GENETIC MATCHING FOR ESTIMATING CAUSAL EFFECTS: A GENERAL MULTIVARIATE MATCHING METHOD FOR ACHIEVING BALANCE IN OBSERVATIONAL STUDIES

Alexis Diamond and Jasjeet S. Sekhon*

Abstract—This paper presents genetic matching, a method of multivariate matching that uses an evolutionary search algorithm to determine the weight each covariate is given. Both propensity score matching and matching based on Mahalanobis distance are limiting cases of this method. The algorithm makes transparent certain issues that all matching methods must confront. We present simulation studies that show that the algorithm improves covariate balance and that it may reduce bias if the selection on observables assumption holds. We then present a reanalysis of a number of data sets in the LaLonde (1986) controversy.

I. Introduction

MATCHING has become an increasingly popular method in many fields, including statistics (Rosenbaum, 2002; Rubin, 2006), economics (Abadie & Imbens, 2006; Dehejia & Wahba, 1999; Galiani, Gertler, & Schargrodsky, 2005), medicine (Christakis & Iwashyna, 2003; Rubin, 1997), political science (Herron & Wand, 2007; Imai, 2005; Sekhon, 2004), sociology (Diprete & Engelhardt, 2004; Morgan & Harding, 2006; Winship & Morgan, 1999), and even law (Epstein et al., 2005; Gordon & Huber, 2007; Rubin, 2001). There is, however, no consensus on how exactly matching ought to be done, how to measure the success of the matching procedure, and whether matching estimators are sufficiently robust to misspecification so as to be useful in practice (Heckman et al., 1998).

Matching on a correctly specified propensity score will asymptotically balance the observed covariates, and it will asymptotically remove the bias conditional on such covariates (Rosenbaum & Rubin, 1983). By covariate balance, we mean that the treatment and control groups have the same joint distribution of observed covariates. The correct propensity score model is generally unknown, and if the model is misspecified, it can increase bias even if the selection on observables assumption holds (Drake, 1993).[1] A misspecified propensity score model may increase the imbalance of some observed variables postmatching, especially if the covariates have nonellipsoidal distributions (Rubin & Thomas, 1992).

Received for publication June 12, 2008. Revision accepted for publication March 14, 2012.

* Diamond: International Finance Corporation; Sekhon: University of California Berkeley and Institute of Governmental Studies.

For valuable comments, we thank Michael Greenstone (the editor), two anonymous reviewers, Alberto Abadie, Henry Brady, Devin Caughey, Rajeev Dehejia, Jens Hainmueller, Erin Hartman, Joseph Hotz, Kosuke Imai, Guido Imbens, Gary King, Walter Mebane Jr., Kevin Quinn, Jamie Robins, Donald Rubin, Phil Schrodt, Jeffrey Smith, Jonathan Wand, Rocío Titiunik, and Petra Todd. We thank John Henderson for research assistance. Matching software that implements the technology outlined in this paper can be downloaded from http://sekhon.berkeley.edu/matching/. All errors are our responsibility.

[1] What we, following the convention in the literature, call the "selection on observables" assumption is actually an assumption that selection is based on observed covariates.

Since the propensity score is a balancing score, covariate imbalance after propensity score matching is a concern. (Rosenbaum and Rubin, 1984) provide an algorithm for estimating a propensity score that involves iteratively checking if matching on the estimated propensity score produces balance. They recommend that the specification of the propensity score be revised until covariate imbalance is minimized.

The importance of iteratively checking the specification of the propensity score model is not controversial in the theoretical literature on matching. Because outcome data are not used in the propensity score, one may consider various models of treatment assignment without introducing sequential testing problems, even if analysts estimate many candidate models and sequentially learn from one specification to the next. This is sometimes considered one of the central benefits of matching (Rubin, 2008). The propensity score is an ancillary statistic for estimating the average treatment effect given the assumption that treatment assignment is ignorable conditional on observed confounders (Hahn, 1998).

The process of iteratively modifying the propensity score to maximize balance is challenging. Applied researchers often fail to report the balance achieved by the propensity score model they settle on. For example, we reviewed all articles in this REVIEW from 2000 to August 2010. In this period, 31 articles discussed matching, and 23 articles presented empirical applications of matching. Of these 23 articles, only 11 provided any measure of covariate balance. Four articles presented difference of means tests for some variables that were included in the propensity score model. And only 1 article presented balance measures for all of the variables that were matched on. Our review of three other leading economics journals found even lower rates of reporting covariate balance.[2] In these three journals, there were 24 empirical applications of matching, of which only 4 presented any balance measures at all. These findings are similar to those found in other disciplines. In a review of the medical literature, Austin (2008) found 47 studies that used propensity score matching but only 2 of them reported standardized measures of covariate balance post-matching.

Our method, genetic matching (GenMatch), eliminates the need to manually and iteratively check the propensity score. GenMatch uses a search algorithm to iteratively check and improve covariate balance, and it is a generalization of propensity score and Mahalanobis distance (MD) matching (Rosenbaum & Rubin, 1985). It is a multivariate matching

[2] For the same time period, we reviewed articles published in the *American Economic Review*, *Journal of Political Economy*, and *Quarterly Journal of Economics*.

The Review of Economics and Statistics, July 2013, 95(3): 932–945

method that uses an evolutionary search algorithm developed by Mebane and Sekhon (1998; Sekhon & Mebane 1998) to maximize the balance of observed covariates across matched treated and control units.

In this paper, we provide a general description of the method, and we evaluate its performance using two simulation studies and an actual data example. In section II, we describe our automated iterative algorithm after a brief background discussion. In section III, we present our two simulation studies. The first simulation study was developed by other researchers to demonstrate the effectiveness of machine learning algorithms for semiparametric estimation of the propensity score (Lee, Lessler, & Stuart, 2010; Setoguchi et al., 2008). We use this study to benchmark GenMatch against two alternative methods proposed in the literature. In the second simulation, we evaluate the performance of GenMatch when the covariates are distributed as they are in a well-known data set, the Dehejia and Wahba (1999) sample of the LaLonde (1986) data set.

In section IV, we reanalyze several data sets that have been explored in the controversy spawned by LaLonde (1986): Dehejia and Wahba (1999, 2002), Dehejia (2005), and Smith and Todd (2001, 2005a, 2005b). We examine these data sets because they are well known and because they offer an opportunity to see if our algorithm is able to improve on the covariate balance found in a literature that has generated a number of propensity score models. We offer concluding comments in section V.

II. Methods

A. *Propensity Score Matching*

In observational studies, variables that affect the response may be distributed differently across treatment groups and so confound the treatment effect (Cochran & Rubin, 1973). Matching assumes selection on observables or, using the conditional independence notation of Dawid (1979), $T \perp\!\!\!\perp U \mid X$, where T denotes the treatment and X and U are observed and unobserved covariates, respectively. This implies that confounding from both observed and unobserved variables can be removed by achieving covariate balance or $T \perp\!\!\!\perp X$.

Matching on the true propensity score adjusts for observed confounders. The propensity score, $\pi(X_i)$, is the conditional probability of assignment to treatment given the covariates:

$$\pi(X_i) \equiv \Pr(T_i = 1 \mid X_i) = E(T_i \mid X_i). \quad (1)$$

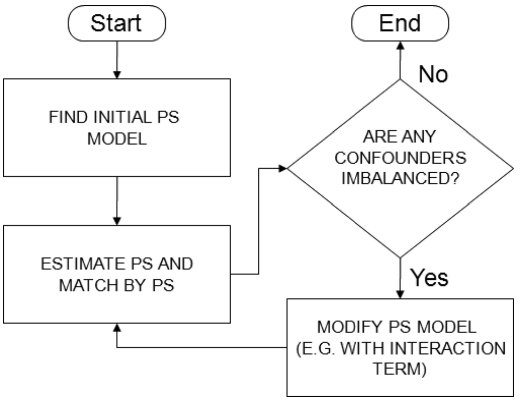
The key property of interest is that treatment assignment and the observed covariates are conditionally independent given the true propensity score, that is,

$$X \perp\!\!\!\perp T \mid \pi(X). \quad (2)$$

Equation (2) is theorem 1 in Rosenbaum and Rubin (1983).

The main implication of equation (2) is that "if a subclass of units or a matched treatment-control pair is homogeneous in $\pi(X)$, then the treated and control units in that subclass or matched pair will have the same distribution of X" (Rosenbaum and Rubin, 1983, p. 44). Matching on the true propensity score results in the observed covariates (X) being asymptotically balanced between treatment and control groups.

FIGURE 1.—FLOWCHART OF ALGORITHM FOR ITERATIVE ESTIMATION OF A PROPENSITY SCORE MODEL



The propensity score is a balancing score: conditioning on the true propensity score asymptotically balances the observed covariates. This property leads to what has been called the propensity score tautology (Ho et al., 2007). Since the propensity score is a balancing score, the estimate of the propensity score is consistent only if matching on this propensity score asymptotically balances the observed covariates. This tautology can be used to judge the quality of an estimated propensity score. If the distributions of observed confounders are not similar after matching on an estimated propensity score, the propensity score must be misspecified or the sample size is too small for the propensity score to remove the conditional bias.

Therefore, it is important to assess covariate balance in the matched sample and modify the propensity score model with the aim of balancing the covariates. Rosenbaum and Rubin (1984) recommend an iterative approach for achieving covariate balance. After propensity score matching, covariate balance is assessed, and the propensity score is modified accordingly. The iteration ends when acceptable balance is achieved, although it is generally desirable to maximize balance without limit. This manual and iterative algorithm is outlined in figure 1. This advice is echoed by others, such as Austin (2008).

A crucial step in this iterative algorithm is evaluating balance. Researchers frequently do not clearly state how they evaluated postmatching covariate balance, and there is no consensus in the literature on how best to measure balance. Rosenbaum and Rubin (1984) recommend using F-ratios to measure individual covariate balance, but alternatives include likelihood ratios, standardized mean differences, eQQ plots, and Kolmogorov-Smirnov (KS) test statistics (Austin 2009).[3]

[3] The KS test statistic, the maximum discrepancy in the eQQ plot, is sensitive to imbalance across the empirical distribution.

Whatever the balance statistic, iterative methods are not only laborious, but there is also no guarantee that overall balance improves after refinement of the propensity score. Moreover, as we noted in section I, applied researchers often/do not follow the iterative approach.

B. Genetic Matching

Mahalanobis distance. Before turning to GenMatch itself, it is useful to discuss Mahalanobis distance (MD) matching because GenMatch is a generalization of this distance metric. MD, rarely used in economics, is more common in other fields, such as statistics. It is a scalar quantity that measures the multivariate distance between individuals in different groups. The MD between the X covariates for two units i and j is

$$\text{MD}(X_i, X_j) = \sqrt{(X_i - X_j)^T S^{-1} (X_i - X_j)},$$

where S is the sample covariance matrix of X and X^T is the transpose of the matrix X. The matrix X may contain not only the observed confounders but also terms that are functions of them (for example, power functions, interactions).

Because MD does not perform well when covariates have nonellipsoidal distributions, Rosenbaum and Rubin (1983) suggested matching on the propensity score, $\pi(X) = P(T = 1 \mid X)$ instead. However, due to sampling variation and nonexact matching, $T \perp\!\!\!\perp X \mid \pi(X)$ may not hold after matching on the propensity score. Rosenbaum and Rubin (1985) therefore recommend that in addition to propensity score matching, one should match on individual covariates by minimizing the MD of X to obtain balance on X. Hence, they argue that the propensity should be included among the covariates X or, alternatively, one may first match on the propensity score and then match based on MD within propensity score strata.

A more general distance metric. The GenMatch algorithm searches a range of distance metrics to find the particular measure that optimizes postmatching covariate balance. Each potential distance metric considered corresponds to a particular assignment of weights W for all matching variables. The algorithm weights each variable according to its relative importance for achieving the best overall balance. One must decide how to measure covariate balance and specify a loss function. GenMatch matches by minimizing a generalized version of Mahalanobis distance (GMD), which has the additional weight parameter W. Formally,

$$\text{GMD}(X_i, X_j, W) = \sqrt{(X_i - X_j)^T (S^{-1/2})^T W S^{-1/2} (X_i - X_j)}, \quad (3)$$

where W is a $k \times k$ positive definite weight matrix and $S^{-1/2}$ is the Cholesky decomposition of S, that is, $S = S^{-1/2}(S^{-1/2})^T$. All elements of W are restricted to 0 except those down the main diagonal, which consists of k parameters that must be chosen.

One may match on the propensity score in addition to the covariates. Therefore, X in equation (3) may be replaced with Z, where Z is a matrix consisting of both the propensity score, $\pi(X)$, and the underlying covariates X.[4] In this case, if optimal balance is achieved by simply matching on the propensity score, then the other variables will be given a zero weight and GenMatch will be equivalent to propensity score matching.[5] Alternatively, GenMatch may converge to giving 0 weight to the propensity score and a weight of 1 to every other variable in Z. This would be equivalent of minimizing the MD. Usually, however, the algorithm will find that neither minimizing the MD nor matching on the propensity score minimizes the loss function and will search for improved metrics that optimize balance.

Generally it is recommended that GenMatch be started with a propensity score if one is available. In all applications in this paper, we provide GenMatch with a fixed simple linear additive propensity score—that is, without any interactions or high-order terms.

An iterative algorithm. GenMatch automates the iterative process of checking and improving overall covariate balance and guarantees asymptotic convergence to the optimal matched sample. It may or may not decrease the bias in the conditional estimates. However, by construction, the algorithm will improve covariate balance, if possible, as measured by the particular loss function chosen to measure balance.

The GenMatch algorithm minimizes any loss function specified by the user, but the choice must be explicit. It is recommended that the loss function include individual balance measures that are sensitive to many forms of imbalance, such as KS test statistics, and not simply difference of means tests. In GenMatch, the default loss function requires the algorithm to minimize overall imbalance by minimizing the largest individual discrepancy, based on p-values from KS tests and paired t-tests for all variables that are being matched on.[6]

The algorithm uses p-values so that results from different tests can be compared on the same scale. Because the sample size is fixed within the optimization, the general concern that p-values depend on sample size does not apply (Imai, King, & Staurt, 2008).

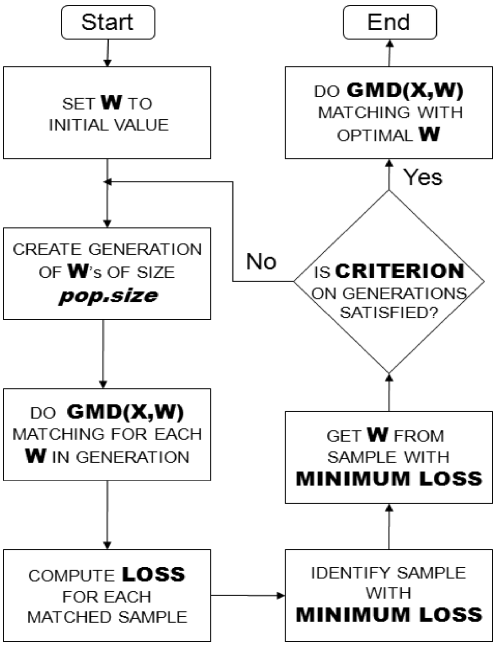
GenMatch uses a genetic search algorithm to choose weights, W, which optimize the loss function specified.

[4] In practice, it may be preferable to match on the propensity score and the covariates after they have been made orthogonal to it. This may be accomplished by regressing each covariate on $\pi(X)$. Moreover, if one is using a propensity score estimated by logistic regression, it may be preferable to match not on the predicted probabilities but on the linear predictor, as the latter avoids compression of propensity scores near 0 and 1.

[5] Technically, the other variables will be given weights just large enough to ensure that the weight matrix is positive definite.

[6] Using p-values may be preferable with the KS bootstrap test since the test statistic is not monotonically related to the p-value when there are point masses in the empirical distribution (Abadie, 2002).

FIGURE 2.—FLOWCHART OF THE GENETIC MATCHING ALGORITHM



The algorithm proposes batches of weights, Ws, and moves toward the batch of weights that maximize overall balance—that is, minimize loss. Each batch is a generation and is used iteratively to produce a subsequent generation with better candidate Ws. The size of each generation is the population size (say, 1,000) and is constant for all generations. Increasing the population size usually improves the overall balance achieved by GenMatch. The algorithm will converge asymptotically in population size (Mebane & Sekhon, 2011; Sekhon & Mebane, 1998). The GenMatch algorithm is summarized in figure 2.

Each W corresponds to a different distance metric, as defined in equation (3). For each generation, the sample is matched according to each metric, producing as many matched samples as the population size. The loss function is evaluated for each matched sample, and the algorithm identifies the weights corresponding to the minimum loss. The generation of candidate trials evolves toward those containing, on average, better Ws and asymptotically converges toward the optimal solution: the one that minimizes the loss function. Further computational details are provided in Sekhon (2011).

The key decisions GenMatch requires the researcher to make are the same that must be made when using any matching procedure: what variables to match on, how to measure postmatching covariate balance, and how exactly to perform the matching.

Matching methods. GenMatch, along with its generalized distance metric, equation (3), and iterative algorithm (figure 2), can be used with any arbitrary matching method. For example, it could be used to conduct nearest-neighbor matching with or without replacement, with or without a caliper, or to conduct optimal full matching (Hansen, 2004; Hansen & Klofer, 2006; Rosenbaum, 1991). Regardless of the precise matching method used, GenMatch will modify the distance metric in an attempt to optimize postmatching covariate balance. The chosen estimand is also arbitrary. Just as the precise model used to estimate the propensity score does not imply a particular matching method or estimand, GenMatch can be used with different matching methods and to estimate different estimands.

In all of the analyses in this paper, we estimate the average treatment effect on the treated (ATT) by one-to-one matching with replacement. Aside from the generalized distance metric and the iterative algorithm, the matching method used is identical to that of Abadie and Imbens (2006).[7] We use matching with replacement because this procedure will result in the highest degree of balance in the (observed) variables and the lowest conditional bias (Abadie & Imbens, 2006). Other matching procedures may, however, result in more efficient estimates.[8]

III. Monte Carlo Experiments

Two sets of Monte Carlo experiments are presented. The first set of simulations has been used in the matching literature by a number of authors to evaluate the behavior of machine learning algorithms for semiparametric estimation of the propensity score. This set was developed by Setoguchi et al. (2008) and subsequently used by Lee et al. (2010). The advantage of these simulations is that they were developed by other researchers for the purposes of judging the relative merits of different matching methods.

We base the second set of simulations on the Dehejia and Wahba (1999) sample of the LaLonde (1986) experimental data set. This set of simulations is focused on determining how GenMatch performs when the covariates are distributed as they are in this well-known data set.

A. *Simulation Study 1: Comparing Machine Learning Algorithms*

We use the same simulation setup as Lee et al. (2010).[9] They find that the best performance is achieved by two different ensemble methods: random forests and boosted classification and regression trees (CART). We compare the performance of GenMatch to these ensemble methods along with a simple linear fixed logistic regression model. There are many alternative methods of estimating the propensity score

[7] Following Abadie and Imbens (2006), all ties are kept and averaged over. The common alternative of breaking ties at random leads to underestimation of the variance of matched estimates.

[8] In results not shown, using other methods, such as matching without replacement, does not change the qualitative conclusions discussed below about the relative advantages of GenMatch.

[9] Lee et al. (2010) modify the Setoguchi et al. (2008) simulations in that the outcome variable is continuous instead of binary, and they use the simulations to evaluate the performance of propensity score weighting.

semiparametrically (Lehrer & Kordas, 2013). We focus on these two ensemble methods because they have been used with these simulations by other authors.

Classification and regression trees are widely used in machine learning and statistics (Breiman et al., 1984). Trees methods estimate a function by recursively partitioning the data, based on covariates, into regions. The sample mean of the outcome variable is equal to the estimated function within a given region. The data are partitioned so as to make the resulting regions as homogeneous as possible. In this way, the prediction error is minimized. Tree methods are insensitive to monotonic functions of the data, and interactions and nonlinearities are naturally approximated by the recursive splits.

CARTs may approximate smooth functions poorly, however, and they are prone to overfitting the data. To overcome these difficulties, various ensemble approaches are used.[10] In these approaches, instead of fitting one large tree, subsamples of the data are taken and multiple trees fit. Each individual tree is set to be weak so as to prevent overfitting. But the trees are combined to form a "committee" that is a powerful learner. In the case of random forests, the original data set is resampled with replacement, as in bootstrapping, and a tree is fit to each bootstrap sample using only a random subset of the available covariates (Breiman, 2001). All of the trees (across bootstrap samples) are then combined to make a prediction for each observation. In the case of boosting, the classification algorithm is repeatedly applied to modified versions of the data (Schapire, 1990). In each pass through the data, the observations are reweighed so as to increase the influence of observations that were previously poorly classified. The predictions from this sequence of classifiers are then combined, often by weighted majority voting.

These simulations consist of ten covariates (X_k, $k = 1, \ldots, 10$): four confounders associated with both treatment and outcome, three treatment predictors, and three outcome predictors.[11] Six covariates ($X_1, X_3, X_5, X_6, X_8, X_9$) are binary, whereas four (X_2, X_4, X_7, X_{10}) are standard normal. Treatment is binary, and the average probability of treatment assignment at the average value of the covariates is ≈ 0.5. There are seven different scenarios that differ in the degree of linearity and additivity in the true propensity score model—that is, the degree to which the propensity score model includes nonlinear (quadratic) terms and interactions. The seven scenarios have the following properties:

A: Additivity and linearity (mean effects only)
B: Mild nonlinearity (one quadratic term)
C: Moderate nonlinearity (three quadratic terms)
D: Mild nonadditivity (three two-way interaction terms)
E: Mild nonadditivity and nonlinearity (three two-way interaction terms and one quadratic term)
F: Moderate nonadditivity (ten two-way interaction terms)
G: Moderate nonadditivity and nonlinearity (ten two-way interaction terms and three quadratic terms)

The continuous outcome Y is always generated by a linear combination of treatment T and the confounders such that $Y = \alpha_k X_k + \gamma T$, where $\gamma = -0.4$. The values of α and further details of both how the outcome is generated and how treatment is assigned in each scenario are provided in the appendix.

We report results for two different data set sizes: $n = 1{,}000$, and $n = 5{,}000$.[12] One thousand simulated data sets were generated for each scenario. The random forest and boosted CART models are implemented using the same software and parameters as Lee et al. (2010). The random forest models are implemented using the randomForest package in `R` with the default parameters (Liaw & Wiener, 2002). The boosted regression trees are implemented using the `twang` package in `R` (Ridgeway, McCaffrey, & Morral, 2010). The parameters used are those recommended by McCaffrey, Ridgeway, and Morral (2004).[13] GenMatch is asked to optimize balance for all ten observed covariates using its default parameters.

Table 1 presents the results. It displays the bias and the root mean squared error (RMSE) of the estimates. Bias is reported as the absolute percentage difference from the true treatment effect of -0.4. When the true propensity score is linear and additive in the covariates, scenario A, all methods perform well. Since in scenario A the fixed logistic regression is correctly specified, it has the smallest absolute bias and the second smallest RMSE of all estimators. In all scenarios, GenMatch has the smallest RMSE. GenMatch also has the smallest bias in all scenarios except for the first, where the correctly specified propensity score model has less bias.

As the scenarios become more nonlinear and nonadditive, the fixed linear additive propensity score performs worse. The random forests method performs well across the scenarios. Aside from scenario A, it has the second-lowest RMSE after

[10] Hastie, Tibshirani, and Friedman (2009) provide a review of modern statistical learning and data mining algorithms.

[11] Note that the treatment predictors do not directly affect the outcome so they are instruments. In the simulations, such predictors are included to make the adjustment task more difficult. In practice, instruments should not be included in the propensity score when assuming selection on observed variables (and not in an OLS regression model if that is used). If there exists any bias because of unobserved confounding and if the relationships between the variables are linear, the inclusion of instruments in the propensity score (or OLS model) will increase asymptotic bias (Bhattacharya & Vogt, 2007; Wooldridge, 2009). In the nonparametric case, the direction of the bias is less straightforward, but increasing asymptotic bias is possible (Pearl, 2012). Of course, in practice, if one has an instrument in an observational study, one should use an instrumental variable estimator.

[12] The results for these two samples sizes are consistent with the results for the other sample sizes that were tried ($n = 500$, $n = 10{,}000$, $n = 20{,}000$).

[13] The parameters are 20,000 iterations and a shrinkage parameter of 0.0005. The shrinkage parameter reduces the loss for any misclassification, and it reduces how quickly the weights in the boosting algorithm change over iterations. A smaller shrinkage parameter results in a slower algorithm, but one that may have better out-of-sample performance because of less overfitting (Buhlmann & Yu, 2003; Friedman, 2001).

TABLE 1.—PERFORMANCE OF MATCHING ESTIMATION METHODS, SIMULATION STUDY 1

		Scenario						
Metric	Method	A	B	C	D	E	F	G
For sample size 1,000								
Absolute bias (percent)	GenMatch	1.64	0.976	1.85	0.042	0.375	0.107	2.39
	Logit	0.395	3.73	12.6	6.51	9.58	8.98	16.8
	RFRST	9.39	6.3	1.9	4.72	1.85	3.9	4.76
	BOOST	23.7	19	11.2	25.9	20.2	23.1	14.5
RMSE	GenMatch	0.0274	0.0259	0.0359	0.0286	0.027	0.0275	0.0334
	Logit	0.0562	0.0574	0.0705	0.0674	0.0698	0.0668	0.0837
	RFRST	0.0626	0.0548	0.06	0.055	0.0521	0.0532	0.0596
	BOOST	0.151	0.132	0.15	0.162	0.135	0.147	0.154
For sample size 5,000								
Absolute bias (percent)	GenMatch	0.694	0.0913	0.55	0.917	0.882	1.11	0.334
	Logit	0.00172	4.64	13.4	6.25	10.3	8.8	16.3
	RFRST	6.63	3.33	3.47	3.29	0.556	4.05	3.45
	BOOST	3.13	7.37	6.29	10.7	8.78	11.3	10.4
RMSE	GenMatch	0.013	0.0117	0.0219	0.0145	0.0136	0.0147	0.0191
	Logit	0.022	0.0278	0.0569	0.0345	0.0469	0.0411	0.0675
	RFRST	0.0321	0.0214	0.0288	0.0241	0.0206	0.0265	0.0251
	BOOST	0.0461	0.0381	0.0535	0.0516	0.0441	0.0532	0.059

GenMatch = Genetic Matching; Logit = logistic regression; RFRST = random forest; BOOST = boosted CART.

GenMatch, and in scenario A, it has the third-lowest RMSE. Boosted CART performs relatively poorly. When the sample size is 1,000, it has the highest RMSE in all seven scenarios and the largest absolute bias in all scenarios except C, where the bias of logistic regression is worse. When $n = 5{,}000$, although boosted CART still performs worse than the other two adaptive methods, its performance improves the most with the increase in sample size. With the larger sample size, boosted CART has lower RMSE than the fixed logistic regression model in scenarios C, E, and G, but it has higher RMSE in scenarios A, B, D, and F. Across scenarios, either boosted CART or fixed logistic regression has the highest RMSE. All methods perform better with the additional data, although the relative performance between methods changes little.

The absolute bias for GenMatch is never large. In the $n = 1{,}000$ case, the largest percentage bias for GenMatch is 2.39% and this occurs in scenario G. For $n = 5{,}000$, the largest GenMatch bias is 1.11% (scenario F). In the $n = 1{,}000$ case, the largest absolute bias for random forest is 9.39% (scenario A); for boosted CART, it is 25.9% (scenario D); and for logistic regression, it is 16.8% (scenario G). For the $n = 5{,}000$ case, the largest absolute bias for random forest is 4.05% (scenario A); for boosted CART, it is 11.3% (scenario F); and for logistic regression, it is 16.3% (scenario G).

Figures 3 and 4 present balance statistics for the confounders in the $n = 1{,}000$ simulations.[14] For each scenario and method, a box plot is provided that displays the distribution of the smallest p-value in each of the 1,000 matched datasets across t-tests and KS tests. In all seven scenarios, GenMatch has the best balance, even in scenario A, where the logistic regression is correctly specified. After GenMatch, either logistic regression or random forests provides the best balance depending on the scenario. Although there is a relationship between the balance observed in the covariates as shown in the figures and the bias estimates in table 1, it is less than perfect. This highlights the problem of choosing how to best measure covariate balance, which remains an open research question.

[14] The balance figures for the $n = 5000$ simulations are similar.

B. *Simulation Study 2: LaLonde Data*

In this simulation study, the distribution of covariates is based on the Dehejia and Wahba (1999) experimental sample of the LaLonde (1986) data.[15] This experiment offers a more difficult case for matching than the previous simulation study. Some of the baseline variables are discrete, and others contain point masses and skewed distributions. None of the covariates, as they are based on real data, have ellipsoidal distributions. The propensity score is not correctly specified, and the mapping between X and Y is nonlinear. The selection into treatment is more extreme than in the previous simulations study. A greater proportion of the data has either a very high or very low probability of receiving treatment. This feature was adopted to be consistent with the observational data set that LaLonde created. The sample is not large, which makes the matching problem more difficult. There are 185 treated and 260 control observations.

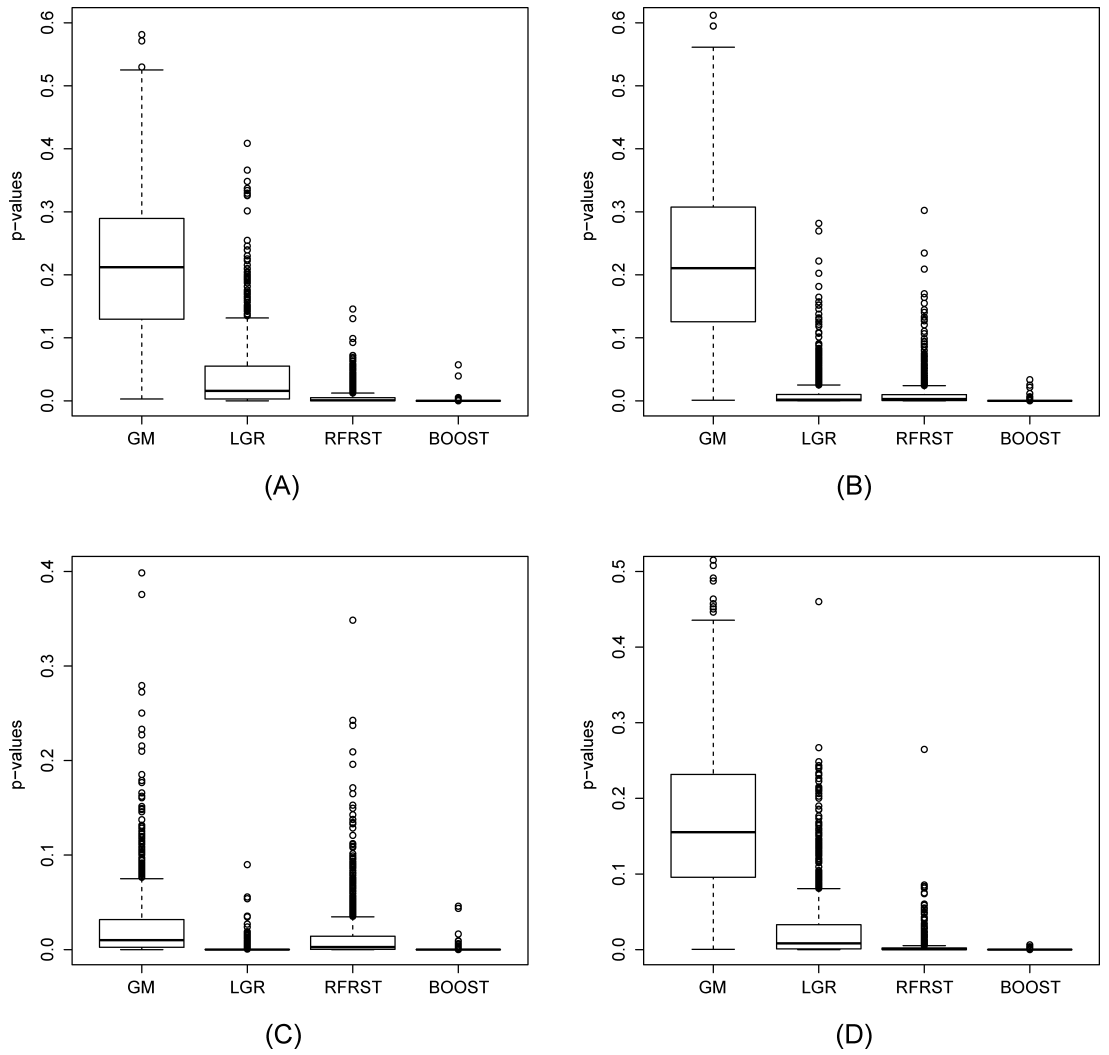
In this simulation, we assume a homogeneous treatment effect of \$1,000. The equation that determines outcomes Y (fictional earnings) is

$$Y = 1000\,T + .1\,\exp[.7\,\log(\text{re74} + .01) + .7\,\log(\text{re75} + 0.01)] + \epsilon,$$

where $\epsilon \sim N(0, 10)$, re74 is real earnings in 1974, re75 is real earnings in 1975, and T is the treatment indicator. The

[15] Adjusting the simulations so that they are based on either the entire LaLonde male sample or the early randomization sample of Smith and Todd (2005a) produces results similar to those presented here.

FIGURE 3.—BALANCE OF MATCHING ESTIMATION METHODS IN SIMULATION STUDY 1, $N = 1{,}000$



Box plots of the smallest p-value of the balance tests in 1,000 draws. Results for simulations A−D. GM = Genetic Matching; LGR = logistic regression; RFRST = random forest; BOOST = boosted CART.

mapping from baseline covariates to Y is obviously nonlinear, and only two of the baseline variables are directly related to Y.

The true propensity score for each observation, π_i, is defined by

$$\pi_i = \text{logit}^{-1}\left[1 + .5\hat{\mu} + .01\ \text{age}^2 - .3\ \text{educ}^2 - .01\ \log(\text{re74} + .01)^2 + .01\ \log(\text{re75} + .01)^2\right], \quad (4)$$

where $\hat{\mu}$ is the linear predictor obtained by estimating a logistic regression model and the dependent variable is the observed treatment indicator in the Dehejia and Wahba (1999) experimental sample of the LaLonde (1986) data. This propensity score is a mix of the estimated propensity score in the Dehejia and Wahba sample plus extra variables in equation (4), because we want to ensure that the propensity model estimated in the Monte Carlo samples is badly misspecified. The linear predictor is

$$\begin{aligned}\hat{\mu} = {} & 1 + 1.428 \times 10^{-4}\text{age}^2 - 2.918 \times 10^{-3}\text{educ}^2 \\ & - .2275\ \text{black} + -.8276\ \text{Hisp} + .2071\ \text{married} \\ & - .8232\ \text{nodegree} - 1.236 \times 10^{-9}\text{re74}^2 \\ & + 5.865 \times 10^{-10}\text{re75}^2 - .04328\ \text{u74} - .3804\ \text{u75},\end{aligned}$$

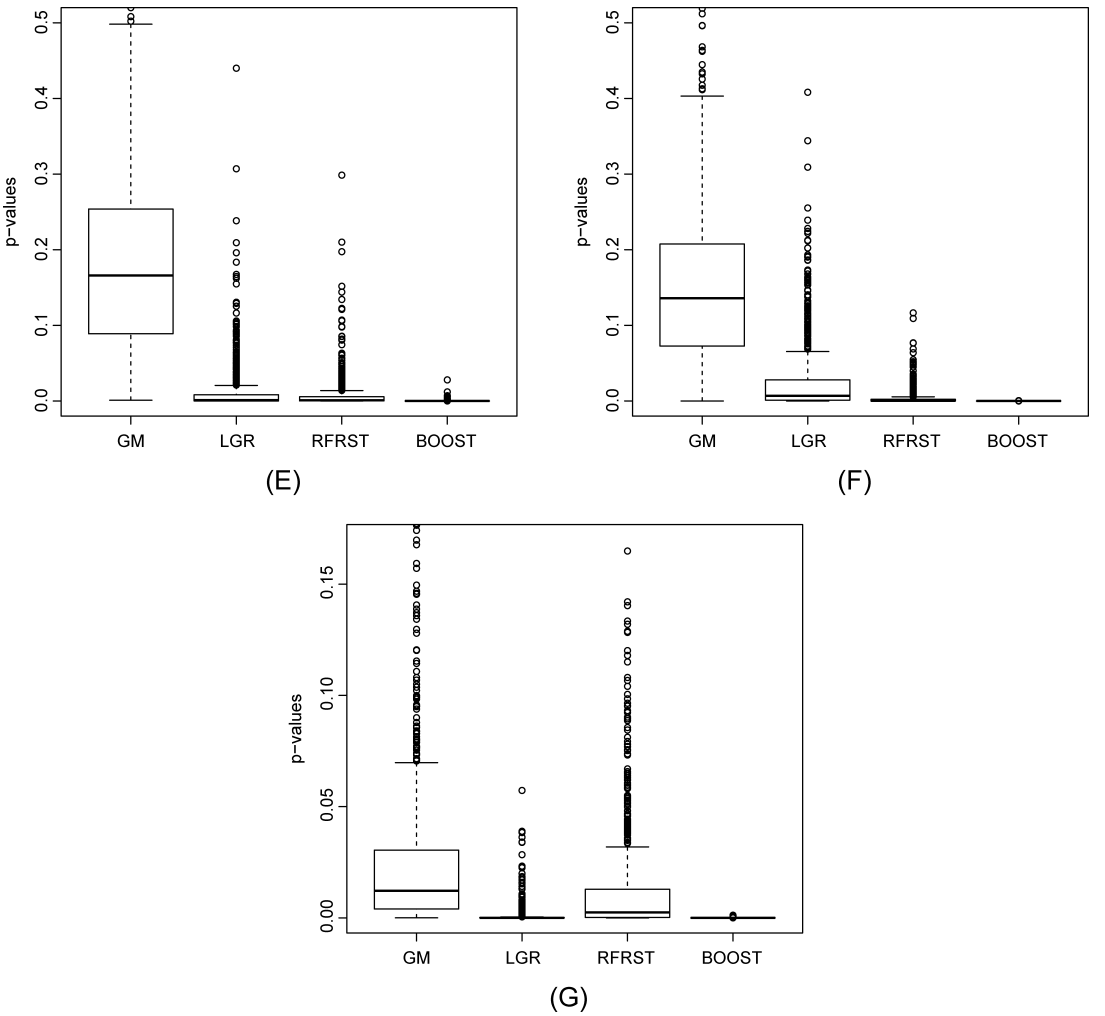
where u74 is an indicator variable for real earnings in 1974 equal to 0 and u75 is an indicator variable for real earnings in 1975 equal to 0.

In each Monte Carlo sample of this experiment, the propensity score is estimated using logistic regression and the following incorrect functional form:

$$\begin{aligned}\hat{\mu}^* = {} & \alpha + \alpha_1\ \text{age} + \alpha_2\ \text{educ} + \alpha_3\ \text{black} + \alpha_4\ \text{Hisp} \\ & + \alpha_5\ \text{married} + \alpha_6\ \text{nodegree} + \alpha_7\ \text{re74} + \alpha_8\ \text{re75} \\ & + \alpha_9\ \text{u74} + \alpha_{10}\ \text{u75}.\end{aligned}$$

Table 2 presents the results for this Monte Carlo experiment based on 1,000 samples. As before, we compare

FIGURE 4.—BALANCE OF MATCHING ESTIMATION METHODS IN SIMULATION STUDY 1, $N = 1{,}000$



Box plots of the smallest p-value of the balance tests in 1,000 draws. Results for simulations E–G. GM = Genetic Matching; LGR = logistic regression; RFRST = random forest; BOOST = boosting.

TABLE 2.—PERFORMANCE OF MATCHING ESTIMATION METHODS, IN SIMULATION STUDY 2

			Bias	RMSE
Estimator	Bias %	RMSE	Bias Genmatch	RMSE Genmatch
GenMatch	4.32	512		
RFRST	81.3	2,223	18.8	4.34
BOOST	103.9	2,492	24	4.87
Logit	51.2	1,832	11.9	3.58
Raw	48.5	1,611	11.2	3.15

GenMatch = Genetic Matching; RFRST = random forest, BOOST = boosted CART; logit = logistic regression; raw = simple mean differences.

GenMatch with random forest and boosted CART, along with the misspecified propensity score.

The raw unadjusted estimate (the sample mean of treated minus the sample mean of controls) has a bias of 48.5% and an RMSE of 1,611. Matching on the fixed logistic regression model increases the absolute bias relative to not adjusting at all to 51.2%, and it also increases the RMSE to 1,832. In contrast, GenMatch has a bias of 4.32% and an RMSE of 512, although it conditions on the same observables. Estimating the propensity score with random forests produces a bias of 81.3% and an RMSE of 2,223. Boosted CART has similar performance with a bias of 103.9% and an RMSE of 2,492.

Of the methods considered, only GenMatch reduces the RMSE relative to the unadjusted estimate. The bias of the other adjustment methods ranges from 11.9 times that of GenMatch for the fixed logistic regression model to 24 times GenMatch for boosted CART. The fixed logistic regression specification has an RMSE of 3.58 times that of GenMatch, while random forests has an RMSE 4.34 times that of GenMatch and boosted CART has an RMSE of 4.87 times that of GenMatch.

This simulation shows that matching methods may perform worse than not adjusting for covariates even when the selection on observables assumption holds. As the sample size increases, the behavior of all of these matching methods will improve relative to the unadjusted estimate since the selection on observables assumption does hold. See, for example, the results of simulation study 1 in section IIIB.

IV. Empirical Example: Job Training Experiment

Following LaLonde (1986), Dehejia and Wahba (1999, 2002; Dehejia 2005) (DW), and Smith and Todd (2001, 2005a, 2005b) (ST), we examine data from a randomized job training experiment, the National Supported Work Demonstration Program (NSW), combined with observational survey data. This data set has been analyzed by DW, ST, and many others, and it has been widely distributed as a teaching tool for use with matching software.

LaLonde's goal was to design a test bed for observational methods. He used the NSW experimental data to establish benchmark estimates of average treatment effects. Then, to create an observational setting, data from the experimental control group were replaced with data from the Current Population Survey (CPS) or, alternatively, the Panel Study of Income Dynamics (PSID). LaLonde's goal was to determine which statistical methods, if any, were able to use the observational survey data to recover the results obtained from the randomized experiment.

We explore whether GenMatch is able to find matched data sets with good balance in the observed covariates. We compare the balance found by GenMatch to that of the propensity score models used in the literature to analyze these data sets.[16]

A. Data

The NSW was a job training program implemented in the mid-1970s to provide work experience for six to eighteen months to individuals facing economic and social disadvantages. Those randomly selected to join the program participated in various types of work. Information on preintervention variables (preintervention earnings, as well as education, age, ethnicity, and marital status) was obtained from initial surveys and Social Security Administration records. In the LaLonde data sample of NSW, baseline data are observed in 1975 and earlier, and the outcome of interest is real earnings in 1978.

There are eight observed possible confounders: age, years of education, real earnings in 1975, a series of variables indicating if the person has a high school diploma, is black, is married, or is Hispanic, and, for a subset of the data, real earnings in 1974.[17] The four dichotomous variables respectively indicate whether the individual is black, Hispanic, married, or a high school graduate.

We analyze three NSW data sets: the LaLonde, DW, and early randomization samples. Following DW, our LaLonde sample consists of only male participants in the original LaLonde analysis. This experimental sample is composed of 297 treated observations and 425 control observations.

Dehejia and Wahba (1999) created the DW sample from the LaLonde sample. They argued that it was necessary to control for more than one year of preintervention earnings in order to make the selection on observables assumption plausible because of Ashenfelter's dip (Ashenfelter, 1978). DW limited themselves to a particular subset of LaLonde's NSW data for which they claimed to either measure 1974 earnings or assumed zero earnings. They used individuals who were randomized before April 1976 and individuals who were randomized later but were known to be unemployed prior to randomization. The DW subset contains 185 treated and 260 control observations.

The early random assignment (early RA) sample was created by Smith and Todd (2005a). Like the DW sample, the early RA sample is a subset of the LaLonde sample for which two years of prior earnings are available. The sample excludes people in the LaLonde data who were randomized after April 1976. This sample was created because ST found the decision to include in the DW sample people randomized after April 1976 only if they had zero earnings in months 13 to 24 before randomization to be problematic. The early RA sample consists of 108 treated and 142 control observations.

LaLonde's nonexperimental estimates were based on two different observational control groups: the Panel Study of Income Dynamics (PSID-1) and Westat's Matched Current Population Survey–Social Security Administration file (CPS-1). Both differ substantially from the NSW experimental treatment group in terms of age, marital status, ethnicity, and preintervention earnings. All mean differences across treated and control groups are significantly different from 0 at conventional significance levels, except the indicator for Hispanic ethnic background.

To bridge the gap between treatment and comparison group preintervention characteristics, LaLonde extracted subsets from PSID-1 and CPS-1 (denoted PSID-2 and -3, and CPS-2 and -3) that he deemed similar to the treatment group in terms of particular covariates.[18] According to LaLonde, these smaller comparison groups were composed of individuals whose characteristics were similar to the eligibility criteria used to admit applicants into the NSW program. Even so, the subsets remain substantially different from the control group and from each other. GenMatch is applied to CPS-1 and PSID-1 because those offer the largest number of controls and because we wish to determine if the matching algorithm itself can find suitable matches, without the help of human preprocessing.

[16] We also used both the random forest and boosted CART algorithms to estimate propensity score models in these data sets, but neither algorithm produced matched data sets with better covariate balance than the best propensity score models proposed in the literature.

[17] The variable that DW call "real earnings in 1974" actually consists of real earnings in months 13 to 24 prior to the month of randomization. For some people, these months overlap with calendar year 1974. For people randomized late in the experiment, these months largely overlap with 1975. See Smith and Todd (2005a) for details.

[18] PSID-2 selects from PSID-1 all men not working when surveyed in 1976; PSID-3 selects from PSID-1 all men not working when surveyed in either 1975 or 1976; CPS-2 selects from CPS-1 all males not working in 1976; CPS-3 selects from CPS-1 all males unemployed in 1976 with 1975 income below the poverty level. CPS-1 has 15,992 observations, CPS-2 has 2,369 observations, and CPS-3 has 429 observations; PSID-1 has 2,490 observations, PSID-2 has 253 observations, and PSID-3 has 128 observations.

TABLE 3.—THE NSW RANDOMIZED EXPERIMENT AND NONEXPERIMENTAL SURVEY DATA

			ATT Estimates		
				95% Confidence Interval	
Data	Method	Balance Measure	Estimate	Lower Bound	Upper Bound
DW Subsample (Benchmark)	Experiment		\$1,794	\$512	\$3,146
CPS-1	GenMatch	fitness value = 0.21	\$1,734	−\$298	\$3,766
PSID-1	GenMatch	fitness value = 0.029	\$1,045	−\$2,354	\$4,454
PSID-2	P score matching	t-test p-value > 0.05	−\$487	−\$3,469	\$2,493
PSID-3	P score matching	t-test p-value > 0.05	−\$1,044	−\$4,688	\$2,600
CPS-2	P score matching	t-test p-value > 0.05	\$705	−\$1,553	\$2,962
CPS-3	P score matching	t-test p-value > 0.05	−\$295	−\$2,745	\$2,155
				95% Confidence Interval	
Data	Method	Balance Measure	Point Estimate	Lower Bound	Upper Bound
Early RA sample (Benchmark)	Experiment		\$2,748	\$764	\$4,733
CPS-1	GenMatch	fitness value = 0.46	\$1,631	−\$831	\$4,093
PSID-1	GenMatch	fitness value = .089	\$1,331	−\$2,007	\$4,670
				95% Confidence Interval	
Data	Method	Balance Measure	Point Estimate	Lower Bound	Upper Bound
Lalonde Sample (Benchmark)	Experiment		\$886	−\$54	\$1,864
CPS-1	GenMatch	fitness value = 0.23	\$281	−\$1,122	\$1,686
PSID-1	GenMatch	fitness value = 0.024	−\$571	−\$2,786	\$1,645
CPS-3	P score matching	t-test p-value > 0.05	−\$1,512	−\$3,748	\$724

Balance was evaluated via the default GenMatch fitness value: the lowest p-value obtained via paired t- and Kolmogorov-Smirnov tests. Propensity score results show different models that achieve balance by conventional standards when only differences of means are examined. The propensity score models achieve poor balance as measured by the KS test (p-values < 0.01) for some of the covariates.

The NSW data and the LaLonde (1986) research design present a difficult evaluation problem. As Smith and Todd (2005a), observed the data do not include a rich set of baseline covariates, the nonexperimental comparison groups are not drawn from the same local labor market as participants, and the dependent variable and the baseline earnings variables are measured differently for participants and nonparticipants. Moreover, the original NSW experiment had four target groups: ex-addicts, ex-convicts, high school dropouts, and long-term welfare recipients. Smith and Todd argue that it is implausible that conditioning on the eight observed variables suffices to make ex-addicts and ex-convicts look like (conditionally) random draws from the CPS or PSID. In addition, there is no single uniquely defined experimental target result, but rather several candidate target estimates, all of which have wide confidence intervals.[19] Much of the prior literature has estimated the experimental treatment effect as the simple difference in the means of outcomes across treatment and control groups, and we do the same. Taking simple differences results in an estimated average treatment effect of \$886 in the LaLonde sample, \$1,794 in the DW subsample, and \$2,748 in the early RA sample. The 95% confidence intervals of all three estimates cover \$900 (see table 3).

B. *Matching Results*

In table 3 we present GenMatch results for six of the observational data sets considered in the literature: the GenMatch estimate for each of the three NSW experimental samples paired with CPS-1 controls and PSID-1 controls. GenMatch was asked to maximize balance in all of the observed covariates, their first-order interactions, and quadratic terms. We also present, for comparison, the results for propensity score matching if, for the given observational data set, there is a propensity score in the literature that obtains balance as measured by difference of means for all of the baseline variables and their first-order interactions. All of the propensity scores reported, however, have baseline imbalances in at least one KS test of p-value < 0.001.

For the DW treatment subsample and the CPS-1 controls, GenMatch finds very good balance on the observables. The smallest observed p-value is 0.21 across both t and KS tests. And in this case, the GenMatch estimate of \$1,734 matches the experimental benchmark of \$1,794 well. However, when the DW treatment is matched to the PSID-1 controls, balance is poor (smallest p-value: 0.029), and the matched estimate is \$1,045. This is still closer to the experimental benchmark than the propensity score estimates we find in the DW sample that have good balance in difference of means.

In the early RA sample, GenMatch is again able to find good covariate balance with the CPS-1 controls: the smallest p-value is 0.46. However, the experimental estimate is \$2,748, while the GenMatch estimate is \$1,631, although the estimates are not significantly different because of the large confidence intervals. When the PSID-1 controls are used, GenMatch finds relatively poor balance: the smallest p-value across the matching set is 0.089. Consequently, the GenMatch estimate of \$1,331 is even further away from the experimental benchmark.

Recall that for both the early RA sample and the DW sample, two years of prior earnings are available. For the LaLonde

[19] One might propose other experimental target estimates produced by matching, regression adjustment, or difference-in-difference estimation. All produce qualitatively similar estimates.

sample, this is not the case, and one would expect the bias to be greatest in this data set. In the LaLonde sample and the CPS-1 controls, GenMatch finds good balance (smallest p-value is 0.23), but the GenMatch estimate is $281 while the experimental benchmark is $886. With the PSID-1 controls, the GenMatch balance is poor (smallest p-value is 0.024), and the GenMatch estimate of −$571 is outside the confidence intervals of the experimental benchmark.[20]

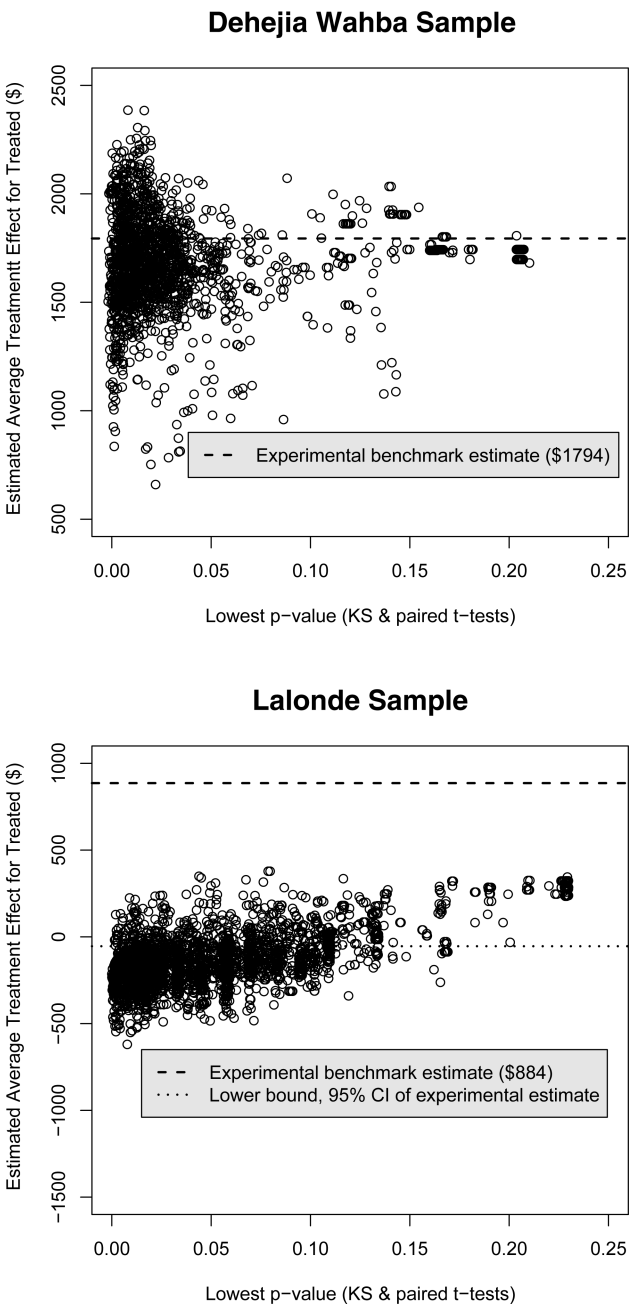
In all cases, GenMatch has better balance than the best propensity score estimates found, since all of the propensity score estimates have at least one KS test with a p-value of less than 0.01, although the p-values for the difference of means tests for all reported propensity score models are greater than 0.05. The GenMatch estimates are less variable than those of the various propensity score models.

Figure 5 shows how the distribution of GenMatch estimates varies with fitness. Fitness is measured as the lowest p-value obtained, after matching, from covariate-by-covariate paired t-tests and KS tests across all covariates, their first-order interactions, and quadratic terms. Each point represents one matched data set, its measure of balance, and its estimate of the causal effect. The universe of possible one-to-one matched data sets with replacement using the CPS-1 controls was sampled and plotted.[21] The figure plots the search space which GenMatch is searching, and GenMatch is able to find the best matched data set in this universe. The upper panel shows the DW sample, with estimates distributed above and below the target experimental result. The 64 best-balancing estimates at the maximum fitness value are all within $52 of the experimental difference in means. As the figure shows, in the DW sample, it is possible to get lucky and produce a reliable result even when balance has not been attained. The figure helps to explain why it is possible for DW to obtain accurate results with propensity scores models that do not achieve a high degree of balance and why it is possible for ST to find propensity scores with an equal degree of balance but estimates that are far from the experimental benchmark. Reliable results are obtained only at the highest fitness values.

The lower panel of figure 5 shows that in the LaLonde sample, all the GenMatch estimates are negatively biased, which may be expected given the omission of earnings in 1974. The figure for the early RA sample, not reported, looks similar to both of these figures. In the early RA sample, as seen in table 3, the bias is greater than in the DW sample but less than in the LaLonde sample.

In this literature, there are numerous propensity score models that achieve weak but conventionally accepted degrees of balance. DW follow a conventional approach to balance testing, checking balance across variables within blocks of a given propensity-score range. The DW papers do not provide detailed information on the degree of balance achieved on each variable. Instead, the authors plot the distributions of treated and control propensity scores and claim overlap. With our replication of Dehejia and Wahba (2002) and Dehejia (2005), it is clear that while their figures indicate overlap and their results satisfy conventional notions of balance, performing paired t-tests and Kolmogorov-Smirnov

FIGURE 5.—RELIABLE ESTIMATES REQUIRE HIGH DEGREE OF BALANCE



Both plots are based on the CPS-1 observational control sample. The top plot uses the Dehejia-Wahba experimental treatment sample, and the bottom plot uses the LaLonde experimental treatment sample.

[20] The GenMatch estimates are substantially similar if GenMatch is asked to optimize slightly different balance measures from the default—for example, if balance is measured as the mean standardized difference in the empirical-QQ plot for each variable.

[21] The figures were generated by Monte Carlos sampling: 1 million random values of W were generated and the unique matches that result from each unique weight matrix plotted.

tests across matched treated and control covariates yields significant p-values.[22]

For example, consider the case that should be most favorable to DW: the DW experimental sample, the control sample with the largest of the nonexperimental control groups (CPS-1), and the most recent propensity score specification from Dehejia (2005).[23] In this case, the dummy variable for High School Degree has a t-test p-value significant at conventional test levels, as does its interaction with Age, Education, and Black. We obtain Kolmogorov-Smirnov p-values less than 0.01 for all nondichotomous covariates: Age, Education, and two years of pretreatment income. Moreover, the ratios of covariate variances across control and treatment groups exceed 2 in several cases. By contrast, the lowest p-value GenMatch obtains in this case is 0.21.

V. Conclusion

The main advantage of GenMatch is that it directly optimizes covariate balance. This avoids the manual process of checking covariate balance in the matched samples and then respecifying the propensity score accordingly. Although there is little disagreement that this process should be followed in principle, it is rarely followed in practice. By using an automated process to search the data for the best matches, GenMatch is able to obtain better levels of balance without requiring the analyst to correctly specify the propensity score. There is little reason for a human to try the multitude of possible models to achieve balance when a computer can do this systematically and faster.

Historically, the matching literature, like much of statistics, has been limited by computational power. In recent years, computationally intensive simulation and machine learning methods have become popular. We think that matching is a case where computational power and machine learning algorithms may help. Our algorithm allows researchers to include their substantive knowledge of the data when choosing the covariates to match on, the measures of balance to use, and the propensity score model to include. It is also possible to start the algorithm with suggested weights, and indeed it is possible for researchers to bound the weights. From this substantive base, the algorithm will search and improve balance if possible given the data.

Open source software that implements GenMatch and a variety of other matching algorithms is available for the `R` programming environment (`R` Development Core Team, 2011). The package, called `Matching`, and is available on the Comprehensive R Archive Network.[24] Details of the software are described in Sekhon (2011). The software allows combining GenMatch with many of other matching methods, such as matching using calipers and matching exactly on some variables.

There are many outstanding questions and issues. There are other ways to generalize Mahalanobis distance, and these should be examined. Our proposed generalization works well in this example and in examples that other researchers have produced (see Sekhon, 2011, for a review). But there is no claim that it is generally the best, especially since it is unclear how to best measure the degree of covariate balance. It is also possible to use alternative optimization methods to search the space of possible solutions. Finally, the estimand has been held fixed in this study. It is possible to adapt the estimand, mostly by dropping observations, so as to maximize covariate balance (Crump et al., 2006).

A number of recent proposed methods use a weighting approach, as opposed to matching, and build in covariate balance. These include auxiliary-to-study tilting (Graham, Campos de Xavier Pinto, & Egel, 2011), which has the benefit of being doubly robust (Robins, Rotnitzky, & Zhao, 1994, 1995), and a proposal to adapt maximum entropy weighting, which has long been used in the survey data literature to match moments using auxiliary information, to the case of estimating treatment effects (Hainmueller, 2012). However, there are open questions about the fragility of such weighting estimators in finite samples when the estimated probabilities of treatment assignment are close to 0 or 1 (Freedman & Berk, 2008; Kang & Schafer, 2007; Porter et al., 2011).

The advantage of any new matching method is limited because of the selection on observables assumption. The plausibility of the assumption must be carefully scrutinized in each application using evidence beyond the statistical method. In observational studies, key identifying assumptions cannot be tested by simulations or proven mathematically. Therefore, more validation studies based on real data are needed to improve observational methods in practice and clarify the conditions in which these methods are appropriate.

[22] Smith and Todd (2005b) also note that the DW propensity score specifications fail some balancing tests other than the one DW rely on.

[23] We have replicated the earlier DW results across their models and data sets, and results are much the same. Their propensity score matching methods do not achieve a very high degree of balance across all the confounders, their interactions, and the quadratic terms.

[24] http://CRAN.R-project.org/package=Matching.

REFERENCES

Abadie, Alberto, "Bootstrap Tests for Distributional Treatment Effect in Instrumental Variable Models," *Journal of the American Statistics Association* 97:457 (2002), 284–292.

Abadie, Alberto, and Guido Imbens, "Large Sample Properties of Matching Estimators for Average Treatment Effects," *Econometrica* 74:1 (2006), 235–267.

Ashenfelter, Orley, "Estimating the Effects of Training Programs on Earnings," this REVIEW 60:1 (1978), 47–57.

Austin, Peter C., "A Critical Appraisal of Propensity Score Matching in the Medical Literature between 1996 and 2003," *Statistics in Medicine* 27:12 (2008), 2037–2049.

——— "Balance Diagnostics for Comparing the Distribution of Baseline Covariates between Treatment Groups in Propensity Score Matched Samples," *Statistics in Medicine* 28:25 (2009), 3083–3107.

Bhattacharya, Jay, and William B. Vogt, "Do Instrumental Variables Belong in Propensity Scores?" NBER technical working paper 343 (2007).

Breiman, Leo, "Random Forests," *Machine Learning* 45:1 (2001), 5–32.

Breiman, Leo, Jerome Friedman, Charles J. Stone, and R. A. Olshen, *Classification and Regression Trees* (New York: Chapman & Hall, 1984).

Bühlmann, Peter, and Bin Yu, "Boosting with the L_2-Loss: Regression and Classification," *Journal of the American Statistical Association* 98:462 (2003), 324–339.

Christakis, Nicholas A., and Theodore I. Iwashyna, "The Health Impact of Health Care on Families: A Matched Cohort Study of Hospice Use by Decedents and Mortality Outcomes in Surviving, Widowed Spouses," *Social Science and Medicine* 57:3 (2003), 465–475.
Cochran, William G., and Donald B. Rubin, "Controlling Bias in Observational Studies: A Review," *Sankhya, Series A* 35:4 (1973), 417–446.
Crump, Richard K., V. Joseph Hotz, Guido W. Imbens, and Oscar A. Mitnik, "Moving the Goalposts: Addressing Limited Overlap in Estimation of Average Treatment Effects by Changing the Estimand," NBER technical working paper 330 (2006).
Dawid, A. Phillip, "Conditional Independence in Statistical Theory," *Journal of the Royal Statistical Society, Series B* 41:1 (1979), 1–31.
Dehejia, Rajeev, "Practical Propensity Score Matching: A Reply to Smith and Todd," *Journal of Econometrics* 125:1–2 (2005), 355–364.
Dehejia, Rajeev, and Sadek Wahba, "Causal Effects in Non-Experimental Studies: Re-Evaluating the Evaluation of Training Programs," *Journal of the American Statistical Association* 94:448 (1999), 1053–1062.
——— "Propensity Score Matching Methods for Nonexperimental Causal Studies," this REVIEW 84:1 (2002), 151–161.
Diprete, Thomas A., and Henriette Engelhardt, "Estimating Causal Effects with Matching Methods in the Presence and Absence of Bias Cancellation," *Sociological Methods and Research* 32:4 (2004), 501–528.
Drake, Christiana, "Effects of Misspecification of the Propensity Score on Estimators of Treatment Effect," *Biometrics* 49:4 (1993), 1231–1236.
Epstein, Lee, Daniel E. Ho, Gary King, and Jeffrey A. Segal, "The Supreme Court during Crisis: How War Affects Only Non-War Cases," *New York University Law Review* 80:1 (2005), 1–116.
Freedman, D. A., and R. A. Berk, "Weighting Regressions by Propensity Scores," *Evaluation Review* 32:4 (2008), 392–409.
Friedman, Jerome H, "Greedy Function Approximation: A Gradient Boosting Machine," *Annals of Statistics* 29:5 (2001), 1189–1232.
Galiani, Sebastian, Paul Gertler, and Ernesto Schargrodsky, "Water for Life: The Impact of the Privatization of Water Services on Child Mortality," *Journal of Political Economy* 113:1 (2005), 83–120.
Gordon, Sandy, and Greg Huber, "The Effect of Electoral Competitiveness on Incumbent Behavior," *Quarterly Journal of Political Science* 2:2 (2007), 107–138.
Graham, Bryan, Cristine Campos de Xavier Pinto, and Daniel Egel, "Efficient Estimation of Data Combination Models by the Method of Auxiliary-to-Study Tilting (AST)," NBER working paper 16928 (2011).
Hahn, Jinyong, "On the Role of the Propensity Score in Efficient Estimation of Average Treatment Effects," *Econometrica* 66:2 (1998), 315–331.
Hainmueller, Jens, "Entropy Balancing: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies," *Political Analysis* 20:1 (2012), 25–46.
Hansen, Ben B., "Full Matching in an Observational Study of Coaching for the SAT," *Journal of the American Statistical Association* 99:467 (2004), 609–618.
Hansen, Ben B., and S. O. Klopfer, "Optimal Full Matching and Related Designs via Network Flows," *Journal of Computational and Graphical Statistics* 15:3 (2006), 609–627.
Hastie, Trevor, Robert Tibshirani, and Jerome Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed: (New York: Springer, 2009).
Heckman, James J., Hidehiko Ichimura, Jeffrey Smith, and Petra Todd, "Characterizing Selection Bias Using Experimental Data," *Econometrica* 66:5 (1998), 1017–1098.
Herron, Michael C., and Jonathan Wand, "Assessing Partisan Bias in Voting Technology: The Case of the 2004 New Hampshire Recount," *Electoral Studies* 26:2 (2007), 247–261.
Ho, Daniel E., Kosuke Imai, Gary King, and Elizabeth A. Stuart, "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference," *Political Analysis* 15:3 (2007), 199–236.
Imai, Kosuke, "Do Get-Out-the-Vote Calls Reduce Turnout? The Importance of Statistical Methods for Field Experiments," *American Political Science Review* 99:2 (2005), 283–300.
Imai, Kosuke, Gary King, and Elizabeth A. Stuart, "Misunderstandings among Experimentalists and Observationalists about Causal Inference," *Journal of the Royal Statistical Society, Series A* 171:2 (2008), 481–502.
Kang, Joseph D. Y., and Joseph L. Schafer, "Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data (with discussion)," *Statistical Science* 22:4 (2007), 523–539.
LaLonde, Robert, "Evaluating the Econometric Evaluations of Training Programs with Experimental Data," *American Economic Review* 76:4 (1986), 604–620.
Lee, Brian, Justin Lessler, and Elizabeth A. Stuart, "Improving Propensity Score Weighting Using Machine Learning," *Statistics in Medicine* 29:3 (2010), 337–346.
Lehrer, Steven F., and Gregory Kordas, "Matching Using Semiparametric Propensity Scores," *Empirical Economics* 44:1 (2013), 13–45.
Liaw, A., and M. Wiener, "Classification and Regression by Random Forest," *R News* 2:3 (2002), 18–22.
McCaffrey, D. F., G. Ridgeway, and A. R. Morral, "Propensity Score Estimation with Boosted Regression for Evaluating Causal Effects in Observational Studies," *Psychological Methods* 9:4 (2004), 403–425.
Mebane, Walter R. Jr., and Jasjeet S. Sekhon, "GENetic Optimization Using Derivatives (GENOUD)," software (1998), http://sekhon.berkeley.edu/rgenoud/.
Mebane, Walter R. Jr., and Jasjeet S. Sekhon, "Genetic Optimization Using Derivatives: The `rgenoud` package for `R`," *Journal of Statistical Software* 42:11 (2011), 1–26.
Morgan, Stephen L., and David J. Harding, "Matching Estimators of Causal Effects: Prospects and Pitfalls in Theory and Practice," *Sociological Methods and Research* 35:1 (2006), 3–60.
Pearl, Judea, "On a Class of Bias-Amplifying Variables That Endanger Effect Estimates," in *Proceedings of the Twenty-Sixth Conference in Artificial Intelligence* (2012), http://adsabs.harvard.edu/egi-bin/bib_queryXiv:1203.3503.
Porter, Kristin E., Susan Gruber, Mark J. van der Laan, and Jasjeet S. Sekhon, "The Relative Performance of Targeted Maximum Likelihood Estimators," *International Journal of Biostatistics* 7:1 (2011), 1–34.
`R` Development Core Team, *`R`: A Language and Environment for Statistical Computing* (Vienna: R Foundation for Statistical Computing, 2011).
Ridgeway, G., D. F. McCaffrey, and A. R. Morral, "Twang: Toolkit for Weighting and Analysis of Nonequivalent Groups," R Package Version 1.0–2 (2010).
Robins, J. M., A. Rotnitzky, and L. P. Zhao, "Estimation of Regression Coefficients When Some Regressors Are Not Always Observed," *Journal of the American Statistical Association* 89:427 (1994), 846–866.
——— "Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data," *Journal of the American Statistical Association* 90:429 (1995), 106–121.
Rosenbaum, Paul R, "A Characterization of Optimal Designs for Observational Studies," *Journal of the Royal Statistical Society, Series B* 53:3 (1991), 597–610.
——— *Observational Studies*, 2nd ed. (New York: Springer-Verlag, 2002).
Rosenbaum, Paul R., and Donald B. Rubin, "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika* 70:1 (1983), 41–55.
——— "Reducing Bias in Observational Studies Using Subclassification on the Propensity Score," *Journal of the American Statistical Association* 79:387 (1984), 516–524.
——— "Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score," *American Statistician* 39:1 (1985), 33–38.
Rubin, Donald B., "Estimating Causal Effects from Large Data Sets Using Propensity Scores," *Annals of Internal Medicine* 127:8S (1997), 757–763.
——— "Using Propensity Scores to Help Design Observational Studies: Application to the Tobacco Litigation," *Health Services and Outcomes Research Methodology* 2:1 (2001), 169–188.
——— *Matched Sampling for Causal Effects* (Cambridge: Cambridge University Press, 2006).
——— "For Objective Causal Inference, Design Trumps Analysis," *Annals of Applied Statistics* 2:3 (2008), 808–840.

Rubin, Donald B., and Neal Thomas, "Affinely Invariant Matching Methods with Ellipsoidal Distributions," *Annals of Statistics* 20:2 (1992), 1079–1093.
Schapire, Rob, "The Strength of Weak Learnability," *Machine Learning* 5:2 (1990), 197–227.
Sekhon, Jasjeet S, "Quality Meets Quantity: Case Studies, Conditional Probability and Counterfactuals," *Perspectives on Politics* 2:2 (2004), 281–293.
———"Matching: Multivariate and Propensity Score Matching with Automated Balance Search," *Journal of Statistical Software* 42:7 (2011), 1–52. Computer program, http://sekhon.berkeley.edu/matching/.
Sekhon, Jasjeet Singh, and Walter R. Mebane Jr., "Genetic Optimization Using Derivatives: Theory and Application to Nonlinear Models," *Political Analysis* 7 (1998), 189–203.
Setoguchi, Soko, Sebastian Schneeweiss, M. Alan Brookhart, Robert J. Glynn, and E. Francis Cook, "Evaluating Uses of Data Mining Techniques in Propensity Score Estimation: A Simulation Study," *Pharmacoepidemiology and Drug Safety* 17:6 (2008), 546–555.
Smith, Jeffrey A., and Petra E. Todd, "Reconciling Conflicting Evidence on the Performance of Propensity Score Matching Methods," *AEA Papers and Proceedings* 91:2 (2001), 112–118.
——— "Does Matching Overcome LaLonde's Critique of Nonexperimental Estimators?" *Journal of Econometrics* 125:1–2 (2005a), 305–353.
——— "Rejoinder," *Journal of Econometrics* 125:1–2 (2005b), 365–375.
Winship, Christopher, and Stephen Morgan, "The Estimation of Causal Effects from Observational Data," *Annual Review of Sociology* 25 (1999), 659–707.
Wooldridge, J., "Should Instrumental Variables Be Used as Matching Variables?" Technical report, Michigan State University (2009).

APPENDIX

Data Generation Model Formulas for Simulation 1

This simulation study is from Lee et al. (2010), which is the same as Setoguchi et al. (2008) except that a continuous outcome is substituted for the binary outcome used in the original study.

All of the true propensity score models are of the form $\Pr[T = 1 \mid X_i] = \frac{1}{(1+\exp\{-\mu\})}$. The linear predictor, μ, varies across the seven experimental conditions as follows.

Scenario A (a model with additivity and linearity):

$$\mu = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7$$

Scenario B (a model with mild nonlinearity):

$$\mu = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_2 X_2^2$$

Scenario C (a model with moderate nonlinearity):

$$\mu = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_2 X_2^2 + \beta_4 X_4^2 + \beta_7 X_7^2$$

Scenario D (a model with mild nonadditivity):

$$\mu = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_1 0.5 X_1 X_3 + \beta_2 0.7 X_2 X_4 + \beta_4 0.5 X_4 X_5 + \beta_5 0.5 X_5 X_6$$

Scenario E (a model with mild nonadditivity and nonlinearity):

$$\mu = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_2 X_2^2 + \beta_1 0.5 X_1 X_3 + \beta_2 0.7 X_2 X_4 + \beta_4 0.5 X_4 X_5 + \beta_5 X_5 X_6$$

Scenario F (a model with moderate nonadditivity):

$$\mu = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_1 0.5 X_1 X_3 + \beta_2 0.7 X_2 X_4 + \beta_3 0.5 X_3 X_5 + \beta_4 0.7 X_4 X_6 + \beta_5 0.5 X_5 X_7 + \beta_1 0.5 X_1 X_6 + \beta_2 0.7 X_2 X_3 + \beta_3 0.5 X_3 X_4 + \beta_4 0.5 X_4 X_5 + \beta_5 0.5 X_5 X_6$$

Scenario G (a model with moderate nonadditivity and nonlinearity):

$$\mu = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_2 X_2^2 + \beta_4 X_4^4 + \beta_7 X_7^2 + \beta_1 0.5 X_1 X_3 + \beta_2 0.7 X_2 X_4 + \beta_3 0.5 X_3 X_5 + \beta_4 0.7 X_4 X_6 + \beta_5 0.5 X_5 X_7 + \beta_1 0.5 X_1 X_6 + \beta_2 0.7 X_2 X_3 + \beta_3 0.5 X_3 X_4 + \beta_4 0.5 X_4 X_5 + \beta_5 0.5 X_5 X_6$$

The coefficients are $\beta_0 = 0, \beta_1 = 0.8, \beta_2 = -0.25, \beta_3 = 0.6, \beta_4 = -0.4, \beta_5 = -0.8, \beta_6 = -0.5$, and $\beta_7 = 0.7$.

The outcome model is:

$$Y = \gamma\, T - 3.85 + 0.3 X_1 + -0.36 X_2 - 0.73 X_3 - 0.2 X_4 + 0.71 X_8 - 0.19 X_9 + 0.26 X_{10}$$

where $\gamma = -0.4$ is the treatment effect.