Paul R. Rosenbaum

Design of Observational Studies

August 5, 2009

Springer

Précis of Design of Observational Studies

by Paul R. Rosenbaum

This précis contains errata, preface, table of contents, the two-page summary chapter, the list of solutions to common problems, and finally a corrected page 78 as discussed in the errata.

A sample chapter is available on the publisher's web page, www.springer.com.

Errata: Design of Observational Studies

Page 78. In expressions (3.14), (3.15) and (3.18), replace each π_j by the corresponding odds $\omega_j = \pi_j / (1 - \pi_j)$. Thanks to Keith Goldfeld.

Typos

Page 50. Bottom of page, second sentence of the new section. This should read: "have to be part of the story." Thanks to Jose Zubizarreta.

Page 52. On line 8, "the collection is $\mathcal{K}$ contains" should read "the collection $\mathcal{K}$ contains" Thanks to Sue Marcus.

Preface

An observational study is an empiric investigation of effects caused by treatments when randomized experimentation is unethical or infeasible. The quality and strength of evidence provided by an observational study is determined largely by its design. Excellent methods of analysis will not salvage a poorly designed study.

The line between design and analysis is easier to draw in practice than it is in theory. In practice, the design of an observational study consists of all activities that precede the examination or use of those outcome measures that will be the basis for the study's conclusions. Unlike experiments, in some observational studies, the outcomes may exist as measurements prior to the design of the study; it is their examination and use, not their existence, that separates design from analysis. Aspects of design include the framing of scientific questions to permit empirical investigation, the choice of a context in which to conduct the investigation, decisions about what data to collect, where and how to collect it, matching to remove bias from measured covariates, strategies and tactics to limit uncertainty caused by covariates not measured, and sample splitting to guide design using individuals who will not be included in the final analysis. In practice, design ends and analysis begins when outcomes are examined for individuals who will be the basis of the study's conclusions. An observational study that begins by examining outcomes is a formless, undisciplined investigation that lacks design.

In theory, design anticipates analysis. Analysis is ever present in design, as any goal is ever present in any organized effort, as a goal is necessary to organize effort. One seeks to ask questions and collect data so that results will be decisive when analyzed. To end well, how should we begin?

Philadelphia, PA *Paul Rosenbaum*
5 August 2009

Contents

Summary: Key Elements of Design

In an observational study, competing theories should make conflicting predictions. Many studies dissipate before they begin from a simple lack of focus. The study is intended to settle something, or at least take a step towards settling something, and for that, there needs to be something definite to settle and some prospect of and means for settling it; see Chapter 4. Some pairs of theories rarely yield conflicting predictions, so one may search for unusual opportunities to contrast them; see §5.1.

An observational study should be structured to resemble a simple experiment. A typical structure is the comparison of a treated and a control group that looked comparable prior to treatment in terms of observed covariates. An experiment is an unusual situation. In a carefully controlled experiment, one of the rarest of things happens: the effects caused by treatments are seen with clarity. Each step away from the experimental template is a step closer to the edge of an abyss; see §1.2 and §12.1.

Adjustments for observed covariates should be simple, transparent, and convincing. The major source of uncertainty about the conclusions of an observational study comes from the possible failure to control for covariates that were not measured. This possibility is raised in virtually every observational study. If you think this possibility will not be raised in evaluating your study, then you are kidding yourself. There is little hope of addressing this major source of uncertainty if the study becomes bogged down in unnecessarily complex, obscure, or unconvincing adjustments for observed covariates. One simple, transparent, and convincing way to adjust for observed covariates is to compare a treated and a control group with similar distributions of the observed covariates. Such a control group may often be constructed with the aid of multivariate matching; see Part II.

The most plausible alternatives to a treatment effect should be anticipated and addressed. In an observational study, it is not possible to address every conceivable alternative to a treatment effect. It is often possible to anticipate several plausible objections to a claim that a comparison of matched treated and control groups estimates the effects of the treatment. Such an objection claims that the comparison

is ambiguous, that it could estimate a treatment effect or it could be distorted by some specific form of bias. With a specific form of bias in mind, design elements can often be added, such as two control groups or unaffected outcomes, that resolve specific ambiguities; see §5.2.

The analysis should address possible biases from unmeasured covariates. Typically, the analysis should include a sensitivity analysis of one form or another; see Chapter 3. A sensitivity analysis asks: how much bias from unmeasured covariates – what magnitude of deviation from random assignment – would need to be present to qualitatively alter the conclusions suggested by the naïve, straightforward comparison of matched treated and control groups? The degree of sensitivity to unmeasured bias is a fact of the matter, something that is determined without ambiguity from the data at hand. Whether or not biases of this magnitude are present remains a matter of reasoned conjecture and responsible debate, but that debate is now informed and constrained by the facts of the matter. A *P*-value does not rule out the possibility that bad luck produced the observed results; rather, it objectively measures how much bad luck would be required to produce the observed results. A sensitivity analysis does not rule out the possibility that unmeasured bias produced the observed results; rather, it objectively measures how much unmeasured bias would be required to produce the observed results.

To the extent possible, observational studies should be designed to be insensitive to biases from unmeasured covariates. To do this, one must know what makes some designs (or data generating processes) sensitive to unmeasured biases and others insensitive. With this knowledge, when faced with a choice, an insensitive design may be chosen. Many factors strongly affect design sensitivity; see Part III.

There should be a plan for a primary analysis. A randomized controlled trial invariably has a protocol detailing a plan for a primary analysis. An observational study also needs such a plan. With design elements intended to resolve otherwise ambiguous comparisons, the study design anticipates one of a few patterns of results, what R.A. Fisher called an 'elaborate theory'; see §19.2. The typical elaborate theory predicts a difference here, near equivalence there, the absence of a trend here, a discontinuity there. For instance, an elaborate theory might predict much higher responses in the treated group than in two control groups, with the two control groups differing negligibly from each other. A planned primary analysis will attempt to examine, possibly confirm, the predictions of an elaborate theory; see Chapter 19. The predictions of an elaborate theory are predictions only if they precede examination of the outcomes; there is no value in an elaborate theory constructed after the fact to accommodate a particular set of data. Therefore, the analysis of an elaborate theory must be a planned analysis. Planning may be aided by sample splitting; see §18.1. A plan for a primary analysis does not preclude unplanned exploratory analyses; rather, it distinguishes planned and unplanned analyses.

Solutions to Common Problems

The matching problem is too large. If a matching problem is too large, divide it into several smaller problems by exact matching for one or more important covariates; see §9.3.

Treated and control groups are too far apart to match. Before trying to solve this problem, make sure the problem is real. Compare boxplots of the propensity scores for treated and control groups. If the boxplots cover much the same range, but some covariates or the propensity score are poorly balanced, consider: (i) tightening the caliper on the propensity score so that, for example, the penalty engages at 10% of the standard deviation rather than 20% (see §8.4), (ii) use penalties to improve balance on one or two stubborn covariates (see §9.2), or (iii) try full matching (see §8.6). Otherwise, if the boxplots exhibit large regions with little or no overlap, consider redefining the study population using a few key covariates (see §3.6). For an example of redefining the study population, see [2].

Treated and control groups overlap, but at some values of x there are too few controls even for 1-to-1 matching. Try full matching; see §8.6.

People are treated at different times. How do I match? Consider risk-set matching; see Chapter 12.

I want to match for a variable with many categories, but there are not enough people in the categories to permit a close match. Try matching with fine balance; see Chapter 10.

I have two control groups. How do I match? There are several options. One approach is to form matched triples by matching one treated group twice, once to one control group, then to the other. Another approach forms matched pairs in an 'incomplete block' design. See Chapter 11. A different problem is to split one control group to form two for a specific purpose; see the discussion of tapered matching in §18.2.

My propensity score model does a poor job of predicting treatment assignment. Not a problem. In a large, completely randomized experiment, a propensity score

model would have great difficulty predicting treatment assignments from covariates precisely because treatment assignment is random and does not depend upon the covariates. The propensity score is intended to fix a specific problem, namely imbalances in observed covariates. If your study does not have that problem, then that is just fine.

How do I judge whether my model for the propensity score is a good model? The propensity score has various uses, and the answer to this question depends upon how the propensity score will be used. In this book, propensity scores are used for matching. When used for matching, propensity scores are a means to an end, namely matched pairs or sets that balance observed covariates. When you have matched pairs or sets that balance observed covariates, the matching for observed covariates is done, and attention shifts to potential bias from unmeasured covariates. In light of this, judging the propensity score model when used in matching is essentially the same task as judging whether the matching has balanced observed covariates; see §9.1.

How do I select covariates to use in the propensity score? This question inverts the means and the end. The proper question is: Which covariates do you wish to balance by matching on the propensity score?

There is a covariate that strongly predicts treatment assignment Z but seems unlikely to matter much for the response R. What should I do? Read about 'seemingly innocuous confounding,' analytical adjustments, and tapered matching in §18.2.

There is a variable that is subsequent to treatment assignment and may have been affected by the treatment, so it is not a covariate, but I feel I should adjust for it anyway. Although this is sometimes reasonable, think long and hard before you do this. If you adjust for a concomitant variable that has been affected by the treatment, you may introduce a bias that would not otherwise have been present; see [5]. If in doubt, it may be best to leave such a variable unmatched, so that analyses with and without analytical adjustments for it are possible; see §18.2. One alternative is to both match and not match for such a variable; see the discussion of tapered matching in §18.2.

I have matched with a variable ratio of controls to treated subjects, and now I want a boxplot. It is not difficult to do, but at the time I write this, current software does not do it, so a few steps are needed. Essentially, you need to compute a weighted empirical distribution function, compute the quantiles from it, and make a boxplot using those quantiles. The R function 'bxp' will help: it will draw boxplots from quantiles you give it. See Note 3 in Chapter 8. The resulting boxplots look like ordinary boxplots, but they are weighted to reflect the variable numbers of controls. Examples are found in [2].

Should I match with a fixed or a variable ratio of controls to treated subjects? The choice is discussed in §8.5. The previous solution, about the boxplot, illustrates the biggest disadvantage of matching with variable controls; simple, straightforward

tasks require special programming effort. Opposed to this, theory strongly suggests that matching with variable controls is an efficient way to produce closer matches; see [4].

People say that a specific unobserved covariate u is strongly related to the outcome, r_C, but I doubt it. To show that a specific unobserved covariate u is not strongly related to the outcome r_C, find two control groups that differ markedly in terms of the unobserved u, and show that outcomes do not greatly differ in these two control groups. See §5.2.2 and §19.3.

People say that a specific unobserved covariate u is strongly related to treatment assignment Z, but I doubt it. To show that a specific unobserved covariate u is not strongly related to treatment assignment, Z, find an outcome known to be unaffected by the treatment that is highly correlated with u, and show that this unaffected outcome has a similar distribution in treated ($Z = 1$) and control ($Z = 0$) groups. See §5.2.4.

I would like to perform a sensitivity analysis, but I do not have matched pairs. It's not difficult. See [6, Chapter 4]. A quick and easy approach for matching with multiple controls uses the stratified Wilcoxon rank sum statistic and Table 1 in [1]. For matched sets with one to four controls (n_i equal to two to five in that table), and for Γ equal to one to four, the Table gives the calculations needed for one matched set; these contributions are summed to produce the sensitivity analysis. For any rank statistic, say the aligned rank statistic, the general calculations in [1, §3] are easy to perform in `R` or with a spreadsheet. An alternative approach with multiple controls uses an m-test [7]. Also, [7, §5] discusses sensitivity analysis with covariance adjustment in matching with multiple controls.

I would like to perform a sensitivity analysis, but my outcomes are binary. Again, it is not difficult. See [6, Chapter 4]. That chapter also discusses outcomes that are censored survival times.

How do I interpret the parameter Γ? The parameter Γ is convenient in that it is a single parameter that can refer to a wide variety of situations. The resulting sensitivity analysis is one-dimensional: just one parameter varies. The sensitivity bounds implicitly refer to an unobserved covariate u that is strongly related to the response, r_C, and that has a controlled relationship with the treatment, Z, the control being provided by the value of Γ. Sometimes, that situation is not the one under discussion, because a very strong relationship between u and r_C isn't plausible. It is possible to reexpress the one-parameter analysis in terms of two parameters, where one parameter controls the relationship between the unobserved covariate u and the response r_C and the other controls the relationship between u and the treatment, Z. It is the same one-dimensional sensitivity analysis, with no new computations, but now with a two-dimensional interpretation; see [8].

I am worried that my findings will be sensitive to small unmeasured biases. Review Chapters 14–17 to see what issues affect the sensitivity of conclusions to

unmeasured biases. Consider sample splitting to guide design decisions that affect design sensitivity; see §18.1 and [3].

I have doses of treatment, but I am not sure whether they are any good. Split the sample and find out; see §18.1 and [3]. Similar advice applies to many other design decisions.

The treatment that interests me is known to be assigned in a very biased fashion. Instead of studying the effect of that treatment, consider studying its differential effect compared with other treatments affected by the same biases; see §5.2.6. This may help in some circumstances.

There have been several studies of the treatment that interests me, but none is convincing. Is it possible to study the treatment again, this time removing one of the several problems that made previous studies unconvincing? See §4.5.

The same problems inevitably occur whenever efforts are made to study the treatment that interests me. Why is the treatment thought to work? What reasons are given? Why is there doubt about the effects of the treatment? What reasons are given? Can an empirical study shed light on whether these reasons are valid? Perhaps the reasons can be supported or refuted, even though direct investigation of the effects is difficult. See §4.6.

I am disappointed by my 'null result.' Do you possess evidence for the absence of a large effect? That might be an important finding; see §19.5. Or do you lack evidence about the magnitude of the effect? That is disappointing, but it is a common occurrence, one that happens now and then to everyone.

References

1. Gastwirth, J.L., Krieger, A.M., Rosenbaum, P.R.: Asymptotic separability in sensitivity analysis. J Roy Statist Soc B **62**, 545–555 (2000)
2. Haviland, A.M., Nagin, D.S., Rosenbaum, P.R.: Combining propensity score matching and group-based trajectory analysis in an observational study. Psychol Methods **12**, 247–267 (2007)
3. Heller, R., Rosenbaum, P.R., Small, D.: Split samples and design sensitivity in observational studies. J Am Statist Assoc **104**, to appear (2009)
4. Ming, K., Rosenbaum, P.R.: Substantial gains in bias reduction from matching with a variable number of controls. Biometrics **56**, 118–124 (2000)
5. Rosenbaum, P.R.: The consequences of adjustment for a concomitant variable that has been affected by the treatment. J Roy Statist Soc A **147**, 656–666 (1984)
6. Rosenbaum, P.R.: Observational Studies (2nd ed.). New York: Springer (2002)
7. Rosenbaum, P.R.: Sensitivity analysis for m-estimates, tests, and confidence intervals in matched observational studies. Biometrics **63**, 456–464 (2007)
8. Rosenbaum, P.R., Silber, J.H.: Amplification of sensitivity analysis in observational studies. J Am Statist Assoc, to appear

Sensitivity analysis model when pairs are matched for observed covariates

The sensitivity analysis model (3.13) is quite general in its applicability [85, Chapter 4], but here its implications for matched pairs are developed [74]. Suppose that two subjects, k and ℓ, with the same observed covariates, $\mathbf{x}_k = \mathbf{x}_\ell$, are paired, with precisely the additional fact that one of them is treated and the other control, $Z_k + Z_\ell = 1$. Then in the representation (3.1), the chance that k is treated and ℓ is control is $\Pr(Z_k = 1, Z_\ell = 0 \,|\, r_{Tk}, r_{Ck}, \mathbf{x}_k, u_k, r_{T\ell}, r_{C\ell}, \mathbf{x}_\ell, u_\ell, Z_k + Z_\ell = 1)$

$$= \frac{\pi_k(1-\pi_\ell)}{\pi_k(1-\pi_\ell)+\pi_\ell(1-\pi_k)}. \tag{3.14}$$

If in addition the sensitivity model (3.13) were true in (3.1), then simple algebra yields

$$\frac{1}{1+\Gamma} \le \frac{\pi_k(1-\pi_\ell)}{\pi_k(1-\pi_\ell)+\pi_\ell(1-\pi_k)} \le \frac{\Gamma}{1+\Gamma}. \tag{3.15}$$

In words, the condition (3.13) becomes a new condition (3.15) on paired individuals where one is treated and the other control, $Z_k + Z_\ell = 1$. If $\Gamma = 1$, then all three terms in (3.15) equal $\frac{1}{2}$, as in the randomized experiment in Chapter 2. As $\Gamma \to \infty$, the lower bound in (3.13) tends to zero and the upper bound tends to one.

Instead of pairing just two individuals, k and ℓ, suppose we pair $2I$ distinct individuals of the L individuals in the population in just this way, insisting that within each pair the two subjects have the same observed covariates and different treatments. Renumber these paired subjects into I pairs of two subjects, $i = 1, 2, \ldots, I$, $j = 1, 2$, so $\mathbf{x}_{i1} = \mathbf{x}_{i2}$, $Z_{i1} = 1 - Z_{i2}$ in each of the I pairs.[14] If (3.1) and (3.13) are true, then the distribution of treatment assignments in the I pairs satisfies

$$Z_{i1}, i = 1, \ldots, I \text{ are mutually independent,} \tag{3.16}$$

$$Z_{i2} = 1 - Z_{i1}, i = 1, \ldots, I, \tag{3.17}$$

$$\frac{1}{1+\Gamma} \le \frac{\pi_k(1-\pi_\ell)}{\pi_k(1-\pi_\ell)+\pi_\ell(1-\pi_k)} \le \frac{\Gamma}{1+\Gamma}, i = 1, \ldots, I. \tag{3.18}$$

---

insist that $\pi_\ell = \Pr(Z_\ell = 1 \,|\, \mathbf{x}_\ell, u_\ell)$. Conversely, if (3.1) and (3.13) were true as they stand, then there is an unobserved covariate $\widetilde{u}_\ell$ such that (3.1) and (3.13) are true with $\pi_\ell = \Pr(Z_\ell = 1 \,|\, \mathbf{x}_\ell, \widetilde{u}_\ell)$; simply take $\widetilde{u}_\ell = \pi_\ell = \Pr(Z_\ell = 1 \,|\, r_{T\ell}, r_{C\ell}, \mathbf{x}_\ell, u_\ell)$.

[14] In a fussy technical sense, the numbering of pairs and people within pairs is supposed to convey nothing about these people, except that they were eligible to be paired, that is, they have the same observed covariates, different treatments, with $2I$ distinct people. Information about people is supposed to be recorded in variables that describe them, such as Z, $\mathbf{x}$, u, r_T, r_C, not in their position in the data set. You can't put your brother-in-law in the last pair just because of that remark he made last Thanksgiving; you have to code him in an explicit brother-in-law variable. Obviously, it is easy to make up subscripts that meet this fussy requirement: number the pairs at random, then number the people in a pair at random. The fussy technical point is that, in going from the L people in (3.1) to the $2I$ paired people, no information has been added and tucked away into the subject numbers — the criteria for pairs are precisely $\mathbf{x}_{i1} = \mathbf{x}_{i2}$, $Z_{i1} + Z_{i2} = 1$ with $2I$ distinct individuals.