12-1-2015

Strong Control of the Familywise Error Rate in Observational Studies that Discover Effect Modification by Exploratory Methods

Jesse Y. Hsu
University of Pennsylvania

José R. Zubizarreta

Dylan S. Small
University of Pennsylvania

Paul R. Rosenbaum
University of Pennsylvania

Recommended Citation
Hsu, J. Y., Zubizarreta, J. R., Small, D. S., & Rosenbaum, P. R. (2015). Strong Control of the Familywise Error Rate in Observational Studies that Discover Effect Modification by Exploratory Methods. *Biometrika, 102* (4), 767-782. http://dx.doi.org/10.1093/biomet/asv034

Strong Control of the Familywise Error Rate in Observational Studies that Discover Effect Modification by Exploratory Methods

Abstract
An effect modifier is a pretreatment covariate that affects the magnitude of the treatment effect or its stability. When there is effect modification, an overall test that ignores an effect modifier may be more sensitive to unmeasured bias than a test that combines results from subgroups defined by the effect modifier. If there is effect modification, one would like to identify specific subgroups for which there is evidence of effect that is insensitive to small or moderate biases. In this paper, we propose an exploratory method for discovering effect modification, and combine it with a confirmatory method of simultaneous inference that strongly controls the familywise error rate in a sensitivity analysis, despite the fact that the groups being compared are defined empirically. A new form of matching, strength-*k* matching, permits a search through more than *k* covariates for effect modifiers, in such a way that no pairs are lost, provided that at most *k* covariates are selected to group the pairs. In a strength-*k* match, each set of *k* covariates is exactly balanced, although a set of more than *k* covariates may exhibit imbalance. We apply the proposed method to study the effects of the earthquake that struck Chile in 2010.

Keywords
design sensitivity, effect modification, integer programming, matched sampling, power of a sensitivity analysis, observational study, sensitivity analysis, truncated product of P-values

Disciplines
Business | Business Analytics | Design of Experiments and Sample Surveys | Management Sciences and Quantitative Methods | Statistical Models | Statistics and Probability

Strong control of the family-wise error rate in observational studies that discover effect modification by exploratory methods

Jesse Y. Hsu[1], José R. Zubizarreta, Dylan S. Small, Paul R. Rosenbaum
University of Pennsylvania and Columbia University

Abstract. An effect modifier is a pretreatment covariate such that the magnitude of the treatment effect or its stability changes with the level of the covariate. Generally, other things being equal, larger treatment effects and less heterogeneous treatment effects are less sensitive to unmeasured biases in observational studies. It is known that when there is effect modification, an overall test that ignores an effect modifier may report greater sensitivity to unmeasured bias than a test that combines results at different levels of the effect modifier. This known combined test reports that there is evidence of an effect somewhere that is insensitive to bias of a certain magnitude, but it does not draw inferences about affected subgroups. If there is effect modification, one would like to identify specific subgroups for which there is evidence of effect that is insensitive to small or moderate biases. In the current paper, we propose an exploratory method for discovering effect modification combined with a confirmatory method of simultaneous inference that strongly controls the family-wise error rate in a sensitivity analysis, despite the fact that the groups being compared are defined empirically. Groups of treatment-control matched pairs are identified using a special version of CART. A new form of matching, strength k matching, permits CART to search through many covariates for effect modifiers, yet no pairs are lost providing CART settles on a tree that uses at most k covariates. In a strength k match, we can build the CART tree using more than k variables, let CART decide which k or fewer variables are the best candidates as effect modifiers, and know that all individuals can be matched exactly for the variables CART selects. We apply the method to study the effects of the powerful earthquake that struck Chile in 2010.

Keywords: Design sensitivity; effect modification; integer programming; matched sampling; power of a sensitivity analysis; observational study; sensitivity analysis; truncated product of P-values

[1]Jesse Y. Hsu is a post-doctoral fellow, Department of Statistics, The Wharton School, University of Pennsylvania Philadelphia, PA 19104-6340 USA and the Center for Outcomes Research, The Children's Hospital of Philadelphia (E-mail: *hsu9@wharton.upenn.edu*). José R. Zubizarreta is assistant professor, Division of Decision, Risk and Operations, Columbia University Business School (E-mail: *jz2313@columbia.edu*). Dylan S. Small is associate professor and Paul R. Rosenbaum is professor, Department of Statistics, The Wharton School, University of Pennsylvania Philadelphia, PA 19104-6340 USA, (E-mail: *dsmall@wharton.upenn.edu* and *rosenbaum@wharton.upenn.edu*). 19 Dec 2013

1 Introduction: effect modification; attentive inference

1.1 Attentive inference: using information that will be fixed anyway

It is common in practice to alter a statistical analysis to reflect features discovered in the data at hand, and statistical theory is constantly trying to catch up with these diverse practices, that is, to appropriately allow for repeated use of the same data. If no care is taken when analyses are selected in light of the data analyzed, then desirable properties of statistical procedures may evaporate: tests with nominal level α may reject true null hypotheses with probability substantially greater than α, and confidence intervals with nominal coverage $1-\alpha$ may cover the true parameter with probability substantially less than $1-\alpha$. There are, of course, many approaches that permit multiple uses of the same data, and "attentive inference" is one of the simplest though least developed of these.

If the observed data are (A, B) and inference will be based on the conditional distribution of B given A, $\Pr(B|A)$, then an inference is "attentive" if the method of inference is selected having examined A without examining B; otherwise, if A is not examined the inference is "inattentive," whereas if A and B are both examined the inference is "not attentive." For instance, if the inference were a hypothesis test and the null distribution of the selected test statistic is derived from the null distribution of B given A, then one could alter the choice of test statistic on the basis of an examination of A alone without altering the level of the test. An important class of such tests are permutation (or randomization) tests. In particular, the only two-sample tests that have level α for all continuous distributions are permutation tests formed by conditioning on the pooled sample order statistics, with an analogous result for two-sample tests stratified for covariates (Lehmann and Romano 2005, §5.8, Theorem 5.8.1); here, A is the order statistic for the two sample test and the stratified order statistic for the stratified two-sample test. In the two-sample problem, Hogg, Fisher and Randles (1975) adaptively select a test statistic on the basis of the tail-behavior of the order statistic, and Jones (1979) takes a parallel approach to testing symmetry about zero in the one-sample problem. There are other forms of adaptive inference that use both A and B, but unlike attentive inference, these forms of adaptive inference must take account of the repeated use of B because they are "not attentive"; see, for instance, Donegani (1991) and Rosenbaum (2012). Expressed informally, attentive inference makes use of information that is freely available for use, whereas adaptive inferences that are "not attentive" must pay a price for adaptation, perhaps a price worth

paying but a price nonetheless.

1.2 The 2010 earthquake in Chile

On 27 February 2010, a powerful earthquake of magnitude 8.8 struck off the coast of Chile. Its epicenter was located off the coast of central Chile, near the country's second largest city, Concepción (USGS 2011a). Depending on the city, during 3 to 6 minutes the earthquake shook the center of the country with massive force. It moved the city of Concepción 3.04 meters to the west (Pollitz et al. 2011). The earthquake was followed by a tsunami and 525 people were killed (Interior 2011), almost 500,000 homes suffered severe damage, and nearly 2,000,000 people were injured (La Tercera 2010). The earthquake was the 4th strongest earthquake in the world in the last 50 years (USGS 2011b).

About two months before the earthquake, the Chilean government had completed its national socioeconomic survey (CASEN). To measure the impact of the 2010 Chilean earthquake, the Chilean government decided to reinterview a subsample of the CASEN following the earthquake, thereby creating rare longitudinal data before and after a major disaster. The Post Earthquake Survey (EPT) was a national longitudinal household survey conducted between May and June 2010, nearly two months after the earthquake. The EPT consisted of 22,456 households out of the 71,460 original households in CASEN 2009. For a description of the EPT see Mideplan (2011).

The effect of the earthquake on posttraumatic stress was analyzed by Zubizarreta, Cerdá and Rosenbaum (2013). In the current paper, we examine the effect of the earthquake on the change in individual work income from before the earthquake to after. In principle, a major earthquake might disrupt existing economic activity, thereby reducing work income, or it might create jobs in construction to repair damage done by the earthquake, so even the direction of the possible effect is in doubt.

We constructed 2106 matched pairs of two individuals, one in a region of Chile severely shaken by the earthquake, the other in a region remote from the earthquake. See Zubizarreta et al. (2013) for discussion of the geologic measure used to define these regions; however, essentially the severely shaken middle of Chile is compared to its north and south. A covariate is a variable measured prior to the earthquake, hence unaffected by the earthquake. The CASEN survey before the earthquake provided many covariates. The matching controlled for covariates from the CASEN: sex, marital status, number of persons in the household, self-reported health problem, self-reported health perception, quartile of

work income (before the earthquake), age, self-reported psychological problems, disability, health insurance status, years of education, employment status, per capita total household income, poverty status, housing status, quality of housing structure, and overcrowding. Figure 1 shows the covariate balance in these pairs for three continuous covariates, namely age in years, education in years and work income in pesos before the earthquake. Obviously, when subgroups are examined, it is important also to check for covariate balance within each subgroup.

The matching introduces a new technique called strength k matching, and this technique is described in detail in §4.2.

We wish to consider six covariates as possible effect modifiers. An effect modifier is essentially a covariate that interacts with the treatment, so that the treatment effect is not constant in size but rather varies with levels of the covariate. Here, that would mean that the effect of the earthquake on the change in work income is larger for some groups of individuals and smaller for others, where the groups are defined by some of the covariates. Hsu et al. (2013) found that effect modifiers affect sensitivity to bias from unmeasured covariates, because larger effects can be less sensitive to unmeasured biases; more precisely, effect modifiers affect the design sensitivity.

For brevity, we refer to these $V = 6$ candidate effect-modifiers as "the basic covariates," specifically: gender (male, female), health problems (yes, no), self-rated health (poor, fair, good), quartile of individual work income in 2009, number of persons in the household (1, 2, 3, 4 or 5, ≥ 6), and marital status (married/cohabiting versus other). Because most people, including especially many women and elderly individuals, did not have individual work income in 2009, the quartiles of work income defined only 3, not 4 groups. The six basic covariates define $2 \times 2 \times 3 \times 3 \times 5 \times 2 = 360$ types of individuals. A total of $I = 2106$ matched pairs were formed, so many of the 360 types of pairs are represented by only a moderate number of pairs. We allow the data to suggest a grouping of types of pairs so the groups have many pairs, and much of the technical work in the paper is concerned with appropriately allowing an analysis for a data-derived grouping of pairs. Until §4, we ignore the possibility of pairs inexactly matched for basic covariates; then §4 describes a simple approach, one actually used here, to incorporate them.

1.3 Outline: Can one strongly control the family-wise error rate in subgroup analyses when the subgroups were discovered empirically using the data?

Section 2 defines notation (§2.1), reviews randomization inference in experiments (§2.2) and sensitivity analysis in observational studies (§2.3), and then reviews the connection between these topics and effect modification (§2.4). The new results are in §3. In §3.1, Proposition 5 shows that a specific form of adaptive identification of effect modifiers does not alter the null sensitivity distribution; that is, it is attentive in the sense of §1.1. In §3.2, Proposition 8 uses the result from §3.1 to perform simultaneous inference with data-dependent groups, strongly controlling the family-wise error rate in a sensitivity analysis. A brief summary of the findings of §3 is given in §3.3. In the earthquake data in §5, six covariates are considered as candidates for effect modification, two covariates are selected to form three subgroups, and the method of §3 is applied. The earthquake example uses a new form of matching, strength $k = 3$ matching, with the consequence that the data may be exactly matched for any $k = 3$ of the six covariates. As reviewed in §2.4, asymptotically the power of a sensitivity analysis is determined by the design sensitivity; however, §6 examines finite-sample power using simulation.

2 Notation and review: randomization inference; sensitivity analysis

2.1 Notation: treatment effects, treatment assignments, observed and unobserved covariates

The data permit the construction of I pairs, $i = 1, \ldots, I$, of two subjects, $j = 1, 2$, one treated with $Z_{ij} = 1$, the other control with $Z_{ij} = 0$, so $Z_{i1} + Z_{i2} = 1$ for each i. Write $\mathbf{Z} = (Z_{11}, Z_{12}, \ldots, Z_{I2})^T$ for the $2I$-dimensional vector containing the Z_{ij}, and write $\mathcal{Z}$ for the set containing the 2^I possible values $\mathbf{z}$ of $\mathbf{Z}$, so $\mathbf{z} \in \mathcal{Z}$ if $\mathbf{z} = (z_{11}, \ldots, z_{I2})^T$ with $z_{i1} + z_{i2} = 1$ and $z_{ij} = 0$ or $z_{ij} = 1$ for each i, j. Conditioning on the event $\mathbf{Z} \in \mathcal{Z}$ is abbreviated as conditioning on $\mathcal{Z}$. Write $|\mathcal{S}|$ for the number of elements of a finite set $\mathcal{S}$; for instance, $|\mathcal{Z}| = 2^I$.

Subject ij has an observed covariate $(\mathbf{x}_{ij}, \mathbf{v}_{ij})$ used in matching and an unobserved covariate u_{ij} that is not controlled by matching. The V-dimensional covariate $\mathbf{v}_{ij}$ consists of V nominal covariates that are of interest as possible effect modifiers. A pair i is exactly matched for $\mathbf{v}_{ij}$ if $\mathbf{v}_{i1} = \mathbf{v}_{i2}$ and inexactly matched if $\mathbf{v}_{i1} \neq \mathbf{v}_{i2}$, and until §4 we assume all pairs are exactly matched. Let $\mathcal{V}$ be the set of possible values $\mathbf{v}$ of $\mathbf{v}_{ij}$, so there are

$|\mathcal{V}|$ possible values of $\mathbf{v}_{ij}$. In §1.2, $\mathbf{v}_{ij}$ contains $V = 6$ nominal covariates, and $\mathbf{v}_{ij}$ has $|\mathcal{V}| = 360$ possible values. Because u_{ij} is not observed, it is quite possible that $u_{i1} \neq u_{i2}$ for many or all i.

Each subject has a potential response r_{Tij} if treated with $Z_{ij} = 1$, a potential response r_{Cij} if assigned to control with $Z_{ij} = 0$, an observed response $R_{ij} = Z_{ij}\, r_{Tij} + (1 - Z_{ij})\, r_{Cij}$ under the treatment actually received, whereas the effect of the treatment, namely $r_{Tij} - r_{Cij}$, is not observed for any subject; see Neyman (1923) and Rubin (1974). Fisher's (1935) sharp null hypothesis of no treatment effect asserts $H_0 : r_{Tij} = r_{Cij}$, $\forall i, j$. Importantly, if H_0 were true, then $R_{ij} = r_{Cij}$ does not change with treatment assignment Z_{ij}, but if H_0 is false then at least some R_{ij} do change with Z_{ij}. Write $\mathcal{F} = \{(r_{Tij}, r_{Cij}, \mathbf{x}_{ij}, \mathbf{v}_{ij}, u_{ij}),$ $i = 1, \ldots, I,\ j = 1, 2\}$. The treated-minus-control pair difference in observed responses in pair i is $Y_i = (Z_{i1} - Z_{i2})\,(R_{i1} - R_{i2})$, and it equals $(Z_{i1} - Z_{i2})\,(r_{Ci1} - r_{Ci2})$ if H_0 is true. Also, write $\mathbf{r}_C = (r_{C11},\, r_{C12}, \ldots, r_{CI2})^T$ and $\mathbf{R} = (R_{11},\, R_{12}, \ldots, R_{I2})^T$ for the vectors of dimension $2I$, and $\mathbf{Y} = (Y_1, \ldots, Y_I)^T$ for the vector of dimension I. Effect modification refers to the possibility that the size of the effect, $r_{Tij} - r_{Cij}$, varies systematically with observed covariates, $(\mathbf{x}_{ij}, \mathbf{v}_{ij})$, and here we are focusing specifically on $\mathbf{v}_{ij}$ as possible effect modifiers.

2.2 Randomization inference in experiments

In a paired randomized experiment, subjects are paired on the basis of observed covariates, $(\mathbf{x}_{ij}, \mathbf{v}_{ij})$, and then a fair coin is flipped independently I times to determine the treatment assignments Z_{i1} with $Z_{i2} = 1 - Z_{i1}$; that is, $\Pr(Z_{ij} = 1 \mid \mathcal{F},\, \mathcal{Z}) = 1/2$ for each i, j and $\Pr(\mathbf{Z} = \mathbf{z} \mid \mathcal{F},\, \mathcal{Z}) = 2^{-I}$ for each $\mathbf{z} \in \mathcal{Z}$. Let $t(\mathbf{Z}, \mathbf{R})$ be a test statistic, that is, a function of the treatment assignment $\mathbf{Z}$ and the observed responses $\mathbf{R}$. The statistic $t(\mathbf{Z}, \mathbf{R})$ may depend also upon the observed covariates, but the notation does not indicate this explicitly. The null distribution of $t(\mathbf{Z}, \mathbf{R})$ under Fisher's H_0 in a paired randomized experiment is its permutation distribution, namely

$$\Pr\{t(\mathbf{Z}, \mathbf{R}) \geq k \mid \mathcal{F},\, \mathcal{Z}\} = \Pr\{t(\mathbf{Z}, \mathbf{r}_C) \geq k \mid \mathcal{F},\, \mathcal{Z}\} = \frac{|\{\mathbf{z} \in \mathcal{Z} : t(\mathbf{z}, \mathbf{r}_C) \geq k\}|}{|\mathcal{Z}|}, \qquad (1)$$

because $\mathbf{R} = \mathbf{r}_C$ if H_0 is true, where $\mathbf{r}_C$ is fixed by conditioning on $\mathcal{F}$, and the distribution of $\mathbf{Z}$ is uniform on $\mathcal{Z}$ in a randomized experiment. For instance, if $t(\mathbf{Z}, \mathbf{R})$ were Wilcoxon's signed rank statistic, then (1) would be its usual exact null distribution.

Similarly, Maritz (1979) proposed testing H_0 using (1) and a suitably defined M-statistic — that is, the quantity equated to zero in defining Huber's (1964) M-estimates — specifically, $t(\mathbf{Z}, \mathbf{R}) = \sum_{i=1}^{I} \psi(Y_i/s)$ where s is a quantile of the $|Y_i|$ and $\psi(\cdot)$ is a monotone increasing odd function, $\psi(d) = -\psi(-d)$, so $\psi(0) = 0$. Under H_0, the pair difference is $Y_i = (Z_{i1} - Z_{i2})(r_{Ci1} - r_{Ci2}) = \pm|r_{Ci1} - r_{Ci2}|$ so $|Y_i| = |r_{Ci1} - r_{Ci2}|$ is fixed by conditioning on $\mathcal{F}$ in (1), so s is also fixed, and $t(\mathbf{Z}, \mathbf{R}) = \sum_{i=1}^{I} \text{sign}(Y_i)\, q_i$, where $q_i = \psi(|r_{Ci1} - r_{Ci2}|/s)$ is fixed by conditioning on $\mathcal{F}$ and $\text{sign}(Y_i) = 1$, 0, or -1 as $Y_i > 0$, $Y_i = 0$, or $Y_i < 0$. As a consequence, under H_0, the distribution (1) is the distribution of the sum of I independent random variables taking the values $\pm\psi(|r_{Ci1} - r_{Ci2}|/s)$ with equal probabilities 1/2 if $|r_{Ci1} - r_{Ci2}| > 0$ or taking the value 0 with probability 1 if $|r_{Ci1} - r_{Ci2}| = 0$.

In a limited sense, Maritz (1979)'s test is an attentive inference: the scale factor, s, used in $t(\mathbf{Z}, \mathbf{R})$ is selected on the basis of the data; however, under H_0, this scale factor s is a function of $\mathcal{F}$ which is fixed in (1), so using the data to determine the scale factor s as a quantile of $|Y_i|$ does not invalidate the exact null distribution (1). This same idea can be put to work on a much larger scale.

2.3 Sensitivity analysis for nonrandom treatment assignment in observational studies

The sensitivity analysis in an observational study imagines that, in the population prior to matching, individuals are independently assigned to treatment or control with unknown probabilities, $\pi_{ij} = \Pr(Z_{ij} = 1 \mid \mathcal{F})$, that may depend upon both the observed covariates $(\mathbf{x}_{ij}, \mathbf{v}_{ij})$ and unobserved covariate u_{ij} as recorded in $\mathcal{F}$. The model says that two subjects ij and $i'j'$ with the same observed covariates, $(\mathbf{x}_{ij}, \mathbf{v}_{ij}) = (\mathbf{x}_{i'j'}, \mathbf{v}_{i'j'})$, may differ in their odds of treatment by at most a factor of $\Gamma \geq 1$, that is, $\Gamma^{-1} \leq \pi_{ij}(1 - \pi_{i'j'}) / \{\pi_{i'j'}(1 - \pi_{ij})\} \leq \Gamma$. It is easy to show that this is equivalent to assuming $\log\{\pi_{ij}/(1 - \pi_{ij})\} = \kappa(\mathbf{x}_{ij}, \mathbf{v}_{ij}) + \gamma u_{ij}$ with $\gamma = \log(\Gamma)$ and $0 \leq u_{ij} \leq 1$ for some unknown function $\kappa(\cdot, \cdot)$; see Rosenbaum (2002, §4) where the proof consists in constructing u_{ij} from π_{ij} and conversely. The distribution of $\mathbf{Z}$ is then restricted to $\mathcal{Z}$ by conditioning on the event $\mathbf{Z} \in \mathcal{Z}$. If pairs are matched for observed covariates $(\mathbf{x}_{ij}, \mathbf{v}_{ij})$ so that $\kappa(\mathbf{x}_{i1}, \mathbf{v}_{i2}) = \kappa(\mathbf{x}_{i2}, \mathbf{v}_{i2})$, then $\Pr(Z_{i1} = 1 \mid \mathcal{F}, Z_{i1} + Z_{i2} = 1) = \exp(\gamma u_{i1}) / \{\exp(\gamma u_{i1}) + \exp(\gamma u_{i2})\}$ and

$$\Pr(\mathbf{Z} = \mathbf{z} \mid \mathcal{F}, \mathcal{Z}) = \prod_{i=1}^{I} \frac{z_{i1} \exp(\gamma u_{i1}) + z_{i2} \exp(\gamma u_{i2})}{\exp(\gamma u_{i1}) + \exp(\gamma u_{i2})}$$

$$= \frac{\exp\left(\gamma \mathbf{z}^T \mathbf{u}\right)}{\sum_{\mathbf{b}\in\mathcal{Z}} \exp\left(\gamma \mathbf{b}^T \mathbf{u}\right)} \text{ for } \mathbf{z} \in \mathcal{Z} \text{ for some } \mathbf{u} = (u_{11}, \ldots, u_{I2})^T \in \mathcal{U}, \tag{2}$$

where $\mathcal{U} = [0,1]^{2I}$ is the $2I$-dimensional unit cube. When $\Gamma = 1$ so $\gamma = 0$, expression (2) equals the randomization distribution, $\Pr(\mathbf{Z} = \mathbf{z} \mid \mathcal{F}, \mathcal{Z}) = 2^{-I}$. Using (2), if γ and $\mathbf{u}$ were known, then under H_0 the distribution of the test statistic $T = t(\mathbf{Z}, \mathbf{R}) = t(\mathbf{Z}, \mathbf{r}_C)$ would be the sum of the probabilities in (2) over the $\mathbf{z}$ in $\{\mathbf{z} \in \mathcal{Z} : t(\mathbf{z}, \mathbf{r}_C) \geq k\}$. The sensitivity analysis asks: How large a departure Γ from randomization must be present to materially alter inferences based on the naive model that claims adjustments for observed covariates $(\mathbf{x}_{ij}, \mathbf{v}_{ij})$ suffice to remove all bias? Each value of $\Gamma \geq 1$ yields an interval of possible P-values or point estimates or endpoints for confidence intervals, and the question is: How large must Γ be if this interval is to be so long as to be uninformative, say permitting both acceptance and rejection of H_0?

The current paper considers analyses of subsets of the I pairs. Let $\mathfrak{s} \subseteq \{1, 2, \ldots, I\}$ be a fixed nonempty subset of $|\mathfrak{s}| \geq 1$ of the I pairs. For instance, if the pairs were exactly matched for gender, the set $\mathfrak{s}$ might consist of the pairs i consisting of two paired women. Much of our concern later on will be with sets of pairs selected on the basis of the data, but the complications introduced by a data-dependent set of pairs are deferred to §2.4 and later, and in the current paragraph $\mathfrak{s}$ is a set of pairs determined a priori, for instance, a planned subgroup analysis for pairs of women. Appending a subscript $\mathfrak{s}$ to a vector such as $\mathbf{Z}$, as in $\mathbf{Z}_\mathfrak{s}$, means the vector of dimension $2|\mathfrak{s}|$ containing those coordinates of $\mathbf{Z}$ corresponding to pairs $i \in \mathfrak{s}$. A similar notation applies to $\mathbf{R}$ as $\mathbf{R}_\mathfrak{s}$, to $\mathcal{F}$ as $\mathcal{F}_\mathfrak{s}$, and to $\mathcal{U}$ as $\mathcal{U}_\mathfrak{s}$; moreover, $H_{0\mathfrak{s}}$ is the hypothesis of no treatment effect for all pairs $i \in \mathfrak{s}$, that is, $H_{0\mathfrak{s}} : r_{Tij} = r_{Cij}$ for $i \in \mathfrak{s}$ and $j = 1, 2$. If, as in §2.2, the test statistic is of the form $T_\mathfrak{s} = t(\mathbf{Z}_\mathfrak{s}, \mathbf{R}_\mathfrak{s}) = \sum_{i\in\mathfrak{s}} \text{sign}(Y_i)\, q_{\mathfrak{s}i}$ where $q_{\mathfrak{s}i} \geq 0$ is a function of $\mathcal{F}_\mathfrak{s}$, then $T_\mathfrak{s}$ is a function of aspects of just the pairs in $\mathfrak{s}$. Define $\overline{\overline{T}}_{\Gamma\mathfrak{s}}$ to be a random variable that is the sum of s independent random variables, the ith random variable being $q_{\mathfrak{s}i}$ with probability $\Gamma/(1+\Gamma)$ and $-q_{\mathfrak{s}i}$ with probability $1/(1+\Gamma)$ providing $q_{\mathfrak{s}i} > 0$, and otherwise the ith random variable is 0 with probability 1 if $q_{\mathfrak{s}i} = 0$. Define $\overline{T}_{\Gamma\mathfrak{s}}$ analogously but with $\Gamma/(1+\Gamma)$ and $1/(1+\Gamma)$ interchanged. Then it is not difficult to show for each fixed $\Gamma = \exp(\gamma)$, as $\mathbf{u}_\mathfrak{s}$ ranges over $\mathcal{U}_\mathfrak{s}$, the unknown distribution $\Pr(T_\mathfrak{s} \geq k \mid \mathcal{F}, \mathcal{Z})$ of $T_\mathfrak{s}$ under $H_{0\mathfrak{s}}$ and (2) is sharply bounded by two known distributions,

$$\Pr\left(\overline{T}_{\Gamma\mathfrak{s}} \geq k \,\middle|\, \mathcal{F}, \mathcal{Z}\right) \leq \Pr\left(T_\mathfrak{s} \geq k \mid \mathcal{F}, \mathcal{Z}\right) \leq \Pr\left(\overline{\overline{T}}_{\Gamma\mathfrak{s}} \geq k \,\middle|\, \mathcal{F}, \mathcal{Z}\right); \tag{3}$$

see Rosenbaum (1987; 2002, §4; 2007). When $0 = \gamma = \log(\Gamma)$, there is equality in (2), and

both bounds in (2) equal the randomization distribution (1). The bounds in (3) are sharp being attained for particular $\mathbf{u}_\mathfrak{s}$ in $\mathcal{U}_\mathfrak{s}$; therefore, the bounds (3) cannot be improved except with additional information about the unobserved $\mathbf{u}_\mathfrak{s}$. The bounds (3) yield bounds on P-values, point estimates and confidence intervals.

Other methods of sensitivity analysis in observational studies are discussed by Cornfield et al. (1959), Gastwirth (1992), Hosman et al. (2010), Liu et al. (2013), Small (2007), Wang and Kreiger (2006), Yanagawa (1984), and Yu and Gastwirth (2005).

2.4 Review: use of effect modifiers when testing the hypothesis of no effect

Hsu et al. (2013, §4) tested Fisher's null hypothesis H_0 of no effect by first dividing the pairs $i \in \{1, \ldots, I\}$ into several groups based on $\mathbf{v}_{ij}$ looking for possible effect modification, that is, larger or more stable treatment effects in some groups than in others. More precisely, $G \geq 1$ mutually exclusive and exhaustive groups of pairs were formed, $\mathfrak{g}_g \subseteq \{1, \ldots, I\}$ with $\mathfrak{g}_g \cap \mathfrak{g}_{g'} = \emptyset$ for $g \neq g'$, and $\{1, \ldots, I\} = \bigcup_{g=1}^{G} \mathfrak{g}_g$. Write $\mathcal{G} = \{\mathfrak{g}_1, \ldots, \mathfrak{g}_G\}$. These groups $\mathfrak{g} \in \mathcal{G}$ were formed in an attentive fashion; that is, when H_0 is true, the groups are a function of $\mathcal{F}$, $\mathcal{Z}$ and not of $\mathbf{Z}$, so the grouping is fixed by conditioning in the null distributions (1), (2) and (3). Specifically, as in Hsu et al. (2013, §4), a function of $|Y_i|$ is regressed on $\mathbf{v}_{i1} = \mathbf{v}_{i2} = \mathbf{v}_i$, say, in some fashion that yields nonoverlapping groups. Under Fisher's H_0, the absolute difference in responses $|Y_i| = |r_{Ci1} - r_{Ci2}|$ is fixed by conditioning on $\mathcal{F}$, as discussed in §2.2, so the grouping produced by the regression of $|Y_i|$ on $\mathbf{v}_i$ is also fixed.

There is some reason to hope that a grouping based on the regression of $|Y_i|$ on $\mathbf{v}_i$ will construct useful groups. If H_0 were false with $Y_i = \rho(\mathbf{v}_i) + \xi_i$ where $\rho(\cdot) \geq 0$ and ξ_i independent and identically distributed with continuous unimodal distribution symmetric about zero, then $|Y_i|$ is stochastically larger than $|Y_{i'}|$ if $\rho(\mathbf{v}_i) > \rho(\mathbf{v}_{i'})$; see Jogdeo (1977, Theorem 2.2). Therefore, the regression of $|Y_i|$ on $\mathbf{v}_i$ may form groups with different typical effects under this simple model.

In the example in §1.2, there are $|\mathcal{V}| = 360$ possible values of $\mathbf{v}_{ij}$. In §4, the rank of $|Y_i|$ is regressed on $\mathbf{v}_{i1} = \mathbf{v}_{i2}$ using the CART regression tree method of Breiman et al. (1983), resulting in three leaves or groups, namely $\mathfrak{g}_1$ consisting of individuals with work income prior to the earthquake, $\mathfrak{g}_2$ consisting of men without work income prior to the earthquake, and $\mathfrak{g}_3$ consisting of women without work income prior to the earthquake. It is not practical to study effect modification with 2106 pairs in 360 groups of pairs, but it

is practical to study effect modification with 2106 pairs in 3 groups of pairs.

Groups $\mathcal{G} = \{\mathfrak{g}_1, \ldots, \mathfrak{g}_G\}$ built in this way are functions of $\mathcal{F}$, $\mathcal{Z}$ when H_0 is true, so the groups are fixed conditionally given $\mathcal{F}$, $\mathcal{Z}$ when H_0 is true. Under model (2) when H_0 is true, a test statistic $T_{\mathfrak{g}} = t(\mathbf{Z}_{\mathfrak{g}}, \mathbf{R}_{\mathfrak{g}}) = \sum_{i\in\mathfrak{g}} \text{sign}(Y_i)\, q_{\mathfrak{g}i}$ for $\mathfrak{g} \in \mathcal{G}$ has the usual bounds on its null distribution, namely (3), because these bounds refer to the conditional distribution given $\mathcal{F}$, $\mathcal{Z}$ when H_0 is true. In particular, in a randomized experiment, model (2) holds with $0 = \gamma = \log(\Gamma)$, so that under H_0, the group-specific statistic $T_{\mathfrak{g}}$ has its usual randomization distribution despite the data-dependent nature of the groups $\mathcal{G} = \{\mathfrak{g}_1, \ldots, \mathfrak{g}_G\}$. Typically, $\mathcal{G}$ is not fixed by conditioning on $\mathcal{F}$, $\mathcal{Z}$ when H_0 is false, and this is the central issue addressed in §3, where H_0 is not assumed to be true.

Hsu et al. (2013, §4) test H_0 by computing a P-value of the form (1) or P-value bound of the form (3) using the pairs in each group $\mathfrak{g}_g$ separately, yielding G independent P-values, and combine them using a generalization of Fisher's method for combining independent P-values, namely the truncated product of P-values of Zaykin et al. (2002). The truncated product uses as its test statistic the product of those P-values that are no larger than a prespecified cutoff $\widetilde{\alpha}$ with $0 < \widetilde{\alpha} \leq 1$, and for $\widetilde{\alpha} = 1$ it is equivalent to Fisher's procedure; see Benjamini and Heller (2008) for simultaneous inference using Fisher's procedure. Hsu et al. show that in the presence of even a small amount of effect modification, this procedure has higher power in a sensitivity analysis and larger design sensitivity than a test that ignores the groups.

So far, the discussion has focused on testing the null hypothesis of no effect H_0 at all, and that hypothesis played a key role in permitting the groups $\mathcal{G} = \{\mathfrak{g}_1, \ldots, \mathfrak{g}_G\}$ to be determined from the data by regressing a function of $|Y_i|$ on $\mathbf{v}_i$. A more interesting question not addressed by Hsu et al. (2013) is whether $H_{0\mathfrak{g}}$ may be tested using (3) when H_0 may be false. If H_0 is false, then there is at least one pair i for which $r_{Ti1} \neq r_{Ci1}$ or $r_{Ti2} \neq r_{Ci2}$ or both, and in this case $\mathcal{G} = \{\mathfrak{g}_1, \ldots, \mathfrak{g}_G\}$ is not a function of $\mathcal{F}$, $\mathcal{Z}$ because $\mathbf{R} \neq \mathbf{r}_C$ in the sense that $\mathbf{r}_C$ is determined by $\mathcal{F}$ but $\mathbf{R}$ varies with $\mathbf{Z}$. In the example, $\mathcal{G} = \{\mathfrak{g}_1, \mathfrak{g}_2, \mathfrak{g}_3\}$ suggested a focus on individuals with work income before the earthquake, men without work income and women without work income. If we reject the null hypothesis H_0 of no effect on anyone, it is not clear from the argument of this section that we can say anything about just one of the groups, say about $H_{0\mathfrak{g}_2}$ for men without work income. Indeed, to reject H_0 is to reject the hypothesis that $\mathcal{G} = \{\mathfrak{g}_1, \ldots, \mathfrak{g}_G\}$ is a function of $\mathcal{F}$, $\mathcal{Z}$. In other words, if H_0 is false in a randomized experiment, then the grouping $\mathcal{G}$ depends on $\mathbf{Z}$: had randomization yielded a different treatment assignment $\mathbf{Z}$, it might also have yielded

different groups $\mathcal{G}$, and the hypothesis $H_{0\mathfrak{g}_2}$ is not even a hypothesis in any conventional sense, because the hypotheses change as the treatment assignments $\mathbf{Z}$ change. This issue is explored in §3.

In large samples, the power of a sensitivity analysis is determined by the design sensitivity (Rosenbaum 2004), and a formula for the design sensitivity of Maritz's (1979) M-test is given in Corollary 1 in Rosenbaum (2013). Other things being equal, the design sensitivity is larger — so the sensitivity analysis has greater power in large samples — when the effect is larger, say the typical Y_i is larger, or when the dispersion of the Y_i is smaller for a given typical size; see Rosenbaum (2004, 2005, 2013). Combining separate P-values within groups $\mathcal{G} = \{\mathfrak{g}_1, \ldots, \mathfrak{g}_G\}$ can increase the power of a sensitivity analysis when either the size or dispersion of the Y_i vary from group to group; see Hsu et al. (2013, §3.3).

3 Strong control of the family-wise error rate with groups constructed from the data

3.1 Data-dependent groups of pairs and null distributions within those groups

To address the issue raised at the end of §2.4, the following conditions are assumed to hold.

Condition 1 *The distribution of* $\mathbf{Z}$ *given* $\mathcal{F}$, $\mathcal{Z}$ *is (2) for a specific* $\gamma = \log(\Gamma) \geq 0$ *and unknown* $\mathbf{u} \in \mathcal{U}$.

Condition 2 *Mutually exclusive and exhaustive groups* $\mathcal{G} = \{\mathfrak{g}_1, \ldots, \mathfrak{g}_G\}$ *of pairs are formed as a function of* $\mathbf{R}$, $\mathbf{v}_i$ *and* $\mathbf{x}_i$, $i = 1, \ldots, I$ *from pairs exactly matched for* $\mathbf{v}_{i1} = \mathbf{v}_{i2} = \mathbf{v}_i$ *and* $\mathbf{x}_{i1} = \mathbf{x}_{i2} = \mathbf{x}_i$.

Here, G and $\mathcal{G}$ are random quantities given $\mathcal{F}$, $\mathcal{Z}$ because H_0 may be false and, if so, then $\mathbf{R} \neq \mathbf{r}_C$ depends upon $\mathbf{Z}$, so the groups in Condition 2 may also depend on $\mathbf{Z}$. For instance, in §2.4, had the earthquake struck different people in Chile, then there might have been $G = 6$ groups, say, rather than the $G = 3$ groups in §2.4, and these groups might have involved different variables in $\mathbf{v}_{ij}$. If the groups $\mathcal{G}$ are random quantities depending upon $\mathbf{Z}$, then taking the groups to be fixed, conditioning on $\mathcal{G}$, may alter the distribution of $\mathbf{Z}$. In particular, condition 2 is satisfied when, as in §2.4 and §5, groups are produced as the leaves of a CART regression tree formed from pairs exactly matched for observed covariates by regressing the rank of $|Y_i|$ on $\mathbf{v}_{i1} = \mathbf{v}_{i2}$.

Let $\mathfrak{h}$ be the union of all of the groups, $\mathfrak{g}_g$, for which there is no treatment effect, that is, the union of those $\mathfrak{g}_g$ such that $r_{Ti1} = r_{Ci1}$ and $r_{Ti2} = r_{Ci2}$ for all $i \in \mathfrak{g}_g$; possibly, $\mathfrak{h} = \emptyset$. Obviously, the investigator does not know $\mathfrak{h}$.

Remark 3 *There may be pairs $i \notin \mathfrak{h}$ with $r_{Ti1} = r_{Ci1}$ and $r_{Ti2} = r_{Ci2}$; however, these pairs are in groups $\mathfrak{g}_g$ that contain at least one individual i' such that either $r_{Ti'1} \neq r_{Ci'1}$ or $r_{Ti'2} \neq r_{Ci'2}$. In other words, $\mathfrak{h}$ is the union of all groups with no effect, not the set of all pairs with no effect. For instance, in §2.4, $\mathfrak{h}$ would exclude all women with no work income prior to the earthquake if at least one such woman ij was affected in the sense that $r_{Tij} \neq r_{Cij}$.*

Remark 4 *As $\mathcal{G}$ is a random quantity, $\mathfrak{h}$ is also a random quantity because $\mathfrak{h}$ is a union of some of the $\mathfrak{g}_g$. Indeed, the set $\mathfrak{h}$ is a function of $\mathcal{F}$, $\mathcal{Z}$, $\mathcal{G}$. Conditionally given $\mathcal{F}$, $\mathcal{Z}$, $\mathcal{G}$, the set $\mathfrak{h}$ is fixed. Conditionally given $\mathcal{F}$, $\mathcal{Z}$, $\mathcal{G}$, if $\mathfrak{h}=\emptyset$ then there are affected pairs in every group $\mathfrak{g} \in \mathcal{G}$ so every $H_{0\mathfrak{g}}$ is false, and false rejection of a true $H_{0\mathfrak{g}}$ cannot occur. Conversely, conditionally given $\mathcal{F}$, $\mathcal{Z}$, $\mathcal{G}$, if $\mathfrak{h}\neq\emptyset$ then some group or groups $\mathfrak{g} \in \mathcal{G}$ contain no affected individuals and false rejection of a true $H_{0\mathfrak{g}}$ is possible. Proposition 5 and its Corollary speak about the distribution of the test statistic $T_\mathfrak{h} = t(\mathbf{Z}_\mathfrak{h}, \mathbf{R}_\mathfrak{h})$ given $\mathcal{F}$, $\mathcal{Z}$, $\mathcal{G}$ where the pairs $i \in \mathfrak{h}$ are all unaffected by the treatment, but the grouping $\mathcal{G}$ itself (and hence also $\mathfrak{h}$) may have been affected by the treatment. Stated informally, Proposition 5 says that the data-dependent grouping $\mathcal{G}$ did not alter the null distribution of $T_\mathfrak{h}$ even when H_0 is false so the argument of §2.4 is inapplicable. To emphasize, $T_\mathfrak{h}$ is computed from the union $\mathfrak{h}$ of all groups $\mathfrak{g}_g$ where there is no treatment effect, and because the investigator does not know $\mathfrak{h}$ she cannot know when she has computed $T_\mathfrak{h}$. Proposition 5 is a step in developing a multiple inference procedure that strongly controls false rejections, as discussed in §3.2.*

Proposition 5 *Assume Conditions 1 and 2. The conditional distribution $\Pr(T_\mathfrak{h} \geq k \mid \mathcal{F}, \mathcal{Z}, \mathcal{G})$ of $T_\mathfrak{h} = t(\mathbf{Z}_\mathfrak{h}, \mathbf{R}_\mathfrak{h})$ given $\mathcal{F}$, $\mathcal{Z}$, $\mathcal{G}$ is sharply bounded by the bounds in (3) with $\mathfrak{s} = \mathfrak{h}$, providing $\mathfrak{h}\neq\emptyset$.*

Proof. Assume $\mathfrak{h}\neq\emptyset$, for otherwise there is nothing to prove. Let $\mathcal{N} \subseteq \{1, \ldots, I\}$ be the set of pairs with no treatment effect, so $r_{Ti1} = r_{Ci1}$ and $r_{Ti2} = r_{Ci2}$ if and only if $i \in \mathcal{N}$, and let $\mathcal{E}$ be the complementary set of affected pairs, $\mathcal{E} = \{1, \ldots, I\} - \mathcal{N}$. Of course, $\mathcal{N} \supseteq \mathfrak{h}\neq\emptyset$, so $\mathcal{N} \neq \emptyset$. Let $\mathbf{z}$ be a possible value of $\mathbf{Z}_\mathfrak{h}$, so $\mathbf{z}$ is a $2|\mathfrak{h}|$-dimensional

vector $\mathbf{z} = \left(z_{11}, z_{12}, \ldots, z_{\ell j}, \ldots, z_{|\mathfrak{h}|,1}, z_{|\mathfrak{h}|,2}\right)^T$ with $z_{\ell j} = 1$ or $z_{\ell j} = 0$ and $z_{\ell 1} + z_{\ell 2} = 1$ for each ℓ. Write $\mathcal{D}$ for the combination of the data $\{(r_{Ci1}, r_{Ci2}, \mathbf{v}_i, \mathbf{x}_i), i \in \mathcal{N}\}$ and the data $(R_{i1}, R_{i2}, Z_{i1}, Z_{i2}, \mathbf{v}_i, \mathbf{x}_i)$, $i \in \mathcal{E}$. Because pairs $i \in \mathcal{N}$ are unaffected with $R_{ij} = r_{Cij}$ for $i \in \mathcal{N}$ and $j = 1, 2$, the grouping $\mathcal{G} = \{\mathfrak{g}_1, \ldots, \mathfrak{g}_G\}$, is a function of $\mathcal{D}$. Because the grouping $\mathcal{G}$ is a function of $\mathcal{D}$, conditioning on $\mathcal{G}$, $\mathcal{D}$, $\mathcal{F}$, $\mathcal{Z}$ is the same as conditioning on $\mathcal{D}$, $\mathcal{F}$, $\mathcal{Z}$. For $i \in \mathcal{E}$, the information in $(R_{i1}, R_{i2}, Z_{i1}, Z_{i2}, \mathbf{v}_i, \mathbf{x}_i)$ that is not in $\mathcal{F}$, $\mathcal{Z}$ is precisely $Z_{i1} = 1 - Z_{i2}$ for $i \in \mathcal{E}$; that is, one could construct $(R_{i1}, R_{i2}, Z_{i1}, Z_{i2}, \mathbf{v}_i, \mathbf{x}_i)$ from $\mathcal{F}$, $\mathcal{Z}$ if one were told Z_{i1}. Putting this all together under (2), the $Z_{i1} = 1 - Z_{i2}$ for $i \in \mathcal{N}$ satisfy

$$\begin{aligned}\Pr\left(Z_{i1} = 1 \,|\, \mathcal{F}, \mathcal{Z}, \mathcal{G}, \mathcal{D}\right) &= \Pr\left(Z_{i1} = 1 \,|\, \mathcal{F}, \mathcal{Z}, \mathcal{D}\right) \\ &= \Pr\left(Z_{i1} = 1 \,|\, \mathcal{F}, \mathcal{Z}\right) = \frac{\exp\left(\gamma u_{i1}\right)}{\exp\left(\gamma u_{i1}\right) + \exp\left(\gamma u_{i2}\right)}\end{aligned}$$

because (i) $\mathcal{G}$ is a function of $\mathcal{D}$, and (ii) the Z_{i1} for $i \in \mathcal{N}$ are conditionally independent of the $Z_{i'1}$ for $i' \in \mathcal{E}$, and apart from $Z_{i'j}$ for $i' \in \mathcal{E}$, the rest of $(R_{i'1}, R_{i'2}, Z_{i'1}, Z_{i'2}, \mathbf{v}_{i'}, \mathbf{x}_{i'})$, $i' \in \mathcal{E}$ is already fixed by conditioning on $\mathcal{F}$, $\mathcal{Z}$. Using (2) again and adding the fact that $\mathfrak{h}$ is fixed by conditioning on $\mathcal{G}$, $\mathcal{F}$, $\mathcal{Z}$ yields

$$\Pr\left(\mathbf{Z}_\mathfrak{h} = \mathbf{z} \,|\, \mathcal{F}, \mathcal{Z}, \mathcal{G}, \mathcal{D}\right) = \prod_{\ell \in \mathfrak{h}} \frac{z_{\ell 1} \exp\left(\gamma u_{\ell 1}\right) + z_{\ell 2} \exp\left(\gamma u_{\ell 2}\right)}{\exp\left(\gamma u_{\ell 1}\right) + \exp\left(\gamma u_{\ell i 2}\right)}. \tag{4}$$

Now, the right side of (4) depends on $\mathcal{G}$, $\mathcal{F}$, $\mathcal{Z}$ because $\mathfrak{h}$ depends upon $\mathcal{G}$, $\mathcal{F}$, $\mathcal{Z}$, but it does not depend on $\mathcal{D}$ given $\mathcal{G}$, $\mathcal{F}$, $\mathcal{Z}$; therefore, (4) equals $\Pr\left(\mathbf{Z}_\mathfrak{h} = \mathbf{z} \,|\, \mathcal{F}, \mathcal{Z}, \mathcal{G}\right)$. It follows that the distribution of $\mathbf{Z}_\mathfrak{h}$, namely $\Pr\left(\mathbf{Z}_\mathfrak{h} = \mathbf{z} \,|\, \mathcal{G}, \mathcal{F}, \mathcal{Z}\right)$, and hence also the distribution $\Pr\left(T_\mathfrak{h} \geq k \,|\, \mathcal{F}, \mathcal{Z}, \mathcal{G}\right)$ is identical to the distribution that produced the bounds in (3), proving the result. ■

Corollary 6 *Assume Condition 2. In a randomized experiment, the conditional distribution of $T_\mathfrak{h} = t\left(\mathbf{Z}_\mathfrak{h}, \mathbf{R}_\mathfrak{h}\right)$ given $\mathcal{F}$, $\mathcal{Z}$, $\mathcal{G}$ is its randomization distribution (i.e., (3) with $\gamma = 0$), providing $\mathfrak{h} \neq \emptyset$.*

3.2 Sensitivity bounds for closed testing with groups built from the data

Let $\mathcal{K} \subseteq \{1, \ldots, G\}$ be a nonempty subset of the groups and let $\mathfrak{k}\left(\mathcal{K}\right) = \cup_{g \in \mathcal{K}} \mathfrak{g}_g \subseteq \{1, \ldots, I\}$ be the indices of the pairs i in the groups $\mathfrak{g}_g$ for $g \in \mathcal{K}$. If the groups $\mathcal{G} = \{\mathfrak{g}_1, \ldots, \mathfrak{g}_G\}$ were fixed a priori, then the hypothesis $H_\mathcal{K}$ could be defined to say that there is no treatment

effect in the pairs $i \in \mathfrak{k}(\mathcal{K})$; that is, $H_{\mathcal{K}}$ asserts that $r_{Tij} = r_{Cij}$ for $j = 1, 2$ for all $i \in \mathfrak{g}_g$ for all $g \in \mathcal{K}$. A test of the a priori hypothesis $H_{\mathcal{K}}$ with a priori groups $\mathcal{G} = \{\mathfrak{g}_1, \ldots, \mathfrak{g}_G\}$ could be based on $T_{\mathfrak{k}(\mathcal{K})}$ in §2.3, and in particular, for each fixed $\Gamma = \exp(\gamma) \geq 1$, a level α test could be constructed using the upper bound in (3), and this would be a conventional randomization test if $\Gamma = 1$. With a priori groups $\mathcal{G} = \{\mathfrak{g}_1, \ldots, \mathfrak{g}_G\}$, the closed testing procedure of Marcus et al. (1976) would reject $H_{\mathcal{K}}$ at level α if and only if it had rejected at level α all $H_{\mathcal{L}}$ with $\mathcal{K} \subseteq \mathcal{L} \subseteq \{1, \ldots, G\}$, and it would strongly control the family-wise error rate, that is, it would falsely reject at least one true $H_{\mathcal{K}}$ with probability at most α no matter which hypotheses $H_{\mathcal{M}}$ are true for $\mathcal{M} \subseteq \{1, \ldots, G\}$. See Hochberg and Tamhane (1987, Chapter 1) for discussion of the family-wise error rate, and see Rosenbaum and Silber (2009a) for discussion in the context of a sensitivity analysis. (Weak control of the family-wise error rate is no longer regarded as adequate, so we do not discuss weak control; it says that the chance of falsely rejecting $H_{\mathcal{K}}$ is at most α if H_0 is true, but if H_0 is false then there are no promises about false rejection of $H_{\mathcal{K}}$.) Does a similar result hold when the groups $\mathcal{G} = \{\mathfrak{g}_1, \ldots, \mathfrak{g}_G\}$ are built using the data subject to Condition 2?

Proposition 8 says that we may apply closed testing using groups constructed attentively from the data at hand, yet strongly control the family-wise error rate in a sensitivity analysis. Setting $\Gamma = 1$ yields the Corollary to Proposition 8

Algorithm 7 *Construct groups $\mathcal{G} = \{\mathfrak{g}_1, \ldots, \mathfrak{g}_G\}$ by a method that satisfies Condition 2. Fix $\Gamma \geq 1$, and for $\mathcal{K} \subseteq \{1, \ldots, G\}$ determine the value $k_{\Gamma,\mathcal{K}}$ from the upper bound in (3) with $\mathfrak{s} = \mathfrak{k}(\mathcal{K})$ as the smallest value such that $\Pr\left(\overline{\overline{T}}_{\Gamma,\mathfrak{k}(\mathcal{K})} \geq k_{\Gamma,\mathcal{K}} \,\middle|\, \mathcal{F}, \mathcal{Z}\right) \leq \alpha$ for a fixed α with $0 < \alpha < 1$. Reject the null hypothesis that all pairs $i \in \mathfrak{k}(\mathcal{K})$ experience no treatment effect if $\overline{\overline{T}}_{\Gamma,\mathfrak{k}(\mathcal{L})} \geq k_{\Gamma,\mathcal{L}}$ for all $\mathcal{L}$ such that $\mathcal{K} \subseteq \mathcal{L} \subseteq \{1, \ldots, G\}$, and assert that this rejection is insensitive to unmeasured biases no larger than Γ.*

Proposition 8 *Assume Condition 1 holds with the specified Γ. The conditional probability given $\mathcal{F}$, $\mathcal{Z}$, $\mathcal{G}$ that Algorithm 7 makes at least one false rejection is at most α.*

Proof. Under the stated conditions, Proposition 5 says the individual tests have conditional level α. The main result in Marcus et al. (1976) says that a closed testing procedure as in Algorithm 7 ensures that the probability of at least one false rejection is at most α providing the component tests have level α. ■

Corollary 9 *In a randomized paired experiment, the conditional probability given $\mathcal{F}$, $\mathcal{Z}$, $\mathcal{G}$ that Algorithm 7 makes at least one false rejection is at most α.*

3.3 Summary: closed testing with groups discovered by exploratory analysis

To summarize, Proposition 8 and Corollary 6 would be relatively straightforward applications of closed testing if the groups $\mathcal{G} = \{\mathfrak{g}_1, \ldots, \mathfrak{g}_G\}$ had been specified a priori; see Hsu et al. (2013, §3.4). However, in §1.2 and §5, a collection of 360 types of individuals were collapsed to $G = 3$ groups using the data, specifically by applying CART to a regression of the rank of $|Y_i|$ on $v_{i1} = v_{i2}$, so the groups were not given a priori. Conditioning on $\mathcal{G} = \{\mathfrak{g}_1, \ldots, \mathfrak{g}_G\}$ to fix the groups, and hence also to fix the null hypotheses, distorts the distributions of some of the Z_{ij} when some $H_{0\mathfrak{g}}$ are true and others are false. Proposition 5 says that, under Condition 2, the distortion of the distribution of Z_{ij} is confined to groups $\mathfrak{g}$ such that $H_{0\mathfrak{g}}$ is false, and as a consequence Proposition 8 and Corollary 6 say that closed testing strongly controls the family-wise error rate among groups selected on the basis of the data.

4 Near exact matching with strength k balance

4.1 Offering CART more variables than can be used in the hope that CART will refuse some of them

In studying effect modification, it is convenient to have treatment-control pairs with the same values of the covariates under study as potential effect modifiers. If a covariate is not exactly matched, if men are sometimes matched to women, then the treated-minus-control pair difference in outcomes Y_i may be associated with gender because of the mismatch on gender rather than because the treatment effect is different for men than for women. Expressed in familiar if imprecise terms, gender may have a main effect and an interaction with the treatment, and when pairs are exactly matched for gender the main effect is removed so the interaction can be seen clearly. In §1.2, it was not possible to match exactly for all $V = 6$ candidate effect modifiers $\mathbf{v}_{ij}$, yet we did not want to lose any pairs because of this.

It is easy to match to balance many covariates, perhaps by matching on the propensity score (Rosenbaum and Rubin 1985), or perhaps using fine balance (Zubizarreta et al. 2011), but even with nominal covariates, it is not possible to match everyone exactly for more than a few covariates because the number of combinations grows exponentially with the number of covariates. So far, we have been ignoring the issue of matching exactly for $\mathbf{v}_{ij}$, but it is a common problem when more than a few covariates are candidates as effect

modifiers. Before discussing the problem abstractly, consider the problem as it occurs in the example in §1.2.

In the earthquake data in §1.2, there were $V = 6$ candidate covariates $\mathbf{v}_{ij}$ that were plausible effect modifiers, yielding $|\mathcal{V}| = 2 \times 2 \times 3 \times 3 \times 5 \times 2 = 360$ types of individuals, yet there were only $I = 2106$ pairs of an exposed and an unexposed individual. In fact, only $1978/2106 = 94\%$ of the pairs are exactly matched for all six basic covariates. However, by design, the six covariates exhibit a new and very strong form of covariate balance, specifically strength 3 balance. There are $\binom{6}{3} = 20$ ways to pick 3 of the six covariates. For each of these 20 choices of three covariates there is a nominal variable formed as all combinations of levels of these three covariates; for instance, gender, marital status and self-rated health combine to yield a nominal variable with $2 \times 2 \times 3 = 12$ levels. In a strength 3 match, each of the $\binom{V}{3} = \binom{6}{3} = 20$ combinations of 3 of the V covariates is exactly balanced: the marginal distribution is the same in treated and control matched pairs. Table 1 illustrates this in the case of gender, marital status and self-rated health, but the same balance occurs for all 20 groups of three of the $V = 6$ basic covariates. One of these 20 choices of 3 covariates had $3 \times 3 \times 5 = 45$ categories, where all six covariates had $2 \times 2 \times 3 \times 3 \times 5 \times 2 = 360$ categories.

More generally, in a strength k match, each of the $\binom{V}{k}$ nominal variables built from k of the V basic variables is perfectly balanced. The term "strength k" match is intended to suggest a (limited) analogy with the orthogonal arrays used to construct fractional factorial designs (Hedayat, Sloane and Stufken 1999).

How can the marginal distributions be identical with pairs that are not perfectly matched? A mismatch in one pair counterbalances a mismatch in another. Subject to the requirement (or constraint) of strength k balance, plus balance requirement on $\mathbf{x}_{ij}$, the matching algorithm maximized the number of pairs that were exactly matched for $\mathbf{v}$, so $\mathbf{v}_{i1} = \mathbf{v}_{i2}$ as often as possible. Specifically, $1978/2106 = 94\%$ of the pairs are exactly matched for all $V = 6$ basic covariates $\mathbf{v}_{ij}$, although 128 pairs could be balanced but not exactly matched. The CART tree and its associated groups were built using the 1978 exact pairs, briefly ignoring the 128 inexact pairs. Because the six covariates are exactly balanced, whenever an inexact match does occur in the 128 inexact pairs, the mismatch is counterbalanced in another inexact pair. For example, there were $15/2106 = 0.007$ pairs in which a treated individual with a health problem was paired to a control without a health problem, but this was counterbalanced by $15/2106 = 0.007$ pairs in which a control with a health problem was paired to a treated individual without one. For detailed display of

how imperfect matches can yield perfect balance; see Zubizarreta et al. (2011, Table 4).

With a strength k match, we build the tree and its associated groups using just the exactly matched pairs. If the resulting tree involves k or fewer variables, then the inexactly matched pairs can be rearranged to be exactly matched for all of the k or fewer variables used in the tree. In the example, the tree selected two variables, gender and income quartile, and the match was strength 3, so it was possible to break the original pairing of the 128 inexactly matched pairs, and pair these 2×128 individuals again to be exactly matched for the groups defined by Figure 2, with the consequence that all of the individuals in the original 2106 pairs were retained.

In a strength k match, we can build the tree using more than k variables, let CART decide which k or fewer variables are the best candidates as effect modifiers, and know that all individuals can be matched exactly for the variables CART selects.

4.2 Details of implementing strength k matching

Implementing strength k matching is straightforward using Zubizarreta (2012)'s `mipmatch` package in `R`. First, the $\binom{V}{k}$ nominal variables formed from k of the V basic variables $\mathbf{v}_{ij}$ are determined, and the match is constrained to perfectly balance all of these. Second, additional balance constraints are imposed on the remaining observed covariates $\mathbf{x}_{ij}$. At this stage, the problem becomes an integer program, a constrained combinatorial optimization problem that `mipmatch` solves. The match maximizes the number of pairs subject to covariate balance constraints. Optionally, one may also use other standard matching techniques also available in `mipmatch`. In the current paper, as in Zubizarreta, Paredes and Rosenbaum (2013), we view matching and pairing as separate tasks: matching selects treated and control groups that exhibit covariate balance, whereas in the current paper the pairing in §4.3 focuses attention on the V nominal candidate effect modifiers.

The remainder of §4.2 describes the construction of the matched earthquake data; however, this material is not used later in the paper and may be skipped. The match was the largest possible match that exhibited certain stipulated and desired properties of covariate balance, a process called "cardinality matching" (Zubizarreta, Paredes and Rosenbaum 2013). We matched:

(i) with exact pair matching for age groups, using 6 age groups, namely [15, 25), [25, 35), [35, 45), [45, 55), [55, 65), and [65,);

(ii) to exactly balance the 20 possible 3-way interactions of sex, married or cohabitant,

number of persons in the household (1, 2, 3, 4 or 5, 6 or more), health problem, health perception (poor, fair, good), and quartile of work income (0, (0, 150000], (150000, ∞));

(iii) to force very similar means or proportions for age, marital status (divorced or widow, single), health problem, hospitalized, psychological problem, disability (self sufficient or low, moderate or severe, no, unknown), health insurance (public, private, other, no, unknown), years of education, employment status (employed, unemployed, inactive), work income, total income, poverty status, housing status (own housing or paying to own it, rented housing, ceded housing, irregular use of housing), housing structure (acceptable, reparable, irreparable), overcrowding (no, medium, critical), and an estimated propensity score. Here, we constrained the differences in means or proportions to be at most 0.1 times their standard deviations before matching; see Rosenbaum and Rubin (1985). This produced 2106 matched pairs meeting the covariate balance properties.

4.3 Form the groups $\mathcal{G} = \{\mathfrak{g}_1, \ldots, \mathfrak{g}_G\}$ using the exactly matched pairs

After matching with strength k, the pairs inexactly matched for the basic variables $\mathbf{v}_{ij}$ are set aside, where pairs $\mathcal{I} \subseteq \{1, \ldots, I\}$ are exactly matched. The remaining pairs in $\{1, \ldots, I\} - \mathcal{I}$ continue to exhibit strength $k = 3$ balance for the covariates in $\mathbf{v}_{ij}$, but individual pairs differ, $\mathbf{v}_{i1} \neq \mathbf{v}_{i2}$ for $i \notin \mathcal{I}$; however, we minimized the number of differences. Building the groups $\mathcal{G} = \{\mathfrak{g}_1, \ldots, \mathfrak{g}_G\}$ uses only the exactly matched pairs in $\mathcal{I}$ with the remaining pairs in $\{1, \ldots, I\} - \mathcal{I}$ briefly set aside.

In the earthquake example in §1.2, 128 of $I = 2106$ pairs are set aside, leaving $|\mathcal{I}| = 1978$ exactly matched pairs. These pairs exactly matched for $\mathbf{v}_{ij}$ are used to determine the values of $\mathbf{v}_{ij}$ that define the boundaries of the groups of pairs. In the earthquake data, the rank of the absolute pair difference $|Y_i|$ was regressed on $\mathbf{v}_{i1} = \mathbf{v}_{i2} = \mathbf{v}_i$ for the exactly matched pairs $i \in \mathcal{I}$, yielding three groups, namely $\mathfrak{g}_1$ consisting of individuals with work income prior to the earthquake, $\mathfrak{g}_2$ consisting of men without work income prior to the earthquake, and $\mathfrak{g}_3$ consisting of women without work income prior to the earthquake.

Replacing $\{1, \ldots, I\}$ by $\mathcal{I}$ in §3, Propositions 5 and 8 apply immediately to the pairs in $\mathcal{I}$. Under model (2), pairs $i \in \mathcal{I}$ are conditionally independent of pairs $i \notin \mathcal{I}$ given $\mathcal{F}$, $\mathcal{Z}$, so groups $\mathcal{G} = \{\mathfrak{g}_1, \ldots, \mathfrak{g}_G\}$ formed using the pairs in $\mathcal{I}$ are conditionally independent given $\mathcal{F}$, $\mathcal{Z}$ of treatment assignments Z_{ij} for $i \notin \mathcal{I}$. Therefore, conditioning on $\mathcal{G} = \{\mathfrak{g}_1, \ldots, \mathfrak{g}_G\}$ does not change the conditional distribution given $\mathcal{F}$, $\mathcal{Z}$ of treatment assignments Z_{ij} for $i \notin \mathcal{I}$ for the unused inexact pairs.

If the groups $\mathcal{G} = \{\mathfrak{g}_1, \ldots, \mathfrak{g}_G\}$ formed using the pairs in $\mathcal{I}$ involve k or fewer of the V basic covariates, then the pairs $i \notin \mathcal{I}$ are perfectly balanced for the k or fewer covariates that define the groups $\mathcal{G}$. The proof of this is immediate: the pairs $i \in \mathcal{I}$ exactly matched for $\mathbf{v}_{ij}$ are certainly balanced for these k covariates, and all of the pairs $\{1, \ldots, I\}$ are balanced for these k covariates because the match is strength k, so the pairs $i \notin \mathcal{I}$ must also be balanced for these k covariates. As a consequence, if the groups $\mathcal{G} = \{\mathfrak{g}_1, \ldots, \mathfrak{g}_G\}$ are determined by k or fewer covariates, it is possible to break the inexact pairing for $i \notin \mathcal{I}$ and pair these pairs again so that they are exactly paired for the k covariates that define the groups, $\mathcal{G}$. In the earthquake example, the $128 = 2106 - 1978 = I - |\mathcal{I}|$ inexact pairs not used in forming the groups are balanced for the $2 \leq k = 3$ covariates that define the three groups, namely gender and quantile of work income before the earthquake. Moreover, because the new pairing uses the same $2 \times 128 = 2 \times (I - |\mathcal{I}|)$ individuals as before, the new pairing has the same balance properties: all $\binom{V}{k} = \binom{6}{3} = 20$ composites of $k = 3$ of the $V = 6$ basic covariates are exactly balanced in the new pairing of the $128 = I - |\mathcal{I}|$ inexactly matched individuals. Therefore, the new match formed by combining the new pairing of $I - |\mathcal{I}| = 128$ pairs with the original exact pairing of $|\mathcal{I}| = 1978$ pairs is also a match of strength $k = 3$. In short, the new pairs are all exactly matched for the $2 \leq 3 = k$ variables that define the groups, yet there is still excellent balance on the remaining coordinates of $\mathbf{v}$. Moreover, the treatment assignments Z_{ij} in these new pairs $i \notin \mathcal{I}$ are conditionally independent of the groups $\mathcal{G} = \{\mathfrak{g}_1, \ldots, \mathfrak{g}_G\}$ given $\mathcal{F}$, $\mathcal{Z}$. If the π_{ij} depend upon $\mathbf{v}_{ij}$ only through the k coordinates of $\mathbf{v}_{ij}$ that define the groups, then Propositions 5 and 8 apply to the I pairs formed by merging the old exact pairing of $1978 = |\mathcal{I}|$ exact pairs combined with the new pairing of $128 = 2106 - 1978 = I - |\mathcal{I}|$ pairs. If the π_{ij} depend upon all the coordinates of $\mathbf{v}_{ij}$, then the bias introduced by the inexact pairs is unlikely to be large, as all $\binom{V}{k}$ composites of k covariates in $\mathbf{v}_{ij}$ are exactly balanced, a degree of balance on observed covariates that is much better than expected by chance under completely randomized treatment assignment.

The re-pairing of the $2 \times 128 = 2 \times (I - |\mathcal{I}|)$ individuals inexactly matched for $\mathbf{v}_{ij}$ used Hansen's (2007) `pairmatch` function in his `optmatch` package in `R`, with a distance that severely penalized mismatches for the three groups defined by Figure 2, but otherwise simply counted the number of mismatches, 0 to 6, on coordinates of $\mathbf{v}_{ij}$; see Rosenbaum (2010, §9.2). Among these 128 re-paired inexact pairs, none differed on the groups in Figure 2, 126 pairs differed on exactly one of the six basic covariates in $\mathbf{v}_{ij}$, and 2 pairs differed on two of the basic covariates.

5 Change in work income following the Chilean earthquake

In the earthquake data, there were 1978 pairs that were exactly matched, $\mathbf{v}_{i1} = \mathbf{v}_{i2} = \mathbf{v}_i$, for the six basic covariates mentioned in §1.2. The outcome is the change in individual work income from before the earthquake to after. Under H_0, we have $R_{ij} = r_{Cij}$, so that R_{ij} is fixed, and $|Y_i| = |R_{i1} - R_{i2}| = |r_{Ci1} - r_{Ci2}|$ is fixed. Using a regression tree (Breiman et al.'s (1984, §8) CART as implemented in `R` by the `rpart` package with the default settings) applied to the 1978 exact match pairs, the rank of $|Y_i|$ was predicted from the value of $\mathbf{v}_i$; see Figure 2. The regression tree formed three groups, $\mathcal{G} = \{\mathfrak{g}_1, \mathfrak{g}_2, \mathfrak{g}_3\}$, a group with at least some work income in 2009 (labeled "p" for positive in Table 2), and two groups with 0 pesos of work income in 2009, namely males and females (labeled "zm" and "zf" respectively in Table 2).

Because the match was of strength $k = 3$ for $\mathbf{v}_{ij}$, the remaining $128 = 2106 - 1978 = I - |\mathcal{I}|$ inexact pairs were balanced for quartile of work income and sex, so is was possible to break the pairing for these 128 pairs, then pair again to control exactly work income and sex, retaining the strength $k = 3$ balance on all of $\mathbf{v}_{ij}$. Now, all $I = 2106$ pairs are exactly matched for the variable that define the groups, $\mathcal{G} = \{\mathfrak{g}_1, \mathfrak{g}_2, \mathfrak{g}_3\}$.

Table 2 contrasts two sensitivity analyses. The first sensitivity analysis ("combined") tests the null hypothesis of no treatment effect at all, H_0, with no attempt to consider subgroups, and does this using an M-test of the type suggested by Maritz (1979) and that is similar to a lightly trimmed mean, with Huber's ψ-function, $\psi(d) = \max\{-1, \min(1, d)\}$ applied to Y_i/s where s is the upper 1% quantile of $|Y_i|$. The second sensitivity analysis uses the groups defined in Figure 2, and calculates the same M-test within each of the three groups, combining their P-values using the truncated product of P-values . The second sensitivity analysis tests no effect at all, H_0, using the truncated product of three P-values for the three groups (zf.zm.p in Table 2), as suggested in Hsu et al. (2013), and if H_0 is rejected then tests hypotheses about subgroups, as developed in Proposition 8.

In Table 2, the truncated product zf.zm.p reports less sensitivity to bias than the combined test, the former being insensitive to $\Gamma = 1.45$, and the latter being sensitive to $\Gamma = 1.25$. Using the device in Rosenbaum and Silber (2009b), an unobserved covariate that doubled the odds of exposure to the treatment ($Z_{i1} - Z_{i2} = 1$) and doubled the odds of a positive pair difference in outcomes ($Y_i > 0$) corresponds with $\Gamma = 1.25$, whereas an unobserved covariate that doubled the odds exposure to the treatment ($Z_{i1} - Z_{i2} = 1$) and tripled the odds of a positive pair difference in outcomes ($Y_i > 0$) corresponds with $\Gamma = 1.4$.

Proposition 8 permits more to be said. In the absence of bias in exposure to the earthquake, $\Gamma = 1$, there is less of an increase in work income in all three groups among those exposed to the earthquake than among those not exposed; however, this finding is sensitive to a bias of $\Gamma = 1.45$ except for men without work income before the earthquake. The strongest evidence of an effect of the earthquake on work income is among men without work income prior to the earthquake: those exposed to the earthquake were less likely to find jobs and have work income after the earthquake than similar men far from the earthquake.

The novel aspect of Table 2 is that the three groups $\mathcal{G} = \{\mathfrak{g}_1, \mathfrak{g}_2, \mathfrak{g}_3\}$ were built using the data at hand, yet Proposition 8 implies that the family-wise error rate has been controlled with data-dependent groups and multiple tests in a sensitivity analysis that allows for a bias of $\Gamma = 1.45$.

6 Simulation: Do attentive groups increase the power of a sensitivity analysis?

Propositions 5 and 8 in §3 concern the family-wise error rate in a sensitivity analysis when groups are attentively determined using the data at hand. The simulation will (i) check the claims of Propositions 5 and 8, (ii) examine the ability of CART regression of $|Y_i|$ on $\mathbf{v}_i$ to identify relevant subgroups, (iii) examine various concepts of power.

We hope to report insensitivity to bias when the association between treatment Z_{ij} and response R_{ij} is produced by an actual treatment effect, not by bias in assigning treatments. Therefore, the power of a sensitivity analysis is evaluated when, unknown to the investigator, the treatment is effective and there is no unmeasured bias. In this situation, the power of an α-level sensitivity analysis performed with a specific value of $\Gamma \geq 1$ is the probability that the upper bound on the P-value will be less than or equal to α when computed with this Γ; see Rosenbaum (2004, 2005, 2010) for detailed discussion. For instance, if a 0.05-level sensitivity analysis performed with $\Gamma = 2$ has power 0.99, then it is very likely that the investigator will be able to assert that the ostensible treatment effect could not be explained away by a bias of magnitude $\Gamma = 2$.

6.1 Structure of the simulation

The structure of the simulation follows.

(i) Six potential effect modifiers. In parallel with the example in §1.2, the simulation considers six covariates $\mathbf{v}$ as potential effect modifiers. Each of these covariates

is binary, and they are six independent Bernoulli trials with probability of success 1/2. Of these, at most two of the covariates interact with the treatment to affect the responses, affecting either the mean or the variance of the pair differences Y_i, but it is left to the regression tree to discover which covariates affect the response. A tree may fail by splitting in the wrong place. A first question addressed by the simulation is the degree to which the trees accurately group pairs.

(ii) Building trees. In all cases, there are $I = 2000$ pairs, and of these $|\mathcal{I}| = 1000$ are used to build the regression tree. For comparison, in §5 there were $I = 2106$ pairs and $|\mathcal{I}| = 1978$ pairs were used to build the tree. The tree is built from the CART regression of the ranks of 1000 absolute pair differences $|Y_i|$ on the $\mathbf{v}_i$ for the 1000 exactly matched pairs, in parallel with §5. The CART was fitted using the `rpart` package in `R` with complexity parameter set to 0.005. The remaining 1000 pairs of the $I = 2000$ pairs were classified using the tree constructed from the first 1000 pairs. Each sampling situation was replicated 5000 times, so an estimated power or an estimated family-wise error rate has standard error of at most $\sqrt{0.5 \times 0.5/5000} = 0.0071$.

(iii) Sampling situations. Table 3 describes nine sampling situations with Normal errors and constant variance of matched pair differences, Y_i, whereas Table 4 permits the variance of Y_i to change with the covariates. Case G is the null case, with all $Y_i \sim N(0, 1)$. Cases M and N have $E(Y_i) = 0$ but the variance changes with the covariates. In all other cases, the average effect is 1/2 averaging over the four equally likely cells defined by the first two covariates, which are the only active covariates. In Table 4, the average variance over the active cases is 1. In case U, the expected effect is constant but the variance changes.

6.2 Evaluating the groups

How can we judge whether groups $\mathcal{G} = \{\mathfrak{g}_1, \ldots, \mathfrak{g}_G\}$ built by CART are in fact good groups? In each sampling situation, let $\mu_i = \mathrm{E}(Y_i)$ and $\sigma_i^2 = \mathrm{var}(Y_i)$, and of course in a simulation we know μ_i and σ_i^2. Remember that in all simulated cases, μ_i and σ_i^2 vary with at most two of the binary covariates, so there are at most four values of each. A tree yields leaves or groups $\mathcal{G} = \{\mathfrak{g}_1, \ldots, \mathfrak{g}_G\}$. Write $\overline{\mu}_g = |\mathfrak{g}_g|^{-1} \sum_{i \in \mathfrak{g}_g} \mu_i$ for the average expectation in group g. We say that a tree is "perfect" if $\mu_i = \overline{\mu}_g$ for every $i \in \mathfrak{g}_g$, for every g; that is, if

the groups always separate pair differences with different expectations. Perfection is too much to expect. For every $\mathcal{G}$, we quantify departures from perfection by

$$\iota_{\mathcal{G}} = \frac{\sum_{g=1}^{G} \sum_{i \in \mathfrak{g}_g} \left(\mu_i - \overline{\mu}_g\right)^2 + \sigma_i^2}{\sum_{g=1}^{G} \sum_{i \in \mathfrak{g}_g} \sigma_i^2},$$

which is the fractional increase in the mean square error from grouping by $\mathcal{G} = \{\mathfrak{g}_1, \ldots, \mathfrak{g}_G\}$ rather than by a perfect grouping. A perfect tree has $\iota_{\mathcal{G}} = 1$. For comparison, we also compute $\iota_{\mathcal{A}}$ where $\mathcal{A}$ is a single group of all the pairs, $\mathcal{A} = \{\mathfrak{g}_1\}$ with $\mathfrak{g}_1 = \{1, \ldots, I\}$.

In Tables 3 and 4, the mean of $\iota_{\mathcal{G}}$ and $\iota_{\mathcal{A}}$ is reported for nine sampling situations, each replicated 5000 times. Without groups, the increase in mean square error ranges from $\iota_{\mathcal{A}} = 1.000$ for the null case G, to $\iota_{\mathcal{A}} = 1.375$ in several other cases. In contrast, the groups formed from the tree $\iota_{\mathcal{G}}$ are typically much better. Tables 3 and 4 also record the fraction of trees that are perfect; however, frequent imperfection is compatible with near perfection, that is $1 \approx \iota_{\mathcal{G}} \ll \iota_{\mathcal{A}}$.

Tables 3 and 4 also record the number of trees, out of 5000 trees, that had a single leaf, so CART produced just one group consisting of all I pairs. For instance, in cases F and G, the pairs are homogeneous, and more than 4000 of the 5000 trees had a single leaf. The good power of CART groups in homogenous cases like F in §6.4 partly reflects CART's typical decision not to form subgroups in homogeneous cases.

6.3 Level of the tests

Propositions 5 and 8 make assertions about the level of certain tests or testing procedures. Specifically, Proposition 5 says that whenever a group of pairs is entirely unaffected, a test with nominal level α will falsely reject with probability at most α, despite the fact that the groups were built using the data. Proposition 8 says that when closed testing is applied with component testing having nominal level α, the family-wise error rate is strongly controlled at α: the chance of falsely rejecting at least one true hypothesis is at most α. Is the simulation in agreement with these assertions?

Tables 3 and 4 record the number of null leaves, that is, the average over 5000 samples of the number of groups of pairs in which all pairs experience no treatment effect. A single tree may have no null leaves, one null leaf, or several null leaves. For example, case B typically had 2.114 null subgroups, while case D typically had 1.106. Tables 3 and 4 record the average over 5000 samples of the total number of leaves, null or not. In the homogenous case F, the average number of leaves or groups was 1.284 with no null leaves.

In Tables 3 and 4, the column "False rejections, All" is the proportion of null leaves in which the hypothesis of no effect was falsely rejected and for $\Gamma = 1$ it is consistently near 0.05. A false rejection cannot occur when every individual is affected by the treatment, and in these cases (F, H, I, S, T and U), the "False rejections" section of the table is blank. When there is no unmeasured bias, as in all the simulated examples, but the sensitivity analysis entertains the possibility of such bias, $\Gamma > 1$, the chance of false rejection is much less than 0.05. The column "Null rejects, family" is the proportion of applications of closed testing that issued in at least one false rejection, that is, a null leaf declared to have been affected. Here too, the 0.05 familywise level appears to have been preserved, consistent with the claim of Proposition 8.

In brief, building the groups by CART regression of $|Y_i|$ on $\mathbf{v}_i$ does not appear to have increased the probability of falsely rejecting a true null hypothesis, consistent with the claims of Propositions 5 and 8.

6.4 Power of the tests

The four columns of Tables 3 and 4 labeled "Rejecting H_0" give the power of four sensitivity analyses when testing no effect at all, H_0. Here, "one" is the combined test in Table 2 and "trunc" is the truncated product, truncated at 0.05. Also, "Fisher" is Fisher's combination of P-values used in place of the truncated product. Finally, "Simes" is the Simes method for combining independent P-values, and it is, by definition, a uniform improvement on use of the Bonferroni inequality. Consistent with asymptotic results in Hsu et al. (2013) about design sensitivity and limiting power of a sensitivity analysis, the combined method "one" is substantially inferior except when the effect is constant in case F. The truncated product and Simes method are similar and often best in terms of power, but they are not uniformly best; see, for instance, the homogeneous case F where the combined method and Fisher's method win by a small margin.

The final column requires some explanation. There are $I = 2000$ pairs in each simulated sample, but only some of these are affected by the treatment. For a pair that is affected, we may score a 1 if that pair is in a group for which the hypothesis of no effect is rejected by closed testing using the truncated product, and we may score a 0 otherwise. The final column, "Reject false H_0", is the average over 5000 replicates of the proportion of 1's among the affected pairs. For example, if this number were 0.5, then we expect half of the affected pairs to be in groups successfully identified by closed testing as "non-null groups."

This section is blank when there is no expected effect, as in cases G, M, and N. In case A, at $\Gamma = 9$, 56.2% of affected pairs were in groups where an effect was found, whereas the combined test found no evidence of an effect. In general, comparing the last column "Reject False H_0" to the column "Power to reject H_0-one" it is seen that the truncated product will often identify specific affected groups by closed testing at values of Γ such that the combined test has virtually no power to detect anything.

Particularly interesting is case U. In case U, the first two covariates affect the variance of Y_i but not its mean. Despite this, and consistent with results in Rosenbaum (2005) and Zubizarreta et al (2014), the single test has inferior power when compare to all of the tree-based methods that focus on subgroups.

In brief, a single test for all pairs is substantially inferior in terms of power in all simulated cases of effect modification, and it has only slightly higher power than the other methods when the effect is constant in case F. Closed testing using the truncated product will often identify affected groups when a single test would accept the null hypothesis of no effect at all.

7 Summary: It is useful to notice effect modification in observational studies

If there is effect modification in an observational study — if the magnitude or stability of an effect varies with measured pretreatment covariates — then the degree of sensitivity to unmeasured biases may vary markedly within subgroups defined by the effect modifiers. There may be stronger evidence of a treatment effect, evidence insensitive to small and moderate biases, in some subgroups than in others. Propositions 5 and 8 permit an empirical search for effect modifiers to be combined with a sensitivity analysis for subgroups that controls the family-wise error rate in the strong sense.

References

Benjamini, Y. and Heller, R. (2008). Screening for partial conjunction hypotheses. *Biometrics* **64** 1215-1222.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and Regression Trees.* New York: Chapman and Hall/CRC.

Cornfield, J., Haenszel, W., Hammond, E., Lilienfeld, A., Shimkin, M., Wynder, E. (1959). Smoking and lung cancer. *J. Nat. Cancer Inst.* **22** 173-203.

Donegani, M. (1991). An adaptive and powerful randomization test. *Biometrika* **78** 930-933.

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.

Hansen, B. B. (2007). Optmatch: flexible, optimal matching for observational studies. *R News* **7** 18-24. (Package `optmatch` in `R`)

Hedayat, A. S., Sloane, N. J. A., and Stufken, J. (1999). *Orthogonal Arrays*. New York: Springer.

Hochberg, Y. and Tamhane, A. C. (1987). *Mutiple Comparison Procedures*. New York: Wiley.

Hogg, R. V., Fisher, D. M. and Randles, R. H. (1975). A two-sample adpative distribution-free test. *J. Am. Statist. Assoc.* **70** 656-661.

Hosman, C. A., Hansen, B. B., and Holland, P. W. H. (2010). The sensitivity of linear regression coefficients' confidence limits to the omission of a confounder. *Ann. Appl Statist.* **4** 849-870.

Hsu, J. Y., Small, D. S. and Rosenbaum, P. R. (2013). Effect Modification and Design Sensitivity in Observational Studies. *J. Am. Statist. Assoc.* **108** 135-148.

Huber, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.* **35** 73-101.

Interior, Subsecretaría del Interior del Gobierno de Chile (2011). Informe final de fallecidos y desaparecidos por comuna. Accessed on December 21, 2013 from http://www.interior.gob.cl/filesapp/listado_fallecidos_desaparecidos_27Feb.pdf

Jogdeo, K. (1977). Association and probability inequalities. *Ann. Statist.* **5** 495-504.

Jones, D. H. (1979). An efficient adpative distribution-free test for location. *J. Am. Statist. Assoc.* **74** 822-828.

La Tercera (2010). Peor tragedia natural de los últimos 50 años deja huella de destrucción en zona centro sur. Accessed on December 21, 2013 from http://diario.latercera.com/2010/02/28/01/contenido/9_25213_9.html

Lehmann, E. L. and Romano, J. P. (2005). *Testing Statistical Hypotheses* (3^{rd} edition). New York: Springer.

Liu, W., Kuramoto, S. J., and Stuart, E. A. (2013). An introduction to sensitivity analysis for unobserved confounding in nonexperimental prevention research. *Prevent. Sci.* **14** 570-580.

Marcus, R., Eric, P. and Gabriel, K. R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* **63** 655-660.

Maritz, J. S. (1979). A note on exact robust confidence intervals for location. *Biometrika* **66** 163-166.

Mideplan (2011). Ficha Técnica Encuesta Post Terromoto. Accessed on December 8, 2011 from http://www.ministeriodesarrollosocial.gob.cl/encuestapostterremoto/index.html

Neyman, J. (1923, 1990). On the application of probability theory to agricultural experiments. *Statist. Sci.* **5** 463-480.

Pollitz FF, Brooks B, Tong X, et al. (2011). Coseismic slip distribution of the February 27, 2010 Mw 8.8 Maule, Chile earthquake. *Geophys. Res. Lett.* 2011; 38(9).

Rosenbaum, P. and Rubin, D. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *Am. Statist.* **39** 33-38.

Rosenbaum, P. R. (1987). Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika* **74** 13-26.

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

Rosenbaum, P. R. (2004). Design sensitivity in observational studies. *Biometrika* **91** 153-164.

Rosenbaum, P. R. (2005). Heterogeneity and causality: Unit heterogeneity and design sensitivity in observational studies. *Am. Statist.* **59** 147-152.

Rosenbaum, P. R. (2007). Sensitivity analysis for m-estimates, tests and confidence intervals in matched observational studies. Biometrics 63 456-464. (`R` package `sensitivitymv`)

Rosenbaum, P. R. and Silber, J. H. (2009a). Sensitivity analysis for equivalence and difference in an observational study of neonatal intensive care units. *J. Am. Statist. Assoc.* **104** 501-511.

Rosenbaum, P. R. and Silber, J. H. (2009b). Amplification of sensitivity analysis in observational studies. *J. Am. Statist. Assoc.* **104** 1398-1405. (`R` package `sensitivitymv`)

Rosenbaum, P. R. (2010). *Design of Observational Studies.* New York: Springer.

Rosenbaum, P. R. (2012). An exact, adaptive test with superior design sensitivity in an observational study of treatments for ovarian cancer. *Ann. App. Statist.* **6** 83-105.

Rosenbaum, P. R. (2013). Impact of multiple matched controls on design sensitivity in observational studies. *Biometrics* **69** 118-127

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Ed. Psych.* **66** 688-701.

Small, D. (2007). Sensitivity analysis for instrumental variables regression with overidentifying restrictions. *J. Am. Statist. Assoc.* **102** 1049-1058.

USGS. (2011a) Largest Earthquakes in the World Since 1900. (http://earthquake.usgs.gov/earthquakes/world

USGS. (2011b) Magnitude 8.8 - Offshore Bio-Bio, Chile. (http://earthquake.usgs.gov/earthquakes/recenteqsw

Wang, L. and Krieger, A. M. (2006). Causal conclusions are most sensitive to unobserved binary covariates. *Statist. Med.* **25** 2257-2271.

Yanagawa, T. (1984). Case-control studies: assessing the effect of a confounding factor. *Biometrika* **71** 191-194.

Yu, B. B., Gastwirth, J. L. (2005). Sensitivity analysis for trend tests: application to the risk of radiation exposure. *Biostatistics* **6** 201-209.

Zaykin, D. V., Zhivotovsky, L. A., Westfall, P. H., and Weir, B. S. (2002). Truncated product method of combining *P*-values. *Genetic Epidemiology* **22** 170-185. (`R` package `sensitivitymv`)

Zubizarreta, J. R., Reinke, C., Kelz, R. R., Silber, J. H., and Rosenbaum, P. R. (2011). Matching for several sparse nominal variables in a case-control study of readmission following surgery. *Am. Statist.* **65** 229-238.

Zubizarreta, J. R. (2012). Using mixed integer programming for matching in an observational study of kidney failure after surgery. *J. Am. Statist. Assoc.* **107** 1360-1371. (`R` software `mipmatch` at http://www-stat.wharton.upenn.edu/~josezubi/)

Zubizarreta, J. R., Cerda, M. and Rosenbaum, P. R. (2013). Effect of the 2010 Chilean earthquake on posttraumatic stress: reducing sensitivity to unmeasured bias through study design. *Epidemiology* **24** 79-87.

Zubizarreta, J. R., Paredes, R. D. and Rosenbaum, P. R. (2014). Matching for balance, pairing for heterogeneity in an observational study of the effectiveness of for-profit and not-for-profit high schools in Chile. *Ann. Applied Statist.*, to appear.

Table 1: Covariate balance for individuals exposed to severe shaking from the earthquake and matched controls. One of 20 strength-3 tables of covariate balance, this one for the 3 covariates gender, marital status and self-rated health. In all cells, the count in the exposed group equals the count in the control group, and the same is true for the other 19 tables (not shown) describing 3 of the 6 balanced covariates.

	Marital status	Married/cohabitating			Other		
Gender	Self-rated health	Poor	Fair	Good	Poor	Fair	Good
Male	Exposed	18	299	167	5	145	47
	Control	18	299	167	5	145	47
Female	Exposed	40	542	339	21	280	203
	Control	40	542	339	21	280	203

Table 2: Contrasting two sensitivity analyses for the change in work income following the earthquake. The "combined" test is a single m-test using all $I = 2106$ pairs with no attempt to discover effect modification. The three tests within groups use the same test within each of the groups defined by the regression tree, namely "zl" for zero-work-income-female, "zm" for zero-work-income-male, and "p" for positive-work-income. These individual P-values are combined using the truncated product of P-values truncated at 0.05, so zf.zm combines the two P-values for pairs with zero work income before the earthquake. Closed testing starts with zf.zm.p, continuing to subhypotheses only if certain rejections take place. When testing the null hypothesis H_0 of no effect at all, the combined test is sensitive at $\Gamma = 1.3$ while the truncated product zf.zm.p is insensitive at $\Gamma = 1.45$. Although the null hypothesis of no effect is rejected in all groups at $\Gamma = 1$, at $\Gamma = 1.45$ no effect is rejected only for men with no work income prior to the earthquake. In each column, the least sensitive P-value bound significant at the 0.05 level in closed testing is in bold.

	Overall tests		Two groups			Individual groups		
Γ	Combined	zf.zm.p	zf.zm	zf.p	zm.p	zf	zm	p
1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	**0.009**
1.1	0.001	0.000	0.000	0.006	0.004	0.002	0.001	0.079
1.2	**0.023**	0.000	0.000	0.018	0.007	0.008	0.002	0.288
1.25	0.067	0.001	0.000	0.031	0.010	**0.015**	0.004	0.437
1.3	0.151	0.002	0.001	0.051	0.013	0.026	0.006	0.589
1.35	0.280	0.004	0.002	0.080	0.018	0.041	0.008	0.724
1.4	0.440	0.037	0.024	1.000	0.024	0.062	0.011	0.829
1.45	0.606	**0.048**	**0.031**	1.000	**0.031**	0.089	**0.015**	0.903
1.5	0.749	0.061	0.040	1.000	0.040	0.123	0.020	0.948
1.6	0.925	0.096	0.065	1.000	0.065	0.210	0.033	0.988

Table 3: Summary of evaluating the groups, level of the tests, and power of the tests for the null hypothesis of no treatment effect with various Γ when matched pair differences have Normal errors and constant variance

Scenario		# Trees	Avg. Leaves		Avg. MSE				False Rejections		Power to Reject H_0				Reject
		1-leaf	Null	Total	Perfect	$\iota_{\mathscr{G}}$	$\iota_{\mathscr{A}}$	Γ	All	Family	one	Fisher	Simes	trunc	False H_0
A	X1=0 X2=0: N(1,1); X1=0 X2=1: N(1,1); X1=1 X2=0: N(0,1); X1=1 X2=1: N(0,1)	0	1.113	2.354	1.000	1.000	1.250	1.0	0.048	0.047	1.000	1.000	1.000	1.000	1.000
								2.5	0.000	0.000	0.701	1.000	1.000	1.000	1.000
								2.6	0.000	0.000	0.436	1.000	1.000	1.000	1.000
								9.0	0.000	0.000	0.000	0.388	0.591	0.598	0.562
B	X1=0 X2=0: N(2,1); X1=0 X2=1: N(0,1); X1=1 X2=0: N(0,1); X1=1 X2=1: N(0,1)	0	2.114	3.132	1.000	1.000	1.749	1.0	0.047	0.047	1.000	1.000	1.000	1.000	1.000
								2.0	0.000	0.000	0.721	1.000	1.000	1.000	1.000
								2.1	0.000	0.000	0.381	1.000	1.000	1.000	1.000
								30.0	0.000	0.000	0.000	0.858	0.999	0.999	0.994
C	X1=0 X2=0: N(3/2,1); X1=0 X2=1: N(0,1); X1=1 X2=0: N(0,1); X1=1 X2=1: N(1/2,1)	0	1.400	3.585	0.292	1.022	1.375	1.0	0.047	0.043	1.000	1.000	1.000	1.000	1.000
								2.3	0.000	0.000	0.817	1.000	1.000	1.000	0.615
								2.5	0.000	0.000	0.290	1.000	1.000	1.000	0.581
								15.0	0.000	0.000	0.000	0.803	0.993	0.997	0.487
D	X1=0 X2=0: N(6/5,1); X1=0 X2=1: N(4/5,1); X1=1 X2=0: N(0,1); X1=1 X2=1: N(0,1)	0	1.106	3.305	0.902	1.002	1.270	1.0	0.044	0.044	1.000	1.000	1.000	1.000	1.000
								2.3	0.000	0.000	0.967	1.000	1.000	1.000	1.000
								2.5	0.000	0.000	0.631	1.000	1.000	1.000	0.999
								10.0	0.000	0.000	0.000	0.564	0.849	0.849	0.409
E	X1=0 X2=0: N(3/2,1); X1=0 X2=1: N(1/2,1); X1=1 X2=0: N(0,1); X1=1 X2=1: N(0,1)	0	1.098	3.295	0.935	1.002	1.375	1.0	0.054	0.054	1.000	1.000	1.000	1.000	1.000
								2.3	0.000	0.000	0.818	1.000	1.000	1.000	0.873
								2.5	0.000	0.000	0.281	1.000	1.000	1.000	0.759
								20.0	0.000	0.000	0.000	0.222	0.826	0.822	0.398
F	X1=0 X2=0: N(1/2,1); X1=0 X2=1: N(1/2,1); X1=1 X2=0: N(1/2,1); X1=1 X2=1: N(1/2,1)	4122	0.000	1.284	1.000	1.000	1.000	1.0			1.000	1.000	1.000	1.000	1.000
								2.8			0.809	0.801	0.777	0.782	0.738
								3.0			0.375	0.365	0.347	0.348	0.327
								3.2			0.081	0.081	0.079	0.079	0.072
G	X1=0 X2=0: N(0,1); X1=0 X2=1: N(0,1); X1=1 X2=0: N(0,1); X1=1 X2=1: N(0,1)	4128	1.285	1.285	1.000	1.000	1.000	1.0	0.049	0.048	0.049	0.049	0.048	0.048	
								1.1	0.002	0.001	0.001	0.001	0.001	0.001	
								1.2	0.000	0.000	0.000	0.000	0.000	0.000	
								1.3	0.000	0.000	0.000	0.000	0.000	0.000	
H	X1=0 X2=0: N(3/5,1); X1=0 X2=1: N(3/5,1); X1=1 X2=0: N(2/5,1); X1=1 X2=1: N(2/5,1)	3094	0.000	1.587	0.263	1.007	1.010	1.0			1.000	1.000	1.000	1.000	1.000
								2.8			0.767	0.822	0.810	0.812	0.640
								3.0			0.321	0.477	0.487	0.487	0.331
								3.2			0.065	0.245	0.277	0.278	0.148
I	X1=0 X2=0: N(3/4,1); X1=0 X2=1: N(3/4,1); X1=1 X2=0: N(1/4,1); X1=1 X2=1: N(1/4,1)	181	0.000	2.335	0.949	1.003	1.062	1.0			1.000	1.000	1.000	1.000	1.000
								2.8			0.577	0.976	0.975	0.975	0.494
								3.0			0.164	0.959	0.959	0.959	0.479
								3.2			0.019	0.955	0.956	0.956	0.474

NOTE: There are six potential effect modifiers, $X1$–$X6$, following six independent Bernoulli trials with probability of success 1/2. At most two of the covariates, say $X1$ and/or $X2$, interact with the treatment to affect the response. There are $I = 2000$ matched pairs, $Y_i = \delta_{00} + \varepsilon_i$ for $X1_i = 0$ and $X2_i = 0$, $Y_i = \delta_{01} + \varepsilon_i$ for $X1_i = 0$ and $X2_i = 1$, $Y_i = \delta_{10} + \varepsilon_i$ for $X1_i = 1$ and $X2_i = 0$, and $Y_i = \delta_{11} + \varepsilon_i$ for $X1_i = 1$ and $X2_i = 1$, where ε_i's are independently drawn from the standard Normal distribution. The statistic is M-statistic. Each situation is sampled 5,000 times. "# Trees, 1-leaf" is number of single-leaf tree among 5,000 replicates. "Avg. Leaves, Null" and "Avg. Leaves, Total" are averaged null and total leaves over 5,000 replicates. The $\iota_{\mathscr{G}}$ and $\iota_{\mathscr{A}}$ quantify departures from perfection, where $\mathscr{G} = \{\mathfrak{g}_1, \ldots, \mathfrak{g}_G\}$ and $\mathscr{A} = \{\mathfrak{g}_1\}$. A perfect tree has $\iota_{\mathscr{G}} = 1$. "False Rejections, All" is the proportion of null leaves in which the hypothesis of no effect was falsely rejected. "False Rejections, Family" is the proportion of applications of closed testing that issued in at least one false rejection. "Power to Reject H_0" gives the power of four sensitivity analyses when testing no effect at all, H_0. Here, "one" is the combined test, "Fisher" is Fisher's combination of P-values, "Simes" is the Simes method for combining independent P-values, and "trunc" is the truncated product. Finally, "Reject, False H_0" is the proportion of pairs in a group for which the hypothesis of no effect is rejected by closed testing using the truncated product, averaging over affected pairs and then 5,000 replicates.

Table 4: Summary of evaluating the groups, level of the tests, and power of the tests for the null hypothesis of no treatment effect with various Γ when matched pair differences have Normal errors and different variances

Scenario				# Trees 1-leaf	Avg. Leaves Null	Avg. Leaves Total	Avg. MSE Perfect	Avg. MSE $\iota_{\mathscr{G}}$	Avg. MSE $\iota_{\mathscr{A}}$	Γ	False Rejections All	False Rejections Family	Power to Reject H_0 one	Power to Reject H_0 Fisher	Power to Reject H_0 Simes	Power to Reject H_0 trunc	Reject False H_0
J		X2=0	X2=1	0	1.131	2.501	1.000	1.000	1.250	1.0	0.045	0.046	1.000	1.000	1.000	1.000	1.000
	X1=0	N(1,2/3)	N(1,4/3)							2.5	0.000	0.000	0.732	1.000	1.000	1.000	1.000
	X1=1	N(0,1)	N(0,1)							2.6	0.000	0.000	0.478	1.000	1.000	1.000	1.000
										9.0	0.000	0.000	0.000	0.491	0.674	0.676	0.568
K		X2=0	X2=1	98	1.164	2.345	0.971	1.007	1.250	1.0	0.050	0.049	1.000	1.000	1.000	1.000	1.000
	X1=0	N(1,2/3)	N(1,2/3)							2.5	0.000	0.000	0.884	0.993	0.993	0.993	0.990
	X1=1	N(0,4/3)	N(0,4/3)							2.6	0.000	0.000	0.700	0.987	0.986	0.986	0.983
										9.0	0.000	0.000	0.000	0.975	0.976	0.976	0.969
L		X2=0	X2=1	0	1.054	2.300	1.000	1.000	1.250	1.0	0.045	0.045	1.000	1.000	1.000	1.000	1.000
	X1=0	N(1,4/3)	N(1,4/3)							2.5	0.000	0.000	0.485	1.000	1.000	1.000	1.000
	X1=1	N(0,2/3)	N(0,2/3)							2.6	0.000	0.000	0.230	1.000	1.000	1.000	1.000
										9.0	0.000	0.000	0.000	0.000	0.001	0.001	0.001
M		X2=0	X2=1	0	2.381	2.381	1.000	1.000	1.000	1.0	0.051	0.049	0.057	0.056	0.051	0.051	
	X1=0	N(0,2/3)	N(0,4/3)							1.1	0.003	0.001	0.001	0.001	0.001	0.001	
	X1=1	N(0,2/3)	N(0,4/3)							1.2	0.000	0.000	0.000	0.000	0.000	0.000	
										1.3	0.000	0.000	0.000	0.000	0.000	0.000	
N		X2=0	X2=1	1286	4.209	4.209	1.000	1.000	1.000	1.0	0.049	0.040	0.051	0.051	0.053	0.054	
	X1=0	N(0,2/3)	N(0,4/3)							1.1	0.009	0.001	0.000	0.000	0.004	0.001	
	X1=1	N(0,4/3)	N(0,2/3)							1.2	0.002	0.000	0.000	0.000	0.000	0.000	
										1.3	0.000	0.000	0.000	0.000	0.000	0.000	
O		X2=0	X2=1	0	1.108	3.304	0.378	1.019	1.375	1.0	0.051	0.051	1.000	1.000	1.000	1.000	1.000
	X1=0	N(3/2,4/3)	N(1/2,2/3)							2.3	0.000	0.000	0.816	1.000	1.000	1.000	0.688
	X1=1	N(0,1)	N(0,1)							2.5	0.000	0.000	0.279	1.000	1.000	1.000	0.687
										20.0	0.000	0.000	0.000	0.000	0.049	0.043	0.021
P		X2=0	X2=1	0	2.121	4.257	0.954	1.001	1.375	1.0	0.051	0.050	1.000	1.000	1.000	1.000	1.000
	X1=0	N(3/2,9/5)	N(1/2,1/5)							2.3	0.000	0.000	0.879	1.000	1.000	1.000	0.980
	X1=1	N(0,1)	N(0,1)							2.5	0.000	0.000	0.380	1.000	1.000	1.000	0.978
										20.0	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Q		X2=0	X2=1	0	1.085	3.282	0.009	1.031	1.375	1.0	0.052	0.052	1.000	1.000	1.000	1.000	1.000
	X1=0	N(3/2,4/3)	N(0,1)							2.3	0.000	0.000	0.820	1.000	1.000	1.000	0.505
	X1=1	N(0,1)	N(1/2,2/3)							2.5	0.000	0.000	0.280	1.000	1.000	1.000	0.505
										15.0	0.000	0.000	0.000	0.100	0.582	0.568	0.276
R		X2=0	X2=1	0	2.093	4.224	0.921	1.002	1.375	1.0	0.050	0.055	1.000	1.000	1.000	1.000	1.000
	X1=0	N(3/2,9/5)	N(0,1)							2.3	0.000	0.000	0.891	1.000	1.000	1.000	0.963
	X1=1	N(0,1)	N(1/2,1/5)							2.5	0.000	0.000	0.392	1.000	1.000	1.000	0.962
										15.0	0.000	0.000	0.000	0.001	0.010	0.008	0.003
S		X2=0	X2=1	4074	0.000	1.294	0.031	1.060	1.062	1.0			1.000	1.000	1.000	1.000	1.000
	X1=0	N(3/4,2/3)	N(3/4,2/3)							2.8			0.785	0.791	0.768	0.771	0.715
	X1=1	N(1/4,4/3)	N(1/4,4/3)							3.0			0.359	0.385	0.373	0.375	0.332
										3.2			0.070	0.113	0.113	0.113	0.082
T		X2=0	X2=1	0	0.000	2.328	1.000	1.000	1.062	1.0			1.000	1.000	1.000	1.000	1.000
	X1=0	N(3/4,4/3)	N(3/4,4/3)							2.8			0.561	1.000	1.000	1.000	0.495
	X1=1	N(1/4,2/3)	N(1/4,2/3)							3.0			0.147	0.995	0.997	0.998	0.489
										3.2			0.019	0.972	0.987	0.990	0.477
U		X2=0	X2=1	0	0.000	4.294	1.000	1.000	1.000	1.0			1.000	1.000	1.000	1.000	1.000
	X1=0	N(1/2,1/4)	N(1/2,3/4)							3.0			0.830	1.000	1.000	1.000	0.390
	X1=1	N(1/2,1)	N(1/2,2)							3.2			0.445	1.000	1.000	1.000	0.332
										3.4			0.127	1.000	1.000	1.000	0.297

NOTE: There are six potential effect modifiers, $X1$–$X6$, following six independent Bernoulli trials with probability of success 1/2. At most two of the covariates, say $X1$ and/or $X2$, interact with the treatment to affect the response. There are $I = 2000$ matched pairs, $Y_i = \delta_{00} + \sigma_{00}\varepsilon_i$ for $X1_i = 0$ and $X2_i = 0$, $Y_i = \delta_{01} + \sigma_{01}\varepsilon_i$ for $X1_i = 0$ and $X2_i = 1$, $Y_i = \delta_{10} + \sigma_{10}\varepsilon_i$ for $X1_i = 1$ and $X2_i = 0$, and $Y_i = \delta_{11} + \sigma_{11}\varepsilon_i$ for $X1_i = 1$ and $X2_i = 1$, where ε_i's are independently drawn from the standard Normal distribution. The statistic is M-statistic. Each situation is sampled 5,000 times. "# Trees, 1-leaf" is number of single-leaf tree among 5,000 replicates. "Avg. Leaves, Null" and "Avg. Leaves, Total" are averaged null and total leaves over 5,000 replicates. The $\iota_{\mathscr{G}}$ and $\iota_{\mathscr{A}}$ quantify departures from perfection, where $\mathscr{G} = \{\mathfrak{g}_1, \ldots, \mathfrak{g}_G\}$ and $\mathscr{A} = \{\mathfrak{g}_1\}$. A perfect tree has $\iota_{\mathscr{G}} = 1$. "False Rejections, All" is the proportion of null leaves in which the hypothesis of no effect was falsely rejected. "False Rejections, Family" is the proportion of applications of closed testing that issued in at least one false rejection. "Power to Reject H_0" gives the power of four sensitivity analyses when testing no effect at all, H_0. Here, "one" is the combined test, "Fisher" is Fisher's combination of P-values, "Simes" is the Simes method for combining independent P-values, and "trunc" is the truncated product. Finally, "Reject, False H_0" is the proportion of pairs in a group for which the hypothesis of no effect is rejected by closed testing using the truncated product, averaging over affected pairs and then 5,000 replicates.

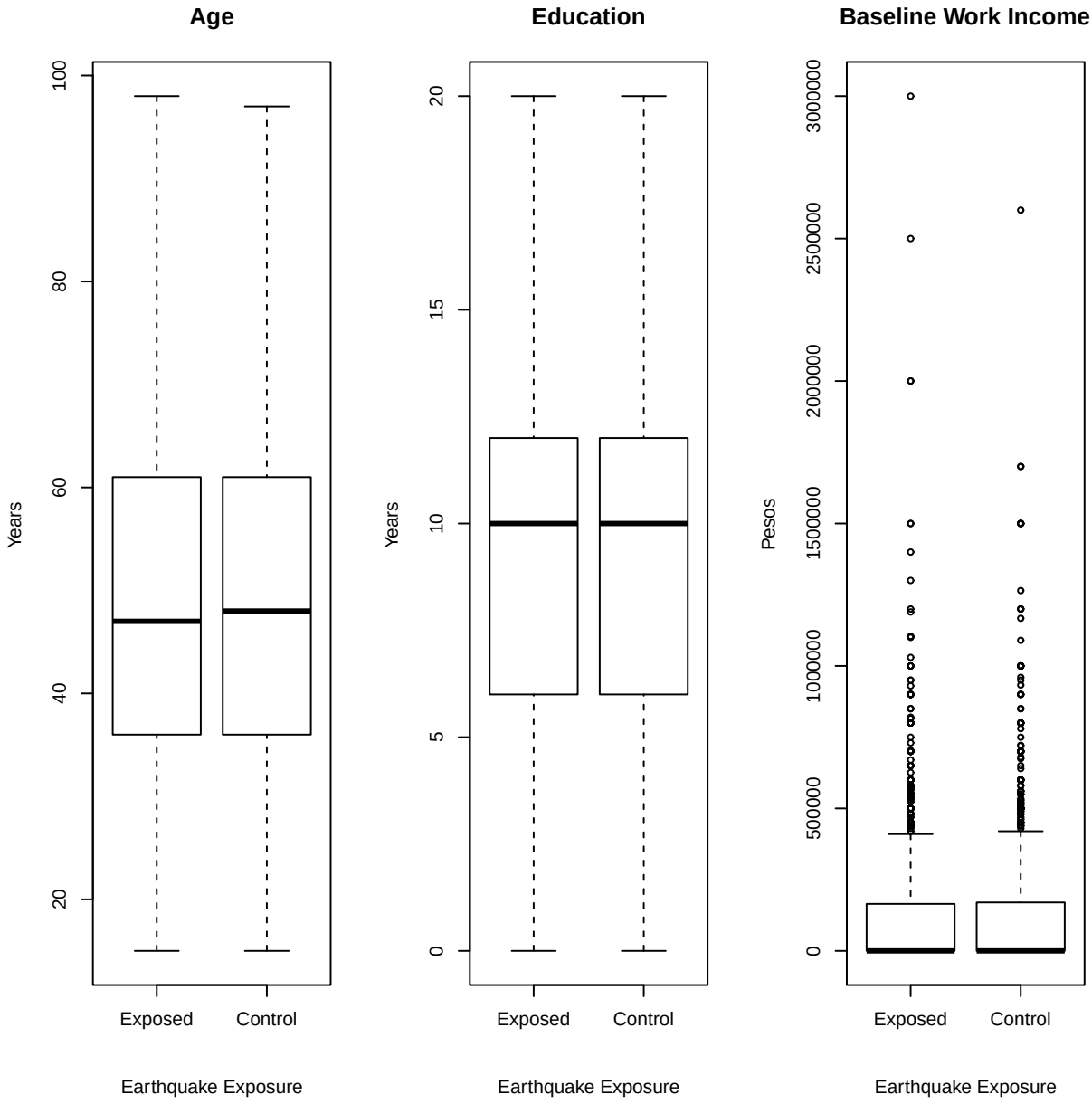

Figure 1: Balance for three continuous covariates in 2009, before the earthquake, in *I* = 2016 matched pairs containing one individual from a severely shaken region of Chile and one control from a region barely touched by the earthquake.

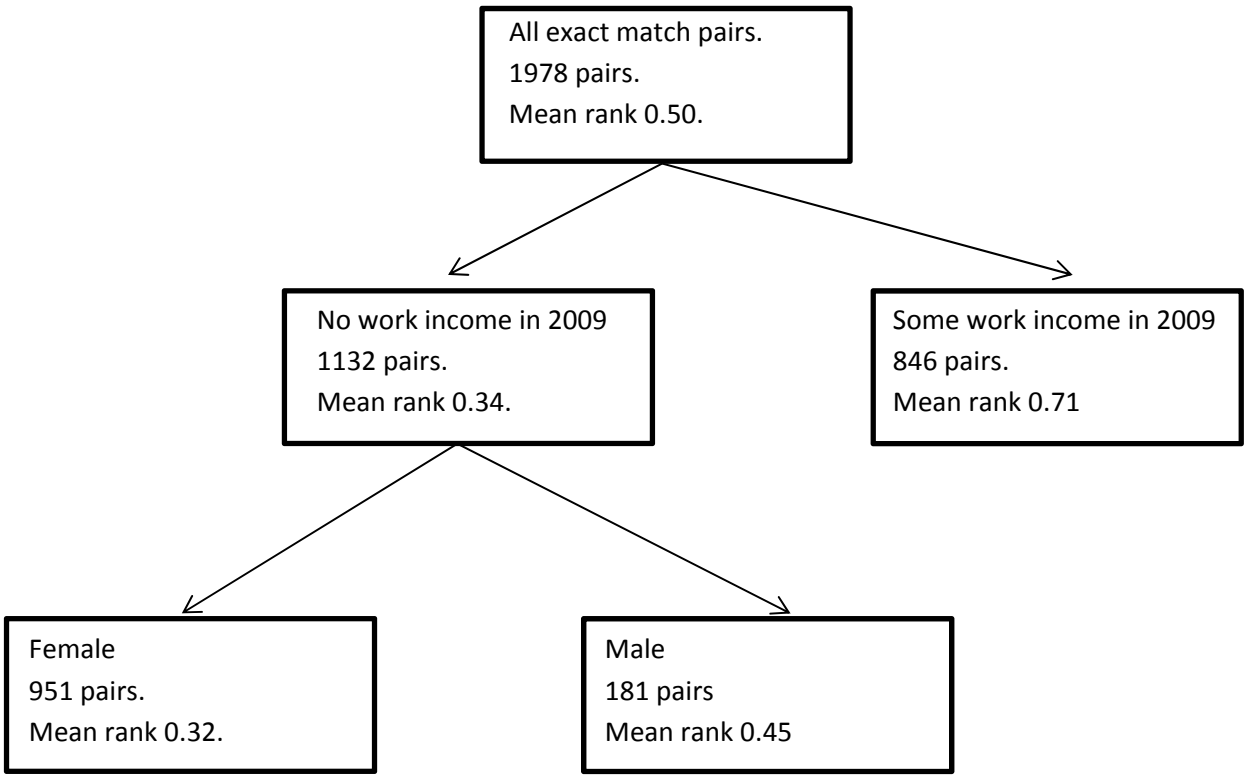

Figure 2: Regression tree built from the ranks of the absolute differences in work income for the 1978 pairs that were exactly matched for all 6 balanced covariates. Ranks were divided by 1978, so that they fall in [0, 1].