Sensitivity of matching estimators to unconfoundedness.
An application to the effect of temporary work on future employment.*

Andrea Ichino
EUROPEAN UNIVERSITY INSTITUTE

Fabrizia Mealli
UNIVERSITY OF FLORENCE

Tommaso Nannicini
EUROPEAN UNIVERSITY INSTITUTE

January 28, 2005

Abstract

Matching estimators are now easy to use and perhaps too many users adopt them without really checking that the conditions for their application are satisfied. In this paper we propose a sensitivity analysis for matching estimators aimed at assessing to what extent the estimates derived under the unconfoundedness assumption are robust with respect to specific failures of this assumption. We also give an example of how this sensitivity analysis can be implemented, based on Italian data collected to evaluate the effect of a temporary work experience on future employment probabilities.

JEL Classification: C2, C8, J6.

*Andrea Ichino is also affiliated with CEPR, CESifo and IZA. Financial support by the Italian Ministry of Welfare and the Tuscany Region is gratefully acknowledged. We also thank “Manpower Italia Spa” for help in data collection and seminar participants at the universities of Venezia, Perugia, the EC workshop on “Temporary Work in Europe”, the 2004 CEA Annual Meeting and the 2004 ZEW Conference on Evaluation Research for insightful comments. Address correspondence to: andrea.ichino@iue.it; mealli@ds.unifi.it; tommaso.nannicini@iue.it.
1 Introduction

Matching estimators have been recently used with increasing frequency in evaluation studies for which a convincing source of exogenous random variation of assignment to treatment does not exist. The debate raised by Dehejia and Wahba (1999) and Smith and Todd (2003) about the relative accuracy of Propensity Score matching methods in observational studies\(^1\) and the availability of free user-friendly software routines\(^2\) have contributed to increase the diffusion of these techniques in the applied work of economists. Matching estimators are now easy to use and perhaps too many users adopt them without really checking that the conditions for their application are satisfied. In particular, the crucial identifying assumption of “unconfoundedness of assignment to treatment” is often accepted without any careful assessment of its plausibility.

Building on Rosenbaum and Rubin (1983b) and Rosenbaum (1987), in this paper we propose a sensitivity analysis for matching estimators aimed at assessing their robustness to failures of the unconfoundedness assumption in a specific evaluation problem. The intuition is simple. Suppose that unconfoundedness is not satisfied given observables but would be satisfied if we could observe an additional binary variable. This binary variable can be simulated in the data and used as an additional matching factor in combination with the preferred matching estimator. A comparison of the estimates obtained with and without matching on this simulated binary variable tells us to what extent the estimator is robust to this specific source of failure of the unconfoundedness assumption. Moreover, the simulated values of the binary


\(^2\)See, for instance, Abadie et al. (2001), Sianesi (2001), Becker and Ichino (2002).
variable can be constructed to capture different hypotheses on the nature of potential confounding factors. This is convenient and instructive particularly when the simulated confounder is designed to mimic the distribution of important observed covariates.

Similar types of sensitivity analysis have been proposed in the literature for other kinds of estimators. For example, Rosenbaum and Rubin (1983b) and recently Imbens (2003) propose a method to assess the sensitivity of Average Treatment Effect (ATE) estimates in parametric regression models. Altonji, Elder and Taber (2005) use a similar idea to assess how strong selection on unobservables would have to be in order to imply that the entire estimated effect should be attributed to selection bias. Also their result is however restricted to a specific parametric setup, i.e. the standard Heckman’s selection model based on the assumption of joint normality of the error terms in the selection and outcome equations. We contribute to this literature by extending this type of analysis to matching estimators of the Average effect of Treatment on the Treated (ATT). Like Rosenbaum (1987), but differently than the above literature, we do not have to necessarily rely on any parametric model. Moreover, and unlike Rosenbaum’s paper, we derive point estimates of the ATT under different possible scenarios of deviation from unconfoundedness.

After a general presentation of the proposed sensitivity analysis in Section 2, we describe in Section 3 how it can be applied using Italian data collected to evaluate the effect of a temporary work experience on future employment probabilities. Policy makers and labor market analysts are becoming increasingly concerned about the growth of temporary employment in Europe. It is typically feared that this growth will generate a “segmentation” of labor markets between a primary sector of protected permanent jobs and a
secondary sector of unprotected temporary work opportunities, in which a relevant fraction of predominantly young workers will remain trapped. Thus, establishing to what extent temporary work is indeed a “trap” or, on the contrary, a possible “springboard” for a transition to permanent employment is becoming a question of increasing importance.

The unavailability of a convincing source of exogenous variation of the selection into temporary jobs for the Italian case, imposes an evaluation strategy based on matching techniques. We perform this evaluation using data specifically collected for this purpose, consisting of the universe of Temporary Work Agency (TWA) workers who had an assignment during the first six months of 2001 in two Italian regions (the “treated” units) and of a sample of similar workers who, at the beginning of this period, were unemployed or employed with a non-permanent contract (the “control” units).

We show that the effect of treatment on the treated for a worker on a TWA assignment is to increase her probability of finding a permanent job 18 months later by 17 percentage points in Tuscany and by 8 percentage points in Sicily. The second effect is, however, only barely significant. These point estimates are large given that the observed probabilities of finding a permanent job in the treated group are respectively 31% and 23% in the two regions. Thus, it is possible to conclude that in Tuscany the probability of a transition to permanent employment for temporary workers would have been equal to a half in the counterfactual case of no assignment, while in Sicily it would have been equal to two thirds.

The sensitivity analysis complements in an important way these results by showing that in Tuscany, but not in Sicily, the estimated ATT is robust to deviations from unconfoundedness caused by binary unobservable factors with characteristics similar to those of observed variables like gender,
education, marital status and previous employment history. Only when the unobservable confounding factor is hypothesized to cause distortions that are significantly larger than the ones caused by observed covariates, the ATT for Tuscany is estimated to be close to zero. For Sicily, the ATT appears not to be robust even in the case of minor failures of unconfoundedness.

2 A framework to assess the sensitivity of matching estimators

2.1 Notation and identification

Consider a standard evaluation problem in which $Y_0$ is the outcome if the unit is exposed to treatment $T = 0$, and $Y_1$ is the outcome if the unit is exposed to treatment $T = 1$. Only one of these two potential outcomes can be observed (i.e., the one corresponding to the treatment the unit received), but the causal effect is defined by their comparison $Y_1 - Y_0$. Thus, causal inference becomes a problem of inference with missing data. We are interested in estimating the ATT, defined as:

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

The selection into treatment can be represented, without loss of generality, as a process of utility maximization:

$$V = f(Z, \epsilon_v) \quad T = I(V > 0)$$ (2)

where $Z$ and $\epsilon_v$ are observed and unobserved characteristics determining the choice, respectively. Analogously, the two potential outcomes can be written as functions of observed ($X$) and unobserved ($\epsilon_y$) pre-treatment variables:

$$Y_1 = g_1(X, \epsilon_y)$$ (3)
The two sets of variables \( X \) and \( Z \) may coincide or overlap to a certain extent.

The evaluation aim is to identify and consistently estimate the ATT. Problems may arise because of the potential association between some of the \( \epsilon_y \) and the treatment indicator \( T \), as determined by the observable and unobservable variables expressed in equation (2).

In this kind of situations, one of the assumptions that allow the identification of the ATT is “unconfoundedness” (Rosenbaum and Rubin, 1983a), which is a special case of ignorable missing mechanism and the rationale behind common estimation strategies such as regression modeling and matching. This assumption does not distinguish between \( X \) and \( Z \), but considers the whole conditioning set of pre-treatment variables \( W = (X, Z) \) and assumes that

\[
(Y_1, Y_0) \perp T \mid W \tag{5}
\]

and

\[
0 < Pr(T = 1 \mid W) < 1. \tag{6}
\]

This means that, conditioning on observed covariates \( W \), treatment assignment is independent of potential outcomes. In other words, the assignment to treatment is random within cells defined by the variables \( W \). Although very strong, the plausibility of this assumption heavily relies on the quality and amount of information contained in \( W \).

Under unconfoundedness, one can identify the average treatment effect on the treated as

\[
E(Y_1 - Y_0 \mid T = 1) = E(E(Y_1 - Y_0 \mid T = 1, W)) = \]

\[
= E(E(Y_1 \mid T = 1, W) - E(Y_0 \mid T = 0, W) \mid T = 1), \tag{7}
\]
where the outer expectation is over the distribution of $W$ in the sub-population of treated individuals.

Often the covariates $W$ are multivalued or continuous, so that some smoothing techniques are in order: under unconfoundedness, several estimation strategies can serve this purpose.\(^3\) One of these is regression modeling. Using regression to “adjust” or “control for” pre-intervention covariates is, in principle, a good strategy, although it has some pitfalls. For instance, if there are many covariates, it can be difficult to find an appropriate specification. Moreover, regression modeling obscures information on the distribution of covariates in the two treatment groups. In principle, one would like to compare individuals that have the same values of all covariates. Unless there is a substantial overlap of the two distributions of covariates, with a regression model one has to rely heavily on model specification (i.e., on extrapolation) for the estimation of treatment effects. It is thus crucial to check how much the two distributions overlap and what is their “region of common support”. When the number of covariates is large, this task is not an easy one. A possible solution is to reduce the problem to a single dimension, by using Propensity Score matching techniques as discussed in the next section.

### 2.2 Matching estimators of the ATT based on the Propensity Score

The Propensity Score is the individual probability of receiving the treatment given the observed covariates: $p(W) = P(T = 1|W)$. Under unconfoundedness, the following results hold (Rosenbaum and Rubin, 1983a): $T$ is independent of $W$ given $p(W)$, and $Y_0$ and $Y_1$ are independent of $T$ given $p(W)$. Note that the Propensity Score satisfies the so-called “balancing property”,

---

\(^3\)For a general review of nonparametric estimation methods of ATT’s under exogeneity, see Imbens (2004).
i.e., observations with the same value of the Score have the same distribution of observable (and possibly unobservable) characteristics irrespective of treatment status; moreover, the exposure to treatment or control status is random for a given value of the Score. These properties allow the use of the Propensity Score as a univariate summary of all $W$. It is enough to check the distribution of the Score in the two groups, and use the Score as the single covariate that needs to be adjusted for. In fact, adjusting for the Score automatically controls for all observed covariates, at least in big samples. As a result, if $p(W)$ is known, the ATT can be estimated as follows:

$$
\tau \equiv E(Y_1 - Y_0 | T = 1) = 
$$

$$
= E(E(Y_1 - Y_0 | p(W), T = 1)) = 
$$

$$
= E(E(Y_1 | p(W), T = 1) - E(Y_0 | p(W), T = 0) | T = 1)
$$

where the outer expectation is over the distribution of $(p(W)|T = 1)$.

Any standard probability model can be used to estimate the Propensity Score. For example, $Pr(T = 1|W) = F(h(W))$, where $F(.)$ is the normal or the logistic cumulative distribution and $h(W)$ is a function of the covariates with linear and higher order terms.\(^4\) Inasmuch as the specification of $h(W)$ which satisfies the balancing property is more parsimonious than the full set of interactions needed to match treated and control units according to observable characteristics, the Propensity Score reduces the dimensionality.

---

\(^4\)In a choice-based sampling scheme, which is the scheme of many empirical studies like our application described in Section 3, the odds ratio of the mis-specified (i.e., choice-based) Propensity Score can be used to implement matching as suggested by Heckman and Todd (1999). The mis-specified odds ratio is monotonically related to the odds ratio of the true Propensity Score, which is itself a monotonic transformation of the Propensity Score. Note, however, that unconfoundedness holds also in the choice-based sample, although the true Propensity Score cannot be consistently estimated. Hence, the choice-based Propensity score can still be used as a balancing score, in order to construct a control group with the same distribution of covariates as the treated group.
problem of matching procedures based on the multidimensional vector $W$.\textsuperscript{5}

The estimation of the Propensity Score is not enough to estimate the ATT of interest using equation (8). In fact, the probability of observing two units with exactly the same value of the Score is in principle zero, since $p(W)$ is a continuous variable. Various methods have been proposed in the literature to overcome this problem.\textsuperscript{6} Here we concentrate on the Nearest Neighbor matching estimator, which we describe formally below. However, the sensitivity analysis proposed in the next section can be implemented for any type of Propensity Score matching algorithm.

Let $D$ be the set of treated units and $C$ the set of control units, and $Y^D_h$ and $Y^C_k$ be the observed outcomes of the treated and control units, respectively. Denote by $C(h)$ the set of control units $k$ matched to the treated unit $h$ with an estimated value of the Propensity Score of $p_h$. Nearest Neighbor matching sets

$$C(h) = \{k \mid k = \text{arg min}_k \| p_h - p_k \| \}, \quad (9)$$

which is a singleton set unless there are multiple nearest neighbors. In practice, the case of multiple nearest neighbors should be rare, particularly if the set of observable characteristics $W$ contains continuous variables.

Denote the number of controls matched with observation $h \in D$ by $N^C_h$ and define the weights $\omega_{hk} = \frac{1}{N^C_h}$ if $k \in C(h)$ and $\omega_{hk} = 0$ otherwise. Then, the formula for the Nearest Neighbor estimator can be written as:

$$\tau^M = \frac{1}{N^D} \sum_{h \in T} \left[ Y^D_h - \sum_{k \in C(h)} \omega_{hk} Y^C_k \right] \quad (10)$$

\textsuperscript{5}It is important to note that the outcome plays no role in the algorithm for the estimation of the Propensity Score. This is equivalent, in this context, to what happens in controlled experiments, in which the design of the experiment has to be specified independently of the outcome.

\textsuperscript{6}See Becker and Ichino (2002) for a summary.
\[
= \frac{1}{ND} \sum_{h \in D} Y^D_h - \frac{1}{ND} \sum_{k \in C(h)} \omega_y Y^C_k
\]

where the weights \( \omega_k \) are defined by \( \omega_k = \Sigma_h \omega_y \). Standard errors can be derived both analytically and by bootstrapping.

### 2.3 Sensitivity analysis

The possibility to identify the ATT using the estimator described above relies crucially on the validity of the unconfoundedness assumption. Parametric selection models, which in principle permit a relaxation of this assumption, are formally identified thanks to other types of non-necessarily preferable hypotheses, as pointed out in several papers (e.g., Little, 1985; Copas and Li, 1997). If such parametric hypotheses cannot be accepted, assuming unconfoundedness is an alternative option, but it requires to check the robustness of results to departures from this assumption. This is the spirit of some methods proposed in the literature. Here, we briefly review them and present an alternative simulation-based method of sensitivity analysis for matching estimators of the ATT.

Rosenbaum and Rubin (1983b) propose to assess the sensitivity of the estimated causal effects (in particular, the Average Treatment Effect) with respect to assumptions about an unobserved binary covariate that is associated with both the treatment and the response. The unobservables are assumed to be summarized by a binary variable in order to simplify the analysis, although similar techniques could be used assuming some other distribution for the unobservables. The central assumption of their analysis is that the assignment to treatment is not unconfounded given the set of
observable variables $W$, i.e.,

$$Pr(T = 1|Y_0, Y_1, W) \neq Pr(T = 1|W)$$

(11)

but unconfoundedness holds given $W$ and an unobserved binary covariate $U$,

$$Pr(T = 1|Y_0, Y_1, W, U) = Pr(T = 1|W, U).$$

(12)

Given these assumptions, that are common to all the other methods of sensitivity presented below, Rosenbaum and Rubin (1983b) suggest to specify four (sets of) parameters that characterize the distribution of $U$ and the association of $U$ with $T$, $Y_1$ and $Y_0$ given observed covariates (or given strata defined by observed covariates). The unobservable $U$ is usually assumed to be independent of the observed covariates, i.e., $Pr(U = 1|W) = Pr(U = 1)$.

After this step, the full-likelihood for $T, Y_0, Y_1, U|W$ is derived and maximized, holding the sensitivity parameters as fixed known values. It is then possible to judge the sensitivity of inference conclusions with respect to certain plausible variations in the assumptions about the association of $U$ with $T$, $Y_0$ and $Y_1$. If conclusions are relatively insensitive over a range of plausible assumptions about $U$, then causal inference is more defensible.

Imbens (2003) applies the same method but expresses the sensitivity parameters in terms of partial R-squared, in order to ease the interpretation of results. Note, however, that the approach followed by these authors use a parametric model as the basis for the estimation of the average treatment effects: specifically, a normal model when the outcome is continuous as in Imbens (2003) and a logistic regression when the outcome is binary as in Rosenbaum and Rubin (1983b). Parametrization is instead not necessary in the sensitivity analysis that we propose in this paper.

Rosenbaum (1987) proposes to assess the sensitivity of significance levels and confidence intervals, rather than the sensitivity of point estimates.
The method involves only one sensitivity parameter (which represents the association of $T$ and $U$), instead of the four (sets of) sensitivity parameters specified in Rosenbaum and Rubin (1983b), so that the joint distribution of $T,Y_1,U|W$ and $T,Y_0,U|W$ is only partially specified. As a consequence, only bounds for significance levels and confidence intervals can be derived: a result that is more general but leads to more conservative statements about the sensitivity of an inference.

Our proposed method aims instead at assessing the sensitivity of point estimates (and specifically the sensitivity of ATT matching estimates). Like Rosenbaum (1987), we do not rely on any parametric model for the outcome, but, unlike his paper, we derive point estimates of the ATT under different possible scenarios of deviation from unconfoundedness. Instead of estimating by maximum likelihood a model for the outcome and the treatment status involving the confounding factor $U$, we impose the values of the parameters that characterize the distribution of $U$. Given these parameters, we then predict a value of the confounding factor for each treated and control subject and we re-estimate the ATT including the simulated $U$ in the set of matching variables. By changing the assumptions about the distribution of $U$, we can assess the robustness of the ATT with respect to different hypotheses on the nature of the confounding factor. Moreover, we can verify whether there exists a set of plausible assumptions on $U$ under which the estimated ATT is driven down to zero by the inclusion of $U$ in the matching set.

More formally, we consider for expositional simplicity the case of binary potential outcomes $Y_0, Y_1 \in \{0, 1\}$, as in the application discussed in Section 3, and we 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 exposure. Assuming that equations 11 and 12 are satisfied, we characterize the distribution of the unobserved binary confounding factor $U$ by specifying the parameters

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

with $i, j = \{0, 1\}$, which give the probability that $U = 1$ in each of the four groups defined by the treatment status and the outcome value. Note that $U$ is considered as any other observed characteristics in $W$; we are only assuming that the distribution of $U$ given $T$ and $Y$ does not vary with $W$, although this could be easily extended to allow for different patterns of association according to the value of observed covariates. In addition, note that the assumption that $Pr(U = 1|T = i, Y = j, W) = Pr(U = 1|T = i, Y = j)$ does not imply that $Pr(U = 1|W) = Pr(U = 1)$, i.e., in the population, $U$, as any other observed covariate, may be correlated with other variables. Of course, if $U$ were fully correlated with one or more of the variables in $W$, it would not represent a problem, but the insurgence of this extreme and uninteresting case can be verified empirically.

Using these parameters and the probability of a given outcome by treatment status $Pr(Y = i|T = j)$, which is observed in the data, we can compute the fraction of subjects with $U = 1$ by treatment status only:

$$p_i \equiv Pr(U = 1|T = i) = \sum_{j=0}^{1} p_{ij} \cdot Pr(Y = j|T = i).$$

Therefore, by setting the parameters $p_{ij}$ appropriately, we can generate a situation in which the fraction of subjects with $U = 1$ is greater among the treated ($p_1 > p_0$) or among the controls ($p_1 < p_0$).

---

7See below, at the end of this section, for a discussion of the more general case of multi-valued or continuous outcome.

8This is never the case in the application of Section 3.
To facilitate the interpretation of these parameters, it may be useful to note that if \( p_{11} > p_{00} \) then \( Pr(Y = 1|T = i, U = 1) > Pr(Y = 1|T = i, U = 0) \), which implies that for subjects exposed to treatment \( T = i \) the confounding factor \( U \) has a positive effect on the outcome. Moreover, if \( p_{1} > p_{0} \) then \( Pr(T = 1|U = 1) > Pr(T = 1|U = 0) \), which implies that \( U \) has a positive effect on the selection into treatment.\(^9\)

Given arbitrary (but meaningful) values of the parameters \( p_{ij} \), our sensitivity analysis proceeds by attributing a value of \( U \) to each subject, according to her belonging to one of the four groups defined by the treatment status and the outcome. We then treat \( U \) as any other observed covariate and, in particular, we include \( U \) in the set of matching variables used to estimate the Propensity Score and to compute the ATT according to the Nearest Neighbor estimator described in Section 2.2. Using a given set of values of the sensitivity parameters, we repeat the matching estimation many times (e.g. \( m = 10,000 \)) and obtain an estimate of the ATT, which is an average of the ATT’s over the distribution of the unobserved \( U \).

To compute a standard error for the proposed ATT estimator, we consider the problem of the unobserved confounding factor \( U \) as a problem of missing data that can be solved by multiply imputing the missing values of \( U \). The only difference between this approach and the standard multiple imputation procedure (see Rubin, 1987; Little and Rubin, 2002) is that here for any given set of values of the sensitivity parameters there is no uncertainty regarding the \( U \)’s data generating process. So, let \( m \) be the number of imputations (i.e., replications) of the missing \( U \)’s, and let \( \hat{ATT}_i \) and \( se_i^2 \) be the point estimate

---

\( \text{To see this note that, using Bayes’ rule and defining } \gamma = Pr(T = 1), Pr(T = 1|U = 1) > Pr(T = 1|U = 0) \text{ can be rewritten as: } \frac{p_{1} \gamma^p_{1} (1-\gamma)^{1-p_{1}}}{(1-p_{1}) \gamma^p_{0} (1-\gamma)^{1-p_{0}}}. \text{ After some simplifications, this condition becomes: } \frac{p_{1}(1-p_{0})}{(1-p_{1}) p_{0}}, \text{ which is always verified if } p_{1} > p_{0}. \text{ Similar steps apply to the case discussed immediately above concerning the effects of } T \text{ on the two outcomes.} \)
and the estimated variance of the ATT estimator at the $i$-th imputed data set, $i = 1, 2, \ldots, m$. The ATT estimate, $\hat{ATT}$, is then obtained (as already explained above) by the average of the $\hat{ATT}_i$'s over the imputations. Let the within-imputation variance be defined as

$$se^2_W = \frac{1}{m} \sum_{i=1}^{m} se^2_i \quad (15)$$

and the between-imputation variance as

$$se^2_B = \frac{1}{m-1} \sum_{i=1}^{m} (\hat{ATT}_i - \hat{ATT})^2. \quad (16)$$

Then, the total variance associated with $\hat{ATT}$ is given by (Rubin 1987):

$$T = se^2_W + (1 + \frac{1}{m})se^2_B. \quad (17)$$

For a large number of replications the statistics $(\hat{ATT} - ATT)T^{1/2}$ is approximately normal (Barnard e Rubin, 1999).

Thus, for any given configuration of the parameters $p_{ij}$, we can retrieve a point estimate of the ATT which is robust to the failure of the unconfoundedness assumption implied by that specific configuration.

Note that the method can be adapted to include multi-valued or continuous outcomes. Indeed, in such cases, it is possible to define:

$$p_{ij} \equiv Pr(U = 1|T = i, I(Y > y^*)) \quad (18)$$

where $I$ is the indicator function and $y^*$ is a chosen typical value, e.g. the median, of the distribution of $Y$. Once the parameters $p_{ij}$ are defined in this way, one can proceed as described above. Alternatively, at the cost of introducing a parametric specification, one can specify a model for the conditional distribution of $U$ given $T$ and $Y$. A possible choice is, for example,
the following:

\[ Pr(U = 1|T = i, Y = y) = \frac{\exp(\alpha_0 + \alpha_1 i + \beta_i y)}{1 + \exp(\alpha_0 + \alpha_1 i + \beta_i y)}, \]  

(19)

where \( \alpha_0, \alpha_1, \beta_0, \) and \( \beta_1 \) become the sensitivity parameters used to calibrate the distribution of \( U \).

Despite its simplicity, this approach has several advantages. First, note that the hypothesized associations of \( U \) with \( Y \) and \( T \) are stated in terms of proportions characterizing the distribution of \( U|T,Y,W \). This permits to avoid a possibly incorrect parametric specification of the distribution of \( Y|T,U,W \), which is instead the strategy adopted by competing types of sensitivity analysis. For example, Altonji et al. (2003) use a standard selection model as a benchmark for their sensitivity analysis. In this way they can obtain some analytical results at the cost of imposing a model that assumes a constant (in the logit scale) treatment effect and a single parameter (the correlation between the error terms in the selection and outcome equations) to characterize both the unobserved selection into treatment and its association with the outcome.

Second, the parameters \( p_{ij} \) and \( p_i \) can be chosen to make the distribution of \( U \) similar to the empirical distribution of observable binary covariates. In this case, the simulation exercise reveals the extent to which matching estimates are robust to deviations from unconfoundedness induced by the impossibility of observing factors similar to the ones used to calibrate the distribution of \( U \).

Third, one can search for the existence of a set of parameters \( p_{ij} \) and \( p_i \) such that if \( U \) were observed the estimated ATT would be driven to zero, and then assess the plausibility of this configuration of parameters. If all the configurations leading to such result could be considered very unlikely,
the exercise would support the validity of the estimates derived under the unconfoundedness assumption.

Finally, our simulation-based sensitivity analysis allows to assess the robustness of matching estimates of the ATT, irrespective of the specific algorithm used to match observations.

2.4 Sensitivity and bounds

The sensitivity analysis proposed in the previous section starts from a point-identifying assumption (unconfoundedness in our case) and then examines how results change as this assumption is weakened in specific ways. A complementary approach, proposed by Manski (1990), consists of dropping the unconfoundedness assumption entirely, and constructing bounds for the treatment effect that rely on either the outcome being bounded or on alternative identifying assumptions. It is useful to clarify with an example the relationship between these two approaches.

Consider the ATT defined as:

\[
ATT = E(Y_1|T = 1) - E(Y_0|T = 1).
\]  

(20)

Since \( Y_0 \) is not observed when \( T = 1 \), the term \( E(Y_0|T = 1) \) cannot be estimated from empirical evidence alone. Nevertheless if \( Y_0 \) is, for example, a binary variable assuming values 1 or 0 (or more generally a bounded variable), one can obtain non-parametric bounds for the ATT, substituting \( E(Y_0|T = 1) \) with its smallest and biggest possible values:

\[
E(Y_1|T = 1) - 1 \leq ATT \leq E(Y_1|T = 1).
\]  

(21)

These bounds can be estimated using sample analogs.

Our sensitivity analysis offers a way to understand what set of assumptions, concerning a potential confounder \( U \), would lead to an ATT equal to
the lower or the upper one of these non-parametric bounds.

It is easy to show that the lower bound is achieved when, among the treated, there are only individuals with $U = 1$, i.e., $Pr(U = 1|T = 1) = 1$, and among the controls all the individuals with $U = 1$ have $Y_0 = 1$, i.e., $Pr(Y_0 = 1|T = 0, U = 1) = 1$. This translates into the following assumptions on the parameters $p_{ij}$:

$$
\begin{align*}
p_{01} &= 1 \\
p_{11} &= 1 \\
p_{10} &= k > 0 \\
p_{00} &= 0.
\end{align*}
$$

The upper bound is instead achieved when, among the treated, there are only individuals with $U = 1$, i.e., $Pr(U = 1|T = 1) = 1$, and among the controls all the individuals with $U = 1$ have $Y_0 = 0$, i.e., $Pr(Y_0 = 1|T = 0, U = 1) = 0$. This translates into the following assumptions on the parameters $p_{ij}$:

$$
\begin{align*}
p_{01} &= 1 \\
p_{11} &= 1 \\
p_{10} &= 0 \\
p_{00} &= k > 0.
\end{align*}
$$

Note that these sets of circumstances are really extreme and thus seem highly implausible. This explains why non-parametric bounds are often uninformative in specific applications. This happens because there exist sets of values of the treatment effect which are within the bounds but correspond to scenarios that are very unlikely, despite being potentially possible. Thanks to reasonable assumptions on the association between confounding factors, treatment status and potential outcomes, a sensitivity analysis like
the one proposed in this paper offers the possibility to restrict the size of non-parametric bounds, eliminating possible but unlikely values of the ATT.

The next section describes an application of this sensitivity analysis to ATT estimates based on Italian data collected to evaluate the effect of a temporary work experience on future employment probabilities.

3 An Application to Temporary Work Agencies in Italy

3.1 The context

The growing share of temporary employment in many European countries raised concerns over the risk of an undesirable labor market “segmentation”. Several studies have indicated the existence of a gap in the working conditions of permanent and temporary employees, particularly in terms of wages and working rights.\textsuperscript{10} Triggered by this gap, public opinion and policy makers have stressed the importance of searching “an appropriate balance between flexibility and security” (European Commission, 2003). While such a balance may not be possible in a cross-sectional sense, it may become possible in an intertemporal sense if temporary jobs are an effective springboard toward permanent employment, as opposed to a trap of endless precariousness.

From a theoretical point of view, there may be two broad reasons why temporary employment could offer a springboard to a stable job. First, more able workers may use TWA jobs to signal their skills by making themselves available for screening. And, second, TWA jobs may provide an occasion to acquire additional human capital (general or specific), social contacts and information about permanent vacancies. Despite these two possibilities, how-

\textsuperscript{10}See the literature survey in Oecd (2002).
ever, temporary employment may also represent a trap of endless precariousness if it gives a “bad signal” of lack of alternatives or low human capital. There is no obvious reason to expect one or the other mechanism to prevail. In the Italian labor market, which is characterized by a high rigidity of standard jobs, firms appear to be interested in TWA employment both to screen workers and to deal with demand fluctuations. At the end of the day, whether TWA employment is a springboard or a trap is ultimately an empirical question.\textsuperscript{11}

In Italy this question originated a very harsh debate\textsuperscript{12} after the approval of the so called “Treu law” (law 196/1997), which legalized and regulated the supply of temporary workers by authorized agencies (which were illegal until then).\textsuperscript{13} After the introduction of this law, TWA employment has rapidly expanded, especially in the north of the country and in the manufacturing sector.\textsuperscript{14}

Despite this rapid expansion and the wide interest in the debate on TWA

---

\textsuperscript{11}Studies in other countries exist but display a wide set of results, depending on the institutional setting and the evaluation strategy. See, among others, Booth, Francesconi and Frank (2002), Guell and Petrongolo (2003), Malo and Munoz-Bullon (2002), Lechner et al. (2000), Zijl et al. (2002). The contribution by Autor and Houseman (2002) distinguishes itself by these observational studies, since it exploits a particular case of random assignment to TWA jobs for welfare clients.

\textsuperscript{12}A debate which has unfortunately degenerated in the terrorist attacks that killed Massimo D’Antona in 1999 and Marco Biagi in 2002, two labor law scholars and consultants of the Ministry of Welfare.

\textsuperscript{13}On the introduction of this kind of non-standard employment in the Italian labor law, see Ichino (2000). The Treu law (and subsequent modifications) states that TWA employment is allowed in all but the following cases: replacement of workers on strike; firms that experienced collective dismissals in the previous 12 months; jobs that require medical vigilance. The law does not set a maximum cumulated duration of TWA assignments or legal motivations for using them, leaving the provision of further regulation to collective bargaining. Collective agreements have typically stipulated that temporary workers cannot exceed 8-15% of normal employees (depending on the sector). Moreover they have constrained the set of acceptable motivations: peak activity; one-off work; need of skills not available within the firm. Firms cannot extend an individual TWA contract for more than four times or for a cumulated period longer than 24 months.

\textsuperscript{14}See Nannicini (2004).
jobs, no convincing evaluation study of their effects has yet been performed in Italy. Our paper is the first one trying to fill in this gap.\textsuperscript{15}

3.2 Data

We collected data for two Italian regions, Tuscany and Sicily, which were among the few remaining areas with incomplete penetration of TWAs in 2000. In these regions, we selected five provinces that already had a temporary work agency (Livorno, Pisa, Lucca, Catania, Palermo) and four that had none (Grosseto, Massa, Messina, Trapani) but were otherwise similar to the previous five in terms of a wide set of economic, demographic and geographic indicators. “Manpower Italia Spa”, a major company operating in the TWA sector with a national market share of about 25\%, gave us access to information about the workers in their files. From this dataset, we extracted workers who were on a TWA mission in one of the nine provinces mentioned above during the first semester of 2001. Hence, the first semester of 2001 was chosen as the “treatment” period, i.e., the period in which treated individuals had their TWA experience. Data collection developed along the following further steps: 1) phone interviews to all temps who were resident in the nine provinces and were in a TWA mission during the first semester of 2001; 2) phone interviews to a random sample of “controls” drawn from the population of the nine provinces, in order to match them with the treated units. Controls were chosen so as to have two characteristics: to be aged between 18 and 40 and not to have a stable job (either an open-ended contract or self-employment) on January 1, 2001. This first screening of potential control subjects might be interpreted as part of the matching strategy, aimed

\textsuperscript{15}This evaluation study is part of a project on TWA employment financed by the Italian Ministry of Welfare and the Tuscany Region. See the final report (Ichino, Mealli and Nannicini, 2004) for further details.
at identifying a common support for the treated and the control units with respect to observable characteristics.

In order to get a sufficient number of controls in each area, we stratified the sample according to the province of residence. Hence, our data collection strategy lead to both choice-based sampling and geographical stratification. Information on the period before January 1, 2001, provided “pre-treatment” variables for both the treated and the control observations, while information on the date of the interview (November 2002) provided “outcome” or “post-treatment” variables. For both the treated and the control units, interviews followed an identical path, asking: a) demographic characteristics; b) family background; c) educational achievements; d) work experience before the treatment period; e) job characteristics during the treatment period; f) work experience from the treatment period to the end of 2002; g) job characteristics at the end of 2002.

After a first analysis of the data, control subjects who were out of the labor force in the treatment period (e.g. students) were dropped from the sample. In fact, these subjects showed characteristics that made them not easily comparable with the treatment units. Notice that this was a conservative choice with respect to the estimated treatment effects, since all these individuals had a very low probability of having a permanent job at the end of 2002. Dropping these observations is another step of the search for a common support between treated and control units. As a result, the final data set used for the empirical evaluation contains treated and control units that

As mentioned in Section 2.2, the matching methods we use are not affected by this feature of the data collection strategy, as long as the distribution of the observed pre-treatment variables is guaranteed to be the same in the treatment and in the control group. It should be taken into account, moreover, that this data collection strategy combines flow sampling for the treated group and stock sampling for the control group. As long as the transition probability from permanent employment to unemployment or non-permanent employment is very low, there is no loss of relevant information from the control group.
are fairly similar, even before starting the identification of nearest neighbors.

To summarize, the treated sample contains individuals who lived in the nine provinces mentioned above and were on a TWA assignment through “Manpower” during the first semester of 2001; while the control sample contains residents in the same nine provinces, aged 18-40, who belonged to the labor force but did not have a stable job. The final dataset contains 2030 individuals: 511 treated and 1519 controls.

Tables 1 and 2 summarize the pre-treatment information available for the individuals in the sample on which Nearest Neighbor matching has been based. Table 1 shows the characteristics of all individuals, while Table 2 is based on information available only for those treated and control units who were employed in the pre-treatment period. These tables report also the average characteristics of an important subset of controls, the “matched controls”. These are the control units that have been used as “nearest neighbors” of at least one treated unit in the Nearest Neighbor matching estimation. Inasmuch as the treated units are more similar to the “matched controls” than to “all controls”, the matching strategy has succeeded in improving the quality of the comparison used to estimate the causal effect of interest. It is evident in these tables that for almost all variables an average treated unit is more similar to an average Nearest Neighbor than to an average control unit, and this is exactly what we would expect from a matching algorithm. However, it is also evident that a gap remains. For this reason, it is particularly important to complement the estimates based on Nearest Neighbor Matching

---

17 Notice that the control sample might include subjects who went through a TWA experience in a period different from the first semester of 2001. This is because the treatment coincides with “TWA assignment during the first semester of 2001”. If the outcome of some control units were affected by a TWA experience in another period, our exercise would result in conservative estimates.

with an appropriate assessment of their robustness with respect to violations of the unconfoundedness assumption. The sensitivity analysis presented in Section 2 has precisely this goal and we show in the next section what we can learn from it in this specific application.

3.3 The estimated causal effects and their sensitivity to unconfoundedness

Tables 3 and 4 display the results for Tuscany. For expositional simplicity, let us say that $U$ measures some unobservable component of ability, which for brevity we call “skill”. Each row of the first four columns of Table 3 contains the four probabilities $p_{ij} = Pr(U = 1|T = i, Y = j)$, with $i, j = \{0, 1\}$, which characterize the binary distribution of skill, by treatment status and outcome, under which the ATT has been estimated. The first subscript denotes the treatment status while the second refers to the outcome status. So, for example, $p_{11}$ indicates the fraction of skilled subjects among those who are treated and find a permanent job after treatment, while $p_{10}$ is the fraction of skilled subjects among those who receive treatment but do not find a job. As explained in Section 2.3, the probabilities $p_{ij}$ are the parameters that we set to characterize the distribution of the confounding factor. Together with the observed probability of a given outcome by treatment status, $Pr(Y = i|T = j)$, they allow to characterize the distribution of skill by treatment status only. This is done in the next two columns, which display $p_{1.} = Pr(U = 1|T = 1)$, i.e. the fraction of skilled subjects among the treated, and $p_{0.} = Pr(U = 1|T = 0)$, i.e. the fraction of the skilled among controls.

The first row of Table 3 describes the baseline situation in which $p_{ij} = 0$ for all $i$ and $j$ and the assumption of unconfoundedness is satisfied because there is no confounding factor. The ATT is estimated to be equal to 0.17,
with a standard error of 0.06. Since the observed probability of transition to a permanent job for the treated subjects in Tuscany is 0.31, the estimated ATT suggests that this probability would be half as large in the counterfactual case of no assignment to treatment.

Exploring the robustness of this estimate with respect to failures of the unconfoundedness assumption is the goal of our sensitivity analysis. The amount of information contained in the set of matching variables on which this estimate is based (see Section 3.2) is fairly large. Specifically it includes variables that are rarely available in standard datasets (e.g., father’s education and job status, final grade in the higher educational degree and complete previous employment history) and that are typically considered crucial for the validity of the unconfoundedness assumption in this type of applications. Nevertheless, despite this rich set of matching variables, it remains questionable whether unconfoundedness is effectively satisfied.

The other rows of Table 3 show how the baseline estimate changes when the binary confounding factor $U$ is calibrated to mimic different observable covariates and is then included in the set of matching variables. The first case sets the distribution of $U$ to be similar to the distribution of gender. Thus, in this example, given that 54% of the subjects who are exposed to treatment and find a job are males, by setting $p_{11} = 0.54$ we impose that an identical fraction of subjects (not necessarily all males) are skilled and therefore are assigned a value of $U$ equal to 1. An analogous interpretation holds for the other probabilities $p_{ij}$ in this row. Note that these assumptions imply that the treated are more skilled than the controls, independently of the outcome, as indicated by the fact that $p_{11} = 0.56$ while $p_{01} = 0.29$. In other words, as stated in Section 2.3, we are assuming that “skilled” subjects (as males) are more likely to be exposed to treatment. Under a deviation from
unconfoundedness with these characteristics, the ATT is estimated equal to 0.14. This estimate differs by only three percentage points with respect to the baseline estimate obtained in the absence of confounding effects, and remains statistically significant even when the standard error is calculated using the more stringent multiple imputation formula in equation 17. The other rows assume that the distribution of $U$ is in turn comparable to the distribution of observable variables like marital status, high school degree, college degree, existence of a previous work experience and high education of the father. In none of these cases, the ATT is estimated to differ by more than 3 percentage points from the baseline estimate, which therefore appears to be fairly robust to these deviations from the assumption of unconfoundedness.

The advantage of such a simulation exercise goes beyond this finding, because it allows us to explore the characteristics of the confounding factor $U$ under which the point estimate of the ATT becomes close to zero. This is done in Table 4 with respect to a grid of values of the derived parameters $d_1 = p_{11} - p_{10}$, $d_0 = p_{01} - p_{00}$ and $s = p_1 - p_0$. The set of interesting situations can be analyzed by looking at the cases in which $d_1 > 0$ (positive effect of $U$ on the outcome for treated units), $d_0 > 0$ (positive effect of $U$ on the outcome for control units) and $s > 0$ (positive effect of $U$ on selection). Thus, the rows of Table 4 describe a grid of cases in which $d_1 \geq d_0 > 0$ where the first weak inequality is meant to capture the plausible situation in which the (positive) outcome effect of $U$ is stronger for treated subjects (or at most equal). Along each row of the table, higher values of $s$ indicate, instead, increasingly strong effects of $U$ on the chances that a subject is selected into treatment. As a reference, the corresponding non-parametric Manski bounds are reported in the footnote to the table.

In the first panel of the table the outcome effect for control subjects is
kept constant at the low level $d_0 = 0.1$. The outcome effect for treated subjects changes instead, between rows, from $d_1 = 0.1$ to $d_1 = 0.7$. In these cases, the fraction of skilled subjects among the treated can be up to thirty percentage points higher than the fraction of skilled subjects among controls ($0.1 \leq s \leq 0.3$) without bringing the ATT down to a level close to zero. In all these cases the point estimate obtained when $U$ is included in the matching set is never smaller than 0.10. In the other panels of the table the outcome effect on control subjects is set to higher values, and in these cases smaller effects of $U$ on selection are sufficient to kill the ATT, i.e., to explain almost entirely the positive baseline estimate of the ATT.

A comparison with the results of Table 3 reveals that the cases in which $U$ is calibrated to match observed characteristics of subjects, correspond to cells close to the top row of Table 4 with $s$ smaller than 0.3. To be precise, all the observed characteristics considered in Table 3 imply outcome effects smaller than 0.1 for both the treated and the control subjects, while marital status is characterized by the greatest effect on selection into treatment: $s = 0.24$. Thus, the comparison between the two tables suggests that even if the unobserved confounding factor had outcome effects on the treated substantially larger than the ones of observed covariates, it would not cause much damage to the estimated ATT. There is instead less room for stronger effects on the side of selection into treatment, since the omission of a variable with an effect similar to the one of marital status would bring the estimate close to the problematic region of “zero ATT”.

While for Tuscany the sensitivity analysis that we have just described conveys an image of robustness of our matching estimator with respect to reasonable failures of unconfoundedness, a quite different picture emerges in Tables 5 and 6 for Sicily. The baseline estimate displays an ATT equal to 8
percentage points, which is barely significant at standard levels. Considering this point estimate, since in Sicily the observed probability of finding a permanent job for the treated is 0.23, the estimated “counterfactual” probability of the same outcome in the case of no assignment is 0.15. However, the sensitivity analysis shows that, for configurations of the parameters that mimic the distribution of important covariates, the baseline point estimate is always killed by the inclusion of \( U \) in the matching set. For instance, in the first simulation, the presence of a confounding factor distributed as gender across the four groups brings the ATT down to 0.01 (s.e. = 0.07).

This result is not simply due to the fact that in Sicily observable covariates have a stronger association with selection into treatment, leading us to generate confounding factors \( U \) that are more influential than the ones considered in the case of Tuscany. In fact, in Table 6, the same grid of simulations performed in Table 4 for Tuscany shows a very different picture. As soon as \( U \) is allowed to have an effect on selection into treatment such that \( s = 0.1 \), the estimated ATT is halved, indicating that at least half of the baseline estimate is explained by the presence of confounding factors with this characteristic. Moreover, when \( U \) is calibrated to have increasingly stronger effects on outcomes and selection into treatment, the estimated ATT approaches zero at a very rapid pace.

To sum up, while in the absence of the sensitivity analysis one could have argued in favor of a positive effect of TWA employment also in Sicily, the simulations described in Tables 5 and 6 reveal that the estimates are clearly not robust to even minor deviations from the assumption of unconfoundedness. This finding has a possible explanation, since in this region the public sector is the primary source of stable positions and this sector has only recently started to recruit through Temporary Work Agencies. In Tuscany, on
the contrary, the private sector is able to create a relevant number of stable positions for which a temporary job experience appears to be an effective springboard.\textsuperscript{19}

4 Conclusions

Matching estimators are used with increasing frequency often without checking that the conditions for their application are satisfied. In this paper we propose a sensitivity analysis for matching estimators aimed at assessing to what extent the unconfoundedness assumption is plausible in a specific evaluation problem.

Our approach is easy to implement and has several advantages. First, the sensitivity analysis can be designed to shed light on the robustness of the estimates with respect to deviations from unconfoundedness similar to the ones that would occur if observed factors were omitted from the matching set. Second, the configuration of sensitivity parameters may be varied so as to search for characteristics of the confounding factor under which the point estimate of the ATT becomes close to zero. Third, the sensitivity analysis does not involve the specification of parametric models and it can assess the robustness of matching estimates irrespective of the algorithm used to match observations.

We describe an application of this approach using data collected in Italy to evaluate the effect of Temporary Work Agency jobs on future employment opportunities. The sensitivity analysis proves to be helpful in identifying the cases in which the estimated effect should be considered not robust to failures of unconfoundedness.

\textsuperscript{19}\textit{See Ichino, Mealli and Nannicini (2004) for further details on these results and on their interpretation.}
References


European Commission, Employment in Europe 2003, DG for Employment and Social Affairs, Brussels.


Heckman J. and Todd P. (1999), Adapting Propensity Score Matching and Selection Models to Choice-based Samples, mimeo, University of Chicago.


Sianesi B. (2001), *Propensity score matching*, United Kingdom Stata Users’ Group Meeting 2001


Table 1: Characteristics of the whole sample

<table>
<thead>
<tr>
<th></th>
<th>TUSCANY</th>
<th></th>
<th>Sicily</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Treated</td>
<td>Matched</td>
<td>All</td>
<td>Treated</td>
</tr>
<tr>
<td>Age</td>
<td>26.5</td>
<td>27.5</td>
<td>29.1</td>
<td>26.8</td>
</tr>
<tr>
<td>Male</td>
<td>0.56</td>
<td>0.41</td>
<td>0.29</td>
<td>0.67</td>
</tr>
<tr>
<td>Single</td>
<td>0.90</td>
<td>0.87</td>
<td>0.66</td>
<td>0.83</td>
</tr>
<tr>
<td>Children</td>
<td>0.09</td>
<td>0.16</td>
<td>0.45</td>
<td>0.20</td>
</tr>
<tr>
<td>Distance</td>
<td>12.2</td>
<td>16.8</td>
<td>25.0</td>
<td>24.4</td>
</tr>
<tr>
<td>Father school</td>
<td>9.3</td>
<td>9.2</td>
<td>8.6</td>
<td>8.7</td>
</tr>
<tr>
<td>Father blue</td>
<td>0.33</td>
<td>0.39</td>
<td>0.43</td>
<td>0.30</td>
</tr>
<tr>
<td>Father active</td>
<td>0.53</td>
<td>0.46</td>
<td>0.37</td>
<td>0.46</td>
</tr>
<tr>
<td>School</td>
<td>12.5</td>
<td>12.7</td>
<td>12.3</td>
<td>12.0</td>
</tr>
<tr>
<td>Grade</td>
<td>75.9</td>
<td>77.1</td>
<td>76.9</td>
<td>74.7</td>
</tr>
<tr>
<td>Training</td>
<td>0.32</td>
<td>0.30</td>
<td>0.28</td>
<td>0.42</td>
</tr>
<tr>
<td>Unemployment</td>
<td>0.38</td>
<td>0.42</td>
<td>0.48</td>
<td>0.42</td>
</tr>
<tr>
<td>Employed 2000</td>
<td>0.35</td>
<td>0.36</td>
<td>0.42</td>
<td>0.34</td>
</tr>
<tr>
<td>Unemployed 2000</td>
<td>0.52</td>
<td>0.53</td>
<td>0.52</td>
<td>0.60</td>
</tr>
<tr>
<td>Out l.force 2000</td>
<td>0.13</td>
<td>0.10</td>
<td>0.05</td>
<td>0.06</td>
</tr>
<tr>
<td>Employed 2001</td>
<td>1.00</td>
<td>0.36</td>
<td>0.36</td>
<td>1.00</td>
</tr>
<tr>
<td>Unemployed 2001</td>
<td>0.00</td>
<td>0.64</td>
<td>0.64</td>
<td>0.00</td>
</tr>
<tr>
<td>Permanent 2002</td>
<td>0.31</td>
<td>0.16</td>
<td>0.17</td>
<td>0.23</td>
</tr>
<tr>
<td>Atypical 2002</td>
<td>0.42</td>
<td>0.36</td>
<td>0.31</td>
<td>0.39</td>
</tr>
<tr>
<td>Unemployed 2002</td>
<td>0.16</td>
<td>0.44</td>
<td>0.45</td>
<td>0.30</td>
</tr>
<tr>
<td>Out l.force 2002</td>
<td>0.11</td>
<td>0.04</td>
<td>0.07</td>
<td>0.07</td>
</tr>
<tr>
<td>N.individuals</td>
<td>281</td>
<td>135</td>
<td>628</td>
<td>230</td>
</tr>
</tbody>
</table>

Note: All variables except age, number of children, distance from the nearest agency (expressed in Km), father’s years of schooling, grade (expressed as a fraction of the highest mark), years of schooling and unemployment period (expressed as a fraction of the transition from school to work) are dummies. “Matched controls” are individuals who belong to the control sample and are used in the Nearest Neighbor Propensity Score matching estimation.
Table 2: Characteristics of subjects who were employed before the treatment

<table>
<thead>
<tr>
<th></th>
<th>TUSCANY Treated</th>
<th></th>
<th>TUSCANY Matched Controls</th>
<th></th>
<th>SICILY Treated Controls</th>
<th></th>
<th>SICILY Matched Controls</th>
<th></th>
<th>SICILY All Controls</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>All</td>
<td></td>
<td></td>
<td>All</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Permanent</td>
<td>0.16</td>
<td>0.26</td>
<td>0.14</td>
<td>0.16</td>
<td>0.36</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Atypical</td>
<td>0.84</td>
<td>0.74</td>
<td>0.86</td>
<td>0.84</td>
<td>0.64</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Blue-collar</td>
<td>0.62</td>
<td>0.39</td>
<td>0.44</td>
<td>0.24</td>
<td>0.22</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White-collar</td>
<td>0.36</td>
<td>0.54</td>
<td>0.54</td>
<td>0.71</td>
<td>0.67</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Self-empl.</td>
<td>0.02</td>
<td>0.07</td>
<td>0.01</td>
<td>0.04</td>
<td>0.10</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Manufact.</td>
<td>0.53</td>
<td>0.23</td>
<td>0.39</td>
<td>0.20</td>
<td>0.15</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Service</td>
<td>0.39</td>
<td>0.67</td>
<td>0.49</td>
<td>0.67</td>
<td>0.70</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>0.08</td>
<td>0.11</td>
<td>0.11</td>
<td>0.13</td>
<td>0.15</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wage</td>
<td>5.2</td>
<td>6.8</td>
<td>5.6</td>
<td>7.6</td>
<td>7.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hours</td>
<td>38.0</td>
<td>33.3</td>
<td>34.5</td>
<td>32.1</td>
<td>31.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N.individuals</td>
<td>98</td>
<td>266</td>
<td>79</td>
<td>45</td>
<td>267</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: All variables, except the hourly wage (expressed in Euros) and the weekly hours of work, are dummies. “Matched controls” are individuals who belong to the control sample and are used in the Nearest Neighbor Propensity Score matching estimation.
Table 3: Sensitivity analysis in Tuscany: effect of “calibrated” confounders

<table>
<thead>
<tr>
<th>U like:</th>
<th>Fraction $U = 1$ by treatment/outcome</th>
<th>Fraction $U = 1$ by treatment</th>
<th>ATT</th>
<th>s.e.</th>
<th>$\Pi$</th>
</tr>
</thead>
<tbody>
<tr>
<td>No confounder</td>
<td>$p_{11}$ $p_{10}$ $p_{01}$ $p_{00}$</td>
<td>$p_1$ $p_0$</td>
<td>0.17</td>
<td>0.06</td>
<td>-</td>
</tr>
<tr>
<td>Male</td>
<td>0.54 0.56 0.32 0.28</td>
<td>0.56 0.29</td>
<td>0.14</td>
<td>0.07</td>
<td>0.70</td>
</tr>
<tr>
<td>Single</td>
<td>0.86 0.92 0.76 0.64</td>
<td>0.90 0.66</td>
<td>0.14</td>
<td>0.07</td>
<td>0.68</td>
</tr>
<tr>
<td>High school</td>
<td>0.75 0.74 0.69 0.71</td>
<td>0.74 0.71</td>
<td>0.16</td>
<td>0.06</td>
<td>0.92</td>
</tr>
<tr>
<td>College</td>
<td>0.14 0.12 0.13 0.19</td>
<td>0.12 0.18</td>
<td>0.15</td>
<td>0.06</td>
<td>0.89</td>
</tr>
<tr>
<td>Prev.employed</td>
<td>0.40 0.33 0.50 0.41</td>
<td>0.35 0.42</td>
<td>0.16</td>
<td>0.06</td>
<td>0.91</td>
</tr>
<tr>
<td>Father educ.</td>
<td>0.34 0.31 0.32 0.27</td>
<td>0.32 0.28</td>
<td>0.16</td>
<td>0.07</td>
<td>0.89</td>
</tr>
</tbody>
</table>

Note: Let $U$ be a binary confounding factor and denote the fraction of $U = 1$ by treatment and outcome as: $p_{ij} = Pr(U = 1|T = i, Y = j)$, with $i, j = \{0, 1\}$. Once these four probabilities are imposed, since $Pr(Y = j|T = i)$ is observed, we can compute: $p_i = Pr(U = 1|T = i) = \sum_{j=0}^{1} p_{ij} \cdot Pr(Y = j|T = i)$. On the basis of these parameters, a value of $U$ is imputed to each individual and the ATT is estimated by Nearest Neighbor Propensity Score matching with $U$ in the set of matching variables. The process is repeated 10,000 times. “ATT” is the average of the simulated ATT’s; “s.e.” is the standard error (calculated as shown in equation 17); $\Pi$ is the percentage of significant ATT (at 5% level) in the 10,000 iterations. The first row shows the ATT estimate with no confounding factor. The “U like” rows show simulations in which $U$ has been calibrated to match the distribution of the corresponding variable.
Table 4: Sensitivity analysis in Tuscany: characterizing “killer” confounders

<table>
<thead>
<tr>
<th></th>
<th>s=0.1</th>
<th>s=0.2</th>
<th>s=0.3</th>
<th>s=0.4</th>
<th>s=0.5</th>
<th>s=0.6</th>
<th>s=0.7</th>
</tr>
</thead>
<tbody>
<tr>
<td>$d_1$=0.1 $d_0$=0.1</td>
<td>0.12</td>
<td>0.11</td>
<td>0.10</td>
<td>0.09</td>
<td>0.08</td>
<td>0.06</td>
<td>0.05</td>
</tr>
<tr>
<td>$d_1$=0.3 $d_0$=0.1</td>
<td>0.12</td>
<td>0.12</td>
<td>0.10</td>
<td>0.11</td>
<td>0.07</td>
<td>0.07</td>
<td>0.04</td>
</tr>
<tr>
<td>$d_1$=0.5 $d_0$=0.1</td>
<td>0.12</td>
<td>0.12</td>
<td>0.10</td>
<td>0.10</td>
<td>0.08</td>
<td>0.02</td>
<td>-0.01</td>
</tr>
<tr>
<td>$d_1$=0.7 $d_0$=0.1</td>
<td>0.13</td>
<td>0.12</td>
<td>0.10</td>
<td>0.05</td>
<td>-0.03</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$d_1$=0.3 $d_0$=0.3</td>
<td>0.09</td>
<td>0.02</td>
<td>0.04</td>
<td>-0.01</td>
<td>0.00</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$d_1$=0.5 $d_0$=0.3</td>
<td>0.11</td>
<td>0.07</td>
<td>0.07</td>
<td>0.02</td>
<td>-0.01</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$d_1$=0.7 $d_0$=0.3</td>
<td>0.10</td>
<td>0.08</td>
<td>0.00</td>
<td>-0.04</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$d_1$=0.5 $d_0$=0.5</td>
<td>0.07</td>
<td>0.01</td>
<td>0.01</td>
<td>-0.03</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$d_1$=0.7 $d_0$=0.5</td>
<td>0.09</td>
<td>0.04</td>
<td>-0.02</td>
<td>-0.02</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$d_1$=0.7 $d_0$=0.7</td>
<td>0.06</td>
<td>-0.02</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

The difference $d_1 = p_{11} - p_{10}$ captures the outcome effect of $U$ for the treated. The difference $d_0 = p_{01} - p_{00}$ captures the outcome effect of $U$ for the controls. The difference $s = p_1 - p_0$ captures the effect of $U$ on the selection into treatment. For further details on the interpretation of these parameters see Table 77 and the corresponding text. All ATT’s are averaged over 10,000 iterations. The baseline estimate without confounder is equal to 0.17 in Tuscany. The Manski bounds are (-0.69, 0.31).
Table 5: Sensitivity analysis in Sicily: effect of “calibrated” confounders

<table>
<thead>
<tr>
<th></th>
<th>No confounder</th>
<th>Male</th>
<th>Single</th>
<th>High school</th>
<th>College</th>
<th>Prev. employed</th>
<th>Father educ.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fraction $U = 1$ by treatment/outcome</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$p_{11}$</td>
<td>0.87</td>
<td>0.87</td>
<td>0.74</td>
<td>0.11</td>
<td>0.11</td>
<td>0.98</td>
<td>0.24</td>
</tr>
<tr>
<td>$p_{10}$</td>
<td>0.61</td>
<td>0.82</td>
<td>0.72</td>
<td>0.06</td>
<td>0.06</td>
<td>0.30</td>
<td>0.25</td>
</tr>
<tr>
<td>$p_{01}$</td>
<td>0.45</td>
<td>0.59</td>
<td>0.76</td>
<td>0.16</td>
<td>0.16</td>
<td>0.62</td>
<td>0.27</td>
</tr>
<tr>
<td>$p_{00}$</td>
<td>0.27</td>
<td>0.47</td>
<td>0.63</td>
<td>0.25</td>
<td>0.25</td>
<td>0.25</td>
<td>0.18</td>
</tr>
<tr>
<td>$p_{1.}$</td>
<td>0.67</td>
<td>0.83</td>
<td>0.72</td>
<td>0.46</td>
<td>0.46</td>
<td>0.46</td>
<td>0.25</td>
</tr>
<tr>
<td>$p_{0.}$</td>
<td>0.29</td>
<td>0.49</td>
<td>0.64</td>
<td>0.64</td>
<td>0.30</td>
<td>0.30</td>
<td>0.19</td>
</tr>
<tr>
<td>ATT</td>
<td>0.01</td>
<td>0.03</td>
<td>0.06</td>
<td>0.06</td>
<td>0.06</td>
<td>0.02</td>
<td>0.06</td>
</tr>
<tr>
<td>s.e.</td>
<td>0.07</td>
<td>0.07</td>
<td>0.06</td>
<td>0.06</td>
<td>0.06</td>
<td>0.06</td>
<td>0.06</td>
</tr>
<tr>
<td>Π</td>
<td>0.13</td>
<td>0.14</td>
<td>0.22</td>
<td>0.26</td>
<td>0.14</td>
<td>0.14</td>
<td>0.21</td>
</tr>
</tbody>
</table>

Let $U$ be a binary confounding factor and denote the fraction of $U = 1$ by treatment and outcome as: $p_{ij} = \Pr(U = 1|T = i, Y = j)$, with $i, j = \{0, 1\}$. Once these four probabilities are imposed, since $\Pr(Y = j|T = i)$ is observed, we can compute: $p_{i.} = \Pr(U = 1|T = i) = \sum_{j=0}^{1} p_{ij} \cdot \Pr(Y = j|T = i)$. On the basis of these parameters, a value of $U$ is imputed to each individual and the ATT is estimated by Nearest Neighbor Propensity Score matching with $U$ in the set of matching variables. The process is repeated 10,000 times. “ATT” is the average of the simulated ATT’s; “s.e.” is the standard error (calculated as shown in equation 17); Π is the percentage of significant ATT (at 5% level) in the 10,000 iterations. The first row shows the ATT estimate with no confounding factor. The “$U$ like” rows show simulations in which $U$ has been calibrated to match the distribution of the corresponding variable.
Table 6: Sensitivity analysis in Sicily: characterizing “killer” confounders

<table>
<thead>
<tr>
<th></th>
<th>s=0.1</th>
<th>s=0.2</th>
<th>s=0.3</th>
<th>s=0.4</th>
<th>s=0.5</th>
<th>s=0.6</th>
<th>s=0.7</th>
</tr>
</thead>
<tbody>
<tr>
<td>d₁=0.1 d₀=0.1</td>
<td>0.04</td>
<td>0.03</td>
<td>0.01</td>
<td>0.00</td>
<td>-0.03</td>
<td>-0.05</td>
<td>-0.05</td>
</tr>
<tr>
<td>d₁=0.3 d₀=0.1</td>
<td>0.04</td>
<td>0.03</td>
<td>0.02</td>
<td>0.01</td>
<td>-0.02</td>
<td>-0.03</td>
<td>-0.05</td>
</tr>
<tr>
<td>d₁=0.5 d₀=0.1</td>
<td>0.05</td>
<td>0.03</td>
<td>0.01</td>
<td>0.01</td>
<td>-0.01</td>
<td>-0.05</td>
<td>-0.08</td>
</tr>
<tr>
<td>d₁=0.7 d₀=0.1</td>
<td>0.06</td>
<td>0.04</td>
<td>0.02</td>
<td>-0.02</td>
<td>-0.15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>d₁=0.3 d₀=0.3</td>
<td>0.03</td>
<td>-0.01</td>
<td>-0.02</td>
<td>-0.10</td>
<td>-0.09</td>
<td></td>
<td></td>
</tr>
<tr>
<td>d₁=0.5 d₀=0.3</td>
<td>0.05</td>
<td>0.01</td>
<td>0.00</td>
<td>-0.04</td>
<td>-0.08</td>
<td></td>
<td></td>
</tr>
<tr>
<td>d₁=0.7 d₀=0.3</td>
<td>0.06</td>
<td>0.03</td>
<td>-0.03</td>
<td>-0.06</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>d₁=0.5 d₀=0.5</td>
<td>0.04</td>
<td>-0.01</td>
<td>-0.03</td>
<td>-0.07</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>d₁=0.7 d₀=0.5</td>
<td>0.06</td>
<td>0.01</td>
<td>-0.04</td>
<td>-0.08</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>d₁=0.7 d₀=0.7</td>
<td>0.04</td>
<td>-0.01</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The difference $d₁ = p_{11} - p_{10}$ captures the outcome effect of $U$ for the treated. The difference $d₀ = p_{01} - p_{00}$ captures the outcome effect of $U$ for the controls. The difference $s = p_{1} - p_{0}$ captures the effect of $U$ on the selection into treatment. For further details on the interpretation of these parameters see Table ?? and the corresponding text. All ATT’s are averaged over 10,000 iterations. The baseline estimate without confounder is equal to 0.08 in Sicily. The Manski bounds are (-0.76, 0.24).