A RECENT OBSERVATIONAL STUDY USED TO ILLUSTRATE RECENT METHODOLOGY FOR SUCH STUDIES

PAUL R. ROSENBAUM

Abstract. A recent observational study is used to illustrate two recent methodological proposals, namely optimal matching with fine balance and sensitivity analysis for uncommon but dramatic responses to treatment.

1. Matching with Fine Balance

1.1. **What is fine balance?** Fine balance constrains an optimal match to exactly balance the marginal distributions of a nominal variable, perhaps one with many levels, placing no restrictions on who is matched to whom. The nominal variable might have many categories, like zip code, or the direct product of several nominal variables. Propensity scores balance covariates stochastically, but this may be inadequate for a nominal variable with many levels. Use in conjunction with calipers on the propensity score, a version of a Mahalanobis distance, possibly penalties for additional constraints.

1.2. **How do you construct an optimal finely balanced match?** The assignment algorithm [1] takes a matrix of distances and finds the pairing of rows and columns that minimizes the total distance within pairs. The problem is not trivial because two rows may want the same column. For an $n \times n$ matrix, it is possible to solve the optimal assignment problem in $O\left(n^3\right)$ arithmetic operations, which is the same order as for multiplying two $n \times n$ matrices in the conventional way. In R, the pairmatch function in Ben Hanson's [4] optmatch package solves the assignment problem. In SAS OR, proc assign solves the assignment problem. Hansen's optmatch package calls Bertsekas [1] Fortran code.

To construct a finely balanced match, the distance matrix is patterned in such a way that the proper numbers of controls are deleted. Proposition 1 in [12] shows that this procedure produces a minimum distance match subject to the constraint that fine balance is achieved. A toy example is given in the other handout. **Software in R**: The balmatch function in the other handout creates the distance matrix, calls Hansen's pairmatch function, outputs the match. It is at: http://www-stat.wharton.upenn.edu/~rosenbap/index.html

2. Sensitivity Analysis For Uncommon But Dramatic Treatment Effects

2.1. **Randomization Inference in a Paired Experiment.** Observed covariate $\mathbf{x}$ and an unobserved covariate u. I pairs, $i = 1, \ldots, I$, of two subjects, $j = 1, 2$, one treated, one

Supported by a grant from the MMS Program at NSF. Reprints, software and copies of overheads at: http://www-stat.wharton.upenn.edu/~rosenbap/index.html. October 2013.

control, matched for $\mathbf{x}$, so $\mathbf{x}_{i1} = \mathbf{x}_{i2}$, but not matched for u, so typically $u_{i1} \neq u_{i2}$. $Z_{ij} = 1$ if j received the treatment in pair i, and $Z_{ij} = 0$ if j received the control, so $Z_{i1} + Z_{i2} = 1$. Subject (i, j) has two potential responses, (r_{Tij}, r_{Cij}), r_{Tij} observed under treatment, $Z_{ij} = 1$, r_{Cij} observed under control, $Z_{ij} = 0$, so the effect of the treatment is $r_{Tij} - r_{Cij}$; Neyman [7] and Rubin [17]. Write $\mathcal{F}$ for $\{(r_{Tij}, r_{Cij}, \mathbf{x}_{ij}, u_{ij}),\ i = 1, \ldots, I,\ j = 1, 2\}$ and $\mathcal{Z}$ for the event $\{Z_{i1} + Z_{i2} = 1,\ i = 1, \ldots, I\}$; then $\mathcal{F}$ and $\mathcal{Z}$ are fixed by conditioning in Fisher's [3] theory of randomization inference. Randomization within pairs ensures $\Pr(Z_{i1} = 1 \mid \mathcal{Z}, \mathcal{F}) = \frac{1}{2}$, $i = 1, \ldots, I$, with independent assignments in distinct pairs. Fisher's sharp null hypothesis of no treatment effect is $H_0 : r_{Tij} = r_{Cij}$, $\forall i, j$. Observed response is $R_{ij} = Z_{ij} r_{Tij} + (1 - Z_{ij}) r_{Cij}$, and the treated-minus-control difference in responses in pair i is $D_i = (2Z_{i1} - 1)(R_{i1} - R_{i2})$, so that $D_i = (2Z_{i1} - 1)(r_{Ci1} - r_{Ci2})$ if H_0 is true. To test H_0 rank $|D_i|$ from 1 to I; then Wilcoxon's signed rank statistic, W, is the sum of the ranks for which $D_i > 0$. If q_i is the rank of $|D_i|$, Stephenson's [21] signed rank statistic S uses $\binom{q_i - 1}{m - 1}$ as a rank score in place of q_i for fixed integer $m \geq 2$, where $\binom{a}{b}$ is defined to be zero if $a < b$, so $S = \sum_{i=1}^{I} \chi(D_i > 0) \cdot \binom{q_i - 1}{m - 1}$, where $\chi(a) = 1$ if a is true, $= 0$ otherwise. S is virtually the same as Wilcoxon's W for $m = 2$.

Conover and Salsburg [2] found the locally most powerful rank test for comparing $r_{Cij} \sim_{iid} F$ to $r_{Tij} \sim_{iid} (1 - p) F + pF^m$ as $I \to \infty$ and $p \to 0$, so that only a small fraction p of treated subjects are affected by the treatment. Here, $F^m = F \times \cdots \times F$ is the distribution of the maximum of m iid observations from F, so of course F^m is stochastically larger than F. See also [6] and [18]. The Conover-Salsburg ranks are a polynomial in q_i of order $m - 1$, are not easy to interpret, but for large I behave in a manner similar to Stephenson's [21] ranks. The advantage of Stephenson's ranks is that they permit the rank test to be inverted to give confidence statements for the number or proportion of extreme responses caused by the treatment [13]. (For matched pair data, as here, take $r_{Cij} - \alpha_i \sim_{iid} F$ to $r_{Tij} - \alpha_i \sim_{iid} (1 - p) F + pF^m$ where α_i is a pair parameter.)

2.2. **Sensitivity to Departures from Random Assignment in Observational Studies.** (i) In the population, before matching, treatment assignments were independent, with unknown probabilities $\pi_{ij} = \Pr(Z_{ij} = 1 \mid \mathcal{F})$, (ii) subjects with same *observed* $\mathbf{x}_{ij}$ may differ in *unobserved* u_{ij} and hence in odds of treatment by factor of $\Gamma \geq 1$,

$$\frac{1}{\Gamma} \leq \frac{\pi_{ij}(1 - \pi_{ik})}{\pi_{ik}(1 - \pi_{ij})} \leq \Gamma,\ \ \forall i, j, k \tag{2.1}$$

and (iii) the distribution of treatments within treated/control matched pairs $\Pr(Z_{i1} = 1 \mid \mathcal{Z}, \mathcal{F})$ is then obtained by conditioning on $Z_{i1} + Z_{i2} = 1$. If $\Gamma = 1$, then $\mathbf{x}_{ij} = \mathbf{x}_{ik}$ ensures $\pi_{ij} = \pi_{ik}$, $i = 1, \ldots, I$, whereupon $\Pr(Z_{i1} = 1 \mid \mathcal{Z}, \mathcal{F}) = \pi_{i1}/(\pi_{i1} + \pi_{i2}) = \frac{1}{2}$, and the distribution of treatment assignments is again the randomization distribution: bias solely due to observed $\mathbf{x}$ can be eliminated by matching on $\mathbf{x}$. If $\Gamma > 1$ in (2.1), then matching on $\mathbf{x}$ may fail to equalize the π_{ij} in pair i. Γ is unknown. A sensitivity analysis calculates, for several values of Γ, the range of possible inferences. How large must Γ be before qualitatively different causal interpretations are possible?

2.3. **Sensitivity Analysis.** If (2.1) and $H_0 : \tau = \tau_0$ are true, then the null distribution of S is unknown but is bounded by two known distributions. Write $\theta = \Gamma/(1+\Gamma)$ so $\theta \geq \frac{1}{2}$ because $\Gamma \geq 1$. Write $\overline{\overline{S}}$ for the sum of I independent random variables taking value $\binom{i-1}{m-1}$ with probability θ and value 0 with probability $1-\theta$, $i = 1, \ldots, I$; also, write $\overline{S}$ for the sum of I independent random variables taking value $\binom{i-1}{m-1}$ with probability $1-\theta$ and value 0 with probability θ. Then (2.1) and $H_0 : \tau = \tau_0$ imply the sharp bounds

$$\Pr\left(\overline{S} \geq s\right) \leq \Pr\left(S \geq s \mid \mathcal{Z}, \mathcal{F}\right) \leq \Pr\left(\overline{\overline{S}} \geq s\right), \quad \forall s; \tag{2.2}$$

e.g., [8, 9, §4]. If $\Gamma = 1$, then equality in (2.2); otherwise bounds (2.2) widen as Γ increases. For testing H_0, the upper bound on the one-sided significance level is at most 0.05 for all π_{ij} satisfying (2.1) if $S \geq \widetilde{s}$ where $0.05 = \Pr\left(\overline{\overline{S}} \geq \widetilde{s}\right)$.

In the example, it appears that only some MO's treated very intensively. Presumably because of this, the results are less sensitive to unobserved biases when S is computed with $m = 5$ or $m = 10$ rather than $m = 2$ for Wilcoxon's test.

If unobserved bias led to a Δ-fold increase in the odds of a positive response, $D_i > 0$, and a Λ-fold increase in the odds of treatment, $Z_{i1} - Z_{i2} = 1$, then this is the same as a bias of $\Gamma = (\Delta\Lambda + 1)/(\Delta + \Lambda)$; see [14]. For instance, $\Gamma = 1.5$ corresponds with $\Delta = 4$, $\Lambda = 2$; $\Gamma = 3$ corresponds with $\Delta = 7$, $\Lambda = 5$; $\Gamma = 1.2$ with $\Delta = 2$, $\Lambda = 1.75$.

2.4. **Design Sensitivity.** The design sensitivity [5, 10, 11, 15, 16] refers to the limiting case, as the number of pairs increases, $I \to \infty$. Add a subscript I to denote quantities, say S_I, computed from a sample of size I. Then, for a given Γ, the maximum significance level for a sample of size I is ≤ 0.05 if $S_I \geq \widetilde{s}_I$ where $0.05 = \Pr\left(\overline{\overline{S}}_I \geq \widetilde{s}_I\right)$.

Suppose the treatment had an effect and there was no bias from the unobserved covariate u; call this the favorable situation. We would not be able to know that we are in the favorable situation from the observable data. The best we could hope to say is that the results are insensitive to bias, that is, for some large Γ we have $S_I \geq \widetilde{s}_I$. Consider a specific model that generated the I observations in the favorable situation. Then $\Pr\left(S_I \geq \widetilde{s}_I\right)$ tends to 0 or 1 as $I \to \infty$ depending upon the value of Γ. More precisely, there is a number, $\widetilde{\Gamma}$, called the design sensitivity, such that $\Pr\left(S_I \geq \widetilde{s}_I\right) \to 1$ for $\Gamma < \widetilde{\Gamma}$ and $\Pr\left(S_I \geq \widetilde{s}_I\right) \to 0$ for $\Gamma > \widetilde{\Gamma}$ as $I \to \infty$; i.e., $\widetilde{\Gamma}$ is the limiting sensitivity to unobserved bias in a favorable situation in which it would be desirable to report that the results are insensitive.

What makes a study design insensitive to unobserved biases? The answer is provided by comparing $\widetilde{\Gamma}$ for different designs. See [10, 11, 15, 16] for some comparisons.

Under the Conover-Salsburg [2] model, $r_{Cij} - \alpha_i \sim_{iid} F$ to $r_{Tij} - \alpha_i \sim_{iid} (1-p)\,F + pF^m$, with F standard Normal, the design sensitivity $\widetilde{\Gamma}$ is larger, sometimes much larger, when the analysis recognizes that only a fraction p of treated subjects respond to treatment by taking $m > 2$ in S_I.

References

[1] Bertsekas, D. P. (1981), "A new algorithm for the assignment problem," *Math. Prog.,* 21, 152-171. M

[2] Conover, W. J. and Salsburg, D. S. (1988), "Locally most powerful tests for detecting treatment effects when only a subset of patients can be expected to 'respond' to treatment," *Biometrics,* 44, 189-196. UD

[3] Fisher, R. A. (1935), *Design of Experiments*, Edinburgh: Oliver and Boyd. CE

[4] Hansen, B. B. (2007), "Optmatch," *R News*, 7, 18-24. M

[5] Hsu, J. Y., Small, D. S., Rosenbaum, P.R. (2013), "Effect modification and design sensitivity in observational studies," *JASA*, 108, 135-148. DS

[6] Lehmann, E. L. (1953), "The power of rank tests," *Ann. Math. Stat.*, 24, 23-43. UD

[7] Neyman, J. (1923, 1990), "On the application of probability theory to agricultural experiments," *Stat. Sci.*, 5, 463-480. CE

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

[9] Rosenbaum, P. R. (2002), *Observational Studies* (2^{nd} ed). NY: Springer. CE, M, NE, SA

[10] Rosenbaum, P. R. (2004), "Design sensitivity in observational studies," *Biometrika*, 91, 153-164. DS

[11] Rosenbaum, P. R. (2005), "Heterogeneity and causality," *Am. Stat.*, 59, 147-152. DS

[12] Rosenbaum, P. R., Ross R. N., and Silber, J. H.. (2007), "Minimum distance matched sampling with fine balance in an observational study of treatment for ovarian cancer," *JASA*, 2007; 102: 75-83. BT, M

[13] Rosenbaum, P. R. (2007), "Confidence intervals for uncommon but dramatic responses to treatment," *Biometrics*, 2007, 63, 1164–1171. BT, CE, SA, UD

[14] Rosenbaum, P. R., Silber, J. H. (2009), "Amplification of sensitivity analysis in observational studies," *JASA*, 104, 1398-1405. SA (Discusses interpretation of Γ.)

[15] Rosenbaum, P. R. (2010), *Design of Observational Studies*, NY: Springer. DS, SA

[16] Rosenbaum, P. R. (2010), "Design sensitivity and efficiency in observational studies," *JASA*, 105, 692-702. DS

[17] Rubin, D. B. (1974), "Estimating causal effects of treatments in randomized and nonrandomized studies," *J. Educ. Psych.*, 66, 688-701. CE

[18] Salsburg, D. S. (1986), "Alternative hypotheses for the effects of drugs in small-scale clinical studies," *Biometrics*, 42, 671-674. UD

[19] Silber, J. H., Rosenbaum, P. R., Polsky, D. , Ross, R. N., Even-Shoshan, O., Schwartz, S., Armstrong, K. A., Randall, T. C. (2007) Does ovarian cancer treatment and survival differ by the specialty providing chemotherapy? *Journal of Clinical Oncology (JCO)*, 2007; 25: 1169-1175. Related editorial: Cannistra, S. A. (2007) Gynecologic oncology or medical oncology: What's in a name? *Journal of Clinical Oncology*, 2007; 25: 1157-1159. 5 related letters and 2 rejoinders from S. Blank, J. Curtin, A. Berchuck, M. Hoffman, U. Iqbal, M. Markham, W. McGuire, J. Silber, P. Rosenbaum, and S. Cannistra. *JCO*, 2007; 25: 1151-1158. BT, NE

[20] Small, D. and Rosenbaum, P.R. (2008), "War and wages: the strength of instrumental variables and their sensitivity to unobserved biases," *JASA*, 103, 924-933. DS

[21] Stephenson, W. R. (1981), "A general class of one-sample nonparametric test statistics based on subsamples," *JASA*, 76, 960-966. UD

[22] Vandenbroucke, J. P. (2004), "When are observational studies as credible as randomized experiments?" *Lancet* 363, 1728-31. NE

Notes: BT=basis for talk, CE=causal effects, DS=design sensitivity, M=matching, NE=natural experiments, SA=sensitivity analysis, UD=uncommon, dramatic effects.

Department of Statistics, University of Pennsylvania, Philadelphia, PA 19104-6430 USA
E-mail address: rosenbaum@wharton.upenn.edu
URL: http://www-stat.wharton.upenn.edu/~rosenbap/index.html