The Stata Journal (2007)
**7**, Number 3, pp. 334–350

# Simulation-based sensitivity analysis for matching estimators

Tommaso Nannicini
Universidad Carlos III de Madrid
Madrid, Spain
tommaso.nannicini@uc3m.es

**Abstract.** This article presents a Stata program (`sensatt`) that implements the sensitivity analysis for matching estimators proposed by Ichino, Mealli, and Nannicini (*Journal of Applied Econometrics*, forthcoming). The analysis simulates a potential confounder to assess the robustness of the estimated treatment effects with respect to deviations from the conditional independence assumption. The program uses the commands for propensity-score matching (`att`*) developed by Becker and Ichino (*Stata Journal* 2: 358–377). I give an example by using the National Supported Work demonstration, widely known in the program evaluation literature.

**Keywords:** st0130, sensatt, sensitivity analysis, matching, propensity score, program evaluation

## 1 Introduction

During the last few years, using matching estimators in evaluation studies of treatment effects has skyrocketed. In particular, two factors have favored the diffusion of these methods in empirical works. First, the findings by Dehejia and Wahba (1999, 2002) about the promising performance of propensity-score matching estimators in observational studies have triggered the attention of theoretical and empirical researchers to these techniques.[1] Even though Dehejia and Wahba make it clear that these estimators do not represent a "magic bullet" and the later literature shows that they are effective only in data contexts satisfying particular conditions, their use is now widespread in applied studies. Second, many free user-friendly software routines have been made available to apply matching estimators. In Stata, Becker and Ichino (2002) provide a suite of commands (`attnd`, `attnw`, `atts`, `attr`, and `attk`) that carry out different propensity-score matching estimators of the average treatment effect on the treated (ATT); Leuven and Sianesi (2003) develop a program (`psmatch2`) that imple-

---

[1] Using data from the influential study by LaLonde (1986), Dehejia and Wahba (1999) show that propensity-score matching estimates are closer to the experimental benchmark than the ones produced by traditional evaluation methods. This apparent "propensity score paradox" (i.e., that these estimators seem to perform better with respect to alternative nonexperimental methods that rely on the same identification assumptions) have contributed to the recent popularity of matching in empirical studies, even though Smith and Todd (2005) have subsequently shown that matching estimators work well only for a specific subsample of the LaLonde data, casting doubts on the generalizability of the results by Dehejia and Wahba. See also, among others, Heckman, Ichimura, and Todd (1997, 1998); Imbens (2004); and Michalopoulos, Bloom, and Hill (2004).

ments full Mahalanobis matching and a variety of propensity-score matching methods; Abadie et al. (2004) develop a command (`nnmatch`) that implements nearest-neighbor matching estimators for average treatment effects.

As a combined result of the above two factors, matching estimators are now widely known and easy to use. And, perhaps, too many users adopt them without carefully discussing whether the conditions for their application are met or how robust the derived estimates are with respect to possible deviations from these conditions. In particular, matching relies on the assumption of conditional independence of potential outcomes and treatment assignment given observables, i.e., on the fact that selection into treatment is driven only by factors that the researcher can observe. This is the so-called conditional independence assumption (CIA), also known as "unconfoundedness" or "selection on observables" in the program evaluation literature.[2] Moreover, Heckman, Ichimura, and Todd (1997) show that, for matching estimators to reduce bias as conventionally measured, it is crucial that 1) the same questionnaire be used for both the treated and control units and 2) the nonexperimental comparison group be drawn from the same local labor market with respect to the treated.[3] In data contexts where the CIA appears plausible and the above conditions are met, matching may be a better strategy to control for observables than regression modeling (if there is no credible source of exogenous variation), since it does not rely on linearity and allows checking whether there is a substantial overlap of the distributions of covariates in the treatment and comparison groups. However, every evaluation strategy that uses matching estimators should contain some (possibly all) of the following steps:

1. To use data where the treated and control units come from the same local market and are asked the same set of questions
2. To discuss (carefully) why the CIA should be verified in the specific context of the evaluation question at hand
3. To test (indirectly) whether the available empirical evidence casts doubt on the plausibility of the CIA[4]
4. To inspect how the observations are distributed across the propensity-score common support and how sensitive the estimates are with respect to using observations in the tails of the common support[5]
5. To assess whether (and to what extent) the estimated average treatment effects are robust to possible deviations from the CIA (e.g., implementing some type of sensitivity analysis)

---

[2] See Imbens (2004) for a review of nonparametric estimation methods under this assumption.

[3] The experimental evidence by Michalopoulos, Bloom, and Hill (2004) reinforces this second point by showing that in-state comparison groups produce less bias than out-of-state groups.

[4] For instance, multiple control groups could be used in this respect (Rosenbaum 1987a).

[5] See Black and Smith (2004) for an excellent example.

The sensitivity analysis proposed by Ichino, Mealli, and Nannicini (2007) allows applied researchers who use matching estimators to tackle the fifth step.[6] The analysis builds on Rosenbaum and Rubin (1983a) and Rosenbaum (1987b) and is based on a simple idea. Suppose that the CIA is not satisfied given observables but would be satisfied if one could observe an additional binary variable. This potential confounder can be simulated in the data and used as an additional covariate in combination with the preferred matching estimator. The comparison of the estimates obtained with and without matching on the simulated confounder shows to what extent the baseline results are robust to specific sources of failure of the CIA, since the distribution of the simulated variable can be constructed to capture different hypotheses on the nature of potential confounding factors. In this article, I summarize this econometric tool and present a program (`sensatt`) that implements it in Stata.

## 2 Propensity-score matching

Consider Rubin's (1974) potential-outcome framework for causal inference, where $Y_1$ represents the outcome if the unit is exposed to treatment $T = 1$, and $Y_0$ is the outcome if the unit is exposed to treatment $T = 0$. Assume also that the average treatment effect of interest is the ATT, defined as follows:

$$E(Y_1 - Y_0|T = 1)$$

Here one possible estimation strategy is to assume that, given a set of observable covariates $W$, the potential outcome for no treatment is independent of treatment assignment:[7]

$$Y_0 \perp\!\!\!\perp T \,|\, W \tag{1}$$

This condition is the CIA. The behavioral assumption behind it is that the potential outcome for no treatment ($Y_0$) does not influence treatment assignment, whereas the possibility that the selection process depends on the treated outcome ($Y_1$) does not have to be ruled out. Although strong, the plausibility of this assumption relies heavily on the quality and amount of information contained in $W$. The CIA is an untestable assumption, since the data are completely uninformative about the distribution of $Y_0$ for treated subjects, but its credibility can be supported or rejected by theoretical reasoning and more evidence.[8] Besides the CIA, a further requirement for identification is the common support or overlap condition, which ensures that for each treated unit there are control units with the same observables:[9]

$$\Pr(T = 1|W) < 1 \tag{2}$$

---

[6] See Becker and Caliendo (2007) for implementing a different sensitivity analysis in Stata.

[7] If the effect of interest were the average treatment effect for the whole population, both potential outcomes should be conditionally independent of treatment assignment: $(Y_1, Y_0) \perp\!\!\!\perp T \,|\, W$.

[8] See Imbens (2004) and Rosenbaum (1987a).

[9] To estimate the average treatment effect, the overlap condition would require that $0 < \Pr(T = 1|W) < 1$.

Under assumptions (1) and (2), within each cell defined by $W$, treatment assignment is random, and the outcome of control subjects can be used to estimate the counterfactual outcome of the treated for no treatment. However, with a high-dimensional vector $W$, this task may be problematic. To deal with the dimensionality problem, one can use the results by Rosenbaum and Rubin (1983b) on the so-called propensity score. The propensity score is the individual probability of receiving the treatment given the observed covariates: $p(W) = P(T = 1|W)$. If the potential outcome $Y_0$ is independent of treatment assignment conditional on $W$, it is also independent of treatment assignment conditional on $p(W)$. The propensity score can thus be used as a univariate summary of all observable variables. As a consequence, if $p(W)$ is known, the ATT can be consistently estimated as follows:

$$\tau_{\text{ATT}} \equiv E(Y_1 - Y_0|T = 1) = \\ E_{\{p(W)|T=1\}}\left[E\{Y_1|p(W), T = 1\} - E\{Y_0|p(W), T = 0\}\right] \quad (3)$$

In practice, $p(W)$ is usually unknown and has to be estimated through some probabilistic model (e.g., probit or logit). Such a model should include all the pretreatment observable variables that influence both the selection into treatment and the outcome. Higher-order or interaction terms should be included in the specification of the model only if they served to make the estimated propensity score satisfy the balancing property, i.e., to have that within each cell of the propensity score the treated and control units have the same distribution of observable covariates.[10] However, the estimation of the propensity score is not enough to estimate the ATT with (3), since the probability of finding 2 observations with the same value of the score is extremely low. Various methods have been proposed in the literature to overcome this problem and match treated and control units on the basis of the estimated propensity score. The program `sensatt` uses three different algorithms: nearest neighbor, kernel, and radius.[11] These methods differ in how they select the control units that are matched to the treated and with respect to the weights they attribute to the selected controls when estimating the counterfactual outcome of the treated: $E\{Y_0|p(W), T = 1\}$. However, they all provide consistent estimates of the ATT under the CIA and the overlap condition.

## 3 Sensitivity analysis

This section borrows from Ichino, Mealli, and Nannicini (2007) and sketches the sensitivity analysis for propensity-score matching estimators that they propose. One of the central assumptions of the analysis is that treatment assignment is not unconfounded given the set of covariates $W$, i.e., that assumption (1) no longer holds. Also, it is assumed that the CIA holds given $W$ and an unobserved binary variable $U$:

---

[10] Usually, the balancing property is tested with reference to first moments.

[11] See Becker and Ichino (2002) for a description of these matching algorithms and the commands that implement them in Stata. See also Caliendo and Kopeinig (2007) for a discussion of the different properties of these and other propensity-score matching algorithms.

$$Y_0 \perp\!\!\!\perp T \mid (W, U)^{12} \tag{4}$$

As long as $U$ is not observed, the outcome of the controls cannot be credibly used to estimate the counterfactual outcome of the treated:

$$E(Y_0|T = 1, W) \neq E(Y_0|T = 0, W)$$

On the contrary, knowing $U$ (together with the observable covariates $W$) would be enough to consistently estimate the ATT as discussed in section 2, since

$$E(Y_0|T = 1, W, U) = E(Y_0|T = 0, W, U)$$

Assumption (4) is common to similar sensitivity analysis proposed in the econometric and statistical literature,[13] but the analysis discussed in this article is the only one that assesses the robustness of point estimates without relying on any parametric model for the outcome equation.

The next step characterizes the distribution of $U$ to simulate this potential confounder in the data. $U$ is assumed to be binary. It is also assumed to be independently and identically distributed in the cells represented by the Cartesian product of the treatment and outcome values. For simplicity, consider the case of binary potential outcomes: $Y_0, Y_1 \in \{0, 1\}$.[14] Also denote with $Y = T \cdot Y_1 + (1 - T) \cdot Y_0$ the observed outcome for a given unit, which is equal to one of the two potential outcomes depending on treatment assignment. The distribution of the binary confounding factor $U$ is fully characterized by the choice of four parameters,

$$p_{ij} \equiv \Pr(U = 1|T = i, Y = j) = \Pr(U = 1|T = i, Y = j, W)$$

with $i, j \in \{0, 1\}$, which give the probability that $U = 1$ in each of the four groups defined by the treatment status and the outcome value.[15] To make the simulation of the potential confounder feasible, we assume binary $U$ and conditional independence of $U$ with respect to $W$. Ichino, Mealli, and Nannicini (2007) present two Monte Carlo exercises showing that these simulation assumptions do not critically affect the results of the sensitivity analysis.

As a final step, given arbitrary (but meaningful) values of the parameters $p_{ij}$, a value of $U$ is attributed to each unit, according to its belonging to one of the four groups defined by the treatment status and the outcome value. The simulated $U$ is then treated as any other observed covariate and is included in the set of matching variables used to estimate the propensity score and to compute the ATT according to the chosen matching

---

[12] Using Rosenbaum's (1987a) terminology, we are moving from $(Y_0|W)$-adjustable treatment assignment in condition 1 to $(Y_0|W, U)$-adjustable treatment assignment in condition 4.

[13] See Rosenbaum and Rubin (1983a); Rosenbaum (1987b, 2002); Gastwirth, Krieger, and Rosenbaum (1998); Imbens (2003); and Altonji, Elder, and Taber (2005).

[14] This assumption will be removed at the end of the section.

[15] Using the parameters $p_{ij}$ and the probabilities of having a positive outcome by treatment status, $\Pr(Y = i|T = j)$, which are observed in the data, one can compute the fraction of subjects with $U = 1$ by treatment status only: $p_{i.} \equiv \Pr(U = 1|T = i) = \sum_{j=0}^{1} p_{ij} \cdot \Pr(Y = j|T = i)$, with $i \in \{0, 1\}$.

estimator (e.g., kernel). By using a given set of values of the sensitivity parameters, the matching estimation is repeated many times (e.g., 1,000) and a simulated estimate of the ATT is retrieved as an average of the ATTs over the distribution of $U$. Thus, for any given configuration of the parameters $p_{ij}$, the sensitivity analysis retrieves a point estimate of the ATT that is robust to the failure of the CIA implied by that particular configuration.[16]

## 3.1 Standard errors

For computing a standard error for the simulated ATT, the imputation of $U$ is considered a normal problem of missing data, which can be solved by repeatedly imputing the missing values of $U$. Let $m$ be the number of imputations of the missing $U$, and let $\widehat{\text{ATT}}_k$ and $\text{se}_k^2$ be the point estimate and the estimated variance of the ATT estimator at the $k$th imputed dataset (with $k = 1, 2, \dots, m$). The simulated ATT, $\widehat{\text{ATT}}$, is obtained by the average of the $\widehat{\text{ATT}}_k$ over the $m$ replications. In this setting, the within-imputation variance is equal to

$$\text{se}_W^2 = \frac{1}{m} \sum_{k=1}^{m} \text{se}_k^2$$

whereas the between-imputation variance is given by

$$\text{se}_B^2 = \frac{1}{m-1} \sum_{k=1}^{m} (\widehat{\text{ATT}}_k - \widehat{\text{ATT}})^2$$

As a consequence, the total variance associated to $\widehat{\text{ATT}}$ can be expressed as follows:

$$\text{se}_T^2 = \text{se}_W^2 + (1 + \frac{1}{m})\text{se}_B^2 \tag{5}$$

For many replications, the statistic $(\widehat{\text{ATT}} - \text{ATT})/\text{se}_T$ is approximately normal. One could also consider either the within-imputation or the between-imputation standard error as the basis for inference. The program `sensatt` allows using all three types. The standard error in (5) leads to conservative inferential conclusions, since it is always greater than the other two alternatives.

## 3.2 Extension to continuous outcomes

The above sensitivity analysis can be easily extended to multivalued or continuous outcomes. Indeed, in such cases, one can define the simulation parameters $p_{ij}$ on the basis of $T$ and a binary transformation of $Y$ (instead of the outcome itself). Define

$$p_{ij} \equiv \Pr(U = 1 | T = i, I(Y > y^*) = j)$$

---

[16] A high sensitivity of the baseline results could also come from a problem of weak identification, should even small variations in the conditioning set cause large changes in the estimates. However, this would reinforce the usefulness of the sensitivity analysis, since the harmful effects of a potential confounder are even more severe in the presence of weak identification.

with $i, j \in \{0, 1\}$, where $I$ is the indicator function and $y^*$ is a chosen typical value of the distribution of $Y$.[17] Once the parameters $p_{ij}$ are set thus, one can implement the sensitivity analysis as described above. Of course, the ATT is still estimated for the multivalued or continuous outcome $Y$.

## 4 Guidelines for implementing the simulations

To implement the sensitivity analysis described in section 3, one must have in mind which kind of potential confounding factors would be useful to simulate in the data. One must answer the following question: which values of the parameters $p_{ij}$ should I choose to learn something useful from the effect of a confounder $U$ like the one associated with the chosen values? I focus on two simulation exercises here. In the first one, the $p_{ij}$ are set so as to let $U$ mimic the behavior of some important covariates. In the second one, a grid of different $p_{ij}$ is built to capture the characteristics of those potential confounders that would drive the ATT estimates to zero or far away from the baseline result. However, the above sensitivity analysis is a flexible tool and its application is not restricted to the exercises suggested here.

Before I discuss these two sensitivity exercises, it is important to understand which kind of potential confounders would represent a real threat for the baseline estimates. Since the treatment is binary, we can assume without loss of generality that the ATT estimated according to the matching strategy outlined in section 2 is positive and significant. In a similar situation, before interpreting the baseline estimate as evidence of a true causal effect of the treatment, we may want to investigate how sensitive this estimate is with respect to the possible existence of an unobservable variable $U$ that affects both the potential outcome $Y_0$ and the selection into treatment $T$ (after controlling for observable covariates $W$). $U$ would be a dangerous confounder (i.e., a confounder whose existence might give rise to a positive and significant ATT estimate even without a true causal effect) if we observed that

$$\Pr(Y_0 = 1|T, W, U) \neq \Pr(Y_0 = 1|T, W) \tag{6}$$

$$\Pr(T = 1|W, U) \neq \Pr(T = 1|W) \tag{7}$$

Expressions (6) and (7)—unlike the parameters $p_{ij}$—both include $W$ and refer to the potential (not observed) outcome for no treatment. Hence, one may be worried that, by simply choosing the parameters $p_{ij}$, it is not possible to simulate a dangerous confounder like the one captured by these expressions. However, Ichino, Mealli, and Nannicini (2007) demonstrate that the following implications hold:

$$p_{01} > p_{00} \Rightarrow \Pr(Y_0 = 1|T = 0, U = 1, W) > \Pr(Y_0 = 1|T = 0, U = 0, W)$$

$$p_{1.} > p_{0.} \Rightarrow \Pr(T = 1|U = 1, W) > \Pr(T = 1|U = 0, W)$$

---

[17] The program `sensatt` allows using four $y^*$: mean, median, 25th, or 75th centile.

As a consequence, by simply assuming that $p_{01} > p_{00}$, one can simulate a confounding factor that has a positive effect on the untreated outcome $Y_0$ (conditioning on $W$). Similarly, by setting $p_{1.} > p_{0.}$,[18] one can simulate a confounding factor that has a positive effect on treatment assignment (conditioning on $W$).

I must address one limitation, however. Following the above reasoning, it would be tempting to interpret the difference $d = p_{01} - p_{00}$ as a measure of the effect of $U$ on the untreated outcome, and the difference $s = p_{1.} - p_{0.}$ as a measure of the effect of $U$ on the selection into treatment. But these two effects should be evaluated after conditioning on $W$, whereas $d$ and $s$ do not account for the association between $U$ and $W$ that shows up in the data. By setting the sensitivity parameters $p_{ij}$, we can control the *sign* but not the *magnitude* of the conditional association of $U$ with $Y_0$ and $T$. To sidestep this shortcoming, we can measure how each chosen configuration of the $p_{ij}$ translates in terms of the effect of $U$ on $Y_0$ and $T$ (conditioning on $W$). The program `sensatt` performs this task in the following way. At every iteration, a logit model of $\Pr(Y = 1|T = 0, U, W)$ is estimated and the average odds ratio of $U$ is reported as the outcome effect of the simulated confounder:[19]

$$\Gamma \equiv \frac{\frac{\Pr(Y=1|T=0,U=1,W)}{\Pr(Y=0|T=0,U=1,W)}}{\frac{\Pr(Y=1|T=0,U=0,W)}{\Pr(Y=0|T=0,U=0,W)}}$$

Similarly, the logit model of $\Pr(T = 1|U, W)$ is estimated at every iteration and the average odds ratio of $U$ is reported as the selection effect of the simulated confounder:

$$\Lambda \equiv \frac{\frac{\Pr(T=1|U=1,W)}{\Pr(T=0|U=1,W)}}{\frac{\Pr(T=1|U=0,W)}{\Pr(T=0|U=0,W)}}$$

By simulating $U$ under the assumptions that $d > 0$ and $s > 0$, we know from the above arguments that both the outcome and selection effects must be positive (i.e., $\Gamma > 1$ and $\Lambda > 1$). Moreover, by displaying the associated $\Gamma$ and $\Lambda$ as an additional output of the sensitivity analysis, we can easily assess the magnitude of these two effects, which end up characterizing the simulated confounder $U$.

## 4.1 A first simulation exercise: calibrated confounders

Keeping in mind the above reasoning, one can pick the parameters $p_{ij}$ (which in turn determine the parameters $p_{i.}$) to make the distribution of $U$ similar to the empirical distribution of important binary covariates (or binary transformations of continuous covariates). Here the simulation exercise reveals the extent to which the baseline estimates are robust to deviations from the CIA induced by the impossibility of observing factors similar to the ones used to calibrate the distribution of $U$. This is a different exercise from the simple removal of an observed variable from the matching set $W$, since

---

[18] After the choice of $p_{01}$ and $p_{00}$, this condition can be imposed by setting $p_{11}$ and $p_{10}$ appropriately.

[19] With continuous outcomes, $\Gamma$ is the odds ratio of $U$ in the logit model $\Pr(Y > y * |T = 0, U, W)$.

in every sensitivity-analysis estimation we still control for all the relevant covariates observed by the econometrician. Of course, this exercise is interesting only when the chosen covariates display $p_{ij}$ that satisfy the conditions $d > 0$ and $s > 0$.

### 4.2 A second simulation exercise: killer confounders

Since the results of the previous exercise may be driven by the particular behavior of the chosen covariates, another simulation exercise is even more instructive. One can search for the existence of a set of parameters $p_{ij}$ such that if $U$ were observed the estimated ATT would be driven to zero, and one can then assess the plausibility of this particular configuration of parameters. If all the configurations leading to such a result could be considered unlikely, the exercise would support the robustness of the estimates derived under the CIA. To reduce the dimensionality problem of the characterization of these killer confounding factors, one could fix at some predetermined values the probability $\Pr(U = 1)$ and the difference $d' = p_{11} - p_{10}$. Since these quantities are not expected to represent a real threat for the baseline estimate, they can be held fixed and the simulated confounder $U$ can be fully described by the differences $d$ and $s$.[20] For instance, one could build a table of simulated ATTs such that $d$ increases by 0.1 along each column, and $s$ increases by 0.1 along each column, looking for those configurations of these two parameters that drive the ATT to zero or far away from the baseline estimate ($d = 0$, $s = 0$).[21] Moreover, when displaying the results of the sensitivity analysis, the values of $d$ and $s$ should be associated with the estimated values of $\Gamma$ and $\Lambda$, respectively. The estimated odds ratios would thus provide a measure of the observed effects of $U$ on the untreated outcome and the selection into treatment, allowing the researcher to discuss the plausibility of the existence of a similar confounder. If only implausible confounders drove the ATT to zero or far away from the baseline estimate, the sensitivity analysis would support the robustness of matching results.[22]

## 5 Syntax

```
sensatt outcome treatment [varlist] [if] [in] [weight] [, alg(att*)
   reps(#) p(varname) p11(#) p10(#) p01(#) p00(#) se(se_type)
   ycent(#) pscore(scorevar) logit index comsup bootstrap]
```

The following remarks should be taken into account:

- The program uses the commands for the propensity-score matching estimation of average treatment effects written by Becker and Ichino (2002): `attnd`, `attnw`,

---

[20] Keeping $\Pr(U = 1)$ and $d'$ fixed, and substituting $\Pr(Y = i|T = j)$ and $\Pr(T = j)$ by their sample analogues, the parameters $d$ and $s$ are enough to characterize the distribution of $U$.

[21] At http://www.tommasonannicini.eu, there is an ancillary Matlab code that, once $d'$ and $\Pr(U = 1)$ are specified and the estimated $\Pr(Y = i|T = j)$ and $\Pr(T = j)$ are given, retrieves all the $p_{ij}$ parameters that can be used to simulate $U$ with $d$ and $s$ varying from 0.1 to 0.6.

[22] See Ichino, Mealli, and Nannicini (2007) for a concrete example.

attk, and attr. Before using sensatt, you should install them and be familiar with their use.

- The treatment must be binary.
- It is important to clean up the dataset before running the program, in particular to delete observations with missing values.

# 6 Options

## 6.1 Options specific to sensatt

alg(att*) specifies the name of the command (i.e., of the matching algorithm) that is used in the ATT estimation. One of the following commands can be specified: attnd, attnw, attk, and attr. The default is attnd.

reps(#) specifies the number of iterations, i.e., how many times the simulation of $U$ and the ATT estimation are replicated. The default is reps(1000).

p(*varname*) specifies the binary variable that is used to simulate the confounder. The parameters $p_{ij}$ used to simulate $U$ are set equal to the ones observed for *varname*. Instead of selecting this option, the user can directly specify the parameters $p_{ij}$.

p11(#), p10(#), p01(#), and p00(#) jointly specify the parameters $p_{ij}$ used to simulate $U$ in the data. Since they are probabilities, they must be between zero and one. For each parameter, the default is zero.

se(*se_type*) allows the user to decide which standard error should be displayed with the simulated ATT. Three *se_type*s are possible: tse uses the total variance in a multiple-imputation setting; wse uses the within-imputation variance; and bse uses the between-imputation variance. The default is tse.

ycent(#) is relevant only with continuous outcomes. It means that $U$ is simulated on the basis of the binary transformation of the outcome $I(Y > y^*)$, where $y^*$ is the #th centile of the distribution of $Y$. Three centiles are allowed: 25, 50, and 75. If ycent(#) is not specified by the user, but the outcome is continuous, $U$ is simulated on the basis of the transformation $I(Y > y^*)$, where $y^*$ is the mean of $Y$.

## 6.2 Options common to attnd, attnw, attk, and attr

pscore(*scorevar*) specifies the name of the user-provided variable containing the estimated propensity score. If this option is not selected, the propensity score is estimated with the specification provided in *varlist*.

logit uses a logit model to estimate the propensity score instead of the default probit model when the option pscore(*scorevar*) is not specified by the user.

index requires the use of the linear index as the propensity score when the option pscore(*scorevar*) is not specified by the user.

`comsup` restricts the computation of the ATT to the region of common support.

`bootstrap` bootstraps the standard errors of the estimated ATTs. This option should not be used in association with the nearest-neighbor algorithm (`attnd`); see Abadie and Imbens (2006).

## 7 Saved results

`sensatt` saves the following in `r()`:

Scalars

| | |
|---|---|
| `r(att)` | simulated ATT |
| `r(se)` | default standard error |
| `r(sew)` | within-imputation standard errors |
| `r(seb)` | between-imputation standard errors |
| `r(yodds)` | estimated outcome effect of the confounder $U$ (odds ratio) |
| `r(todds)` | estimated selection effect of the confounder $U$ (odds ratio) |

## 8 Example

Per Becker and Ichino (2002), I use data from Dehejia and Wahba (1999), which are publicly available at http://www.nber.org/%7Erdehejia/nswdata.html. The data come from LaLonde's (1986) well-known evaluation of nonexperimental evaluation methods, which combines the treated units from a randomized study of the National Supported Work (NSW) training program with nonexperimental comparison groups drawn from public surveys. As mentioned in section 1, Dehejia and Wahba use this dataset to show that propensity-score matching estimates are closer to the experimental benchmark than those produced by traditional evaluation methods. I restrict my example to the comparison group drawn from the Panel Study of Income Dynamics (PSID-1).[23] The outcome of interest is continuous and is represented by the postintervention real earnings (`RE78`). The treatment indicator (`T`) coincides with the participation to the NSW treated group. Control variables are age (`age`), education (`educ`), black dummy (`black`), Hispanic dummy (`hisp`), marital status (`marr`), real earnings in 1975 (`RE75`), and real earnings in 1974 (`RE74`).[24] At the end, there are 185 observations in the treated group and 2,490 in the control group. For this subsample of the NSW treated group, the experimental estimate of the ATT is 1,794 (with a standard error equal to 633). I focus on the nearest-neighbor matching estimate, which is the default in `sensatt`. Assume that we want to calculate this estimate and assess its robustness with respect to a potential confounder that behaves like an important observed covariate: the probability of being nonemployed in 1974 (`U74`). The following three Stata outputs are produced by running `sensatt` with the above specification of the propensity score and simulating $U$ to mimic the variable `U74`:

---

[23] See LaLonde (1986) and Dehejia and Wahba (1999, 2002) for more data details.

[24] Throughout this example, to replicate Becker and Ichino's results, which in turn replicate those by Dehejia and Wahba, the propensity-score specification also includes the following variables: squared education (`educ2`), squared earnings in 1974 (`RE742`), squared earnings in 1975 (`RE752`), and the interaction of the black dummy with a dummy for nonemployment in 1974 (`blackU74`).

```
. sensatt RE78 T age age2 educ educ2 marr black hisp RE74 RE75 RE742 RE752
> blackU74, p(U74) r(100) comsup logit

*** THIS IS THE BASELINE ATT ESTIMATION (WITH NO SIMULATED CONFOUNDER).

 The program is searching the nearest neighbor of each treated unit.
 This operation may take a while.

ATT estimation with Nearest Neighbor Matching method
(random draw version)
Analytical standard errors

---------------------------------------------------------
n. treat.   n. contr.         ATT    Std. Err.          t
---------------------------------------------------------

      185          57    1667.644     2113.592      0.789

---------------------------------------------------------
Note: the numbers of treated and controls refer to actual
nearest neighbour matches
```

First, `sensatt` shows the ATT calculated by the command for propensity-score matching that has been selected (`attnd` in this example). Correctly, the above estimate is the same of the example by Becker and Ichino, and it is close to the nearest-neighbor matching estimate in Dehejia and Wahba's original paper, which is equal to 1,691 (with a standard error of 2,209). The baseline ATT point estimate is close to the experimental benchmark, even though the standard error is high. The fact that we can compare the nonexperimental estimates with this unbiased benchmark makes the sensitivity analysis useless. But let us assume that this is not the case, and we would like to assess the robustness of the above matching estimate. After the simple step of reproducing the output by `attnd`, the program moves on and simulates the confounder $U$ to retrieve the associated ATT:

```
*** THIS IS THE SIMULATED ATT ESTIMATION (WITH THE CONFOUNDER U).
The probability of having U=1 if T=1 and Y=1 (p11) is equal to:     0.78
The probability of having U=1 if T=1 and Y=0 (p10) is equal to:     0.70
The probability of having U=1 if T=0 and Y=1 (p01) is equal to:     0.02
The probability of having U=1 if T=0 and Y=0 (p00) is equal to:     0.15
The probability of having U=1 if T=1 (p1.) is equal to:     0.71
The probability of having U=1 if T=0 (p0.) is equal to:     0.09

 The program is iterating the ATT estimation with simulated confounder.
 You have chosen to perform 100 iterations. This step may take a while.
```

The iteration step can be time consuming, especially when, unlike in this example, one selects the `bootstrap` option to calculate the standard error of the chosen propensity-score matching estimator. At the end of the iteration step, `sensatt` displays the simulated ATT, as well as the outcome and selection effects of $U$:

```
ATT estimation with simulated confounder
General multiple-imputation standard errors

       ATT    Std. Err.    Out. Eff.   Sel. Eff.

 2122.280    3548.766        0.132      15.946

Note: Both the outcome and the selection effect
are odds ratios from logit estimations.
```

As one would expect, the simulated ATT (2,122) is even greater than the baseline estimate (1,668) since, even though the selection effect of the confounder is large, the outcome effect is negative (i.e., $d < 0$). One may want to test the robustness of the baseline ATT with respect to a confounder that is more dangerous (i.e., a confounder $U$ such that both $d > 0$ and $s > 0$) but still behaves like other relevant observable variables. Let us run `sensatt` with the confounder $U$ calibrated to mimic the constructed variable `young` (i.e., being below the 75th centile of the age distribution):

```
. sensatt RE78 T age age2 educ educ2 marr black hisp RE74 RE75 RE742 RE752
> blackU74, p(young) comsup logit;
  (output omitted)
*** THIS IS THE SIMULATED ATT ESTIMATION (WITH THE CONFOUNDER U).
The probability of having U=1 if T=1 and Y=1 (p11) is equal to:     1.00
The probability of having U=1 if T=1 and Y=0 (p10) is equal to:     0.97
The probability of having U=1 if T=0 and Y=1 (p01) is equal to:     0.75
The probability of having U=1 if T=0 and Y=0 (p00) is equal to:     0.72
The probability of having U=1 if T=1 (p1.) is equal to:     0.97
The probability of having U=1 if T=0 (p0.) is equal to:     0.73

 The program is iterating the ATT estimation with simulated confounder.
 You have chosen to perform 100 iterations. This step may take a while.

ATT estimation with simulated confounder
General multiple-imputation standard errors

       ATT    Std. Err.    Out. Eff.   Sel. Eff.

 1593.286     3006.708        1.149      19.643

Note: Both the outcome and the selection effect
are odds ratios from logit estimations.
```

Here the simulated ATT is lower, but the potential confounder kills only by a small amount the baseline estimate. The sensitivity analysis is telling us that the existence of a confounder $U$ behaving like the young dummy might account for nearly 5% of the baseline estimate: $(1{,}668 - 1{,}593)/1{,}668 = 0.05$. Since the outcome is continuous, one may want to check whether the sensitivity conclusions depend on the fact that $U$ is simulated on the basis of the binary transformation of $Y$ that uses the mean of the outcome (see sec. 3). In the following Stata output, $U$ is again simulated to mimic `young`, but the parameters $p_{ij}$ refer to the binary transformation of $Y$ that uses the median of the outcome. Moreover, the between-imputation standard error is shown, to use only the variability of the simulated ATT across iterations.

```
. sensatt RE78 T age age2 educ educ2 marr black hisp RE74 RE75 RE742 RE752
> blackU74, p(young) ycent(50) se(bse) comsup logit;
```

*(output omitted)*

```
ATT estimation with simulated confounder
Between-imputation standard errors

-----------------------------------------------
      ATT    Std. Err.    Out. Eff.   Sel. Eff.
-----------------------------------------------
 1526.429      867.942        1.179      21.676
-----------------------------------------------
Note: Both the outcome and the selection effect
are odds ratios from logit estimations.
```

Using the median instead of the mean of $Y$ does not affect the results of the sensitivity analysis, since the simulated ATT is close to the previous one. On the contrary, the between-imputation standard error is much lower than the default one. However, the sensitivity conclusions should be drawn more as the comparison of the point estimates than as the significance of the simulated ATT.

The above simulations convey an image of robustness of the nearest neighbor matching estimate equal to 1,668. This image, however, might be produced by the particular characteristics of the covariates used to simulate $U$ (`U74` and `young`), rather than by the fact that the baseline ATT is robust to possible deviations from the CIA. Similar sensitivity conclusions, however, arise from the second simulation exercise proposed in section 4. For brevity, I do not calculate a table like the one suggested in the discussion about the search for killer confounders. Two simple examples will suffice. Assume that you wish to simulate $U$ according to the following parameters: $p_{11} = 0.8$, $p_{10} = 0.8$, $p_{01} = 0.6$, and $p_{00} = 0.3$. We expect this potential confounder to represent a real threat for the baseline estimate and to be associated with large selection and outcome effects ($s = 0.34 > 0$ and $d = 0.3 > 0$).

```
. sensatt RE78 T age age2 educ educ2 marr black hisp RE74 RE75 RE742 RE752
> blackU74, p11(0.8) p10(0.8) p01(0.6) p00(0.3) se(bse) comsup logit;
```

*(output omitted)*

```
*** THIS IS THE SIMULATED ATT ESTIMATION (WITH THE CONFOUNDER U).
The probability of having U=1 if T=1 and Y=1 (p11) is equal to:     0.80
The probability of having U=1 if T=1 and Y=0 (p10) is equal to:     0.80
The probability of having U=1 if T=0 and Y=1 (p01) is equal to:     0.60
The probability of having U=1 if T=0 and Y=0 (p00) is equal to:     0.30
The probability of having U=1 if T=1 (p1.) is equal to:     0.80
The probability of having U=1 if T=0 (p0.) is equal to:     0.46

 The program is iterating the ATT estimation with simulated confounder.
 You have chosen to perform 100 iterations. This step may take a while.
```

```
ATT estimation with simulated confounder
Between-imputation standard errors

      ATT    Std. Err.    Out. Eff.   Sel. Eff.

 1588.864     1093.235        3.502       9.757

Note: Both the outcome and the selection effect
are odds ratios from logit estimations.
```

On the contrary, even though $U$ is associated with large selection and outcome effects ($\Lambda = 9.8$ and $\Gamma = 3.5$), the simulated ATT is still close to the baseline estimate. But when $U$ is simulated so that it displays a large outcome effect, the ATT is driven closer to zero:

```
. sensatt RE78 T age age2 educ educ2 marr black hisp RE74 RE75 RE742 RE752
> blackU74, p11(0.8) p10(0.8) p01(0.6) p00(0.1) se(bse) comsup logit;
  (output omitted)
*** THIS IS THE SIMULATED ATT ESTIMATION (WITH THE CONFOUNDER U).
The probability of having U=1 if T=1 and Y=1 (p11) is equal to:     0.80
The probability of having U=1 if T=1 and Y=0 (p10) is equal to:     0.80
The probability of having U=1 if T=0 and Y=1 (p01) is equal to:     0.60
The probability of having U=1 if T=0 and Y=0 (p00) is equal to:     0.10
The probability of having U=1 if T=1 (p1.) is equal to:     0.80
The probability of having U=1 if T=0 (p0.) is equal to:     0.36

 The program is iterating the ATT estimation with simulated confounder.
 You have chosen to perform 100 iterations. This step may take a while.

ATT estimation with simulated confounder
Between-imputation standard errors

      ATT    Std. Err.    Out. Eff.   Sel. Eff.

  274.110     2640.772       13.220      31.413

Note: Both the outcome and the selection effect
are odds ratios from logit estimations.
```

To let $U$ explain about 84% of the baseline estimate ($(1{,}668 - 274)/1{,}668 = 0.84$), such a confounder must have a large effect on both the outcome and the selection into treatment. More precisely, $U$ must increase the relative probability of having $Y$ above the mean ($T = 1$) by a factor greater than 13 (31). The presence among unobservable factors of a confounder with similar characteristics can be considered implausible in the present setting (where the set of matching variables $W$ is rich). These simple simulation exercises support the robustness of the matching estimate.

## 9 Acknowledgments

I thank Sascha Becker, Andrea Ichino, Fabrizia Mealli, participants of the 2006 North American Stata Users Group meeting, and an anonymous referee for useful suggestions.

## 10 References

Abadie, A., D. Drukker, J. Leber Herr, and G. W. Imbens. 2004. Implementing matching estimators for average treatment effects in Stata. *Stata Journal* 4: 290–311.

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

Altonji, J. G., T. E. Elder, and C. R. Taber. 2005. Selection on observed and unobserved variables: Assessing the effectiveness of Catholic schools. *Journal of Political Economy* 113: 151–184.

Becker, S., and M. Caliendo. 2007. Sensitivity analysis for average treatment effects. *Stata Journal* 7: 71–83.

Becker, S. O., and A. Ichino. 2002. Estimation of average treatment effects based on propensity scores. *Stata Journal* 2: 358–377.

Black, D., and J. Smith. 2004. How robust is the evidence on the effects of college quality? Evidence from matching. *Journal of Econometrics* 121: 99–124.

Caliendo, M., and S. Kopeinig. 2007. Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys.* Forthcoming.

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

———. 2002. Propensity score matching methods for non-experimental causal studies. *Review of Economics and Statistics* 84: 151–161.

Gastwirth, J. L., A. M. Krieger, and P. R. Rosenbaum. 1998. Dual and simultaneous sensitivity for match pairs. *Biometrika* 85: 907–920.

Heckman, J., H. Ichimura, and P. Todd. 1997. Matching as an econometric evaluation estimator: Evidence from evaluating a job training program. *Review of Economic Studies* 64: 605–654.

———. 1998. Matching as an econometric evaluation estimator. *Review of Economic Studies* 65: 261–294.

Ichino, A., F. Mealli, and T. Nannicini. 2007. From temporary help jobs to permanent employment: What can we learn from matching estimators and their sensitivity? *Journal of Applied Econometrics.* Forthcoming.

Imbens, G. W. 2003. Sensitivity to exogenieity assumptions in program evaluation. *American Economic Review Papers and Proceedings* 2: 126–132.

———. 2004. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics* 86: 4–29.

LaLonde, R. 1986. Evaluating the econometric evaluations of training programs. *American Economic Review* 76: 604–620.

Leuven, E., and B. Sianesi. 2003. PSMATCH2: Stata module to perform full Mahalanobis and propensity score matching, common support graphing, and covariate imbalance testing. Boston College Department of Economics, Statistical Software Components. Downloadable from http://ideas.repec.org/c/boc/bocode/s432001.html.

Michalopoulos, C., H. S. Bloom, and C. J. Hill. 2004. Can propensity-score methods match the findings from a random assignment evaluation of mandatory welfare-to-work programs? *Review of Economics and Statistics* 86: 156–179.

Rosenbaum, P. R. 1987a. The role of a second control group in an observational study. *Statistical Science* 2: 292–306.

———. 1987b. Sensitivity analysis to certain permutation inferences in matched observational studies. *Journal of the Royal Statistical Society, Series B* 45: 212–218.

———. 2002. *Observational Studies*. 2nd ed. New York: Springer.

Rosenbaum, P. R., and D. B. Rubin. 1983a. Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society, Series B* 45: 212–218.

———. 1983b. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70: 41–55.

Rubin, D. B. 1974. Estimating causal effects of treatments in randomised and nonrandomised studies. *Journal of Educational Psychology* 66: 688–701.

Smith, J., and P. Todd. 2005. Does matching overcome Lalonde's critique of nonexperimental estimators? *Journal of Econometrics* 125: 305–353.

## About the author

Tommaso Nannicini is a visiting assistant professor of economics at Carlos III University of Madrid, Madrid, Spain. Improved versions of the program may become available on the web site http://www.tommasonannicini.eu.