}(W_n - 1/2)| \geq z_{(1-\alpha)/2}\sqrt{v(1/2)}$ where z_α is the α quantile of $N(0,1)$. The power of the test is produced by the fact that $\theta \neq 1/2$ if H_0 is violated and hence $\sqrt{n}(W_n - \frac{1}{2})$ has a mean that diverges to either $+\infty$ or $-\infty$. Let Φ be the distribution function of $N(0,1)$. Then Theorem 1 implies that the asymptotic power of the test is

$$1 - \Phi\left\{\sqrt{n}(\tfrac{1}{2} - \theta)v(\theta)^{-1} + v(\tfrac{1}{2})v^{-1}(\theta)z_{(1-\alpha)/2}\right\} + \Phi\left\{\sqrt{n}(\tfrac{1}{2} - \theta)v(\theta)^{-1} - v(\tfrac{1}{2})v^{-1}(\theta)z_{(1-\alpha)/2}\right\}$$

which converges to 1 as $n \to \infty$ regardless of $\theta < 1/2$ or $\theta > 1/2$. Hence, the test is consistent.

4. Semiparametric Extensions to the Multiple Covariate Situation

In the proposed adjusted Mann-Whitney test, the averaging with respect to X as well as the undersmoothing can alleviate some impacts of the dimensionality of X. However, if the covariates' dimensionality is high, a semiparametric extension of (4) will reduce the impacts of the dimensionality in X and hence improve the performance of the test.

We note that conditioning on $\pi_m(X_m)$, S_m and Y_m are independent (Rosenbaum and Rubin, 1983). Thus, a dimension reduction can be achieved by replacing X_{mi} by univariate $t_{mi} = \pi_m(X_{mi})$ in the formulation of the adjusted Mann-Whitney test statistic W_n. Suppose that it is reasonable to assume parametric models $\pi_m(x;\beta_m)$, for instance the logistic models, for the propensity functions, where β_m are unknown parameters. Let $\hat{\beta}_m$ be the maximum likelihood estimates (MLE) based on the binary log-likelihood

$$\ell_m(\beta_m) = \sum_{i=1}^{n_m}\Big[S_{mi}\log\{\pi_m(X_{mi},\beta_m)\} + (1 - S_{mi})\log\{1 - \pi_m(X_{mi},\beta_m)\}\Big]. \tag{12}$$

Let $\hat{t}_{mi} = \pi_m(X_i;\hat{\beta}_m)$ for pooled covariates $\{X_i\}_{i=1}^n$ and $m = 1,2$ respectively, then a semiparametric version of the adjusted Mann-Whitney statistic (4) is

$$T_n = \frac{1}{n^2 n_1 n_2}\sum_{l=1}^{n}\sum_{k=1}^{n_2}\sum_{j=1}^{n}\sum_{i=1}^{n_1}\frac{I(Y_{1i} \leq Y_{2k})K_{h_1}(\hat{t}_{1i} - \hat{t}_{1j})K_{h_2}(\hat{t}_{2k} - \hat{t}_{2l})S_{1i}S_{2k}}{\hat{\eta}_1(\hat{t}_{1j})\hat{\eta}_2(\hat{t}_{2l})} \tag{13}$$

where $\hat{\eta}_m(t) = n_m^{-1}\sum_{i=1}^{n_m} S_{mi}K_{h_m}(t_{mi} - t)$ is a kernel estimator of $\pi_m(x)f_m(t)$, $f_m(t)$ is the density of the transformed random variable $t = \pi_m(X)$ and K is now a univariate kernel function with bandwidth h_m. We assume in this section the following condition:

C4: The missing propensity function takes the parametric form $P(S_{mi} = 1|X_{mi}, Y_{mi}) = \pi_m(X_{mi}, \beta_m)$ that is bounded away from 0 and twice continuously differentiable in β_m.

We note that theory for maximum likelihood estimate implies that $\hat{\beta}_m$ by maximizing (12) is $\sqrt{n}$-consistent (Newey and McFadden, 1994). We define two projections of the second sample survival function and the first sample distribution function with respect to the conditional distributions given the propensity functions as

$$\psi_1(X) = \int \bar{F}_2(y)dF_1\{y|\pi_1(X;\beta_{01})\} \text{ and } \psi_2(X) = \int F_1(y)dF_2\{y|\pi_2(X;\beta_{02})\}. \tag{14}$$

The first order approximation of T_n is presented in the following lemma, which resembles that in Lemma 1 for W_n.

Lemma 2. Under Conditions C1 - C4, as $\min\{n_1, n_2\} \to \infty$,

$$\begin{aligned} T_n - \theta = n_1^{-1}\sum_{i=1}^{n_1}\left[\frac{S_{1i}}{\pi_1(X_{1i})}\left\{\bar{F}_2(Y_{1i}) - \psi_1(X_{1i})\right\}\right] + n_2^{-1}\sum_{k=1}^{n_2}\left[\frac{S_{2k}}{\pi_2(X_{2k})}\left\{F_1(Y_{2k}) - \psi_2(X_{2k})\right\}\right] \\ + n^{-1}\sum_{j=1}^{n}\{\psi_1(X_j) + \psi_2(X_j) - 2\theta\} + o_p(n^{-1/2}). \end{aligned} \tag{15}$$

Since S_{mi} and Y_{mi} are conditionally independent given $\pi_m(X_{mi})$ (Rosenbaum and Rubin, 1983), all terms in (15) are uncorrelated. Define

$$u_1^2(X) = \int \bar{F}_2^2(y)dF_1(y|X) - \psi_1^2(X) \text{ and } u_2^2(X) = \int F_1^2(y)dF_2(y|X) - \psi_2^2(X).$$

Let

$$v_p^2(\theta) = E\left[\rho_1\pi_1^{-1}(X)u_1^2(X) + \rho_2\pi_2^{-1}(X)u_2^2(X) + \{\psi_1(X) + \psi_2(X) - 2\theta\}^2\right]. \tag{16}$$

The following theorem provides the asymptotic normality of T_n.

Theorem 2. Under Conditions C1-C4, as $\min\{n_1, n_2\} \to \infty$, $\sqrt{n}(T_n - \theta) \xrightarrow{d} N\{0, v_g^2(\theta)\}$.

We can compare T_n given by (13) to propensity adjusted Mann-Whitney test statistic

$$R_n = n_1^{-1}n_2^{-1}\sum_{i=1}^{n_1}\sum_{j=1}^{n_1}\frac{I(Y_{1i} \le Y_{2j})S_{1i}S_{2j}}{\hat{t}_{1i}\hat{t}_{2j}}. \tag{17}$$

Though both T_n and R_n utilize estimated propensity functions from parametric models, they differ substantially in utilizing information in the base-line covariates X_i. Like W_n,

T_n is more active in pursuing such information, whileas R_n is more passive, only through the missing propensities. A variance comparison between R_n and T_n is not attainble in the presence of missing values, sine R_n has a leading variance contribution from the estimated parameters in the propensity function (Cheung, 2005) while T_n does not have such terms due to the kernel smoothing. However, a comment can be made in the absence of missing values, where R_n is equivalent to W_{0n}. By replicating Remark 1 to Theorem 1, we can show that T_n is more efficient than W_{0n} (R_n when there is no missing data), indicating the benefit by incorporating common baseline covariate information. The simulation studies reported in Section 6 contain numerical comparisons between the two tests, which lend support to this view.

The semiparametric statistic T_n in (13) is attractive in reduced covariate dimensionality, and hence it overcomes the difficulty of W_n with the dimensionality of X in the presence of missing values or observational studies. However we note that $v_p^2(\theta) \geq v^2(\theta)$ by comparing (16) and (7), because $\xi_1(X)$ and $\xi_2(X)$ are the minimum variance unbiased predictors of $\bar{F}_2(Y_1)$ and $F_1(Y_2)$ conditioning on X respectively. This illustrates the connection between the nonparametric adjusted test (4) and the semiparametric extension (13). Having said these, we would like to voice caution. It should be emphasized that the result is asymptotic, for n being sufficiently abundant so that the dimensionality is not an issue for the fully nonparametric W_n. As we will demonstrate in the simulation study, in finite sample situation, the dimensionality is an issue for the fully nonparametric test based on W_n.

The above discussion suggests room for improving the propensity function based semiparametric extension (13). How to obtain a better projection than $\psi_m(x)$ in (14) motivates us to consider working towards $F_m(y|g_m(X;\gamma_m))$ for a proper general index function $g_m(X;\gamma_m)$ with parameter γ_m. The index function can be a working regression model postulated on the complete data. We provide examples of such working models in the simulation section. We note that the idea here is related to approximating conditional distribution function by dimensional reduction considered in Hall and Yao (2005); see also Hu et al. (2010) and Hu et al. (2011) for dimensional reduction via the kernel smoothing for inference on the mean and the distribution function with missing data incorporating parametric models for the propensities. The parameters γ_m in the index functions can be estimated based on the m-th sample via several methods, including the maximum likelihood estimation (Newey and McFadden, 1994), general methods of moments (Hansen, 1982), and the minimum distance approach in Hall and Yao (2005).

This leads us to a new semiparametric test statistic by incorporating $\hat{z}_{mi} = g_m(X_i;\hat{\gamma}_m)$ for pooled covariates $\{X_i\}_{i=1}^n$. Because Y_m and S_m are not conditionally independent given

$g_m(X;\gamma_m)$, it is necessary to modify the definition (2) with propensity weighting. Let $\hat{\pi}_{mi} = \pi_m(X_{mi}, \hat{\beta}_m)$, then

$$\hat{F}_m(y|z) = \hat{f}_m^{-1}(z)\left\{n_m^{-1}\sum_{i=1}^{n_m} I(Y_{mi} \le y)K_{h_m}(z - z_{mi})\frac{S_{mi}}{\hat{\pi}_{mi}}\right\} \tag{18}$$

estimates the conditional distribution $F_m(y|z)$ consistently, where $\hat{f}_m(z) = n_m^{-1}\sum_{i=1}^{n} K(z - z_{mi})S_{mi}/\hat{\pi}_{mi}$. Then following the same steps constructing W_n, we define

$$Z_n = n^{-2}n_1^{-1}n_2^{-1}\sum_{l=1}^{n}\sum_{k=1}^{n_2}\sum_{j=1}^{n}\sum_{i=1}^{n_1}\frac{I(Y_{1i} \le Y_{2k})K_{h_1}(z_{1i} - z_{1j})K_{h_2}(z_{2k} - z_{2l})S_{1i}S_{2k}}{\hat{f}_1(z_{1j})\hat{f}_2(z_{2l})\hat{\pi}_{1i}\hat{\pi}_{2k}}. \tag{19}$$

We assume the following additional condition for the general semiparametric extension using.

C5: There exist limits γ_{01} and γ_{02} such that the estimator $\hat{\gamma}_m$ based on the m-th sample is $\sqrt{n}$-consistent to γ_{0m}. And $g_m(x;\gamma_m)$, $m = 1, 2$, is continuously twice differentiable in γ_m with bounded first partial derivative in a neighborhood of γ_{0m}.

We note that the $\sqrt{n}$-consistency of $\hat{\gamma}_m$ in Condition C4 is a mild requirement that is satisfied by a range of estimation approaches including the maximum likelihood (Newey and McFadden, 1994), general methods of moments (Hansen, 1982), and the minimum distance approach in Hall and Yao (2005). Denote the conditional expectations analogous to (14) by

$$\phi_1(X) = \int \bar{F}_2(y)dF_1\{y|g_1(X;\gamma_{01})\} \text{ and } \phi_2(X) = \int F_1(y)dF_2\{y|g_2(X;\gamma_{02})\}.$$

The first order approximation of Z_n is presented in the following lemma.

Lemma 3. Under Conditions C1 - C5,

$$\begin{aligned} Z_n - \theta = {} & n_1^{-1}\sum_{i=1}^{n_1}\left[\frac{S_{1i}}{\pi_1(X_{1i})}\left\{\bar{F}_2(Y_{1i}) - \phi_1(X_{1i})\right\}\right] + n_2^{-1}\sum_{k=1}^{n_2}\left[\frac{S_{2k}}{\pi_2(X_{2k})}\left\{F_1(Y_{2k}) - \phi_2(X_{2k})\right\}\right] \\ & + n^{-1}\sum_{j=1}^{n}\{\phi_1(X_j) + \phi_2(X_j) - 2\theta\} + o_p(n^{-1/2}). \end{aligned} \tag{20}$$

If S_{mi} and Y_{mi} are conditionally independent given $g_m(X_{mi})$, then all terms in (20) are uncorrelated. The conditional independence holds when $g_m(X)$ is chosen to be the propensity function (Rosenbaum and Rubin, 1983); otherwise correlations among terms in (20) generally exist. Define

$$w_1^2(X) = v_1^2(X) + \{\xi_1(X) - \phi_1(X)\}^2 \text{ and } w_2^2(X) = v_2^2(X) + \{\xi_2(X) - \phi_2(X)\}^2$$

where $v_1^2(X)$ and $v_2^2(X)$ are given in (5). Let

$$\begin{aligned} v_g^2(\theta) = {} & E\left[\rho_1\pi_1^{-1}(X)w_1^2(X) + \rho_2\pi_2^{-1}(X)w_2^2(X) + \{\phi_1(X) + \phi_2(X) - 2\theta\}^2\right] \\ & + 2E\left[\{\xi_1(X) + \xi_2(X) - \phi_1(X) - \phi_2(X)\}\{\phi_1(X) + \phi_2(X) - 2\theta\}\right]. \end{aligned} \tag{21}$$

The following theorem provides the asymptotic normality of Z_n.

Theorem 3. Under the conditions C1-C5, as $\min\{n_1, n_2\} \to \infty$ $\sqrt{n}(Z_n - \theta) \xrightarrow{d} N\{0, v_g^2(\theta)\}$.

Comparing the variances $v_g^2(\theta)$ and $v^2(\theta)$ given by (21) and (7) respectively, we find that

$$v_g^2(\theta) - v^2(\theta) \geq E\left[\sqrt{\rho_1 - 1}\{\xi_1(X) - \phi_1(X)\} - \sqrt{\rho_2 - 1}\{\xi_2(X) - \phi_2(X)\}\right]^2 \geq 0. \qquad (22)$$

When $g_m(X; \gamma_m)$ is appropriately chosen such that $\xi_m(X) = \phi_m(X)$ for $m = 1, 2$, then the variance of Z_n is minimized. Intuitively it makes sense, because in such case $g_m(x; \gamma_m)$ achieves the minimum variance unbiased predictors of $\bar{F}_2(Y_1)$ and $F_1(Y_2)$ conditioning on X respectively. This observation can also be suggested by the fact that the more $g_m(X)$ is relevant to the conditional distribution of $F_m(Y)$ given X, the more improvement in the variance of Z_n can be achieved. Again, we stress that the variance comparison is only valid asymptotically and Z_n is more advantageous in practice. Therefore, Lemma 3 and Theorem 3 illustrate that Z_n successfully combines the merits of W_n and T_n in efficiency and convenience for multivariate covariates.

5. Bootstrap Calibration

To implement the proposed adjusted Mann-Whitney test based on W_n, we need to approximate the distributions of W_n, T_n and Z_n under $H_0 : F_1 = F_2$. We will only present the bootstrap for W_n and that for T_n or Z_n is available by replacing all the conditioning variables X to $\pi_m(X)$ or $g_m(X)$. One approach is to estimate the asymptotic variance $v^2(1/2)$ under H_0. However, $v^2(1/2)$ as implied from (7) involves many unknown functions including the missing propensities $\pi_1(x)$ and $\pi_2(x)$, the marginal distributions F_1 and F_2, the common density f of the covariates as well as the conditional distributions $F_1(y|x)$ and $F_2(y|x)$. This makes any direct plugging-in estimation of $v^2(1/2)$ rather involved and is prone to error.

We consider a bootstrap approximation to the null distribution of W_n. The challenge for the bootstrap in the current context is how to generate resamples (X_i^*, Y_i^*, S_i^*) which meet two requirements:

(i) the resampled outcomes Y^* under the treatment and control have the same marginal distribution to satisfy H_0;

(ii) the underlying conditional distributions $F_1(y|x)$ and $F_2(y|x)$, the distribution of the covariate X and the missing propensities are respected by the resamples.

A seemingly straightforward solution was to pool two samples together and then to draw resamples with replacement from the combined sample randomly as some conventional bootstrap approaches do. While this creates a scenario of the null hypothesis, it may fail to respect the conditional distributions $F_m(y|x)$ and the missing propensities $\pi_m(x)$ respectively.

Recall that $\hat{F}_1$ and $\hat{F}_2$ are estimators to the distributions of the outcome variables F_1 and F_2 given by (3), and let

$$\hat{G}(y) = n^{-1}\left\{n_1\hat{F}_1(y) + n_2\hat{F}_2(y)\right\}. \tag{23}$$

The proposed bootstrap procedure consists of the following steps:

1. Obtain $(X^*_{mi}, S^*_{mi}, Y^*_{mi})$ by sampling with replacement in the original sample m for $m = 1, 2$ and $i = 1, 2, \ldots, n_m$ respectively.

2. Let $U_{mi} = \hat{F}_m(Y^*_{mi})$ and replace Y^*_{mi} by $\tilde{Y}^*_{mi} = \hat{G}^{-1}(U_{mi})$ where the inverse function is defined by $\hat{G}^{-1}(u) = \sup\{y : \hat{G}(y) \leq u\}$.

3. Calculate W^*_n by (4) based on $\{(X^*_{mi}, S^*_{mi}, \tilde{Y}^*_{mi})\}_{i=1}^{n_m}$ for $m = 1$ and 2.

4. Repeat Steps 1-3 B times for a large integer B, obtain the test statistics based on the resamples, and order them such that $W^*_{n1} \leq W^*_{n2} \leq \ldots W^*_{nB}$.

Step 1 draws resamples with replacement from the two original samples respectively. This maintains the joint distributions of (X, Y, S) and hence the conditional distributions and the missing mechanisms in the original samples. This step maintains the underlying conditional distributions $F_m(y|x)$, but F_1 and F_2 may be different. Step 2 replaces the response variable by inverting the estimated marginal distribution of Y based on the pooled sample, which results in $\tilde{Y}^*_{mi}$ having the same marginal distribution, and hence having H_0 maintained. The latter is explicitly outlined in Appendix A.4.

Let $c_{\alpha/2} = W^*_{n[B\alpha/2+1]}$ and $c_{1-\alpha/2} = W^*_{n[B(1-\alpha/2)+1]}$ be, respectively, the $\alpha/2$ and $1-\alpha/2$ level empirical quantiles of the resampled test statistics $\{W^*_{nb}\}_{b=1}^{B}$. The proposed bootstrap test rejects H_0 if $W_n \notin (c_{\alpha/2}, c_{1-\alpha/2})$. Let $\mathcal{F}_n$ be the σ-field generated by $\{(X_{mi}, S_{mi}, Y_{mi})\}_{i=1}^{n_{mi}}$ for $m = 1, 2$. A justification to the bootstrap calibration is provided in the following theorem whose proof is given in the Appendix.

Theorem 4. Under Conditions C1-C3 and H_0, the conditional distribution of $\sqrt{n}(W^*_n - 1/2)/v(1/2)$ given $\mathcal{F}_n$ converges in distribution to $N(0,1)$ almost surely, as $\min\{n_1, n_2\} \to \infty$.

Theorem 4 confirms the validity of the bootstrap procedure in approximating the limiting distribution of the test statistic. A similar bootstrap procedure can be applied to the semiparametric extensions of the proposed approach to obtain the critical values for implementing the tests.

6. Simulation Studies

We conducted extensive simulations to demonstrate the merits of the proposed adjusted Mann-Whitney test and its semiparametric extensions. The simulations evaluated the performance of the nonparametrically adjusted Mann-Whitney test based on W_n, the semiparametrically adjusted tests T_n and Z_n with an index function linear in all covariates. When implementing T_n, parameters in the propensity functions were estimated by maximizing binary likelihood functions. For parameters in the working linear function in Z_n, least squares estimates were obtained by minimizing $\sum_{i=1}^{n_1} S_{1i}\{1-\tilde{F}_2(Y_{1i})-X_{1i}^T\gamma_1-\gamma_{0,1}\}^2$ and $\sum_{i=1}^{n_2} S_{2i}\{\tilde{F}_1(Y_{2i})-X_{2i}^T\gamma_2-\gamma_{0,2}\}^2$ respectively for unknown parameters $\gamma_{0,m}$ and $\gamma_m = (\gamma_{1,m},\ldots\gamma_{d,m})^T$, $m=1,2$. Those initial estimates $\tilde{F}_m(y)$ in the least squares were obtained by weighted empirical distributions $\tilde{F}_m(y) = n_m^{-1}\sum_{i=1}^{n_m} I(Y_{mi}\le y)S_{mi}/\pi_m(X_{mi};\hat{\beta}_m)$.

We compared the proposed adjusted tests with two testing procedures in missing data problems. One is based on the propensity weighted Mann-Whitney statistic R_n in (17), which is an extension of a method in Cheung (2005). The other is based on the adjusted mean comparison:

$$\tilde{t}_n = \sqrt{n}|\hat{\mu}_1-\hat{\mu}_2| \tag{24}$$

where $\mu_m^{-1} = n_m^{-1}\sum_{i=1}^{n_m} Y_{mi}/\hat{\pi}_m(X_{mi};\beta_m)$ is the propensity adjusted estimation for the mean of Y. Clearly, (24) is an extension of the t-test for missing data with covariates. We chose the correctly specified parametric model for the missing propensity function for (17) and (24) so that they would perform under the most ideal conditions. We also obtained results for two impractical Oracle tests: the classical Mann-Whitney and the two sample t tests by accessing to the missing values in Y to gain benchmarks for power of the tests.

A d-variate product kernel was employed throughout the simulation when implementing the proposed fully nonparametric test statistics W_n in (4); and a univariate kernel was used for the semiparametric statistics T_n and Z_n in (13) and (19). The Gaussian kernel was chosen as the univariate kernel and was used to generate a d-dimensional product kernel. The bandwidths were chosen by the cross-validation method (Hall et al., 2004) then divided by 2 for undersmoothing. To evaluate the robustness of the test against the choices of bandwidths, we evaluated the tests statistics at two additional bandwidths, being 10% larger or smaller. The results were largely similar and hence are not reported here. In the data generating process, we set $n_1 = n_2 = n$ without loss of generality, and in particular, $n = 50$ and 80 respectively. The covariates $X_{mi} = (X_{mi,1},\ldots,X_{mi,d})^T$ were a d-dimensional random vector for $m = 1,2$. We assigned d from 1 to 4 to examine impacts of covariates' dimension. In all simulations the number of replications was 1000 and the bootstrap was repeated for 100 time to obtain the critical values.

We first experimented Gaussian distribution for the outcome variable Ys. In this experiment, we generated $(Y_{mi}, X_{mi}^T)^T \sim N(\mu_m, \Sigma)$ independently for $m = 1, 2$, where $\Sigma = (\sigma_{ij})_{i,j=1,\dots,d+1}$. Here $\sigma_{ii} = s^2$ for $i = 1, \dots, d+1$, $\sigma_{ij} = 0.6s^2$ for $i \neq j$, and $s = 0.3$. When assessing the sizes of tests, both μ_1 and μ_2 were set to be 0; and when assessing the power, μ_1 was made zero and μ_2 was set to be 0.1 and 0.15 respectively generating two scenarios, Power$_1$ and Power$_2$ respectively. We then varied the sample size n in combination with different missing data models. We considered two cases of missing at random: 1) the propensities functions were the same for both samples (MAR1); and 2) two different propensity functions in the two samples (MAR2). In both mechanisms, the propensity functions assumed a parametric form

$$P(S = 1|Y, X) = \pi(X) = \theta_0 + \theta_1 X_1 + \cdots + \theta_d X_d + \theta_{d+1} X_1^2$$

with parameter $\theta = (\theta_0, \theta_1, \dots, \theta_{d+1})$. For MAR1, θ was set to be $(1.25, 1/\sqrt{d}, \dots, 1/\sqrt{d}, -3.0)^T$ for both samples. For MAR2, θ was set $(-0.5, 1/\sqrt{d}, \dots, 1/\sqrt{d}, 0)^T$ for one sample and for the other the same as MAR1. We note that dividing $\sqrt{d}$ in the parameter values assignments was to ensure that the missing propensities were at a similar level in average with respect to different dimensions to allow comparable results across d. In average about 25% responses under MAR1 were missed, while those for MAR2 were about 60% and 25% respectively.

To gain further empirical evidence, we experimented another simulation design where the responses between the two samples had different distributions, unlike the previous setting (Gaussian setting) where both were Gaussian distributed. Under the design, both the covariate and the response in the first sample were kept the same as in the previous Gaussian cases but with $\mu = 0$; and the missing values were governed by the MAR1 and MAR2 respectively. In the second sample, the distribution of X_{mi} and the missing propensity were identical to the first sample, but Y followed a centralized Gamma distribution with the shape parameter $\alpha = 2.0$ and the scale parameter $\beta = 1.6$. This was attained by a $d+1$ dimensional Gaussian copula such that

$$P(Y < y, X_1 < x_1, \dots, X_d < x_d) = \Phi_{d+1}\{\Phi^{-1}(u_1), \dots, \Phi^{-1}(u_{d+1}); \Sigma\},$$

where $u_1 = P(Y < y)$, $u_2 = P(X_1 < x_1), \dots, u_{d+1} = P(X_d < x_d)$, $\Phi_{d+1}(x_1, \dots, x_{d+1}; \Sigma)$ was the same $d+1$ dimensional normal distribution used in the Gaussian setting with the same covariance Σ used there.

Table 1 reports the mean and standard deviation of the test statistics under H_0 for the Gaussian data. The empirical size and power for the tests with 5% nominal significance are provided in Tables 2 and 3, respectively alone with the impractical Mann-Whiteney test and the two sample t-test by obtaining inside information on the missed Ys. Table 4 contains the empirical power for the setting where the outcomes were Gaussian and Gamma distributed respectively. We observe from Table 1 that there was a clear effect of the dimensionality on W_n with the mean deviating from 1/2 more and the variance increased as d was increased, which was also the case for the variance of the propensity weighted test staitistic R_n. The variance of R_n was consistently larger than that of W_n, T_n and Z_n. This foreshadowed a different test performance between the proposed and the propensity weighted tests. In contrast, the variance of the semiparametric T_n and Z_n were not sensitive to d, indicating the practical merits of the semiparametric extensions.

Table 2 indicates that all the tests considered had reasonable empirical size, which was especially the case for the two semiparametric tests. The slightly larger size distortion for the test based on W_n under $d \geq 3$ reflected the larger standard deviation in the mean from 1/2 as reported in Table 1. A deeper reason was the curse of dimension as Condition C3 was not met for $d = 4$ and just barely for $d = 3$, which was the motivation for proposing the semiparametric adjustments T_n and Z_n. The performance of the semiparametric adjusted Mann-Whitney tests was very encouraging for both the size and power, and across different dimensions. We observe from Table 3 that the proposed nonparametric and semiparametric tests were much more powerful than the direct propensity adjusted Mann-Whitney test based on R_n and the covariate adjusted t-test for almost all the Gaussian simulation settings where the covariates and outcomes were all Gaussian, despite the settings were not that favorable to the proposed Mann-Whitney tests. Table 4 shows that, when the two outcome distributions were different, the powers of the proposed tests based on W_n, T_n and Z_n were much better than those of the tests based on R_n and $\tilde{t}_n$. As expected, both the adjusted t-test and the impractical Oracle t-test broke down completely. Both Tables 3 and 4 show that the semiparametric test based on Z_n (with the working linear function) was consistently more powerful than that of the test based on T_n using propensity function. And both semiparametric tests were consistently better than the tests based on R_n and $\tilde{t}_n$. Both Tables 3 and 4 also reveal that the powers from proposed nonparametric and semiparametric tests were quite reasonable in comparison to the power of the Oracle Mann-Whitney test based on W_{0n}.

7. A Data Analysis

In this section, we apply the proposed tests to a data set obtained in an economic observational study, which allowed us to demonstrate how to apply the proposed tests for observational studies. The original data were considered in Lalonde (1986). We use a subset of the original data considered in Dehejia and Wahba (1999), Imbens (2004) and Qin et al.

(2008). Dehejia and Wahba (1999) considered propensity score match for comparison of two means, and Imbens (2004) conducted inverse probability weighting for the mean difference. The datasets `NSWRE74_CONTROL.TXT` and `NSWRE74_TREATED.TXT` can be obtained at `http://www.nber.org/~rdehejia/nswdata.html`. The dataset contains 445 individuals, 185 of whom participated in a training program and 260 did not. We are interested in the effect of the training program on earning in 1978. The covariates available for both groups (trained and not trained) include age, years of education, indicators of African-American and Hispanic-American, marital and degree statuses, and earnings in 1975. A comparison of the mean earnings of the two groups was considered in Qin et al. (2008). We consider here testing for the equality of the earning distributions. As advocated at the end of Section 2, in the formulation of the adjusted Mann-Whitney statistic W_n, we assign $S = 1$ for all the 185 individuals participated in the training program regarding them as "respondents" while assigning $S = 0$ for the rest of 260 individuals regarding them as "non-respondents" (missing outcomes). Similarly, in the second sample we treat the observations from 260 individuals not participated in the training program as "respondents" with $S = 1$ while regarding the other 185 individuals as "non-respondents" with $S = 0$.

Figure 1 displays the histograms of the earnings in 1978 for the trained and control groups, which conveys that both groups have a significant portion of members whose earnings were 0. The percentages of zero earnings were 35.4% and 24.3% in the control and trained groups respectively, which constitutes a quite sharp difference between the two groups. A direct application of the naive Mann-Whitney statistic, that ignored the pre-treatment covariates, on the earnings gave a p-value 0.011 and thus concluded a significant difference in the distributions of the earnings between the two groups. However, conditioning on earnings greater than zero, the distributions seem to be close to each other in Figure 1. This is confirmed by an application of the Mann-Whitney statistic on those with earnings greater zero, which gave a p-value of 0.374. In other words, the latter test could not reject the hypothesis that the distributions were the same for those with earnings greater than zero. However, both tests failed to reflect the observational nature of the data. In addition, we also observe from Figure 1 that the distributions of the earnings in 1978 are clearly not symmetric, indicating that t test may be less powerful in this case.

To gain more insights on the dataset and to reconcile the conflicting testing results mentioned above, we first apply the kernel estimator (2) to estimate the earning distributions F_1 and F_2 in 1978, adjusted with respect to covariate effect and missing values. The kernel estimates are plotted in Figure 2, where the line between the two estimated distributions of the two groups is the pooled estimator in (23). From the estimated CDF, we can see that almost all quantiles of the trained group are consistently larger than those of the control group. Then we apply the proposed test statistics with adjustments to the covariate in comparing the earning distributions. We assume that individuals participated in the training program with a propensity function that depends on covariates. We use the product Gaussian kernel for smoothing the continuous covariates: the age, years of education and earnings in 1975. The bandwidths were chosen by the same approach as in the simulation study. The proposed bootstrap procedure was implemented to obtain the critical value of the test statistic with $B = 100$. The resulting test statistic $W_n = 0.397$, which was less than the second smallest value, but greater than the smallest one, of the bootstrapped statistics. Hence the p-value was between 0.02 and 0.04 for a two sided test.

To apply the semiparametric test based on (13), we use the logistic model for the propensities of both groups. All covariates were included in the model with an additional quadratic term of age, which is suggested in Dehejia and Wahba (1999). Then we apply (13) to obtain T_n using the estimated propensity function. The bandwidth was chosen by cross-validation and then divided by 2. The same bootstrap procedure was applied to calculate the critical value for T_n. The resulting test statistic $T_n = 0.401$ and the p-value was between 0.06 and 0.08. We then apply the working linear function approach using the same set of covariates as in the propensity function to get Z_n, and get the test statistic $Z_n = 0.391$ and p-value between 0.02 and 0.04. We find that the conclusions of the proposed tests are largely consistent with each other. Comparing the p-values of the proposed tests to that of the Mann-Whitney test that ignored the pre-treatment covariates, we observe substantial differences which clearly indicates the impact of the adjustment. This suggested that an adjustment to the covariate effect is important for analyzing data from observational studies.

Acknowledgements

We are very grateful to the Editor, AE, and two referees for their insightful comments and constructive suggestions that have greatly improved this paper. Chen acknowledges support from Center for Statistical Science at Peking University, and Tang acknowledges research support from National University of Singapore Academic research grants.

Appendix: Technical Details

A.1 Proof of Lemma 1

We start with an expansion for the Mann-Whitney statistic W_n which is used in proving Theorems 1 and 2. The subscript m in all following expressions takes value 1 and 2, indicating the first and second sample. Results from kernel regression and density estimation (Härdle, 1990; Fan and Gijbels, 1996) indicate that $E\{\hat{\eta}_m(x)\} = \eta_m(x) + O(h_m^2)$, where h_m is the bandwidth used in $\hat{\eta}_m(x)$. We recall by its definition $\eta_m(x) = \pi_m(x) f(x)$. Applying Taylor's expansion, we have

$$1/\hat{\eta}_m(x) = 1/\eta_m(x) - 1/\eta^2(x)\{\hat{\eta}_m(x) - \eta_m(x)\} + o_p(n^{-1/2}). \tag{25}$$

Define $\pi_{mi} = \pi_m(X_{mi})$, $\alpha_{1ik} = \pi_{1i}^{-1}\pi_{2k}^{-1}$,

$$\begin{aligned}
\alpha_{2ik} &= \pi_{2k}^{-1}\left\{n^{-1}\sum_{j=1}^{n} K_{h_1}(X_{1i} - X_j)\eta_1^{-1}(X_j) - \pi_{1i}^{-1}\right\},\\
\alpha_{3ik} &= \pi_{1i}^{-1}\left\{n^{-1}\sum_{l=1}^{n} K_{h_2}(X_{2k} - X_l)\eta_2^{-1}(X_l) - \pi_{2k}^{-1}\right\},\\
\alpha_{4ik} &= \pi_{2k}^{-1}\left\{n^{-1}\sum_{j=1}^{n} K_{h_1}(X_{1i} - X_j)\{\hat{\eta}_1(X_j) - \eta_1(X_j)\}\eta_1^{-2}(X_j)\right\} \text{ and}\\
\alpha_{5ik} &= \pi_{1i}^{-1}\left\{n^{-1}\sum_{l=1}^{n} K_{h_2}(X_{2k} - X_l)\{\hat{\eta}_2(X_l) - \eta_2(X_l)\}\eta_2^{-2}(X_l)\right\}.
\end{aligned}$$

Then, let $V_{ik} = I(Y_{1i} \le Y_{2k})S_{1i}S_{2k}$, we have by substituting (25) into W_n defined by (4),

$$\begin{aligned}
W_n &= n_1^{-1}n_2^{-1}\sum_{i=1}^{n_1}\sum_{k=1}^{n_2} V_{ik}\left\{n^{-1}\sum_{j=1}^{n} K_{h_1}(X_{1i} - X_j)\hat{\eta}_1^{-1}(X_j)\right\}\left\{n^{-1}\sum_{l=1}^{n} K_{h_2}(X_{2k} - X_l)\hat{\eta}_2^{-1}(X_j)\right\}\\
&= W_{n1} + W_{n2} + W_{n3} - W_{n4} - W_{n5} + o_p(n^{-1/2}) \qquad (26)
\end{aligned}$$

where $W_{na} = n_1^{-1}n_2^{-1}\sum_{i=1}^{n_1}\sum_{k=1}^{n_2} V_{ik}\alpha_{aik}$ for $a = 1, \ldots, 5$. Here we note that the second equation is just a re-organization of the terms as two-sample U- or V- statistics, and the $o_p(n^{-1/2})$ term in (26) is from the approximation (25).

We note that W_{n1} is a two-sample U-statistic, while $W_{n2}, \ldots, W_{n5}$ are all related to two-sample V-statistics (Serfling, 1980) after symmetrizing the summations. Let $O_{mi} = (X_{mi}, Y_{mi}, S_{mi})$ for $m = 1, 2$ and $i = 1, \cdots, n_m$, and define the projected statistic

$$\tilde{W}_{n1} = E(W_{n1}) + \sum_{m=1}^{2}\sum_{j=1}^{n_m}\{E(W_{n1}|O_{mj}) - E(W_{n1})\}.$$

Then by applying the theory of U-statistics (Hoeffding, 1948; Serfling, 1980; Koroljuk and Borovskich, 1994),

$$W_{n1} - E(W_{n1}) = \{\tilde{W}_{n1} - E(\tilde{W}_{n1})\}\{1 + o_p(1)\}. \tag{27}$$

Clearly, $E(W_{n1}) = \int F_1(y)dF_2(y) = \theta$, and it is straightforward to show that

$$\tilde{W}_{n1} = \theta + n_1^{-1}\sum_{i=1}^{n_1}\left\{\frac{\bar{F}_2(Y_{1i})S_{1i}}{\pi_1(X_{1i})} - \theta\right\} + n_2^{-1}\sum_{k=1}^{n_2}\left\{\frac{F_1(Y_{2k})S_{2k}}{\pi_2(X_{2k})} - \theta\right\} \tag{28}$$

where $\bar{F}(y)$ is the survival function defined to be $1 - F(y)$.

As for W_{n2}, we define two kernels of two sample V-statistics by

$$h_1(O_{1i}, O_{1j}; O_{2k}) = 1/2\{V_{ik}\pi_{2k}^{-1}K_h(X_{1i} - X_{1j})\eta_1^{-1}(X_{1j}) + V_{jk}\pi_{2k}^{-1}K_h(X_{1j} - X_{1i})\eta_1^{-1}(X_{1i})\},$$
$$h_2(O_{1i}; O_{2k}, O_{2j}) = 1/2\{V_{ik}\pi_{2k}^{-1}K_h(X_{1i} - X_{2j})\eta_1^{-1}(X_{2j}) + V_{ij}\pi_{2j}^{-1}K_h(X_{1i} - X_{2k})\eta_1^{-1}(X_{2k})\},$$

Then the first part of W_{n2} can be written as

$$W_{n2}^{(1)} = n^{-1}n_1^{-1}n_2^{-1}\left\{\sum_{i=1}^{n_1}\sum_{j=1}^{n_1}\sum_{k=1}^{n_2} h_1(O_{1i}, O_{1j}; O_{2k}) + \sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\sum_{k=1}^{n_2} h_1(O_{1i}; O_{2j}, O_{2k})\right\}. \quad (29)$$

By the V-statistics theory (Serfling, 1980), a V-statistic is equivalent in the first order to the U-statistic with the same kernel. Hence, by the projection method and note that

$$E\{h_1(O_{1i}, O_{1j}; O_{2k})|O_{2k}\} = \frac{F_1(Y_{2k})S_{2k}}{\pi_2(X_{2k})}, E\{h_2(O_{1i}; O_{2j}, O_{2k})|O_{1i}\} = \frac{\bar{F}_2(Y_{1i})S_{1i}}{\pi_1(X_{1i})},$$
$$E\{h_1(O_{1i}, O_{1j}; O_{2k})|O_{1i}\} = 1/2\left\{\frac{\bar{F}_2(Y_{1i})S_{1i}}{\pi_1(X_{1i})} + \int \bar{F}_2(y)dF_1(y|X_{1i})\right\} \text{ and}$$
$$E\{h_2(O_{1i}; O_{2j}, O_{2k})|O_{2k}\} = 1/2\left\{\frac{F_1(Y_{2k})S_{2k}}{\pi_2(X_{2k})} + \int F_1(y)dF_2(y|X_{2k})\right\}.$$

Note that $E(W_{n2}^{(1)}) = \theta$ and the projection of the second part of W_{n2} is the same as (28). Applying the same argument on W_{n3}, we obtain the following approximations to W_{n2} and W_{n3},

$$\tilde{W}_{n2} = n^{-1}\sum_{j=1}^{n}\{\xi_1(X_j) - \theta\} \text{ and } \tilde{W}_{n3} = n^{-1}\sum_{j=1}^{n}\{\xi_2(X_j) - \theta\}. \quad (30)$$

Applying the same approach for the V-statistics in W_{n4} and W_{n5}, we have the projected statistics

$$\tilde{W}_{n4} = n_1^{-1}\sum_{i=1}^{n_1}\left\{\frac{S_{1i}}{\pi_1(X_{1i})}\xi_1(X_{1i}) - \theta\right\} \text{ and } \tilde{W}_{n5} = n_2^{-1}\sum_{k=1}^{n_2}\left\{\frac{S_{2k}}{\pi_2(X_{2k})}\xi_2(X_{2k}) - \theta\right\}. \quad (31)$$

Then Lemma 1 follows by combining (28), (30) and (31).

A.2 Proof of Lemma 2

For simplicity in presentation, we let $\beta = (\beta_1^T, \beta_2^T)^T$ to be the combined unknown parameters in $\pi_m(x;\beta)$ for $m = 1$ and 2. Because $\{X_j\}_{j=1}^n = (X_{11}, \dots, X_{1n_1}, X_{21}, \dots, X_{2n_2})$ are independent and identically distributed (iid), $\{t_{1j}\}_{j=1}^n$ and $\{t_{2j}\}_{j=1}^n$ are also iid. We note that $\hat{t}_{mj} \xrightarrow{p} t_{ij}$ as $n \to \infty$ and the approximation of $\hat{t}_{mj}$ is given by Taylor's expansion

$$\hat{t}_{mj} = \pi_m(X_j;\hat{\beta}) = \pi_m(X_j, \beta_0) + \pi'_m(X_j;\tilde{\beta})(\hat{\beta} - \beta_0) = t_{mj} + \pi'_m(X_j;\tilde{\beta})(\hat{\beta} - \beta_0). \quad (32)$$

where $\tilde{\beta} \xrightarrow{p} \beta_0$ as $n \to \infty$. We now consider generic kernel smoothing taking the following form with $\hat{t}_{mj}$ as smoother and Ω as a generic observable random variable:

$$\phi_{mi} = n_m^{-1} \sum_{j=1}^{n_m} K_{h_m}(\hat{t}_{mi} - \hat{t}_{mj})\Omega_j = n_m^{-1} \sum_{j=1}^{n_m} K_{h_m}(t_{mi} - t_{mj} + \hat{t}_{mi} - t_{mi} - \hat{t}_{mj} + t_{mj})\Omega_j.$$

$$= n_m^{-1} \sum_{j=1}^{n_m} \left\{ K_{h_m}(t_{mi} - t_{mj}) + K'_{ij}(\pi'_{mi} - \pi'_{mj})(\hat{\beta} - \beta_0) \right\} \Omega_j.$$

where $K'_{ij} = K'_h\{t_{mi} - t_{mj} + \tau_{ij}\}$ with $\tau_{ij} \xrightarrow{p} 0$ and $\pi'_{mi} = \pi'(X_i; \tilde{\beta})$ as in (32). Clearly because the smoothing is targeted at t_{mi}, we have

$$n_m^{-1} \sum_{j=1}^{n_m} \left\{ K'_{ij}\{\pi'_{mi} - \pi'_{mj}\}(\hat{\beta} - \beta_0) \right\} \Omega_j = f'(X_{mi})(\pi'_{mi} - \pi'_{mi})(\hat{\beta} - \beta_0)E(\Omega|t_{mi})\{1 + o_p(1)\}.$$

Because by the assumption C4 that $\hat{\beta} - \beta_0$ is $\sqrt{n}$-consistent, we conclude

$$\phi_{mi} = n_m^{-1} \sum_{j=1}^{n_m} K_{h_m}(t_{mi} - t_{mj})\Omega_j + o_p(n^{-1/2}). \tag{33}$$

In other words, using the estimated covariate as smoother brings in ignorable impact comparing to using the corresponding true values. Then the remaining steps of proving Lemma 2 are exactly replicating those in proving Lemma 1 by replacing those X_{mj} by $\pi_m(X_{mj}, \beta_0)$.

A.3 Proof of Lemma 3

The proof in A.2 already shows that smoothing at an estimated index value is first order equivalent to that at the truth. We now show that the impact due to estimating β_m in the propensity function is also negligible in $\hat{F}_m(y|z)$.

$$\hat{f}_m(z) = n_m^{-1} \sum_{i=1}^{n_m} K(z - Z_{mi}) S_{mi} \pi_{mi}^{-1}\{1 - \pi_{mi}^{-1}(\hat{\pi}_{mi} - \pi_{mi})\} + o_p(n^{-1/2})$$

$$= n^{-1} \sum_{i=1}^{n_m} K(z - Z_{mi}) S_i \pi_{mi}^{-1} - f_m(z)\pi_{mi}^{-1}(\hat{\pi}_{mi} - \pi_{mi}) + o_p(n^{-1/2})$$

Let $\hat{b}_m(y, z) = n_m^{-1} \sum_{i=1}^{n_m} I(Y_{mi} \le y) K(z - Z_{mi}) S_{mi} \hat{\pi}_{mi}^{-1}$ and denote its probability limit by $b_m(y, z)$, it follows similarly that

$$\hat{b}_m(y, z) = n_m^{-1} \sum_{i=1}^{n_m} I(Y_{mi} \le y) K(z - Z_{mi}) S_{mi} \pi_{mi}^{-1} - b_m(y, z)\pi_{mi}^{-1}(\hat{\pi}_{mi} - \pi_{mi}) + o_p(n^{-1/2}).$$

Then substituting the above expressions into the following expansion,

$$\hat{F}_m(y|z) = \hat{b}_m(y, z)\hat{f}_m^{-1}(z) = \hat{b}_m(y, z) f_m^{-1}(z)[1 - f_m^{-1}(z)\{\hat{f}_m(z) - f_m(z)\}] + o_p(n^{-1/2}),$$

and note that the $\hat{\pi}_m$ terms exactly cancel each other. We note that this result is similar to the finding in Wang et al. (1998). The rest proof of Lemma 3 is repeating the proof of Lemma 1 by replacing X_{mi} by $g_m(X_{mi}; \gamma_{0m})$.

A.4 Proof of Theorem 4

The same projection method as in proving Lemma 1 is applicable to derive the asymptotic conditional distribution of W_n^*, with all the probability limits taken with respect to the empirical distribution. In particular, $\sqrt{n}\{W_n^* - E(W_n^*|\mathcal{F}_n)\}/v^* \xrightarrow{d} N(0,1)$ $a.s.$ where $(v^*)^2 = \lim_{n\to\infty} n\text{var}(W_n^*|\mathcal{F}_n)$. Let $\hat{\lambda}_m(x) = n^{-1}\sum_{j=1}^n K_{h_m}(x - X_j)\frac{\hat{\eta}_m^*(X_j)-\hat{\eta}_m(X_j)}{\hat{\eta}_m^2(X_j)}, \hat{\eta}_m^*(x) = n_m^{-1}\sum_{j=1}^{n_m} K_{h_m}(x - X_{mj}^*)S_{mj}^*$ and $\hat{\gamma}_m(x) = n_m^{-1}\sum_{j=1}^{n_m} \frac{K_{h_m}(x-X_{mj}^*)}{\hat{\eta}_m(X_{mj}^*)}$. By repeating the steps in proving Lemma 1, we can establish an expansion of W_n^* resembling (26) given $\mathcal{F}_n$ as $W_n^* = W_{n1}^* + W_{n2}^* + W_{n3}^* - W_{n4}^* - W_{n5}^* + o_p(n^{-1/2})$ where for $m = 1, 2$,

$$W_{n1}^* = n_1^{-1}n_2^{-1}\sum_{i=1}^{n_1}\sum_{k=1}^{n_2}\frac{I(\tilde{Y}_{1i}^* \le \tilde{Y}_{2k}^*)S_{1i}^*S_{2k}^*}{\hat{\pi}_1(X_{1i}^*)\hat{\pi}_2(X_{2k}^*)}, \hat{\pi}_m(x) = \hat{\eta}_m(x)/\hat{f}_m(x),$$

$$W_{n2}^* = n_1^{-1}n_2^{-1}\sum_{i=1}^{n_1}\sum_{k=1}^{n_2}\frac{I(\tilde{Y}_{1i}^* \le \tilde{Y}_{2k}^*)S_{1i}^*S_{2k}^*}{\hat{\pi}_2(X_{2k}^*)}\left\{n^{-1}\sum_{j=1}^{n}\frac{K_{h_1}(X_{1i}^* - X_j^*)}{\hat{\eta}_1(X_j^*)} - \hat{\gamma}_1(X_{1i}^*)\right\},$$

$$W_{n4}^* = n_1^{-1}n_2^{-1}\sum_{i=1}^{n_1}\sum_{k=1}^{n_2}\frac{I(\tilde{Y}_{1i}^* \le \tilde{Y}_{2k}^*)S_{1i}^*S_{2k}^*\hat{\lambda}_1(X_{1i}^*)}{\hat{\pi}_2(X_{2k}^*)},$$

and W_{3n}^* and W_{5n}^* are respectively the second sample version of W_{2n}^* and W_{4n}^* by switching indices i and k other than those in the index function.

The crucial implication of the proposed bootstrap procedure is $\hat{G}(\tilde{Y}_{mi}^*) = U_{mi} = \hat{F}_m(Y_{mi}^*)$ for $m = 1, 2$. The joint distribution of each sample is respected in the following sense,

$$\begin{aligned} P(\tilde{Y}_{mi}^* \le \tilde{y}, X_{mi}^* < x, S_{mi}^* = 1) &= P(\hat{F}_m^{-1}\{\hat{G}(\tilde{Y}_{mi}^*)\} \le \hat{F}_m^{-1}\{\hat{G}(\tilde{y})\}, X_{mi}^* < x, S_{mi}^* = 1) \\ &= P(Y_{mi}^* < y_m, X_{mi}^* < x, S_{mi}^* = 1) \text{ for } m = 1, 2. \end{aligned} \tag{34}$$

Here y_m and $\tilde{y}$ are connected such that y_m is the $\hat{G}(\tilde{y})$-th estimated quantile in the m-th sample, for every $\tilde{y}$ in its sample space. It is clear that $\tilde{Y}_{1i}^*$ and $\tilde{Y}_{2i}^*$ follow the same marginal distribution $\hat{G}$. Under the null hypothesis the joint distribution is exactly preserved because

$$|y_m - \tilde{y}| = |\hat{F}_m^{-1}\{\hat{G}(\tilde{y})\} - \tilde{y}| \xrightarrow{p} |F^{-1}\{F(\tilde{y})\} - \tilde{y}| = 0 \text{ as } n_m \to \infty.$$

Then (34) implies as $n_m \to \infty$, $P(\tilde{Y}_{mi}^* \le \tilde{y}, X_{mi}^* < x, S_{mi}^* = 1) \xrightarrow{p} P(Y_{mi} < y, X_{mi} < x, S_{mi} = 1)$. Therefore, as $n \to \infty$, $E(W_n^*|\mathcal{F}_n) = 1/2 + o(n^{-1/2})$ $a.s.$ It remains to show that under the null hypothesis, $\lim_{n\to\infty} n\text{var}(W_n^*|\mathcal{F}_n) \to v^2(1/2)$ $a.s.$ This is because conditioning on $\mathcal{F}_n$ and from (34), we have $\omega_2^*(x) = \int \hat{F}_1(\tilde{y})d\hat{F}_2(\tilde{y}|x) = \int \hat{G}(\tilde{y})d\hat{F}_2(\tilde{y}|x) = \int \hat{F}_1(y_1)d\hat{F}_2(\tilde{y}|x)$. As $n \to \infty$, under the null hypothesis $y_m \to \tilde{y}$ $a.s.$ for $m = 1, 2$, and hence $\omega_2^*(x) \to \int F_1(y)dF_2(y|x) = \xi_2(x)$ $a.s.$ Similarly $\int\{1 - \hat{F}_1(\tilde{y})\}d\hat{F}_1(\tilde{y}|x) \to \xi_1(x)$ $a.s.$ Theorem 3 then follows similarly as proving Theorem 1.

References

Bilker, W. and Wang, M.-C. (1996), "A semiparametric extension of the Mann-Whitney test for randomly truncated data," *Biometrics*, 52, 10–20.

Breslow, N. (2003), "Are statistical contributions to medicine undervalued?" *Biometrics*, 59, 1–8.

Cheng, P. E. (1994), "Nonparametric-estimation of mean functionals with data missing at random," *Journal of the American Statistical Association*, 89, 81–87.

Cheung, Y. K. (2005), "Exact two-sample inference with missing data," *Biometrics*, 61, 524–531.

Chu, C. K. and Chen, K. F. (1995), "Nonparametric regression estimates using misspecified binary responses," *Biometrika*, 82, 315–325.

Dehejia, R. H. and Wahba, S. (1999), "Causal effects in nonexperimental studies: reevaluation the evaluation of training programs," *Journal of the American Statistical Association*, 94, 1053–1062.

Fan, J. and Gijbels, I. (1996), *Local Polynomial Modeling and Its Applications*, Chapman and Hall, London.

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

Hall, P., Racine, J., and Li, Q. (2004), "Cross-validation and the estimation of conditional probability densities," *Journal of the American Statistical Association*, 99, 1015–1026.

Hall, P. and Yao, Q. (2005), "Approximating conditional distribution functions using dimension reduction," *The Annals of Statistics*, 33, 1404–1421.

Hansen, L. P. (1982), "Large sample properties of generalized method of moments estimators," *Econometrica*, 50, 1029–1054.

Härdle, W. (1990), *Applied Nonparametric Regression*, Cambridge: Cambridge University Press.

Hirano, K., Imbens, G. W., and Ridder, G. (2003), "Efficient estimation of average treatment effects using the estimated propensity score," *Econometrica*, 71, 1161–1189.

Hoeffding, W. (1948), "A class of statistics with asymptotically normal distribution," *The Annals of Mathematical Statistics*, 19, 293–325.

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

Hu, Z., Follmann, D. A., and Qin, J. (2010), "Semiparametric dimension reduction estimation for mean response with missing data," *Biometrika*, 97, 305–319.

— (2011), "Dimension reducted kernel estimation for distribution function with incomplete data," *Journal of Stat*, 141, 3084–3093.

Imbens, G. (2004), "Nonparametric estimation of average treatment effects under exogeneity: a review," *Review of Economics and Statistics*, 86, 4–30.

Korn, E. L. and Baumrind, S. (1998), "Clinician preferences and the estimation of causal treatment differences," *Statistical Science*, 13, 209–235.

Koroljuk, V. S. and Borovskich, Y. V. (1994), *Theory of U-Statistics*, Kluwer, Dordrecht.

Kuk, A. Y. C. (1993), "A kernel method for estimating finite population functions using auxiliary information," *Biometrika*, 80, 385–392.

Lalonde, R. J. (1986), "Evaluating the econometric evaluations of training programs with experimental data," *American Economic Review*, 76, 604–620.

Little, R. and Rubin, D. (2002), *Statistical Analysis With Missing Data*, Wiley, 2nd ed.

Matloff, N. S. (1981), "Use of regression functions for improved estimation of means," *Biometrika*, 68, 685–689.

Newey, W. K. and McFadden, D. (1994), "Large sample estimation and hypothesis testing," *Handbook of Econometrics, Vol 4, ed. by R. Engle and D. McFadden. New York: North Holland.*

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

Rosenbaum, P. R. (2002), *Observational Studies*, Springer-Verlag: New York.

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.

Rubin, D. B. (1976), "Inference and missing values (with discussion)," *Biometrika*, 63, 581–592.

Serfling, R. J. (1980), *Approximation Theorems of Mathematical Statistics*, John Wiley.

Tsiatis, A. A. (2006), *Semiparametric Theory and Missing Data*, Springer-Verlag: New York.

Wang, C. Y., Wang, S., Gutierrez, R. G., and Carroll, R. J. (1998), "Local linear regression for generalized linear models with missing data," *The Annals of Statistics*, 26, 1028–1050.

Table 1: Empirical means and standard deviations (SDs) of R_n, W_n, T_n (propensity function based), Z_n (working linear function based), given by (17), (4), (13) and (19) respectively, for the Gaussian distributed responses under H_0.

			MAR1				MAR2			
n	d		R_n	W_n	T_n	Z_n	R_n	W_n	T_n	Z_n
50	1	Mean	0.504	0.495	0.490	0.496	0.502	0.495	0.496	0.497
	2		0.507	0.487	0.491	0.496	0.504	0.490	0.492	0.499
	3		0.497	0.473	0.489	0.491	0.502	0.479	0.488	0.489
	4		0.503	0.468	0.488	0.493	0.502	0.466	0.485	0.488
50	1	SD	0.080	0.070	0.075	0.071	0.084	0.074	0.085	0.076
	2		0.087	0.074	0.076	0.071	0.092	0.086	0.086	0.078
	3		0.088	0.079	0.076	0.071	0.096	0.089	0.087	0.079
	4		0.092	0.082	0.077	0.073	0.103	0.091	0.086	0.080
80	1	Mean	0.506	0.496	0.495	0.503	0.505	0.499	0.495	0.499
	2		0.507	0.487	0.491	0.495	0.507	0.488	0.491	0.504
	3		0.507	0.484	0.490	0.495	0.510	0.482	0.488	0.501
	4		0.506	0.478	0.488	0.493	0.503	0.474	0.485	0.495
80	1	SD	0.051	0.048	0.050	0.046	0.065	0.057	0.063	0.057
	2		0.055	0.052	0.049	0.047	0.067	0.063	0.065	0.059
	3		0.058	0.055	0.052	0.048	0.069	0.065	0.065	0.058
	4		0.060	0.056	0.049	0.046	0.072	0.068	0.064	0.057

Table 2: Empirical sizes ($\times 10^2$) of the proposed nonparametrically and semiparametrically adjusted Mann-Whitney tests based on W_n, T_n (propensity function) and Z_n (working linear function), the tests based on the covariate adjusted R_n and $\tilde{t}_n$ and the Oracle test W_{0n} and the two sample t test. The outcome distributions are Gaussian and $\alpha = 0.05$.

n	d	t	W_{0n}	R_n	$\tilde{t}_n$	W_n	T_n	Z_n	R_n	$\tilde{t}_n$	W_n	T_n	Z_n
		5.1	4.9	MAR1					MAR2				
50	1	-	-	5.1	4.7	5.2	5.4	4.8	5.8	5.5	5.4	5.2	5.2
	2	-	-	6.2	6.2	4.4	4.9	5.4	6.0	6.1	3.6	4.9	4.8
	3	-	-	5.7	6.0	4.1	4.5	5.4	6.3	6.5	3.3	5.2	5.3
	4	-	-	5.9	6.2	3.5	5.4	5.2	6.7	6.9	3.0	4.6	4.5
		4.8	5.3	MAR1					MAR2				
80	1	-	-	5.4	4.4	4.6	4.4	4.8	5.8	5.7	4.6	4.3	4.6
	2	-	-	5.4	5.6	4.0	4.6	5.2	5.8	5.4	3.6	4.7	5.4
	3	-	-	6.2	5.9	3.6	4.2	4.9	4.6	5.3	3.5	4.6	5.2
	4	-	-	5.7	5.8	3.3	4.5	4.9	5.2	5.6	3.1	4.5	5.5

Table 3: Empirical powers ($\times 10^2$) of the proposed nonparametrically and semiparametrically adjusted Mann-Whitney tests based on W_n, T_n (propensity function) and Z_n (working linear function), the tests based on the covariate adjusted R_n and $\tilde{t}_n$ and the Oracle test W_{0n} and the two sample t test. The outcome distributions are Gaussian and $\alpha = 0.05$.

n	d		t	W_{0n}	R_n	$\tilde{t}_n$	W_n	T_n	Z_n	R_n	$\tilde{t}_n$	W_n	T_n	Z_n
50		Power$_1$	39.6	38.4	MAR1					MAR2				
	1		-	-	29.8	31.4	36.4	32.1	36.6	21.2	22.4	24.6	23.2	24.8
	2		-	-	27.3	30.4	34.6	32.4	35.9	18.3	19.9	21.2	24.1	25.8
	3		-	-	26.3	31.0	31.4	32.9	36.0	16.1	18.5	20.4	24.6	25.9
	4		-	-	23.1	23.5	26.9	31.5	35.9	14.2	16.1	17.9	22.2	25.3
		Power$_2$	72.0	70.6	MAR1					MAR2				
	1		-	-	53.6	58.8	68.4	62.2	68.1	34.8	40.2	48.2	45.3	48.2
	2		-	-	56.5	57.4	65.4	62.0	67.8	34.9	41.1	41.2	45.1	48.1
	3		-	-	56.5	59.0	60.2	62.9	68.3	33.6	39.8	40.1	44.7	48.1
	4		-	-	53.9	54.7	58.2	63.7	68.7	32.0	32.4	37.3	43.7	48.2
80		Power$_1$	55.4	53.2	MAR1					MAR2				
	1		-	-	46.4	48.6	52.4	50.8	52.6	36.1	38.0	39.8	37.8	39.6
	2		-	-	46.6	46.8	50.4	49.8	52.4	34.4	34.8	35.6	37.0	38.9
	3		-	-	41.4	43.6	44.8	48.9	52.1	32.2	32.8	31.4	37.6	39.2
	4		-	-	39.6	43.8	41.4	48.2	52.6	27.3	29.7	29.0	37.2	39.4
		Power$_2$	87.8	85.2	MAR1					MAR2				
	1		-	-	77.0	78.0	82.6	79.6	82.6	52.6	60.0	67.2	64.2	67.8
	2		-	-	76.0	78.4	81.0	78.8	82.8	53.8	60.6	62.6	64.9	67.9
	3		-	-	76.8	77.2	76.8	77.9	82.5	53.6	61.0	56.8	63.9	67.8
	4		-	-	75.9	76.9	72.1	77.1	81.9	52.9	59.8	55.2	63.0	68.1

Table 4: Empirical powers ($\times 10^2$) the proposed nonparametrically and semiparametrically adjusted Mann-Whitney tests based on W_n, T_n (propensity function) and Z_n (working linear function), the tests based on the covariate adjusted R_n and $\tilde{t}_n$ and the Oracle test W_{0n} and the two sample t test. The outcome distributions are Gaussian and Gamma, and $\alpha = 0.05$.

n	d	t	W_{0n}	R_n	$\tilde{t}_n$	W_n	T_n	Z_n	R_n	$\tilde{t}_n$	W_n	T_n	Z_n
50		4.8	33.2	MAR1					MAR2				
	1	-	-	19.4	8.8	27.8	23.6	28.1	16.4	4.6	22.1	17.7	23.6
	2	-	-	17.5	7.6	24.9	22.2	27.8	15.4	4.8	19.0	17.9	22.9
	3	-	-	15.3	5.1	20.9	24.1	27.9	14.7	6.3	16.7	17.7	23.3
	4	-	-	13.6	3.0	17.8	23.8	27.7	13.1	3.9	15.9	17.5	23.2
80		5.3	42.6	MAR1					MAR2				
	1	-	-	25.6	5.6	37.4	35.6	38.6	24.2	4.8	35.6	32.6	36.1
	2	-	-	24.4	4.4	37.0	35.9	39.1	21.4	4.2	33.4	33.6	36.2
	3	-	-	23.0	7.2	33.4	35.9	38.3	17.6	4.0	30.2	35.1	37.2
	4	-	-	21.2	7.7	29.1	34.2	38.9	17.5	3.8	28.9	34.1	37.0

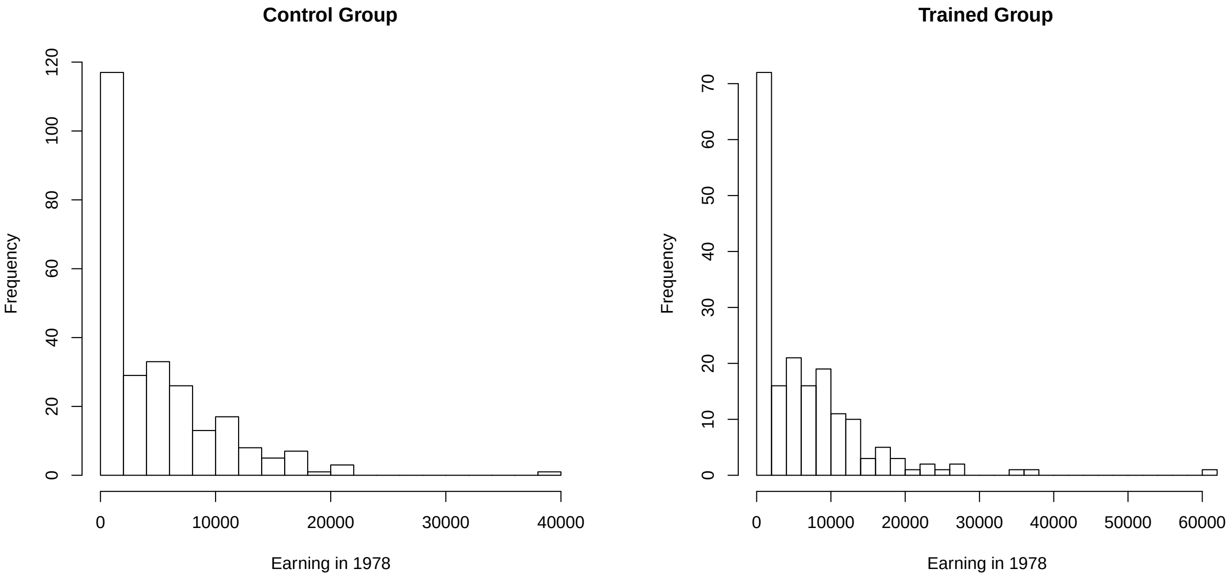

Figure 1: Histograms of the earnings in 1978

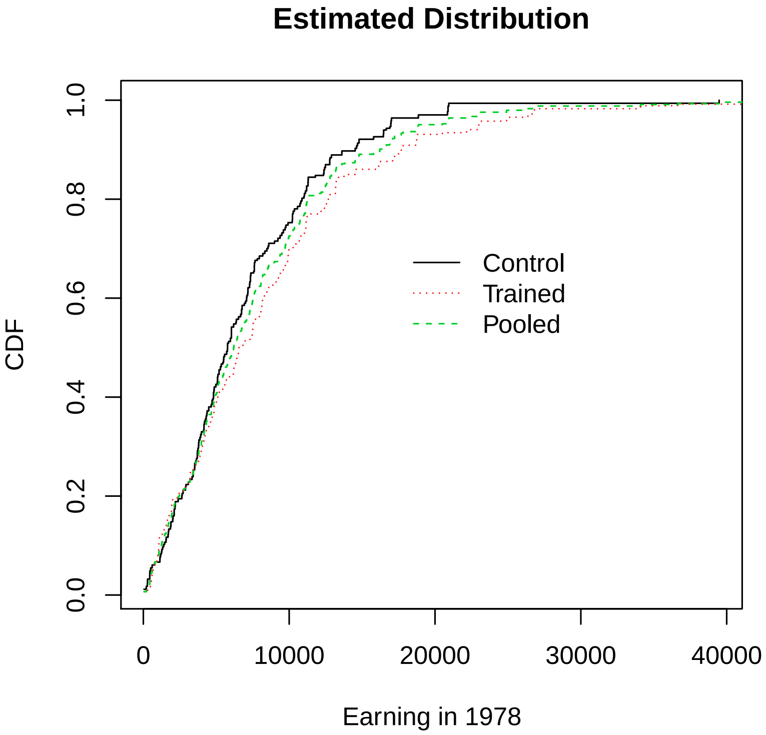

Figure 2: The estimated CDFs of the earnings in 1978 for the trained group, control group and the pooled samples.