Combining matching and linear regression: Introducing a mathematical framework and software for simulations, diagnostics and calibration

Alireza S. Mahani

Scientific Computing Group, Sentrana Inc, Washington, DC, USA.

Mansour T. A. Sharabiani

National Heart and Lung Institute, Imperial College London, UK.

Summary. Matching and linear regression are sometimes combined to achieve double-robustness in causal inference. In real-world scenarios, however, there are ambiguities in quantifying the impact of this combination on treatment effect bias as well as concerns over the paradoxical effects of matching on bias and variance. We, therefore, present (i) a theoretical and mathematical framework, and (ii) a set of diagnostic and calibration tools for quantifying the combined effect of matching and linear regression for finite samples with arbitrary covariate distributions. Our theorems confirm that perfect matching eliminates bias due to covariate omission in regression, and prove that balanced distribution of covariates included in regression model minimizes variance for a given data size. The accompanying software, R package MatchLinReg, allows applied statisticians to visualize the combined impact of matching and linear regression on bias, variance and mean-squared-error, and to calibrate matching parameters to minimize the error. Also, the incorporated power analysis tool quantifies the potential reduction in study power due to reduced data size after matching. As an additional benefit, our closed-form expressions for bias and variance can replace Monte Carlo techniques for conducting accurate and generalizable simulation studies much faster.

Keywords: causal inference, observational studies, linear regression, bias, variance, power analysis

1. Introduction

Theoretical work on causal inference in observational studies has largely focused on large-sample, bias-removal properties of techniques such as matching and regression (Abadie and Imbens, 2006, 2011). In applied settings, however, treatment - and even control - groups are finite and sometimes small, asymptotic properties are not realized, and variance is as important as bias in contributing towards total estimation error. This is particularly true in medical applications such as studying the effectiveness of a novel in-patient heart procedure (Kereiakes et al., 2000), where not only randomized treatment assignment is unfeasible due to ethical or logistical concerns, but data collection is costly and time-consuming, and thus early identification of treatment effectiveness is highly desirable.

In particular, the general recognition by researchers of the added benefit of combining matching and regression to achieve the so-called 'double-robustness' (Rubin, 1973; Carpenter, 1977; Rubin, 1979; Robins and Rotnitzky, 1995; Heckman et al., 1997; Rubin and Thomas, 2000; Glazerman et al., 2003; Abadie and Imbens, 2006; Ho et al., 2007; Abadie and Imbens, 2011) has yet to translate into a set of guidelines and tools that allow for their effective utilization by practitioners. For example, what caliper size should be used in matching? What is the impact of the choice of matching technique, e.g. Mahalanobis distance matching vs. propensity score matching (PSM), on treatment effect (TE) estimation? What terms should be included in matching (e.g. in the logistic regression model used for generating propensity scores)? How do such decisions affect not only TE estimation bias but also variance? At what point is study power diminished beyond usefulness when matching is applied more and more restrictively to reduce bias? Simulation studies have shed light on some of these questions empirically (Austin et al., 2007; Austin, 2009; Gayat et al., 2012), and providing a theoretical foundation can facilitate generalizable conclusions from such empirical work (Imbens, 2004).

In this paper, we develop the mathematical framework for quantifying the combined effect of matching and linear regression on TE estimation bias and variance in finite samples with arbitrary covariate distributions. Similar to the approach of Ho et al. (2007), we consider matching as a non-parametric pre-processing step prior to regression adjustment. We cast the model misspecification problem as a covariate omission problem, and derive closed-form expressions for TE estimation bias and variance in linear regression adjustment. These expressions are useful in three capacities:

1. Theory: Using the normalized form of bias expression, we prove that perfect matching on included and omitted covariates eliminates bias. Furthermore, perfect matching on included covariates minimizes variance, given a fixed number of treatment and control observations. The variance-minimization property of matching provides theoretical support for the notion that study power is not lost as rapidly when sample size is reduced due to matching, compared to random data loss.

2. Simulations: The closed-form expressions for bias and variance can be efficient and accurate substitutes for Monte Carlo simulations for studying matching and linear regression. We have used these expressions for most of the simulations presented in this paper, and expect that future empirical studies in the field will similarly benefit from this efficient simulation tool.

3. Diagnostic and calibration software: Our expressions can be utilized in a diagnostic capacity to analyze bias-variance trade-off resulting from various choices of matching and regression parameters, and to quantify the impact of matching on study power. By utilizing such diagnostic tools - embedded in our recently-released open-source R package MatchLinReg (Mahani and Sharabiani, 2015) - practitioners will be able to make better-informed decisions during the design and analysis of observational studies.

The rest of this paper is organized as follows. We continue by reviewing previous work on combining matching and regression for causal inference, and presenting simulation results using real data sets to highlight the benefits and complexities associated with combining matching and regression, which mo-

tivated our research. In Section 2, we present our mathematical framework. We outline the approach and assumptions of the paper, formally define the problem, derive equations for TE bias and variance estimation, and prove theorems regarding the impact of matching on bias and variance. (Lengthy proofs can be found in Supplementary Material.) In Section 3, we use the developed framework to create diagnostic and calibration tools. We begin with exploratory diagnostics such as relative squared bias reduction due to single omitted covariates, and comparison of normalized single-covariate biases, and proceed to generate an aggregate measure of normalized bias using constrained bias estimation methodology. We then combine normalized bias and variance to arrive at normalized MSE, using a conversion factor called 'omitted R-squared'. Minimizing MSE allows us to calibrate the matching parameters. Finally, in Section 4 we provide a summary of contributions in this paper, and outline pointers for promising future research directions. In addition to mathematical details, Supplementary Material contains the list of acronyms, description of data sets, and overview of MatchLinReg R package.

1.1. Previous work

There is a rich body of literature on causal inference for observational studies, and some have focused - partially or wholly - on combined use of matching and regression adjustment. Rubin (1973) derive expressions for bias reduction for a single confounder, considering parallel and non-parallel response surfaces. Carpenter (1977) analyze the performance of nearest-neighbor matching, and its combination with regression adjustment. While their primary focus is on normally-distributed covariates, they present an Analysis-of-covariance (ANCOVA) equation for variance, which is equivalent to our standard expression (Eq. 26b). Rubin (1979) study the effect of various approaches for combining matching and regression adjustment for controlling bias, and conclude that covariate matching combined with regression on paired differences performs best. Rubin and Thomas (1992, 1996) study bias reduction resulting from PSM for normally-distributed covariates. Some of their expressions rely on calculating the correlation between covariates and outcome. Rubin and Thomas (2000) study the bias-removal impact of combining PSM with regression adjustment and conclude that the combination has superior performance to regression alone. Ho et al. (2007) use real data sets to illustrate that using matching as a non-parametric pre-processing step reduces model dependence in parametric causal inference. While cautioning that increased variance from data loss can outweigh reduced bias in matching, they also point out that 'no precise rule exists for how to make these [bias-variance tradeoff] choices.' Schafer and Kang (2008) use simulations based on a real data set to compare the performance of various combinations of ANCOVA and PSM techniques. Hosman et al. (2010) present methods for analyzing sensitivity of regression to confounder omission, potentially combined with propensity-score-based stratification. Abadie and Imbens (2011) prove that adding a bias-correcting step such as least-squares regression to nearest-neighbor matching renders it $N^{1/2}$-consistent. Their analysis is asymptotic in nature, however, and is focused on matching with replacement, a common assumption in econometric

literature. When the unconfoundedness assumption (see Section 2.1) is inadequate, other approaches for double-robustness must be adopted such as instrumental variables. See, for example, Tan (2010); Rotnitzky et al. (2012).

1.2.   Complexities of TE estimation: Looking beyond double-robustness

A common approach for combining matching and regression in statistical analysis is to use matching as a non-parametric pre-processing step before regression (Ho et al., 2007). This approach allows practitioners to apply the familiar regression tools to their problem once matching is done, although it has been argued that issues such as calculation of standard errors must take into consideration the matching step (Austin, 2007). In linear regression with a treatment indicator binary variable and other (adjustment) covariates included, the coefficient of treatment indicator equals TE (Imbens, 2004). Including adjustment covariates in regression helps remove residual bias due to imperfect matching. A special case is when no adjustment covariates are used in regression; we call this method the 'simple difference' method, since it is equivalent to calculating and subtracting the mean outcome in treatment and control groups to obtain TE. As an alternative view, matching helps reduce sensitivity of regression adjustment to model misspecification (Section 2.4). This mutually-reinforcing effect of combining matching and regression for TE estimation has been referred to as 'double-robustness' in the literature (Stuart, 2010).

Figure 1 illustrates the combined effect of matching and regression for TE estimation, using Monte Carlo simulations based on lalonde and lindner data sets. (See Supplementary Material for a description of these data sets.) For each data set, noise variance as well as coefficients of original variables were extracted from linear regression on full data sets, and a quadratic term ($re74^2$ for lalonde and $ejecfrac^2$ for lindner) was added to the generative model, with its coefficient chosen so as to achieve an omitted R-squared (R_o^2) of 5.9% and 3.3% for lalonde and lindner, respectively, amounting to 40% of R-squared from regression on main effects in each data set. R_o^2 quantifies the fraction of variation in outcome attributable to covariates that are not included in the regression model. (See Section 3.2 for mathematical definition.) In order to simulate regression misspecification through covariate omission (Section 2.2), regressions used in Monte Carlo simulations did not include this quadratic term. As expected, matching and regression each significantly reduce bias (compared to simple difference method) in both data sets, and combining them further reduces bias.

However, mean squared error (MSE) is the sum of squared bias and variance:

$$MSE = (bias)^2 + variance \tag{1}$$

While matching reduces TE bias in regression adjustment caused by covariate omission, yet it does so by discarding data points and thus reducing data size, leading to increased variance. As Figure 1 shows, while in lalonde data set the bias-reduction effect of adding matching to regression more than offsets the increased variance such that MSE is reduced, the net effect for lindner is adverse, such that regression alone is the best option. This represent one of the key challenges in calibrating
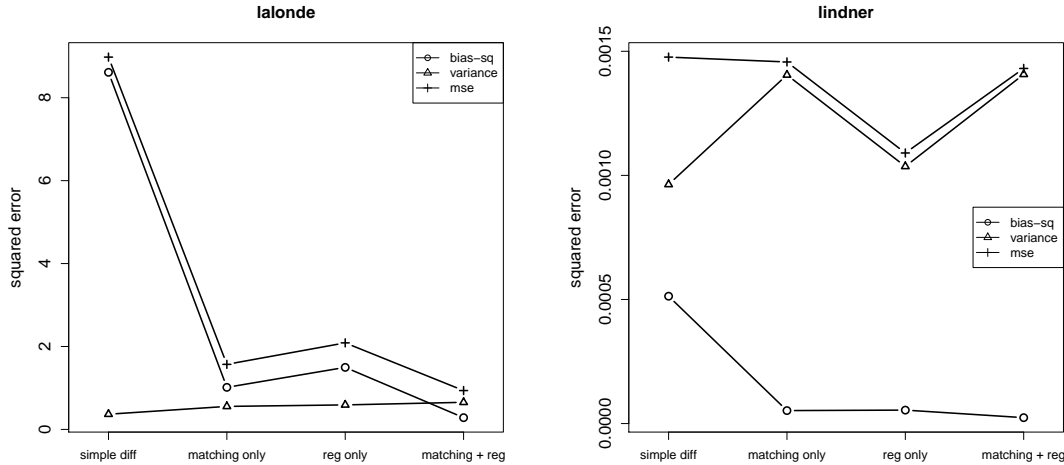
Fig. 1: Combining matching and regression for TE estimation in lalonde (left) and lindner (right) data sets. Error components (squared bias, variance, MSE) are compared using four approaches: 1) simple difference, 2) matching followed by simple difference, 3) regression adjustment, and 4) matching followed by regression adjustment. A caliper of 1.5 (lalonde) and 0.2 (lindner) was used for matching. Numbers are based on 10,000 Monte Carlo simulation. Omitted covariate is $re74^2$ for lalonde and $ejecfrac^2$ for lindner, with coefficient of this omitted covariate chosen so as to achieve an R_o^2 of 5.9% and 3.3% for lalonde and lindner, respectively, which represents 40% of R-squared of main-effects-only regression model.

the matching-plus-regression combination, i.e. striking the right balance between bias reduction and variance increase so as to minimize MSE for TE.

Variance considerations are not the only source of complexity. Matching affects bias indirectly and only by adjusting the distribution of covariates across treatment and control groups. In determining the ultimate impact of matching on TE bias, we face several complexities: 1) Bias induced by an omitted covariate is not proportional to its imbalance (Figure 2, top row); 2) matching does not improve balance for all covariates, although it tends to do so for the most imbalanced covariates (Figure 2, middle row); 3) as a result of 1 and 2, matching does not improve bias due to all omitted covariates, but again it tends to reduce the largest biases (Figure 2, bottom row); 4) Finally, in generating Figure 2 we assumed only one omitted covariate at a time, but in reality multiple such covariates can be missing from our regression model. What combination of these unknown terms, and their impact on TE bias, should be considered in how we combine regression and matching? The mathematical framework and diagnostics software developed in this paper seek to address the above challenges.

Note that Equal-Percent-Bias-Reduction (EPBR) property of matching techniques (Rubin, 1976) such as PSM does not solve the complexities associated with TE bias: 1) EPBR is valid only when covariate distributions follow restrictions such as being discriminant mixtures of ellipsoidally symmetric distributions (Rubin and Stuart, 2006), 2) EPBR only works for linear combinations of covariates,

and not their nonlinear functions such as interactions and powers.

2.   Framework

In this section, we develop the mathematical framework for analyzing the combined effect of matching and linear regression for TE estimation. We start with the underlying assumptions and notation.

2.1.   Approach and assumptions

Ignorability of treatment assignment: Also referred to as 'unconfoundedness' (Imbens, 2004) and 'conditional independence' (Lechner, 1999, 2002), this assumption states that, conditioned on the covariates available in our data set, treatment assignment is random (Rosenbaum and Rubin, 1983). In the context of a standard linear model (Section 2.2), this assumption means that all covariates in the generative model are derivatives (e.g. powers, interaction terms, splines, etc) of a core set of covariates included in the data set. This, in turn, means that perfect matching on the core set of covariates would automatically lead to perfect matching on the full set of covariates in the generative model.

Matching as pre-processing for regression adjustment: Similar to Ho et al. (2007), we use matching as a pre-processing step before parametric estimation of TE using linear regression. Combined with the parallel response surface assumption (see below), linear regression on the full data set produces TE as the coefficient of treatment indicator variable (Imbens, 2004). Our approach can be considered an extension of two-group ANCOVA (Rubin, 1973; Carpenter, 1977; Quade, 1982) to include the impact of regression misspecification parameterically and through the covariate omission concept. Notable alternatives include regression on matched pair differences (Rubin, 1979), and separate regressions on treatment and control groups (Imbens, 2004). Using linear regression for parametric TE estimation allows us to derive closed-form expressions for TE bias and variance (Section 2.3) for finite samples and without any assumptions about the distribution of covariates in the treatment and control groups, or how their covariate matrices are related. Matching followed by a simple-difference calculation of TE, i.e. without regression adjustment, is a special case of the above framework, but with no adjustment covariates included in the regression model. In this case, the regression model has only two coefficients: intercept and TE.

Parallel response surfaces: This assumption states that the difference between functions describing mean outcome for treatment and control groups is independent of the covariate vector. In other words, TE is constant for all observations, conditional and marginal/average TE are the same, and independent of sample. In a linear regression setting, this assumption means the generative model does not contain any interaction terms between treatment indicator and adjustment covariates, and correspondingly we do not include any such interaction terms in the regression model. While an important restriction, the parallel response surface assumption allows us to focus on developing the mathematical framework without concern for what relative weights to attach to TE errors in different
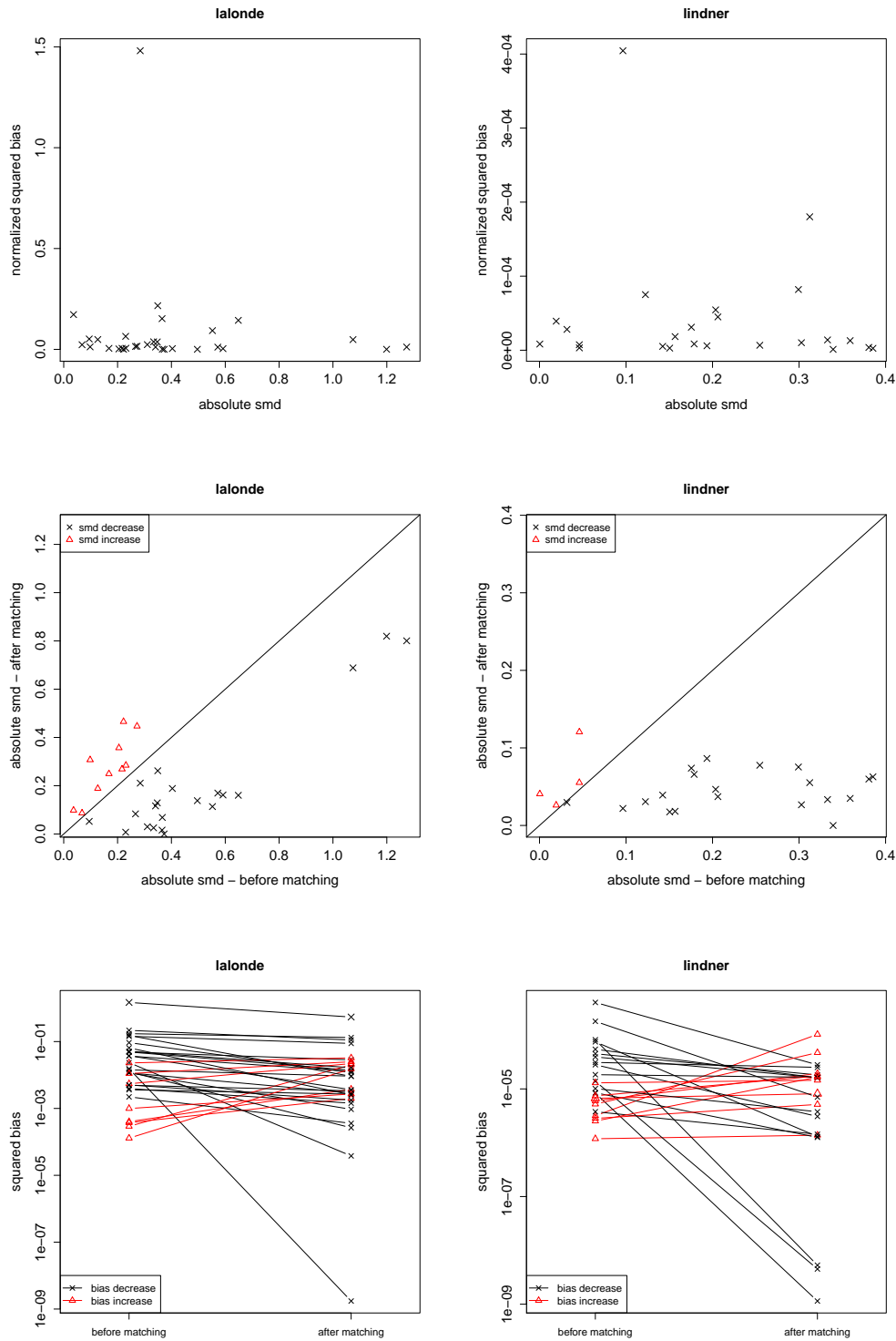
Fig. 2: Complex relationship between matching, covariate imbalance, and omission bias for lalonde (left) and lindner (right) data sets. Matching followed by linear regression is used for all plots. Outcome variable definitions, matching algorithm and parameters, and choice of coefficients for omitted covariates are identical to Figure 1. Equations of Section 2.3 are used for bias calculations. Top row: squared bias vs. absolute SMD for all second-order interaction terms (missing from regression model). Middle row: Comparison of SMD before and after matching for omitted terms. Bottom row: Comparison of squared bias, before and after matching, for omitted terms.

strata.

Covariate omission as model misspecification: Matching reduces potential TE estimation bias caused by a misspecified regression model. In the linear regression framework - where terms can be nonlinear in original covariates - misspecification manifests itself in the form of covariate omission, i.e. one or more terms included in data generation being absent from the regression model. Importantly, including nuisance terms in regression, i.e. terms that were not part of the generative model, does not induce bias, but only increases variance.

Exclusion of outcome variable from analyses: It is possible to use the outcome variable - implicitly or explicitly - during the matching calibration process. For example, one may focus on achieving better covariate balance for those covariates that show a stronger correlation with outcome. While careful use of outcome variable can enhance the quality of entire modeling process including matching, in the hands of inexperienced practitioners it can be lead to overfit models with poor generalizability in real-world (Babyak, 2004). We have therefore chosen to develop our diagnostic and calibration tools (Section 3) with complete exclusion of outcome variable from consideration. In addition to preventing overfit creep, this approach also makes our framework and tools applicable for study design analyses, where outcome data is unavailable (Stuart, 2010).

2.2.   Problem definition

In this paper, we focus on causal inference for a *Standard Linear Model* (SLM), which consists of two components: a data generation (generative) model and a coefficient estimation (regression) model:

1. Generative model: Outcome ($\mathbf{y}$) is continuous and its mean is a linear function of covariates ($\mathbf{X}$):

$$\mathrm{E}[\mathbf{y}] = \mathbf{X}\boldsymbol{\beta}. \tag{2}$$

Data generation has an i.i.d. noise distribution:

$$\mathrm{cov}[\mathbf{y}] = \sigma_0^2\,\mathbf{I}. \tag{3}$$

Data-generation covariates consist of a binary treatment indicator ($\mathbf{w}$), a unit vector ($\mathbf{1}$), as well as a set of K adjustment covariates ($\mathbf{Z}$), none of which include interactions with treatment indicator (parallel response surface):

$$\mathbf{X} = \begin{bmatrix} \mathbf{w} & \mathbf{1} & \mathbf{Z} \end{bmatrix}. \tag{4}$$

$\mathbf{w}$ is 1 for treatment observations (N_t cases) and 0 for controls (N_c cases), with total observations being $N = N_t + N_c$. Correspondingly, the vector of coefficients ($\boldsymbol{\beta}$) consists of TE (τ), intercept (β_0) and vector of coefficients for adjustment covariates ($\boldsymbol{\gamma}$):

$$\boldsymbol{\beta} \equiv \begin{bmatrix} \tau & \beta_0 & \boldsymbol{\gamma}^t \end{bmatrix}^t, \tag{5}$$

where $\boldsymbol{\gamma}^t$ stands for transpose of $\boldsymbol{\gamma}$. For SLM, average TE for treated (ATT), sample average TE (ATE) and average TE for controls (ATC) are all the same and equal to the coefficient of treatment indicator variable, i.e. τ. We refer to this entity simply as TE (treatment effect).

2. Regression model: Ordinary-Least-Squares (OLS) is used to estimate model coefficients, including τ. However, only a subset of covariates ($\mathbf{X}^i$) are included in the regression model. This subset includes $\mathbf{w}$ and $\mathbf{1}$, as well as K^i adjustment covariates ($\mathbf{Z}^i$), which we call 'included adjustment covariates' or simply included covariates. The remaining K^o ($= K - K^i$) adjustment covariates ($\mathbf{Z}^o$) are called 'omitted adjustment covariates', or simply omitted covariates:

$$\mathbf{Z} \equiv \left[\begin{array}{cc} \mathbf{Z}^i & \mathbf{Z}^o \end{array} \right], \tag{6}$$

$$\mathbf{X}^i \equiv \left[\begin{array}{ccc} \mathbf{w} & \mathbf{1} & \mathbf{Z}^i \end{array} \right], \tag{7}$$

$$\mathbf{X} = \left[\begin{array}{cc} \mathbf{X}^i & \mathbf{Z}^o \end{array} \right]. \tag{8}$$

The vector of coefficients for adjustment covariates is correspondingly partitioned into included ($\boldsymbol{\gamma}^i$) and omitted ($\boldsymbol{\gamma}^o$) vectors:

$$\boldsymbol{\gamma} = \left[\begin{array}{cc} \boldsymbol{\gamma}^{i,t} & \boldsymbol{\gamma}^{o,t} \end{array} \right]^t, \tag{9}$$

$$\boldsymbol{\beta}^i \equiv \left[\begin{array}{ccc} \tau & \beta_0 & \boldsymbol{\gamma}^{i,t} \end{array} \right]^t, \tag{10}$$

$$\boldsymbol{\beta} = \left[\begin{array}{cc} \boldsymbol{\beta}^{i,t} & \boldsymbol{\gamma}^{o,t} \end{array} \right]^t. \tag{11}$$

OLS estimation of coefficients leads to:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{i,t}\mathbf{X}^i)^{-1}\,\mathbf{X}^{i,t}\mathbf{y}, \tag{12}$$

with

$$\hat{\boldsymbol{\beta}} \equiv \left[\begin{array}{ccc} \hat{\tau} & \hat{\beta}_0 & \hat{\boldsymbol{\gamma}^i}^{t} \end{array} \right]^t. \tag{13}$$

Note that omitted covariates ($\boldsymbol{\gamma}^o$) are not estimated, since they are not included in the regression model.

2.3. TE Bias and variance equations

As mentioned earlier, TE estimation bias in regression adjustment is caused by model misspecification, which is manifested as covariate omission in our framework, i.e. when $\mathbf{Z}^o \neq \mathbf{0}$, or equivalently when $\mathbf{X} \neq \mathbf{X}^i$. It is easy to check that a correctly-specified model is unbiased:

$$\mathbf{X}^i = \mathbf{X} \implies \mathrm{E}[\hat{\boldsymbol{\beta}}] = (\mathbf{X}^t\mathbf{X})^{-1}\,\mathbf{X}^t\,\mathrm{E}[\mathbf{y}] = (\mathbf{X}^t\mathbf{X})^{-1}\,\mathbf{X}^t(\mathbf{X}\boldsymbol{\beta}) = \boldsymbol{\beta}, \tag{14}$$

where we have combined Eqs. 2 and 12. With missing covariates, we have:

$$\mathrm{E}[\hat{\boldsymbol{\beta}}] - \boldsymbol{\beta}^i = (\mathbf{X}^{i,t}\mathbf{X}^i)^{-1}\mathbf{X}^{i,t}\left(\mathbf{X}^i\boldsymbol{\beta}^i + \mathbf{Z}^o\boldsymbol{\gamma}^o\right) - \boldsymbol{\beta}^i = (\mathbf{X}^{i,t}\mathbf{X}^i)^{-1}(\mathbf{X}^{i,t}\mathbf{Z}^o)\,\boldsymbol{\gamma}^o. \tag{15}$$

As for variance, we have the following standard expression:

$$\mathrm{cov}[\hat{\boldsymbol{\beta}}] = \sigma_0^2\,(\mathbf{X}^{i,t}\mathbf{X}^i)^{-1}. \tag{16}$$

(In the case of repeated observations, e.g. in matching with replacement, the bias expression remains valid, while the variance expression must be modified. See Supplementary Material.) Given our

convention regarding the order of covariates in Eq. 13, TE bias and variance are the first / top-left-corner elements of the bias vector / covariance matrix, respectively, given in Eqs. 15 and 16.

Next, we rewrite the above expressions in a way that highlights the role of covariate imbalance and the effect of matching. Since Eqs. 15 and 16 both contain $(\mathbf{X}^{i,t}\mathbf{X}^i)^{-1}$, we begin by transforming this matrix. From Equation 7, we have:

$$\mathbf{X}^{i,t}\mathbf{X}^i = \begin{bmatrix} \mathbf{w}^t \\ \mathbf{1}^t \\ \mathbf{Z}^{i,t} \end{bmatrix} \begin{bmatrix} \mathbf{w} & \mathbf{1} & \mathbf{Z}^i \end{bmatrix} \tag{17}$$

$$= \begin{bmatrix} N_t & N_t & \mathbf{w}^t\mathbf{Z}^i \\ N_t & N_t + N_c & \mathbf{1}^t\mathbf{Z}^i \\ \mathbf{Z}^{i,t}\mathbf{w} & \mathbf{Z}^{i,t}\mathbf{1} & \mathbf{Z}^{i,t}\mathbf{Z}^i \end{bmatrix}, \tag{18}$$

where we have taken advantage of $\mathbf{w}^t\mathbf{w} = \mathbf{1}^t\mathbf{w} = \mathbf{w}^t\mathbf{1} = N_t$ and $\mathbf{1}^t\mathbf{1} = N_t + N_c$. Since $\mathbf{X}^{i,t}\mathbf{X}^i$ is symmetric, so is its inverse. To calculate TE bias and variance, we are only concerned with the first row/column of $(\mathbf{X}^{i,t}\mathbf{X}^i)^{-1}$. In Supplementary Material, we prove that:

$$(\mathbf{X}^{i,t}\mathbf{X}^i)^{-1} = \begin{pmatrix} \frac{1}{N_t} + \frac{1}{N_c} + \boldsymbol{u}^{i,t}\mathbf{A}^{-1}\boldsymbol{u}^i & -\frac{1}{N_c} + \frac{1}{N_c}(\boldsymbol{p}-\boldsymbol{q})^t\mathbf{A}^{-1}\boldsymbol{u}^i & \boldsymbol{u}^{i,t}\mathbf{A}^{-1} \\ -\frac{1}{N_c} + \frac{1}{N_c}(\boldsymbol{p}-\boldsymbol{q})^t\mathbf{A}^{-1}\boldsymbol{u}^i & \dots & \dots \\ \mathbf{A}^{-1}\boldsymbol{u}^i & \dots & \dots \end{pmatrix}, \tag{19}$$

where $\boldsymbol{u}^i$ is the vector of mean differences of included covariates, $\boldsymbol{p}$ and $\boldsymbol{q}$ are vector of sums of included covariates across treatment/all observations, and $\mathbf{A}$ is the pooled, within-group covariance matrix for included covariates:

$$\boldsymbol{u}^i \equiv \frac{1}{N_c}(\mathbf{1}-\mathbf{w})^t\mathbf{Z}^i - \frac{1}{N_t}\mathbf{w}^t\mathbf{Z}^i, \tag{20}$$

$$\boldsymbol{p} \equiv \mathbf{Z}^{i,t}\mathbf{w}, \tag{21}$$

$$\boldsymbol{q} \equiv \mathbf{Z}^{i,t}\mathbf{1}, \tag{22}$$

$$\mathbf{A} \equiv (N_t - 1)\operatorname{cov}(\mathbf{Z}^i)_T + (N_c - 1)\operatorname{cov}(\mathbf{Z}^i)_C. \tag{23}$$

Within-group covariance matrices are formally defined as follows:

$$\operatorname{cov}(\mathbf{Z}^i)_T \equiv \frac{1}{N_t - 1}\mathbf{Z}_T^{i,t}(\mathbf{I}_T - \frac{1}{N_t}\mathbf{1}_T\mathbf{1}_T^t)\mathbf{Z}_T^i, \tag{24}$$

$$\operatorname{cov}(\mathbf{Z}^i)_C \equiv \frac{1}{N_c - 1}\mathbf{Z}_C^{i,t}(\mathbf{I}_C - \frac{1}{N_c}\mathbf{1}_C\mathbf{1}_C^t)\mathbf{Z}_C^i, \tag{25}$$

where $\mathbf{Z}_T^i$ and $\mathbf{Z}_C^i$ are the row subsets of included adjustment covariate matrices for treatment and control groups, respectively. $\mathbf{I}_T$ and $\mathbf{I}_C$ are identity matrices of dimensions N_t and N_c, and $\mathbf{1}_T$ and $\mathbf{1}_C$ are unit vectors of length N_t and N_c, respectively.

TE bias is the first element of the bias vector in Eq. 15, which is the result of multiplying the first row of $(\mathbf{X}^{i,t}\mathbf{X}^i)^{-1}\mathbf{X}^{i,t}\mathbf{Z}^o$ by $\boldsymbol{\gamma}^o$, while TE variance is simply σ_0^2 multiplied by the top-left-corner element of $(\mathbf{X}^{i,t}\mathbf{X}^i)^{-1}$ (Eq. 16). Using Equation 19, we obtain the following expressions for bias and

variance:

$$\delta = \left\{ \left(\frac{1}{N_t} + \frac{1}{N_c} + \boldsymbol{u}^{i,t} \mathbf{A}^{-1} \boldsymbol{u}^i \right) \mathbf{w}^t + \left(-\frac{1}{N_c} + \frac{1}{N_c} (\boldsymbol{p} - \boldsymbol{q})^t \mathbf{A}^{-1} \boldsymbol{u}^i \right) \mathbf{1}^t \right.$$
$$\left. + \boldsymbol{u}^{i,t} \mathbf{A}^{-1} \mathbf{Z}^{i,t} \right\} \mathbf{Z}^o \boldsymbol{\gamma}^o. \tag{26a}$$

$$\sigma^2 = \sigma_0^2 \left(\frac{1}{N_t} + \frac{1}{N_c} + \boldsymbol{u}^{i,t} \mathbf{A}^{-1} \boldsymbol{u}^i \right). \tag{26b}$$

We refer to Equations 26a and 26b as the standard equations for bias and variance. As shown in Supplementary Material, the above expressions can be transformed into normalized forms:

$$\delta = \left[\frac{N_t N_c}{(N_t + N_c)(N_t + N_c + 1)} \right]^{1/2} \left\{ \boldsymbol{\rho}^{o,t} + \boldsymbol{\rho}^{i,t} \boldsymbol{\Lambda}^{-1} (\boldsymbol{\rho}^i \boldsymbol{\rho}^{o,t} - \boldsymbol{\Phi}^{io}) \right\} (\boldsymbol{\Sigma}^{o1/2} \boldsymbol{\gamma}^o), \tag{27a}$$

$$\sigma^2 = \sigma_0^2 (\frac{1}{N_t} + \frac{1}{N_c})(1 + \boldsymbol{\rho}^{i,t} \boldsymbol{\Lambda}^{-1} \boldsymbol{\rho}^i) = \sigma_{min}^2 (1 + \boldsymbol{\rho}^{i,t} \boldsymbol{\Lambda}^{-1} \boldsymbol{\rho}^i) \tag{27b}$$

where $\boldsymbol{\rho}^i / \boldsymbol{\rho}^o$ are the vectors of treatment-covariate correlations for included/omitted covariates, $\boldsymbol{\Phi}^{io}$ is the cross-correlation matrix between included and omitted covariates, $\boldsymbol{\Lambda}$ is the weighted, pooled correlation matrix, $\boldsymbol{\Sigma}^i$ and $\boldsymbol{\Sigma}^o$ are the diagonal matrices of variances for included and omitted covariates respectively, and σ_{min}^2 is the minimum achievable variance (Theorem 2). These symbols are mathematically defined as:

$$\boldsymbol{\rho}^i \equiv [\, \mathrm{corr}(\mathbf{w}, \mathbf{z}_{,k}^i) \,]_{k=1,\dots,K^i} \tag{28}$$

$$\boldsymbol{\rho}^o \equiv [\, \mathrm{corr}(\mathbf{w}, \mathbf{z}_{,k}^o) \,]_{k=1,\dots,K^o} \tag{29}$$

$$\boldsymbol{\Phi}^{io} \equiv [\, \mathrm{corr}(\mathbf{z}_{,k_1}^i, \mathbf{z}_{,k_2}^o) \,]_{k_1=1,\dots,K^i, k_2=1,\dots,K^o} \tag{30}$$

$$\boldsymbol{\Lambda} \equiv \boldsymbol{\Sigma}^{i,-1/2} \mathbf{A} \, \boldsymbol{\Sigma}^{i,-1/2} / (N_t + N_c - 1) \tag{31}$$

$$\boldsymbol{\Sigma}^i \equiv [\, \mathrm{cov}(\mathbf{z}_{,k_1}^i, \mathbf{z}_{,k_2}^i) \, \delta_{k_1,k_2} \,]_{k_1,k_2=1,\dots,K^i} \tag{32}$$

$$\boldsymbol{\Sigma}^o \equiv [\, \mathrm{cov}(\mathbf{z}_{,k_1}^o, \mathbf{z}_{,k_2}^o) \, \delta_{k_1,k_2} \,]_{k_1,k_2=1,\dots,K^o} \tag{33}$$

$$\sigma_{min}^2 \equiv \sigma_0^2 \left(\frac{1}{N_t} + \frac{1}{N_c} \right) \tag{34}$$

The normalized equations make it clear that TE variance is independent of shifting and scaling of covariates. More importantly, the normalized bias expression of Eq. 27a makes the role of covariate balance and matching immediately discernible (Theorem 1).

The vector of mean differences, $\boldsymbol{u}^i$, can be normalized to obtain a vector of 'standardized mean difference' (SMD) values for each included covariate, $\mathbf{d}^i$. These vectors are proportional to the vector of treatment-covariate correlations, $\boldsymbol{\rho}^i$:

$$\mathbf{d}^i \equiv -\boldsymbol{\Sigma}^{i-1/2} \boldsymbol{u}^i, \tag{35}$$

$$\boldsymbol{\rho}^i = - \left(\frac{N_t N_c}{(N_t + N_c)(N_t + N_c - 1)} \right)^{1/2} \boldsymbol{\Sigma}^{i-1/2} \boldsymbol{u}^i, \tag{36}$$

$$\boldsymbol{\rho}^i = \left(\frac{N_t N_c}{(N_t + N_c)(N_t + N_c - 1)} \right)^{1/2} \mathbf{d}^i. \tag{37}$$

Eq. 36 is proven in Supplementary Material, and Eq. 37 follows from Eqs. 35 and 36. Similar relationships hold for the corresponding vectors of omitted covariates, $\boldsymbol{u}^o$, $\mathbf{d}^o$ and $\boldsymbol{\rho}^o$. A corollary of Eq. 37 is the following:

COROLLARY 1. *If a covariate is balanced across treatment and control groups, it is uncorrelated with the treatment indicator variable.*

PROOF. Balance for covariate k means its SMD, d_k, is zero. This mean, according to Eq. 37, that ρ_k is zero. □

2.4. Impact of matching

The impact of matching on TE estimation bias and variance is captured in the following two theorems.

THEOREM 1. (Matching and TE bias) *In a SLM, TE estimation is unbiased if all included and omitted covariates have equal means across treatment and control groups.*

PROOF. The premise means $\mathbf{d}^i = \mathbf{d}^o = \mathbf{0}$. This, according to Corollary 1, means that $\boldsymbol{\rho}^i = \boldsymbol{\rho}^o = \mathbf{0}$. From Eq. 27a, we conclude that $\delta = 0$. □

While matching reduces TE estimation bias, it does so by discarding data and thus may increase variance. As discussed in the literature (Ho et al., 2007; Stuart, 2010), matching does not increase TE variance as rapidly as random subsampling of data would. The following theorem formalizes this intuitive and empirical notion.

THEOREM 2. (Matching and TE variance) *Among all data sets with a given treatment (N_t) and control (N_c) group size, generated from a SLM with noise variance of σ_0^2, a data set with balanced distribution of included covariates achieves the lowest (potentially misspecified) OLS-based estimation variance for TE, and this minimum variance equals $\sigma_0^2 (1/N_c + 1/N_t)$.*

PROOF. Since within-group covariance matrices, $\text{cov}(\mathbf{Z}^i)_T$ and $\text{cov}(\mathbf{Z}^i)_C$, are positive semi-definite (Gentle, 2007), their weighted sum, $\mathbf{A}$, in Eq. 23 is also positive semi-definite. Therefore, $\mathbf{A}^{-1}$ is positive definite (with eigenvalues being inverse of eigenvalues for $\mathbf{A}$). Therefore, by definition of positive-definiteness, $\boldsymbol{u}^{i,t}\mathbf{A}^{-1}\boldsymbol{u}^i > 0$, $\forall \boldsymbol{u}^i \neq \mathbf{0}$. The equality happens when $\boldsymbol{u}^i = \mathbf{0}$, i.e. when included covariates are balanced, leading to $\sigma^2 = \sigma_{min}^2$. As shown in Supplementary Material, sampling with replacement increases variance, and thus the minimum variance established in this theorem applies to matching with replacement as well. □

A few observations are worth mentioning with regards to the bias/variance expressions and the resulting theorems:

1. For TE estimation to be unbiased, not only included but also omitted covariates must be balanced. Since we don't know what omitted covariates are (otherwise we would include them in regression and eliminate bias), we generally cannot ensure they are balanced. An exception is perfect matching where, e.g. each treatment observation is perfectly matched - with respect to all original covariates - with a control observation. But thanks to ignorability assumption, we are ensured that all omitted covariates are also functions of included covariates. Therefore, perfect matching on included covariates

would automatically lead to perfect matching on omitted covariates. But perfect matching is often impractical, either because some covariates are continuous, or because perfect matching leads to an excessively small sample size and hence large variance, or both. Rubin (1973) similarly observe that matching on covariates in a linear regression or ANCOVA analysis does not guarantee bias removal for nonlinear response surfaces.

2. In the absence of perfect matching, we must accept the fact that TE will be biased, and instead do our best to minimize it. Again, our main challenge is that, according to Eq. 27a, bias is a function of included as well as omitted covariates. This represents a key difference between bias and variance equations, since variance, according to Eq. 27b only depends on included covariates. In other words, quantifying bias is a more difficult proposition than quantifying variance. In Section 3.2, we will present a framework for quantifying bias that is independent of outcome variable.

3. In Theorem 2, we showed that a data set that is balanced with respect to included covariates achieves σ^2_{min}. A second way to achieve minimum variance is to simply include no adjustment covariates in the regression, i.e. only have intercept and treatment indicator. This is the simple difference method mentioned in Section 1.2. As we saw in Figure 1, while this approach has low variance, using it with an unmatched data often produces a large enough bias that more than offsets the low variance and leads to large MSE.

2.5. TE bias and orthogonality

Neither the standard (Eqs. 26a and 26b) nor the normalized (Eq. 27a and 27b) expressions for bias make an intuitive concept clear: An omitted covariate that is very 'similar' to the included covariates will not induce much bias, since its contribution to the outcome will be captured by the coefficient of the included covariate acting as a surrogate for the coefficient of omitted covariate. This concept can be formally stated in the following theorem:

THEOREM 3. (TE Bias and orthogonality) *Projections of omitted covariates onto the subspace spanned by included adjustment covariates and intercept variable produce zero TE bias, i.e. only components of omitted covariates orthogonal to the aforementioned subspace contribute towards bias.*

PROOF. Equation 15 can be expanded in terms of columns of $\mathbf{Z}$:

$$\mathrm{E}[\hat{\boldsymbol{\beta}}] - \boldsymbol{\beta} = (\mathbf{X}^{i,t}\mathbf{X}^i)^{-1}\mathbf{X}^t \sum_{k=1}^{K^o} \mathbf{z}^o_{,k}\gamma_k. \tag{38}$$

where $\mathbf{z}^o_{,k}$ is the k'th column of $\mathbf{Z}^o$ and γ_k is its coefficient. To prove our theorem, and recalling our convention for arranging covariates in $\mathbf{X}^i$ according to Eq. 7, we simply prove that the matrix operator $(\mathbf{X}^{i,t}\mathbf{X}^i)^{-1}\mathbf{X}^{i,t}$ acting on any one of K columns of $\mathbf{Z}^i$, i.e. $\mathbf{z}^i_{,k}$, produces a K-dimensional vector whose first element is zero. This is easy to see since, by definition, $(\mathbf{X}^{i,t}\mathbf{X}^i)^{-1}\mathbf{X}^{i,t}\mathbf{X}^i = \mathbf{I}_{K^i+2}$, where $\mathbf{I}_{K^i+2}$ is the identity matrix of dimensions $K^i + 2$. Therfore, $(\mathbf{X}^{i,t}\mathbf{X}^i)^{-1}\mathbf{X}^{i,t}\mathbf{z}^i_{,k}$ equals the $(k+2)$'th column of the identity matrix $\mathbf{I}_{K+2}$, whose first two elements will always be zero. Therefore, omission of any linear combination of $\mathbf{z}^i_{,k}$'s will produce zero TE bias.   □
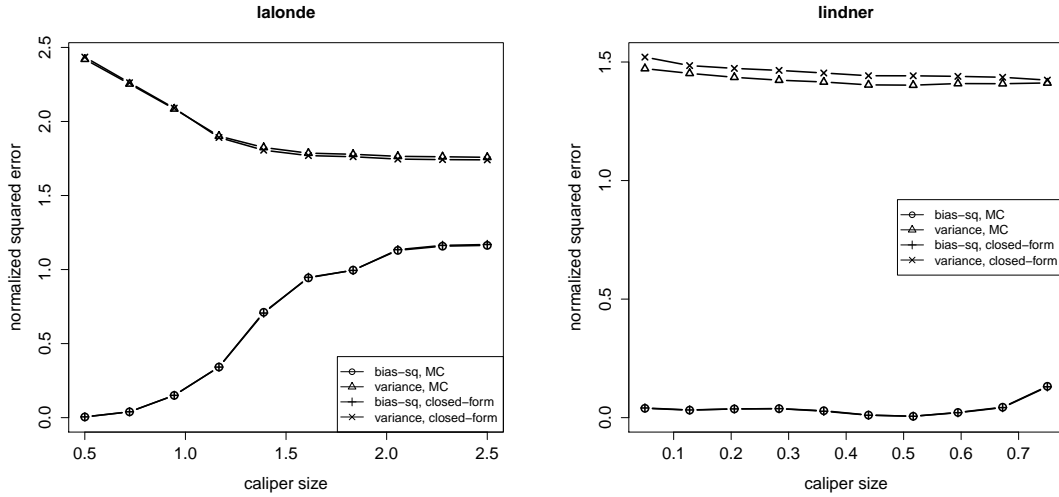
Fig. 3: Comparison of TE bias and variance calculations using Monte Carlo simulations (10,000 iterations per data point) and closed-form expressions of our framework for lalonde (left) and lindner (right) data sets. MC calculations took ~10-15min, while closed-form calculations took less than 0.1sec. A combination of matching and linear regression was used. Data generation and matching parameters are the same as Figure 1.

This theorem will be used in Section 3.2 to develop the constrained bias estimation approach.

3. Applications

In Section 2, we developed the mathematical framework for quantifying TE bias and variance, as well as the impact of matching, in a SLM (Section 2.2). In addition to providing a theoretical basis for understanding the combined use of matching and linear regression, this framework can be utilized towards developing diagnostic and calibration tools for causal inference. Such applications are the subject of this section.

3.1. Fast, accurate and generalizable simulations

The first application of the closed-form expressions for bias and variance (Eqs. 26a, 26b, 27a, and 27b), is fast and accurate simulations of the combined effect of matching and variance without requiring lengthy Monte Carlo runs. Figure 3 compares the bias and variance calculations using our closed-form expressions - encoded in the R package MatchLinReg (see Supplementary Material) - as well as Monte Carlo simulations, using 10,000 iterations to generate each point. While closed-form expressions produce results nearly instantaneously, the MC approach took several minutes on an average laptop, despite the data sets being quite small (fewer than 1000 observations). For larger data sets, the advantage of our framework becomes more prominent. We see that even 10,000 Monte Carlo runs can produce results that are somewhat different from theoretical values, particularly for variance. Our efficient simulation tool allows for exhaustive exploration of data sets and parameter spaces in future research, ultimately leading to better empirical rules for combining matching and linear regression.

In addition to speed of simulations, an equally-important contribution of our framework is that it provides for better generalizability of simulation results since we have a theoretical basis for what are the drivers of each error component. For example, Eqs. 26a and 26b make it clear that TE bias and variance do not depend on coefficients of adjustment covariates included in the regression model. Therefore, this aspect of data sets can be safely ignored while comparing their simulation results across studies. The enhanced generalizability of simulations resulting from utilizing our framework directly addresses the requirement stated in Imbens (2004). This paper is the first beneficiary of the framework, as all the simulations in the rest of this paper utilize the aforementioned closed-form expressions.

3.2.   Quantifying bias

A key challenge in assessing the benefits of matching is that TE bias is a function of the unknown, omitted covariates and their coefficients (Eqs. 26a and 27a) and hence any bias-removal impact of matching is similarly unknown, and hard to quantify. Naively, it might seem like a reasonable idea to use 'data exploration' to identify such omitted covariates and the strength of their impact, e.g. using pairwise correlation analysis of various candidate terms with outcome. Such explorations, however, can lead to overfitted models, particularly in the hands of inexperienced practitioners (Babyak, 2004; Ho et al., 2007). Furthermore, if we could somehow learn what the omitted covariates are, we could simply include them in the regression model and eliminate the bias that way, rather than using matching. Instead of using outcome data, we opt for a dual approach:

1. Bias-oriented analysis: We calculate maximum TE bias, normalized by the upper bound on the amount of unexplained variation in outcome that is due to omitted covariates, and where the vector of contributions from omitted covariates lies within a pre-specified subspace. This normalized bias is then combined with TE variance over a range of values for the bias-to-variance conversion factor, which we call 'omitted R-squared'. This analysis begins with a set of diagnostic tools (Sections 3.2.1, 3.2.2 and 3.2.3), and culminates in a calibration tool (Section 3.3). As it emphasizes bias, this analysis favors matching.

2. Variance-oriented analysis: We assume the other extreme with regards to the strength of impact of omitted covariates on outcome, namely that it is zero. Assuming a correct model specification - and hence zero bias - we calculate the reduced study power resulting from the choice of matching parameters in previous analysis (Section 3.4). This represents a worst-case-scenario as far as the negative impact of matching on TE estimation error.

Practitioners must weight the benefits of matching (from the first analysis) against the worst-case-scenario (from the second analysis) to determine the best course of action. An interesting area for future research is to combine these two facets into a unified framework, possibly using risk-based Bayesian methods to offer an even more prescriptive path towards matching calibration for practitioners.

With the above roadmap in mind, we begin developing a methodology for quantifying TE bias in misspecified regression. Eq. 26a can be re-written as follows:

$$\delta = \mathbf{g}^t \, \mathbf{Z}^o \boldsymbol{\gamma}^o = \sum_{k=1}^{K^o} \mathbf{g}^t (\gamma_k^o \, \mathbf{z}_{,k}^o), \tag{39}$$

where

$$\mathbf{g} \equiv \left(\frac{1}{N_t} + \frac{1}{N_c} + \boldsymbol{u}^{i,t} \mathbf{A}^{-1} \boldsymbol{u}^i \right) \mathbf{w} + \left(-\frac{1}{N_c} + \frac{1}{N_c} (\boldsymbol{p} - \boldsymbol{q})^t \mathbf{A}^{-1} \boldsymbol{u}^i \right) \mathbf{1} + \mathbf{Z}^i \mathbf{A}^{-1} \boldsymbol{u}^i, \tag{40}$$

and $\mathbf{g}$ is a vector of length N. Eq. 39 means that the contribution of each omitted covariate towards TE bias is 1) additive, 2) proportional to its coefficient, and 3) dependent on how parallel it is with $\mathbf{g}$.

While we do not know exactly what the omitted covariates are (otherwise they would have been included), we can often form a reasonable candidate set, e.g. by forming interactions (second-order or higher) and powers of the included covariates. Given the additive property of bias, studying the impact of matching on bias due to each potentially-omitted covariate is a meaningful first step.

3.2.1. Single-covariate relative squared bias reduction

A first diagnostic tool is to compare (squared) bias caused by a candidate omitted covariate, before and after matching. Such a relative bias reduction metric will be independent of γ_k^o:

$$\frac{\delta_{\tau,k}^{2,i} - \delta_{\tau,k}^{2,f}}{\delta_{\tau,k}^{2,i}} = \frac{||\mathbf{g}^t \, \mathbf{z}_{,k}^{o,i}||^2 - ||\mathbf{g}^t \, \mathbf{z}_{,k}^{o,f}||^2}{||\mathbf{g}^t \, \mathbf{z}_{,k}^{o,i}||^2} \tag{41}$$

where i and f superscripts refer to before and after matching, respectively. This quantity (or its square) can be produced for all candidate omitted terms, producing plots similar to those shown in Figure 4. Ideally, we would like matching to eliminate, or at least reduce, bias for all terms, but the figure shows that this may not happen.

3.2.2. Comparison of single-covariate normalized biases

While relative bias reduction is independent of omitted coefficients ($\boldsymbol{\gamma}^o$), yet it does not provide any comparison of the magnitude of biases induced by each omitted covariate. Matching may not significantly reduce the bias due to an omitted covariate, yet if the overall bias contributed by that covariate is likely to be small, lack of bias reduction by matching for that covariate becomes unimportant.

Since each term's contribution is proportional to its coefficient (γ_k^o), we must somehow determine these coefficients. As emphasized before, we opt for an approach that does not use the outcome variable. If omitted covariates have been orthogonalized with respect to $\{\mathbf{1}, \mathbf{Z}^i\}$ (the subspace of included adjustment covariates plus intercept), their mean squared contribution to outcome is roughly the same as the change in R-squared of the regression model, multiplied by σ_0^2. This motivates us to define a parameter called 'omitted R-squared': it is the ratio of mean squared contribution to outcome for an omitted covariate (or a set of covariates), after orthogonalization, to noise variance:

$$R_o^2 = ||\mathbf{Z}_\perp^o \boldsymbol{\gamma}^o||^2 / \sigma_0^2, \tag{42}$$
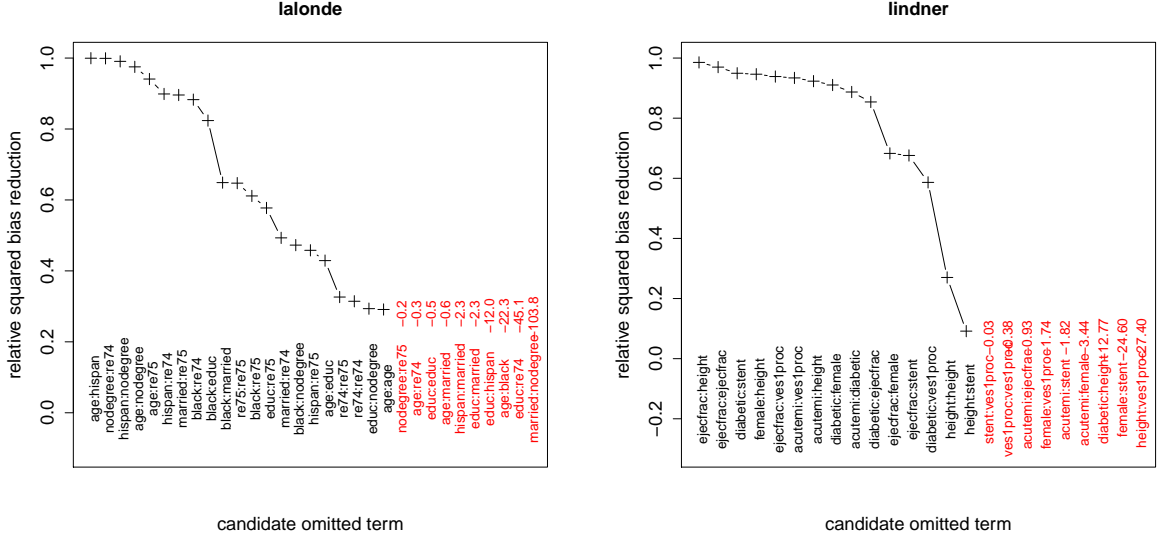
Fig. 4: Relative squared bias reduction (due to matching) for all second-order terms as candidate omitted covariates. Results for lalonde (left) and lindner (right) data sets are shown. Regression adjustment using all main effects ('before') is supplemented by PSM as a pre-processing step ('after'), using caliper size of 1.5 for lalonde and 0.2 for lindner. Omitted terms are sorted by decreasing relative bias reduction. Negative numbers (on the right of each plot) reflect terms for which matching has increased squared bias.

where $\mathbf{Z}_\perp^o$ is the component of $\mathbf{Z}^o$ (each column, to be precise) orthogonal to the subspace spanned by $\{\mathbf{1}, \mathbf{Z}^i\}$. If included covariates are all linear terms (in original covariates), then one can also think of R_o^2 as the degree of nonlinearity in the data-generation process. We must reiterate that R_o^2 is not calculated using outcome data, and the terms R-squared is used loosely to describe this parameter. Rather, this parameter must be determined based on a domain expert's knowledge of the underlying mechanisms involved as well as experience and future simulation studies. We will use this parameter in Section 3.3 to put bias on the same scale as variance and combine them to arrive at total MSE for TE.

Given the above, we propose a constrained bias estimation approach where we estimate bias, subject to the constraint that mean squared contribution from omitted covariates equals a given value. In fact, we can canonically set this value to 1 and estimate a normalized bias value:

$$\delta^n = \sqrt{N}\, \mathbf{g}^t\, \mathbf{Z}_\perp^o \boldsymbol{\gamma}^o \,/\, ||\mathbf{Z}_\perp^o \boldsymbol{\gamma}^o|| \tag{43}$$

Normalized bias has an intuitive interpretation: it is the fraction of unexplained (or omitted) variation in outcome (or 'signal') that turns into TE bias. For a perfectly-matched data set, this fraction is zero.

Within this general approach, several variations can be conceived of. We begin with the simplest approach, where we calculate and compare normalized bias for each of the candidate omitted covariates. In other words, we permit $\boldsymbol{\gamma}^o$ to have exactly one non-zero element. This produces K^o

normalized bias numbers, δ_k^n:

$$\delta_k^n = \sqrt{N}\, \mathbf{g}^t\, \gamma_k^o \mathbf{z}_{\perp,k}^o \,/\, ||\gamma_k^o \mathbf{z}_{\perp,k}^o||,$$
$$= \pm \sqrt{N}\, \mathbf{g}^t\, \mathbf{z}_{\perp,k}^o / ||\mathbf{z}_{\perp,k}^o||. \tag{44}$$

We can now produce and compare these single-covariate, normalized biases for a set of candidate omitted terms, perhaps before and after matching, and for different calibrations of matching to identify nonlinearities that can potentially induce large residual bias, even after matching. Since absolute values are more important than signs, we focus on squared bias, a metric which is also dimensionally compatible with variance and total MSE.

An example is provided in Figure 5, where single-covariate normalized square bias has been plotted for a handful of second-order interactions as a function of matching caliper size. We see that matching does manage to reduce the normalized bias for the biggest offenders. A notable exception is the *female:stent* interaction in the lindner data set, where matching has raised the bias to relatively significant level after matching ($\sim 1\%$ of omitted variance). Also, note that the overall normalized bias levels are significantly higher in lalonde data set, compared to lindner. In other words, a larger fraction of omitted signal translates into TE bias in lalonde data set. Figure 5 also illustrates that impact of matching on covariate imbalance (measured via absolute mean difference) does not have a simple relationship with bias reduction.

3.2.3. Estimating aggregate bias

While producing and examining normalized squared biases for potentially-omitted terms is insightful, it stops short of producing a single, aggregate measure of bias, which we would need to combine with variance to arrive at an estimated MSE. We consider three aggregation methods below. They all produce a normalized bias estimate (Eq. 44), and are all based on maximizing normalized bias within a given, eligible subspace. They differ in how this eligible subspace is constructed.

Single-covariate maximization: Eligible subspace is one of K^o candidate omitted covariates. The aggregate bias estimate is simply the maximum value of normalized bias terms due to each candidate, omitted covariate, $\{\delta_{\tau,k}^n\}$.

Covariate-subspace maximization: Here the eligible subspace is the entire covariate subspace that is orthogonal to $\{\mathbf{1}, \mathbf{Z}^i\}$, and the aggregate bias estimate is the linear combination of omitted covariates - after orthogonalization - that maximizes bias, subject to normalization constraint. This bias estimate is bound to be greater than or equal to the output from single-covariate maximization approach. The direction of $\mathbf{Z}^o \boldsymbol{\gamma}^o$ must be parallel to the projection of $\mathbf{g}$ in the omitted covariate subspace, after orthogonalization ($\mathbf{g}_\parallel$):

$$\delta = \sqrt{N}\, \mathbf{g}^t \mathbf{g}_\parallel / (\mathbf{g}_\parallel^t \mathbf{g}_\parallel)^{1/2} \tag{45}$$

Absolute maximization: The eligible subspace is the entire subspace orthogonal to $\{\mathbf{1}, \mathbf{Z}^i\}$, and the aggregate bias estimate in this case is the absolute maximum normalized bias achievable. Here the
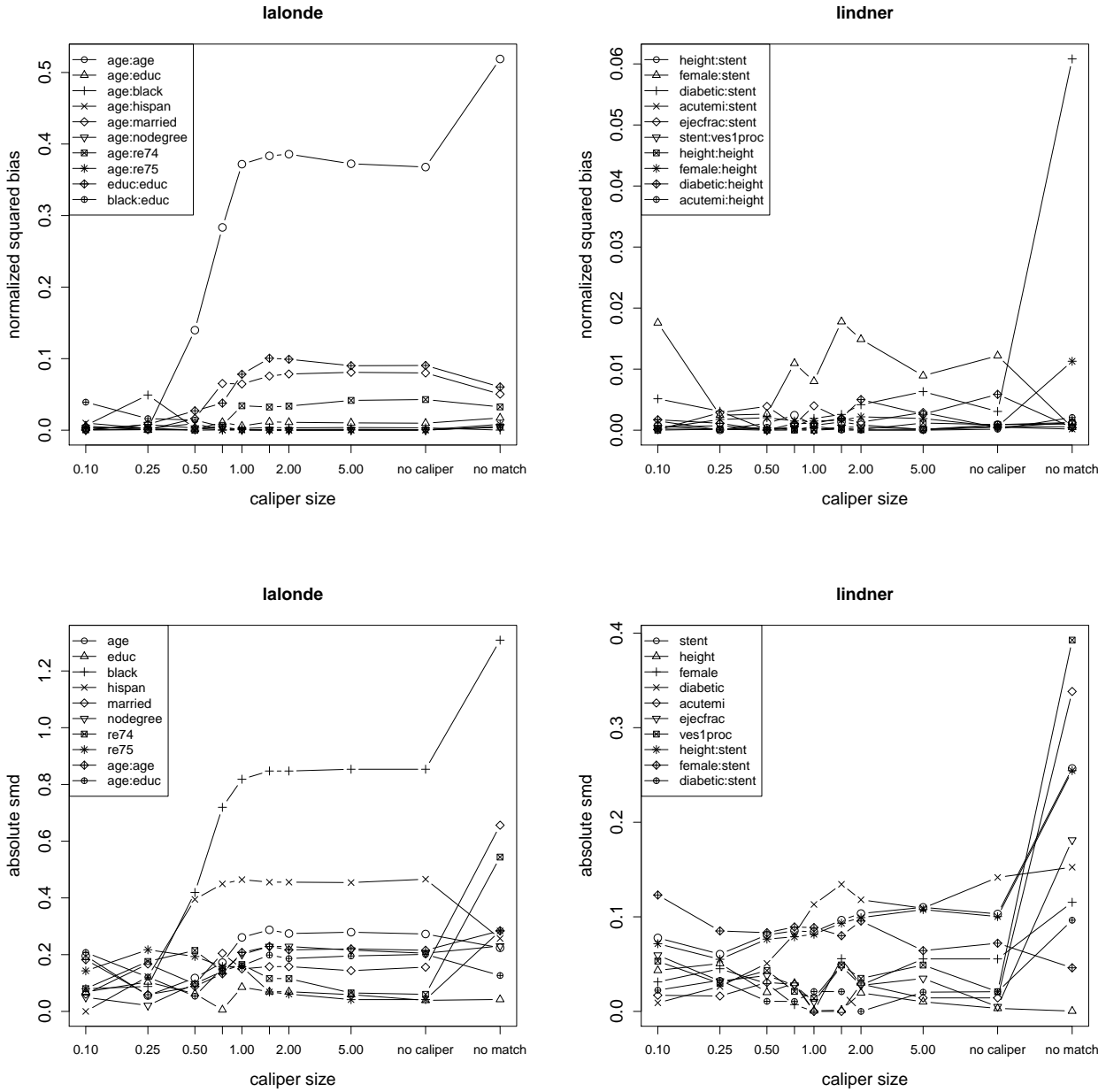
Fig. 5: Top row: Comparison of normalized squared biases due to omission of 10 second-order terms from regression adjustment for lalonde (left) and lindner (right) data sets. Horizontal axis is caliper size used in PSM, with the right-most point on each plot corresponding to no matching. Bottom row: Impact of matching caliper size on absolute mean differences for same 10 interaction terms.
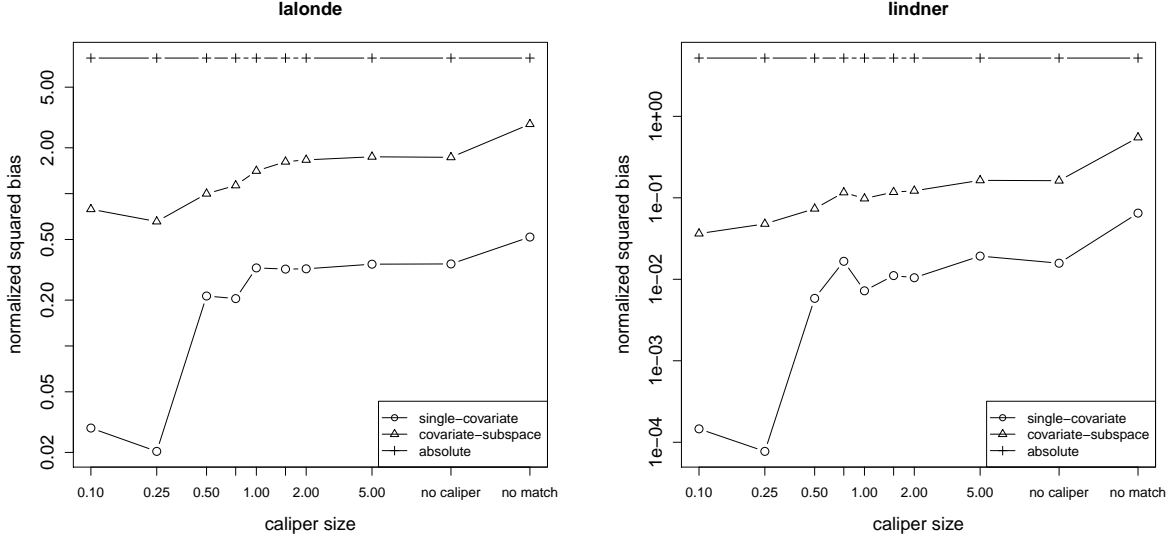
Fig. 6: Impact of matching caliper size on TE normalized squared bias using single-covariate maximization, covariate-space maximization, and absolute maximization for lalonde (left) and lindner (right) data sets. Set of candidate omitted covariates includes second-order and third-order interactions among original covariates, as well as step functions using cuts at 25%, 50% and 75% percentiles for continuous covariates with at least 20 distinct values.

direction of $\mathbf{Z}^o\boldsymbol{\gamma}^o$ is simply parallel to $\mathbf{g}$:

$$\delta = \mathbf{g}^t(\frac{\mathbf{g}}{||\mathbf{g}||}\sqrt{N}) = \sqrt{N}||\mathbf{g}||, \tag{46}$$

$$= \sqrt{N}\,\frac{\sigma}{\sigma_0}. \tag{47}$$

Figure 6 compares the bias estimated for these three methods, along with impact of matching on bias reduction for each method. Of these three measures, we believe the covariate-subspace approach is the most suitable approach. While absolute maximization is too extreme as far as bearing no connection to adjustment covariates, single-covariate maximization approach is too rigid in focusing on a very limited set of possible combinations of candidate omitted covariates. Interestingly, bias estimate using absolute maximization does not change with matching. This property is proven in Supplementary Material.

While the absolute maximization approach is independent of the set of candidate omitted covariates considered, $\mathbf{Z}^o$, the other two approaches - single-covariate maximization and covariate-subspace maximization - rely on the choice of $\mathbf{Z}^o$. While this topic requires further research, we propose that a good starting point is interaction terms and binary (step) functions formed from discretizing continuous variables. In generating Figures 6 and 7, we have used second-order and third-order interaction terms, as well as step functions formed at 25%, 50% and 75% percentiles of continuous variables with more than 20 distinct values. Increasing the number of covariates, e.g. by adding to maximum order of interaction terms, will have diminishing value since it approaches the absolute maximum approach.

However, this is not a case of overfitting and obtaining extreme bias numbers, thanks to the constraint put on the size of the omitted signal.

It is possible to replace maximization with another operator - such as averaging - in each of the above methods. The fact that using the maximum operator tends to exaggerate the impact of bias motivates us to present a counter-view in which we assess the damage caused by matching towards variance increase and the resulting loss of study power, assuming a total absence of bias. This is described in Section 3.4.

3.3. Combining bias and variance

In Section 3.2 we developed the constrained bias estimation method for arriving at 'normalized squared bias', i.e. the ratio of squared bias to omitted signal. On the other hand, Eqs. 26b and 27b allow us to calculate 'normalized variance', i.e. the ratio of TE variance to σ_0^2, in a straightforward manner. In order to combine normalized bias and variance to arrive at MSE, we must have an 'exchange rate'. We propose an intuitive parameter, 'omitted R-squared', defined as the ratio of omitted signal to generative noise. If we define 'normalized MSE' as the ratio of MSE to σ_0^2, we have:

$$\text{normalized MSE} = \text{normalized variance} + R_o^2 \times \text{normalized squared bias} \qquad (48)$$

Figure 7 shows normalized MSE as a function of matching caliper size for various values of R_o^2, for lalonde and lindner data sets. For small values of R_o^2, variance dominates bias and thus optimal caliper size (that minimizes MSE) tends to be large (or we may choose no caliper, or no matching) (top row). As R_o^2 gets larger, bias and variance become comparable and thus the bias-removal impact of matching outweighs its variance increase; thus optimal caliper size shifts towards smaller values (second row). For even larger R_o^2 (third row), bias dominates variance and even smaller caliper sizes become optimal. This trend is seen more clearly in the bottom row, where optimal caliper size has been plotted as a function of R_o^2. These plots can be considered the most prescriptive of all our analyses so far: they suggest a particular calibration (caliper size here) assuming different values of omitted R-squared. Same type of analysis can be applied to other aspects of matching, e.g. what terms to include in matching, PSM vs. Mahalanobis matching, etc.

How should practitioners go about selecting a value (or a range of values) for omitted R-squared? The answer is, partly domain expertise, and partly experience. In particular, if outcome data is available, practitioners can calculate R-squared for a main-effect-only regression model, and then reason about how much they expect nonlinear effects to improve upon this R-squared. For example, assume that the main-effect R-squared is 14%, and that we believe nonlinear effects contribute about 30% towards explaining data variance. Then we can reasonably assume that $R_o^2 = 30\% \times 14\% \sim 4\%$. Some researchers and practitioners may approve of such a controlled use of outcome data.
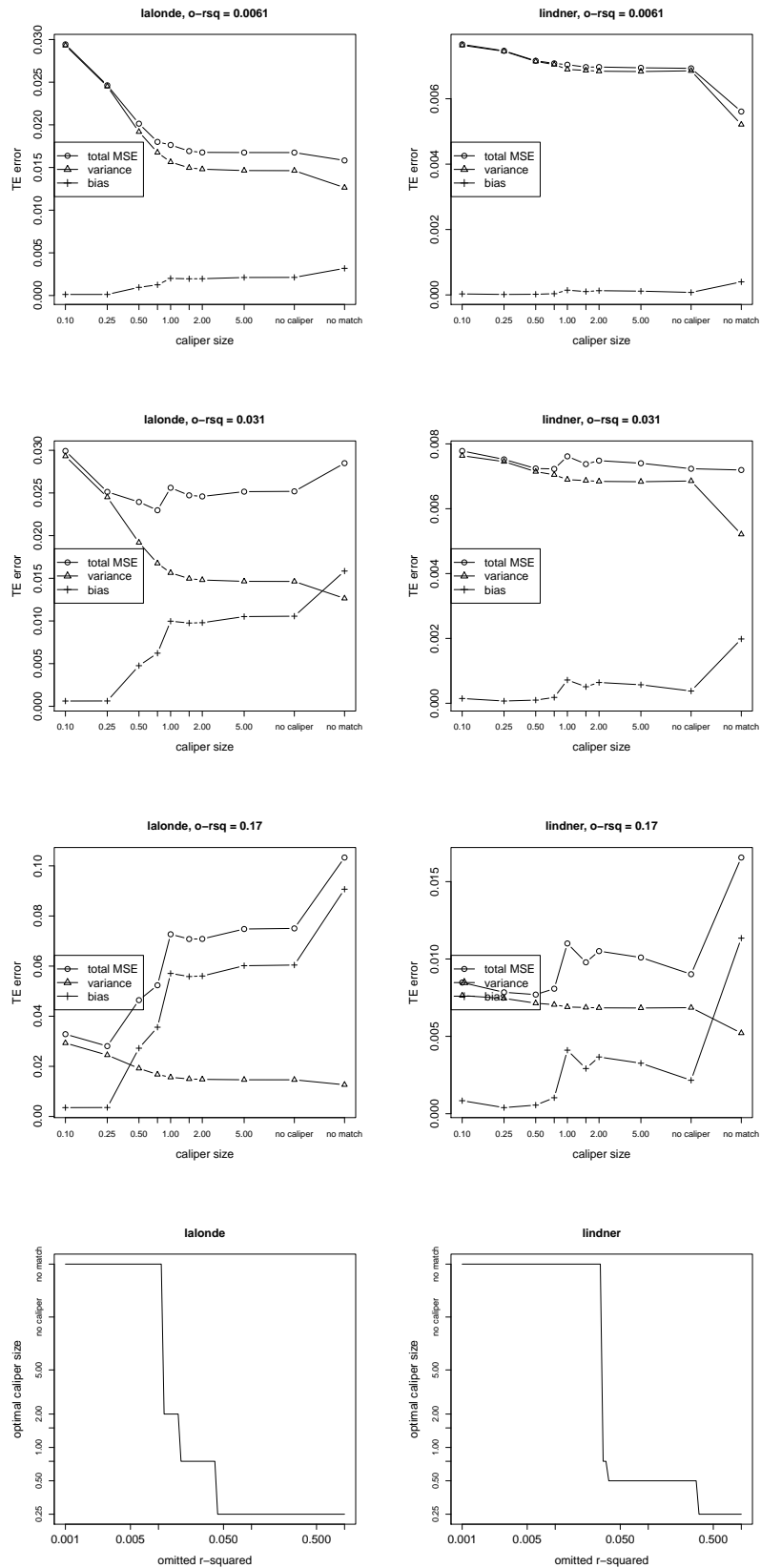
Fig. 7: Matching calibration by combining bias and variance and minimizing MSE, for lalonde (left) and lindner (right) data sets. Top three rows: TE error and its components as a function of matching caliper size, for three levels of omitted R-squared (0.5%, 3%, and 17% starting from top). Bottom row: Optimal caliper size as a function of omitted R-squared (R_o^2). Set of candidate omitted covariates is identical to that of Figure 6.
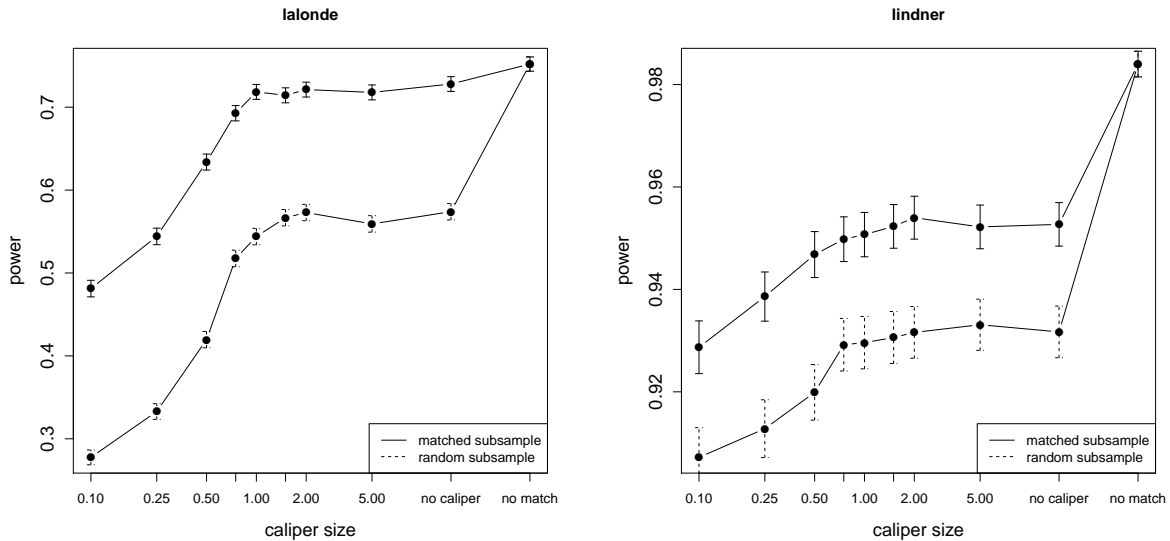
Fig. 8: Impact of matching caliper size on study power for lalonde (left) and lindner (right) data sets (mean +/-2 stdev). Effect size is defined as ratio of TE to standard deviation of noise (σ_0). An effect size of 0.3 is used for both plot. 10000 Monte Carlo iterations were used to generate each data point.

3.4. Power analysis

In covariate-subspace maximization approach to constrained bias estimation (Section 3.2.3), we erred on high bias estimation in our bias-variance trade-off analysis by using the maximum operator on the eligible subspace. In this section, we take the opposite view by assuming no bias, and assess the negative impact of matching in terms of variance increase, or equivalently study power decrease.

In power analysis for TE estimation using linear regression adjustment, definition of effect size must reflect our focus on TE, i.e. the coefficient τ (Eq. 5). A natural definition is τ/σ_0, i.e. the ratio of TE to standard deviation of generative noise. Based on this definition, we have implemented a Mont-Carlo based approach to power calculation (function **mlr.power** in R package MatchLinReg). Figure 8 shows study power as a function of caliper size for lalonde and lindner data sets. Effect size is set to 0.3 for both plot. In each case, we used 10,000 Monte Carlo simulations to generate mean and standard deviation values for study power. For comparison, we have also plotted study power when data is subsampled randomly with the same sizes as the matched case from treatment and control groups. The significantly higher study power for matched subsamples is an illustration of the minimum-variance property of matching (Theorem 2). When designing clinical trials, power calculations based on observational studies must be performed after matching to increase the likelihood that TE bias has been significantly reduced and thus power calculations are valid. Our software can be of use in such design settings.

4. Discussion

In this paper, we developed a framework for quantifying the combined effect of matching - as a non-parametric pre-processing technique - and linear regression for estimating TE for causal inference. In addition to providing a theoretical basis for the impact of matching on TE bias and variance in a misspecified linear regression, the applicability of our framework is strengthened by four attributes: 1) a finite-sample focus (as opposed to large-sample or asymptotic analysis) makes our results directly applicable to real-world problems with small control and/or treatment groups, 2) quantifying not just bias but also variance of TE estimation allows us to minimize total MSE, which is the ultimate measure of estimation accuracy in a single experiment, 3) exclusion of the outcome variable from the equations ensures that the use of our diagnostic and calibration tools by the broad community of practitioners does not lead to overfit results that have questionable prognostic value, and 4) capturing the entire framework along with calibration and diagnostic tools in an open-source software, R package MatchLinReg, allows researchers and practitioners to utilize the currently provided set of tools for their observational research, extend - and experiment with - the framework, and implement new diagnostic and calibration functions that better suit their application domains.

Some of the limitations and assumptions of the current work provide for interesting research opportunities. Below we highlight two key areas: *1) Generalized Linear Models (GLMs):* The current framework is focused on, and takes full advantage of, linear models with a continuous (unbounded) outcome variable. This facilitated the derivation of closed-form expressions for bias and variance. Maximum-Likelihood (ML) estimation of GLM coefficients does not generally contain a closed-form solution. The nonlinear link function creates complex interactions between the coefficients of adjustment covariates (not just omitted but also included) and TE bias and variance. Given the significant percentage of applied studies that utilize non-linear models, including logistic regression, Poisson regression and survival models, it is important to generalize this framework to include such common applications. *2) Large regressions as alternative to matching:* An alternative to matching for removing covariate omission bias is to simply include all candidate omitted terms in the regression model. Any nuisance covariates, i.e. those not present in generative model but included in the regression model, do not contribute to bias - as we can simply assume their corresponding coefficients are zero - but only increase variance. It may appear that increased variance due to inclusion of all candidate terms may overwhelm the regression, thus becoming an unrealistic candidate. However, our early simulations and analysis suggest that this may not be the case, and this topic deserves further research.

References

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

— (2011) Bias-corrected matching estimators for average treatment effects. *Journal of Business &*

Economic Statistics, **29**.

Austin, P. C. (2007) Propensity-score matching in the cardiovascular surgery literature from 2004 to 2006: a systematic review and suggestions for improvement. *The Journal of Thoracic and Cardiovascular Surgery*, **134**, 1128–1135.

— (2009) Some methods of propensity-score matching had superior performance to others: Results of an empirical investigation and monte carlo simulations. *Biometrical Journal*, **51**, 171–184.

Austin, P. C., Grootendorst, P. and Anderson, G. M. (2007) A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a monte carlo study. *Statistics in medicine*, **26**, 734–753.

Babyak, M. A. (2004) What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. *Psychosomatic medicine*, **66**, 411–421.

Carpenter, R. (1977) Matching when covariables are normally distributed. *Biometrika*, **64**, 299–307.

Gayat, E., Resche-Rigon, M., Mary, J.-Y. and Porcher, R. (2012) Propensity score applied to survival data analysis through proportional hazards models: a monte carlo study. *Pharmaceutical statistics*, **11**, 222–229.

Gentle, J. E. (2007) *Matrix algebra: theory, computations, and applications in statistics*. Springer Science & Business Media.

Glazerman, S., Levy, D. M. and Myers, D. (2003) Nonexperimental versus experimental estimates of earnings impacts. *The Annals of the American Academy of Political and Social Science*, **589**, 63–93.

Heckman, J. J., Ichimura, H. and Todd, P. E. (1997) Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *The review of economic studies*, **64**, 605–654.

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**, 199–236.

Hosman, C. A., Hansen, B. B. and Holland, P. W. (2010) The sensitivity of linear regression coefficients'confidence limits to the omission of a confounder. *The Annals of Applied Statistics*, 849–870.

Imbens, G. W. (2004) Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and statistics*, **86**, 4–29.

Kereiakes, D. J., Obenchain, R. L., Barber, B. L., Smith, A., McDonald, M., Broderick, T. M., Runyon, J. P., Shimshak, T. M., Schneider, J. F., Hattemer, C. R. et al. (2000) Abciximab provides cost-effective survival advantage in high-volume interventional practice. *American heart journal*, **140**, 603–610.

Lechner, M. (1999) Earnings and employment effects of continuous gff-the-job training in east germany after unification. *Journal of Business & Economic Statistics*, **17**, 74–90.

— (2002) Program heterogeneity and propensity score matching: An application to the evaluation of active labor market policies. *Review of Economics and Statistics*, **84**, 205–220.

Mahani, A. S. and Sharabiani, M. T. (2015) *MatchLinReg: Combining matching and linear regression for causal inference.* URLhttp://CRAN.R-project.org/package=MatchLinReg. R package version 0.7.0.

Quade, D. (1982) Nonparametric analysis of covariance by matching. *Biometrics*, 597–611.

Robins, J. M. and Rotnitzky, A. (1995) Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, **90**, 122–129.

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.

Rotnitzky, A., Faraggi, D. and Schisterman, E. (2012) Doubly robust estimation of the area under the receiver-operating characteristic curve in the presence of verification bias. *Journal of the American Statistical Association*.

Rubin, D. B. (1973) The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics*, 185–203.

— (1976) Multivariate matching methods that are equal percent bias reducing, i: Some examples. *Biometrics*, 109–120.

— (1979) Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association*, **74**, 318–328.

Rubin, D. B. and Stuart, E. A. (2006) Affinely invariant matching methods with discriminant mixtures of proportional ellipsoidally symmetric distributions. *The Annals of Statistics*, 1814–1826.

Rubin, D. B. and Thomas, N. (1992) Affinely invariant matching methods with ellipsoidal distributions. *The Annals of Statistics*, 1079–1093.

— (1996) Matching using estimated propensity scores: relating theory to practice. *Biometrics*, 249–264.

— (2000) Combining propensity score matching with additional adjustments for prognostic covariates. *Journal of the American Statistical Association*, **95**, 573–585.

Schafer, J. L. and Kang, J. (2008) Average causal effects from nonrandomized studies: a practical guide and simulated example. *Psychological methods*, **13**, 279.

Stuart, E. A. (2010) Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics*, **25**, 1.

Tan, Z. (2010) Bounded, efficient and doubly robust estimation with inverse weighting. *Biometrika*, **97**, 661–682.