

# Sensitivity Analysis for m-Estimates, Tests, and Confidence Intervals in Matched Observational Studies

Paul R. Rosenbaum

Department of Statistics, University of Pennsylvania, 473 Huntsman Hall, Philadelphia,  
Pennsylvania 19104-6340, U.S.A.

*email:* rosenbaum@stat.wharton.upenn.edu

**SUMMARY.** Huber's m-estimates use an estimating equation in which observations are permitted a controlled level of influence. The family of m-estimates includes least squares and maximum likelihood, but typical applications give extreme observations limited weight. Maritz proposed methods of exact and approximate permutation inference for m-tests, confidence intervals, and estimators, which can be derived from random assignment of paired subjects to treatment or control. In contrast, in observational studies, where treatments are not randomly assigned, subjects matched for observed covariates may differ in terms of unobserved covariates, so differing outcomes may not be treatment effects. In observational studies, a method of sensitivity analysis is developed for m-tests, m-intervals, and m-estimates: it shows the extent to which inferences would be altered by biases of various magnitudes due to nonrandom treatment assignment. The method is developed for both matched pairs, with one treated subject matched to one control, and for matched sets, with one treated subject matched to one or more controls. The method is illustrated using two studies: (i) a paired study of damage to DNA from exposure to chromium and nickel and (ii) a study with one or two matched controls comparing side effects of two drug regimes to treat tuberculosis. The approach yields sensitivity analyses for: (i) m-tests with Huber's weight function and other robust weight functions, (ii) the permutational *t*-test which uses the observations directly, and (iii) various other procedures such as the sign test, Noether's test, and the permutation distribution of the efficient score test for a location family of distributions. Permutation inference with covariance adjustment is briefly discussed.

**KEY WORDS:** m-estimate; Noether's estimate; Permutation test; Randomization test.

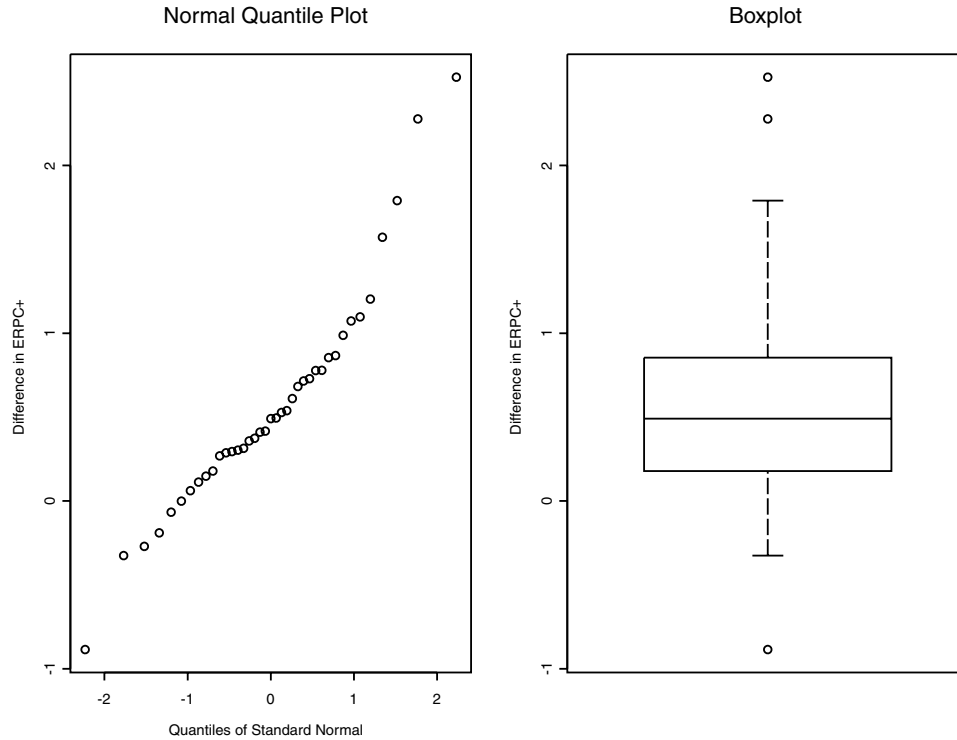
## 1. Introduction and Example: DNA Damage Among Welders

Experiments with cells suggest that chromium and nickel can damage DNA. Werfel et al. (1998) matched 39 welders exposed to nickel and chromium to 39 unexposed controls, matching for age and smoking habits. They measured DNA strand breakages and DNA-protein crosslinks in lymphocytes using several techniques, including relative DNA elution rates through polycarbonate filters with proteinase K (or ERPC+). Figure 1 displays differences in ERPC+, welder minus control, for the 39 matched pairs. Most differences are positive, consistent with greater DNA damage among welders, and the center of the distribution resembles a Gaussian distribution, but its tails are thicker.

This is an observational study, not an experiment. Although matched for age and smoking, in the absence of random assignment, there is nothing to ensure that welders and controls are comparable in terms of covariates  $u$  that were not measured. If it had been a randomized experiment, then a robust inference could be based on a suggestion of Maritz (1979, 1995, Section 2.8) for randomization inference related to Huber's m-estimator; see Section 3. If that technique were applied to Figure 1 with a constant effect  $\tau$ , the one-sided significance level for testing  $H_0 : \tau = 0$  would be  $1.04 \times 10^{-6}$ ,

the one-sided 95% confidence interval would be  $\tau \geq 0.35$ , and the point estimate would be  $\hat{\tau} = 0.49$ . How sensitive are these randomization inferences to possible departures from random assignment? How would departures of various magnitudes alter the inferences?

Methods exist for sensitivity analysis when using rank tests and R-estimates in observational studies, for example, when using Wilcoxon's signed rank test and the associated Hodges–Lehmann estimate; see Rosenbaum (1993, 2002a, 2003) and Gastwirth, Krieger, and Rosenbaum (2000). There are close links between R-estimates and Huber's m-estimates; see Maritz (1979, 1995, Section 2.8) and Jureková and Sen (1996, Section 7.2). The family of m-estimates is highly flexible, including as special cases least squares, maximum likelihood estimation for location families of distributions, and a variety of robust estimates. Unlike rank statistics, m-statistics are typically designed to be continuous functions of data and parameters, and so may be more useful in describing coarse, heavily tied data, as in the example in Section 4.1 where the responses are integers between 3 and 11. In the current article, the method of sensitivity analysis for R-estimates is extended for use with m-estimates, using certain connections suggested by Maritz (1979).



**Figure 1.** Matched pair differences  $D_i$  in ERPC+, a measure of DNA damage: normal quantile plot and boxplot.

Section 2 defines notation and reviews a model for sensitivity analysis. The case of matched pairs is simpler, so it is discussed first in Section 3. The adjustments needed for matching with a variable number of controls are discussed in Section 4, with a second example in Section 4.1. The extension to permutation inference with covariance adjustment is discussed in Section 5.

## 2. Notation and Review

### 2.1 Notation for Matched Pairs or Matched Sets

The  $i$ th of  $I$  matched sets contains  $n_i \geq 2$  subjects,  $j = 1, \dots, n_i$ , of whom one receives treatment, denoted  $Z_{ij} = 1$ , and the others receive control, denoted  $Z_{ij} = 0$ , so that  $1 = \sum_{j=1}^{n_i} Z_{ij}$  for  $i = 1, \dots, I$ , with  $N = \sum n_i$  subjects in total. The  $j$ th subject in set  $i$  has an observed covariate  $\mathbf{x}_{ij}$  and an unobserved covariate  $u_{ij}$ , and the sets were matched for the observed covariate but not for the unobserved covariate, so  $\mathbf{x}_{ij} = \mathbf{x}_{i,j'}$  for all  $i, j, j'$ , but possibly  $u_{ij} \neq u_{i,j'}$ . In Section 1,  $I = 39$ ,  $n_i = 2$  for all  $i$ ,  $N = 78$ , and  $\mathbf{x}_{ij}$  described subject  $(i, j)$ 's age and smoking behavior. The  $j$ th subject in set  $i$  would exhibit response  $r_{Tij}$  if this subject received treatment,  $Z_{ij} = 1$ , and would exhibit response  $r_{Cij}$  if this subject received control,  $Z_{ij} = 0$ , so the response actually observed from this subject is  $R_{ij} = Z_{ij}r_{Tij} + (1 - Z_{ij})r_{Cij}$ , and the effect of the treatment on this subject is  $r_{Tij} - r_{Cij}$ ; for example, Neyman (1923) and Rubin (1974). Because each subject receives either treatment or control, either  $r_{Tij}$  or  $r_{Cij}$  is observed but not both, so the effect  $r_{Tij} - r_{Cij}$  cannot be calculated. Write  $\mathbf{R} = (R_{11}, R_{12}, \dots, R_{I, n_I})^T$

and  $\mathbf{Z} = (Z_{11}, Z_{12}, \dots, Z_{I, n_I})^T$ . Write  $|A|$  for the number of elements in a set  $A$ . Let  $\Omega$  be the set of the  $|\Omega| = \prod_{i=1}^I n_i$  possible values  $\mathbf{z} = (z_{11}, z_{12}, \dots, z_{I, n_I})^T$  of  $\mathbf{Z}$ , so each  $\mathbf{z} \in \Omega$  has binary coordinates with  $1 = \sum_{j=1}^{n_i} z_{ij}$  for each  $i$ . Write  $\mathcal{F} = \{(r_{Tij}, r_{Cij}, \mathbf{x}_{ij}, u_{ij}), i = 1, \dots, I, j = 1, \dots, n_i\}$ , and write  $\mathcal{Z}$  for the event  $\mathcal{Z} = \{1 = \sum_{j=1}^{n_i} Z_{ij}, i = 1, \dots, I\}$ .

In a randomized experiment, one subject in each matched set  $i$  is picked at random to receive treatment, each having probability  $1/n_i$ , with independent assignments in distinct matched sets, so each  $\mathbf{z} \in \Omega$  has probability  $1/|\Omega|$ . In Fisher's (1935) theory, this is the only probability distribution used in randomization inference. In this approach, the treatment assignment,  $\mathbf{Z}$ , is a random variable, as are quantities that depend on  $\mathbf{Z}$ , such as  $\mathbf{R}$ , but quantities that do not involve  $\mathbf{Z}$ , such as  $\mathcal{F}$ , are fixed features of the finite population of  $N$  subjects; for example, Welch (1937). To say that a quantity is *fixed* is to say that all probabilities are implicitly conditional probabilities given their values. For example, in a randomized experiment,  $\Pr(Z_{ij} = 1 | \mathcal{F}, \mathcal{Z}) = 1/n_i$  and  $\Pr(\mathbf{Z} = \mathbf{z} | \mathcal{F}, \mathcal{Z}) = 1/|\Omega|$  for each  $\mathbf{z} \in \Omega$ ; that is, having formed the sets of given sizes  $n_i$  matched for  $\mathbf{x}_{ij}$ , treatments were assigned at random, and neither unobserved covariates  $u_{ij}$  nor potential responses under alternative treatments ( $r_{Tij}, r_{Cij}$ ) would predict the treatment assignment in a matched set.

The treatment has a constant effect if  $r_{Tij} - r_{Cij} = \tau$  for all  $i, j$ , so that  $R_{ij} = r_{Cij} + Z_{ij}\tau$ . Hypotheses of constant effect,  $H_0 : \tau = \tau_0$ , are often of interest and often lead to parsimonious descriptions, but they are not essential to randomization inference or to sensitivity analysis; see Section 3.5 and Rosenbaum (2002a, Section 5, 2003).

2.2 Review of a Model for Sensitivity Analysis

The model for sensitivity analysis (Rosenbaum, 1993, 2002a, Section 4, 2003) says failure to match on the unobserved covariate  $u_{ij}$  may have led to biased treatment assignments, so matched subjects with the same  $\mathbf{x}$  may differ in their chances  $\Pr(Z_{ij} = 1 | \mathcal{F}, \mathcal{Z})$  of receiving the treatment by at most a factor of  $\Gamma = \exp(\gamma) \geq 1$  because they differ in terms of  $u_{ij}$ ,

$$\frac{1}{\Gamma} \leq \frac{\Pr(Z_{ij} = 1 | \mathcal{F}, \mathcal{Z})}{\Pr(Z_{ik} = 1 | \mathcal{F}, \mathcal{Z})} \leq \Gamma, \quad i = 1, \dots, I, 1 \leq j, k \leq n_i, \tag{1}$$

with independent assignments in distinct matched sets. For instance, if  $\Gamma$  were two, then in place of random assignment, one subject  $j$  in a matched set  $i$  might be twice as likely as another  $k$  to receive the treatment, even though they have the same observed covariates,  $\mathbf{x}_{ij} = \mathbf{x}_{ik}$ , because they differ in terms of the unobserved covariate,  $u_{ij} \neq u_{ik}$ .

It is easy to show that the model (1) can be written in the equivalent form (2):

$$\Pr(\mathbf{Z} = \mathbf{z} | \mathcal{F}, \mathcal{Z}) = \prod_{i=1}^I \frac{\exp\left(\gamma \sum_{j=1}^{n_i} z_{ij} u_{ij}\right)}{\sum_{k=1}^{n_i} \exp(\gamma u_{ik})}$$

for each  $\mathbf{z} \in \Omega$ , with  $0 \leq u_{ik} \leq 1$ . (2)

That (2) implies (1) is immediate because (2) implies  $\Pr(Z_{ij} = 1 | \mathcal{F}, \mathcal{Z}) / \Pr(Z_{ik} = 1 | \mathcal{F}, \mathcal{Z}) = \exp\{\gamma(u_{ij} - u_{ik})\}$ . That independence between matched sets and (1) implies (2) follows by writing  $\pi_{ij} = \Pr(Z_{ij} = 1 | \mathcal{F}, \mathcal{Z})$ ,  $\pi_{i,\min} = \min_{1 \leq j \leq n_i} \pi_{ij}$  and defining the unobserved  $u_{ij}$  to be  $u_{ij} = \{\log(\pi_{ij}) - \log(\pi_{i,\min})\} / \gamma$ ; for details, see Rosenbaum (2002a, Section 4.2). If  $\Gamma = 1$  so  $\gamma = 0$ , then (2) is the randomization distribution  $\Pr(\mathbf{Z} = \mathbf{z} | \mathcal{F}, \mathcal{Z}) = 1/|\Omega|$ , whereas if  $\Gamma > 1$ , then (2) is unknown because  $u_{ij}$  is unknown, but the extent of the deviation from the randomization distribution is controlled by  $\Gamma$ . For a fixed value of  $\Gamma \geq 1$ , there is a range of possible inferences, say a range of possible significance levels or a range of possible point estimates. A sensitivity analysis computes this range for several values of  $\Gamma \geq 1$ , thereby indicating the magnitude  $\Gamma$  of unobserved bias that would need to be present to materially alter the conclusions.

In thinking about  $\Gamma$ , two considerations may be helpful. First,  $\gamma = \log(\Gamma)$  is a logit regression coefficient, in the following sense. If in the population before matching, the treatment assignments  $Z$  are mutually independent with  $\log\{\Pr(Z = 1 | \mathbf{x}, u) / \Pr(Z = 0 | \mathbf{x}, u)\} = \lambda(\mathbf{x}) + \gamma u$ , where  $\lambda(\cdot)$  is any function and  $0 \leq u \leq 1$ , and subjects are matched exactly for  $\mathbf{x}$ , then (2) is the conditional distribution of treatment assignments  $\mathbf{Z}$  given  $1 = \sum_{j=1}^{n_i} Z_{ij}$ ,  $i = 1, \dots, I$ ; see Rosenbaum (2002a, Section 4.2) for details. Second, under this same logit model with a binary  $u$ , the parameter  $\Gamma$  is the odds ratio for  $\Pr(Z = a | \mathbf{x}, u = b)$ , for  $a, b = 0, 1$ , and by Bayes theorem,  $\Gamma$  is also the odds ratio for  $\Pr(u = a | \mathbf{x}, Z = b)$ , so it closely resembles the prevalence ratio  $\Pr(u = 1 | \mathbf{x}, Z = 1) / \Pr(u = 1 | \mathbf{x}, Z = 0)$  used in the first sensitivity analysis by Cornfield et al. (1959).

Observational studies vary markedly in their sensitivity to unobserved biases; see Rosenbaum (2002a, Section 4), where a study of smoking as a cause of lung cancer is insensitive to  $\Gamma = 5$ , but a study of coffee as a cause of myocardial infarction is sensitive to  $\Gamma = 1.3$ . Other methods for sensitivity analysis in observational studies are discussed by Cornfield et al. (1959), Gastwirth (1992), Lin, Psaty, and Kronmal (1998), Robins (1999), Robins, Rotnitzky, and Scharfstein (1999), Copas and Eguchi (2001), and Imbens (2003).

3. Matched Pairs

3.1 Paired  $m$ -Tests

The treatment effect is assumed to be an additive constant,  $r_{Tij} - r_{Cij} = \tau$ , and the matched sets are pairs,  $n_i = 2$  for all  $i$ , so that, as in Section 1, attention focuses on the treated-minus-control difference in observed responses,  $D_i = \tau + \epsilon_i$ , where  $\epsilon_i = (2Z_{i1} - 1)(r_{C11} - r_{C12})$ .

Consider testing  $H_0 : \tau = \tau_0$ . Let  $h_{\tau_0}$  be a function of  $|D_i - \tau_0|$ ,  $i = 1, \dots, I$ ; Maritz (1979) suggests  $h_{\tau_0} = \text{median } |D_i - \tau_0|$ . The  $m$ -test proposed by Maritz (1979) uses

$$T_{\tau_0} = \sum_{i=1}^I \text{sign}(D_i - \tau_0) \psi\left(\frac{|D_i - \tau_0|}{h_{\tau_0}}\right), \tag{3}$$

where  $\psi(\cdot)$  is an odd function,  $\psi(d) = -\psi(-d)$ , which is non-negative  $\psi(d) \geq 0$  for  $d \geq 0$ . Section 3.2 discusses  $\psi(\cdot), h_{\tau_0}$ , and the relationship between  $T_{\tau_0}$  and Huber's  $m$ -estimates.

If the hypothesis is false, if  $\tau \neq \tau_0$ , then in (3) the three quantities  $\text{sign}(D_i - \tau_0)$ ,  $|D_i - \tau_0|$ , and  $h_{\tau_0}$ , are all random variables; their values fluctuate with the treatment assignment  $\mathbf{Z}$ . The situation is simpler if  $H_0$  is true. Write  $\kappa_i = r_{C11} - r_{C12}$  and  $\boldsymbol{\kappa} = (\kappa_1, \dots, \kappa_I)^T$ , so  $D_i = (2Z_{i1} - 1)\kappa_i + \tau$ , and  $|D_i - \tau| = |\kappa_i|$ . Define  $\psi_i = \psi(|D_i - \tau|/h_{\tau}) = \psi(|\kappa_i|/h_{\tau})$ ; then  $\boldsymbol{\kappa}$  and  $h_{\tau}$  and  $\psi_i$  are functions of  $\mathcal{F}$  and so are fixed, not varying with  $\mathbf{Z}$ .

Consider, first, the situation developed by Maritz (1979), that is, a randomized experiment. If  $H_0 : \tau = \tau_0$  were true, then  $T_{\tau_0} = T_{\tau}$  where  $T_{\tau} = \sum_{i=1}^I \text{sign}\{(2Z_{i1} - 1)\kappa_i\} \psi_i$  is the sum of  $I$  independent random variables that take values  $\psi_i$  or  $-\psi_i$  each with probability  $\frac{1}{2}$  if  $\kappa_i \neq 0$  and take value 0 with probability 1 if  $\kappa_i = 0$ . Of course,  $\psi_i = 0$  if  $\kappa_i = 0$  because  $\psi(\cdot)$  is an odd function. This defines the exact null distribution of  $T_{\tau_0}$ , whose null expectation and variance may be used as  $I \rightarrow \infty$  in a Normal approximation (Maritz 1979, Section 4). A confidence set is formed by inverting the test, and a point estimate is obtained by solving the estimating equation  $T_{\tau_0} = 0$ .

Now consider sensitivity analysis for an observational study under (2). Write  $\bar{T}$  for the sum of  $I$  independent random variables,  $i = 1, \dots, I$ , that take value  $\psi_i$  with probability  $1/(1 + \Gamma)$  and value  $-\psi_i$  with probability  $\Gamma/(1 + \Gamma)$  if  $\kappa_i \neq 0$  and that take value 0 with probability 1 if  $\kappa_i = 0$ . In parallel, write  $\bar{\bar{T}}$  for the sum of  $I$  independent random variables that take value  $\psi_i$  with probability  $\Gamma/(1 + \Gamma)$  and value  $-\psi_i$  with probability  $1/(1 + \Gamma)$  if  $\kappa_i \neq 0$  and that take value 0 with probability 1 if  $\kappa_i = 0$ . Then under model (2), an argument parallel to that in Rosenbaum (2002a, Section 4) demonstrates:

$$\Pr(\bar{T} \geq k) \leq \Pr(T_{\tau} \geq k) \leq \Pr(\bar{\bar{T}} \geq k) \text{ for all } k. \tag{4}$$

In words, under  $H_0 : \tau = \tau_0$ , the null distribution of  $T_{\tau_0} = T_{\tau}$  is unknown because  $u_{ij}$  was not observed, but for each fixed  $\Gamma \geq 1$ , the null distribution is bounded by two known exact distributions (4). The bounds in (4) are tight: they are attained for particular distributions satisfying (1) or particular  $u_{ij}$  satisfying (2). For several values of  $\Gamma$ , the sensitivity analysis uses (4) to obtain bounds on inference quantities, such as significance levels, point estimates, or confidence intervals, thereby indicating the quantity of unobserved bias that would need to be present to materially alter the conclusions of the study.

The expectation and variance of  $\bar{T}$  and  $\bar{\bar{T}}$  in (4) are easily seen to be:

$$E(\bar{T}) = \frac{1-\Gamma}{1+\Gamma} \sum_{i=1}^I \psi_i, \quad E(\bar{\bar{T}}) = \frac{\Gamma-1}{1+\Gamma} \sum_{i=1}^I \psi_i,$$

$$\text{var}(\bar{T}) = \text{var}(\bar{\bar{T}}) = \frac{4\Gamma}{(1+\Gamma)^2} \sum_{i=1}^I \psi_i^2. \quad (5)$$

The  $\psi(\cdot)$ 's typically used with  $m$ -estimates are bounded, and  $\bar{T}$  and  $\bar{\bar{T}}$  are sums of independent bounded random variables, so by the central limit theorem as  $I \rightarrow \infty$ ,

$$\frac{\bar{T} - \frac{1-\Gamma}{1+\Gamma} \sum_{i=1}^I \psi_i}{\sqrt{\frac{4\Gamma}{(1+\Gamma)^2} \sum_{i=1}^I \psi_i^2}} \xrightarrow{D} N(0, 1) \quad \text{and}$$

$$\frac{\bar{\bar{T}} - \frac{\Gamma-1}{1+\Gamma} \sum_{i=1}^I \psi_i}{\sqrt{\frac{4\Gamma}{(1+\Gamma)^2} \sum_{i=1}^I \psi_i^2}} \xrightarrow{D} N(0, 1), \quad (6)$$

and approximations to the bounds in (4) are easily calculated.

### 3.2 Choice of $\psi$ and $h_{\tau}$ for Confidence Intervals and Tests

Where Huber (1964) emphasized point estimates, Maritz (1979) emphasized tests and confidence intervals. The differences are small, but consequential. Roughly speaking, when examining the large sample properties of estimates, one can safely assume that one is working near the true value of  $\tau$ , but one would like tests and confidence intervals to also work properly when testing  $H_0 : \tau = \tau_0$  for values of  $\tau_0$  far from the true  $\tau$ .

Consider, first, the scale measure,  $h_{\tau}$ . Maritz uses a scale estimate  $h_{\tau_0}$ , such as  $h_{\tau_0} = \text{median } |D_i - \tau_0|$ , computed in terms of absolute deviations  $|D_i - \tau_0|$  from the hypothesized value  $\tau_0$  in  $H_0 : \tau = \tau_0$ . With this scale measure, the exact null distributions for  $T_{\tau_0}$  have a conceptually simple form described in Section 3.1, and the exact null moments of  $T_{\tau_0}$  have simple formulas (5). A more common scale estimate in  $m$ -estimation is  $\text{median } |D_i - \tilde{\tau}|$  where  $\tilde{\tau}$  is the sample median of the  $D_i$ , but then Section 3.1 is not applicable, because  $T_{\tau}$  is not the sum of  $I$  independent random variables taking values  $\psi_i$  or  $-\psi_i$ . (In detail, if  $H_0 : \tau = \tau_0$  is true, then  $|D_i - \tau_0|$

is fixed, not varying with  $Z_{i1}$ , whereas  $\tilde{\tau}$  depends on all the  $D_k$ 's, so every  $|D_i - \tilde{\tau}|$  may change when any  $Z_{k1}$  changes.)

Maritz (1979) suggested using one of Huber's (1964, Section 4) original proposals  $\tilde{\psi}(\cdot)$  for  $\psi(\cdot)$ , namely  $\tilde{\psi}(y) = -1$  for  $y \leq -1$ ,  $\tilde{\psi}(y) = y$  for  $-1 < y < 1$ ,  $\tilde{\psi}(y) = 1$  for  $y \geq 1$ , together with  $h_{\tau_0} = \text{median } |D_i - \tau_0|$ , and he emphasized that with these definitions  $T_{\tau_0}$  is monotone decreasing as a function of  $\tau_0$ . This version of  $T_{\tau_0}$  was used in Section 1. With these definitions,  $T_{\tau_0}$  gives weights proportional to  $D_i - \tau_0$  to half of the pairs, specifically the pairs with  $D_i$ 's closest to the hypothesized value  $\tau_0$ , and absolutely larger but bounded weights of  $\pm 1$  to the other half of the pairs. A reader content with this definition of  $T_{\tau_0}$  could skip the remainder of this subsection, which discusses alternative definitions.

More generally, define  $h_{\tau_0}^{\ell}$  to be the  $\ell$ th of the  $|D_i - \tau_0|$  when they are sorted from smallest to largest. (Formally, to ensure that  $T_{\tau_0}$  is always well defined, if  $h_{\tau_0}^{\ell} = 0$ , define  $\tilde{\psi}\{y/h_{\tau_0}^{\ell}\} = \text{sign}(y)$ , noticing that  $h_{\tau_0}^{\ell} = 0$  for at most finitely many values of  $\tau_0$ .) Using  $\tilde{\psi}(\cdot)$  with this scale measure  $h_{\tau_0}^{\ell}$  makes  $T_{\tau_0}$  monotone decreasing in  $\tau_0$  for every choice of  $\ell$ ,  $1 \leq \ell \leq I$ . Several choices of  $\ell$  yield familiar procedures. Specifically, (i)  $\ell \doteq (I+1)/2$  yields the definition in Maritz (1979), (ii)  $\ell = 1$  yields the sign test, and (iii)  $\ell = I$  yields the permutational  $t$ -test. If these three procedures are used in the example of Section 1, the large sample approximation to the null randomization distribution for testing  $H_0 : \tau = 0$  yields standardized deviates and approximate one-sided significance levels of: (i) 4.74 and  $1.04 \times 10^{-6}$  for the proposal of Maritz, (ii) 4.32 and  $7.68 \times 10^{-6}$  for the sign test (or exactly  $7.15 \times 10^{-6}$  using the binomial), and (iii) 4.10 and  $2.05 \times 10^{-5}$  for the permutational  $t$ -test; so in this one example, (i) yielded the smallest approximate significance level. With the same measure of scale,  $h_{\tau_0}^{\ell}$ , but with a different definition of  $\psi(\cdot)$ , specifically  $\tilde{\psi}(y) = -1$  for  $y \leq -1$ ,  $\tilde{\psi}(y) = 0$  for  $-1 < y < 1$ ,  $\tilde{\psi}(y) = 1$ , for  $y \geq 1$ , the statistic  $T_{\tau_0}$  becomes Noether's (1973) statistic. In the absence of ties, Noether's statistic,  $B_{\tau_0} = (T_{\tau_0} + I - \ell + 1)/2$ , counts the number of positive differences among the  $I - \ell + 1$  differences with the largest  $|D_i - \tau_0|$ , and  $B_{\tau_0}$ 's null randomization distribution is binomial with sample size  $I - \ell + 1$  and probability of success  $\frac{1}{2}$ , so it generalizes the sign test, but for some  $\ell$  is more efficient for Normal data. For Noether's statistic, the sensitivity bounds in (4) turn out to be based on the binomial with sample size  $I - \ell + 1$  and probabilities of success  $1/(1+\Gamma)$  and  $\Gamma/(1+\Gamma)$ . Roughly speaking, if  $h_{\tau_0}^{\ell}$  and Huber's  $\tilde{\psi}(\cdot)$  are used, then  $T_{\tau_0}$  is the sum of a scaled mean and a sign statistic, or more precisely, the mean of the  $\ell - 1$  scaled differences  $(D_i - \tau_0)/h_{\tau_0}^{\ell}$  with the smallest  $|D_i - \tau_0|$  plus Noether's statistic for the  $I - \ell + 1$  observations with the largest  $|D_i - \tau_0|$ .

Popular in practice are redescending  $m$ -estimates, in which  $\psi(y)$  is not monotone, such as the biweight given by  $\psi(y) = y(1 - y^2)^2$  for  $|y| \leq 1$  and  $\psi(y) = 0$  for  $|y| > 1$ . Huber (1981, p. 103) expresses reservations about redescending  $\psi(\cdot)$ 's. For tests of  $H_0 : \tau = \tau_0$ , the difficulty is that if  $\tau_0$  is far from the true  $\tau$ , then a redescending  $\psi(\cdot)$  may attach very little weight to differences,  $D_i$ , that are close to  $\tau$  but far from  $\tau_0$ . Figure 2 contrasts Huber's  $\tilde{\psi}(\cdot)$  to the biweight for the randomization inference in Section 1, plotting  $\{T_{\tau_0} - E(T_{\tau_0})\}/(\text{var}(T_{\tau_0}))^{1/2}$  against  $\tau_0$ . There are horizontal lines at 0 and  $\pm 1.96$ . The

**Table 1**

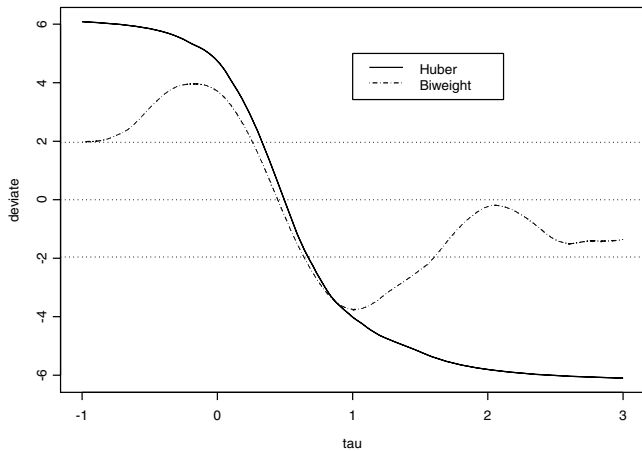
Sensitivity analysis for Welder data using an  $m$ -estimator and an  $m$ -test. The table displays the range of possible inferences for biases of various sizes, as measured by  $\Gamma$

	$\Gamma = 1$	$\Gamma = 2$	$\Gamma = 3$	$\Gamma = 4$
Maximum P-value	$1.04 \times 10^{-6}$	0.0014	0.017	0.057
Minimum P-value	$1.04 \times 10^{-6}$	$<10^{-10}$	$<10^{-10}$	$<10^{-10}$
Minimum $\hat{\tau}$	0.49	0.34	0.25	0.18
Maximum $\hat{\tau}$	0.49	0.67	0.78	0.86
Minimum 95% CI	$\tau \geq 0.36$	$\tau \geq 0.18$	$\tau \geq 0.07$	$\tau \geq -0.01$

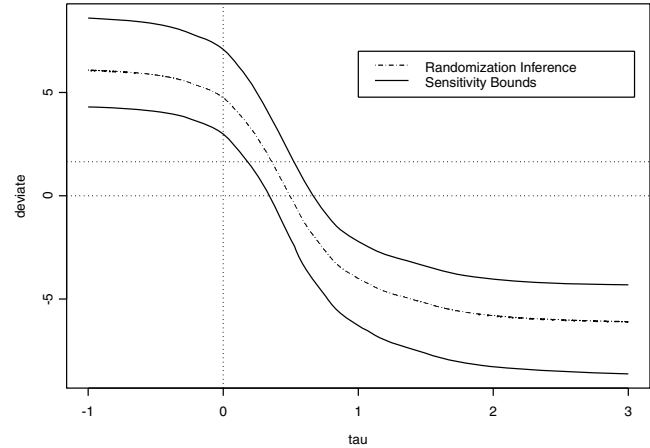
curves cut the horizontal line at 0 at the point estimates; e.g., for Huber's  $\tilde{\psi}(\cdot)$  at  $\tau_0 = 0.49$ , as in Section 1. The approximate 95% confidence sets are the sets of values of  $\tau_0$  for which the curve is between  $\pm 1.96$ . For Huber's  $\tilde{\psi}(\cdot)$ , the confidence set is an interval. In contrast, the biweight would need some additional rule besides the value of the standardized deviate to realize that  $\tau_0 = 2$  is not a plausible hypothesis; compare with Figure 1. Although one could certainly patch up the situation with redescending  $m$ -estimates, there is some convenience, and perhaps little loss, in using a  $T_{\tau_0}$  that is monotone. Among such estimates, Andrews et al. (1972, p. 240) favor a 25% trimmed mean, which is close to Maritz's proposal, namely Huber's  $\tilde{\psi}(\cdot)$  with  $\ell = 0.5 \times I$ .

3.3 Example of Sensitivity Analysis Using  $m$ -Estimates, Tests, and Confidence Intervals

The sensitivity analysis for the welder data in Section 1 is displayed in Table 1 and depicted for  $\Gamma = 1$  and  $\Gamma = 2$  in Figure 3. The  $m$ -estimate uses the suggestions of Maritz (1979), as in Section 1, scaling by  $h_{\tau_0} = \text{median } |D_i - \tau_0|$  and Huber's  $\tilde{\psi}(\cdot)$ .



**Figure 2.** Welder data: standardized deviate for testing  $H_0 : \tau = \tau_0$  under random assignment plotted against  $\tau_0$ , for Huber and Biweight  $\psi(\cdot)$ . Horizontal lines at 0 and  $\pm 1.96$  define the point estimates and the approximate 95% confidence sets.



**Figure 3.** Sensitivity analysis for the welder data: standardized deviate for testing  $H_0 : \tau = \tau_0$  plotted against  $\tau_0$ , for randomization inference,  $\Gamma = 1$ , and for the sensitivity bounds at  $\Gamma = 2$ , using Huber's  $\tilde{\psi}(\cdot)$ . Horizontal lines are at 0 and 1.65 for point estimates and approximate, one-sided 95% confidence sets. Vertical line at 0 for testing  $H_0 : \tau = 0$ .

For  $\Gamma = 1$ —that is, for randomization inference as in Section 1—there is a single null distribution for  $T_{\tau_0}$ , yielding a single, one-sided significance level,  $1.04 \times 10^{-6}$  for testing  $H_0 : \tau = 0$ , a single point estimate,  $\hat{\tau} = 0.49$ , and a single one-sided 95% confidence interval  $\tau \geq 0.36$ . In other words, the bounds in (4) are equal and are depicted in Figure 3 by the dashed curve. The dashed curve cuts the horizontal line at zero at the point estimate,  $\hat{\tau} = 0.49$ ; it cuts the horizontal line at 1.65 at the endpoint of the one-sided 95% confidence interval,  $\tau \geq 0.36$ ; it cuts the vertical line at zero at the standardized deviate 4.74 for testing  $H_0 : \tau = 0$  yielding an approximate one-sided significance level of  $1.04 \times 10^{-6}$ .

If the unobserved bias were of magnitude  $\Gamma = 2$ , then two subjects matched for the observed covariates,  $\mathbf{x}_{i1} = \mathbf{x}_{i2}$ , might differ in their chances of exposure to the treatment by a factor of  $\Gamma = 2$  because they differ in terms of an unobserved covariate  $u_{i1} \neq u_{i2}$ . In this case, the null distribution of  $T_{\tau_0}$  is unknown but is bounded by two known distributions, exactly bounded by (4). The solid curves in Figure 3 depict the standardized deviates for the bounds (6). For testing  $H_0 : \tau = 0$ , there is a range of possible standardized deviates, from 2.98 to 7.09, where the two solid curves cut the vertical line at 0 in Figure 3, yielding a range of one-sided significance levels from 0.0014 to  $<10^{-10}$  in Table 1. There is also a range of possible point estimates,  $\hat{\tau}$ , from 0.34 to 0.67, depicted in Figure 3 by the cutting of the horizontal line at 0 by the two solid curves. Similarly, there is a range of possible endpoints for the one-sided confidence interval, the minimum endpoint being  $\tau \geq 0.18$ , where the lower solid curve in Figure 3 cuts the horizontal line at 1.65.

In short, to explain away the observed results in Section 1 as a bias due to failure to control for an unobserved covariate  $u_{ij}$ , that covariate would have to be associated with a  $\Gamma = 4$  fold increase in the chance of a career as a welder and also be

**Table 2**

Upper bounds on the one-sided significance level testing no effect using several  $m$ -tests

	$\Gamma = 1$	$\Gamma = 2$	$\Gamma = 3$
Permutational $t$ -test, $\ell = I$	0.0000205	0.0037	0.023
$\ell \doteq 29 \doteq 0.75 \times I$	0.0000027	0.0020	0.018
Maritz, $\ell = 20 \doteq 0.5 \times I$	0.0000010	0.0014	0.017
Sign test, $\ell = 1$	0.0000077	0.0087	0.083

a strong predictor of DNA elution rates, and even with a bias of this magnitude, the minimum point estimate is 0.18. A bias of  $\Gamma = 3$  might explain as much as half the estimated effect, from  $\hat{\tau} = 0.49$  to as small as  $\hat{\tau} = 0.25$ , but the significance level testing no effect would still be no larger than 0.017. This is a fairly high degree of insensitivity to hidden bias: only fairly large biases could explain as not causal the observed association between treatment and outcome.

Table 2 contrasts four  $m$ -tests of  $H_0 : \tau = 0$ , specifically,  $\ell = I$  for the permutational  $t$ -test, for  $\ell = 20 \doteq 0.5 \times I$  as proposed by Maritz (1979),  $\ell = 1$  for the sign test, and for an intermediate case,  $\ell = 29 \doteq 0.75 \times I$ . In this one situation, the original proposal of Maritz (1979) was least sensitive to unobserved bias, and neither the permutation  $t$ -test nor the sign test dominated the other. The relative performance of different  $m$ -estimates for  $\Gamma = 1$  varies with the distribution of  $\epsilon_i$ ; see Huber (1981, Section 3), Andrews et al. (1972).

#### 3.4 A Check on the Normal Approximation

How accurate are the bounds on significance levels obtained using the Normal approximation (6)? For matched pairs, exact bounds could be obtained from a permutation distribution (2) with  $2^I$  sign changes in the  $I$  pairs. With  $I = 20$  pairs, there are  $2^I = 2^{20} = 1,048,576$  permutations. As a check on the approximation, 100 samples,  $D_1, \dots, D_{20} \stackrel{i.i.d.}{\sim} N(1, 1)$  were drawn and Maritz (1979)'s version of the test was calculated, as in Section 3.3. For  $\Gamma = 1, 2$ , and 3, both the exact upper bounds in (4) and the approximations in (6) were calculated, yielding 600 significance levels, 100 exact levels for each of  $\Gamma = 1, 2$ , and 3, and 100 approximate levels for each of  $\Gamma = 1, 2$ , and 3. The three Pearson correlations between 100 exact and 100 approximate significance levels for  $\Gamma = 1, 2$ , and 3 were each above 0.999. For  $\Gamma = 1, 2$ , and 3, the medians of the absolute differences between the 100 exact and 100 approximate significance levels were, respectively 0.00021, 0.0047, and 0.0072. The approximate significance levels were slightly conservative: for  $\Gamma = 1, 2$ , and 3, respectively, 100/100, 92/100, and 72/100 approximations were larger than the exact values. Among all approximate significance levels less than 0.05, the single largest absolute difference was between an exact 0.0239 and an approximate 0.0418 for  $\Gamma = 3$ . Even for just  $I = 20$  pairs, the approximation was adequate.

#### 3.5 Alternatives to the Model of an Additive Treatment Effect

Under the additive model in Section 3.1,  $r_{Tij} - r_{Cij} = \tau$ , so  $D_i = \tau + \epsilon_i$ , where  $\epsilon_i = (2Z_{i1} - 1)(r_{C11} - r_{C12})$ . A general model has  $2I$  effect parameters,  $r_{Tij} - r_{Cij} = \tau_{ij}$ , so the observed response is  $R_{ij} = r_{Cij} + \tau_{ij}Z_{ij}$ . Consider the  $2I$  dimensional hypothesis  $H_0 : \tau_{ij} = \tau_{0ij}$ ,  $i = 1, \dots, I$ ,  $j = 1, 2$ . Using the data and the hypothesis,  $H_0$ , calculate adjusted responses,

$R_{ij} - \tau_{0ij}Z_{ij}$ , and from these, calculate treated-minus-control differences in adjusted responses in pair  $i$ ,  $D_i - Z_{i1}\tau_{0i1} - Z_{i2}\tau_{0i2}$ , which would equal  $\epsilon_i = (2Z_{i1} - 1)(r_{C11} - r_{C12})$  if the hypothesis were true, so  $H_0$  may be tested, exactly as before, by applying the methods in Section 3.1 to these adjusted differences. There is no difficulty in principle with testing such a  $2I$  dimensional hypothesis, but it is tedious to specify  $2I$  dimensional hypotheses, so common practice assumes the  $2I$  parameters  $\tau_{ij}$  are functions of one or a few other parameters. For instance, the hypothesis  $H_0 : \tau_{ij} = (\rho_0 - 1)r_{Cij}$  is the hypothesis of a multiplicative effect,  $H_0 : r_{Tij} = \rho_0 r_{Cij}$ , and then  $R_{ij} - \tau_{0ij}Z_{ij} = R_{ij}/\rho_0^{Z_{ij}}$ .

Alternatively, assignment to treatment or to control may manipulate the dose of an active agent, so  $ij$  receives dose  $w_{Tij}$  if treated,  $Z_{ij} = 1$ , and dose  $w_{Cij}$  if control,  $Z_{ij} = 0$ , so the observed dose is  $W_{ij} = Z_{ij}w_{Tij} + (1 - Z_{ij})w_{Cij}$ . Effect is proportional to dose if  $\tau_{ij} = \beta(w_{Tij} - w_{Cij})$ . If  $H_0 : \beta = \beta_0$  were true, then  $R_{ij} - \beta_0 W_{ij} = r_{Tij} - \beta_0 w_{Tij} = r_{Cij} - \beta_0 w_{Cij} = a_{ij}$ , say, is constant, not varying with  $\mathbf{Z}$ , and can be calculated from the observed data for all  $2I$  subjects, so again the methods in Section 3.1 apply. See Greevy et al. (2004) for discussion, an example, and related references.

With certain statistics, including the sign test and Noether's test, the general hypothesis  $H_0 : \tau_{ij} = \tau_{0ij}$  may be inverted to form interpretable confidence statements for 'attributable effects'; see, for instance, Rosenbaum (2003) in the case of Wilcoxon's signed rank test.

## 4. Multiple Controls

### 4.1 Example: Chromosome Aberrations from Drugs Used to Treat Tuberculosis

Isoniazid (H), rifampicin (R), and pyrazinamide (Z) are used to treat tuberculosis, but may cause genetic damage. Rao, Gupta, and Thomas (1991) compared chromosome aberrations for patients treated with two standard regimes: (i) HRZ consisting of daily doses of 300 mg of isoniazid, 450 mg of rifampicin, and 1.5 g of pyrazinamide, and (ii) H2R2Z2 consisting of twice weekly doses of 600 mg of isoniazid, 900 mg of

**Table 3**

Frequency of aberrant metaphases after 2 months of treatment in matched pairs or triples, and their scores  $q_{ij}$  for testing  $H_0 : \tau = 0$

$i$	HRZ	H2R2Z2	H2R2Z2	$q_{i1}$	$q_{i2}$	$q_{i3}$
1	5	5	5	0.000	0.000	0.000
2	11	5		0.500	-0.500	
3	8	5	5	0.667	-0.333	-0.333
4	8	3		0.500	-0.500	
5	10	7	8	0.555	-0.444	-0.111
6	8	5	3	0.667	-0.111	-0.555
7	10	5		0.500	-0.500	
8	8	4		0.500	-0.500	
9	8	7	3	0.444	0.222	-0.667
10	6	3		0.500	-0.500	
11	4	6		-0.333	0.333	
12	9	4		0.500	-0.500	
13	11	3		0.500	-0.500	
14	8	8		0.000	0.000	
15	8	6		0.333	-0.333	

rifampicin, and 1.5 g of pyrazinamide, so the dose was much higher with HRZ. Fifteen patients received HRZ and 20 received H2R2Z2, and the frequency of aberrant metaphases was determined before treatment and after 2 months of treatment. Using the method in Ming and Rosenbaum (2001), HRZ patients were matched to one or two H2R2Z2 patients to minimize the total squared difference in pretreatment frequencies. Here, HRZ is labeled the “treatment condition,”  $I = 15$ , and  $n_i = 2$  for ten pairs,  $n_i = 3$  for five triples, and within a pair or a triple, pretreatment frequencies were identical or differed by one. Posttreatment frequencies are in Table 3.

4.2 Notation and Method for Multiple Controls

With matched pairs, there were  $2^I$  possible treatment assignments that changed the signs of  $\epsilon_i = (2Z_{i1} - 1)(r_{C1} - r_{C2})$ . With matched sets, there are  $\prod_{i=1}^I n_i$  possible treatment assignments,  $\mathbf{Z}$ , or  $2^{10} \times 3^5$  possible  $\mathbf{Z}$  for Table 3. There are  $n_i - 1 \geq 1$  differences between the one treated subject in set  $i$  and the  $n_i - 1$  controls in set  $i$ , and  $\mathbf{D} = (D_{11}, \dots, D_{1, n_i-1}, D_{21}, \dots, D_{I, n_I-1})$  is of dimension  $\sum(n_i - 1)$ . For instance, for  $i = 1$  with  $n_1 = 3$  in Table 3, there are two treatment-minus-control differences,  $D_{11}$  and  $D_{12}$ , and these would be  $D_{11} = R_{11} - R_{12}$  and  $D_{12} = R_{11} - R_{13}$  if, as in Table 3, the first patient received HRZ,  $Z_{i1} = 1$ , but they would have been  $D_{11} = R_{12} - R_{11}$  and  $D_{12} = R_{12} - R_{13}$  if the second patient had received HRZ,  $Z_{i2} = 1$ , and they would have been  $D_{11} = R_{13} - R_{11}$  and  $D_{12} = R_{13} - R_{12}$  if the third patient had received HRZ,  $Z_{i3} = 1$ . Write  $\mathcal{N}_{ij} = \{1, 2, \dots, j - 1, j + 1, \dots, n_i\}$  and for  $1 \leq j \leq n_i, k \in \mathcal{N}_{ij}$ , write  $\kappa_{ijk} = r_{Cij} - r_{Cik}$ , noting that  $\kappa_{ijj}$  is not defined and that  $\kappa_{ijk} = -\kappa_{ikj}$  so  $|\kappa_{ijk}| = |\kappa_{ikj}|$ . One of the  $n_i$  subjects in matched set  $i$  receives treatment; specifically, it is subject number  $J_i = \sum_{j=1}^{n_i} j \cdot Z_{ij}$  in set  $i$ . The  $n_i - 1$  treated-minus-control differences in set  $i$  are  $R_{i, J_i} - R_{ik}$  for  $k \in \mathcal{N}_{i, J_i}$ , which equal  $\kappa_{i, J_i, k} + \tau$  for  $k \in \mathcal{N}_{i, J_i}$ .

The measure of scale,  $h_{\tau_0}$ , needs to be generalized. For  $1 \leq j < k \leq n_i$ , at the true  $\tau$ , the  $\binom{n_i}{2}$  comparisons  $|\kappa_{ijk}| = |r_{Cij} - r_{Cik}| = |(R_{ij} - \tau Z_{ij}) - (R_{ik} - \tau Z_{ik})|$  yield the  $\binom{n_i}{2}$  unsigned values of  $|\kappa_{ijk}| = |\kappa_{ikj}|$ . Write  $\kappa_{ijk}^{(\tau_0)} = |(R_{ij} - \tau_0 Z_{ij}) - (R_{ik} - \tau_0 Z_{ik})|$ , noting that if  $H_0 : \tau = \tau_0$  were true, then  $\kappa_{ijk}^{(\tau_0)} = \kappa_{ijk}$ . When testing  $H_0 : \tau = \tau_0$ , let  $h_{\tau_0}^\ell$  be the  $\ell$ th of the  $\sum \binom{n_i}{2}$  absolute differences  $|\kappa_{ijk}^{(\tau_0)}|, 1 \leq j < k \leq n_i$ , when sorted from smallest to largest; for instance, their median. If each  $n_i = 2$  for matched pairs, this agrees with the definition in Section 3. In Table 3, there were ten pairs,  $n_i = 2$ , and five triples,  $n_i = 3$ , so there are  $10 \binom{2}{2} + 5 \binom{3}{2} = 10 + 15 = 25$  absolute differences  $|\kappa_{ijk}^{(\tau_0)}|, 1 \leq j < k \leq n_i$  for each hypothesis  $H_0 : \tau = \tau_0$ , and for testing  $H_0 : \tau = 0$ , their median is  $h_0^{13} = 3$ .

The generalized statistic  $T_{\tau_0}$  will now be defined. As in (3),  $T_{\tau_0}$  is a (weighted) sum of terms  $sign(D_{im}) \psi(|D_{im}|/h_{\tau_0}^\ell)$ , summing over all of the  $\sum(n_i - 1)$  treated-minus-control differences. The m-test of the hypothesis  $H_0 : \tau = \tau_0$  is based on the statistic

$$T_{\tau_0} = \sum_{i=1}^I \frac{1}{n_i} \sum_{j=1}^{n_i} Z_{ij} \sum_{k \in \mathcal{N}_{ij}} sign(\kappa_{ijk}^{(\tau_0)}) \psi \left( \frac{|\kappa_{ijk}^{(\tau_0)}|}{h_{\tau_0}^\ell} \right) = \sum_{i=1}^I \sum_{j=1}^{n_i} Z_{ij} q_{ij}, \tag{7}$$

where

$$q_{ij} = \frac{1}{n_i} \sum_{k \in \mathcal{N}_{ij}} sign(\kappa_{ijk}^{(\tau_0)}) \psi \left( \frac{|\kappa_{ijk}^{(\tau_0)}|}{h_{\tau_0}^\ell} \right). \tag{8}$$

For each  $i, 0 = \sum_{j=1}^{n_i} q_{ij}$ , because  $sign(\kappa_{ijk}^{(\tau_0)}) \psi(|\kappa_{ijk}^{(\tau_0)}|/h_{\tau_0}^\ell) = -sign(\kappa_{ikj}^{(\tau_0)}) \psi(|\kappa_{ikj}^{(\tau_0)}|/h_{\tau_0}^\ell)$  and they cancel when they appear in  $\sum_{j=1}^{n_i} q_{ij}$ . Table 3 displays the  $q_{ij}$  in (8) when testing  $H_0 : \tau = 0$ , scaling by the median,  $h_0^{13} = 3$ , and using Huber’s  $\tilde{\psi}(\cdot)$ .

In (7), matched sets are assigned weights  $1/n_i$  that would be optimal under a simple model in a randomized experiment using a ‘mean,’ that is using  $\psi(y) = y$  and  $h_{\tau_0}^\ell = 1$ . The one treated subject in matched set  $i$  is used in  $n_i - 1$  correlated, treated-minus-control differences with  $n_i - 1$  distinct controls. More precisely, in a randomized experiment, if  $r_{Cij} = \mu_i + \xi_{ij}$  where the  $\xi_{ij}$  are i.i.d., symmetric about zero, with finite variance  $\sigma^2$ , and if  $\psi(y) = y$ , then  $(n_i - 1)^{-1} \sum_{j=1}^{n_i} Z_{ij} \sum_{k \in \mathcal{N}_{ij}} sign(\kappa_{ijk}^{(\tau_0)}) \psi(|\kappa_{ijk}^{(\tau_0)}|)$  is  $\tau - \tau_0 + (n_i - 1)^{-1} \sum_{j=1}^{n_i} Z_{ij} \sum_{k \in \mathcal{N}_{ij}} (\xi_{ij} - \xi_{ik})$ , with expectation  $\tau - \tau_0$  and variance  $\sigma^2 \{1 + 1/(n_i - 1)\} = \sigma^2 n_i / (n_i - 1)$ ; so weighting these  $I$  unbiased estimates of  $\tau - \tau_0$  inversely as their variance implies weights proportional to  $(\frac{n_i - 1}{n_i})(\frac{1}{n_i - 1}) = \frac{1}{n_i}$  in (7).

The sensitivity analysis for  $T_{\tau_0}$  uses the straightforward computations in Gastwirth et al. (2000, Section 3.1). Unlike the situation in Section 3, with multiple controls instead of pairs, there is generally no random variable  $\bar{T}$  that is both a possible null distribution of  $T_{\tau_0}$  and stochastically as large as or larger than the null distribution of  $T_{\tau_0}$ , in the sense of (4); however, as  $I \rightarrow \infty$ , there is a possible null distribution of  $T_{\tau_0}$  that provides a sharp upper bound on  $\Pr(T_{\tau_0} \geq k)$  asymptotically. The bound is found by finding the pattern of  $u_{ij}$ ’s in (2) that maximizes the expectation of  $T_{\tau_0}$ , and if that pattern is not unique, then selecting one that maximizes the variance of  $T_{\tau_0}$ ; asymptotically, this maximizes the upper tail probability. Specifically, let  $q_{i(1)} \leq q_{i(2)} \leq \dots \leq q_{i(n_i)}$  be the ordered  $q_{ij}$  in (8) for matched set  $i$ . As shown in Gastwirth et al. (2000, Section 3.1), in matched set  $i$ , the largest null expectation of  $\sum_{j=1}^{n_i} Z_{ij} q_{ij}$  under (2) is

$$\eta_{\Gamma i} = \max_{a \in \{1, \dots, n_i - 1\}} \left\{ \frac{\sum_{j=1}^a q_{i(j)} + \Gamma \sum_{j=a+1}^{n_i} q_{i(j)}}{a + \Gamma(n_i - a)} \right\}. \tag{9}$$

Let  $A_i \subseteq \{1, \dots, n_i - 1\}$  be the set of  $a$ ’s that produce the maximum in (9); then for  $a \in A_i$ , find the largest variance of  $\sum_{j=1}^{n_i} Z_{ij} q_{ij}$  as

$$\nu_{\Gamma i}^2 = \max_{a \in A_i} \left\{ \frac{\sum_{j=1}^a q_{i(j)}^2 + \Gamma \sum_{j=a+1}^{n_i} q_{i(j)}^2}{a + \Gamma(n_i - a)} - \eta_{\Gamma i}^2 \right\}. \tag{10}$$

where (10) is a maximum over  $A_i$  not over  $\{1, \dots, k\}$ . As  $I \rightarrow \infty$ , for  $k > \sum \eta_{\Gamma i}$ , a sharp upper bound on  $\Pr(T_{\tau_0} \geq k)$  is obtained as  $1 - \Phi\{(k - \sum \eta_{\Gamma i}) / \sqrt{\sum \nu_{\Gamma i}^2}\}$ , where  $\Phi(\cdot)$  is the

standard Normal distribution. If each set  $i$  is a pair,  $n_i = 2$ , then  $-q_{i(1)} = q_{i(2)} = \psi_i/2$ , and this upper bound on  $\Pr(T_{\tau_0} \geq k)$  is identical to (6).

#### 4.3 Sensitivity Analysis for the Example with Multiple Controls

In Section 4.1, in the absence of unobserved bias ( $\Gamma = 1$ ), the one-sided significance level for testing the null hypothesis of no effect is 0.00028, and the one-sided 95% confidence interval for an additive effect is  $\tau \geq 1.98$ . An unobserved  $u_{ij}$  that doubled the odds of exposure to HRZ rather than H2R2Z2 ( $\Gamma = 2$ ) could yield, at most, a significance level of 0.0090 for testing no effect, and the 95% confidence interval is at least  $\tau \geq 0.94$ . For  $\Gamma = 3$  and 4, the bounds on the significance level are, respectively, 0.032 and 0.065, and the bounds on the confidence interval are  $\tau \geq 0.25$  and  $\tau \geq -0.29$ , so a quite large bias of  $\Gamma = 4$  could just barely explain the observed association.

### 5. Covariance Adjustment of Matched Pairs or Sets

The methods in Sections 3 and 4 may be combined with covariate adjustment as described in Rosenbaum (2002b). In that approach,  $H_0 : \tau = \tau_0$  is tested by calculating adjusted responses,  $R_{ij} - Z_{ij} \tau_0$ , regressing these on covariates to obtain residuals, and applying a permutation test to these residuals. The test is inverted to obtain confidence intervals and point estimates. If  $H_0 : \tau = \tau_0$  is true in a randomized experiment, this test has its nominal level, and a sensitivity analysis may be applied in an observational study; see Rosenbaum (2002b). In simulated randomized experiments, Greevy (2004, Section 4) found the method had the correct level and could meaningfully increase power or shorten confidence intervals, even when the covariance adjustment model was incorrect. Instead of using a rank test, as in Rosenbaum (2002b), the permutation distribution of the m-test based on  $T_{\tau_0}$  may be used instead; the logic underlying the procedure is unchanged.

The method will be applied to the example in Section 4.1, adjusting for baseline frequency by matching and covariance adjustment by m-estimation with Huber weights, as in Splus, without alignment. Specifically, to test  $H_0 : \tau = \tau_0$ , the adjusted responses,  $R_{ij} - Z_{ij} \tau_0$ , were regressed on the baseline frequencies, and the method in Section 4 was applied to their residuals. Because the matching in Section 4.1 was very tight, with a maximum difference of 1, the covariance adjustment barely altered the results in Section 4.3: with  $\Gamma = 1$ , the significance level for testing no effect became 0.00033 (as opposed to 0.00028 in Section 4.3) and the one-sided 95% confidence interval became  $\tau \geq 1.99$  (as opposed to  $\tau \geq 1.98$  in Section 4.3); for  $\Gamma = 3$  the confidence interval became  $\tau \geq 0.16$  (as opposed to  $\tau \geq 0.25$  in Section 4.3). One expects that covariance adjustment may shorten confidence intervals in other contexts in which important covariates are only partially controlled by matching.

### 6. Discussion

Huber's m-estimates include least squares estimates, maximum likelihood estimates, and a large class of robust estimates that give extreme observations a controlled level of influence. In particular, m-estimates can be designed to be smooth functions of observed data, so they may handle coarse, heavily tied data more gracefully than do R-estimates. Using

certain ideas of Maritz (1979, 1995, Section 2.8), a sensitivity analysis for m-estimates, tests, and confidence intervals was proposed for use in matched observational studies to address the possibility that an important unobserved covariate was not controlled by matching on observed covariates. The method is applicable to matched pairs, to matched sets with one or more controls, and to situations which combine matching with covariance adjustment.

#### ACKNOWLEDGEMENTS

This work was supported by grant SES-0345113 from the U.S. National Science Foundation.

#### REFERENCES

- Andrews, D. F., Bickel, P. J., Hampel, F. R., Huber, P. J., Rogers, W. H., and Tukey, J. W. (1972). *Robust Estimates of Location*. Princeton: Princeton University Press.
- Copas, J. and Eguchi, S. (2001). Local sensitivity approximations for selectivity bias. *Journal of the Royal Statistical Society B* **63**, 871–896.
- Cornfield, J., Haenszel, W., Hammond, E., Lilienfeld, A., Shimkin, M., and Wynder, E. (1959). Smoking and lung cancer. *Journal of the National Cancer Institute* **22**, 173–203.
- Fisher, R. A. (1935). *The Design of Experiments*. Edinburgh: Oliver & Boyd.
- Gastwirth, J. L. (1992). Methods for assessing the sensitivity of statistical comparisons used in Title VII cases to omitted variables. *Jurimetrics* **33**, 19–34.
- Gastwirth, J. L., Krieger, A. M., and Rosenbaum, P. R. (2000). Asymptotic separability in sensitivity analysis. *Journal of the Royal Statistical Society B* **62**, 545–555.
- Greevy, R. A. (2004). Noncompliance, covariance adjustment, and matching in randomized controlled trials. Ph.D. dissertation. University of Pennsylvania, Philadelphia.
- Greevy, R. A., Silber, J., Cnaan, A., and Rosenbaum, P. R. (2004). Randomization inference with imperfect compliance in the ACE-inhibitor after anthracycline randomized trial. *Journal of the American Statistical Association* **99**, 7–15.
- Huber, P. J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics* **35**, 73–101.
- Huber, P. J. (1981). *Robust Statistics*. New York: Wiley.
- Imbens, G. W. (2003). Sensitivity to exogeneity assumptions in program evaluation. *American Economic Review* **93**, 126–132.
- Jureková, J. and Sen, P. K. (1996). *Robust Statistical Procedures*. New York: Wiley.
- Lin, D. Y., Psaty, B. M., and Kronmal, R. A. (1998). Assessing the sensitivity of regression results to unmeasured confounders in observational studies. *Biometrics* **54**, 948–963.
- Maritz, J. (1979). Exact robust confidence intervals for location. *Biometrika* **66**, 163–166.
- Maritz, J. S. (1995). *Distribution-Free Statistical Methods*. London: Chapman & Hall.
- Ming, K. and Rosenbaum, P. (2001). Optimal match with variable controls by the assignment algorithm. *Journal of Computational and Graphical Statistics* **10**, 455–463.

- Neyman, J. (1923). On the application of probability theory to agricultural experiments: Essay on principles, Section 9. Reprinted in *Statistical Science* **5**, 463–480.
- Noether, G. (1973). Some distribution-free confidence intervals for the center of a symmetric distribution. *Journal of the American Statistical Association* **68**, 716–719.
- Rao, V. V. N. G., Gupta, E. V. V., and Thomas, I. M. (1991). Chromosomal aberrations in tuberculosis patients before and after treatment with short-term chemotherapy. *Mutation Research* **259**, 13–19.
- Robins, J. M. (1999). Association, causation and marginal structural models. *Synthese* **121**, 151–179.
- Robins, J. M., Rotnitzky, A., and Scharfstein, D. (1999). Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference. In *Statistical Models in Epidemiology*, E. Halloran and D. Berry (eds), pp. 1–94. New York: Springer.
- Rosenbaum, P. R. (1993). Hodges-Lehmann point estimates of treatment effect in observational studies. *Journal of the American Statistical Association* **88**, 1250–1253.
- Rosenbaum, P. R. (2002a). *Observational Studies*, 2nd edition. New York: Springer.
- Rosenbaum, P. R. (2002b). Covariance adjustment in randomized experiments and observational studies (with discussion). *Statistical Science* **17**, 286–327.
- Rosenbaum, P. R. (2003). Exact confidence intervals for non-constant effects by inverting the signed rank test. *American Statistician* **57**, 132–138.
- Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* **66**, 688–701.
- Welch, B. L. (1937). On the z-test in randomized blocks. *Biometrika* **29**, 21–52.
- Werfel, U., Langen, V., Eickhoff, I., Schoonbrood, J., Vahrenholz, C., Brauksiepe, A., Popp, W., and Norpoth, K. (1998). Elevated DNA strand breakage frequencies in lymphocytes of welders exposed to chromium and nickel. *Carcinogenesis* **19**, 413–418.

Received July 2005. Revised September 2006.

Accepted September 2006.