# A Simulation-Based Sensitivity Analysis for Matching Estimators

Tommaso Nannicini
Universidad Carlos III de Madrid

**Abstract.** This article presents a Stata program (`sensatt`) that implements the sensitivity analysis for propensity-score matching estimators proposed by Ichino, Mealli and Nannicini (2006). The analysis builds on Rosenbaum and Rubin (1983a) and Rosenbaum (1987a), and simulates a potential confounder in order to assess the robustness of the estimated treatment effects with respect to deviations from the Conditional Independence Assumption (CIA). The program makes use of the commands for propensity-score matching (`att*`) developed by Becker and Ichino (2002). An example is given by using the National Supported Work (NSW) demonstration, widely known in the program evaluation literature.

**Keywords:** propensity score, matching, sensitivity analysis, program evaluation.

## 1 Introduction

During last years, the utilization of 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 utilization is now widespread in applied studies. Secondly, a lot of free user-friendly software routines have been made available in order 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 implements 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.

[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., the fact that these estimators seem to perform better with respect to alternative non-experimental 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 very 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).

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 only driven 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, in order 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, 2) the non-experimental 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 (assuming that there is no credible source of exogenous variation), since it does not rely on linearity and allows to check whether it exists a substantial overlap of the distributions of covariates in the treatment and comparison groups. However, every evaluation strategy making use of matching estimators should contain some of the following steps (possibly, all of them).

**First step.** To use data where the treated and control units come from the same local market and are asked the same set of questions.

**Second step.** To discuss (carefully) why the CIA should be verified in the specific context of the evaluation question at hand.

**Third step.** To test (indirectly) whether the available empirical evidence casts doubt on the plausibility of the CIA.[4]

**Fourth step.** To inspect how the observations are distributed across the propensity-score common support and how sensitive the estimates are with respect to the utilization of observations in the tails of the common support.[5]

**Fifth step.** 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).

The sensitivity analysis proposed by Ichino, Mealli and Nannicini (2006) allows applied researchers who make use of matching estimators to tackle the fifth step. The analysis builds on Rosenbaum and Rubin (1983a) and Rosenbaum (1987a), and it is based on a simple idea. Suppose that the CIA is not satisfied given observables but would

[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] See Rosenbaum (1987b) and Imbens (2004).

[5] See Black and Smith (2004).

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 a specific source 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 give a short summary of this econometric tool[6] 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:

$$E(Y_1 - Y_0 | T = 1). \tag{1}$$

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

$$Y_0 \perp\!\!\!\perp T \,|\, W. \tag{2}$$

This condition is the CIA. The behavioral assumption behind it is that the potential outcome in case of no treatment ($Y_0$) does not influence treatment assignment, while the possibility that the selection process depends on the treated outcome ($Y_1$) does not have to be ruled out.[8] Although very strong, the plausibility of this assumption heavily relies on the quality and amount of information contained in $W$. Note that 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/rejected by theoretical reasoning and additional evidence.[9] 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:[10]

$$Pr(T = 1|W) < 1. \tag{3}$$

Under assumptions (2) and (3), 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 in case of no treatment. However, with a high dimensional vector $W$, this task may be problematic. To deal with the dimensionality problem, one

[6] See Ichino, Mealli and Nannicini (2006) for further details and an empirical application.

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

[8] See Heckman and Smith (1998) for further discussion.

[9] See the references for the *third step* of a correct matching strategy mentioned in Section 1.

[10] To estimate the ATE, the overlap condition would require: $0 < Pr(T = 1|W) < 1$.

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:

$$\tau_{ATT} \equiv E(Y_1 - Y_0|T = 1) = E_{\{p(W)|T=1\}}[E(Y_1|p(W), T = 1) - E(Y_0|p(W), T = 0)] \quad (4)$$

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 pre-treatment 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.[11] However, the estimation of the propensity score is not enough to estimate the ATT using equation (4), since the probability of finding two observations with exactly 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` makes use of three different algorithms: nearest neighbor; kernel; radius.[12] These methods differ from each other with respect to the way 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 (2006) and briefly 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 (2) no longer holds. In addition, it is assumed that the CIA holds given $W$ and an unobserved binary variable $U$:

$$Y_0 \perp\!\!\!\perp T \,|\, (W, U).^{13} \quad (5)$$

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

[12] 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 (2006) for a discussion of the different properties of these and other propensity-score matching algorithms.

[13] Using Rosenbaum's (1987b) terminology, we are moving from $(Y_0|W)$-adjustable treatment assignment in condition 2 to $(Y_0|W, U)$-adjustable treatment assignment in condition 5.

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

$$E(Y_0|T=1,W) \neq E(Y_0|T=0,W). \tag{6}$$

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:

$$E(Y_0|T=1,W,U) = E(Y_0|T=0,W,U). \tag{7}$$

Note that assumption (5) is common to similar sensitivity analysis proposed in the econometric and statistical literature,[14] 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.[15]

The subsequent step consists in characterizing the distribution of $U$, in order to simulate this potential confounder in the data. As said, $U$ is assumed to be binary. In addition, it is assumed to be i.i.d. distributed in the cells represented by the Cartesian product of the treatment and outcome values. For expositional simplicity, consider the case of binary potential outcomes: $Y_0, Y_1 \in \{0,1\}$.[16] 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:

$$p_{ij} \equiv Pr(U=1|T=i,Y=j) = Pr(U=1|T=i,Y=j,W) \tag{8}$$

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.[17] Note that, in order to make the simulation of the potential confounder feasible, two simplifying assumptions are made: 1) binary $U$, 2) conditional independence of $U$ with respect to $W$. Ichino, Mealli and Nannicini (2006) 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 subject, according to her/his 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 estimator (e.g., kernel). Using a given set of values of the sensitivity parameters, the matching estimation is repeated many times (e.g., $1,000$)

[14] See Rosenbaum and Rubin (1983a), Rosenbaum (1987a, 2002), Imbens (2003), and Altonji, Elder and Taber (2005).

[15] See Ichino, Mealli and Nannicini (2006) for a discussion of how their method contributes to the literature mentioned in the previous footnote.

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

[17] Using the parameters $p_{ij}$ and probabilities of positive potential outcomes 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 $i \in \{0,1\}$.

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 which is robust to the failure of the CIA implied by that particular configuration.

**Standard errors**. In order to compute a standard error for the simulated ATT, the imputation of $U$ is considered as a normal problem of missing data, which can be solved by multiply imputing the missing values of $U$. Let $m$ be the number of imputations of the missing $U$, and let $A\hat{T}T_k$ and $se_k^2$ be the point estimate and the estimated variance of the ATT estimator at the $k$-th imputed data set (with $k = 1, 2, \ldots, m$). The simulated ATT, $A\hat{T}T$, is obtained by the average of the $A\hat{T}T_k$ over the $m$ replications. In this setting, the within-imputation variance is equal to

$$se_W^2 = \frac{1}{m} \sum_{k=1}^{m} se_k^2, \tag{9}$$

while the between-imputation variance is given by

$$se_B^2 = \frac{1}{m-1} \sum_{k=1}^{m} (A\hat{T}T_k - A\hat{T}T)^2. \tag{10}$$

As a consequence, the total variance associated to $A\hat{T}T$ can be expressed as:

$$se_T^2 = se_W^2 + (1 + \frac{1}{m}) se_B^2. \tag{11}$$

For a large number of replications, the statistic $(A\hat{T}T - ATT)/se_T$ is approximately normal. Alternatively, one could consider either the within-imputation or the between-imputation standard error as the basis for inference. The standard error in equation (11) leads to conservative inferential conclusions, since it is always greater than the other two alternatives. Remind, however, that the results of this simulation-based sensitivity analysis should be judged more on the basis of the distance between point estimates associated to different $p_{ij}$, rather than the significance level of the simulated ATTs.

**Extension to continuous outcomes.** The above sensitivity analysis can be easily extended to multi-valued or continuous outcomes. Indeed, in such cases, it is possible to define the simulation parameters $p_{ij}$ on the basis of $T$ and a binary transformation of $Y$ (instead of the outcome itself). Define:

$$p_{ij} \equiv Pr(U = 1 | T = i, I(Y > y^*) = j), \tag{12}$$

with $i, j \in \{0, 1\}$, where $I$ is the indicator function and $y^*$ is a chosen typical value of the distribution of $Y$.[18] Once the parameters $p_{ij}$ are set in this way, one can implement the sensitivity analysis as described above. Of course, the ATT is still estimated for the multi-valued or continuous outcome $Y$.

[18] The program `sensatt` allows for the utilization of four $y^*$: mean, median, 25th or 75th centile.

# 4 Guidelines for the implementation of the simulations

In order 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. In other words, one must answer the following question: which values of the parameters $p_{ij}$ should I choose in order to learn something useful from the effect of a confounder $U$ like the one associated to the chosen values? In this section, two simulation exercises are proposed. 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, in order to capture the characteristics of those potential confounders that would drive the ATT estimates to zero. Note, however, that the above sensitivity analysis is a flexible tool and its application is not restricted to the exercises suggested here.

Before discussing 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 $W$). As a matter of fact, $U$ would be a “dangerous” confounder (i.e., a confounder whose existence might give rise to a positive and significant ATT estimate even in the absence of a true causal effect) if we observed that:

$$Pr(Y_0 = 1|T, W, U) \neq Pr(Y_0 = 1|T, W), \tag{13}$$

$$Pr(T = 1|W, U) \neq Pr(T = 1|W). \tag{14}$$

Note that expressions (13) and (14) - unlike the parameters $p_{ij}$ - both include $W$ and refer to the potential (not observed) outcome in case of 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 (2006) demonstrate that the following implications hold:

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

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

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 $p_{1.} > p_{0.}$,[19] one can simulate a confounding factor that has a positive effect on treatment assignment.

There is a limitation, however, that must be addressed. Following the above reasoning, it would be tempting to interpret the difference $d = p_{01} - p_{00}$ as a measure of the

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

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 $W$, while $d$ and $s$ do not consider the association between $U$ and $W$ that shows up in the data. In other words, 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 $p_{ij}$ 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:

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

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 $U$.

**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 $p_{i.}$) in order to make the distribution of $U$ similar to the empirical distribution of important binary covariates (or binary transformations of continuous covariates). In this case, 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 in every sensitivity-analysis estimation we still control for all the relevant covariates observed by the econometrician. Of course, this exercise is interesting when the chosen covariates display $p_{ij}$ that satisfy the conditions: $d > 0$ and $s > 0$.

**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 then assess the plausibility of this particular configuration of parameters. If all the configurations leading to such result could be considered very unlikely, the exercise would support the robustness of the estimates derived under the CIA. In order 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 ($d = 0$, $s = 0$). Moreover, when displaying the results of the sensitivity analysis, the values of $d$ and $s$ should be associated to the estimated values of $\Gamma$ and $\Lambda$, respectively. In this way, the estimated odds ratios would 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. As said, 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.[21]

# 5 Syntax

```
sensatt outcome treatment [varlist] [weight] [if exp] [in range] [ , 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 makes use of the commands for the propensity-score matching estimation of average treatment effects written by Becker and Ichino (2002): `attnd`, `attnw`, `attk`, `attr`. Before using `sensatt`, you should install them and be familiar with their utilization.
- 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 that are 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`, `attr`. The default is `attnd`.

[20] Note that, 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] See Ichino, Mealli and Nannicini (2006) for a concrete example.

`p(varname)` indicates the binary variable which 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.

`reps(#)` specifies the number of iterations, i.e., how many times the simulation of $U$ and the ATT estimation are replicated. The default is 1,000.

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

`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, 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 that are 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.

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

`bootstrap` bootstraps the standard errors of the estimated ATTs.

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

# 7 Saved results

The `sensatt` command saves in `r()`:

Scalars

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

## 8 An example

Following Becker and Ichino (2002), I use data from Dehejia and Wahba (1999), which are publicly available at the website: http://www.nber.org/%7Erdehejia/nswdata.html. The data come from LaLonde's (1986) well known evaluation of non-experimental evaluation methods, which combines the treated units from a randomized study of the National Supported Work (NSW) training program with non-experimental comparison groups drawn from public surveys. As mentioned in Section 1, Dehejia and Wahba use this data set 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).[22] The outcome of interest is continuous and is represented by the post-intervention 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*), real earnings in 1974 (*RE74*).[23] 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 non-employed in 1974 (*U74*). The following three Stata outputs are produced by running `sensatt` with the above specification of the propensity score and simulating $U$ so as to mimic the variable *U74*:

```
. 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
```

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

[23] Throughout this example, in order to replicate Becker and Ichino's results, which in turn replicate those by Dehejia and Wahba, the propensity-score specification includes also 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 non-employment in 1974 (*blackU74*).

First of all, 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 very 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 very close to the experimental benchmark, even though the standard error is very high. The fact that we can compare the non-experimental 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 $U$ in order 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 very 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.
-----------------------------------------------

 1931.535    3674.773        0.699      15.016

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

The simulated ATT (1,931) is even greater that the baseline estimate (1,668) since, even though the selection effect of the confounder is very large, the outcome effect is negative (i.e., $d < 0$ or $\Gamma < 1$). One may want to test the robustness of the baseline ATT with respect to a confounder that is more "dangerous" (i.e., $U$ such that both $\Gamma > 1$ and $\Lambda > 1$) but still behaves like other relevant observable covariates. Let us run sensatt with the confounder $U$ calibrated in order to mimic the variable *hisp*:

```
. sensatt RE78 T age age2 educ educ2 marr black hisp RE74 RE75 RE742 RE752
> blackU74, p(hisp) 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:     0.11
The probability of having U=1 if T=1 and Y=0 (p10) is equal to:     0.06
The probability of having U=1 if T=0 and Y=1 (p01) is equal to:     0.03
The probability of having U=1 if T=0 and Y=0 (p00) is equal to:     0.04

The probability of having U=1 if T=1 (p1.) is equal to:     0.06
The probability of having U=1 if T=0 (p0.) is equal to:     0.03

 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.
-----------------------------------------------

 1452.600    2247.901        1.093       2.009

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

In this case, the simulated ATT is lower, but the potential confounder "kills" only by a small amount the baseline estimate. In other terms, the sensitivity analysis is telling us that the existence of a confounder $U$ behaving like the Hispanic dummy might account for nearly 13% of the baseline estimate: $(1,668 - 1,453)/1,668 = 0.13$. Not surprisingly, the outcome and selection effects of a similar confounder are also quite small. Since the outcome is continuous, one may want to check whether the results of the sensitivity analysis 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 Section 3). In the following Stata output, $U$ is again simulated in order to mimic *hisp*, 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 showed, in order 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(hisp) ycent(50) se(bse) comsup logit;
 (output omitted)
ATT estimation with simulated confounder
Between-imputation standard errors
-----------------------------------------------
      ATT    Std. Err.    Out. Eff.   Sel. Eff.
-----------------------------------------------

 1453.177     531.893        1.064       1.935

-----------------------------------------------
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 very 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 in terms of the comparison of point estimates, rather than in terms of 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 *hisp*),[24] 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 the sake of brevity, I do not calculate a table like the one suggested in the discussion about the search for "killer" confounders. Two simple examples will be enough to illustrate the point. Simulate $U$ according to the following parameters: $p_{11} = 0.8$, $p_{10} = 0.8$, $p_{01} = 0.6$, $p_{00} = 0.3$. We expect this potential confounder to represent a real threat for the baseline estimate, and to be associated to large selection and outcome effects (note that: $s = 0.34 > 0$ 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
General multiple-imputation standard errors
-----------------------------------------------
      ATT    Std. Err.    Out. Eff.   Sel. Eff.
-----------------------------------------------

 1639.271    1114.289       1.560       9.887

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

On the contrary, even though $U$ is associated to quite large selection and outcome effects ($\Lambda = 9.9$ and $\Gamma = 1.6$), the simulated ATT is still very close to the baseline estimate. Only when $U$ is simulated so that it displays a very (and implausibly) large outcome effect, the ATT is driven closer to zero:

[24] The utilization of other covariates produced similar results.

```
. 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
General multiple-imputation standard errors
-----------------------------------------------
      ATT    Std. Err.    Out. Eff.   Sel. Eff.
-----------------------------------------------

  339.721    2624.902        3.246      33.519

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

To let the potential confounder $U$ explain about 80% of the baseline estimate $((1,668-340)/1,668=0.80)$, such a confounder must have a very large effect on the untreated outcome and the selection into treatment. More precisely, $U$ must increase the relative probability of having $Y=1$ ($T=1$) by a factor greater than 3 (30). 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 quite rich). At the end of the day, these simple simulation exercises support the robustness of the nearest neighbor matching estimate.

## 9 References

Abadie A., Drukker D., Leber Herr J. and Imbens G.W. (2004),"Implementing Matching Estimators for Average Treatment Effects in Stata", *The Stata Journal*, Vol.4, 3, 290-311.

Altonji J.G., Elder T.E., and Taber C.R. (2005),"Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools", *Journal of Political Economy*, Vol.113, 1, 151-184.

Becker S. and Ichino A. (2002),"Estimation of average treatment effects based on Propensity Scores", *The Stata Journal*, Vol.2, 4, 358-377.

Caliendo M. and Kopeinig S. (2006), "Some Pratical Guidance for the Implementation of Propensity Score Matching", *Journal of Economic Surveys*, forthcoming.

Dehejia R.H. and Wahba S. (1999), "Causal Effects in Non-Experimental Studies: Re-Evaluating the Evaluation of Training Programs", *Journal of the American Statistical Association*, Vol.94, 448, 1053-62.

Dehejia R.H. and Wahba S. (2002), "Propensity Score-Matching Methods for Non-experimental Causal Studies", *The Review of Economics and Statistics*, 84(1), 151-161.

Heckman J.J., Ichimura H. and Todd P. (1997), "Matching As An Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme", *The Review of Economic Studies*, 64, 605-654.

Heckman J.J., Ichimura H. and Todd P. (1998), "Matching As An Econometric Evaluation Estimator", *The Review of Economic Studies*, 65, 261-294.

Heckman J. and Smith J. (1998), *Evaluating the Welfare State*, NBER Working Paper No.6542.

Ichino A., Mealli F. and Nannicini T. (2006), *From Temporary Help Jobs to Permanent Employment: What Can We Learn from Matching Estimators and their Sensitivity?*, IZA Discussion Paper No.2149.

Imbens G.W. (2003)"Sensitivity to Exogeneity Assumptions in Program Evaluation", *AEA Papers and Proceedings*, 93, 2, 126-132.

Imbens G.W. (2004), "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review", *The Review of Economics and Statistics*, 86(1), 4-29.

LaLonde R. (1986), "Evaluating the Econometric Evaluations of Training Programs", *American Economic Review*, 76(4), 604-620.

Leuven E. and Sianesi B. (2003), *psmatch2: Stata module to perform full Mahalanobis and propensity score matching, common support graphing, and covariate imbalance testing*, available at: http://ideas.repec.org/c/boc/bocode/s432001.html.

Michalopoulos C., Bloom H.S. and Hill C.J. (2004), "Can Propensity-Score Methods Match the Findings from a Random Assignment Evaluation of Mandatory Welfare-to-Work Programs?", *The Review of Economics and Statistics*, 86(1), 156-179.

Rosenbaum P. (1987a), "Sensitivity Analysis to Certain Permutation Inferences in Matched Observational Studies", *Biometrika*, 74, 1, 13-26.

Rosenbaum P. (1987b), "The Role of a Second Control Group in an Observational Study", *Statistical Science*, 2, 3, 292-306.

Rosenbaum P. (2002), *Observational studies*, New York, Springer Verlag.

Rosenbaum P. and Rubin D. (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.

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

Rubin D. (1974), "Estimating Causal Effects of Treatments in Randomised and Non-Randomised Studies", *Journal of Educational Psychology*, 66, 688-701.

Smith J. and Todd P. (2005), "Does Matching Overcome Lalonde's Critique of Non-experimental Estimators?", *Journal of Econometrics*, 125, 305-353.

### About the Author

Tommaso Nannicini is Visiting Professor of Economics at Carlos III University of Madrid (Spain). Please address any correspondence to: tommaso.nannicini@uc3m.es.