DISCUSSION PAPER SERIES

IZA DP No. 2542

mhbounds - Sensitivity Analysis for Average Treatment Effects

Sascha O. Becker
Marco Caliendo

January 2007

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

mhbounds – Sensitivity Analysis for Average Treatment Effects

Sascha O. Becker
CES, University of Munich and IZA

Marco Caliendo
DIW Berlin and IZA

Discussion Paper No. 2542
January 2007

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
E-mail: iza@iza.org

IZA Discussion Paper No. 2542
January 2007

ABSTRACT

mhbounds – Sensitivity Analysis for Average Treatment Effects

Matching has become a popular approach to estimate average treatment effects. It is based on the conditional independence or unconfoundedness assumption. Checking the sensitivity of the estimated results with respect to deviations from this identifying assumption has become an increasingly important topic in the applied evaluation literature. If there are unobserved variables which affect assignment into treatment and the outcome variable simultaneously, *a hidden bias* might arise to which matching estimators are not robust. We address this problem with the bounding approach proposed by Rosenbaum (2002), where mhbounds allows the researcher to determine how strongly an unmeasured variable must influence the selection process in order to undermine the implications of the matching analysis.

JEL Classification: C40

Keywords: matching, treatment effects, sensitivity analysis, unobserved heterogeneity

Corresponding author:

Marco Caliendo
Department of Public Economics
DIW Berlin
Königin-Luise-Str. 5
14195 Berlin
Germany
E-mail: mcaliendo@diw.de

1 Introduction

Matching has become a popular method to estimate average treatment effects. It is based on the conditional independence or unconfoundedness assumption which states that the researcher should observe all variables simultaneously influencing the participation decision and outcome variables. Clearly, this is a strong identifying assumption and has to be justified case-by-case.[1] Hence, checking the sensitivity of the estimated results with respect to deviations from this identifying assumption becomes an increasingly important topic in the applied evaluation literature.

If there are unobserved variables which simultaneously affect assignment into treatment and the outcome variable, a 'hidden bias' might arise to which matching estimators are not robust (Rosenbaum, 2002). Since it is not possible to estimate the magnitude of selection bias with non-experimental data, we address this problem with the bounding approach proposed by Rosenbaum (2002).[2] The basic question to be answered is whether or not inference about treatment effects may be altered by unobserved factors. In other words, one wants to determine how strongly an unmeasured variable must influence the selection process in order to undermine the implications of the matching analysis. It should be noted that the bounding approach does not test the unconfoundedness assumption itself, because this would amount to testing that there are no (unobserved) variables that influence the selection into treatment. Instead, Rosenbaum bounds provide evidence on the degree to which any significance results hinge on this untestable assumption. Clearly, if the results turn out to be very sensitive, the researcher might have to think about the validity of his/her identifying assumption and consider alternative estimation strategies. DiPrete and Gangl (2004) provide an ado-file (`rbounds`) which allows the researcher to test sensitivity for continuous outcome variables, whereas our module `mhbounds` focusses on the case of binary outcome variables (e.g. employment vs. unemployment), which are frequently used in the evaluation literature.[3] Recent applications of this approach can be found in Aakvik (2001) or Caliendo, Hujer and Thomsen (2005). We outline this approach briefly in Section 2, an extensive discussion can be found in Rosenbaum (2002) and Aakvik (2001). Section 3 presents the syntax and Section 4 the options of `mhbounds`. Finally, in Section 5 we illustrate the module with some examples. It should be noted, that the aim of this paper is not to present or discuss the estimation of treatment effects with matching estimators. Instead we assume that the reader is familiar with this literature. Good overviews can be found in Heckman, Ichimura, Smith and Todd (1998), Imbens (2004) or Smith and Todd (2005). Stata programs to estimate treatments effects are provided by Becker and Ichino (`att*`, 2002), Leuven and Sianesi (`psmatch2`, 2003) and Abadie et al. (`nnmatch`, 2004).

1. Caliendo and Kopeinig (2006) provide a survey of the necessary steps when implementing (propensity score) matching methods.

2. See the paper by Ichino, Mealli, and Nannicini (2006) for a related approach and the ado-package `sensatt` by Nannicini (2006) for an implementation in Stata.

3. Clearly, `mhbounds` is also applicable to binary transformations of the outcome variable in the case of continuous outcomes.

2 Sensitivity Analysis with Rosenbaum Bounds

Checking the sensitivity of estimated treatment effects has become an increasingly important topic in the applied evaluation literature (see Caliendo and Kopeinig (2006) for a recent survey of different methods to do so). Here, we are interested what happens when there are deviations from the underlying identifying conditional independence assumption.

2.1 The Model

Let us assume that the participation probability is given by $P_i = P(x_i, u_i) = P(D_i = 1 \mid x_i, u_i) = F(\beta x_i + \gamma u_i)$, where x_i are the observed characteristics for individual i, u_i is the unobserved variable and γ is the effect of u_i on the participation decision. Clearly, if the study is free of hidden bias, γ will be zero and the participation probability will solely be determined by x_i. However, if there is hidden bias, two individuals with the same observed covariates x have differing chances of receiving treatment. Let us assume we have a matched pair of individuals i and j and further assume that F is the logistic distribution. The odds that individuals receive treatment are then given by $\frac{P_i}{(1-P_i)}$ and $\frac{P_j}{(1-P_j)}$, and the odds ratio is given by:

$$\frac{\frac{P_i}{1-P_i}}{\frac{P_j}{1-P_j}} = \frac{P_i(1-P_j)}{P_j(1-P_i)} = \frac{\exp(\beta x_i + \gamma u_i)}{\exp(\beta x_j + \gamma u_j)}. \tag{1}$$

If both units have identical observed covariates - as implied by the matching procedure - the x-vector cancels out implying that:

$$\frac{\exp(\beta x_i + \gamma u_i)}{\exp(\beta x_j + \gamma u_j)} = exp[\gamma(u_i - u_j)]. \tag{2}$$

But still, both individuals differ in their odds of receiving treatment by a factor that involves the parameter γ and the difference in their unobserved covariates u. So, if there are either no differences in unobserved variables ($u_i = u_j$) or if unobserved variables have no influence on the probability of participating ($\gamma = 0$), the odds ratio is one, implying the absence of hidden or unobserved selection bias. It is now the task of sensitivity analysis to evaluate how inference about the programme effect is altered by changing the values of γ and $(u_i - u_j)$. We follow Aakvik (2001) and assume for the sake of simplicity that the unobserved covariate is a dummy variable with $u_i \in \{0, 1\}$. Rosenbaum (2002) shows that (1) implies the following bounds on the odds-ratio that either of the two matched individuals will receive treatment:

$$\frac{1}{e^\gamma} \leq \frac{P_i(1-P_j)}{P_j(1-P_i)} \leq e^\gamma. \tag{3}$$

Both matched individuals have the same probability of participating only if $e^\gamma = 1$. Otherwise, if for example $e^\gamma = 2$, individuals who appear to be similar (in terms of x)

could differ in their odds of receiving the treatment by as much as a factor of 2. In this sense, e^γ is a measure of the degree of departure from a study that is free of hidden bias (Rosenbaum, 2002).[4]

2.2 The MH Test Statistic

For binary outcomes, Aakvik (2001) suggests using the Mantel and Haenszel (MH, 1959) test statistic. To do so, some additional notation is needed. We observe the outcome y for both participants and non-participants. If y is unaffected by different treatment assignments, treatment d is said to have no effect. If y is different for different assignments, then the treatment has some positive (or negative) effect. To be significant, the treatment effect has to cross some test statistic $t(d, y)$. The MH non-parametric test compares the successful number of individuals in the treatment group against the same expected number given the treatment effect is zero. Aakvik (2001) notes that the MH test can be used to test for no treatment effect both within different strata of the sample and as a weighted average between strata. Under the null-hypothesis of no treatment effect, the distribution of y is hypergeometric. We notate N_{1s} and N_{0s} as the numbers of treated and non-treated individuals in stratum s, where $N_s = N_{0s} + N_{1s}$. Y_{1s} is the number of successful participants, Y_{0s} is the number of successful non-participants, and Y_s is the number of total successes in stratum s. The test-statistic Q_{MH} follows asymptotically the standard normal distribution and is given by:

$$Q_{MH} = \frac{|Y_1 - \sum_{s=1}^{S} E(Y_{1s})| - 0.5}{\sqrt{\sum_{s=1}^{S} Var(Y_{1s})}} = \frac{|Y_1 - \sum_{s=1}^{S} (\frac{N_{1s} Y_s}{N_s})| - 0.5}{\sqrt{\sum_{s=1}^{S} \frac{N_{1s} N_{0s} Y_s (N_s - Y_s)}{N_s^2 (N_s - 1)}}}. \tag{4}$$

To use such a test-statistic, we first have to make the individuals in the treatment and control groups as similar as possible, because this test is based on random sampling. Since this is done by our matching procedure, we can proceed to discuss the possible influences of $e^\gamma > 1$. For fixed $e^\gamma > 1$ and $u \in \{0, 1\}$, Rosenbaum (2002) shows that the test-statistic Q_{MH} can be bounded by two known distributions. As noted already, if $e^\gamma = 1$ the bounds are equal to the 'base' scenario of no hidden bias. With increasing e^γ, the bounds move apart reflecting uncertainty about the test-statistics in the presence of unobserved selection bias. Two scenarios are especially useful. Let Q_{MH}^+ be the test-statistic given that we have overestimated the treatment effect and Q_{MH}^- the case where we have underestimated the treatment effect. The two bounds are then given by:

$$Q_{MH}^+ = \frac{|Y_1 - \sum_{s=1}^{S} \widetilde{E}_s^+| - 0.5}{\sqrt{\sum_{s=1}^{S} Var(\widetilde{E}_s^+)}} \tag{5}$$

4. A related approach can be found in Manski (1990, 1995) who proposes 'worst-case bounds' which are somewhat analogous to letting $e^\gamma \to \infty$ in a sensitivity analysis.

and

$$Q^{-}_{MH} = \frac{|Y_1 - \sum_{s=1}^{S} \widetilde{E}^{-}_{s}| - 0.5}{\sqrt{\sum_{s=1}^{S} Var(\widetilde{E}^{-}_{s})}} \tag{6}$$

where $\widetilde{E_s}$ and $Var(\widetilde{E_s})$ are the large sample approximations to the expectation and variance of the number of successful participants when u is binary and for given γ.[5]

3 Syntax

`mhbounds` computes Mantel-Haenszel bounds to check sensitivity of estimated average treatment effects on the treated.

`mhbounds` *outcome* [*if*], `gamma(numlist)` [`treated(newvar) weight(newvar) support(newvar) stratum(newvar) stratamat`]

4 Options

`gamma(numlist)` is a compulsory option and asks users to specify the values of $\Gamma = e^{\gamma} \geq 1$ for which to carry out the sensitivity analysis. Estimates at $\Gamma = 1$ (no hidden bias) are included in the calculations by default.

`treated(varname)` specifies the name of the user-provided treatment variable; If no name is provided, mhbounds expects `_treated` from `psmatch` or `psmatch2`.

`weight(varname)` specifies the name of the user-provided variable containing the frequency with which the observation is used as a match; if no name is provided, mhbounds expects `_weight` from `psmatch` or `psmatch2`.

`support(varname)` specifies the name of the user-provided common support variable. If no name is provided, `mhbounds` expects `_support` from `psmatch` or `psmatch2`.

`stratum(varname)` specifies the name of the user-provided variable indicating strata. Aakvik (2001) notes that the Mantel-Haenszel test can be used to test for no treatment effect both within different strata of the sample and as a weighted average between strata. This option is particularly useful when used after stratification matching, using, e.g. `atts`.

`stratamat`, in combination with `stratum(varname)` keeps in memory not only the matrix `outmat` containing the overall/combined test statistics, but also the matrices

5. The large sample approximation of $\widetilde{E^{+}_{s}}$ is the unique root of the following quadratic equation: $\widetilde{E}^{2}_{s}(e^{\gamma} - 1) - \widetilde{E}_{s}[(e^{\gamma} - 1)(N_{1s} + Y_s) + N_s] + e^{\gamma} Y_s N_{1s}$, with the addition of $max(0, Y_s + N_{1s} - N_s \leq \widetilde{E_s} \leq min(Y_s, N_{1s}))$ to decide which root to use. $\widetilde{E}^{-}_{s}$ is determined by replacing e^{γ} by $\frac{1}{e^{\gamma}}$. The large sample approximation of the variance is given by: $Var(\widetilde{E}_s) = \left(\frac{1}{\widetilde{E_s}} + \frac{1}{Y_s - \widetilde{E_s}} + \frac{1}{N_{1s} - \widetilde{E_s}} + \frac{1}{N_s - Y_s - N_{1s} + \widetilde{E_s}}\right)^{-1}$.

outmat_j containing the strata-specific test statistics, $j = 1, ...,$#strata.

4.1 Typical Examples

1. Running mhbounds after psamtch2
   - psmatch2 college, outcome(wage) pscore(pscore) caliper(.25) common noreplacement
   - mhbounds wage, gamma(1 (0.05) 2) [performs sensitivity analysis at Gamma = 1,1.05,1.10,...,2.]
2. Running mhbounds with user-defined treatment-, weight- and support-indicators
   - mhbounds outcome, gamma(1 (0.05) 2) treated(mytreat) weight(myweight) support(mysupport)
3. Running mhbounds with user-defined treatment-, weight- and support-indicators with different strata in the population
   - mhbounds outcome, gamma(1 (0.05) 2) treated(mytreat) weight(myweight) support(mysupport) stratum(mystratum) stratamat

Please note that mhbounds is suited for k-nearest neighbor matching without replacement and for stratification matching.

5 Examples

To illustrate mhbounds we give two examples, where the first one is taken from the book of Rosenbaum (2002) and the second one relates to the well known and much discussed studies by Lalonde (1986), Dehejia and Wahba (1999) and Smith and Todd (2005).

5.1 Rosenbaum Example

The first example is given in Rosenbaum (2002, Table 4.11, p. 130) and comes from a medical study of the possible effects of the drug allopurinol as a cause of rash (Boston Collaborative Drug Surveillance Program, 1972). The treatment in this case is the use of the drug ($D \in \{0, 1\}$) and the binary outcome variable is to have a rash or not ($Y \in \{0, 1\}$). Table 1 summarises the available data from a case-referent study, where treated and control group are already comparable and we distinguish two strata of the population ($S = 1$ for males and $S = 2$ for females).

A first look at the distribution of outcomes between treated and control units would suggest that the treatment in fact has a positive effect on the outcome variable, since, e.g. $\frac{5}{33} \approx 15\%$ of the treated males have an outcome of 1 whereas this is true for only $\frac{36}{645} \approx 6\%$ of the control individuals. In order to replicate the example we generate a sample of individuals according to the distribution of D and Y in Table 1.

Table 1: Illustrative Example

Stratum		$Y_i = 0$	$Y_i = 1$
$S_i = 1$ (Males)	$D_i = 1$	33	5
	$D_i = 0$	645	36
$S_i = 2$ (Females)	$D_i = 1$	19	10
	$D_i = 0$	518	58

Source: Rosenbaum (2002), p. 130.

```
.
. clear
. set obs 719
obs was 0, now 719
. gen s = 1
. gen d = _n<=38
. gen out = _n<=5
. replace out = 1 if _n>38&_n<75
(36 real changes made)
. save s1.dta, replace
file s1.dta saved
.
. clear
. set obs 605
obs was 0, now 605
. gen s = 2
. gen d = _n<=29
. gen out = _n<=10
. replace out = 1 if _n>29&_n<88
(58 real changes made)
. save s2.dta, replace
file s2.dta saved
.
. append using s1.dta
. gen myweight = 1
. gen mysupport = 1
. bys s: tab out d
```

```
-> s = 1
```

out	d 0	1	Total
0	645	33	678
1	36	5	41
Total	681	38	719

```
-> s = 2
```

out	d 0	1	Total
0	518	19	537
1	58	10	68
Total	576	29	605

```
.
```

Since we have two strata (males and females) in the population we are going to use the `stratum` option of `mhbounds`. Furthermore, we specify that we are interested in the sensitivity of the results up to a situation where $\Gamma = e^{\gamma} = 8$. Since the data is already matched, we do not have to run any of the available matching routines in Stata. However, in order for `mhbounds` to work we have to define a treatment indicator (`treated`), the weight assigned to each individual of both groups (`weight`) and furthermore identify the individuals who are within the region of common support (`support`). To keep the example simple, we assume equal weights and that all the individuals lie within the common support region.

```
. mhbounds out, gamma(1 (1) 8) treated(d) weight(myweight) support(mysupport) s
> tratum(s)
Mantel-Haenszel (1959) bounds for variable out
Gamma         Q_mh+     Q_mh-     p_mh+     p_mh-
-------------------------------------------------
    1        4.18665   4.18665   .000014   .000014
    2        1.80445   7.05822   .035581   8.4e-13
    3        .515322   9.09935   .303164         0
    4        .074087   10.7675   .470471         0
    5        .787917   12.2124   .215372         0
    6        1.37611   13.5046   .084394         0
    7        1.87943   14.6841   .030093         0
    8        2.32133   15.7759   .010134         0
Gamma : odds of differential assignment due to unobserved factors
Q_mh+ : Mantel-Haenszel statistic (assumption: over-estimation of treatment eff
> ect)
Q_mh- : Mantel-Haenszel statistic (assumption: under-estimation of treatment ef
> fect)
p_mh+ : significance level (assumption: over-estimation of treatment effect)
p_mh- : significance level (assumption: under-estimation of treatment effect)
```

In a study free of hidden bias, i.e. where $\Gamma = 1$, the Q_{MH} test-statistic is 4.19 and would constitute strong evidence that the use of allopurinol causes rash. If we have a positive (unobserved) selection, in the sense that if those most likely to use the drug, also have a higher probability to get rash, then the estimated treatment effects overestimate the true treatment effect. The reported test-statistic Q_{MH} is then too high and should be adjusted downwards. Hence, we will look at Q^{+}_{mh} and p^{+}_{mh} in the Stata output. The upper bounds on the significance levels for $\Gamma = 1, 2$, and 3 are 0.0001, 0.036, and 0.30 (see also Rosenbaum (2002, p.131)). The study is insensitive to a bias that would double the

odds of exposure to allopurinol but sensitive to a bias that would triple the odds. Our example also highlights, that in some applications the significance level on the bounds might fall first and than rise again. If we look, e.g. at the situation for $\Gamma = 8$, we get a significance level p^+_{mh} of .0101 indicating a significant effect once again. It should be clear, that this second significant value of p^+_{mh} indicates a significant negative treatment effect. This is due to the fact, that we assume a large positive unobserved heterogeneity which turns our previously significant positive treatment effect into a negative one.

5.2 The NSW Data Revisited

To illustrate `mhbounds` in a more common evaluation environment, we use the data also used by Dehejia and Wahba (1999) and Smith and Todd (2005). It is well known that the first study was very influential to promote matching as an evaluation method, whereas the second one raised some doubts on the reliability of the results in non-experimental evaluation settings.

The data come from Lalonde's (1986) evaluation of non-experimental evaluation methods and combines treated units from a randomized study of the National Supported Work (NSW) training program with non-experimental comparison groups from surveys as the Panel Study of Income Dynamics (PSID) or the Current Population Survey (CPS).[6] We restrict the sample to the experimental treatment group ($n = 185$) and the PSID control group ($n = 2490$). The outcome of interest in DW99 are the post-intervention real earnings in 1978 ($RE78$). Since we are interested in binary outcomes, we define a new outcome variable *employment* taking the value of 1 if the individual had positive real earnings in 1978 and 0 otherwise. The distribution of the outcome variable is the following:

```
. tab employment d
```

employment	d 0	1	Total
0	286	45	331
1	2,204	140	2,344
Total	2,490	185	2,675

To make the samples comparable we use propensity score matching by running `psmatch2` on the same specification as DW99.

```
. psmatch2 d age age2 education educ2 married black hispanic re74 re75 re742 re
> 752 blacku74, logit out(employment) noreplacement
Logistic regression                               Number of obs   =       2675
                                                  LR chi2(12)     =     935.35
                                                  Prob > chi2     =     0.0000
Log likelihood = -204.97537                       Pseudo R2       =     0.6953
------------------------------------------------------------------------------
```

6. The data are available at Dehejia's website: http://www.nber.org/∼rdehejia/nswdata.html.

```
          d |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
------------+----------------------------------------------------------------
        age |   .3316904   .1203295     2.76   0.006     .0958489    .5675318
       age2 |  -.0063668   .0018554    -3.43   0.001    -.0100033   -.0027303
  education |   .8492683   .3477041     2.44   0.015     .1677807    1.530756
      educ2 |  -.0506202   .0172492    -2.93   0.003     -.084428   -.0168124
    married |  -1.885542   .2993282    -6.30   0.000    -2.472214   -1.298869
      black |   1.135973   .3517793     3.23   0.001      .446498    1.825447
   hispanic |    1.96902   .5668567     3.47   0.001     .8580017    3.080039
       re74 |  -.0001059   .0000353    -3.00   0.003     -.000175   -.0000368
       re75 |  -.0002169   .0000414    -5.24   0.000     -.000298   -.0001357
       re742 |   2.39e-09   6.43e-10     3.72   0.000     1.13e-09    3.65e-09
       re752 |   1.36e-10   6.55e-10     0.21   0.836    -1.15e-09    1.42e-09
    blacku74 |   2.144129   .4268089     5.02   0.000     1.307599    2.980659
       _cons |  -7.474742   2.443502    -3.06   0.002    -12.26392   -2.685566
------------------------------------------------------------------------------

Note: 22 failures and 0 successes completely determined.
There are observations with identical propensity score values.
The sort order of the data could affect your results.
Make sure that the sort order is random before calling psmatch2.

-------------------------------------------------------------------------------
        Variable     Sample |    Treated     Controls   Difference         S.E.
>    T-stat
-------------------------------------------------------------------------------
      employment  Unmatched | .756756757   .885140562  -.128383805   .024978843
>     -5.14
                        ATT | .756756757   .664864865   .091891892   .047025406
>      1.95
-------------------------------------------------------------------------------

Note: S.E. for ATT does not take into account that the propensity score is esti
> mated.

            | psmatch2:
 psmatch2:  |   Common
 Treatment  |  support
assignment  | On suppor |     Total
------------+-----------+----------
 Untreated  |     2,490 |     2,490
   Treated  |       185 |       185
------------+-----------+----------
     Total  |     2,675 |     2,675
```

What can be seen from the output is that we get a significant positive treatment effect on the treated of 0.0919. That is the employment rate of participants is 9.2%-points higher when compared to matched control group members. Since `psmatch2` automatically produces the variables `_treated`, `_weight`, and `_support` we do not have to specify those when using `mhbounds`.

```
. mhbounds employment, gamma(1 (0.05) 1.5)

Mantel-Haenszel (1959) bounds for variable employment

Gamma         Q_mh+     Q_mh-     p_mh+     p_mh-
----------------------------------------------------
    1        1.83216   1.83216   .033464   .033464
 1.05        1.62209   2.04761   .052392   .020299
```

```
  1.1        1.41978    2.2511   .077836    .01219
 1.15        1.22673   2.44599   .109961   .007223
  1.2        1.04213   2.63301   .148677   .004232
 1.25        .865226   2.81282   .193457   .002455
  1.3        .695397   2.98601   .243403   .001413
 1.35        .532076   3.15309   .297337   .000808
  1.4        .374766   3.31449   .353917   .000459
 1.45        .223022   3.47064   .411759    .00026
  1.5        .076449   3.62189   .469531   .000146
Gamma : odds of differential assignment due to unobserved factors
Q_mh+ : Mantel-Haenszel statistic (assumption: over-estimation of treatment eff
> ect)
Q_mh- : Mantel-Haenszel statistic (assumption: under-estimation of treatment ef
> fect)
p_mh+ : significance level (assumption: over-estimation of treatment effect)
p_mh- : significance level (assumption: under-estimation of treatment effect)
```

Under the assumption of no hidden bias ($\Gamma = 1$), the Q_{MH} test-statistic gives a similar result, indicating a significant treatment effect. The two bounds in the output table can be interpreted in the following way: The Q^{+}_{MH} statistic adjusts the MH statistic downward for the case of positive (unobserved) selection. For the given example, positive selection bias occurs when those most likely to participate tend to have higher employment rates even in the absence of participation and given that they have the same x-vector as the individuals in the comparison group. This leads to an upward bias in the estimated treatment effects. The Q^{-}_{MH} statistic adjusts the MH statistic downward for the case of negative (unobserved) selection. In other examples, the treatment effects at $\Gamma = 1$ might be insignificant and the bounds tell us at which degree of unobserved positive or negative selection the effect would become significant.

Given the positive estimated treatment effect, the bounds under the assumption that we have under-estimated the true treatment effect (Q^{-}_{MH}) are somewhat less interesting. The effect is significant under $\Gamma = 1$ and becomes even more significant for increasing values of Γ if we have under-estimated the true treatment effect. However, looking at the bounds under the assumption that we have over-estimated the treatment effect, i.e. Q^{+}_{MH}, reveals that already at relatively small levels of Γ, the result becomes insignificant. To be more specific, with a value of $\Gamma = 1.1$ the result would not be significant at the 5%-significance level any more, with $\Gamma = 1.15$ it is even not significant at the 10%-significance level, since the p-value is 0.109961. Clearly, based on these findings one would be careful when interpreting the results.

However, it should be noted that these are worst-case scenarios. Hence, a critical value of $\Gamma = 1.15$ does not mean that unobserved heterogeneity exists and that there is no effect of treatment on the outcome variable. This result only states that the confidence interval for the effect would include zero if an unobserved variable caused the odds ratio of treatment assignment to differ between the treatment and comparison groups by 1.15. One should keep in mind, that this test cannot directly justify the unconfoundedness assumption. Hence, we cannot state whether the CIA does (not) hold for the given setting (including inter alia the used data, the chosen covariates and the specification of the propensity score). What we can say is, that the results are quite

sensitive to possible deviations from the identifying unconfoundedness assumption and hence, some caution when interpreting the results is advisable.

6 Saved Results

`mhbounds` produces the matrix `outmat` containing the Mantel-Haenszel test statistics for all values of Γ specified by the user. When the option `stratamat` is specified in conjunction with `stratum(varname)`, `mhbounds` keeps in memory not only the matrix `outmat` containing the overall/combined test statistics, but also the matrices `outmat_j` containing the strata-specific test statistics, $j = 1, ...,$#strata.

7 Acknowledgments

We thank Tommaso Nannicini for very useful suggestions.

8 References

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

Aakvik, A. (2001), "Bounding a Matching Estimator: The Case of a Norwegian Training Program", *Oxford Bulletin of Economics and Statistics* 63(1), 115-143.

Becker, S.O. and Ichino, A. (2002), "Estimation of average treatment effects based on propensity scores", *Stata Journal* 2(4), 358-377.

Caliendo, M., Hujer, R. and Thomsen, S. (2005), "The Employment Effects of Job Creation Schemes in Germany", IZA-Discussion Paper No. 1512, Bonn.

Caliendo, M. and Kopeinig, S. (2006), "Some Practical Guidance for the Implementation of Propensity Score Matching", IZA Discussion Paper No. 1588, forthcoming in *Journal of Economic Surveys*.

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

DiPrete, T. and Gangl, M. (2005), "Assessing Bias in the Estimation of Causal Effects: Rosenbaum Bounds on Matching Estimators and Instrumental Variables Estimation with Imperfect Instruments", *Sociological Methodology* 35.

Heckman, J.J., Ichimura, H., Smith, J. and Todd, P. (1998), "Characterizing Selection Bias Using Experimental Data," *Econometrica*, 66(5), 1017-1098.

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, Bonn.

Leuven, E. and Sianesi, B. (2003), "`PSMATCH2`: Stata module to perform full Mahalanobis and propensity score matching, common support graphing, and covariate imbalance testing", *http://ideas.repec.org/c/boc/bocode/s432001.html.*

Mantel, N. and Haenszel, W. (1959), "Statistical Aspects of the Analysis of Data from Retrospective Studies of Disease", *Journal of the National Cancer Institute* 22(4), 719-748.

Manski, C. (1990), "Nonparametric Bounds on Treatment Effects", *American Economic Review* 80, 319-323.

Manski, C. (1995), *Identification Problems in the Social Sciences*, (Cambridge: Harvard University Press).

Nannicini, T. (2006), "`sensatt`: A Simulation-Based Sensitivity Analysis for Matching Estimators".

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

Smith, J. and Todd, P. (2005), "Does Matching Overcome LaLonde's Critique of Non Experimental Estimators?", *Journal of Econometrics* 125(1-2), 305-353.

About the authors

Sascha O. Becker is Assistant Professor at the Center for Economic Studies (CES) at the Ludwig-Maximilians-University Munich, Germany. He is also affiliated with CESifo, ifo and IZA.

Marco Caliendo is Senior Research Associate at the German Institute for Economic Research (DIW Berlin) and Research Fellow of the IZA, Bonn and the IAB, Nuremberg, Germany.

Revised and improved versions of the programs may become available in the future on our web pages (http://www.sobecker.de and http://www.caliendo.de).

Please address any correspondence to Sascha O. Becker (so.b@gmx.net) or Marco Caliendo (mcaliendo@diw.de).