SMITH, H. L. (1997). Matching with multiple controls to estimate treatment effects in observational studies. *Sociological Methodology* **27** 325–353.

SOMMER, A. and ZEGER, S. L. (1991). On estimating efficacy from clinical trials. *Statistics in Medicine* **10** 45–52.

URY, H. (1975). Efficiency of case-control studies with multiple controls per case: Continuous or dichotomous data. *Biometrics* **31** 643–649.

VENABLES, W. N. and RIPLEY, B. D. (1994). *Modern Applied Statistics with S-Plus*. Springer, New York.

WELCH, B. L. (1937). On the z-test in randomized blocks and Latin squares. *Biometrika* **29** 21–52.

WILCOXON, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin* **1** 80–83.

WILK, M. B. (1955). The randomization analysis of a generalized randomized block design. *Biometrika* **42** 70–79.

ZHAO, C., VODICKA, P., SRAM, R. J. and HEMMINKI, K. (2000). Human DNA adducts of 1,3-butadiene, an important environmental carcinogen. *Carcinogenesis* **21** 107–111.

Comment

J. Angrist and G. Imbens

1. INTRODUCTION

Paul Rosenbaum has been an articulate and tireless advocate of randomization inference (RI) as a “reasoned basis for inference” when assessing treatment effects. In this paper and previous work he has extended the scope for RI beyond the traditional field of randomized trials into the much messier world of observational studies. The current paper provides a characteristically lucid discussion of the use of RI in observational studies, where the possibility of overt biases commonly motivates covariance adjustment. The paper discusses an approach based on propensity-score style conditioning on sufficient statistics, incorporates regression adjustment into an RI framework and offers an extension to research designs involving instrumental variables (IV). An especially interesting feature of his discussion of IV is the link to the recent literature on weak instruments, where standard inference based on normal approximations to sampling distributions is often inaccurate. Rosenbaum also discusses the use of sensitivity analyses.

Although the intellectual case for RI is attractive, model-based population inference remains the method of choice in our field of economics and in many fields involving the analysis of social statistics. In particular, regression is an enduring empirical workhorse. At the same time, recent years have seen a number of steps toward a more agnostic use of regression models as fitting devices that summarize causal relationships without being assumed to accurately represent functional relationships. We argue that the *conceptual* gap between the use of regression for RI and the use of regression with population inference has largely been closed. On the other hand, practical issues, such as the accuracy of confidence interval coverage using asymptotic arguments in finite samples, are unresolved. We hope that the current paper will stimulate additional research comparing the operational characteristics of RI with the characteristics of other methods. The purpose of this comment is to point out links to related work by economists and to highlight areas where the RI/population-model distinction seems to us to be sharpest.

2. AGNOSTIC REGRESSION

A compelling conceptual feature of RI is that it is closely tied to the notion of a randomized experiment. A primary virtue of experiments is their simplicity and transparency. In principle, with a randomized trial, no adjustments are required: with a large enough sample, the estimated treatment effects will be invariant to the selection of variables used for adjustment and to the method used to implement the adjustment. In practice, however, randomization may leave chance imbalances, and experiments are typically analyzed with some kind of regression adjustment or matching strategy to control for covariates. Moreover, in observational studies, where treatment assignment is almost always confounded with covariates, adjustment is essential.

If treatment is indeed confounded with covariates, the most important research design issue is whether the

Joshua Angrist is Professor, Department of Economics, Massachusetts Institute of Technology, Cambridge, Massachusetts (e-mail: angrist@mit.edu). Guido Imbens is Professor, Department of Economics, University of California, Berkeley, California (e-mail: imbens@econ.berkeley.edu).

covariate information at hand is adequate to remove bias. This is a question Rosenbaum has addressed in his extensive work on sensitivity analysis. Once covariates have been selected, however, a number of implementation options are available. These include matching, regression and matching on the propensity score. In Section 2.3 of the paper, Rosenbaum suggested covariate adjustment be implemented by using regression to provide an "algorithmic fit." He implicitly contrasts this "model-free" use of regression with earlier papers cited in his outline (Section 1.2), where distribution-free methods are applied to regression models based on a more literal view.

The first point we would like to make is that adoption of an agnostic view of regression is not central to the distinction between RI and population models. An agnostic view of regression is appropriate for any mode of inference. This is illustrated in Angrist (1998), which is concerned with estimating the effects of military service on the post-service civilian earnings of volunteer soldiers. For any military applicant observed after application, define random variables to represent what the applicant would earn had he served in the military and what the applicant would earn had he not served in the military. Denote these two potential outcomes by $\mathbf{Y}_0$ and $\mathbf{Y}_1$ and denote veteran status by a dummy variable $\mathbf{D}$. Treatment assignment is assumed to be ignorable conditional on a covariate vector $\mathbf{X}$, which summarizes the criteria used by the military to select soldiers from the pool of applicants. Angrist (1998) computed treatment effects using the regression of $\mathbf{Y}$ on a saturated model for $\mathbf{X}$ and the treatment dummy $\mathbf{D}$,

$$\mathbf{Y} = \alpha + \beta_x + \delta_r \mathbf{D} + \varepsilon, \tag{1}$$

where β_x is a main effect for each possible value taken on by the discrete covariate vector $\mathbf{X}$ and ε is an error term defined as the difference between $\mathbf{Y}$ and the population regression of $\mathbf{Y}$ on $\mathbf{X}$ and $\mathbf{D}$. The population regression coefficient δ_r can be written

$$\delta_r = E\{(\mathbf{D} - E[\mathbf{D}|\mathbf{X}])\mathbf{Y}\}/E\{(\mathbf{D} - E[\mathbf{D}|\mathbf{X}])\mathbf{D}\},$$

which in turn can be shown to be

$$E\{E[\mathbf{Y}_1 - \mathbf{Y}_0|\mathbf{X}]w(\mathbf{X})\},$$

where

$$w(\mathbf{X}) = \frac{P[\mathbf{D}=1|\mathbf{X}](1 - P[\mathbf{D}=1|\mathbf{X}])}{E\{P[\mathbf{D}=1|X](1 - P[\mathbf{D}=1|\mathbf{X}])\}}.$$

Thus, the population regression coefficient and its sample analog provide a weighted average of the covariate-specific treatment effects, $E[\mathbf{Y}_1 - \mathbf{Y}_0|\mathbf{X}]$, with weights given by the conditional variance of treatment in each covariate cell. The regression equation (1) plays the role of a computational device in the spirit of Rosenbaum's "algorithmic fit." In particular, the conditional expectation function $E[\mathbf{Y}|\mathbf{D}, \mathbf{X}]$ is not restricted to be linear and the individual treatment effects are not restricted to be constant. Note also that there is no extrapolation in this saturated example. In other words, values of $\mathbf{X}$ where the probability of treatment is 0 or 1 do not figure in the estimand.

The previous example uses the discreteness of covariates to provide a simple agnostic interpretation of regression estimates. More generally, however, it is common in many applications to view regression as providing the best linear approximation to an unrestricted conditional mean function (see, e.g., Chamberlain, 1984, or Goldberger, 1991), as providing an average derivative (Angrist and Krueger, 1999) or as an average arc slope (Yitzhaki, 1996).

We can make a similar point with reference to the Hodges and Lehmann (1962, 1963) model discussed at the end of Rosenbaum's Section 7. An important special case of the Hodges–Lehmann estimation strategy Rosenbaum describes, and one likely to have special appeal for practitioners, amounts to estimating a regression with treatment status and a full set of match-set indicators on the right hand side. In this case, regression estimates a weighted average of set-specific treatment effects, with each effect weighted by the conditional variance of treatment in the match set. Thus, regression provides a natural summary statistic for causal relationships. In our view, this statistic has much to recommend it (computational simplicity and efficiency for constant effects) and is easily compared to previous research results using regression. Again, however, there is no need to take the regression model literally, although auxiliary assumptions such as random sampling and linearity may matter for inference.

3. INFERENCE PROBLEMS

As the above discussion suggests, we do not see a sharp distinction between the use of regression in the manner described by Rosenbaum and the application of this tool in much modern empirical work. Still a choice remains: as Rosenbaum shows, inference with reference to a population agnostic regression function of the type described above can be carried out in a RI framework instead of using traditional population models. In our view, the question of whether RI provides substantially more accurate inference is at the heart of the RI/population-model trade-off. The right standards for making this choice seem to us to be the usual ones for alternative statistical procedures, the accuracy of nominal significance levels and statistical power in the scenario of interest.

With independent data and using sample sizes common in the cross-section empirical studies we are familiar with, it seems very likely that normal approximations to sampling distributions are acceptably accurate. In such cases, RI may be conceptually appealing, but will generate inferences that differ little in practice from population models. Of course, if outcome distributions are particularly skewed or if sample sizes are unusually small, there are likely to be some differences and RI may well be more accurate, at least under the simple null hypothesis of no effect.

An especially fruitful field for the application of RI seems likely to be cross-sectional settings with dependent data such as a group-randomized trial (GRT). Here, the need to estimate correlation structures makes inference challenging. A similarly important setting in economics, where GRT's are still rare, is the estimation of treatment effects for treatments that vary at a group level such as a city or state, with the analysis using data on microunits such as individuals or firms. The Card and Krueger study Rosenbaum discusses is one such application. The standard population model for inference in such cases implicitly uses a "design effect" to adjust standard errors for dependence within groups (Moulton, 1986), but these models are restrictive, imposing an equicorrelated structure that may not be accurate. Modern variations on the design effect approach, such as Liang and Zeger's (1986) generalized estimating equations, base inference on an asymptotic argument that requires a large number of groups for accuracy. In many such studies, there are only a few groups. Randomization inference sidesteps the need to estimate the dependence structure and appears to have good operating characteristics even in settings with few groups (for recent evidence on this point in GRTs, see Braun and Feng, 2001; Bertrand, Duflo and Mullainathan, 2001, similarly assess the accuracy of RI for state-level interventions).

4. SENSITIVITY ANALYSES

In a series of papers, Rosenbaum has developed an approach to sensitivity analyses for observational studies. Even after adjusting for overt biases, researchers remain unsure as to whether there are hidden biases. In some cases additional information such as instrumental variables may reduce the likelihood of hidden bias. In many cases, however, there are no plausible instruments. Sensitivity analysis is an approach to investigating the robustness of inferences in such settings. In the framework Rosenbaum has developed, a single parameter, Γ, captures the effect of hidden biases. The parameter Γ summarizes the degree to which the assignment mechanism is assumed to deviate from an experiment where treatment status and potential outcomes are independent. This type of sensitivity analysis is rare in economics and should be more widely used.

Two related procedures for sensitivity analysis that have gotten some attention from economists are the use of bounds and the exploration of sensitivity to observed covariates. Manski (1990) suggested an approach based on bounding the range of treatment effects consistent with the data, while imposing few assumptions beyond restrictions on the support of random variables such as 0–1 and discreteness. In some cases, these bounds can be derived by taking Rosenbaum-style sensitivity analyses to extremes. In other words, by varying the sensitivity parameter over the whole real line, one can obtain the range of values of the parameter of interest that is consistent with the observed data. A second form of sensitivity analysis works as follows. Estimate treatment effects using all available covariates and then explore the impact of omitting covariates one at a time or of dropping specific subsets (see, e.g., Altonji, Elder and Taber, 2000). Invariance to the set of control variables naturally boosts confidence in a causal interpretation of the estimated effects. This approach can be fitted into Rosenbaum's framework by using the correlation between observed covariates and outcomes to calibrate the sensitivity parameter Γ.

5. EXTENSION TO INSTRUMENTAL VARIABLES

A particularly interesting application of Rosenbaum's approach to RI arises in instrumental variables settings. Instrumentals variables methods were originally developed for the estimation of simultaneous equations models by Wright (1928) and Haavelmo (1944), but are increasingly used to solve the problem of hidden bias that has been at the center of Rosenbaum's work (see, Angrist and Krueger, 1999, for examples).

The key assumption in such applications is that the instrumental variables are not correlated with hidden sources of bias and that they affect the outcome solely through their effect on the treatment of interest. A leading example is that of randomized experiments with one-sided noncompliance. Assuming that individuals who do not take the treatment despite being assigned to it are not affected by their assignment, then random assignment to treatment is an instrumental variable for the effect of treatment on the outcome.

In econometric studies, inference with instrumental variables is typically based on large-sample approximations to the sampling distribution derived from a population model. Simple IV estimands are given by the ratio of two differences, with the denominator equal to the difference in average exposure to the treatment by assignment. The normal approximation can be poor when the difference in average exposure by treatment assignment in the denominator is small, that is, when noncompliance is high. In addition, the standard asymptotic approximation can be highly misleading when a single coefficient is estimated with many instrumental variables using two-stage least squares (a procedure for combining alternative instruments to produce a single estimate; see, e.g., Bound, Jaeger and Baker, 1995).

A number of alternatives to standard asymptotic arguments have been proposed for models with weak instruments and/or many instruments. Bekker (1994) suggested asymptotic approximations based on an alternative parameter sequence with the number of instruments increasing with the sample size, and Chamberlain and Imbens (1996) discussed Bayesian methods using hierarchical models for this case. Staiger and Stock (1997) discussed asymptotic approximations based on a correlation between the instruments and the treatment that vanishes as the sample size increases. Rosenbaum's work provides a new and elegant approach to the weak/many instruments problem. His approach leads to exact confidence intervals based on RI, regardless of the number or power of the instruments. In fact, in related work, Imbens and Rosenbaum (2001) showed that RI is the only way to obtain exact confidence intervals for IV estimates.

Finally, at the end of Section 6.3, Rosenbaum suggests an important check for IV coherence or what econometricians would call a specification check. Rosenbaum notes that instruments that have a strong association with outcomes, but a weak or nonexistent association with the causal variable of interest (the "endogenous regressor" in econometric parlance) cannot possibly satisfy the assumptions motivating IV estimation in the first place. Such simple coherence checks should be a routine part of IV analyses. We should also note, however, that in Rosenbaum's RI setup, this scenario may be manifested by empty confidence intervals. Although empty confidence intervals may not be unwelcome when the model is misspecified, a less attractive implication is that when confidence intervals are narrow, one cannot distinguish the possibility that the inferences regarding the effect of interest are precise from the possibility that the underlying model is not compatible with the data.

6. CONCLUSION

Rosenbaum argues persuasively for RI as a conceptual framework and a practical tool. He has shown here and in other work that the scope for RI is much wider than previously noted and extends to observational studies with overt and hidden biases. He has suggested specific methods for implementing these ideas that make them readily applicable. We look forward to seeing more applications of these methods in economics and further discussion and evidence on the relative merits of RI and strategies based on population inference. At a minimum, the use and exploration of such methods promotes recognition of the value of an approach to observational studies that uses the language and methods of the randomized trial as a guiding principle.

Comment

Jennifer Hill

1. INTRODUCTION

Paul Rosenbaum has contributed an extremely helpful paper that consolidates nearly two decades of research on a class of nonparametric approaches to causal inference in the context of observational studies. Rosenbaum first reminds the reader of the use of permutation tests with data from randomized experiments, and then he presents and justifies extensions for application to observational study data. This presentation elucidates the similarities with and differences

Jennifer Hill is Professor, Columbia University School of Social Work, New York, New York 10025 (e-mail: jh1030@columbia.edu).