The Case Against Matching

Michael K. Miller

George Washington University

July 9, 2013

Abstract

Matching has become a common technique in the social sciences, but there is widespread confusion regarding its purposes. This article evaluates the three most prominent justifications for matching: to improve causal inference, to reduce model dependence, and to prevent bias from model misspecification. First, I argue that matching offers no causal leverage or advantage in dealing with selection relative to regression alone. Second, I show that matching increases model dependence and considerably widens the opportunity for data-mining. Claims to the contrary have mistakenly ignored the sensitivity of estimates to choices concerning the match itself. Third, matching can reduce treatment bias when misspecification combines with covariate imbalance. However, I argue this use is justified only after passing a suggested test for misspecification and only as a last resort when parametric modeling fails.

1 Introduction

Matching is an increasingly popular empirical technique throughout the social sciences, but its advantages are widely misunderstood. Numerous methodological articles have debated the appropriate techniques for matching, with a dizzying array of suggested methods

Assistant Professor, Department of Political Science, George Washington University. E-mail: Michael.Miller@anu.edu.au. Thanks to participants at ANU's econometrics workshop for helpful comments.

(e.g., Rosenbaum and Rubin 1983; Morgan and Harding 2006; Iacus et al. 2012; Diamond and Sekhon forthcoming). The focus of this article is not *how* matching should be done, but *why*. Given the many proponents of matching (Morgan and Harding 2006; Ho et al. 2007; Stuart and Rubin 2007), what is missing from the current literature is a critical perspective on its various justifications.

In brief, matching is most commonly used to test the effect of a binary treatment variable.[1] With observational data, treated and control cases usually differ in various ways besides treatment status, leading to a range of inferential problems. Matching selects out a new sample in which treated and control units are as similar as possible on a set of observed covariates. It serves to preprocess, or "prune," the data prior to running an estimator like regression. Thus, matching is not an alternative to regression so much as an optional first step designed to improve it (Ho et al. 2007).[2]

What exactly are the benefits of matching? Based on a survey of recent political science research, I find a great deal of confusion regarding its purposes, some of it shared by methodologists. This paper discusses the three most prominent justifications for matching, finding two to be erroneous and the last in need of caveats.

First, by far the most popular justification for matching is that it addresses the non-random selection of the treatment. In this view, comparing similar treated and control cases approximates an experimental design, making matching a powerful method of causal inference. As most methodologists recognize, this view is mistaken. Matching has no advantage relative to regression for proving causation or dealing with endogeneity, since matching can only account for observed covariates. I discuss what distinguishes matching from true methods of causal inference.

Second, largely due to Ho et al. (2007), an increasingly popular claim is that matching reduces model dependence, or the sensitivity of estimates to the (often arbitrary) modeling choices made by the researcher. However, Ho et al.'s (2007) support for this is flawed because

[1] There has been some work on matching for continuous treatments (Imai and van Dyk 2004), but applications remain relatively rare. I focus on matching for binary treatments.

[2] For brevity's sake, I will use "regression" to stand in for "OLS or a similar parametric estimator."

it ignores the model dependence generated by the choice of how to match. By this I do not mean the choice of matching procedure, but the very quantity of "balance" that matching tries to optimize. Taking this into account, I show that matching *increases* model dependence and widens the opportunity for data-mining. This is supported by a re-examination of one of Ho et al.'s (2007) empirical examples, as well as Monte Carlo simulations.

Third, matching guards against bias from model misspecification. Although rarely used by researchers as a justification, this is a true advantage of matching. However, this benefit is not without its costs. Matching does not fix misspecification, but simply constructs the sample so that the treatment estimate is unaffected by it. Therefore, I suggest two caveats: (a) to avoid its considerable negatives, matching should not be used unless a proposed test indicates that misspecification is actually present, and (b) matching should only be used as a last resort when attempts to correctly specify the model fail. In sum, matching offers a narrow and compromised benefit that is far afield from how it is currently justified by researchers.

2 An Overview of Matching

2.1 The Fundamentals

Suppose we have N observations (indexed by i), some of which receive a binary treatment T. This may be a vaccination, a visit from a campaign worker, or attendance at a job training program. To conceptualize the causal effect of T on an outcome Y, it is instructive to think in terms of *potential outcomes*. Define Y_{i1} as the outcome that i would have if it receives the treatment, and Y_{i0} as the outcome i would have otherwise.

Researchers are typically interested in the average treatment effect (ATE), defined as

$$\tau = E[Y_{i1} - Y_{i0}]. \tag{1}$$

The "fundamental problem of causal inference" is that we only observe one of the quantities Y_{i1} or Y_{i0} for each unit (Holland 1986). Therefore, estimating the ATE requires inferring what Y_i would have been had its treatment assignment been different.

With observational data, we cannot simply compare average outcomes of the treated and untreated units, since they have likely selected non-randomly into the treatment. To make headway, it is common to assume that a set of observed covariates $\mathbf{X}$ are the only variables that predict both the outcome and selection into treatment. This is often termed "unconfoundedness" or "selection-on-observables" in the matching literature, and is equivalent to an assumption of "no omitted variables" in the regression context. Formally, we assume that, conditional on $\mathbf{X}$, the potential outcomes are unrelated to the actual treatment assignment.

If we additionally assume independence across units[3] and "common overlap" (*i.e.*, a non-zero probability of treatment for every observed value of $\mathbf{X}$), we can easily show that the following is unbiased for the ATE:

$$\tau|\mathbf{X} = E[Y_i|T_i = 1, \mathbf{X}] - E[Y_i|T_i = 0, \mathbf{X}], \tag{2}$$

which includes no counterfactuals or potential outcomes. A sample ATE can then be calculated by averaging $\tau|\mathbf{X}$ over the sample distribution of $\mathbf{X}$.[4]

Matching attempts to limit the sample to units with values of $\mathbf{X}$ shared by units with a different treatment assignment. Looking at (2), this allows for a direct comparison of outcome differences to calculate $\tau|\mathbf{X}$. As a result, we do not have to posit a functional form for $E[Y_i|T_i, \mathbf{X}]$, whereas regression assumes that $E[Y_i|T_i, \mathbf{X}]$ can be estimated from a linear model. With continuous data, matching treated and control units with identical values on $\mathbf{X}$ (*exact matching*) is usually not feasible. Other techniques therefore match units that are as close as possible, with different methods mainly varying by how they define closeness.

Ultimately, the goal of matching is to "prune" the data and create a new sample with treated and control units that are as similar as possible on $\mathbf{X}$. The degree of similarity is referred to as *balance*, which can also be measured in different ways. Perfect balance is reached

[3] Specifically, we assume that the potential outcomes for i are independent of treatment assignment for all other units.

[4] Alternatively, we can calculate the average treatment effect on the treated (ATT) by averaging $\tau|\mathbf{X}$ over the sample distribution of $\mathbf{X}|(T_i = 1)$.

when the distributions of covariates are equivalent across the treated and control groups, or

$$p(\mathbf{X}|T = 0) = p(\mathbf{X}|T = 1), \tag{3}$$

where p is the sample probability. Although this ideal is rarely obtainable, matching attempts to improve balance as much as possible.

2.2 Forms of Matching

Methodologists have developed numerous analytical techniques to construct optimally balanced samples. Since my concern is the purpose of matching, I will not cover these varying methods in great detail.[5] However, it will be helpful to describe two of the most popular matching techniques, as I use them in empirical examples below.

The first, and older, technique is called *propensity score matching* (PSM) (Rosenbaum and Rubin 1983). It works by first estimating a probit or logit equation that predicts treatment assignment from $\mathbf{X}$. This generates a probability $p(\mathbf{X_i})$ for each unit called the propensity score. PSM then matches treated and control units that are as similar as possible on the propensity score. The key advantage of this method is that it simplifies the difficult multidimensional problem of matching on $\mathbf{X}$ to a comparison of a scalar probability.

A newer method is called *coarsened exact matching* (CEM) (Iacus et al. 2012). It works by first choosing several cutpoints for each variable contained in $\mathbf{X}$, which thereby classifies each value into one of multiple ranges. CEM then matches the treated and control units that are within the same ranges for every variable in $\mathbf{X}$. By doing this, CEM ensures that matched units have similar values and combinations of every variable.

Although there are many other varieties of matching, the goal is always the same: maximizing balance. In fact, despite the name, matching does not require the pairing of observations at all. This is merely a convenient way of finding a balanced sample.

[5] For more thorough reviews of various matching techniques and practical issues with matching, see Morgan and Harding (2006), Ho et al. (2007), Stuart and Rubin (2007), and Sekhon (2009).

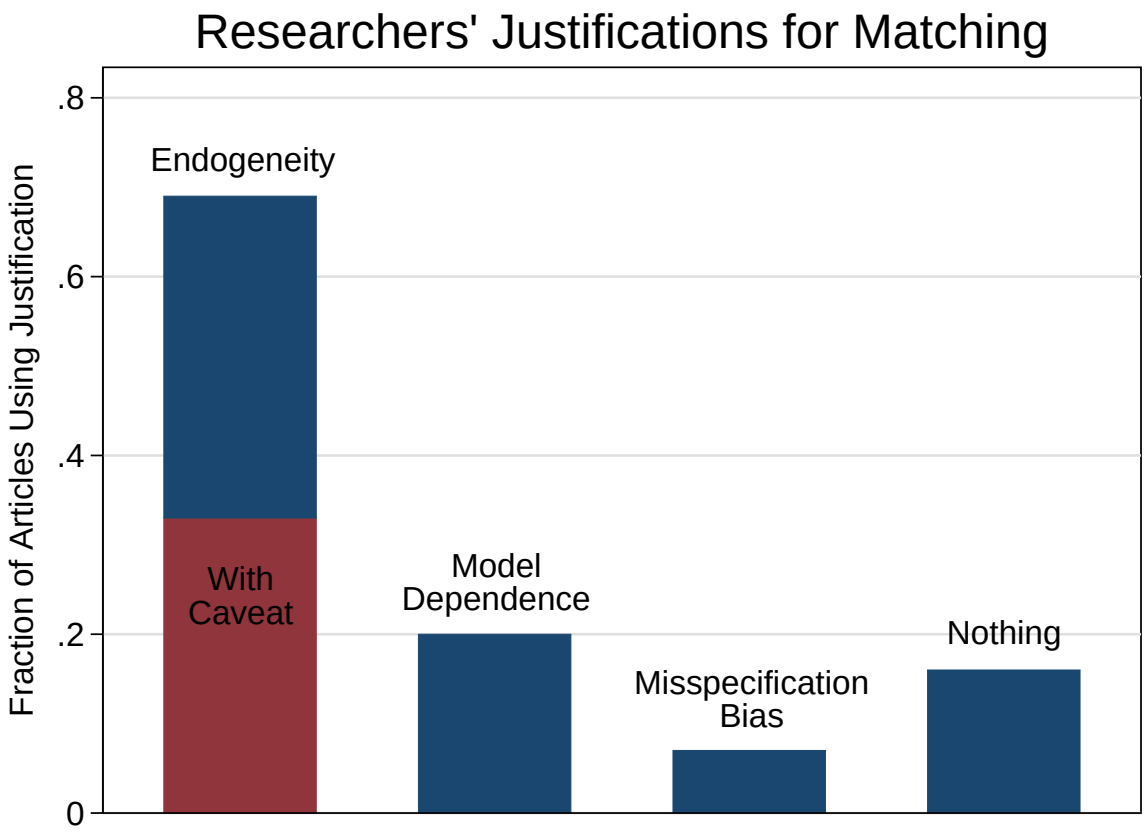

Fig. 1: Based on a survey of 61 articles using matching published since 2009, the figure shows the fraction of articles that reference each of the three justifications for matching analyzed in this article, as well as the fraction that offer no justification.

Finally, it must be stressed that matching is *not* a method of estimation. Rather, it is a way of constructing a new sample that can then be analyzed by an applicable estimator.[6] The original method was to simply compute a difference of means between the treated and control units after matching. Ho et al. (2007) instead recommend running a parametric estimator (such as regression) that controls for $\mathbf{X}$, which helps to adjust for any remaining imbalance. Since this is the most widespread current practice, I focus my discussion on this technique.

2.3 Justifications for Matching

The remainder of this article critically appraises the value of matching, but it is instructive to first take stock of how applied political scientists understand and justify the technique.

[6] Some apply the term "matching" to various weighting estimators and techniques like kernel matching (see Morgan and Harding 2006). However, I limit my discussion to matching as sample pruning.

To gauge this, I surveyed every article using matching published since 2009 in seven top political science journals.[7] The survey comprises 61 articles, of which 34 present matching as the primary technique and 27 as a robustness check.[8] The specific methods used are a balanced mix of propensity score matching, coarsened exact matching, and genetic matching, with nearly all of the articles using a parametric estimator after matching.

The papers vary widely in the explanations given for matching, as well as the depth of the discussions. Figure 1 displays the fraction of articles that include each of the three justifications analyzed in this article, as well as the fraction omitting any substantive explanation.

The most common justification by far is that matching is useful for dealing with endogeneity, specifically the non-random selection of the treatment. A form of this argument is included in roughly 70% of the articles. A common trope is that matching ensures the comparison of similar treated and control cases. Many authors then claim that any remaining difference is thereby causal in nature, often making an analogy with randomized experiments. A central divide among these articles is whether they acknowledge that matching can only adjust for *observed* covariates, and thus cannot eliminate omitted variable bias. Unfortunately, fewer than half of these papers mention this caveat.

About one in five articles mention the concept of "model dependence," with a reference usually given to Ho et al. (2007). However, it is almost never explained what exactly this means. Only 4 of 61 articles note that matching helps to prevent bias from model misspecification.[9] Finally, one in six articles give no substantive explanation at all for matching, although these are mainly those using it as a robustness check.

The results of this literature survey indicate a significant confusion concerning the benefits of matching. As I will argue, matching has no advantage relative to parametric techniques for dealing with selection or proving causation. Further, I show that matching generally increases

[7] The seven journals are *American Political Science Review*, *American Journal of Political Science*, *Journal of Politics*, *International Organization*, *Comparative Political Studies*, *Political Research Quarterly*, and *International Studies Quarterly*.

[8] A file with a list of these articles and codings of their justifications for matching, the types of matching used, and other details is available upon request.

[9] About 30% mention that matching is a non-parametric technique, but almost none explain why this is advantageous.

model dependence, invalidating its second most common justification. Finally, matching can guard against bias from model misspecification, which is rarely given as an explicit justification. However, I will ultimately recommend that researchers use matching as a last resort for dealing with this problem.

3 First Purpose: Matching for Causal Inference

3.1 Matching, Regression, and Endogeneity

Many researchers treat matching as either an easy way to allay endogeneity concerns or a powerful causal technique akin to leveraging a natural experiment.[10] In reality, matching offers no causal leverage relative to regression alone. This section discusses some of the mistaken assumptions underlying this confusion and then explains what distinguishes matching from true methods of causal inference.

As discussed, most scholarly articles justify matching as a way to deal with the non-random selection of the treatment. This is true in the sense that one can reduce selection bias by matching on an endogenous variable. However, one can get the same benefit by controlling for the variable in regression. Methodological results on matching's unbiased estimation of the treatment all rely on a "selection-on-observables" claim, which happens to be the exact same assumption of "no omitted variables" made for regression. Nothing about matching helps to account for the variables we do not control for. Yet the challenge of endogeneity is precisely the difficulty of controlling for every confounder.

The notion that a matched sample approximates experimental data is another widespread fallacy. Even methodologists are prey to making this analogy.[11] If one inspects a full sample and finds that a treatment is uncorrelated with a range of covariates, this is suggestive (but

[10] According to Kam and Palmer (2008), "the matching process mimics random assignment" (613) and "allows us to analyze observational data through the lens of an experimental design" (620). Kelley (2011: 1540) employs "matching to modify the observational data so that they approximate experimental data, which randomly assign observed units to treatment and control groups." Cao (2009: 1118) argues that matching allows a researcher to "differentiate statistical association from causal effect."

[11] For instance, Stuart and Rubin (2007) refer to matching as a "quasi-experimental design."

not definitive) evidence that the treatment is "as if randomly" assigned.[12] Matching improves balance, so that it *looks* as if the treatment was assigned randomly. However, it only looks that way by construction. It does not follow that the matched sample has any other property of randomly assigned data. In other words, the fact that randomness leads to balance does not imply that making a sample balanced leads to randomness. Indeed, a number of studies have shown that matching is usually unable to replicate experimental benchmarks when applied to observational data (LaLonde 1986; Smith and Todd 2005; Arceneaux et al. 2006).

3.2 Causal Inference and Substantive Insight

True methods of causal inference are based on leveraging a substantive argument about the data, by which I mean a claim that utilizes knowledge about the process that generated it. For instance, instrumental variables require making an untestable claim that the instruments only affect Y through their effect on an endogenous variable X. If the argument for this exclusion restriction is lacking, instruments are useless for proving causation.[13] Regression discontinuity requires knowing that an arbitrary threshold is used to assign otherwise similar units to a treatment. Natural experiments exploit a source of variation in the world that for substantive reasons we claim to be "as if random" (Dunning 2008). In essence, causal methods leverage qualitative insights for quantitative purposes, which is the only way to leap from correlation to causation (Sekhon 2009; Keele and Minozzi forthcoming).

Matching, in contrast, does not utilize any substantive argument. It is purely a data manipulation technique and therefore gives us no leverage in proving causation. One is certainly free to make causal claims about a matching estimate, but this requires making the case that one has adjusted for every relevant confounder. Again, this is the same difficult task required for regression.

Although methodologists generally recognize the limitations of matching, they have almost certainly contributed to the confusion by referring to matching as a method of causal inference. This is especially pervasive in the titles of papers on matching. For instance, Sekhon's

[12] This is in fact a common method of validating natural experiments (Dunning 2008).
[13] In fact, without this argument instruments cannot even establish *correlation*.

(2009) article in the *Annual Review of Political Science* is subtitled "Matching methods for causal inference." Similarly, the titles of Ho et al. (2007), Morgan and Harding (2006), Stuart and Rubin (2007), and King et al. (2011) all use "causal inference" or "casual effects" in reference to matching.

Matching is not a method of causal inference.[14] This misunderstanding is damaging to empirical social science, as it offers researchers an easy route to erroneously deflect concerns about endogeneity. Countless papers proceed by acknowledging a serious selection problem, then claiming to address the issue with a matching design. This false sense of security in turn reduces the incentive to pursue causal techniques like instruments and natural experiments. As we will see, there are other potential advantages for matching, but authors cannot justify using it to deal with selection issues.

4 Second Purpose: Matching to Reduce Model Dependence

Increasing emphasis in both the applied and methodological literatures has been placed on matching's value for reducing model dependence. In large part, this is due to the influence of Ho et al. (2007), which recommends matching prior to parametric estimation chiefly for this benefit. As of this writing, this article has 842 cites on Google Scholar and is the most popular reference in recent political science research using matching.

Given some evident confusion in the literature, it is important to clearly define what Ho et al. (2007) mean by "model dependence." Quoting King and Zeng (2006: 135), they define model dependence "as the difference, or distance, between the predicted outcome values from any two *plausible* alternative models" (their emphasis). As the authors note, researchers face numerous modeling choices concerning which variables to include, how they should be transformed (e.g., polynomial functions, ordinal categories), interaction terms, and the general functional form. Model dependence is a measure of the *sensitivity* of the treatment estimate to

[14] One might claim that matching counts as causal inference because it can reduce bias from misspecification, but this waters down the term far too much. Matching may improve an estimate of correlation, but cannot help with the leap to causation. The term "*causal* inference" should be reserved for methods that reinforce causal claims.

these choices. Reducing this sensitivity is desirable for at least two reasons. First, it reduces the arbitrariness of any particular estimate. Second, it limits the opportunity for data-mining. If similar models yield widely divergent outcomes, researchers can easily pick out the models with the desired conclusions and significance levels.

Thus, Ho et al. (2007) argue that matching reduces the sensitivity of estimates to modeling choices. This section critiques this seminal article's support for this claim. I argue in contrast that matching generally increases model dependence and considerably widens the opportunity for data-mining. This is supported by results from empirical and simulated data.

4.1 Ho et al.'s Argument and Why Matching Actually Increases Model Dependence

Ho et al. (2007) emphasize the use of matching as a precursor to running a parametric model like regression. Thus, the analysis they recommend consists of two stages. First, the data is pruned through matching to improve balance on a set of covariates. Second, a parametric estimator is applied to the matched sample, controlling for the same covariates or a large subset.

The key idea of Ho et al. (2007) is that by "preprocessing" the data, matching leads the second-stage estimator to have low sensitivity to the model specification. To the extent that matching improves balance, it breaks the relationship between the treatment and the distribution of the covariates. This is consequential since the treatment estimate is only influenced by control variables that are correlated with the treatment. Thus, a more highly balanced sample implies that the specific variables included in the parametric estimator have a reduced effect on the treatment estimate.[15] Ho et al. (2007) illustrate this with two empirical examples drawn from the literature. For each, they create a matched sample using a comprehensive set of covariates. They then look at the variation in estimates from applying a succession of parametric models with differing subsets of the covariates as controls. They find that this variation is much lower for the matched sample compared to the full sample.

[15] In the limit, the treated and control units are exactly matched on all covariates, in which case the specification of the parametric estimator becomes irrelevant.

The shortcoming of this argument is that it ignores the dependence of the estimate on the first stage of the analysis, in which the actual matching occurs. When using matching, the result one gets is necessarily a product of the two stages in combination. If some uncertain choice in the matching stage has a major effect on the treatment estimate, this imposes an arbitrariness on the result and opens opportunities for data-mining just as if model dependence was generated by the parametric estimator. Thus, to the extent that choices concerning how to match influence the results, this must be factored in to the measure of model dependence. The two empirical examples in Ho et al. (2007) fail to do this by holding the matched sample fixed and only considering the model dependence of the second-stage estimator. I expand on this for one of these examples below.

Ho et al. (2007) briefly acknowledge and discount the concern that the match itself generates model dependence. In essence, they claim this is not an issue because the matching process does not involve substantive modeling choices. It is worth quoting them in full:

> "[T]he matching literature offers a large number of possible and seemingly ad hoc procedures. From one perspective, we might be concerned about the sensitivity of our results to changes in this process, just as we have been concerned with the sensitivity of our causal effect estimates to parametric modeling assumptions. This is not normally viewed as a major issue since the right procedure is the one that maximizes balance (with n as large as possible), no matter how many procedures we try. By applying this criterion in a disciplined way... no choices are open to the analyst" (Ho et al. 2007: 232).

This is mistaken, as the quantity that the "right" matching procedure aims to maximize is itself determined by the researcher. Most centrally, *balance is defined in reference to a specific set of covariates that the researcher has chosen*.[16] This is in fact the most consequential choice made when matching, but Ho et al. (2007) treat it as a fixed quantity. In reality, the choice of what to match on is just as uncertain as the choice of what to control for in parametric

[16] It is incorrect to respond that a researcher should simply lump in "all" covariates. Some choice is always made about what to include and/or what data to collect. Further, this is counterproductive. Adding in irrelevant variables will either reduce the size of the matched sample or increase imbalance on some relevant variables, defeating the purpose of matching (Ho et al. 2007: 217; King et al. 2011).

estimation. In addition, researchers make subjective choices like the following: (1) whether to match on certain transformations and/or interactions of the covariates, (2) whether to prioritize certain covariates (e.g., by seeking exact matches on the most important variables), (3) the specific metric of balance to maximize, and (4) how to prioritize balance versus sample size.[17] All of these choices define at root what the matching procedure is aiming for, and therefore cannot be ignored when measuring model dependence.

Moreover, even with a set goal of what to maximize, a researcher makes additional choices that can substantially affect the empirical results. Among these are: (1) the matching method, the varieties of which often produce similarly balanced samples, (2) whether to perform an initial pruning step to get common support,[18] (3) which post-match estimator to use, and (4) which covariates to include in the post-match estimator. Note that Ho et al. (2007) only consider the effect of (4). Since the methodological literature has yet to converge on a consistent set of guidelines, substantial leeway is left to the analyst. At minimum, this plethora of modeling options dramatically increases the ability to data-mine. In effect, matching provides researchers with the opportunity to determine both the set of control variables and (to a non-trivial degree) the sample itself. In addition, there is the choice over whether to match in the first place. If regression fails to give an analyst his or her desired outcome, matching represents a whole other domain of models to search through.

It remains to show that the results from matching are indeed sensitive to the choices concerning how to match. A nice illustration of this comes from a recent scholarly debate that uses matching to determine whether college education increases political participation. Since this is a classic case in which numerous complex factors influence both sides of the equation, Kam and Palmer (2008) use PSM to match individuals with and without a college education on 81 covariates. Surprisingly, they find no relationship between college experience and various measures of political participation. However, two response articles illustrate the sensitivity of this conclusion to the specific matching procedure. Henderson and Chatfield (2011) re-run the

[17] See King et al. (2011) for a discussion of this tradeoff.

[18] For instance, King and Zeng (2006) recommend first removing controls with covariates outside of the convex hull of the treated units' covariates.

analysis using 766,642 different subsets of the 81 covariates for matching. They find considerable variation in the resulting estimates for college education, with 76.4% of models finding a significantly positive effect on participation. Mayer (2011) uses the same set of variables as Kam and Palmer (2008), but adopts the "genetic matching" procedure introduced by Diamond and Sekhon (forthcoming). He finds that it achieves superior balance and restores the positive effect of college on participation.

This suggests that Ho et al. (2007) considerably underestimate the model dependence of matching by ignoring the specification choices that enter into the first stage. However, it still remains unclear whether accounting for this implies that matching increases model dependence relative to regression alone. For instance, Henderson and Chatfield (2011) do not compare the model sensitivity of matching to estimation without matching. In the remainder of this section, I confirm that matching generally increases model dependence.

4.2 An Empirical Illustration

I begin by replicating and extending one of Ho et al.'s (2007) empirical illustrations. The authors consider a model in Koch (2002) that tests how the visibility of U.S. House candidates (proxied by campaign expenditures) affects citizen perceptions of their ideology. Ho et al. (2007) limit their replication to Republican male candidates and utilize the six remaining covariates (including party ideology and the respondent's political awareness). Using PSM with all six variables, they match the 350 less visible candidates with an equal number of more visible candidates. With this matched sample, they then repeatedly estimate the effect of visibility using OLS, controlling in turn for all possible non-empty subsets of the six covariates. This produces 63 separate treatment estimates. This procedure is then repeated for the full (non-matched) sample.

Ho et al. (2007) find that the dispersion of estimates is quite large when using the full sample, whereas the matched sample produces a similar treatment estimate regardless of the regression model. However, this fails to capture the true model dependence of matching, as the authors hold constant the set of variables used to match. This artificially reduces the

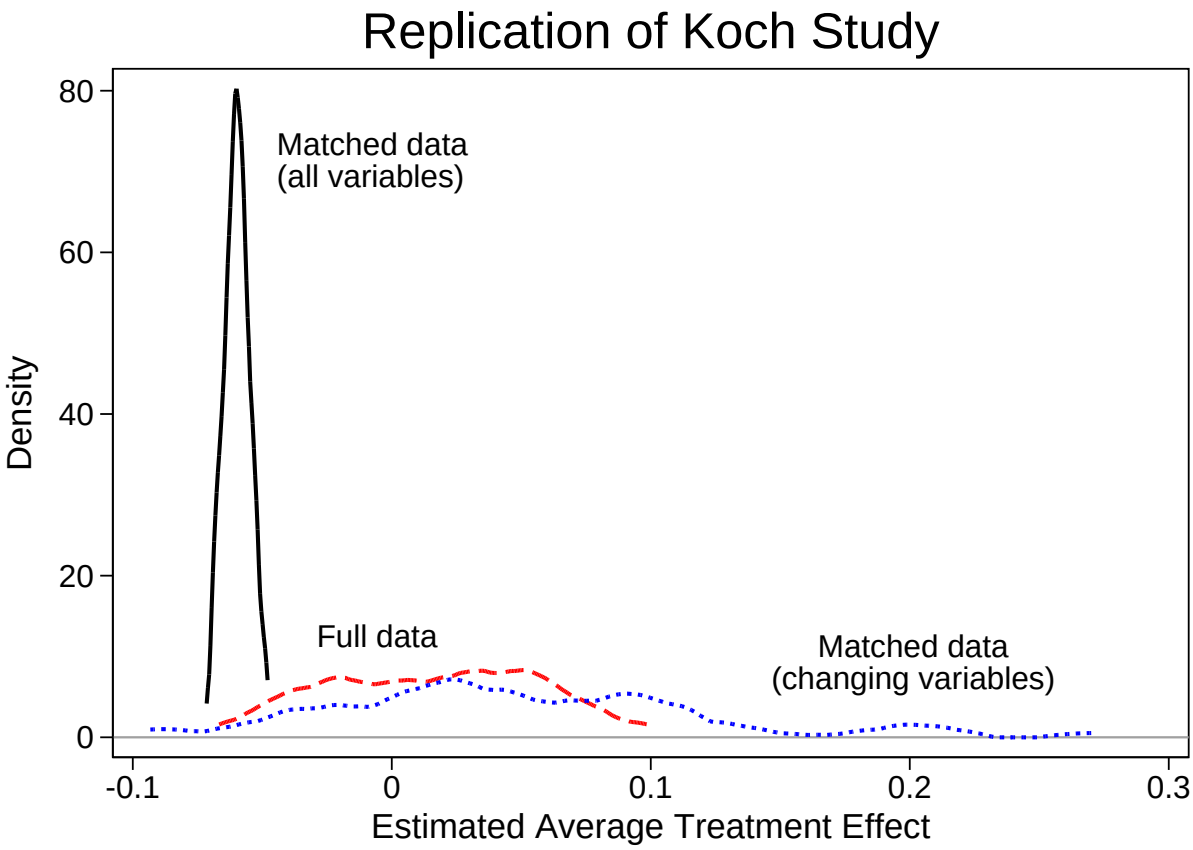

Fig. 2: Kernel density plots of treatment estimates across 63 distinct covariate combinations with the Koch (2002) data. The red dashed line summarizes estimates for the full data, the solid black line for a fixed sample matched on all six covariates, and the dotted blue line for changing samples matched on the differing covariates. As shown in Ho et al. (2007), the estimates are less model-dependent for the fixed matched sample compared to the full sample. However, when the covariates used in the matching stage also change, matching is found to be more model-dependent.

variance of the matching estimates because all six variables are always being accounted for, unlike when regression is used alone. In reality, the researcher chooses the set of matching variables and this fact needs to be incorporated into the calculation of model dependence.

To create a better comparison, I replicated Ho et al.'s (2007) findings for the single matched sample and the full sample. I then extended the analysis by looking at the distribution of estimates when the set of variables used in the matching stage also changes. For each of the 63 combinations of variables, I produced a different matched sample using the equivalent PSM procedure. The treatment is then estimated using OLS, controlling for the same set of covariates.

Figure 2 displays kernel density plots of the 63 treatment estimates for the three different procedures. As seen in Ho et al. (2007), estimates from the full data vary much more widely than those from the fixed matched data (using all variables). However, when the variables used in the matching stage also change, this conclusion is reversed: *matching displays higher sensitivity to the model specification*. In fact, the range of estimates is double that for the full data, and the standard deviation is 75% greater. Further, note that the matching specification chosen by Ho et al. (2007) represents an extreme outlier among the set of possible matches.

4.3 Simulations of Model Dependence

One flaw in using empirical data is that it is unclear what exactly is generating model dependence. The estimate may be changing across models for substantive reasons. Specifically, omitting relevant controls may be generating selection bias, leading the treatment estimate to appropriately change.

Simulated data can address a different question: How sensitive is matching to the addition of a variable that really is irrelevant? This is closer in spirit to the notion of model dependence as sensitivity to arbitrary modeling choices. If a variable truly has no effect on the outcome, an estimation strategy with low model dependence should not vary much based on its inclusion. With this in mind, I use simulations to compare the model dependence of PSM, CEM, and regression without matching.[19] For both matching procedures, I run regression after matching (with the same set of covariates).

A Simple Example

Consider a binary treatment T and two covariates, X_1 and X_2, which are generated from uncorrelated standard normals.[20] T is generated from a binomial distribution with

$$P(T = 1) = \Phi(X_1 + X_2 + \mu), \tag{4}$$

[19] For PSM, I use `psmatch2` in Stata with nearest-neighbor matching. For CEM, I use `cem` in Stata with the default settings.
[20] Results look virtually identical if X_1 and X_2 are correlated.

where μ is a standard-normal error and Φ is the cumulative of the normal distribution (*i.e.*, a probit function). The outcome is generated as

$$Y = 2T + X_1 + X_2 + \varepsilon, \tag{5}$$

where ε is a standard-normal error. Thus, both X_1 and X_2 are correlated with the treatment and the outcome.

In every model, I control for X_1 and X_2, producing unbiased estimates of the effect of T. However, I then calculate the sensitivity of this estimate to the addition of X_1^2 as a covariate (in both the matching and regression stages). Note that after X_1 is controlled for, this variable is unrelated to both the treatment and the outcome, so its inclusion should not change the estimated treatment effect.

For 1,000 draws of simulated data (with pre-match sample sizes of 300), Figure 3 displays how the treatment estimates change for PSM, CEM, and regression alone (OLS). Regression is compared against PSM in the top panel and CEM in the bottom panel. Measured along the horizontal axis is the original treatment estimate with X_1 and X_2 as the only controls. Measured along the vertical axis is the estimate for the same sample, but with X_1^2 added.

If model dependence is low, the two estimates should be very similar and the plotted points should stick closely to the displayed equality line. This is indeed the case with OLS, for which the average absolute difference between models is less than 0.005. The estimates from matching are much more model-dependent, as seen by the larger shifts when the X_1^2 term is added. The average change in the treatment estimate using PSM is 21 times the change using regression alone. For CEM, the average change is 6 times as large.

Model Dependence by Number of Covariates

I now show that matching's level of model dependence is partly a function of how many additional covariates are included. I allow for between 0 and 4 covariates in addition to X_1. Just like X_2 in the previous example, each additional covariate is generated from a standard-normal and is correlated with both the treatment and the outcome. Each is included in both

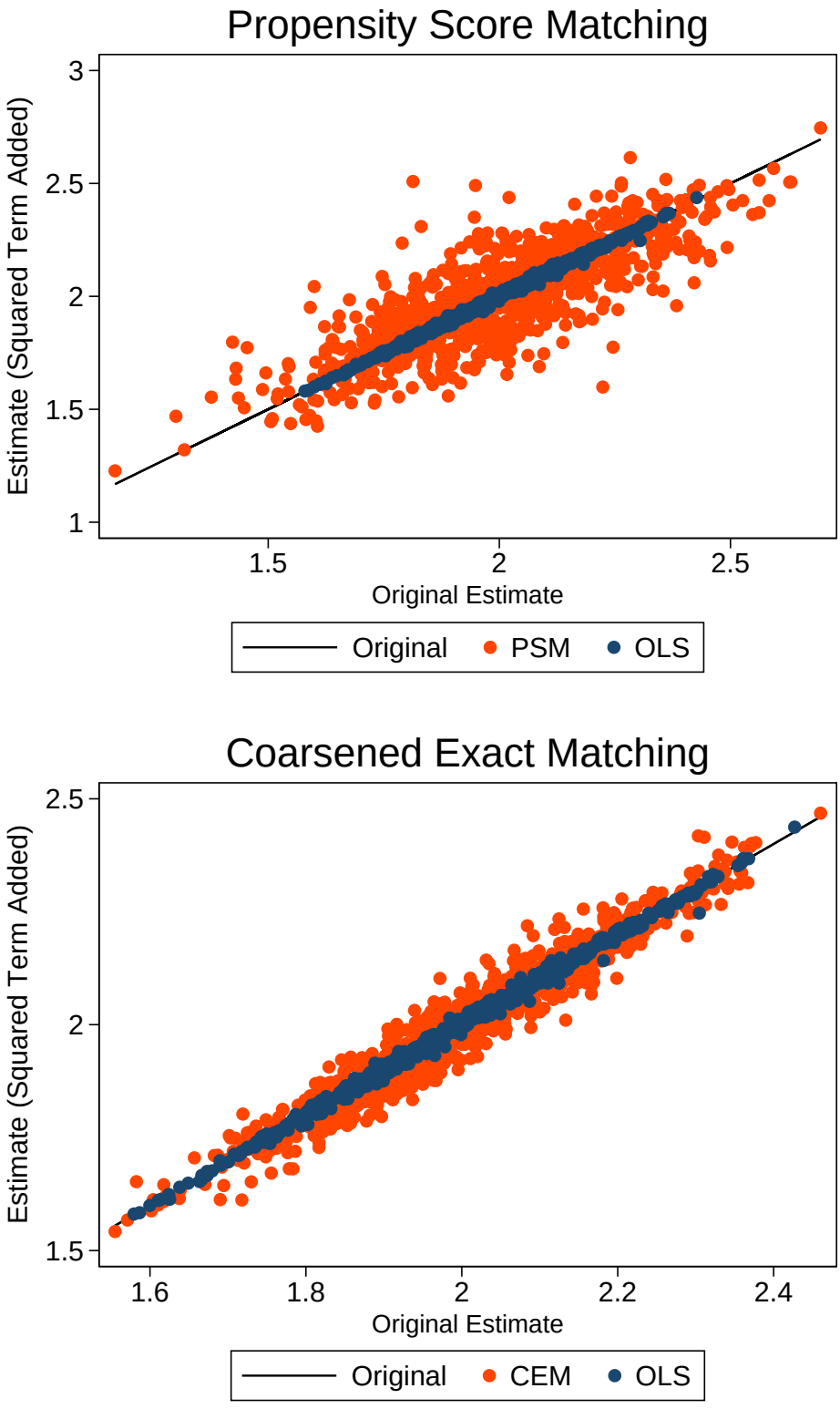

Fig. 3: Plots of model sensitivity, captured by the variation in treatment estimates when an irrelevant term is added (in both the matching and regression stages). Using simulated data, regression alone (OLS) is compared to propensity score matching in the top panel and to coarsened exact matching in the bottom panel. The initial estimate is measured along the horizontal axis. The estimate for the same sample but with an unneeded squared term added is measured along the vertical axis. Regression is shown to be much less sensitive to the model specification than either matching method.

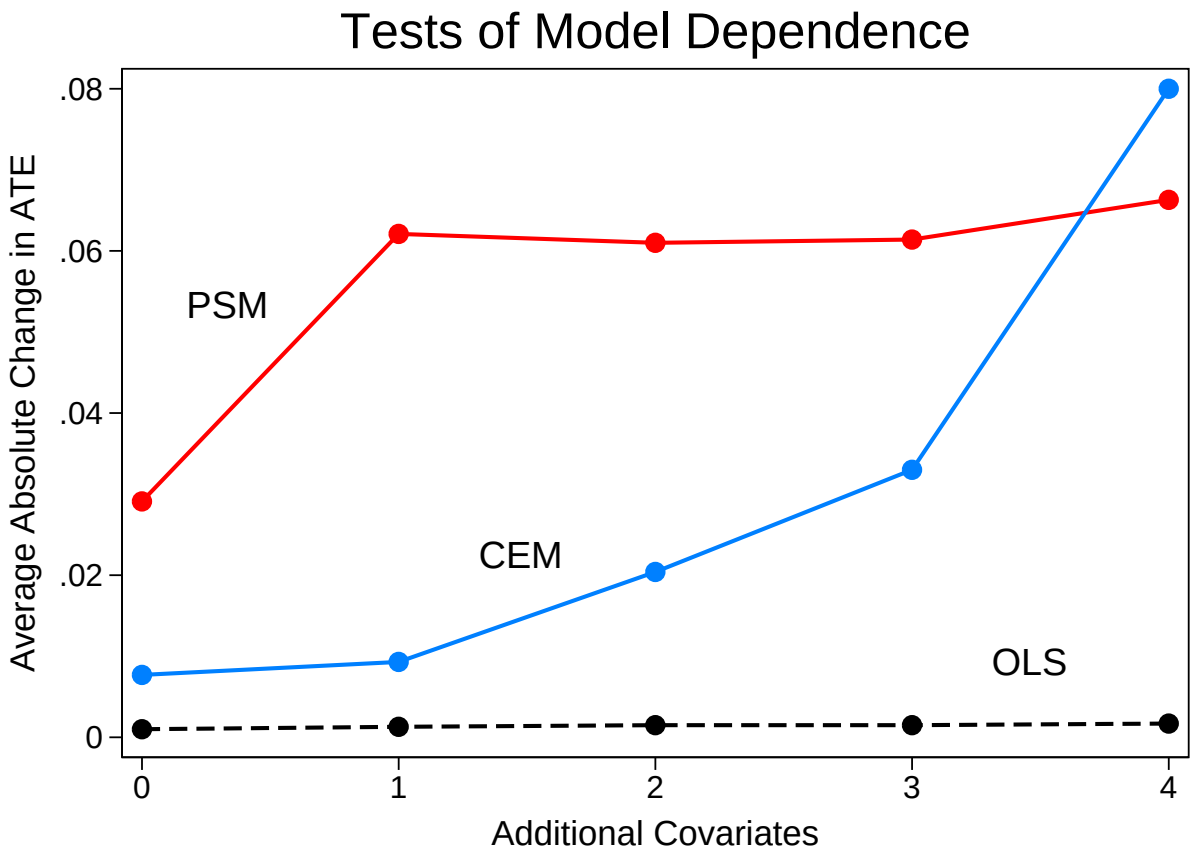

Fig. 4: A plot of model sensitivity, measured by the average variation in treatment estimates when an irrelevant squared term is added (in both the matching and regression stages). Using simulated data, regression alone (OLS) is compared with two matching methods, varying the number of additional covariates in the model (besides X_1). Both matching methods are more sensitive to the model specification, especially with a longer list of covariates.

the matching and regression stages. For 1,000 simulations for each covariate number (with pre-match sample sizes of 1,000), I calculate the estimated treatment effect with and without an irrelevant X_1^2 term. This fairly large sample size represents a conservative test, as the model dependence of matching is exacerbated by smaller sample sizes.

For PSM, CEM, and OLS alone, Figure 4 displays the average absolute change in the estimated ATE, with the number of additional covariates on the horizontal axis. We can plainly see that OLS remains remarkably insensitive to the inclusion of the squared term, regardless of the number of additional covariates. Both types of matching are considerably more model-dependent, varying from about 8 to 50 times OLS's average change. The two matching procedures also display distinct patterns. For 0–3 additional covariates, PSM is the most

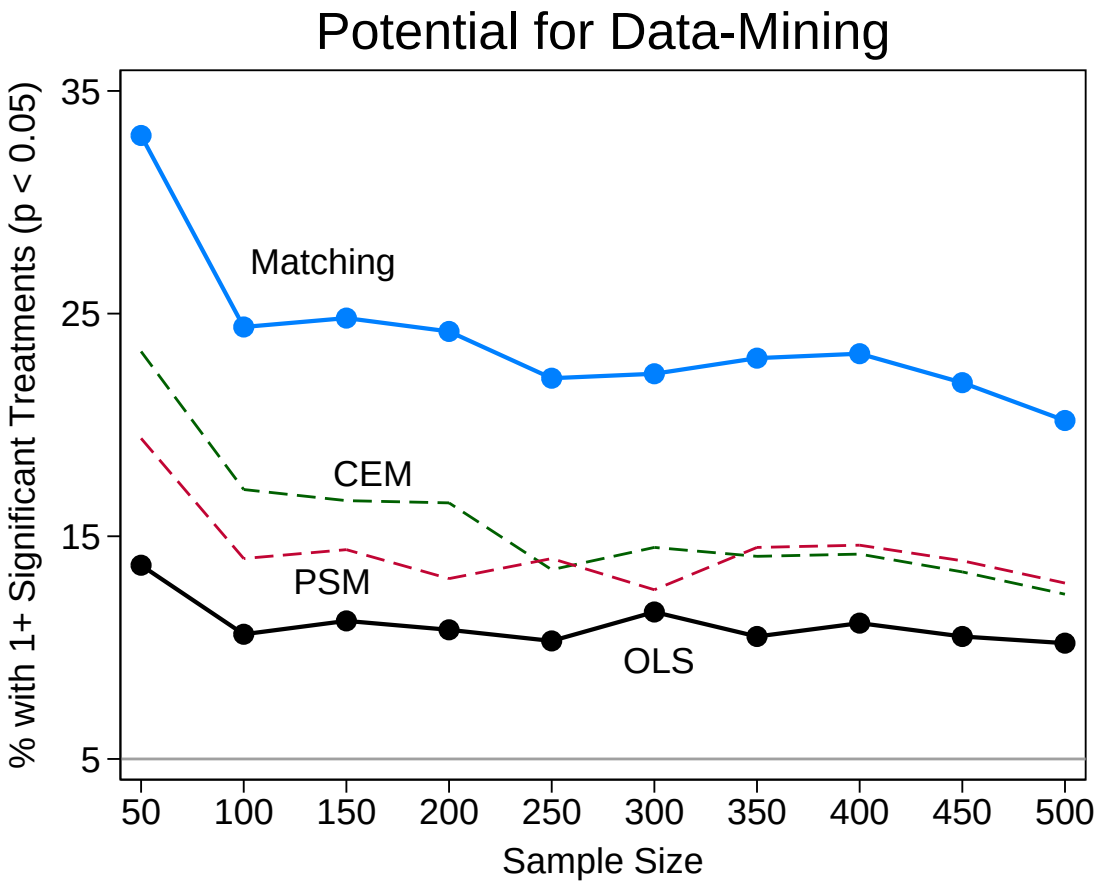

Fig. 5: A plot of the vulnerability of matching and regression to data-mining. This figure uses simulated data with a true treatment effect of 0. For varying sample sizes, it shows the likelihood of finding at least one significant treatment effect among 12 alternative models. This likelihood is consistently higher for both matching techniques, and is more than twice as high if one can choose between matching techniques.

model-dependent, with a jump when moving from 0 to 1, but little change when adding further covariates.[21] In contrast, the model dependence of CEM increases exponentially with the number of additional covariates. In large part, this arises because the sample size steadily declines when CEM is unable to find matches belonging to the same value ranges across all variables.

[21] This arises because PSM simplifies the multi-dimensional covariate data into a scalar propensity score and matches on that. This works much like CEM when a single covariate is involved (the 0 *additional* covariates case) because the propensity functions as a one-to-one transformation of the variable. When the data becomes multi-dimensional, the procedure's simplification works similarly well regardless of the number of covariates.

4.4 Data-Mining

To compare the ability to data-mine using matching and regression, I consider simulations in which the true treatment effect is 0. The potential for data-mining is a function of the likelihood that one can find at least one significant treatment estimate among a set of related models. If this likelihood is high for a particular estimation strategy, researchers can exploit this fact and present only the model(s) that yield significance. Again, I compare PSM, CEM, and regression alone (OLS).

Suppose X_1 and X_2 are generated from uncorrelated standard normals. T is drawn from a binomial with

$$P(T = 1) = \Phi(X_1 + \mu), \tag{6}$$

where μ is a standard-normal error. The outcome is generated as $Y = X_2 + \varepsilon$, where ε is a standard-normal error. Since neither X_1 nor X_2 is correlated with both the treatment and outcome, the inclusion and transformations of these variables in a model ought to be irrelevant to the estimate of T.

I consider 12 separate models for each estimation strategy, allowing for the inclusion of X_1, X_2, their squared terms, and their interaction. To make this set of models more realistic, I assume that if X_j^2 or the interaction term is included, X_j is as well. Since matching requires at least one covariate to match on, I ignore the model with no covariates. Again, for PSM and CEM, the set of variables is used in both the matching and regression stages.

Figure 5 compares the likelihood of finding at least one significant treatment using the three methods. Since one is able to pick between different matching strategies, I also show the likelihood of finding a significant effect using *either* PSM or CEM (Matching). Significance is defined by $p < 0.05$, using the t-value adjusted for degrees of freedom. The sample size is varied along the horizontal axis, with 1,000 simulations for each point shown.

As seen, regression displays the lowest potential for data-mining. For sample sizes of 100 or more, its likelihood of finding a significant treatment across the 12 models is about 10%. Note that this is not much higher than the theoretical 5% probability of finding a significant effect

at random from a single model. For all sample sizes, the ability to data-mine is higher for both PSM and CEM. The latter is particularly prone for small sample sizes, with a likelihood as high as 23%. If one can choose between PSM and CEM, the chance of finding a significant treatment is more than double the chance for regression alone for all sample sizes.

4.5 Discussion

To sum up, past claims about matching and model dependence have ignored the role of specification choices in how the match itself is constructed. Researchers determine how balance is defined and optimized, and can further choose among a range of matching procedures. Moreover, matching estimates are more sensitive to the choice of covariates and the addition of irrelevant variables. This arises because matching on a variable both accounts for its correlation with the treatment and, unlike regression, alters the sample. Further, matching is considerably more prone to data-mining, both because of its higher model dependence and the larger range of specification options.

A related point is that the low model dependence of the parametric estimator after matching ironically *increases* matching's vulnerability to data-mining. This is because the best defense against data-mining is the expectation that authors include several robustness checks when presenting their results, which makes it harder to select out an unrepresentative finding. However, by fixing the matched sample and varying the parametric specification, it becomes much easier to present a range of models with the same result, creating an illusion of robustness. This is illustrated by looking back to Figure 2. Using a single matched sample, Ho et al. (2007) find that the treatment estimate is highly robust to varying the parametric model, which might wrongly increase our confidence that the true effect is about -0.06. Yet one could pick any estimate generated by a different matching specification (traced by the blue dotted line) and produce a similarly tight distribution of parametric models, instead making an estimate of 0.1 or 0.2 look highly robust.

How do we guard against this problem? Most importantly, researchers should drop the presumption that matching inoculates estimates against model dependence. Further, as Hen-

derson and Chatfield (2011) argue, analysts who use matching should show the robustness of the findings to alterations of *both* the matching and parametric specifications. Lastly, more methodological work is needed on the conditions in which model dependence is most severe, how it can be detected in specific cases, and how to adjust estimation uncertainty in the presence of multiple plausible models.

5 Third Purpose: Matching to Reduce Misspecification Bias

5.1 Matching and Misspecification

A neglected virtue of matching (at least in the applied literature) is that it lessens the reliance on functional form assumptions connecting the observed covariates to the outcome. When applied to an exactly balanced sample, a parametric estimator will yield an unbiased treatment effect regardless of the true functional form (assuming the selection-on-observables assumption is satisfied). In other words, matching guards against bias from model misspecification in the parametric estimator.

In contrast, regression assumes a linear function from $\mathbf{X}$ to the outcome Y. This is likely to generate a biased treatment estimate if two conditions are met: (1) imbalance on $\mathbf{X}$, and (2) a nonlinear relationship between $\mathbf{X}$ and Y.

To see why this combination generates treatment bias, suppose that $Y = \alpha T + g(\mathbf{X}) + \varepsilon$, where $g(\mathbf{X})$ is a nonlinear function. If we apply linear regression controlling for $\mathbf{X}$, we will estimate some set of coefficients $\hat{\beta}$ for $\mathbf{X}$, which will cause the quantity $g(\mathbf{X}) - \mathbf{X}\hat{\beta}$ to enter the error term. The treatment estimate $\hat{\alpha}$ will be biased if T is correlated with $g(\mathbf{X}) - \mathbf{X}\hat{\beta}$, since the unexplained residual will be attributed to the treatment. This correlation is only possible if there is imbalance on $\mathbf{X}$, which matching attempts to eliminate.

As an example, consider the top panel of Figure 6. We see a very clear quadratic relationship between X and Y, but suppose we only control for a linear X term. The result will approximate a horizontal line. We also see that the treated units dominate the extreme values of X. Failing to control for X^2 will lead these extreme values to have an unexplained

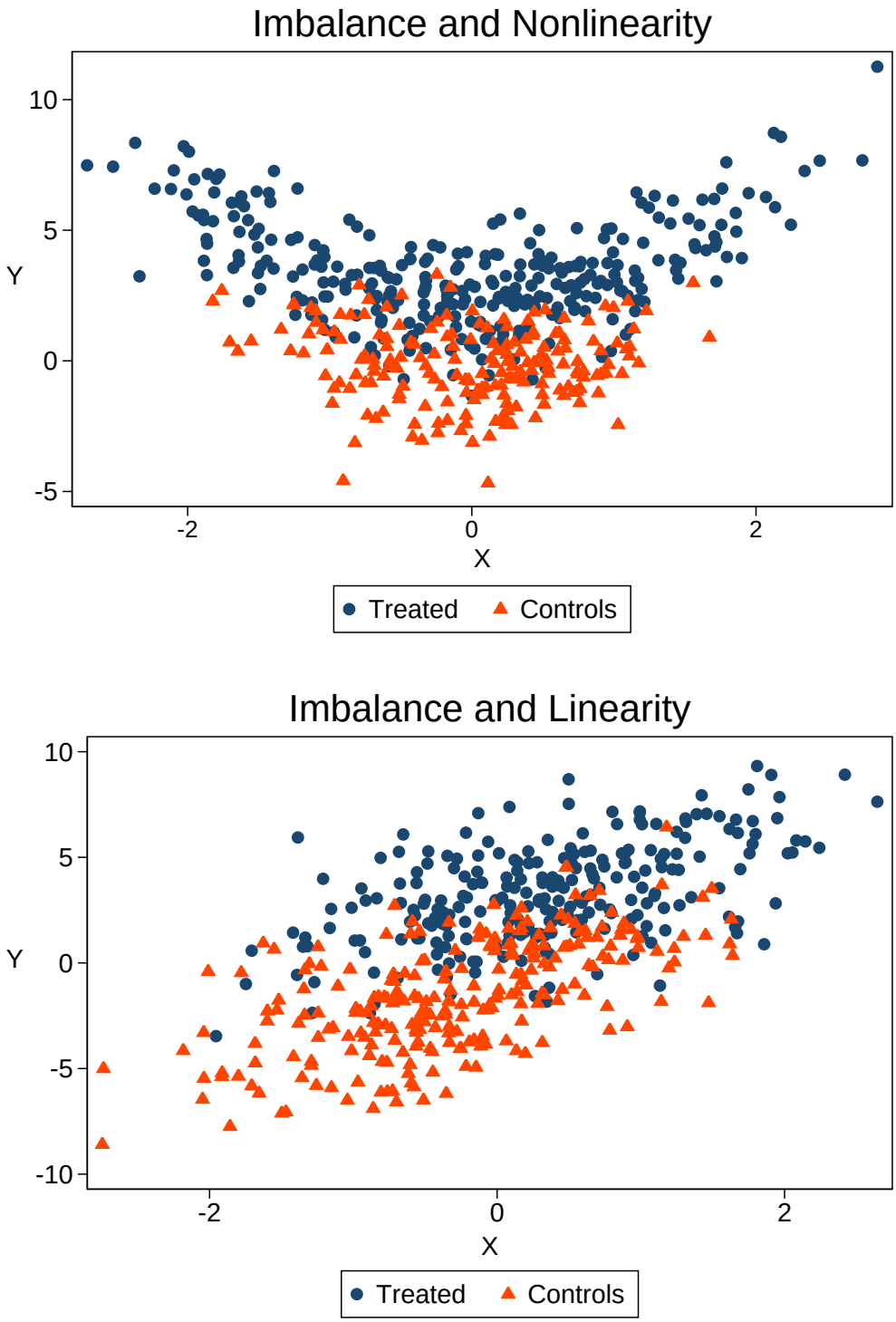

Fig. 6: Simulated data showing a highly nonlinear pattern in the top panel and a linear pattern in the bottom panel. Although both samples are imbalanced, this will only cause treatment bias for the top panel. Matching can guard against this bias, but researchers should perform a test to distinguish between these two types of samples and should first attempt to model the nonlinear nature of the data.

positive residual, which will then mainly be associated with treated units. This will produce an overestimate of the treatment effect. A properly matched sample, however, will remove the treated units with extreme values of X since they have no comparable controls. The uncorrected nonlinearity will still be present, but it will no longer confound the treatment estimate since the nonlinear component will no longer be correlated with the treatment. In other words, matching prevents bias by invalidating condition (1).

Matching provides an attractive method of guarding against misspecification bias, but I suggest two caveats. First, I recommend a regression-based test for whether matching is justified by the data. Researchers often resort to matching whenever imbalance is present, but this is not sufficient to generate bias. For instance, the sample in the bottom panel of Figure 6 is imbalanced on X, but matching is unnecessary since the true model is linear. I suggest how to directly verify that nonlinearities are present in the data. Second, matching is best regarded as a last resort for addressing model misspecification. It is preferable to try to correctly model the data using parametric or semi-parametric methods rather than to accept misspecification and treat the covariates as nuisances.

5.2 A Regression-Based Test

Treatment bias is a significant concern when imbalance is combined with data nonlinearities. This situation may then warrant the use of matching. However, it is curious that methodologists have not posited a method of detecting when this situation is present. Rather, many analysts justify matching based purely on the presence of covariate imbalance. What is really needed to justify matching is a way to distinguish between the types of data shown in the top and bottom panels of Figure 6. How do we tell the difference?

Of course, the best way to detect misspecification is to closely analyze the data. For instance, one would hope that inspecting the sample at the top of Figure 6 would clue the researcher into the presence of nonlinearity. Suppose, however, this has been attempted and we desire a simple and direct test.

I recommend the following procedure. Run a chosen matching method that divides the sample into matched and unmatched units, but do not discard any data. Retain the full sample and create a dummy variable called `unmatched`, as well as the dummy interaction term `unmatched*treatment`. Finally, run a regression controlling for the treatment, the full set of covariates, and the two dummies.[22]

The basic idea is that the coefficient magnitudes for the two dummies test for the combination of imbalance and data nonlinearity. The coefficient on `unmatched` indicates the difference between matched and unmatched controls net of a linear function of the covariates. Similarly, the coefficient on `unmatched*treatment` indicates the difference between the average treatment effects in the matched and unmatched samples net of a linear function of the covariates. Thus, the coefficients test for the presence of a nonlinear component $g(\mathbf{X}) - \mathbf{X}\hat{\beta}$ that is associated with either the unmatched controls or treated units. Since the unmatched units are precisely those with covariate values that are predictive of treatment status, larger coefficients imply a stronger association between data nonlinearities and treatment. This is exactly what generates misspecification bias.

To illustrate this logic further, consider applying this procedure to the top panel of Figure 6. The unmatched data will include the treated units at extreme values of X and possibly some controls at middle values of X.[23] Using the full sample for the regression, it is obvious that the coefficient on `unmatched*treatment` will be very large, indicating the underlying nonlinearity. For the bottom panel, the unmatched data will include treated units at large X and controls at small X. Since a linear function of X fits these units just as well as the matched data, there will be no residual relationship between Y and the unmatched data. Therefore, the coefficients on the dummies will be approximately 0, indicating an absence of nonlinearity and no need for matching.

In many cases, the dummies will also correct for misspecification bias in parallel with matching since they control for the nonlinear component that otherwise confounds the treat-

[22] To be clear, this is not a general test for misspecification, since it depends on a particular match. Rather, it is a test of the self-consistency of using the matched sample.

[23] If no controls are left unmatched, the procedure would have to omit either `unmatched` or `unmatched*treatment` from the regression.

ment effect. In fact, the only reason estimates differ between matching and regression with the dummies is that the latter incorporates the unmatched data in estimating the covariates' coefficients. Thus, if we add interactions between `unmatched` and each covariate, the regression will recover the exact treatment effect from matching. Although cumbersome, this procedure provides more information than matching and keeps the sample fixed, which is useful for comparing models.

In sum, before using a matching procedure, an analyst should ascertain whether the data actually warrants it by running this regression procedure and inspecting the two dummies' coefficients. A reasonable indicator is the significance of the F-test for the two coefficients. If this test cannot reject the null, matching is likely unnecessary and detrimental.[24] If the null is rejected, model misspecification is implicated. Even in this case, I argue that matching should only be used as a last resort when attempts to correctly model the data fail.

5.3 Matching as a Last Resort

Matching does not fix misspecification in any sense nor necessarily limit the sample to ranges of $\mathbf{X}$ in which the outcome is approximately linear. Rather, *matching accepts misspecification for* $\mathbf{X}$, but adjusts the sample so that this problem does not interfere with the treatment estimate. If one is solely concerned with estimating the treatment effect, this is not a major concern. However, for many purposes, this may be too much to sacrifice. If nonlinearity is present, the researcher should attempt to model it, with matching as a last resort if this proves impossible.

A critical point is that there does exist a tradeoff between resorting to matching and improving the model specification. Detecting nonlinearities necessitates using the full sample, since matching tends to remove the units with extreme values on $\mathbf{X}$ that are the most informative for this task.

[24] If the outcome model actually is linear, matching will only serve to reduce efficiency. Using simulated data, Glynn and Quinn (2010) find greater bias and mean squared error when using PSM relative to OLS. This is true for most of their specifications, even some nonlinear ones, but the advantage of OLS is greatest when the outcome function is linear.

Improving the specification is valuable for several reasons. First, we often care about inference on more than just the treatment. If the model is misspecified, matching guards against treatment bias, but virtually guarantees that estimates for the covariates will be biased. Further, nonlinear patterns in the data may be theoretically informative. Second, misspecification makes the model-based calculation of certain quantities more error-prone or impossible (King and Roberts 2012). For instance, if we are interested in predicting Y for particular values of T and $\mathbf{X}$, misspecification can add considerable error to this calculation, even for $\mathbf{X}$ in an area of common support. Third, even if we resort to matching, an improved specification can further reduce treatment bias if there is imbalance remaining in the matched sample. This concern is precisely why Ho et al. (2007) recommend running a parametric estimator after matching. This estimator will serve its function more effectively the better the specification is.

My perspective parallels that of King and Roberts (2012), who caution against the widespread use of robust standard errors in parametric estimation. Their reasoning is that a divergence of robust and classical standard errors is a strong indication that the model is misspecified, which hampers the confidence we should have in the point estimates. They therefore recommend improving the model specification until the two standard errors no longer diverge.

Similarly, the ideal is for researchers to specify their models well enough that matching is no longer needed for an unbiased treatment estimate. They can then avoid the considerable negatives of matching, such as the loss of information, increased model dependence, and uncertain standard errors (see Morgan and Harding 2006). The test discussed in the last subsection provides a reasonable guideline for when misspecification is sufficiently reduced. Similarly, one can directly confirm that the regression and matching estimates no longer differ. The key point is that it's much better to arrive at this estimate by faithfully modeling the data rather than matching and accepting model misspecification.

6 Conclusion

Despite being a well-known and increasingly popular empirical technique, matching's advantages and shortcomings are still widely misunderstood. This article critically evaluated three of the most prominent justifications for matching, leading to the following conclusions.

First, despite the many articles claiming otherwise, matching provides no causal leverage or advantage for dealing with selection relative to regression alone. Causal inference necessarily relies on substantive knowledge regarding the data. Applied scholars should stop justifying matching as a cure for endogeneity and methodologists should avoid using the term "causal inference" in reference to it.

Second, matching generally increases model dependence. Ho et al.'s (2007) well-known case to the contrary is flawed because it ignores the model sensitivity generated by choices over the matching procedure itself. In particular, matching is much more prone to data-mining (relative to regression alone) as it opens up numerous consequential yet uncertain choices to the researcher. To counter this, analysts using matching should pay greater attention to the robustness of their estimates to variations in the matching procedure.

Third, the real advantage of matching is that it guards against misspecification bias in the treatment estimate. However, to justify matching in this way, researchers should directly test for the presence of misspecification, as imbalance is not sufficient to cause bias. I suggested a regression-based procedure that tests whether unexplained variation is present in the unmatched data. Lastly, matching is problematic in that it accepts misspecification in the service of recovering an unbiased treatment effect. To avoid this compromise, researchers should first attempt to model the correct functional form, with matching used only as a last resort.

References

Arceneaux, Kevin, Alan S. Gerber, and Donald P. Green. 2006. Comparing experimental and matching methods using a large-scale voter mobilization experiment. *Political Analysis* 14: 37-62.

Cao, Xun. 2009. Networks of intergovernmental organizations and convergence in domestic economic policies. *International Studies Quarterly* 53(4): 1095-1130.

Diamond, Alexis, and Jasjeet S. Sekhon. Forthcoming. Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. *Review of Economics and Statistics*.

Dunning, Thad. 2008. Improving causal inference: Strengths and limitations of natural experiments. *Political Research Quarterly* 61(2): 282-93.

Glynn, Adam N., and Kevin M. Quinn. 2010. An introduction to the augmented inverse propensity weighted estimator. *Political Analysis* 18: 36-56.

Henderson, John, and Sara Chatfield. 2011. Who matches? Propensity scores and bias in the causal effects of education on participation. *Journal of Politics* 73(3): 646-58.

Ho, Daniel E., Kosuke Imai, Gary King, and Elizabeth A. Stuart. 2007. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis* 15: 199-236.

Holland, Paul W. 1986. Statistics and causal inference. *Journal of the American Statistical Association* 81: 945-60.

Iacus, Stefano M., Gary King, and Giuseppe Porro. 2012. Causal inference without balance checking: Coarsened exact matching. *Political Analysis* 20(1): 1-24.

Imai, Kosuke, and David A. van Dyk. 2004. Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistical Association* 99: 854-66.

Kam, Cindy D., and Carl L. Palmer. 2008. Reconsidering the effects of education on political participation. *Journal of Politics* 70(3): 612-31.

Keele, Luke, and William Minozzi. Forthcoming. How much is Minnesota like Wisconsin? Assumptions and counterfactuals in causal inference with observational data. *Political Analysis*.

Kelley, Judith. 2011. Do international election monitors increase or decrease opposition boycotts? *Comparative Political Studies* 44(11): 1527-56.

King, Gary, Richard Nielsen, Carter Coberley, James E. Pope, and Aaron Wells. 2011. Comparative effectiveness of matching methods for causal inference. Working paper.

King, Gary, and Margaret Roberts. 2012. How robust standard errors expose methodological problems they do not fix. Working paper.

King, Gary, and Langche Zeng. 2006. The dangers of extreme counterfactuals. *Political Analysis* 14: 131-59.

Koch, Jeffrey M. 2002. Gender stereotypes and citizens' impressions of House candidates' ideological orientation. *American Journal of Political Science* 46: 453-62.

LaLonde, Robert J. 1986. Evaluating the econometric evaluations of training programs with

experimental data. *American Economic Review* 76(4): 604-20.

Mayer, Alexander K. 2011. Does education increase political participation? *Journal of Politics* 73(3): 633-45.

Morgan, Stephen L., and David J. Harding. 2006. Matching estimators of causal effects: Prospects and pitfalls in theory and practice. *Sociological Methods & Research* 35(1): 3-60.

Rosenbaum, Paul R., and Donald B. Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70: 41-55.

Sekhon, Jasjeet S. 2009. Opiates for the matches: Matching methods for causal inference. *Annual Review of Political Science* 12: 487-508.

Smith, Jeffrey A., and Petra E. Todd. 2005. Does matching overcome LaLonde's critique of nonexperimental estimators? *Journal of Econometrics* 125: 305-53.

Stuart, Elizabeth A., and Donald B. Rubin. 2007. Best practices in quasi-experimental designs: Matching methods for causal inference. In *Best Practices in Quantitative Social Science*, ed. Jason W. Osborne. Thousand Oaks, CA, Sage Publications, pp.155-76.