Beyond Balance:
A Placebo Test for Matching Estimators *

Alexis Diamond[†]
Program on Political Economy & Gov't
Harvard University

Jens Hainmueller[‡]
Department of Government
Harvard University

April 16, 2006

[*] Paper prepared for Presentation at the 65th National Conference of the Midwest Political Science Association 2006. This paper is a work-in-progress. Please do not cite without permission from the authors. Comments are highly appreciated.

[†] adiamond@fas.harvard.edu

[‡] jhainm@fas.harvard.edu

Abstract

Matching has become a popular method of causal inference but there is no consensus as to how covariate balance obtained via matching ought to be evaluated. We present a new diagnostic called the placebo test that involves comparing two "placebo" sets of control units matched to the same set of treated units. The placebo test is designed to aid estimation of the average treatment effect for the treated units, providing information about the extent to which matching (1) reduces bias due to matching discrepancies on observed characteristics and (2) reduces the variance of treatment effect estimates associated with this bias. The placebo test also checks robustness of the matching procedure across multiple models and samples. Importantly, the placebo test does not require outcomes data for the treated group and thus is blind to the answer (the estimated treatment effect), consistent with Rubin (2001)'s call for impartiality in causal inference research design. We probe the plausibility of the placebo test using the National Supported Work Demonstration (NSW) job training dataset of Dehejia and Wahba (1997). We also use our test to validate the Diamond and Sekhon (2005) balance criterion (based on paired t-test and Kolmogorov-Smirnov p-values) and show that a criterion based on the more conventional unpaired t-test does not appear to have desirable properties.

1 Introduction

Matching has become an increasingly popular method of causal inference in many fields, but there is no consensus as to precisely how covariate balance obtained via matching ought to be evaluated. We present a new diagnostic called the placebo test that involves comparing two "placebo" sets of control units matched to the same set of treated units. The placebo test is designed to aid estimation of the average treatment effect for the treated units, providing information about the extent to which matching (1) reduces bias due to matching discrepancies on observed characteristics and (2) reduces the variance of treatment effect estimates associated with this bias. This diagnostic test also checks the robustness of the matching procedure across multiple models and samples.

The idea behind the placebo test is highly intuitive. Diamond and Sekhon (2005) show that in a classic observational setting with real-world data, ATT estimates converge as balance improves, and ultimately, at the very highest balance levels, collapse to a small neighborhood of the experimental benchmark treatment effect. This result is consistent with a wealth of theory and empirical evidence showing that improving the balance of observed confounders across treatment and control groups can reduce the bias and variance of causal estimates.

One could, at least in theory, exploit this feature of matching-based analysis to claim that adequate balance has been achieved when the best-balancing estimates converge to a small neighborhood of results, but this stopping rule involves looking at the answers which taints the impartiality of the research design. Moreover, there is no empirical evidence showing that the same convergence behavior would be observed were the study rerun with different control units and different matching-models.[1]

Our proposed placebo test exploits the idea that improvements in balance should be associated with a convergence around the true result, but allows one to remain blind to the answer, consistent with Rubin (2001)'s call for impartiality in causal inference research design. The basic idea is that the convergence of ATT estimates described above is induced by the reduction in matching discrepancies that comes with improvements in balance, and so it should be possible to observe this convergence by examining the matched control units (because the set of matched treated units remains the same).

We implement our diagnostic by randomly splitting the control group into two "placebo" subgroups and then using each to run a matching procedure that randomly searches for the subset of each placebo subgroup that best matches the treated units. As the search proceeds and balance improves, we find that the two matched control subsets become increasingly similar. At the highest balance levels—as measured per Diamond and Sekhon (2005)—the differences between average outcomes across matched control subgroups converge to a small neighborhood around zero, identifying the point at which

[1] Throughout this paper, when we discuss different control units and multiple control groups we assume that the definition of the control intervention is defined the same way for all units.

the answers are converging *without actually revealing the answers themselves.*

Others have proposed similar diagnostics involving multiple control groups from entirely different data sources to test for hidden bias (Campbell 1969; Rosenbaum 1984, 1987, 2001, 2002: sec.8, Lu and Rosenbaum 2004, Shadish et al. 2002)[2] but to our knowledge, our placebo diagnostic is the first to show how—in a given dataset—improving balance also reduces the bias and variance of causal estimates. Thereby our placebo test can help to establish a balance threshold for a particular data-set. Because our diagnostic requires running a random-search matching algorithm in both placebo subgroups, it also checks robustness of estimates across datasets and differently-matched control groups.

We probe the plausibility of the placebo test using the National Supported Work Demonstration (NSW) job training dataset of Dehejia and Wahba (1997) and show that, as balance improves, the mean squared error of our estimates declines along with the differences between average outcomes across matched control subgroups. We also use our diagnostic procedure to validate the Diamond and Sekhon (2005) balance criterion (based on paired t- and Kolmogorov-Smirnov p-values) and show that a different criterion based on the unpaired t-test does not appear to have desirable properties.

Section 1 provides a brief background on matching and explains the random-search matching algorithm used for the placebo test. Section 2

[2] These tests can be interpreted as a methodological substitute to the formal sensitivity tests for hidden bias reviewed in Rosenbaum (2002:sec. 4).

describes the diagnostic test itself. Section 3 explains how we probed the plausibility of the test by analyzing the NSW dataset. Section 4 concludes with a discussion of the limitations of our diagnostic test and suggestions for future research.

2 Matching

The Rubin Causal model, the predominant framework for causal inference throughout the sciences, defines a causal effect in terms of the difference in potential outcomes under treatment and control for the same set of units (Rubin 1974, 1978; Holland 1986). The fundamental problem of causal inference is that for each unit, only one potential outcome is observed. In this paper, we focus on estimating ATT, the average treatment effect for the treated units. For these units, only the potential outcomes under treatment are observed. Treated units' potential outcomes under control are missing. Matching involves identifying the control units that are most like the treated, and then using the observed outcomes of these matched control units to impute the missing data.

2.1 Matching in the Rubin Causal Model

To formally characterize the Rubin model, we follow Imbens (2003) which describes the simple case of two interventions (treatment and control), one of which is assigned at a single point in time to N individuals randomly drawn from a large population. Let $Y_i(1)$ denote the potential outcome for individual i following treatment, and $Y_i(0)$ denote the potential outcome for that individual in the absence of treatment. Let W be a treatment indicator: 1 when i is in the treatment regime and 0 otherwise. We also assume SUTVA—the stable unit treatment value assumption (Rubin 1980; Zhang and Rubin 2003), which requires the independence of potential outcomes and treatment assignments for all units.

The observed outcome for observation i is $Y_i = W_i \cdot Y_i(1) + (1 - W_i) \cdot Y_i(0)$, and the effect of treatment for unit i may be defined as $\tau_i = Y_i(1) - Y_i(0)$. ATT is defined as $\frac{1}{N_T} \sum_{i:W=1} [Y_i(1) - Y_i(0)]$ and is often considered to be an important estimand because analysts and policymakers tend to care about the average effect of the treatment on those receiving the treatment. When estimating ATT via matching, assignment to treatment is typically assumed to be:

Assumption 1 (Unconfounded) $\mathbf{Pr}(W|X,Y) = \mathbf{Pr}(W|X)$ *for all possible* W*,* Y*, and pretreatment confounders* X*.*

Assumption 2 (Probability) $0 < Pr(W_i = 1|X) < 1$ *for all* i*,* t*, and all possible* X*.*

Together, unconfoundedness and probability constitute strongly ignorable assignment, and (with SUTVA) allow for the estimation of any causal effect for which the data is available (Rubin 1974, 1976a,1976b, 1978; Rosenbaum and Rubin 1983, 1985). The challenge of raw observational data is that the true model of assignment is unknown and strong ignorability rarely, if ever, obtains for the entire sample. Note that strong ignorability of assignment requires conditioning on pretreatment covariates X, which implies that the assumption only holds when covariate distributions are balanced across treated and control groups. Thus, all matching-based methods share the same basic goal—to identify the subgroup of control units most similar to the treated—and the same burden of evidence: to demonstrate that adequate balance has been achieved across all observed confounders.

2.2 Matching Discrepancies and Covariate Balance

Perfect covariate balance is only achieved in the case of exact matching when the matched units have the exact same X. But unless exact matches are available for all (treated) units matching will be inexact and discrepancies in X across matched treated and control groups will generally induce biased causal estimates (Abadie and Imbens 2006). To show this we follow the exposition of Rubin and Imbens (2006: chap. 12) and define N^* as the number of matched pairs, equal to N_t if we match only the treated units. We also define a matched pair as (l_{ti}, l_{ci}) where i indexes the match. Let $l_i \in 1, \ldots, N$ denote the index of the unit that was originally matched, so that $l_i = l_{ti}$ when a treated unit is matched to a control or $l_i = l_{ci}$ when a control unit is matched to a treated unit. Let $W^*_{l_i}$ indicate the treatment status of the unit originally matched to produce the pair, and let $X^*_{l_i}$, $Y^*_{l_i}(0)$ and $Y^*_{l_i}(1)$ be the covariate vector and potential outcomes for this same unit.

In this notation, the unit-level treatment effect is then equal to $\tau_i(X^*_{l_i}) = Y^*_{l_i}(1) - Y^*_{l_i}(0)$, estimated using the observed outcomes for the two units of the matched pair:

$$\tau_i(X^*_{l_i}) = W^*_{l_i(1)} \cdot (Y^*_{l_i}(1) - Y_{l_{ci}}(0)) + (1 - W^*_{l_i(1)}) \cdot (Y_{l_{ti}}(1) - Y^*_{l_i}(0))$$

Here we only match the treated units where $W^*_{l_i} = 1$ so $\tau(X^*_{l_i}) = Y^*_{l_i}(1) - Y_{l_{ci}}(0)$. In the ideal case of exact matching both units of each pair would have covariates equal to $X^*_{l_i}$. With inexact matching, however, only covariates for the treated unit $X_{l_{ti}}$ equal $X^*_{l_i}$ and the difference in covariate values between the two units in the pair is called the matching discrepancy, $D^*_{l_i} = X_{l_{ti}} - X_{l_{ci}}$.

In the case of exact matches when $D^*_{l_i} = 0$ the expected difference in observed outcomes within each pair is equal to the treatment effect conditional on $X^*_{l_i}$. So for treated units the expected value of the *true* unit level treatment effect is:

$$\tau(X^*_{l_i}) = E[Y_i(1)|X_i = X_{l_{ti}}] - E[Y_i(0)|X_i = X_{l_{ti}}]$$

But with non-zero matching discrepancies, the expected value of our estimator for the unit level treatment effect equals:

$$\begin{aligned} E[\hat{\tau}(X^*_i)|X_{l_{ti}}, X_{l_{ci}}] &= E[Y_{l_{ti}} - Y_{l_{ci}}|X_{l_{ti}}, X_{l_{ci}}] \\ &= E[Y_i(1)|X_i = X_{l_{ti}}] - E[Y_i(0)|X_i = X_{l_{ci}}] \end{aligned}$$

and so for treated units with $W^*_{l_i} = 1$:

$$E[\hat{\tau}(X^*_i)] \;\; = \;\; \tau(X^*_{l_i}) + E[Y_i(0)|X_i = X_{l_{ti}}] - E[Y_i(0)|X_i = X_{l_{ci}}]$$

The difference of the last two terms is the expression of the bias of the matching estimator that arises due to the discrepancy between the X of the treated and the X of its matched control unit. With better balance comes smaller matching discrepancies and smaller bias—and as the bias shrinks, the variance associated with the bias also shrinks.

2.3 Common Standards of Covariate Balance

Despite the popularity of matching-based methods, the literature lacks a commonly agreed-upon metric for covariate balance. Consequently, there is also no agreement about the degree of balance required for reliable inference.[3] In practice, most researchers conduct univariate unpaired t-tests to compare covariate means across treatment and control groups and are satisfied if less than 5% of their p-values are statistically significant at the 0.05 level. Another popular way to test balance is to examine standardized differences between groups (Rubin and Rosenbaum 1985), which is defined as the difference in means across the matched samples, scaled by the square root of the matched samples' average variance. Diamond and Sekhon (2005) recom-

[3] As Smith and Todd (2005b:371) put it: "The most obvious limitation at present is that multiple versions of the balancing test exist in the literature, with little known about the statistical properties of each one or of how they compare to one another given particular types of data."

mend both the paired t-test combined with the Kolmogorov-Smirnov (KS) test be performed across all covariates, two-way interaction, and quadratic terms, and suggest that very high p-values may be required for reliable causal inference in nonexperimental settings. Finally, in a recent paper Ho et. al. (2006) argue that hypothesis tests are inappropriate for assessing balance. Instead, Ho et. al. (2006) suggest examining the full empirical densities for each matching variable and propensity scores using quantile-quantile plots.

Apart from univariate tests the literature also suggests various multivariate balance tests, including the Hotelling T^2 test of the joint null of equal means of all covariates, the multivariate (bootstrapped) Kolmogorov-Smirnov (KS) test, the Chi-Square null deviance tests based on the estimated assignment probabilities, and regression-based tests for joint insignificance. Little is known about which of these tests is preferable under what conditions, how the multivariate tests relate to their univariate counterparts, and the degree of balance required for reliable inference.

2.4 Iterative Random Search Matching Algorithm

The matching method employed in this paper is a special case of genetic matching (Diamond and Sekhon 2005), an affinely invariant matching algorithm designed to maximize the lowest univariate KS and paired t-test p-values. Genetic matching uses the following generalized distance measure when matching each treated unit to the nearest control unit,

$$d(X_i, X_j) \quad = \left\{ (X_i - X_j)' \left(S^{-1/2}\right)' C S^{-1/2} (X_i - X_j) \right\}^{\frac{1}{2}}$$

where $S^{1/2}$ is the Cholesky decomposition of S (the variance-covariance matrix of X) and C is a $k \times k$ positive-definite diagonal weight matrix, with k main diagonal elements of C that must be chosen. Following Diamond and Sekhon (2005,) we match on the linear predictor of the estimated propensity score, $Pr(W_i = 1)$, as well as the covariates X once they have been adjusted so as to be uncorrelated with the linear predictor.[4] Diamond and Sekhon (2005) utilize an evolutionary algorithm called GENOUD (Mebane and Sekhon 1998; Sekhon and Mebane 1998) that selects the $k - 1$ free elements of C to minimize a measure of the maximum observed discrepancy between matched treated and control covariates at every iteration of optimization. Loss is defined as the minimum p-value observed across a series of balance tests performed on distributions of matched baseline covariates.[5]

Instead of this particular genetic matching algorithm, we employed a simple random search over the space of C's diagonal elements.[6] Each random sample (drawn from a standard uniform distribution) populates the diagonal

[4] Adjustment is accomplished by regressing each covariate on the estimated linear predictor, $X_k = \hat{\alpha} + \hat{\mu} + \hat{\epsilon}_k$, where k indexes the covariate number.

[5] In fact, loss is actually defined lexicographically; so two different sets of matching results produced the same lowest p-values, the algorithm would then compare the next lowest p-values.

[6] This random search is a special case of GENOUD whereby optimization occurs within a single generation and the only permutation operator enabled is the one that allows random search. See Sekhon and Mebane (1998) for more details.

of the W matrix, which identifies the distance measure and allows matching to proceed. This process is repeated thousands of times for different randomly generated weights, and each time we record the identities of the matched control units and the balance test output (expressed as KS, paired-t, and unpaired-t test p-values.) We opted for a random search algorithm, as opposed to an "intelligent" evolutionary algorithm because we wanted to evaluate the results of a broad range of weight matrices producing a wide spectrum of balance results, and the standard genetic matching algorithm favors weight matrices that minimize the Diamond and Sekhon (2005) loss function. In our case, random search was sufficient to identify weight matrices producing an extremely high degree of balance.[7]

3 The Placebo Diagnostic Test for ATT

The more similar the matched treated and control groups, the better the matched controls represent the treated in the absence of treatment, and the smaller the bias (and variance) induced by matching discrepancies. In theory, at some high level of balance, matching a given set of treated units should produce similar results regardless of the chosen control group. Any such similarity would be driven entirely by the similarity—across control groups—of mean matched-control outcomes.

The placebo test exploits this idea by randomly dividing the controls into

[7] Subsequently we implemented the standard GENOUD optimization and obtained even better-balancing results that were substantively consistent with our existing findings.

two placebo samples A and B, thereby producing two completely different control groups with very similar features. A and B are composed of different units, but (by construction) there should be no systematic differences between their distributions of observed and unobserved confounders. We then implement the iterative random search matching algorithm described in the prior section two times—once for each treatment and placebo-group pairing. For each pairing, the algorithm repeatedly populates the weight matrix T times (with different random samples from a multivariate standard uniform distribution), performs the matching exercise, and records the identities of matched controls, their mean outcomes (post-treatment earnings), and the balance output. Iterations proceed until the random search is judged to have sufficiently explored the search space and produced results across a broad range of balance levels.[8]

For each of the two placebo groups, this produces a vector that records the mean outcomes for each set of matched controls denoted $\overline{Y}_{a,j}$ with $j = 1, ..., T$ indexing the iterations of the algorithm and $\overline{Y}_{b,k}$ with $k = 1, ..., T$ respectively. We also get a second vector that contains the measure of balance quality associated with each $\overline{Y}_{a,j}$ and $\overline{Y}_{b,k}$. Recall that our measure of balance quality is the lowest p-value from Kolmogorov-Smirnov (KS) tests and either the paired or the unpaired t-test for all raw covariates, two-way interaction, and quadratic terms. We denote these vectors of balance measures by $p_{a,j}$

[8] This can require many thousands of evaluations and days of computation time, even on a fast dedicated server.

and $p_{b,k}$ respectively.

Finally, for all j and k we bin the $\overline{Y}_{a,j}$ and $\overline{Y}_{b,k}$ according to their corresponding p-values into intervals denoted by I_l with $l = 1, ..., \lambda$. In our analysis, we chose $\lambda = 5$, and selected intervals that we were curious about and that represent conventional thresholds for statistical significance. For the runs with the paired t-test we defined the intervals as $I_1 = [0, 0.01], I_2 = (0.01, 0.05], I_3 = (0.05, 0.10], I_4 = (0.10, 1], I_5 = (0.15, 1]$. Since much higher lowest p-values can be achieved using the more lenient unpaired t-tests instead of the paired t-test, we defined the intervals for the former as $\tilde{I}_1 = [0, 0.05], \tilde{I}_2 = (0.05, 0.15], \tilde{I}_3 = (0.15, 0.30], \tilde{I}_4 = (0.30, 0.1], \tilde{I}_5 = (0.50, 1]$.[9]

The last step of the test is then to compute and examine, within each bin, the differences in the mean outcomes between the sets of matched controls from both placebo groups. For the runs with the paired t-test:

$$\delta_l = \overline{Y}_{a,j} - \overline{Y}_{b,k} \, \forall \, (j, k) : p_{a,j} \wedge p_{b,k} \in I_l$$

and for the runs with the unpaired t-test:

$$\delta_l = \overline{Y}_{a,j} - \overline{Y}_{b,k} \, \forall \, (j, k) : p_{a,j} \wedge p_{b,k} \in \tilde{I}_l$$

The logic of our placebo test rests on the idea that the distribution of δ can

[9] The binning scheme we chose is somewhat arbitrary. We confirmed our substantive results were robust across several different binning schemes.

provide information about the quality of the covariate balance achieved and the reliability of the ATT estimate. As long as there are still considerable differences in the mean outcomes between the two matched placebo control groups, at least one of these matched groups is not a valid representation of what would have happened to the treatment group in the absence of the treatment. Under these circumstances, better balance is necessary to assure robust, reliable results.

Using the distribution of δ as a balance measure is appealing for several reasons. First, since a test based on δ never involves looking at the outcomes for the treated units it is "blind to the answer". Second, in contrast to conventional balance tests based on a set of particular X, our placebo test implies a check of the robustness of the matching procedure across multiple models and samples because it involves specifying the combination of X and propensity scores repeatedly in different ways.[10] Third, the placebo test can help to establish a balance threshold for a particular data-set. If a researcher finds that as balance improves the mean of δ moves towards zero, and yet considerable bias remains, then the conclusion would be that better balance or alternative adjustments are required.

[10] The elements of δ result from different (iterative) random draws populating the W weight matrix, and it is this weight matrix that determines which control units get matched.

4 Empirical Assessment of the Placebo Test

4.1 Data

To probe the plausibility of our hypotheses we utilize the National Supported Work Demonstration Program dataset (Dehejia and Wahba (1997; 1999; 2002), Lalonde (1986), Smith and Todd (2001; 2005a; 2005b), which derives from a randomized job training experiment implemented in the mid-1970s. This is the classic dataset that scholars have used repeatedly over several decades to evaluate methods and tools for causal inference. The standard NSW research design is to use the experimentally-derived result as a benchmark ATT estimate, and then attempt to recover this result after replacing the randomized control group with Current Population Survey (CPS) data. The goal is to construct an inferential challenge akin to what analysts face in nonexperimental settings, such that there is a known benchmark answer allowing one to evaluate the reliability of a given estimator.

The dataset is described in detail in Dehejia and Wahba (1995; 1997). It includes covariate information on individuals' age, sex, race, marital status, education, and two years of pretreatment annual income. There are 185 treated units and 15,992 CPS control units. As discussed above, running the placebo test involved randomly dividing the control group into two placebo subgroups, running the random-search matching algorithm for each subgroup, and recording balance-related output based on KS, paired-t, and unpaired-t test p-values for all covariates, squared terms, and two-way interactions. We ran more than 5000 evaluations for each subgroup before halting the procedure and examining the results.

4.2 Results

We first examined whether the distribution of δ within a particular bin was a good signal about the quality of the covariate balance associated with the values in that bin.[11] Covariate balance was first measured by taking the lowest p-value obtained across all Kolmogorov-Smirnov (KS) and paired t-tests. Figure 1 shows the distribution of δ within our p-value intervals, and two features become immediately apparent.

First, the figure conveys strong convergence of the distribution of δ towards zero (both mean and variance) as we go from lower to higher balance quality bins. For example, consider the black line which represents the density distribution of δ in the bin of lowest p-value $[0, 0.01]$. At this low balance level, the distribution of δ, i.e. the differences in mean outcome between the two matched control groups, is fairly wide and centered far to the right of zero. Compare this to the blue line, which represents the density distribution of δ in the bin of lowest p-value $(0.10, 1]$. The mean of this distribution is closer to zero and the variance has decreased. Yet, only in the highest balance bin with lowest p-value of (0.15, 1] does the distribution of δ center in on the close neighborhood of zero. This placebo test result suggests that

[11] Each bin contained at least 100 evaluations. Bins associated with the worst balance estimates had more than 10000 evaluations.

ATT estimates are robust and adequate balance has obtained only for these evaluations with the very highest p-values. Figure 2, which plots the empirical cumulative distribution function of the absolute values of δ for each balance bin makes the convergence in the distribution of δ at higher balance levels even clearer.

The second feature apparent in figure 1 is that—consistent with the findings in Diamond and Sekhon (2005)— our results suggest that a very high balance standard is needed for reliable causal inference (at least in this data-set). The distribution of δ centers around zero only at the highest balance quality bin of lowest p-values $(0.15, 1]$. Note that this refers to the lowest p-value across all KS and *paired* t-tests and thus constitutes a much higher balance hurdle than conventionally demonstrated in the matching literature.

Hitherto, virtually all matching papers only rely on *unpaired* t-tests to evaluate covariate balance. Usually these tests are restricted to the variables used in the matching (no interactions or quadratic terms are tested) and a lowest p-value of higher than 0.05 or 0.1 is considered the cutoff point for sufficient balance. Such a balance standard seems much too lenient and does not allow for reliable causal inference given that considerable differences between the mean outcomes in the two control groups remain.

Figure 1 clarifies this point by plotting the distribution of δ for each lowest p-value bin, this time using the lowest p-value across KS and *unpaired* t-tests as our measure of covariate balance between the treatment group and each of the particular control groups. There is no strong indication of convergence in the distribution of δ towards zero, even for the highest quality bins. Consider the orange line, for example, which plots the distribution of δ in the interval of lowest p-values of $(0.05, 0.15]$, the balance level commonly regarded as sufficient in the literature. Strikingly, the distribution of δ still exhibits a very large variance and the distribution is centered far to the right of zero. Clearly, at this level of balance reliable causal inference seems impossible given the considerable differences observed in the mean outcome between the two control groups.

But even for the highest balance quality bin (with lowest p-values in the interval $(0.5, 1]$, which is astoundingly high by this metric) the distribution of δ is still not centered around zero. The empirical cumulative distribution functions of the absolute values of δ displayed in figure 4 confirms the lack of convergence. As balance (measured by the lowest p-value across KS and unpaired t-tests) improves, the distribution of δ does not collapse over zero. Taken together, these findings suggest that (1) the cutoff commonly used for the lowest p-value is much too low, and that (2) the unpaired t-test does not represent an effective balance criterion.[12]

Thus far, we have shown evidence to support our claims that:

1. the distribution of δ is a good proxy for the quality of the covariate balance achieved between treatment group and each of the particular

[12] Note that in contrast to the usual balance standard used in the literature here we also incorporate the KS test and check balance across all interactions and quadratic terms. Were we to only use the unpaired t-test across the raw covariates the convergence in δ would probably look even worse.

control groups (as measured by the lowest p-value across all KS and paired t-tests).

2. unpaired t-tests do not provide a valid balance criterion.

The last remaining step in validating our placebo test is to show that the convergence in the distribution of δ also gets us closer to the true answer, by which we mean the ATT estimate obtained from the experimental dataset.

Evidence for this claim is provided in table 1 which summarizes the link between mean squared error and the convergence of the distribution of $|\delta|$. Here, MSE is computed as the average squared deviation of the experimental benchmark ATT estimate from the ATT estimates obtained by comparing the mean outcome of the treatment group with the mean outcome of each of the matched control groups contained in each bin.

The upper panel shows the results using the paired t-test, the scenario for which we previously found convergence in the distribution of δ. The convergence is again evident in the table, as the mean and the standard deviation of $|\delta|$ monotonically decreases from lower to higher balance quality bins. Consistent with our argument, the MSE is also monotonically decreasing, and we find a very strong positive correlation between the MSE and the means and variances of $|\delta|$ across the bins. The high correlation suggests that the $|\delta|$ of estimates in a bin contains information about the extent to which those estimates are biased.

The lower panel shows the same type of information using the unpaired

t-test. Here, again we see evidence that the unpaired t-test does not exhibit the nice properties of the paired test. Also, consistent with our argument, we see similarly high positive correlation between the MSE and the mean of $|\delta|$ across the bins even though the mean and standard deviation of $|\delta|$ is no longer monotonically increasing across the bins.

5 Conclusion

We present the placebo test as a new diagnostic instrument intended to supplement, not replace, existing methods for evaluating balance. We agree with Ho et. al. (2006), who caution against relying on one measure to assess balance. We certainly think it is wise, whenever possible, to look at the entire empirical distribution of the matching covariates, especially those believed to produce the most significant confounding.

Moreover, we recognize that the placebo test will not be easy to implement in all cases. When the dataset is such that it is difficult to find good matches using all the controls, finding good matches using only half the controls may well prove impossible. In the NSW empirical example, this was not such a serious problem because there were far more controls than treated units. Were the numbers of treated and control more nearly equal, the matching exercise would have proven much more difficult.[13]

Our final caveat is that it is always dangerous to draw conclusions from

[13] In practice, we have found it difficult to match effectively across many covariates when the ratio of controls to treated is less than 3-to-1.

one empirical case. The NSW example was intended as a plausibility probe. The next step is to perform additional analyses vis-a-vis Monte Carlo experiments and other real-world datasets for which a target "correct" answer is known.

These issues aside, we believe that our proposed diagnostic offers several important benefits. First, it does not require looking at the answers, which means that the test can be designed, performed, and repeated with honesty and impartiality. Second, the test represents a check on the robustness of one's estimates across different datasets and matching models. Third, as we show, examining the distribution of δ as balance improves can provide information about the extent to which matching reduces the bias and variance of treatment effect estimates. Thus, our placebo test is the first diagnostic that allows to establish a balance threshold for a particular data-set.

References

Abadie, A. and Imbens, G. (2006), 'Large Sample Properties of Matching Estimators for Average Treatment Effects', *Econometrica* **74**(2), 235–267.

Campbell, D. (1969), *Artifact in Behavioral Research*, New York: Academic, chapter Artifact and Control, pp. 351–382.

Dehejia, R. (2005), 'Practical propensity score matching: a reply to Smith and Todd', *Journal of Econometrics* **125**(1-2), 355364.

Dehejia, R. and Wahba, S. (1997), *Econometric Methods for Program Evaluation*, Ph. D. Dissertation, Harvard University, chapter Causal Effects in Non-Experimental Studies: Re-Evaluating the Evaluation of Training Programs.

Dehejia, R. and Wahba, S. (1999), 'Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs', *Journal of the American Statistical Association* **94**, 1053–1062.

Ho, D., Imai, K., King, G. and Stuart, E. (2006), Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. Working Paper.

Holland, P. W. (1986), 'Statistics and Causal Inference', *Journal of the American Statistical Association* **81**(396), 945–960.

Imbens, G. (2003), Semiparametric estimation of average treatment effects under exogeneity: A review. MIMEO, Berkley University.

LaLonde, R. J. (1986), 'Evaluating the Econometric Evaluations of Training Programs with Experimental Data', *American Economic Review* **76**, 604–620.

Lu, B. and Rosenbaum, P. R. (2004), 'Optimal matching with two control groups', *Journal of Computational and Graphical Statistics* **13**, 422–434.

Mebane, W. R. J. and Sekhon, J. S. (1998), Genetic optimization using derivatives (genoud). software package. http://sekhon.polisci.berkeley.edu/rgenoud/.

Rosenbaum, P. R. (1987), 'The role of a second control group in an observational study (with discussion)', *Statistical Science* **2**, 292–316.

Rosenbaum, P. R. (2001), 'Stability in the absence of treatment', *Journal of the American Statistical Science Association* **96**, 210–219.

Rosenbaum, P. R. (2002), *Observational Studies*, 2nd edn, New York: Springer-Verlag.

Rosenbaum, P. R. and Rubin, D. B. (1983), 'The Central Role of the Propensity Score in Observational Studies for Causal Effects', *Biometrika* **70**(1), 41–55.

Rosenbaum, P. R. and Rubin, D. B. (1985), 'The bias due to imcomplete matching', *Biometrics* **41**(1), 103–116.

Rubin, D. B. (1974), 'Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies', *Journal of Educational Psychology* **66**, 688–701.

Rubin, D. B. (1976*a*), 'Multivariate Matching Methods that are Equal Percent Bias Reducing, I: Some Examples', *Biometrics* **32**(1), 109–120.

Rubin, D. B. (1976*b*), 'Multivariate Matching Methods that are Equal Percent Bias Reducing, II: Some Examples', *Biometrics* **32**(1), 121–132.

Rubin, D. B. (1978), 'Bayesian Inference for Causal Effects: The Role of Randomization', *Annals of Statistics* **6**(1), 34–58.

Rubin, D. B. (1980), 'Discussion of "Randomization Analysis of Experimental Data in the Fisher Randomization Test" by D. Basu', *Journal of the American Statistical Association* **75**(371), 591–593.

Rubin, D. B. (2001), 'Using Propensity Scores to Help Design Observational Studies: Application to the Tobacco Litigation', *Health Services & Outcomes Research Methodology* **2**, 169188.

Rubin, D. B. and Imbens, G. (2002), *Causal Inference*, Unpublished Manuscript.

Sekhon, J. S. and Mebane, W. R. (1998), 'Genetic Optimization Using Derivatives: Theory and Application to Nonlinear Models', *Political Analysis* **7**, 189–213.

Shadish, W. R., Cook, T. D. and Campbell, D. T. (2002), *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*, Boston: Houghton-Mifflin.

Smith, J. and Todd, P. (2001), 'Reconciling conflicting evidence on the performance of propensity-score matching methods', *American Economic Review* **91**(2), 112–118.

Smith, J. and Todd, P. (2005*a*), 'Does matching overcome LaLonde's critique of nonexperimental estimators?', *Journal of Econometrics* **125**(1-2), 305–353.

Smith, J. and Todd, P. (2005*b*), 'Rejoinder', *Journal of Econometrics* **125**(1-2), 365–375.

Zhang, J. and Rubin, D. (2003), 'Estimation of causal effects via principal stratification when some outcomes are truncated by "death"', *Journal of Educational and Behavioral Statistics* **28**(4), 353–368.

6 Figures

Figure 1:

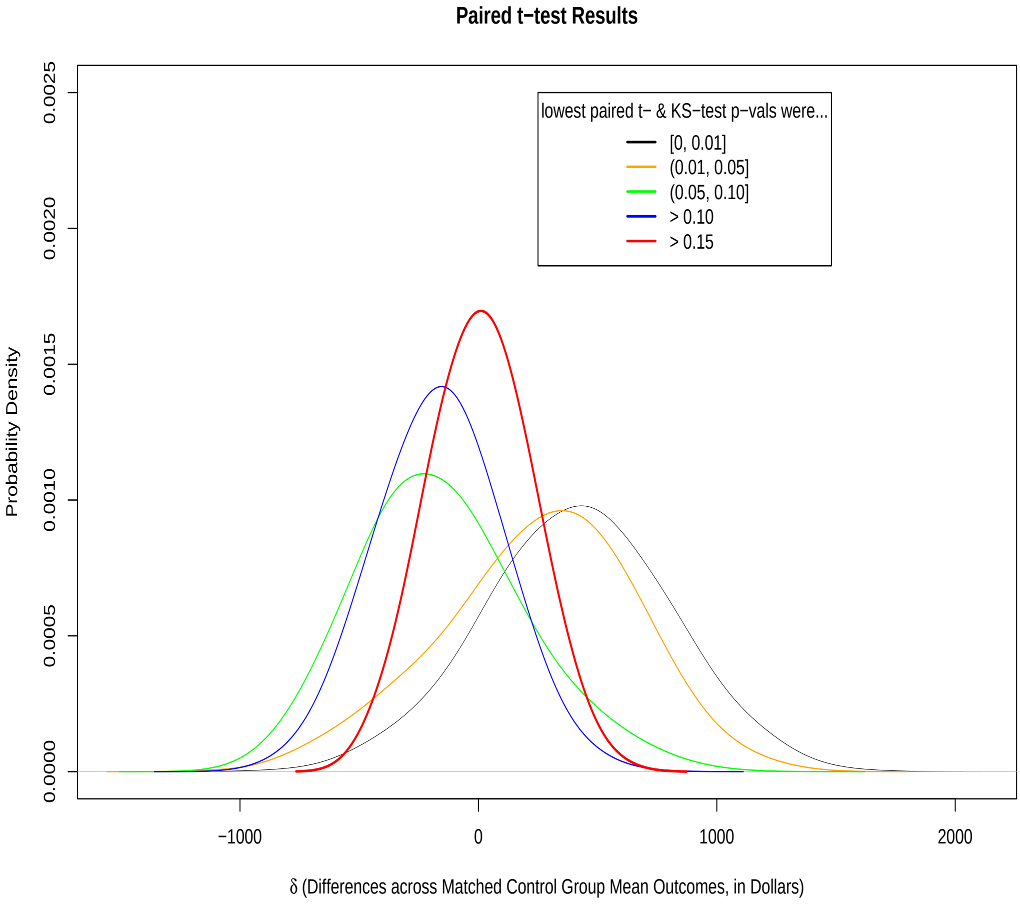


Figure 2:

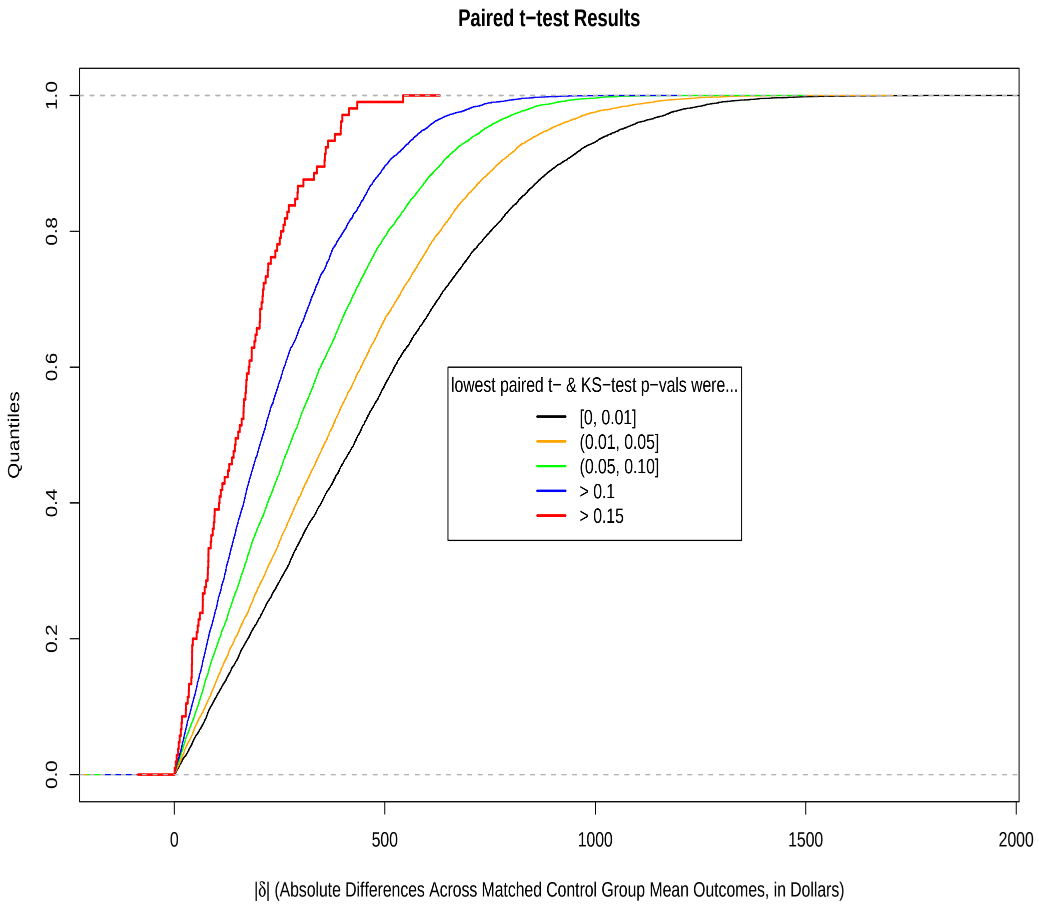

Paired t−test Results
Quantiles
0.0
0.2
0.4
0.6
0.8
1.0
0
500
1000
1500
2000
|δ| (Absolute Differences Across Matched Control Group Mean Outcomes, in Dollars)
lowest paired t− & KS−test p−vals were...
[0, 0.01]
(0.01, 0.05]
(0.05, 0.10]
> 0.1
> 0.15

Figure 3:

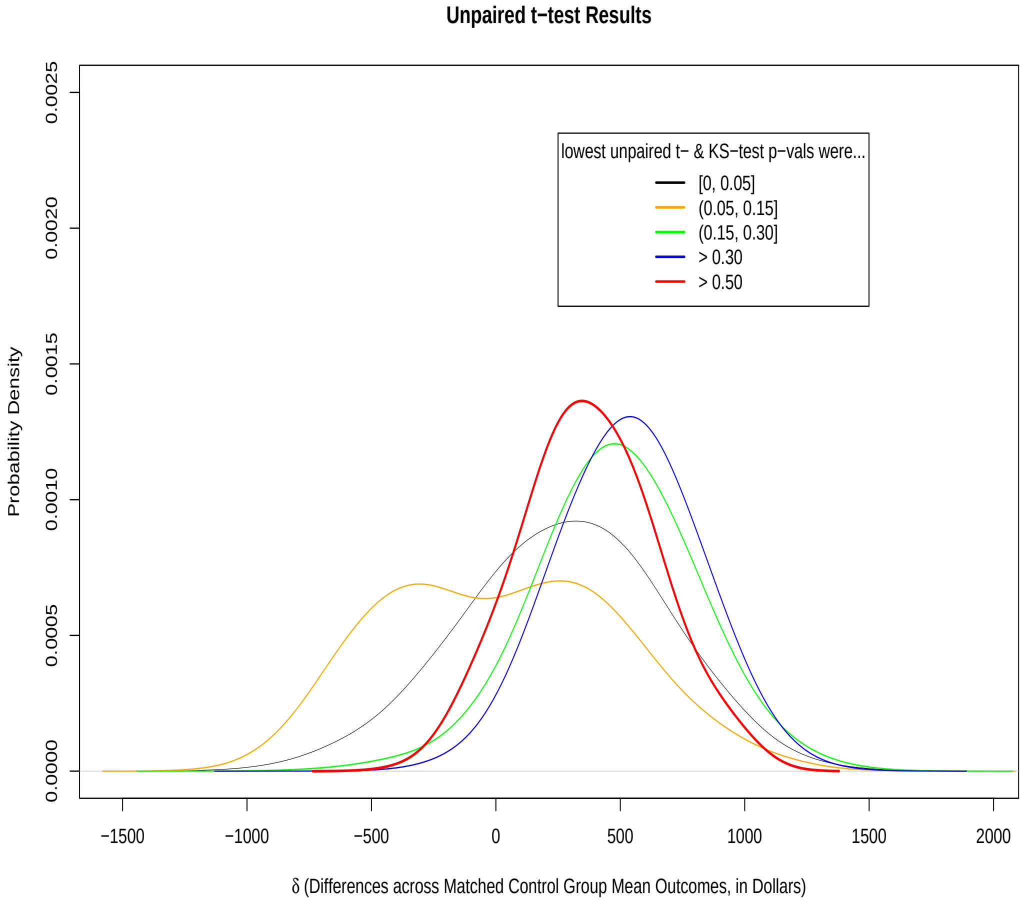
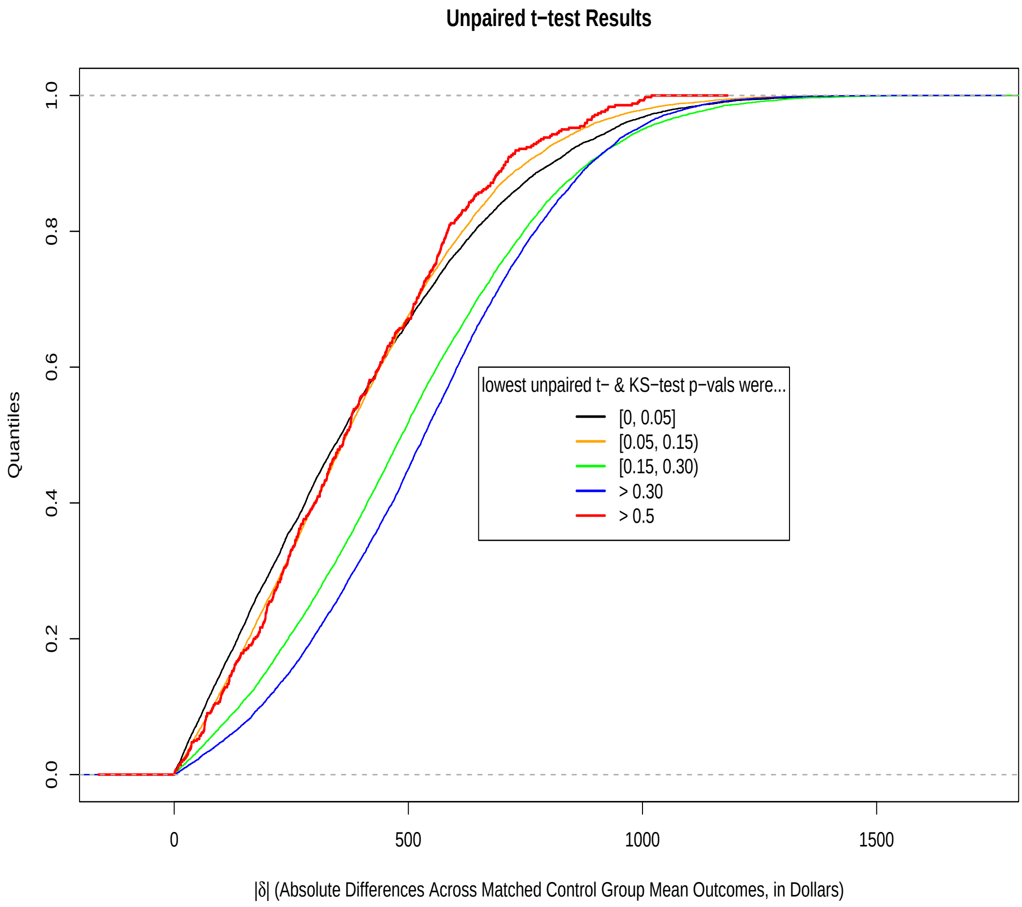

Unpaired t−test Results
Probability Density
0.0000
0.0005
0.0010
0.0015
0.0020
0.0025
−1500
−1000
−500
0
500
1000
1500
2000
δ (Differences across Matched Control Group Mean Outcomes, in Dollars)
lowest unpaired t− & KS−test p−vals were...
[0, 0.05]
(0.05, 0.15]
(0.15, 0.30]
> 0.30
> 0.50

Figure 4:


Unpaired t−test Results
Quantiles
0.0
0.2
0.4
0.6
0.8
1.0
0
500
1000
1500
|δ| (Absolute Differences Across Matched Control Group Mean Outcomes, in Dollars)
lowest unpaired t− & KS−test p−vals were...
[0, 0.05]
[0.05, 0.15)
[0.15, 0.30)
> 0.30
> 0.5

7 Tables

Table 1: Mean Squared Error and the Distribution of δ at Different Balance Levels.

Kolmogorov-Smirnov and Paired T-Tests

Bin I_l	1	2	3	4	5
p-value interval	$[0, 0.01]$	$(0.01, 0.05]$	$(0.05, 0.10]$	$(0.10, 1]$	$(0.15, 1]$
MSE	120683	105812	95750	95678	64448
$\lvert\overline{\delta}_I\rvert$	471	398	320	247	162
$SD\ \lvert\delta_I\rvert$	315	269	223	181	118

$Cor(|\overline{\delta}_I|, MSE)$ 0.95

Kolmogorov-Smirnov and Unpaired T-Tests

Bin $\tilde{I}_l$	1	2	3	4	5
p-value interval	$[0, 0.05]$	$(0.05, 0.15]$	$(0.15, 0.30]$	$(0.30, 1]$	$(0.50, 1]$
MSE	110647	80309	108379	144329	91552
$\lvert\overline{\delta}_{\tilde{I}}\rvert$	400	398	504	538	388
$SD\ \lvert\delta_I\rvert$	283	257	283	265	237

$Cor(|\overline{\delta}_{\tilde{I}}|, MSE)$ 0.81

$|\overline{\delta}_I|$ and $SD\ |\delta_I|$ denote the mean and the standard deviation of the absolute value of the difference of the mean outcomes between each set of matched controls from the two placebo groups computed in each bin I_l or $\tilde{I}_l$ of lowest p-values. For each matched control group, the lowest p-value was obtained by comparing it to the experimental treatment group with respect to all raw covariates plus their two-way interactions and quadratic terms. MSE is the mean squared error of ATT estimates, in each bin, from the experimental benchmark ATT estimate.