Robust Post-Matching Inference

Alberto Abadie Jann Spiess

MIT Microsoft Research

January 2019

First version: December 2015

Abstract

Nearest-neighbor matching is a popular nonparametric tool to create balance between treatment and control groups in observational studies. As a preprocessing step before regression, matching reduces the dependence on parametric modeling assumptions. In current empirical practice, however, the matching step is often ignored in the calculation of standard errors and confidence intervals. In this article, we show that ignoring the matching step results in asymptotically valid standard errors if matching is done without replacement and the regression model is correctly specified relative to the population regression function of the outcome variable on the treatment variable and *all* the covariates used for matching. However, standard errors that ignore the matching step are not valid if matching is conducted with replacement or, more crucially, if the second step regression model is misspecified in the sense indicated above. Moreover, correct specification of the regression model is not required for consistent estimation of treatment effects with matched data. We show that two easily implementable alternatives produce approximations to the distribution of the post-matching estimator that are robust to misspecification. A simulation study and an empirical example demonstrate the empirical relevance of our results.

Alberto Abadie, Department of Economics, MIT, abadie@mit.edu. Jann Spiess, Microsoft Research New England, jspiess@stanford.edu. We thank Gary King and seminar participants at Harvard for helpful comments. Financial support by the NSF through grant SES 0961707 is gratefully acknowledged.

1 Introduction

Matching methods are widely used to create balance between treatment and control groups in observational studies. Oftentimes, matching is followed by a simple comparison of means between treated and nontreated (Cochran, 1953; Rubin, 1973; Dehejia and Wahba, 1999). In other instances, however, matching is used in combination with regression or with other estimation methods more complex than a simple comparison of means. The combination of matching in a first step with a second-step regression estimator brings together parametric and nonparametric estimation strategies and, as demonstrated in Ho et al. (2007), reduces the dependence of regression estimates on modeling decisions. Moreover, matching followed by regression allows the estimation of elaborate models, such as those that include interaction effects and other parameters that go beyond average treatment effects.

In this article, we develop valid standard error estimates for regression after matching. The asymptotic properties of average treatment effect estimators that employ a simple comparison of mean outcomes between treated and nontreated after matching on covariates are well understood (Abadie and Imbens, 2006). However, studies that employ regression models after matching usually ignore the matching step when performing inference on post-matching regression coefficients. We show that this practice is not generally valid if the second step regression is misspecified in the sense we make precise below. We provide standard error formulas that are robust to misspecification for regression coefficient estimators applied to matched samples (with matching done without replacement). First, we show that standard errors that are clustered at the level of the matches are valid under misspecification. Second, we show that a nonparametric block bootstrap that resamples matched pairs or matched sets, as opposed to resampling individual observations, also yields valid inference under misspecification. Furthermore, we show that standard errors that ignore the matching step can both under- or overestimate the variation of post-matching estimates. The procedures proposed in this article are straightforward to implement with standard statistical software.

We will consider the following setup. Let W be a binary random variable represent-

ing exposure to the treatment or condition of interest (e.g., smoking), so $W = 1$ for the treated, and $W = 0$ for the nontreated. Y is a random variable representing the outcome of interest (e.g., forced expiratory volume) and X is a vector of covariates (e.g., gender or age). We will study the problem of estimating how the treatment affects the outcomes of the individuals in the treated population (that is, those with $W = 1$). In particular, we will analyze the properties of a two-step (first matching, then regression) estimator often used in empirical practice. This estimation strategy starts with an unmatched sample, $\mathcal{S}$, from which treated units and their matches are extracted to create a matched sample, $\mathcal{S}^*$. Matching is done without replacement and on the basis of the values of X. Then, using data for the matched sample only, the researcher runs a regression of Y on Z, where Z is a vector of functions of W and X (e.g., individual variables plus interactions). We aim to obtain valid inferential methods for the coefficients of this regression, possibly under misspecification. To be precise, by "misspecification" we mean that there is no version of the conditional expectation of Y given W and X that follows the functional form employed in the second-step estimator. For example, as explained below, a difference in means between treated and nontreated in the second step would be "misspecified" if the conditional expectation of Y given X and W depends on X. To simplify the exposition, here we have described a setting where Z depends only on the treatment, W, and on the covariates used in the matching stage, X. Our general framework in Section 2 allows Z to depend on other covariates not in X.

A special case of our setup is that of the standard matching estimator for the average treatment effect on the treated, which is given by the regression coefficient on treatment W in a regression of Y on $Z = (1, W)'$. In this sense, our article generalizes the standard theory for matching estimators. However, the framework allows for richer analysis, such as the analysis of linear interaction effects of the treatment with a given covariate, $Z = (1, W, WX', X')'$.

To illustrate the implications of our results, consider the simple case when $Z = (1, W)'$. As we mentioned in the previous paragraph, in this setting, the sample regression coeffi-

cient on W corresponds to the simple matching estimator often employed in applied studies, which is based on a post-matching comparison of means between treated and nontreated. Under well-known conditions this estimator is consistent for the average effect of the treatment on the treated (see, e.g., Abadie and Imbens, 2012), irrespective of the true form of the expectation of Y given W and X. Notice, however, that even in this simple scenario, our results imply that regression standard errors that ignore the matching step are not valid in general. While the expectation of Y given W always admits a linear version given that W is binary, a linear regression of Y on $Z = (1, W)'$ will be misspecified *relative to the regression of* Y *on* W *and* X, unless Y is mean-independent of X given W over a set of probability one.

The rest of the article is organized as follows. Section 2 starts with a detailed description of the setup of our investigation. We then characterize the parameters estimated by the two-step procedure described above. We show that these parameters coincide with the regression coefficients in a regression of Y on Z in a population for which the distribution of matching covariates X in the control group has been modified to coincide with that of the treated. Under selection on observables, that is, if treatment is as good as random conditional on X, post-matching regression estimands coincide with the population regression coefficients in an experiment where the treatment is randomly assigned in a population that has the same distribution of X as the treated. We next establish consistency with respect to this vector of parameters, show asymptotic normality, and describe the asymptotic variance of the post-matching estimator. In Section 3, we discuss different ways of constructing standard errors. Based on the results of Section 2, we show that standard errors that ignore the matching step are not generally valid if the regression model is misspecified in the sense indicated above, while clustered standard errors or an analogous block bootstrap procedure yield valid inference. Section 4 presents simulation evidence, which confirms our theoretical results. Section 5 applies our results to the analysis of the effect of smoking on pulmonary function. In this application, both matching before regression and the use of the robust standard errors proposed in this article substantially affect empirical

results. Section 6 concludes.

The appendix contains the proofs of our main results. A supplementary appendix contains proofs of intermediate results and two extensions. In particular, the standard errors derived in this article are valid for unconditional inference. Alternatively, one could perform inference conditional on the values of the regressors, X and W, in the sample. Notice that, in this case, the first step matches are fixed. We discuss this alternative setting in the supplementary appendix, where we show that, for the conditional case, the usual regression standard errors are not generally valid, but valid standard errors can be calculated using the formulas in Abadie et al. (2014). Also, for concreteness and following the vast majority of applied practice, we restrict our analysis to linear regression after matching. In the supplementary appendix we provide an extension of our result to general M-estimation after matching.

2 Post-Matching Inference

In this section, we discuss the asymptotic distribution of the least squares estimator obtained from a linear regression of Y on Z after matching on observables X.

2.1 Post-Matching Least Squares

Consider a standard binary treatment setting along the lines of Rubin (1974) with potential outcomes $Y(1)$ and $Y(0)$, of which we only observe $Y = Y(W)$ for treatment status $W \in \{0, 1\}$. Let S be a set of observed covariates.

We will assume that the data consist of random samples of treated and nontreated. This assumption could be easily relaxed, and we adopt it only to simplify the discussion.

Assumption 1 (Random sampling). $\mathcal{S} = \{(Y_i, W_i, S_i)\}_{i=1}^{N}$ *is a pooled sample obtained from* N_1 *and* N_0 *independent draws from the population distribution of* (Y, S) *for the treated* $(W = 1)$ *and nontreated* $(W = 0)$*, respectively, so* $N = N_0 + N_1$.

Let $\mathcal{S}^* \subseteq \mathcal{S}$ be the matched sample generated by matching each treated unit, i, to M nontreated units, $\mathcal{J}(i)$ without replacement. Specifically, consider an $(m \times 1)$ vector of

covariates $X = f(S) \in \mathcal{X} \subseteq \mathbb{R}^m$, along with some distance metric $d : \mathcal{X} \times \mathcal{X} \to [0, \infty)$ on the support $\mathcal{X}$ of the covariates. Then, the sets of matches, $\mathcal{J}(i) \subseteq \{j; W_j = 0\}$ for all treated units are chosen to minimize the sum of the matching discrepancies

$$\sum_{i=1}^{N} W_i \sum_{j \in \mathcal{J}(i)} d(X_i, X_j),$$

where every nontreated unit appears in at most one set of matches. That is, matching is done without replacement. For simplicity, we omit in our notation the dependence of $\mathcal{J}(i)$ on N and M.

The matched sample, $\mathcal{S}^*$, has size $n = (M + 1)N_1$. We use a double subscript notation to refer to the observations in the matched sample. For instance, $Y_{n1}, \ldots, Y_{nn}$ refers to the values of the outcome variable for the units in $\mathcal{S}^*$, with analogous notation for other variables. Within the matched sample, observations will be rearranged so that the first N_1 observations are the treated units.

Let $Z = g(W, S)$ be a $(k \times 1)$ vector of functions of (W, S), and let $\widehat{\beta}$ be the vector of sample regression coefficients obtained from regressing Y on Z in the matched sample,

$$\begin{aligned} \widehat{\beta} &= \underset{b \in \mathbb{R}^k}{\operatorname{argmin}} \ \frac{1}{n} \sum_{i=1}^{n} (Y_{ni} - Z_{ni}' b)^2 \\ &= \left(\frac{1}{n} \sum_{i=1}^{n} Z_{ni} Z_{ni}' \right)^{-1} \frac{1}{n} \sum_{i=1}^{n} Z_{ni} Y_{ni}. \end{aligned} \tag{1}$$

In Section 2.3 we will introduce a set of assumptions under which $\widehat{\beta}$ exists and is unique with probability approaching one.

As we mentioned above, when $Z = (1, W)'$ the regression coefficient on W in the matched sample is given by

$$\begin{aligned} \widehat{\tau} &= \frac{1}{N_1} \sum_{i=1}^{n} W_{ni} Y_{ni} - \frac{1}{M N_1} \sum_{i=1}^{n} (1 - W_{ni}) Y_{ni} \\ &= \frac{1}{N_1} \sum_{i=1}^{N} W_i \left(Y_i - \frac{1}{M} \sum_{j \in \mathcal{J}(i)} Y_j \right), \end{aligned}$$

which is the usual matching estimator for the average effect of the treatment on the treated.

2.2 Characterization of the Estimand

Before we study the sampling distribution of $\widehat{\beta}$, we first characterize its population counterpart, which we will denote by β. That is, our first task is to obtain a precise description of the nature of the parameters estimated by $\widehat{\beta}$. Although post-matching regressions are often used in empirical practice, to the best of our knowledge, the precise nature of post-matching estimands has not been previously derived.

The goal of matching is to change the distribution of the covariates in the sample of nontreated units, so that it reproduces the distribution of the covariates among the treated. In order to do so it is necessary that the support of the matching variables, X, for the treated is inside the support for the nontreated.

Assumption 2 (Support condition). *Let* $\mathcal{X}_1 = \text{supp}(X|W=1)$ *and* $\mathcal{X}_0 = \text{supp}(X|W=0)$*, then*

$$\mathcal{X}_1 \subseteq \mathcal{X}_0.$$

We now describe the population distribution targeted by the matched sample, $\mathcal{S}^*$. Let $P(\cdot|W=1)$ and $P(\cdot|W=0)$ be the *matching source* distributions of (Y,S) from where the treated and nontreated samples in $\mathcal{S}$ are respectively drawn, and let $E[\cdot|W=1]$ and $E[\cdot|W=0]$ be the corresponding expectation operators. For given $P(\cdot|W=1)$ and $P(\cdot|W=0)$ and a given number of matches, M, we define a *matching target* distribution, P^*, over the triple (Y,S,W), as follows:

$$P^*(W=1) = \frac{1}{1+M},$$

and for each measurable set, A,

$$P^*((Y,S) \in A|W=1) = P((Y,S) \in A|W=1),$$

and

$$P^*((Y,S) \in A|W=0) = E[P((Y,S) \in A|W=0,X)|W=1].$$

That is, in the matching target distribution: (i) treatment is assigned in the same proportion as in the matched sample; (ii) the distribution of (Y,S) among the treated is the

same as in the matching source; (iii) the distribution of (Y, S) among the nontreated is generated by integrating the conditional distribution of (Y, S) given X and $W = 0$ over the distribution of X given $W = 1$, in the matching source. As a result, under the matching target distribution, the distribution of X given $W = 0$ coincides with the distribution of X given $W = 1$.

Under regularity conditions stated below, estimation on the matched sample, $\mathcal{S}^*$, asymptotically recovers parameters of the matching target distribution, P^*, in which the treated and nontreated have the same distribution of X, but possibly different outcome and covariate distributions conditional on X. As a result, comparisons of outcomes between treated and nontreated in the matched sample, $\mathcal{S}^*$, produce the controlled contrasts of the Oaxaca-Blinder decomposition (Oaxaca, 1973; Blinder, 1973; and DiNardo et al., 1996). More generally, under regularity conditions, regression coefficients of Y on Z in the matched sample, $\mathcal{S}^*$, asymptotically recover the analogous regression coefficients in the target population:

$$\begin{aligned} \beta &= \underset{b \in \mathbb{R}^k}{\operatorname{argmin}} \; E^*[(Y - Z'b)^2] \\ &= (E^*[ZZ'])^{-1} E^*[ZY]. \end{aligned} \tag{2}$$

Matching methods are often motivated by a selection-on-observables assumption, that is, by the assumption that treatment assignment is as good as random conditional on observed covariates. To formalize the assumption of selection on observables and its implications in our framework, consider source populations expressed this time in terms of potential outcomes and covariates, $Q(\cdot|W = 1)$ and $Q(\cdot|W = 0)$, which represent the distributions of $(Y(1), Y(0), S)$ given $W = 1$ and $W = 0$, respectively. These distributions are defined in such a way that $P(\cdot|W = 1)$ and $P(\cdot|W = 0)$ can be obtained by integrating out $Y(0)$ from $Q(\cdot|W = 1)$ and $Y(1)$ from $Q(\cdot|W = 0)$, respectively. For given $Q(\cdot|W = 1)$ and $Q(\cdot|W = 0)$, selection on observables means

$$(Y(1), Y(0), S)|X, W = 1 \sim (Y(1), Y(0), S)|X, W = 0$$

almost surely with respect to the distribution of $X|W = 1$. That is, the joint distribution of covariates and potential outcomes is independent of treatment assignment conditional

on the matching variables. Because in this article we focus on causal parameters defined for a population with distribution of the matching variables equal to $X|W = 1$, for our purposes it is enough that the selection-on-observables assumption holds for the distribution of $(Y(0), S)$ only,

$$(Y(0), S)|X, W = 1 \sim (Y(0), S)|X, W = 0. \tag{3}$$

Proposition 1 (Estimand under selection on observables). *Suppose that Assumption 2 holds and that β, as defined in Equation (2), exists and is finite. Then if selection on observables, as defined in Equation (3), holds, the coefficients β are the same as the population coefficients that would be obtained from a regression of Y on Z in a setting where:*

(a) $(Y(1), Y(0), S)$ has distribution $Q(\cdot|W = 1)$,

(b) treatment is randomly assigned with probability $1/(M + 1)$.

This result formalizes the notion that matching under selection on observables allows researchers to reproduce an experimental setting under which average treatment effects can be easily evaluated through a least squares regression of Y on Z. The results in this article, however, apply to the general estimand β in Equation (2), regardless of the validity of the selection-on-observables assumption.

2.3 Consistency and Asymptotic Normality

In this section, we will establish large sample properties of $\widehat{\beta}$, as $N_1, N_0 \to \infty$ with $N_0 \geq MN_1$. Throughout this article, we will assume that the sum of matching discrepancies vanishes quickly enough to allow asymptotic unbiasedness and root-n consistency:

Assumption 3 (Matching discrepancies).

$$\frac{1}{\sqrt{N_1}} \sum_{i=1}^{N} W_i \sum_{j \in \mathcal{J}(i)} d(X_i, X_j) \xrightarrow{p} 0.$$

Abadie and Imbens (2012) derive primitive conditions for Assumption 3. Of course, in concrete empirical settings, the adequacy of matching should not rely on asymptotic results.

Instead, the quality of the matches needs to be evaluated for each particular sample (e.g., using normalized differences as in Abadie and Imbens, 2011).

For any real matrix A, let $\|A\| = \sqrt{\text{tr}(A'A)}$ be the Euclidean norm of A. The next assumption collects regularity conditions on the conditional moments of (Y, Z) given (X, W).

Assumption 4 (Well-behavedness of conditional expectations). *For $w = 0, 1$, and some $\delta > 0$,*

$$E[\|Z\|^4|W = w, X = x] \quad \textit{and} \quad E[\|Z(Y - Z'\beta)\|^{2+\delta}|W = w, X = x]$$

are uniformly bounded on $\mathcal{X}_w$. Furthermore,

$$E[ZZ'|X = x, W = 0], \quad E[ZY|X = x, W = 0] \quad \textit{and} \quad \text{var}(Z(Y - Z'\beta)|X = x, W = 0)$$

are componentwise Lipschitz in x with respect to $d(\cdot, \cdot)$.

To ensure the existence of $\widehat{\beta}$ with probability approaching one as $n \to 0$, we assume invertibility of the Hessian, $H = E^*(ZZ')$. Notice that

$$H = \frac{E\Big[E[ZZ'|X, W = 1] + ME[ZZ'|X, W = 0]\Big|W = 1\Big]}{1 + M}. \tag{4}$$

Assumption 5 (Linear independence of regressors). *H is invertible.*

The next proposition establishes the asymptotic distribution of $\widehat{\beta}$.

Proposition 2 (Asymptotic distribution of the post-matching estimator). *Under Assumptions 1 to 5,*

$$\sqrt{n}(\widehat{\beta} - \beta) \xrightarrow{d} \mathcal{N}(0, H^{-1}JH^{-1}),$$

where

$$J = \frac{\text{var}\Big(E[Z(Y - Z'\beta)|X, W = 1] + ME[Z(Y - Z'\beta)|X, W = 0]\Big|W = 1\Big)}{1 + M} + \frac{E\Big[\text{var}(Z(Y - Z'\beta)|X, W = 1] + M\text{var}(Z(Y - Z'\beta)|X, W = 0)\Big|W = 1\Big]}{1 + M}$$

and H is as defined in Equation (4).

All proofs are in the appendix.

3 Post-Matching Standard Errors

In the previous section, we established that

$$\sqrt{n}(\widehat{\beta} - \beta) \xrightarrow{d} \mathcal{N}(0, H^{-1}JH^{-1})$$

for the post-matching estimator obtained from a regression of Y on Z within the matched sample $\mathcal{S}^*$. In this section, our goal is to estimate the asymptotic variance, $H^{-1}JH^{-1}$.

3.1 OLS Standard Errors Ignoring the Matching Step

Ho et al. (2007) argue that matching can be seen as a preprocessing step, prior to estimation, so the matching step can be ignored in the calculation of standard errors. Here, we consider commonly applied Eicker–Huber–White (EHW or "sandwich") standard error estimates for i.i.d. data (Eicker, 1967; Huber, 1967; White, 1980a,b, 1982). EHW standard errors are robust to misspecification.

OLS (EHW) standard errors can be computed as the square root of the main diagonal of the matrix $\widehat{H}^{-1}\widehat{J}_r\widehat{H}^{-1}/n$, where

$$\widehat{H} = \frac{1}{n}\sum_{i=1}^{n} Z_{ni}Z'_{ni} \tag{5}$$

and

$$\widehat{J}_r = \frac{1}{n}\sum_{i=1}^{n} Z_{ni}(Y_{ni} - Z'_{ni}\widehat{\beta})^2 Z'_{ni}. \tag{6}$$

The following proposition derives the probability limit of $\widehat{J}_r$ with data from a matched sample.

Proposition 3 (Convergence of J_r). *Suppose that Assumptions 1 to 5 hold. Assume also that*

$$E[Z(Y - Z'\beta)^2 Z'|X = x, W = 0]$$

is Lipschitz on $\mathcal{X}_0$ and

$$E[Y^4|X = x, W = w]$$

is uniformly bounded on $\mathcal{X}_w$ *for all* $w \in \{0,1\}$*. Then,* $\widehat{J}_r \xrightarrow{p} J_r$*, where*

$$J_r = \frac{E\Big[E[Z(Y - Z'\beta)^2 Z'|X, W = 1] + ME[Z(Y - Z'\beta)^2 Z'|X, W = 0]\Big|W = 1\Big]}{1 + M}.$$

Notice that $J_r = E^*[Z(Y - Z'\beta)^2 Z]$. That is, J_r is equal to the inner matrix of the EHW asymptotic variance when data are i.i.d. with distribution P^*. However, since the matched sample $\mathcal{S}^*$ is not an i.i.d. sample from P^*, $\widehat{J}_r$ is not generally consistent for J. The difference between the limit of the OLS standard errors $\widehat{H}^{-1}\widehat{J}_r\widehat{H}^{-1}$ and the actual asymptotic variance $H^{-1}JH^{-1}$ is given by $H^{-1}\Delta H^{-1}$, where

$$\Delta = \frac{-ME\big[\Gamma_0(X)\Gamma_1(X)' + \Gamma_1(X)\Gamma_0(X)'|W = 1\big] - (M-1)ME\big[\Gamma_0(X)\Gamma_0(X)'|W = 1\big]}{M+1}, \tag{7}$$

and

$$\Gamma_w(x) = E\big[Z(Y - Z'\beta)|X = x, W = w\big],$$

for $w = 0, 1$.

Therefore, bias in the estimation of the variance may arise when $\Gamma_0(X) \neq 0$. The following example provides a simple instance of this bias.

Example 1: *Inconsistency of OLS standard errors*
Assume the sample is drawn from

$$Y = \tau W + X + \varepsilon, \tag{8}$$

where X is a scalar, $E[X] = E[\varepsilon] = 0$, and W and X are independent of ε. Assume that we match the values of X for N_1 treated units to N_1 untreated units ($M = 1$) without replacement. Let $j(i)$ be the index of the untreated observation that serves as a match for treated observation i. For simplicity, suppose that all matches are perfect, so $X_i = X_{j(i)}$, for every treated unit i so we can ignore potential biases generated by matching discrepancies. Within the matched sample, $\mathcal{S}^*$, we run a linear regression of Y on $Z = (1, W)'$ to obtain the regression coefficient on W,

$$\widehat{\tau} = \frac{1}{N_1}\sum_{i=1}^{N} W_i(Y_i - Y_{j(i)}).$$

$\widehat{\tau}$ is the usual matching estimator for the average effect of the treatment on the treated. Notice that $Y_i - Y_{j(i)} = \tau + \varepsilon_i - \varepsilon_{j(i)}$. Because variation in X is taken care of through matching, all variation in $\widehat{\tau}$ comes through the error terms. Because $n = 2N_1$, it follows that

$$n \operatorname{var}(\widehat{\tau}) = 4\operatorname{var}(\varepsilon).$$

Consider now the residuals of the OLS regression of Y_{ni} on a constant and W_{ni} in the matched sample:

$$\widehat{\varepsilon}_{ni} = Y_{ni} - \widehat{\mu} - \widehat{\tau} W_{ni} \approx X_{ni} + \varepsilon_{ni},$$

where $\widehat{\mu}$ is the intercept of the sample regression line. For this simple case, the OLS (EHW) variance estimator for $\widehat{\tau}$ is

$$n \, \widehat{\operatorname{var}}(\widehat{\tau}) = \frac{4}{n} \sum_{i=1}^{n} \widehat{\varepsilon}_{ni}^2 \approx 4\big(\operatorname{var}(X) + \operatorname{var}(\varepsilon)\big).$$

That is, in this example, OLS standard errors overestimate the variance of $\widehat{\tau}$ because they do not take into account the correlation generated by X between the regression residuals of the treated units and their match. □

The following example shows, however, that OLS standard errors that ignore the matching step may also *underestimate* the variance.

Example 2: *Underestimation of the variance*
In the same setting as Example 1, assume that data is generated by

$$Y = \tau W + X - 2WX + \varepsilon. \tag{9}$$

The post-matching estimator of τ from a regression of Y on $(1, W)'$ is

$$\widehat{\tau} = \frac{1}{N_1} \sum_{i=1}^{n} W_i (Y_i - Y_{j(i)}).$$

In this case, $Y_i - Y_{j(i)} = \tau - 2X + \varepsilon_i - \varepsilon_{j(i)}$. Therefore,

$$n \operatorname{var}(\widehat{\tau}) = 8\operatorname{var}(X) + 4\operatorname{var}(\varepsilon).$$

OLS standard errors are based on residuals,

$$\widehat{\varepsilon}_{ni} = Y_{ni} - \widehat{\mu} - \widehat{\tau} W_{ni} \approx X_i - 2W_{ni}X_{ni} + \varepsilon_{ni} = \begin{cases} -X_{ni} + \varepsilon_{ni} & \text{if} \quad W_{ni} = 1, \\ X_{ni} + \varepsilon_{ni} & \text{if} \quad W_{ni} = 0. \end{cases}$$

As a result, we obtain

$$n\widehat{\text{var}}(\widehat{\tau}) \approx 4\big(\text{var}(X) + \text{var}(\varepsilon)\big).$$

In this example, the OLS variance estimator does not take into account the heterogeneity in the treatment effects generated by X, underestimating the variance of $\widehat{\tau}$. □

OLS standard errors would be valid in examples 1 and 2 if the specifications for the post-matching regressions included the terms containing X in equations (8) and (9), respectively. Indeed, OLS standard errors are generally valid if the regression is correctly specified in a specific sense defined in the following result.

Proposition 4 (Validity of OLS standard errors under correct specification). *Assume that the post-matching regression,*

$$Y = Z'\beta + \varepsilon,$$

is correctly specified with respect to the conditional distribution of Y given (Z, X, W). That is, with $E[\varepsilon|Z, X, W] = 0$. Then, $J_r = J$, and the EHW variance estimator, $\widehat{H}^{-1}\widehat{J}_r\widehat{H}^{-1}$, is consistent for the asymptotic variance of $\sqrt{n}(\widehat{\beta} - \beta)$.

Notice, however, that correct specification is precisely the condition under which matching would not be required to obtain a consistent estimator of β, since direct estimation without matching would be valid. Moreover, a correct specification (in the sense defined above) of the post-matching regression is not required for consistent estimation of causal parameters. For example, under regularity conditions, a simple difference in means between the treated and a matched sample of untreated units is consistent for the average effect of the treatment on the treated. Moreover, consistent estimators of the variance exist for the simple difference in means. These variance estimators are different from the OLS variance estimator, and do not rely on correct specification of the post-matching regression (see Abadie and Imbens, 2006).

Finally, Equation (7) implies that the conditions of Proposition 4 can be slightly weakened to require only that the regression function is correctly specified among the non-treated, in the sense that $E[\varepsilon|Z, X, W = 0] = 0$. This is because for the estimators studied in this article, matching affects only the distribution of the covariates for the non-treated. In addition, for the special case $M = 1$, it is sufficient that the regression function is correctly specified among the treated, in the sense that $E[\varepsilon|Z, X, W = 1] = 0$.

3.2 Match-Level Clustered Standard Errors

We have shown that OLS standard errors are not generally valid for the post-matching least squares estimator. In this section, we will demonstrate that, when matching is done without replacement, clustered standard errors (Liang and Zeger, 1986; Arellano, 1987) can be employed to obtain valid estimates of the standard deviation of post-matching regression coefficients. In particular, we will consider standard errors clustered at the level of the match sets.

Consider an estimator of the asymptotic variance of $\widehat{\beta}$ given by $\widehat{H}^{-1}\widehat{J}\widehat{H}^{-1}$, where $\widehat{H}$ is as in Equation (5) and $\widehat{J}$ is given by the clustered variance formula applied to the match sets,

$$\widehat{J} = \frac{1}{n}\sum_{i=1}^{n} W_i\Big(Z_i(Y_i - Z_i'\widehat{\beta}) + \sum_{j\in\mathcal{J}(i)} Z_j(Y_j - Z_j'\widehat{\beta})\Big) \times \Big(Z_i(Y_i - Z_i'\widehat{\beta}) + \sum_{j\in\mathcal{J}(i)} Z_j(Y_j - Z_j'\widehat{\beta})\Big)'.$$

Clustered standard errors can be readily implemented using standard statistical software. The next result shows that match-level clustered standard errors are valid in large samples for the post-matching estimator (provided matching is done without replacement).

Proposition 5 (Validity of clustered standard errors). *Under the assumptions of Proposition 3 we obtain that*

$$\widehat{J} \xrightarrow{p} J.$$

In particular, the clustered estimator of the variance is consistent, i.e.,

$$\widehat{H}^{-1}\widehat{J}\widehat{H}^{-1} - n\text{var}(\widehat{\beta}) \xrightarrow{p} 0.$$

The intuition behind this result is that matching on covariates makes regression errors statistically dependent among units in the same match sets, $\{i\} \cup \mathcal{J}(i)$, $i = 1, \ldots, N_1$. Standard errors clustered at the level of the match set take this dependency into account.

3.3 Matched Bootstrap

Proposition 5 shows that clustered standard errors are valid for the asymptotic variance of the post-matching estimator. In this section, we show that a clustered version of the nonparametric bootstrap (Efron, 1979) is also valid. This version of the bootstrap relies on resampling of match sets instead on individual observations.

Recall that we reordered the observations in our sample, so that the first N_1 observations are the treated. Consider the nonparametric bootstrap that samples treated units together with their M matches partners from $\mathcal{S}^*$ to obtain

$$\widehat{\beta}^* = \left(\frac{1}{n}\sum_{i=1}^{n} V_{ni} Z_{ni} Z_{ni}'\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} V_{ni} Z_{ni} Y_{ni}$$

where $(V_{n1}, \ldots, V_{nN_1})$ has a multinomial distribution with parameters $(N_1, (1/N_1, \ldots, 1/N_1))$, and $V_{nj} = V_{ni}$ if $j > N_1$ and $j \in \mathcal{J}(i)$. In this bootstrap procedure, N_1 units are drawn at random with replacement from the N_1 treated sample units. Untreated units are drawn along with their treated match. Effectively, the matched bootstrap samples matched sets of one treated unit and M untreated units. The next proposition shows validity of the matched bootstrap.

Proposition 6 (Validity of the matched bootstrap). *Under the assumptions of Proposition 5, we have that*

$$\sup_{r \in \mathbb{R}^s} \left| P\left(\sqrt{n}(\widehat{\beta}^* - \widehat{\beta}) \leq r \middle| \mathcal{S}\right) - P(\mathcal{N}(0, H^{-1}JH^{-1}) \leq r) \right| \xrightarrow{p} 0.$$

Proposition 6 shows that the bootstrap distribution provides an asymptotically valid approximation of the limiting distribution of the post-matching estimator, but that does

not necessarily imply that the associated bootstrap variance is an asymptotically valid estimate of the variance of the estimator. Indeed, the analysis of the bootstrap variance is complicated by the fact that, in forming the bootstrap estimate $\widehat{\beta}^*$, the empirical analog

$$\widehat{H}^* = \frac{1}{n}\sum_{i=1}^{n} V_{ni} Z_{ni} Z_{ni}'$$

of the Hessian H for a given bootstrap draw may be badly conditioned or even non-invertible, which happens with positive probability at any given sample size. To circumvent this issue, we fix constants $c > 0$ and $\alpha \in (0, 1/2)$ and consider the alternative bootstrap estimator

$$\tilde{\beta}^* = \begin{cases} \widehat{\beta}^* & \text{if } \|\widehat{H}^* - \widehat{H}\| \leq c/n^\alpha, \\ \widehat{\beta} & \text{otherwise.} \end{cases}$$

In words, this modified bootstrap estimator coincides with the matched bootstrap estimator whenever the bootstrap Hessian, $\widehat{H}^*$, is close to $\widehat{H}$ in the full matched sample. For the other bootstrap draws, the modified bootstrap estimator is equal to the post-matching estimator. The threshold is chosen such that, as the sample size grows, the two bootstrap estimators coincide with probability approaching one.

We establish that $\tilde{\beta}^*$ allows for valid inference in large samples, including the consistent estimation of standard errors:

Proposition 7 (Validity of bootstrap standard errors). *Under the assumptions of Proposition 5 and $E[\|Z\|^8|W = w, X = x]$ uniformly bounded on $\mathcal{X}_w$, the bootstrap distribution given by $\tilde{\beta}^*$ is valid in the sense of Proposition 6, and yields a valid estimate of the asymptotic variance of $\widehat{\beta}$, i.e.*

$$n\text{var}(\tilde{\beta}^*|\mathcal{S}) \xrightarrow{p} H^{-1}JH^{-1}$$

as $n \to \infty$.

4 Simulations

In this section, we study the performance of the post-matching standard error estimators from Section 3 in a simulation exercise using two data generating processes (DGP).

4.1 DGP1: Robustness to Misspecification

Let $\mathcal{U}(a, b)$ be the Uniform distribution on $[a, b]$. We generate data according to

$$Y = WX + 5X^2 + \varepsilon,$$

where $X|W = 1 \sim \mathcal{U}(-1, 1)$, $X|W = 0 \sim \mathcal{U}(-1, 2)$ and $\varepsilon \sim \mathcal{N}(0, 1)$. We sample $N_1 = 50$ treated and $N_0 = 200$ nontreated units. We first match treated and untreated units on the covariates, X, without replacement and with $M = 1$ match per treated unit. We consider the following post-matching regression specifications.

Specification 1:

$$Y = \alpha + \tau_0 W + \tau_1 WX + \beta_1 X + \varepsilon$$

Specification 2:

$$Y = \alpha + \tau_0 W + \tau_1 WX + \beta_1 X + \beta_2 X^2 + \varepsilon$$

Specification 2 is correct relative to the conditional expectation $E[Y|X, W]$, while specification 1 is not. Regression estimands can always be seen as L_2 approximations to $E[Y|W, X]$, regardless of the specification adopted for estimation (see, e.g., White, 1980b). For our simulation results, we will focus on estimators of τ_0 and τ_1, the regression coefficients on terms involving W. For the DGP in this simulation (DPG1), $\tau_0 = 0$ and $\tau_1 = 1$ under the matching target distribution.

Table 1 reports the results of the simulation exercise. In a regression that uses the full sample without matching, the estimates of τ_0 and τ_1 are biased under misspecification (specification 1), while they are valid under correct specification (specification 2). After matching, both specifications yield valid estimates for τ_0 and τ_1. However, OLS standard error estimates are inflated under misspecification, while average clustered and matched bootstrap standard errors (with 1000 bootstrap draws) closely approximate the standard deviation of $\widehat{\tau}_0$ and $\widehat{\tau}_1$. Under correct specification (specification 2), all standard error estimates perform well.

Table 1: Monte Carlo results for DGP1 (10000 iterations)

(a) Target parameter: coefficient $\tau_0 = 0$ on W

	full sample		post-matching		average standard error		
specification	mean of $\widehat{\tau}_0$	std. of $\widehat{\tau}_0$	mean of $\widehat{\tau}_0$	std. of $\widehat{\tau}_0$	OLS	cluster	bootstrap
1	−0.85	0.404	0.00	0.204	0.359	0.197	0.199
2	0.00	0.165	0.00	0.204	0.196	0.196	0.199

(b) Target parameter: coefficient $\tau_1 = 1$ on the interaction WX

	full sample		post-matching		average standard error		
specification	mean of $\widehat{\tau}_1$	std. of $\widehat{\tau}_1$	mean of $\widehat{\tau}_1$	std. of $\widehat{\tau}_1$	OLS	cluster	bootstrap
1	−4.00	0.646	0.99	0.358	0.728	0.340	0.348
2	1.00	0.286	1.00	0.356	0.337	0.338	0.346

4.2 DGP2: High Treatment-Effect Heterogeneity

In the simulation in the previous section, OLS standard errors overestimate the variation of the post-matching estimator under misspecification. In this section, we present an example in which OLS standard errors are too small. We generate data according to

$$Y = WX + 20WX^2 - 10X^2 + \varepsilon$$

with $\varepsilon \sim \mathcal{N}(0, 1)$ as above. According to this data-generating process (DGP2), the conditional treatment effect is non-linear with

$$E[Y|W = 1, X] - E[Y|W = 0, X] = X + 20X^2.$$

Sample sizes, matching settings, and regression specifications are as in DGP1. Notice that both regression specifications are now misspecified, as they cannot capture non-linear conditional treatment effects. Like in Section 4.1, regression coefficients represent the parameters of an L_2 approximation to $E[Y|W, X]$ over the distribution of (W, X) in Proposition 1. Di-

rect calculations yield $\tau_0 = 20/3$ and $\tau_1 = 1$ for both specifications in the matching target distribution.

Table 2: Simulation results for 10,000 Monte Carlo iterations for DGP2

(a) Target parameter: coefficient $\tau_0 = 6.67$ on W

	full sample		post-matching		average standard error		
specification	mean of $\widehat{\tau}_0$	std. of $\widehat{\tau}_0$	mean of $\widehat{\tau}_0$	std. of $\widehat{\tau}_0$	OLS	cluster	bootstrap
1	8.25	0.754	6.55	0.883	0.630	0.869	0.897
2	6.70	0.857	6.55	0.883	0.630	0.869	0.897

(b) Target parameter: coefficient $\tau_1 = 1$ on the interaction WX

	full sample		post-matching		average standard error		
specification	mean of $\widehat{\tau}_1$	std. of $\widehat{\tau}_1$	mean of $\widehat{\tau}_1$	std. of $\widehat{\tau}_1$	OLS	cluster	bootstrap
1	11.00	1.209	1.01	1.950	1.330	1.848	1.932
2	1.90	1.877	1.01	1.950	1.330	1.848	1.933

Table 2 presents the results of the simulation exercise for DGP2. The large heterogeneity in conditional treatment effects is not captured by either regression specification, and OLS standard errors that ignore the matching step underestimate the variation of the post-matching estimator. In contrast, the robust standard errors proposed in this article closely reflect the variability of the post-matching estimators.

5 Application

This section reports the results of an empirical application where we look at the effect of smoking on the pulmonary function of youths. The application is based on data originally collected in Boston, Massachusetts, by Tager et al. (1979, 1983), and subsequently described and analyzed in Rosner (1995) and Kahn (2005). The sample contains 654 youth, $N_1 = 65$ who have ever smoked regularly ($W = 1$) and $N_0 = 589$ who never smoked regularly

($W = 0$). The outcome of interest is the subjects' forced expiratory volume (Y), ranging from 0.791 to 5.793 liters per second (ℓ/sec). In addition, we use data on age (X_1, ranging from 3 to 19 with the youngest ever-smoker aged 9) and gender (X_2, with $X_2 = 1$ for males and $X_2 = 0$ for females).

The use of matching to study the causal effect of smoking is motivated by the likely confounding effects of age and gender. For instance, while the causal effect of smoking on respiratory volume is expected to be negative, older children are more likely to smoke and have a larger respiratory volume, which induces a positive association between smoking and respiratory volume.

We first match every smoker in the sample to a non-smoker ($M = 1$), without replacement, based on age (X_1) and gender (X_2). Within the resulting matched sample of 65 smokers and 65 non-smokers, we run linear regressions with the following specifications:

Specification 1:

$$Y = \alpha + \tau_0 W + \varepsilon.$$

Specification 2:

$$Y = \alpha + \tau_0 W + \beta_1 X_1 + \beta_2 X_2 + \varepsilon.$$

Specification 3:

$$\begin{aligned} Y = \alpha + \tau_0 W + \tau_1 W(X_1 - E[X_1]) + \tau_2 W(X_2 - E[X_2]) \\ + \beta_1(X_1 - E[X_1]) + \beta_2(X_2 - E[X_2]) + \varepsilon. \end{aligned}$$

The first specification yields the matching estimator for the average treatment effect τ_0 as the regression coefficient on W, while the second adds linear controls in X_1 and X_2. The third specification also includes interaction terms of smoking with age and gender.

Table 3 reports regression estimates of τ_0, τ_1 and τ_2 along with standard errors (regression coefficients on terms not involving W are omitted from the Table 3 for brevity). The first specification demonstrates the confounding problem in this application. Without controlling for age and gender, there is a positive correlation between smoking and forced

Table 3: OLS and post-matching estimates for the smoking data set

dependent variable: forced expiratory volume

	explanatory variables								
	smoker			smoker×age			smoker×male		
	coeff.	std. error		coeff.	std. error		coeff.	std. error	
		OLS	clust		OLS	clust		OLS	clust
Specification 1:									
OLS	.711	.099							
post-matching	−.066	.132	.095						
Specification 2:									
OLS	−.154	.104							
post-matching	−.077	.104	.096						
Specification 3:									
OLS	.495	.187		−.182	.036		.461	.193	
post-matching	−.077	.102	.093	−.092	.054	.038	−.021	.249	.212

expiratory function. After matching on age and gender, the sign of the regression coefficient on smoking becomes negative. In this specification, the clustered standard error for the post-matching estimate is considerably smaller than the corresponding OLS standard error.

Specification 2 includes linear controls for age and gender. The sign and magnitude of the OLS estimate of the coefficient on the smoker variable changes substantially between specifications 1 and 2, while the magnitude of the post-matching estimate stays roughly constant. This result illustrates the higher robustness across specifications of the post-matching estimator relative to OLS (Ho et al., 2007). When specification 2 is adopted for regression, the sign of the coefficient on the smoker variable is not affected by matching, and clustered standard errors are similar to OLS standard errors. Both findings are consistent with the adopted regression specification moving closer towards the correct specification of $E[Y|W, X_1, X_2]$.

In specification 3, which includes interactions between the smoker variable and age and gender, the use of matching and the use of robust standard errors matters for the substantive results of the analysis. First, notice that the coefficient on the interaction

of gender with treatment is large, significant and positive without matching, suggesting that the effect of smoking is more severe for girls than for boys. After matching, the sign changes, and the estimated coefficient is small and insignificant. This suggests that the large interaction finding with OLS for this coefficient is caused by misspecification. Second, in the post-matching regression we find a negative estimate for the interaction of treatment with age. With OLS standard errors, this effect is not significant (at the 5% level). The robust standard errors proposed in this article are smaller (conceivably, because of large coefficient heterogeneity) and result in a rejection of the null hypothesis of a zero interaction coefficient between smoker and age (at the 5% level).

6 Conclusion

This article establishes valid inference in linear regression after nearest-neighbor matching without replacement. OLS standard errors that ignore the matching step are not generally valid if the regression specification is incorrect relative to the expectation of the outcome conditional on the treatment and the matching covariates. Notice, however, that using a correct specification relative to $E[Y|W, X]$ is not necessary to consistently estimate treatment parameters after matching. For example, a simple difference in means can identify the average treatment effect in a matched sample.

We propose two alternatives – standard errors clustered at the match level and an analogous block bootstrap – that are robust to misspecification and easily implementable with standard statistical software. A simulation study and an empirical example demonstrate the usefulness of our results.

To conclude, we outline potential extensions of our results. First, in this article we discuss only matching without replacement, and the results do not directly carry over to matching with replacement as in Abadie and Imbens (2006). Matching with replacement (that is, allowing nontreated units to be used as a match more than once) creates additional dependencies between match sets that are not reflected in OLS standard errors or in the robust standard errors proposed in this article. In addition, our analysis applies to the case when matching is done directly on the covariates, avoiding substantial complications

created by the presence of nuisance parameters in the matching step when matching is done on the estimated propensity score (see Rosenbaum and Rubin, 1983; Abadie and Imbens, 2016). Finally, our analysis assumes that the quality of matches is good enough for matching discrepancies not to bias the asymptotic distribution of the post-matching regression estimator. Post-matching regression adjustments may, in practice, help eliminate the bias as in the bias-corrected matching estimator in Abadie and Imbens (2011). These are angles that we do not explore in this article and interesting avenues for future research.

References

Abadie, A. and Imbens, G. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74(1):235–267.

Abadie, A. and Imbens, G. (2011). Bias-corrected matching estimators for average treatment effects. *Journal of Business & Economic Statistics*, 29(1):1–11.

Abadie, A. and Imbens, G. (2016). Matching on the estimated propensity score. *Econometrica*, 84(2):781–807.

Arellano, M. (1987). Computing robust standard errors for within-groups estimators. *Oxford Bulletin of Economics and Statistics*, 49(4):431–434.

Blinder, A. S. (1973). Wage discrimination: Reduced form and structural estimates. *Journal of Human Resources*, 8(4):436–455.

Cochran, W. G. (1953). Matching in analytical studies. *American Journal of Public Health and the Nation's Health*, 43(6 Pt 1):684–691.

Dehejia, R. H. and Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association*, 94(448):1053–1062.

DiNardo, J., Fortin, N., and Lemieux, T. (1996). Labor market institutions and the distribution of wages, 1973-1992: A semiparametric approach. *Econometrica*, 64(5):1001–1044.

Efron, B. (1979). Bootstrap methods: Another look at the Jackknife. *The Annals of Statistics*, 7(1):1–26.

Eicker, F. (1967). Limit theorems for regressions with unequal and dependent errors. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 59–82.

Ho, D. E., Imai, K., King, G., and Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*, 15(3):199–236.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 221–233.

Kahn, M. (2005). An exhalent problem for teaching statistics. *The Journal of Statistical Education*, 13(2).

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22.

Oaxaca, R. (1973). Male-female wage differentials in urban labor markets. *International Economic Review*, 14(3):693–709.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Rosner, B. (1995). *Fundamentals of Biostatistics*. Duxbury Press.

Rubin, D. B. (1973). Matching to remove bias in observational studies. *Biometrics*, 29(1):159–183.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688.

Tager, I. B., Weiss, S. T., Muñoz, A., Rosner, B., and Speizer, F. E. (1983). Longitudinal study of the effects of maternal smoking on pulmonary function in children. *New England Journal of Medicine*, 309(12):699–703.

Tager, I. B., Weiss, S. T., Rosner, B., and Speizer, F. E. (1979). Effect of parental cigarette smoking on the pulmonary function of children. *American Journal of Epidemiology*, 110(1):15–26.

White, H. (1980a). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48(4):817–838.

White, H. (1980b). Using least squares to approximate unknown regression functions. *International Economic Review*, 21(1):149–170.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1):1–25.

Appendix: Proofs

Preliminary lemmas A.1 and A.2 and propositions A.1-A.3 are in a supplementary appendix.

Proof of Proposition 1. Let $E_{Q(\cdot|W=1)}$ and $E_{Q(\cdot|W=0)}$ be expectation operators for $Q(\cdot|W=1)$ and $Q(\cdot|W=0)$. Notice first that for any measurable function q,

$$E_{Q(\cdot|W=1)}[q(Y(1),S)] = E[q(Y,S)|W=1] \tag{A.1}$$

The result holds also replacing $W=1$ with $W=0$, and after conditioning on X. In particular,

$$E_{Q(\cdot|W=0)}[q(Y(0),S)|X] = E[q(Y,S)|X,W=0]. \tag{A.2}$$

The regression coefficient in the population defined by (a), (b) is the minimizer of

$$\frac{1}{M+1}E_{Q(\cdot|W=1)}[(Y(1)-g(1,S)'b)^2] + \frac{M}{M+1}E_{Q(\cdot|W=1)}[(Y(0)-g(0,S)'b)^2].$$

Notice that,

$$\begin{aligned} E_{Q(\cdot|W=1)}[(Y(1) - g(1,S)'b)^2] &= E[(Y - g(1,S)'b)^2|W=1] \\ &= E^*[(Y - Z'b)^2|W=1], \end{aligned}$$

where the first equality follows from Equation (A.1) and the second equality follows from the definitions of $P^*(\cdot|W=1)$ and Z. Similarly,

$$\begin{aligned} E_{Q(\cdot|W=1)}[(Y(0) - g(0,S)'b)^2] &= E_{Q(\cdot|W=1)}[E_{Q(\cdot|W=1)}[(Y(0) - g(0,S)'b)^2|X]] \\ &= E_{Q(\cdot|W=1)}[E_{Q(\cdot|W=0)}[(Y(0) - g(0,S)'b)^2|X]] \\ &= E[E[(Y - g(W,S)'b)^2|X, W=0]|W=1] \\ &= E^*[(Y - Z'b)^2|W=0]. \end{aligned}$$

In the last equation, the first equality follows from the law of iterated expectations, the second equality follows from selection on observables, the third equality follows from (A.2) and (A.1), and the last equation follows from the definition of $P^*(\cdot|W=0)$. Therefore, we obtain

$$\begin{aligned} &\frac{1}{M+1} E_{Q(\cdot|W=1)}[(Y(1) - g(1,S)'b)^2] + \frac{M}{M+1} E_{Q(\cdot|W=1)}[(Y(0) - g(0,S)'b)^2] \\ &\qquad = \frac{1}{M+1} E^*[(Y - Z'b)^2|W=1] + \frac{M}{M+1} E^*[(Y - Z'b)^2|W=0] \\ &\qquad = E^*[(Y - Z'b)^2], \end{aligned}$$

which implies the result of the proposition. □

Proof of Proposition 2. By Lemma A.1,

$$\frac{1}{n} \sum_{i \in \mathcal{S}^*} Z_i Z_i' \xrightarrow{p} H;$$

by Lemma A.2,

$$\widehat{H} \sqrt{n} \left(\widehat{\beta} - \beta \right) = \sqrt{n} \left(\frac{1}{n} \sum_{i \in \mathcal{S}^*} (Z_i Y_i - Z_i Z_i' \beta) \right) \xrightarrow{d} \mathcal{N}(\mathbf{0}, J),$$

where we note that

$$E[ZY - ZZ'\beta | W = 0, X = x]$$

is Lipschitz. Hence,

$$\sqrt{n}\left(\widehat{\beta} - \beta\right) = \overbrace{\widehat{H}^{-1}}^{\xrightarrow{p} H^{-1}} \underbrace{\widehat{H}\sqrt{n}\left(\frac{1}{n}\sum_{i \in \mathcal{S}^*}(Z_i Y_i - Z_i Z_i'\beta)\right)}_{\xrightarrow{d}\mathcal{N}(\mathbf{0}, J)} \xrightarrow{d} \mathcal{N}(\mathbf{0}, H^{-1}JH^{-1}).$$

□

Proof of Proposition 3. We have that

$$\begin{aligned}
\widehat{J}_r &= \frac{1}{n}\sum_{i=1}^{n} Z_i(Y_i - Z_i'\widehat{\beta})^2 Z_i' \\
&= \frac{1}{n}\sum_{i=1}^{n} Z_i(Y_i - Z_i'\beta)^2 Z_i' + \frac{1}{n}\sum_{i=1}^{n} Z_i\Big((Y_i - Z_i'\widehat{\beta})^2 - (Y_i - Z_i'\beta)^2\Big) Z_i'.
\end{aligned}$$

Notice that

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} Z_i\Big((Y_i - Z_i'\widehat{\beta})^2 - (Y_i - Z_i'\beta)^2\Big) Z_i' \\
&\qquad = (\widehat{\beta} - \beta)'\left(\frac{1}{n}\sum_{i=1}^{n} Z_i(Z_i'Z_i)Z_i'(\widehat{\beta} + \beta) - 2\frac{1}{n}\sum_{i=1}^{n} Z_i(Z_i'Z_i)Y_i\right).
\end{aligned}$$

By assumption, the functions

$$E[\|Z\|^4 | X = x, W = w] \quad \text{and} \quad E[|Y|^4 | X = x, W = w]$$

are uniformly bounded on $\mathcal{X}_w$, for $w = 0, 1$. By Hölder's Inequality, this implies finiteness of

$$E\left[\left\|\frac{1}{n}\sum_{i=1}^{n} Z_i Z_i' Z_i Z_i'\right\|\right] \quad \text{and} \quad E\left[\left\|\frac{1}{n}\sum_{i=1}^{n} Z_i Z_i' Z_i Y_i'\right\|\right].$$

Then, for $\epsilon \in (0, 1/2)$, by Markov's Inequality, we obtain

$$\frac{1}{n}\sum_{i=1}^{n} Z_i((Y_i - Z_i'\widehat{\beta})^2 - (Y_i - Z_i'\beta)^2) Z_i'$$

$$= n^{1/2-\epsilon}(\widehat{\beta}-\beta)' \left(\frac{\sum_{i=1}^n Z_i(Z_iZ_i')Z_i'/n}{n^{1/2-\epsilon}}(\widehat{\beta}+\beta) - \frac{2\sum_{i=1}^n Z_i(Z_iZ_i')Y_i/n}{n^{1/2-\epsilon}} \right) \xrightarrow{p} 0.$$

As a result,

$$\widehat{J}_r = \frac{1}{n}\sum_{i=1}^n Z_i(Y_i - Z_i'\beta)^2 Z_i' + o_p(1),$$

and the claim follows from Lemma A.1. □

Proof of Proposition 4. Under correct specification, we find that

$$\begin{aligned}\Gamma_W(X) &= E[Z(Y-Z'\beta)|W,X] = E[Z\varepsilon|W,X] \\ &= E[E[Z\varepsilon|Z,W,X]|W,X] \\ &= E[Z\underbrace{E[\varepsilon|Z,W,X]}_{=0}|W,X] = 0.\end{aligned}$$

□

Proof of Proposition 5. First, note that

$$\begin{aligned}\widehat{J} =& \frac{1}{n}\sum_{W_i=1}\left(Z_i(Y_i-Z_i'\beta)+\textstyle\sum_{j\in\mathcal{J}(i)}Z_j(Y_j-Z_j'\beta)\right)\left(Z_i(Y_i-Z_i'\beta)+\textstyle\sum_{j\in\mathcal{J}(i)}Z_j(Y_j-Z_j'\beta)\right)' \\ &+ o_P(1),\end{aligned}$$

where we replace $\widehat{\beta}$ by β analogous to the proof of Proposition 3.

Write

$$G = Z(Y-Z'\beta) \qquad\qquad \Gamma_w(x) = E[Z(Y-Z'\beta)|W=w,X=x].$$

Note that $\Gamma_0(x)$ is Lipschitz on $\mathcal{X}$, and that G_i has uniformly bounded fourth moments. We decompose

$$\begin{aligned}\widehat{J} =& \frac{1}{n}\sum_{W_i=1}\left(G_i+\textstyle\sum_{j\in\mathcal{J}(i)}G_j\right)\left(G_i+\textstyle\sum_{j\in\mathcal{J}(i)}G_j\right)' + o_P(1) \\ =& \frac{1}{n}\sum_{W_i=1}\left(\Gamma_1(X_i)+M\Gamma_0(X_i)\right)\left(\Gamma_1(X_i)+M\Gamma_0(X_i)\right)' \\ &+ \frac{1}{n}\sum_{i\in\mathcal{S}^*}\left(G_i-\Gamma_{W_i}(X_i)\right)\left(G_i-\Gamma_{W_i}(X_i)\right)'\end{aligned}$$

$$
\begin{aligned}
&+ \frac{1}{n} \sum_{W_i=1} \sum_{\ell \neq \ell' \in \mathcal{J}(i) \cup \{i\}} (G_\ell - \Gamma_{W_\ell}(X_\ell)) \left(G_{\ell'} - \Gamma_{W_{\ell'}}(X_{\ell'})\right)' \\
&+ \frac{1}{n} \sum_{W_i=1} \Big((\Gamma_1(X_i) + M\Gamma_0(X_i)) \Big(G_i - \Gamma_1(X_i) + \textstyle\sum_{j \in \mathcal{J}(i)} (G_j - \Gamma_0(X_j))\Big)' \\
&\quad + \Big(G_i - \Gamma_1(X_i) + \textstyle\sum_{j \in \mathcal{J}(i)} (G_j - \Gamma_0(X_i))\Big) (\Gamma_1(X_i) + M\Gamma_0(X_j))' \Big) + o_P(1).
\end{aligned}
$$

Here, the o_P terms absorb the deviation due to using $\widehat{\beta}$ instead of β, as well as the matching discrepancies in the conditional expectations.

The first sum is i.i.d. with

$$
\begin{aligned}
&\frac{1}{n} \sum_{W_i=1} (\Gamma_1(X_i) + M\Gamma_0(X_i)) (\Gamma_1(X_i) + M\Gamma_0(X_i))' \\
&\xrightarrow{p} \frac{E\left[(\Gamma_1(X) + M\Gamma_0(X))(\Gamma_1(X) + M\Gamma_0(X))' | W = 1\right]}{1+M} \\
&= \frac{\mathrm{var}(\overbrace{\Gamma_1(X) + M\Gamma_0(X)}^{E[\cdot|W=1]=\mathbf{0}} | W = 1)}{1+M},
\end{aligned}
$$

while the second is a martingale with

$$
\begin{aligned}
&\frac{1}{n} \sum_{i \in \mathcal{S}^*} (G_i - \Gamma_{W_i}(X_i)) (G_i - \Gamma_{W_i}(X_i))' \\
&\xrightarrow{p} \frac{E[\mathrm{var}(Z(Y - Z'\beta) | W = 1, X) + M \mathrm{var}(Z(Y - Z'\beta) | W = 0, X) | W = 1]}{1+M}
\end{aligned}
$$

by Lemma A.1. Under appropriate reordering of the individual increments, all other sums can be represented as averages of mean-zero martingale increments; since the second moments of the increments are uniformly bounded, they vanish asymptotically. □

Proof of Proposition 6. Write

$$
\widehat{H}^* = \frac{1}{n} \sum_{i=1}^{n} V_{ni} Z_{ni} Z_{ni}'.
$$

Note first that

$$
H^{-1} \sqrt{n} (\widehat{H}^* (\widehat{\beta}^* - \beta) - \widehat{H}(\widehat{\beta} - \beta)) = H^{-1} \sqrt{n} \left(\frac{1}{n} \sum_{i=1}^{n} (V_{ni} - 1) Z_{ni} (Y_{ni} - Z_{ni}' \beta) \right)
$$

$$\xrightarrow{d} \mathcal{N}(\mathbf{0}, H^{-1}JH^{-1}),$$

conditional on $\mathcal{S}$, by Proposition A.2. Now,

$$\begin{aligned}
&\sqrt{n}(\widehat{\beta}^* - \widehat{\beta}) \\
&= (\widehat{H}^*)^{-1}H(H^{-1}\sqrt{n}(\widehat{H}^*(\widehat{\beta}^* - \beta) - \widehat{H}^*(\widehat{\beta} - \beta)) \\
&= \underbrace{(\widehat{H}^*)^{-1}H}_{\xrightarrow{p}\mathbb{I}}(H^{-1}\sqrt{n}(\widehat{H}^*(\widehat{\beta}^* - \beta) - \widehat{H}(\widehat{\beta} - \beta))) + \underbrace{((\widehat{H}^*)^{-1}\widehat{H} - \mathbb{I})}_{\xrightarrow{p}\mathbb{O}}\sqrt{n}(\widehat{\beta} - \beta) \\
&\xrightarrow{d} \mathcal{N}(\mathbf{0}, H^{-1}JH^{-1}),
\end{aligned}$$

conditional on $\mathcal{S}$, where we have used that $\widehat{H}^* - \widehat{H} \xrightarrow{p} \mathbb{O}$ conditional on $\mathcal{S}$. □

In the proof of the consistency of bootstrap standard errors (Proposition 7), we will use an auxiliary result on the relationship of the expectation of the limit and the limit of expectations. Specifically, the following result shows that *conditional* convergence in distributions implies that *conditional* moments can only deviate towards the tails. The case where all σ-algebras are trivial (minimal) recovers the standard result that $\liminf_{n\to\infty} E|X_n| \geq E|X|$ for $X_n \xrightarrow{d} X$.

Proof of Proposition 7. First, $P(\tilde{\beta}^* = \widehat{\beta}^*|\mathcal{S}) \geq P(\|\widehat{H}^* - \widehat{H}\| \leq \frac{c}{n^\alpha}|\mathcal{S}) \xrightarrow{p} 1$ as $n \to \infty$. Indeed, since Z has bounded conditional eighth moments, we also have that $E[\|ZZ'\|^4|W = w, X = s]$ is uniformly bounded in X_w. It follows with Proposition A.2 that

$$\sup_{r\in\mathbb{R}^{(\dim Z)^2}} \left|P(\sqrt{n}\,\mathrm{vec}(\widehat{H}^* - \widehat{H}) \leq r|\mathcal{S}) - P(\mathcal{N}(\mathbf{0}, \Sigma_H) \leq r)\right| \xrightarrow{p} 0$$

as $n \to \infty$ and thus in particular

$$P(n^\alpha\|\widehat{H}^* - \widehat{H}\| \leq c|\mathcal{S}) \xrightarrow{p} 1$$

for all $\alpha \in (0, 1/2), c > 0$.

Second, since for $\tilde{A} \cap B = A \cap B$ generally

$$|P(A) - P(\tilde{A})| \leq \underbrace{|P(A \cap B) - P(\tilde{A} \cap B)|}_{=0} + \underbrace{|P(A \cap B^c) - P(\tilde{A} \cap B^c)|}_{\leq P(B^c)} \leq 1 - P(B),$$

for $\Phi(r) = P\left(\mathcal{N}(0, H^{-1}JH^{-1}) \leq r\right)$ we have specifically that

$$\begin{aligned}
&\sup_{r\in\mathbb{R}^s}\left|P\left(\sqrt{n}(\tilde{\beta}^* - \widehat{\beta}) \leq r\middle|\mathcal{S}\right) - \Phi(r)\right| \\
&\leq \sup_{r\in\mathbb{R}^s}\left(\left|P\left(\sqrt{n}(\widehat{\beta}^* - \widehat{\beta}) \leq r\middle|\mathcal{S}\right) - \Phi(r)\right| + \underbrace{\left|P\left(\sqrt{n}(\widehat{\beta}^* - \widehat{\beta}) \leq r\middle|\mathcal{S}\right) - P\left(\sqrt{n}(\tilde{\beta}^* - \widehat{\beta}) \leq r\middle|\mathcal{S}\right)\right|}_{\leq 1 - P(\tilde{\beta}^* = \widehat{\beta}^*|\mathcal{S})}\right) \\
&\leq \underbrace{\sup_{r\in\mathbb{R}^s}\left|P\left(\sqrt{n}(\widehat{\beta}^* - \widehat{\beta}) \leq r\middle|\mathcal{S}\right) - \Phi(r)\right|}_{\xrightarrow{p}0} + \underbrace{1 - P(\tilde{\beta}^* = \widehat{\beta}^*|\mathcal{S})}_{\xrightarrow{p}0} \xrightarrow{p} 0.
\end{aligned}$$

This shows that this alternative bootstrap is valid in the sense of Proposition 6.

Third, for the bootstrap variance, we find

$$\begin{aligned}
\widehat{\beta}^* - \widehat{\beta} &= \left(\widehat{H}^*\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n V_{ni}Z_{ni}Y_{ni} - \widehat{H}^*\widehat{\beta}\right) \\
&= \left(\widehat{H}^*\right)^{-1}\frac{1}{n}\sum_{i=1}^n V_{ni}Z_{ni}(Y_{ni} - Z_{ni}'\widehat{\beta}) \\
&= \underbrace{\widehat{H}^{-1}\frac{1}{n}\sum_{i=1}^n V_{ni}Z_{ni}(Y_{ni} - Z_{ni}'\widehat{\beta})}_{=\widehat{\Delta}^*} + \underbrace{\left(\left(\widehat{H}^*\right)^{-1} - \widehat{H}^{-1}\right)\frac{1}{n}\sum_{i=1}^n V_{ni}Z_{ni}(Y_{ni} - Z_{ni}'\widehat{\beta})}_{=\widehat{R}^*}
\end{aligned}$$

Note first that since $\frac{1}{n}\sum_{i=1}^n Z_{ni}(Y_{ni} - Z_{ni}'\widehat{\beta}) = 0$ and thus $n\text{var}\left(\frac{1}{n}\sum_{i=1}^n V_{ni}Z_{ni}(Y_{ni} - Z_{ni}'\widehat{\beta})\middle|\mathcal{S}\right) = \widehat{J}$,

$$n\text{var}\left(\widehat{\Delta}^*\middle|\mathcal{S}\right) = \widehat{H}^{-1}n\text{var}\left(\frac{1}{n}\sum_{i=1}^n V_{ni}Z_{ni}(Y_{ni} - Z_{ni}'\widehat{\beta})\middle|\mathcal{S}\right)\widehat{H}^{-1} = \widehat{H}^{-1}\widehat{J}\widehat{H}^{-1} \xrightarrow{p} H^{-1}JH^{-1},$$

which is a valid estimate of the asymptotic variance of $\widehat{\beta}$. However, the remainder term $\widehat{R}^*$ generally does not have a bounded second moment since $\widehat{H}^*$ is badly conditioned for some bootstrap draws.

To show that $\tilde{\beta}^*$ yields valid standard errors, we collect a number of preliminary results. Consider the random variables $\widehat{\Delta}^*$ and $\tilde{\Delta}^* = \widehat{\Delta}^* 1_{n^\alpha\|\widehat{H}^* - \widehat{H}\| \leq c}$. $\sqrt{n}\widehat{\Delta}^*$ converges in distribution to $\mathcal{N}(\mathbf{0}, \Sigma)$ with $\Sigma = H^{-1}JH^{-1}$, conditional on $\mathcal{S}$, by Proposition A.2. Since $P(\tilde{\Delta}^* = \widehat{\Delta}^*|\mathcal{S}) \xrightarrow{p} 1$, the same holds true for $\sqrt{n}\tilde{\Delta}^*$ by the above argument. Also, we have

established that

$$E\left(\sqrt{n}\widehat{\Delta}^*\middle|\mathcal{S}\right)=0, \qquad \operatorname{var}\left(\sqrt{n}\widehat{\Delta}^*\middle|\mathcal{S}\right)\xrightarrow{p}\Sigma$$

and thus $E[n\|\widehat{\Delta}^*\|^2|\mathcal{S}]\xrightarrow{p}\operatorname{tr}(\Sigma)$. Since $E[n\|\tilde{\Delta}^*\|^2|\mathcal{S}]\leq E[n\|\widehat{\Delta}^*\|^2|\mathcal{S}]$, and $n\|\tilde{\Delta}^*\|^2$ and $n\|\widehat{\Delta}^*\|^2$ have the same weak limit (with expectation $\operatorname{tr}(\Sigma)$) by the continuous mapping theorem, $E[n\|\tilde{\Delta}^*\|^2|\mathcal{S}]\xrightarrow{p}\operatorname{tr}(\Sigma)$ by Proposition A.3. Consequently,

$$E[n\|\widehat{\Delta}^*\|^2|\mathcal{S}]-E[n\|\tilde{\Delta}^*\|^2|\mathcal{S}]=P(n^\alpha\|\widehat{H}^*-\widehat{H}\|>c|\mathcal{S})\;E[n\|\widehat{\Delta}^*\|^2|n^\alpha\|\widehat{H}^*-\widehat{H}\|>c,\mathcal{S}]\xrightarrow{p}0. \tag{A.3}$$

Next, note that for conformable random variables A,B if $\operatorname{var}(A|\mathcal{S})\xrightarrow{p}\Sigma$, $E[\|B\|^2|\mathcal{S}]\xrightarrow{p}0$ then $\operatorname{var}(A+B|\mathcal{S})\xrightarrow{p}\Sigma$. Indeed,

$$\begin{aligned}&|(\operatorname{var}(A+B|\mathcal{S})-\operatorname{var}(A|\mathcal{S}))_{ij}|=|\operatorname{cov}(A_i,B_j|\mathcal{S})+\operatorname{cov}(A_j,B_i|\mathcal{S})+\operatorname{cov}(B_i,B_j|\mathcal{S})|\\&\leq\sqrt{\operatorname{var}(A_i|\mathcal{S})}\sqrt{\operatorname{var}(B_j|\mathcal{S})}+\sqrt{\operatorname{var}(A_j|\mathcal{S})}\sqrt{\operatorname{var}(B_i|\mathcal{S})}+\sqrt{\operatorname{var}(B_i|\mathcal{S})}\sqrt{\operatorname{var}(B_j|\mathcal{S})}\xrightarrow{p}0.\end{aligned}$$

Hence, setting $A=\sqrt{n}\widehat{\Delta}^*$ and $B=\sqrt{n}(\tilde{\beta}^*-\widehat{\beta}-\widehat{\Delta}^*)$, to establish the desired result $\operatorname{var}(\sqrt{n}(\tilde{\beta}^*-\widehat{\beta})|\mathcal{S})\xrightarrow{p}H^{-1}JH^{-1}$ it suffices to show that

$$E\left[n\|\tilde{\beta}^*-\widehat{\beta}-\widehat{\Delta}^*\|^2\middle|\mathcal{S}\right]\xrightarrow{p}0 \tag{A.4}$$

as $n\to\infty$.

Towards establishing (A.4), note first that whenever $n^\alpha\|\widehat{H}^*-\widehat{H}\|\leq c$ then also

$$\begin{aligned}\|(\widehat{H}^*)^{-1}-\widehat{H}^{-1}\|&=\|(\widehat{H}^*)^{-1}(\widehat{H}-\widehat{H}^*)\widehat{H}^{-1}\|\\&\leq\|(\widehat{H}^*)^{-1}\|\;\|\widehat{H}-\widehat{H}^*\|\;\|\widehat{H}^{-1}\|\\&\leq\lambda_{\min}^{-1}(\widehat{H}^*)\;\lambda_{\min}^{-1}(\widehat{H})\;\|\widehat{H}-\widehat{H}^*\|\;\dim(Z)\end{aligned}$$

where

$$\begin{aligned}\lambda_{\min}(\widehat{H}^*)=\lambda_{\min}(\widehat{H}+\widehat{H}^*-\widehat{H})&=\min_{\|x\|=1}x'(\widehat{H}+\widehat{H}^*-\widehat{H})x\\&\geq\min_{\|x\|=1}x'\widehat{H}x+\min_{\|x\|=1}x'(\widehat{H}^*-\widehat{H})x\geq\lambda_{\min}(\widehat{H})-\|\widehat{H}^*-\widehat{H}\|\end{aligned}$$

and thus

$$\begin{aligned}\|(\widehat{H}^*)^{-1} - \widehat{H}^{-1}\| &\leq (\lambda_{\min}(\widehat{H}) - \|\widehat{H}^* - \widehat{H}\|)^{-1}\ \lambda_{\min}^{-1}(\widehat{H})\ \|\widehat{H}^* - \widehat{H}\|\ \dim(Z) \\ &\leq (\lambda_{\min}(\widehat{H}) - cn^{-\alpha})^{-1}\ \lambda_{\min}^{-1}(\widehat{H})\ cn^{-\alpha}\ \dim(Z). \qquad \text{(A.5)}\end{aligned}$$

If follows that

$$\begin{aligned}&E\left[n\|\tilde{\beta}^* - \widehat{\beta} - \widehat{\Delta}^*\|^2 \middle| \mathcal{S}\right] \\ &= P(n^\alpha\|\widehat{H}^* - \widehat{H}\| \leq c|\mathcal{S})\ E[n\|\overbrace{\tilde{\beta}^*}^{=\widehat{\beta}^*} - \widehat{\beta} - \widehat{\Delta}^*\|^2 | n^\alpha\|\widehat{H}^* - \widehat{H}\| \leq c, \mathcal{S}] \\ &\quad + P(n^\alpha\|\widehat{H}^* - \widehat{H}\| > c|\mathcal{S})\ E[n\|\underbrace{\tilde{\beta}^*}_{=\widehat{\beta}} - \widehat{\beta} - \widehat{\Delta}^*\|^2 | n^\alpha\|\widehat{H}^* - \widehat{H}\| > c, \mathcal{S}] \\ &= P(n^\alpha\|\widehat{H}^* - \widehat{H}\| \leq c|\mathcal{S})\ E[n\overbrace{\|\widehat{R}^*\|^2}^{\leq\|(\widehat{H}^*)^{-1} - \widehat{H}^{-1}\|^2\|\frac{1}{n}\sum_{i=1}^n V_{ni}Z_{ni}(Y_{ni} - Z_{ni}'\widehat{\beta})\|^2} | n^\alpha\|\widehat{H}^* - \widehat{H}\| \leq c, \mathcal{S}] \\ &\quad + P(n^\alpha\|\widehat{H}^* - \widehat{H}\| > c|\mathcal{S})\ E[n\|\widehat{\Delta}^*\|^2 | n^\alpha\|\widehat{H}^* - \widehat{H}\| > c, \mathcal{S}] \\ &\overset{(A.5)}{\leq} (\underbrace{\lambda_{\min}(\widehat{H})}_{\xrightarrow{p}\lambda_{\min}(H)>0} - cn^{-\alpha})^{-1}\ \lambda_{\min}^{-1}(\widehat{H})\ cn^{-\alpha}\ \dim(Z) \\ &\quad \underbrace{P(n^\alpha\|\widehat{H}^* - \widehat{H}\| \leq c|\mathcal{S})\ E[\|n^{-1/2}\textstyle\sum_{i=1}^n V_{ni}Z_{ni}(Y_{ni} - Z_{ni}'\widehat{\beta})\|^2 | n^\alpha\|\widehat{H}^* - \widehat{H}\| \leq c, \mathcal{S}]}_{\leq E[\|\frac{1}{\sqrt{n}}\sum_{i=1}^n V_{ni}Z_{ni}(Y_{ni} - Z_{ni}'\widehat{\beta})\|^2|\mathcal{S}] = \mathrm{tr}(\widehat{J}) \xrightarrow{p} \mathrm{tr}(J)} \\ &\quad + \underbrace{P(n^\alpha\|\widehat{H}^* - \widehat{H}\| > c|\mathcal{S})\ E[n\|\widehat{\Delta}^*\|^2 | n^\alpha\|\widehat{H}^* - \widehat{H}\| > c, \mathcal{S}]}_{\overset{(A.3)}{\xrightarrow{p}} 0} \\ &\xrightarrow{p} 0.\end{aligned}$$

Hence, $\mathrm{var}(\sqrt{n}(\tilde{\beta}^* - \widehat{\beta})|\mathcal{S})$ and $\mathrm{var}(\sqrt{n}\widehat{\Delta}^*|\mathcal{S})$ have the same probability limit $H^{-1}JH^{-1}$, which is also the asymptotic variance of $\widehat{\beta}$. □