Statistical Science
2002, Vol. 17, No. 3, 286–327

Covariance Adjustment in Randomized Experiments and Observational Studies

Paul R. Rosenbaum

Abstract. By slightly reframing the concept of covariance adjustment in randomized experiments, a method of exact permutation inference is derived that is entirely free of distributional assumptions and uses the random assignment of treatments as the "reasoned basis for inference." This method of exact permutation inference may be used with many forms of covariance adjustment, including robust regression and locally weighted smoothers. The method is then generalized to observational studies where treatments were not randomly assigned, so that sensitivity to hidden biases must be examined. Adjustments using an instrumental variable are also discussed. The methods are illustrated using data from two observational studies.

Key words and phrases: Covariance adjustment, matching, observational studies, permutation inference, propensity score, randomization inference, sensitivity analysis.

1. RANDOMIZATION INFERENCE AND COVARIANCE ADJUSTMENT

1.1 Introduction: The Role of Randomization in Inference

Calling randomization the "reasoned basis for inference" in experiments, Fisher (1935) showed that exact inferences about the effects caused by treatments could be based solely on distributions created by the physical act of randomization, without assumptions. Since then, an extensive literature has shown that various commonly used procedures, such as Wilcoxon's (1945) rank sum test, may be viewed as randomization tests (e.g., Lehmann, 1999), and many other procedures, such as analysis of variance, may be viewed as approximations to randomization tests (e.g., Kempthorne, 1952, Section 8). Much less has been written about randomization inference for covariance adjustment—Cox (1956) is one exception—in part because of computational difficulties that once seemed insurmountable, but today look rather modest.

Paul R. Rosenbaum is Robert G. Putzel Professor, Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19104-6340 (e-mail: rosenbaum@wharton.upenn.edu).

In addition to providing a basis for exact, distribution-free inference in randomized experiments, the theory of randomization inference is helpful in clarifying the greater uncertainty that is present in observational studies of treatment effects, where treatments are not randomly assigned (Rosenbaum, 1995). It is possible to quantify the added uncertainty in observational studies only if analyses of randomized experiments are explicit about the role that randomization plays in inference. It is, of course, possible to incorporate randomization in inference in other ways; for instance, Rubin (1978) developed the important role that randomization plays in Bayesian inference.

1.2 Outline

The current paper develops a theory of randomization inference for covariance adjustment in completely randomized experiments in Section 2, extends this to observational studies free of hidden bias in Section 3 and then discusses covariance adjustment of matched pairs in observational studies in Section 4. Sensitivity to hidden bias is discussed in Section 5. The use of instrumental variables is discussed in Section 6. More complex forms of matching are discussed in Section 7.

In the case of simple rank tests, such as Wilcoxon's (1945) rank sum and signed rank tests, textbooks often present parallel discussions of randomization

inference and inferences derived from independent and identically distributed sampling of an infinite population. For instance, in his first four chapters, Lehmann (1999) discussed randomization inference and infinite population models in alternate chapters. Also, Lehmann (1986, Section 5.10, Theorem 5.6, page 231) showed that, to be distribution-free, a test must effectively be a randomization test.

There is an extensive literature on nonparametric and distribution-free methods for regression, but this literature typically uses population models rather than randomization inference. Adichie (1978) tested nonparametric hypotheses about a subset of linear regression coefficients by applying conventional nonparametric tests, such as the signed rank test, to residuals from a reduced regression model. See also Quade (1967), Koul (1970), Jureckova (1971), Jaeckel (1972), Kraft and van Eeden (1972) and McKean and Hettmansperger (1978); see Adichie (1984, Section 3) and Hajek, Sidak and Sen (1999, Section 10.1.2) for surveys of this literature. In this approach (1) it is assumed that the regression model is "correct," that is, the model generated the observed data, (2) an estimate of the reduced model coefficients is needed that has convergence at rate $\sqrt{n}$, where n is the sample size, (3) and only asymptotic results are obtained. The randomization theory of covariance adjustment is different. The reduced model is simply a fit, not a stochastic model, and it need not be "correct" in any sense—rather, it is hoped, but not needed, that the residuals from the fit are more stable than the responses themselves. An exact distribution theory is available, and neither the level of tests nor the coverage of confidence intervals requires $\sqrt{n}$ convergence, so for example, the covariance adjustment may use a smoother with a different rate of convergence. Although large sample approximations are useful in the randomization theory, the needed approximations are simply the usual, simple large sample approximations for the rank sum or signed rank statistics.

The relationship between randomization and covariance adjustment has been discussed from several perspectives. Cox (1956) discussed a form of weighted randomization that led to estimates of mean squares associated with covariance adjustment that are unbiased over the randomization distribution. Robinson (1973a) showed that conventional least squares analysis of covariance may be approximately justified by random assignment of treatments, rather than assuming linear models with random normal errors. Puri and Sen (1969) derived the randomization distribution of nonparametric analysis of covariance by conditioning in an infinite population model; this setup then forms a natural framework for asymptotic approximations to the randomization distribution. Box and Guttman (1966) and Hooper (1989) combined random errors and random assignment of treatments. Gabriel and Hall (1983) performed randomization tests with a restricted set of treatment assignments. Gail, Tan and Piantadosi (1988) discussed the randomization distribution of a statistic motivated by fitting a generalized linear model. Raz (1990) applied randomization inference to regressions using smoothers.

The use of randomization in experiments has its critics; see, for instance, Harville (1975). He argued that in laboratory experiments, where the units are transistors or cell cultures, selection biases are likely to be small and randomization should be replaced by optimal design. Whether or not that is true of laboratory experiments, in studies of human subjects in medicine, public health, economics and public policy, substantial selection biases are often plausible if not likely, and preventing bias through random assignment is a central concern.

1.3 An Example: DNA Damage from an Occupational Hazard

This section introduces the first of two examples that will be used to illustrate methods. Table 1 contains data from an observational study by Zhao, Vodicka, Sram and Hemminki (2000) of a specific alteration of human DNA possibly caused by occupational exposure to the chemical 1,3-butadiene, which is used to produce a variety of polymers. At a chemical operation in the Czech Republic, they compared 15 exposed males who worked with 1,3-butadiene to 11 male controls who worked in the heat production unit. Blood samples yielded DNA from lymphocytes and the DNA adduct N-1-(2,3,4-trihydroxybutyl)-adenine (N-1-THB-Ade) was measured in adducts per 10^9 nucleotides. There are three covariates: age, smoker and cigarettes per day.

The original study compared N-1-THB-Ade levels among exposed and control workers using Wilcoxon's rank sum test, finding significantly higher levels among exposed workers. In performing this analysis, the rank sum statistic was compared with its usual null distribution, which is the correct distribution if the treatment or exposure has no effect and subjects are randomly assigned to treatment or control. Of course, random assignment was not used here, because it would be unethical to expose workers to an environmental

TABLE 1
Human DNA adducts for workers exposed to 1,3-*butadiene and controls*

Group	Age	Smoking	Cigarettes/day	*N*-1-THB-Ade*
Exposed	57	S	15	0.3
	50	S	20	0.5
	28	S	15	1.0
	59	S	40	0.8
	23	S	20	1.0
	49	S	15	12.5
	49	S	2	0.3
	24	S	5	4.3
	45	NS	0	1.5
	48	NS	0	0.1
	38	NS	0	0.3
	44	NS	0	18.0
	43	NS	0	25.0
	44	NS	0	0.3
	57	NS	0	1.3
Control	36	S	10	0.1
	20	S	20	0.1
	31	S	10	2.3
	50	S	25	3.5
	31	NS	0	0.1
	54	NS	0	0.1
	54	NS	0	1.8
	55	NS	0	0.5
	44	NS	0	0.1
	49	NS	0	0.2
	51	NS	0	0.1

Source: Zhao et al. (2000).

* *N*-1-(2,3,4-trihydroxybutyl)-adenine in adducts per 10^9 nucleotides.

hazard as part of a controlled experiment. Moreover, 1,3-butadiene is contained in cigarette smoke, and more than half of the exposed workers were smokers, while fewer than half of the controls were smokers. Notice, for instance, that the two highest *N*-1-THB-Ade levels among controls were found among the four smokers. Three alternative strategies for adjusting for the covariates will be considered. As it turns out, the original analysis by Zhao et al. (2000) holds up well, agreeing with the adjusted analyses, so the three observed covariates cannot explain the higher levels of *N*-1-THB-Ade among exposed workers.

Common covariance-adjustment models involve additive treatment effects, but an additive effect will not adequately describe the *N*-1-THB-Ade levels in Table 1. In the current paper, following conventional practice with extremely skewed data, the *N*-1-THB-Ade levels will be transformed by taking logs, so that additive models on the log scale become multiplicative models on the original scale. Here, logs are quite successful in reducing asymmetry and, of course, they do not change rank tests of no effect. Nonetheless, logs shift the focus of attention in a way that is, perhaps, undesirable. Specifically, logs amplify the small, perhaps unimportant, variations in low *N*-1-THB-Ade levels and they subdue the large, perhaps important, variations in extremely high *N*-1-THB-Ade levels. An alternative method of analysis for data of this sort, without transformations, is discussed in Rosenbaum (1999a).

The second example, discussed in Section 4.2, will be used to illustrate additional techniques, including instrumental variables and sensitivity analyses for unobserved covariates.

2. COVARIANCE ADJUSTMENT IN RANDOMIZED EXPERIMENTS

2.1 Treatments, Responses under Alternative Treatments, Random Assignment

There are n subjects, $j = 1, \ldots, n$, and subject j has two potential responses: the response r_{Tj} that would be observed if j were assigned to treatment, and the response r_{Cj} that would be observed if j were assigned to control (Neyman, 1923; Rubin, 1974, 1977). The effect caused by giving the treatment in place of the control is a comparison of r_{Tj} and r_{Cj} such as $r_{Tj} - r_{Cj}$, but such an effect can never be calculated from observed data, because subject j receives either treatment, displaying response r_{Tj}, or control, displaying response r_{Cj}, but r_{Tj} and r_{Cj} are never jointly observed for the same subject j. In addition, subject j has a vector $\mathbf{x}_j$ of covariates describing j prior to treatment.

Of the n subjects, m are selected at random to receive the treatment; the remaining $n - m$ are assigned to control. That is, each of the $\binom{n}{m}$ possible treatment assignments has the same probability, namely $\binom{n}{m}^{-1}$. Write $Z_j = 1$ if subject j receives the treatment and $Z_j = 0$ if subject j receives the control, so that $\sum_{j=1}^{n} Z_j = m$. Notice that r_{Tj} is observed if $Z_j = 1$ and r_{Cj} is observed if $Z_j = 0$, so the observed response of subject j is $R_j = Z_j r_{Tj} + (1 - Z_j) r_{Cj}$.

In randomization inference (Fisher, 1935), the only stochastic quantities are those that involve the random assignment of treatments, Z_j, so that randomization creates all of the distributions used for inference, and randomization forms "the reasoned basis for inference" in Fisher's words. Specifically, the potential responses and covariates, $(r_{Tj}, r_{Cj}, \mathbf{x}_j)$, $j = 1, \ldots, n$, are fixed features of this finite population of n subjects. In contrast, the observed response R_j of subject j changes

with the random treatment assignment Z_j, so R_j is not fixed. In this view of inference in experiments, the exact inference is the randomization inference derived from the randomization distribution of statistical quantities. In this view, parametric distributions, such as the Normal, the t-distribution and the F-distribution, are never models for data; rather, they are approximations to randomization distributions—they are good approximations to the extent that they reproduce randomization inferences with reduced computational effort (e.g., Welch, 1937; Wilk, 1955; Cox, 1956, 1958; Kempthorne, 1955; Robinson, 1973a, b).

The treatment effect is additive if there is a constant τ, such that $r_{Tj} - r_{Cj} = \tau$ for $j = 1, \ldots, n$; in this case, control responses, r_{Cj}, vary from one subject j to another, but for every subject, the treatment raises the response by the same amount τ. Because r_{Tj} and r_{Cj} are never jointly observed, in terms of observable quantities, the additive model asserts that the distribution of treated responses r_{Tj} is shifted upward by τ when compared to the distribution of control responses r_{Cj}, and the magnitude of the shift does not vary with the covariates, a common nonparametric model (Lehmann, 1999). Analysis of covariance often assumes an additive treatment effect and often takes inference about τ as the goal. Nonadditive effects are possible and consequential in some contexts; see Section 6 for one nonadditive model and see Rosenbaum (1999a) for another.

Write $\mathbf{Z} = (Z_1 \ldots Z_n)^T$, $\mathbf{R} = (R_1 \ldots R_n)^T$, $\mathbf{r}_C = (r_{C1} \ldots r_{Cn})^T$ and so forth, and write $\mathbf{X}$ for the matrix with n rows $\mathbf{x}_j^T$, $j = 1, \ldots, n$. Recall that $\mathbf{Z}$ and $\mathbf{R}$ are observed, but $\mathbf{r}_C$ is not. Notice that if the treatment effect is additive, then the vector of adjusted responses $\mathbf{R} - \tau\mathbf{Z} = \mathbf{r}_C$ is fixed, not varying with the random treatment assignment $\mathbf{Z}$.

2.2 Randomization Inference Ignoring the Covariate

Randomization inference about, say, an additive treatment effect τ, uses the randomization distribution of a statistic $t(\mathbf{Z}, \mathbf{R} - \tau, \mathbf{Z}) = t(\mathbf{Z}, \mathbf{r}_C)$. Notice that if τ were known, the distribution of $t(\mathbf{Z}, \mathbf{R} - \tau\,\mathbf{Z}) = t(\mathbf{Z}, \mathbf{r}_C)$ would be known since $\mathbf{r}_C$ is fixed and $\mathbf{Z}$ has a known distribution created by the randomization.

For instance, a familiar statistic, $t(\mathbf{Z}, \mathbf{R} - \tau\mathbf{Z})$, is Wilcoxon's rank sum statistic in which the adjusted responses $\mathbf{R} - \tau\mathbf{Z}$ are ranked from 1 to n, with average ranks for ties, and the ranks of the treated ($Z_i = 1$) subjects are summed to yield the value of the rank sum statistic $t(\mathbf{Z}, \mathbf{R} - \tau\mathbf{Z})$. The null hypothesis $H_0 : \tau = \tau_0$ is tested by computing $t(\mathbf{Z}, \mathbf{R} - \tau_0\,\mathbf{Z})$ and asking whether it falls in the tail of the randomization distribution, which for the Wilcoxon's rank sum without ties is the distribution of the sum of m numbers randomly selected from $\{1, \ldots, n\}$. Write q_j for the rank of $R_j - \tau_0 Z_j$, so that $t(\mathbf{Z}, \mathbf{R} - \tau_0\,\mathbf{Z}) = \mathbf{q}^T\mathbf{Z}$ for the rank sum statistic, where $\mathbf{q} = (q_1 \ldots q_n)^T$. Under the null hypothesis, $H_0 : \tau = \tau_0$, $R_j - \tau_0 Z_j = r_{Cj}$ and $\mathbf{q}$ is fixed. A two-sided 95% confidence interval for τ is found by testing every value τ_0 and retaining in the interval the values not rejected by such a two-sided 0.05-level test (Lehmann, 1963; Moses, 1965). The Hodges–Lehmann (1963) point estimate of τ is found by equating the statistic $t(\mathbf{Z}, \mathbf{R} - \tau\mathbf{Z})$ to its expectation under the randomization distribution, namely $m(n+1)/2$ for Wilcoxon's rank sum, and solving for τ, with small allowance for the discreteness of the rank sum as a function of τ. For the rank sum, the Hodges–Lehmann estimate turns out to equal the median of all pairwise differences between the m observed treated responses and the $n - m$ observed control responses. See Lehmann (1998) for detailed discussion of these standard methods.

In the example in Section 1.3, the Wilcoxon's rank sum statistic for testing no effect is 242.5, allowing for ties. With the given pattern of ties, the null randomization distribution of the rank sum has expectation 202.5 and variance 362.62, yielding the standardized deviate $\frac{242.5-202.5}{\sqrt{362.62}} = 2.10$, so the null hypothesis of no treatment effect would not be plausible if these data had been observed in a randomized experiment. Testing hypotheses $H_0 : \tau = \tau_0$ on the log scale in a one-sided 0.05-level test leads to a one-sided 95% confidence interval of $\tau \geq 0.41$ or a multiplicative effect of $e^\tau \geq 1.51$ or a 51% increase. This confidence interval would be appropriate in a randomized experiment.

As is well known, appropriate randomization inferences about τ may be drawn ignoring the covariate. However, adjustment for chance imbalances in the covariate using covariance adjustment may increase the efficiency of the inference.

2.3 Using Covariates in Fitting Potential Control Responses

Although $\mathbf{r}_C$ is not observed, imagine for a moment using some algorithm that fits $\mathbf{r}_C$ using $\mathbf{X}$, yielding a vector of residuals $\mathbf{e}$. For instance, one might fit $\mathbf{r}_C$ using $\mathbf{X}$ by least squares linear regression, by robust linear regression (Huber, 1981), by rank linear regression (Jaeckel, 1972) or by using a smoother such as Lowess (Cleveland, 1979). The specific fitting algorithm used is of practical importance, but it does not affect

the logical structure of the argument presented here. Write $\widetilde{\varepsilon}(\cdot)$ for the function that creates residuals from $\mathbf{r}_C$ and $\mathbf{X}$, so $\widetilde{\varepsilon}(\mathbf{r}_C) = \mathbf{e}$, where for notational simplicity the dependence on $\mathbf{X}$ is not explicit in the notation.

Notice that there is no stochastic model here, just an algorithmic fit, because in randomization inference, $\mathbf{r}_C$ and $\mathbf{X}$ are fixed quantities that do not vary with the random treatment assignment $\mathbf{Z}$. Hence, $\widetilde{\varepsilon}(\mathbf{r}_C) = \mathbf{e}$ is a fixed vector computed from the fixed quantities $\mathbf{r}_C$ and $\mathbf{X}$, not a random variable or a by-product of estimation. If the randomization had picked a different treatment assignment $\mathbf{Z}$, yielding different observed responses $\mathbf{R}$, the quantity $\widetilde{\varepsilon}(\mathbf{r}_C) = \mathbf{e}$ is not changed. Notice also that $\widetilde{\varepsilon}(\mathbf{r}_C) = \mathbf{e}$ cannot be computed because $\mathbf{r}_C$ is not observed.

Viewed as a batch of n fixed numbers, the residuals $\mathbf{e}$ may be much more stable and less dispersed than the responses under control $\mathbf{r}_C$, because much of the variation in $\mathbf{r}_C$ may be captured by the covariates $\mathbf{X}$. Although randomization inference using the responses themselves and ignoring the covariates yields tests with the correct level and confidence intervals with the correct coverage rate, more precise inference might have been possible if the variation in fitted values had been removed. In other words, one would like to use the residuals $\mathbf{e}$ in place of the control responses $\mathbf{r}_C$ in performing the randomization test, believing $\mathbf{e}$ to be less dispersed; however, neither $\mathbf{e}$ nor $\mathbf{r}_C$ is observed.

2.4 Randomization Inference with Covariance Adjustment

Suppose that we wish to test the hypothesis $H_0 : \tau = \tau_0$ using an exact, randomization inference, but adjusting by covariance adjustment for $\mathbf{X}$. Given what has been said, the procedure is straightforward. Calculate the adjusted responses $\mathbf{R} - \tau_0\mathbf{Z}$, which equal $\mathbf{r}_C$ when the null hypothesis is true. Then compute $\widetilde{\varepsilon}(\mathbf{R} - \tau_0\mathbf{Z}) = \mathbf{e}_0$, say, which equals $\widetilde{\varepsilon}(\mathbf{r}_C) = \mathbf{e}$ when the null hypothesis is true. Under the null hypothesis, $H_0 : \tau = \tau_0$, the residuals $\mathbf{e}_0 = \mathbf{e}$ are both fixed, not varying with the treatment assignment $\mathbf{Z}$ and known. From the residuals, calculate a test statistic $t(\mathbf{Z}, \mathbf{e}_0)$ and test the null hypothesis by comparing the test statistic to its randomization distribution.

For instance, one might test $H_0 : \tau = \tau_0$ by computing the adjusted responses $\mathbf{R} - \tau_0\mathbf{Z}$, fitting the linear model to these adjusted responses, say, using a fitting algorithm yielding an m-estimate, then finding the residuals $\mathbf{e}_0$ and applying Wilcoxon's rank sum statistic $t(\mathbf{Z}, \mathbf{e}_0)$ to these residuals, so $t(\mathbf{Z}, \mathbf{e}_0)$ is the sum of the ranks of the residuals for treated subjects. If the null hypothesis is true and if the residuals $\mathbf{e}_0$ are untied, the exact randomization distribution of $t(\mathbf{Z}, \mathbf{e}_0)$ is simply the distribution of the sum of m numbers randomly selected from $\{1, \ldots, n\}$. With ties, one uses average ranks, obtaining a slightly more complex exact distribution.

Gail, Tan and Piantadosi (1988) took a similar approach to testing the hypothesis of no effect, by fitting a generalized linear model and using the randomization distribution of the associated test statistic. Residuals $\widetilde{\varepsilon}(\mathbf{R} - \tau_0\mathbf{Z})$ need not be obtained from a linear fit. Instead, the residuals might result from a smoother, such as Cleveland's (1979) Lowess. See Raz (1990) for discussion of randomization tests of no effect after smoothing.

If one uses linear least squares with a constant term to obtain residuals and if the test statistic is simply the sum of the residuals $t(\mathbf{Z}, \mathbf{e}_0) = \mathbf{Z}^T\mathbf{e}_0$, then the Hodges–Lehmann estimate is the usual least squares estimate of a covariance adjusted difference between groups. To see this, write $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}$, so that $\mathbf{e}_0 = (\mathbf{I} - \mathbf{H})(\mathbf{R} - \tau_0\mathbf{Z})$, which equals $\mathbf{e} = (\mathbf{I} - \mathbf{H})\mathbf{r}_C$ if the hypothesis $H_0 : \tau = \tau_0$ is true. If $\mathbf{X}$ contains a constant term, the mean of the fixed residuals, $\frac{1}{n}\mathbf{e}^T\mathbf{1}$, is zero. Hence, in a randomized experiment, the expectation of $\mathbf{Z}^T\mathbf{e}$ is $E(\mathbf{e}^T\mathbf{Z}) = \mathbf{e}^T E(\mathbf{Z}) = \mathbf{e}^T(\frac{m}{n}\mathbf{1}) = 0$. The Hodges–Lehmann estimate $\widehat{\tau}$, equates $t(\mathbf{Z}, \mathbf{e}_0) = \mathbf{Z}^T\mathbf{e}_0 = \mathbf{Z}^T(\mathbf{I} - \mathbf{H})(\mathbf{R} - \tau_0\mathbf{Z})$ to its expectation at the true τ (here 0) and solves $0 = \mathbf{Z}^T(\mathbf{I} - \mathbf{H})(\mathbf{R} - \widehat{\tau}\,\mathbf{Z})$ for $\widehat{\tau}$, which is $\widehat{\tau} = (\mathbf{Z}^T(\mathbf{I} - \mathbf{H})\mathbf{R})/(\mathbf{Z}^T(\mathbf{I} - \mathbf{H})\,\mathbf{Z})$, which is the usual least squares estimate (see Seber, 1977, Section 3.7).

In the examples in this paper, residuals are obtained from linear regressions fitted using Huber's (1981) m-estimates with the weight function he proposed, as implemented in S-Plus; see Venables and Ripley [1994, page 215, glm(·, family = robust)]. Consider again the example of Section 1.3. To test the null hypothesis of no effect, the logs of the N-1-THB-Ade levels are regressed on age, a binary smoking variable and cigarettes per day, and the rank sum test is applied to the residuals, yielding a rank sum of 241 with no ties. Because there are no ties, standard formulas for moments and the tabulated exact distribution may be used. The null expectation of the rank sum is 202.5 and the null variance is 371.25, yielding a standardized deviate of 1.998. If this were a randomized experiment, the null hypothesis would remain implausible even after covariance adjustment for the three covariates. The hypothesis $H_0 : \tau = \tau_0$ is tested by subtracting τ_0 from

the logs of the N-1-THB-Ade levels for treated subjects, refitting the model with these adjusted responses and applying the rank sum test to the residuals. Testing hypotheses $H_0: \tau = \tau_0$ in this way leads to a one-sided 95% confidence interval of $\tau \geq 0.29$ or a multiplicative effect of $e^{\tau} \geq 1.34$ or a 34% increase. Again, these calculations would be appropriate in a randomized experiment in which every subject has the same chance of receiving the treatment. What can be done if the chance of receiving the treatment varies with covariates?

3. COVARIANCE ADJUSTMENT IN OBSERVATIONAL STUDIES WITH AN UNKNOWN PROPENSITY SCORE

3.1 Unknown Assignment Probabilities in Observational Studies

An observational study resembles an experiment to the extent that the goal is to estimate the effects caused by a treatment by comparing treated and control units. However, in an observational study, the units are not randomly assigned to treatment groups, and the groups may not have been comparable prior to treatment (Cochran, 1965). Visible, recorded pretreatment differences are called *overt bias* and are removed by adjustments, such as matching, perhaps in combination with covariance adjustment. Unobserved pretreatment differences are called *hidden bias* and must be studied by other means, such as sensitivity analysis. The current section and Section 4 discuss adjustments for overt biases assuming hidden biases are absent, whereas Section 5 discusses sensitivity analysis for hidden biases.

In an observational study, unlike an experiment, the treatment Z_j is not randomly assigned (Cochran, 1965), so that the $\pi_j = \Pr(Z_j = 1)$ may vary with j and are unknown. The argument in Section 2 is inapplicable in this case. Suppose treatments were assigned independently with unknown π_j, so that $\Pr(\mathbf{Z} = \mathbf{z}) = \prod_{j=1}^{n} \pi_j^{z_j} (1 - \pi_j)^{1 - z_j}$. If one knew that the π_j were equal, then the conditional distribution $\Pr(\mathbf{Z} = \mathbf{z} | \sum Z_j = m)$ would equal the randomization distribution in Section 2.

The study is said to be *free of hidden bias* if the π_j, though unknown, are known to be functions of the observed covariates $\mathbf{x}_j$ alone. If instead the π_j are functions of both the observed covariates $\mathbf{x}_j$ and also relevant unobserved covariates u_j, then there is *hidden bias* due to u_j. When there is no hidden bias, adjustment for observed covariates $\mathbf{x}_j$ permits inference about treatment effects, but hidden bias must be addressed in other ways; see Section 5.

3.2 Conditional Permutation Tests

This section briefly reviews a method discussed in Rosenbaum (1984), where formal results, algorithms and examples may be found; see also Robins, Mark and Newey (1992) and Robins and Ritov (1997) for related developments. Suppose the study is free of hidden bias and, moreover, $\log(\pi_j/(1 - \pi_j)) = \lambda^T \mathbf{x}_j$, where the first coordinate of $\mathbf{x}_j$ is always 1, so that the first coordinate of λ is a constant term. Then $\mathbf{X}^T\mathbf{Z}$ is sufficient for λ (Cox, 1970) and the conditional distribution $\Pr(\mathbf{Z} = \mathbf{z} | \mathbf{X}^T\mathbf{Z})$ is a known distribution, free of the unknown parameter λ, and it gives a known exact null distribution for a test statistic $t(\mathbf{Z}, \mathbf{R} - \tau_0\mathbf{Z}) = t(\mathbf{Z}, \mathbf{r}_C)$. In a sense, this test performs a version of covariance adjustment, because on the conditional sample space with $\mathbf{X}^T\mathbf{Z}$ fixed, the least squares adjusted estimate $\widehat{\tau} = (\mathbf{Z}^T(\mathbf{I} - \mathbf{H})\mathbf{R})/(\mathbf{Z}^T(\mathbf{I} - \mathbf{H})\mathbf{Z})$ is a linear function of the unadjusted total in the treated group $\mathbf{Z}^T\mathbf{R}$.

When the test statistic has the form $t(\mathbf{Z}, \mathbf{r}_C) = \mathbf{q}^T\mathbf{Z}$, where $\mathbf{q}$ is a function of $\mathbf{r}_C$, the conditional permutation test, which rejects when $\mathbf{q}^T\mathbf{Z}$ is in the upper tail of the distribution obtained from $\Pr(\mathbf{Z} = \mathbf{z} | \mathbf{X}^T\mathbf{Z})$, is the same as the exact, uniformly most powerful unbiased test of $H_0: \theta = 0$ in the model $\log(\pi_j/(1 - \pi_j)) = \lambda^T \mathbf{x}_j + \theta q_j$. Of course, $\theta = 0$ when the hypothesis is true ($H_0: \tau = \tau_0$), because $\log(\pi_j/(1 - \pi_j)) = \lambda^T \mathbf{x}_j$ and $\mathbf{R} - \tau_0\mathbf{Z} = \mathbf{r}_C$ is constant, not varying with $\mathbf{Z}$. However, if the hypothesis is false ($\tau \neq \tau_0$), then $\mathbf{R} - \tau_0\mathbf{Z} = \mathbf{r}_C + (\tau - \tau_0)\mathbf{Z}$ and the adjusted responses $\mathbf{R} - \tau_0\mathbf{Z}$ will help to predict $\mathbf{Z}$. When the sample size is moderately large, the uniformly most powerful unbiased test of $H_0: \theta = 0$ may be replaced by one of the several more familiar and computationally simpler tests associated with maximum likelihood estimation of the logit model $\log(\pi_j/(1 - \pi_j)) = \lambda^T \mathbf{x}_j + \theta q_j$. The large sample procedures have an advantage compared to the most powerful unbiased test: they are less affected by the degree of discreteness of $\mathbf{X}$.

Consider, again, the example in Section 1.3, and assume in this section that the study is free of hidden bias and the propensity score follows a logit model, $\log(\pi_j/(1 - \pi_j)) = \lambda^T \mathbf{x}_j$. The hypothesis of no effect, $H_0: \tau = 0$, is tested by Wilcoxon's rank sum using a logit regression of exposure to the treatment Z_j, on a constant term, the three covariates in $\mathbf{x}_j$ and the ranks q_j of the N-1-THB-Ade levels. The ratio of the coefficient of q_j to its approximate standard error is 1.89, leading to rejection of the null hypothesis in a one-sided 0.05-level test. In other words, the ranks of the N-1-THB-Ade levels predict exposure

to treatment Z_j adjusting for covariates, so it is not plausible that the treatment has no effect. Similarly, the hypothesis $H_0:\tau=\tau_0$ is tested by subtracting τ_0 from the responses of treated subjects, ranking the adjusted responses $\mathbf{R}-\tau_0\mathbf{Z}$ and testing $H_0:\theta=0$ in the logit regression $\log(\pi_j/(1-\pi_j))=\lambda^T\mathbf{x}_j+\theta q_j$. Repeating this for each τ_0 gives a one-sided, 95% confidence interval of $\tau\geq 0.41$ or $e^\tau\geq 1.51$ or a 51% increase. This turns out to be nearly the same as the unadjusted estimate in Section 2.2. However, unlike the test performed in Section 2.2, the test performed in this section did not assume the π_j's are known and constant, not varying with $\mathbf{x}_j$.

3.3 Conditional Permutation Tests with Covariance Adjustment

Sections 2.4 and 3.2 offer two different methods of incorporating covariance adjustment into a randomization test such as Wilcoxon's rank sum test. Either may be used in a randomized experiment, but only the method of Section 3.2 is appropriate in an observational study free of hidden bias. The methods may be combined. Assuming the null hypothesis $H_0:\tau=\tau_0$ for the purpose of testing it, the fixed residuals $\widetilde{\varepsilon}(\mathbf{R}-\tau_0\mathbf{Z})=\widetilde{\varepsilon}(\mathbf{r}_C)$ are computed, and from them, the ranks $\mathbf{q}$, which are used in the logit model in Section 3.2 to test $H_0:\theta=0$, are computed. Under the assumptions in Section 3.2, the ranks $\mathbf{q}$ do not help to predict $\mathbf{Z}$ when $H_0:\tau=\tau_0$ is true, but will vary systematically with $\mathbf{Z}$ when the hypothesis is false ($\tau\neq\tau_0$), so that $\mathbf{R}-\tau_0\mathbf{Z}=\mathbf{r}_C+(\tau-\tau_0)\mathbf{Z}$.

In the example in Section 1.3, the hypothesis of no effect, $H_0:\tau=0$, yields a deviate of 1.92 for testing $H_0:\theta=0$ in the logit regression, so no effect is not plausible. Testing hypotheses of the form $H_0:\tau=\tau_0$ in this way yields a 95% one-sided confidence interval of $\tau\geq 0.29$ or $e^\tau\geq 1.34$ or at least a 34% increase, similar to Section 2.4. Although the four tests in Sections 2.2, 2.4 and 3.2 and the current section, produced similar results in this example, and they formed two pairs of tests with nearly identical results, this is a feature of the data in Section 1.3 and is not to be expected in general.

4. MATCHED OBSERVATIONAL STUDIES WITHOUT HIDDEN BIASES

4.1 Matching and Covariance Adjustment

In matching, treated and control subjects are paired so that they are similar before treatment on some aspects of observed covariates. One strategy for matching with desirable properties is to match on the propensity score, that is, the conditional probability of exposure to treatment given observed covariates; see Rosenbaum and Rubin (1983) for a discussion. This sort of matching will be combined with covariance adjustment in the current section, assuming the study is free of hidden bias.

In simulations, Rubin (1973, 1979) showed that covariance adjustment of matched pair differences is more robust to model misspecification than covariance adjustment alone and has greater statistical efficiency than matching alone. In a practical example, Dehejia and Wahba (1999) demonstrated the hazards of relying on models alone, without matching or stratification. Methods for constructing matched samples have been discussed by Rosenbaum and Rubin (1985), Rosenbaum (1989, 1991a), Gu and Rosenbaum (1993), Ming and Rosenbaum (2000) and Li, Propert and Rosenbaum (2001). Implementation in SAS has been discussed by Bergstralh, Kosanke and Jacobsen (1996) and Ming and Rosenbaum (2001). Matching also facilitates the incorporation of thick description into quantitative studies (Rosenbaum and Silber, 2001).

4.2 An Example: Effects of Increasing the Minimum Wage

Economic theory predicts that raising the minimum wage will depress employment. Card and Krueger (1994, 1995) examined this prediction when New Jersey raised its minimum wage by about 20% from \$4.25 to \$5.05 an hour on 1 April 1992. They looked at the change in the number of full time equivalent employees from before the wage increase to after the increase at fast food restaurants such as Burger King and Wendy's, comparing New Jersey to adjacent eastern Pennsylvania, where the minimum wage was not increased. Although starting wages increased substantially in New Jersey, when compared to Pennsylvania, there was a negligible change in the number of employees.

From their data, 66 pairs of restaurants, one from New Jersey, the other from eastern Pennsylvania, were examined in Rosenbaum (1999b, Table 2). The pairs were matched for chain and starting wage before the increase. For example, the first pair consisted of two Burger Kings, one in New Jersey, the other in eastern Pennsylvania, both paying a starting wage of \$4.25 an hour before the increase. However, the pairs were not matched for two other, less important covariates, namely whether the store was company owned and the number of hours the store was open on a weekday before the wage increase. As it turns out, in the first

pair, neither Burger King was company owned and the New Jersey restaurant was open 17 hours while the Pennsylvania restaurant was open 16.5 hours. As an illustration of the methods of the current paper, the 66 pairs will be reanalyzed, making additional covariance adjustments for company ownership and hours open. As seen in the boxplots in Rosenbaum (1999b, Figure 1), the data contain several outliers—perhaps due to survey respondents who misunderstood questions needed to compute full time equivalent employment—and so robust or nonparametric methods are needed here. Aspects of this study are discussed in Rosenbaum (1999c). The data in Table 1 are being used to illustrate methodology, not to reach conclusions about the effects of the minimum wage, which would require consideration of issues beyond the scope of this paper.

4.3 Notation for an Observational Study with Matched Pairs

In the observational study, there are I matched pairs, with one treated subject and one control in pair i, $i = 1, \ldots, I$. In the example, $I = 66$. Quantities defined in Section 2 for a completely randomized experiment are essentially unchanged except that they now have a second subscript i indicating the pair. For instance, the jth subject in matched set i has covariates $\mathbf{x}_{ij}$, *with no coordinate for a constant term*, and would exhibit response r_{Cij} if assigned to control or response r_{Tij} if assigned to treatment, the treatment effect being additive if $r_{Tij} - r_{Cij} = \tau$ for all i, j. In the example, $\mathbf{x}_{ij}$ is two-dimensional: it records a binary variable indicating company ownership (1 for company owned; 0 otherwise) and number of hours open on a weekday before the wage increase. In the example, for jth restaurant in the ith pair, the bivariate potential responses (r_{Tij}, r_{Cij}) record the after-minus-before change in the number of full time equivalent employees if the minimum wage were increased to \$5.05, recorded in r_{Tij}, and if the minimum wage were not increased, recorded in r_{Cij}, where, of course, r_{Tij} is observed for New Jersey restaurants and r_{Cij} is observed for Pennsylvania restaurants. Also, $Z_{ij} = 1$ if the jth subject in matched set i received the treatment or $Z_{ij} = 0$ if this subject received the control, with $\sum_{j=1}^{2} Z_{ij} = 1$ for each i. In the example, $Z_{ij} = 1$ for a New Jersey restaurant and $Z_{ij} = 0$ for a restaurant in eastern Pennsylvania.

Write $V_i = Z_{i1} - Z_{i2}$, so $V_i = 1$ if the first subject in the pair is the treated subject and $V_i = -1$ if the second subject is the treated subject. Write $y_{Ci} = r_{Ci1} - r_{Ci2}$ and $\mathbf{d}_i = \mathbf{x}_{i1} - \mathbf{x}_{i2}$, both of which are fixed, not varying with the treatment assignment. Under the model of an additive treatment effect, the difference in observed responses, $R_{i1} - R_{i2} = Y_i$ say, is a random variable which equals $(r_{Ci1} + \tau Z_{i1}) - (r_{Ci2} + \tau Z_{i2}) = y_{Ci} + \tau V_i$, whereas $Y_i - \tau V_i = y_{Ci}$ is fixed. For the example, Table 2 records the observable quantities $R_{i1} - R_{i2}$, $\mathbf{d}_i = \mathbf{x}_{i1} - \mathbf{x}_{i2}$, where for convenient display the restaurants are renumbered, $j = 1, 2$, so the New Jersey restaurant is always first. Write $\mathbf{Y} = (Y_1 \ldots Y_I)^T$, $\mathbf{y}_C = (y_{C1} \ldots y_{CI})^T$, $\mathbf{V} = (V_1 \ldots V_I)^T$ and $\mathbf{D}$ for the matrix with I rows consisting of the $\mathbf{d}_i$, $i = 1, \ldots, I$.

TABLE 2
Employment and wages from Card and Krueger: 66 *Pairs of New Jersey and Pennsylvania restaurants*

Pair	Chain	Employees Y_i	Company owned d_{i1}	Hours open d_{i2}	Wage change L_i	Sheets
1	BK	12.50	0	0.5	0.65	(230,521)
2	BK	13.00	0	0.0	0.80	(258,45)
3	BK	−5.00	0	1.0	0.87	(168,48)
4	BK	20.50	0	−0.5	0.80	(267,41)
5	BK	−2.25	1	1.5	0.80	(91,435)
6	BK	3.50	0	−1.5	0.30	(105,476)
7	BK	−17.50	0	−2.0	0.30	(340,501)
8	BK	5.00	0	0.5	−1.20	(385,477)
9	BK	9.50	0	0.5	0.80	(66,40)
10	BK	3.00	0	0.5	0.80	(38,37)
11	BK	5.50	0	3.0	0.80	(200,430)
12	BK	−1.00	0	−1.0	0.80	(68,471)
13	BK	−0.50	0	2.5	0.15	(202,472)
14	BK	−12.50	0	−1.0	0.80	(418,434)
15	BK	9.50	0	−1.0	0.78	(249,522)
16	BK	16.00	0	−0.5	0.80	(84,42)
17	BK	3.00	0	1.0	0.55	(64,475)
18	BK	1.50	0	−3.0	0.70	(70,478)
19	BK	24.75	1	1.5	−0.05	(213,432)
20	BK	−19.50	0	−2.5	0.50	(2,450)
21	BK	−18.25	0	−4.0	0.55	(172,503)
22	BK	15.00	0	−1.0	0.55	(71,448)
23	BK	8.50	0	−1.0	0.75	(114,473)
24	BK	−17.50	0	−3.0	0.05	(156,449)
25	BK	4.00	1	3.5	1.00	(89,474)
26	BK	−12.75	1	−9.0	0.05	(298,451)
27	BK	−4.50	0	−2.0	0.75	(371,469)
28	BK	4.50	0	−3.0	0.05	(409,468)
29	BK	18.75	0	−3.0	0.18	(152,445)
30	BK	3.00	0	−5.5	0.30	(85,470)
31	KFC	−1.50	1	2.0	0.80	(278,438)
32	KFC	−2.50	0	−2.5	0.80	(216,51)
33	KFC	−7.00	−1	0.5	0.05	(185,454)
34	KFC	−4.00	−1	0.0	0.55	(158,485)
35	KFC	−0.50	0	−1.5	0.50	(159,50)
36	KFC	2.50	−1	2.0	0.80	(318,407)

TABLE 2
Continued

Pair	Chain	Employees Y_i	Company owned d_{i1}	Hours open d_{i2}	Wage change L_i	Sheets
37	KFC	5.50	0	1.5	0.05	(299,458)
38	KFC	−6.50	−1	0.0	0.30	(30,483)
39	KFC	3.50	−1	1.0	0.05	(31,455)
40	KFC	−3.25	0	0.0	−0.20	(274,481)
41	RR	8.25	0	0.0	0.70	(351,459)
42	RR	0.50	−1	−1.0	0.55	(366,492)
43	RR	−20.00	1	−1.0	0.80	(325,514)
44	RR	6.50	−1	−1.5	0.30	(101,511)
45	RR	19.25	0	1.0	0.14	(225,516)
46	RR	3.50	0	−9.0	0.80	(6,509)
47	RR	−2.00	0	0.0	0.35	(34,487)
48	RR	−5.00	0	4.5	0.30	(78,462)
49	RR	−3.00	1	2.0	0.80	(349,489)
50	RR	−6.00	0	3.0	0.30	(164,515)
51	RR	3.25	0	0.5	0.05	(163,490)
52	RR	14.00	0	1.0	0.30	(161,496)
53	RR	−9.50	0	2.0	0.05	(33,495)
54	RR	10.00	1	0.0	0.30	(190,488)
55	RR	19.00	−1	3.0	0.15	(77,493)
56	WE	0.50	0	1.0	1.05	(226,59)
57	WE	4.00	0	0.0	0.70	(195,443)
58	WE	−0.50	0	0.0	0.80	(310,60)
59	WE	−6.50	0	3.0	0.30	(247,444)
60	WE	−44.00	0	0.0	0.80	(142,441)
61	WE	−16.00	−1	−0.5	0.60	(104,57)
62	WE	6.50	−1	4.0	0.55	(248,58)
63	WE	−6.25	−1	−0.5	0.05	(166,498)
64	WE	13.25	−1	1.5	0.05	(82,499)
65	WE	12.50	0	−0.5	0.25	(406,56)
66	WE	4.00	−1	1.0	0.30	(36,61)

Notes: The restaurants are denoted BK, Burger King; KFC, Kentucky Fried Chicken; RR, Roy Rogers; WE, Wendys. Data are from Card and Krueger (1995, page 18). Sheet Numbers are Card and Krueger's restaurant identification numbers. More detail for these pairs is given in Table 1 of Rosenbaum (1999b).

In close parallel with Section 2.3, consider a fit to the potential control responses r_{Cij}, with a pair effect plus a linear term in $\mathbf{x}_{ij}$, say $\alpha_i + \beta\mathbf{x}_{ij}$. Following Rubin (1973, 1979), the pair effect α_i is eliminated by differencing the two responses in pair i. Then the difference between the two control responses in pair i is $y_{Ci} = r_{Ci1} - r_{Ci2}$, which is fitted by $(\alpha_i + \beta\mathbf{x}_{i1}) - (\alpha_i + \beta\mathbf{x}_{i2}) = \beta\mathbf{d}_i$ with no constant term. Differencing within pairs is effectively a special case of alignment within blocks, as discussed by Hodges and Lehmann (1962), and the general case of alignment will be discussed in Section 7.

Write Ω for the set of possible values of $\mathbf{Z} = (z_{11}\, z_{12}\, \ldots\, z_{I2})^T$, that is, the set containing the 2^I vectors $\mathbf{Z} = (z_{11}\, z_{12}\, \ldots\, z_{I2})^T$ of dimension $2I$ with $z_{ij} \in \{0, 1\}$, and $\sum_{j=1}^{2} z_{ij} = 1$ for all i. Each such $\mathbf{Z}$ corresponds with a unique $\mathbf{V}$ in an obvious way.

4.4 Treatment Assignment without Hidden Biases and with Overt Biases Balanced by Matching

For observational studies, two situations are considered: one here and the other in Section 5. In a randomized, matched study, a treatment assignment would be picked at random from Ω so each $\mathbf{z} \in \Omega$ would have probability 2^{-I}. Suppose instead that treatments were independently assigned with unknown probabilities π_{ij} that are functions of the observed covariates $\mathbf{x}_{ij}$ and then the pairs are formed based on the observed covariates such that $Z_{i1} + Z_{i2} = 1$, but $\pi_{i1} = \pi_{i2}$ for $i = 1, \ldots, I$. For instance, if π_{ij} is a function of $\mathbf{x}_{ij}$, then one way to produce such pairs is to match exactly for $\mathbf{x}_{ij}$, in which case the pairs are homogeneous in $\mathbf{x}_{ij}$ and covariance adjustment is superfluous. However, if π_{ij} is a function of $\mathbf{x}_{ij}$, then π_{ij} is the propensity score, and another way to produce such pairs is to match on the propensity scores π_{ij}, in which case the pairs will typically be heterogeneous in $\mathbf{x}_{ij}$, and further covariance adjustments for chance imbalances in $\mathbf{x}_{ij}$ may be useful.

If $\pi_{i1} = \pi_{i2}$ for $i = 1, \ldots, I$, then the conditional distribution of $\mathbf{Z}$ given $Z_{i1} + Z_{i2} = 1$, $i = 1, \ldots, I$, equals the randomization distribution; that is, it is uniform on Ω, with each $\mathbf{z} \in \Omega$ having probability 2^{-I} (see Rosenbaum, 1984, 1995, Section 3, for a discussion and the elementary proof). Notice in particular that given $Z_{i1} + Z_{i2} = 1$, the chance that $Z_{i1} = 1$ is $\pi_{i1}/(\pi_{i1} + \pi_{i2})$, which is $\frac{1}{2}$ if $\pi_{i1} = \pi_{i2}$. In other words, if it suffices to adjust for the observed covariates $\mathbf{x}_{ij}$ and the matching controls the probability of treatment given $\mathbf{x}_{ij}$, even if the pairs are not matched for $\mathbf{x}_{ij}$ itself, then the matching creates the randomization distribution. This situation is assumed in Section 4. In Section 5, π_{ij} will be assumed to be a function of $\mathbf{x}_{ij}$ and an unobserved covariate u_{ij}, so π_{ij} is not the propensity score given $\mathbf{x}_{ij}$, and matching on $\mathbf{x}_{ij}$ alone or functions of $\mathbf{x}_{ij}$ such as the propensity score will not typically balance u_{ij}.

4.5 Matched Pairs and the Signed Rank Test

To test the hypothesis $H_0: \tau = \tau_0$, use the hypothesized τ_0 to calculate the adjusted response differences $\mathbf{Y} - \tau_0\mathbf{V}$, which equal the fixed vector of response differences under control $\mathbf{y}_C$ if the null hypothesis is true. Use some form of regression with no constant term to fit the adjusted response differences $\mathbf{Y} - \tau_0\mathbf{V}$ using $\mathbf{D}$,

obtaining residuals $\mathbf{e}_0 = \varepsilon(\mathbf{Y} - \tau_0\mathbf{V}) = (e_{01} \ldots e_{0I})^T$, which under the null hypothesis equal $\varepsilon(\mathbf{y}_C)$, a fixed quantity not varying with $\mathbf{V}$. Let q_i be the rank of $|e_i|$ with average ranks for ties, let $s_{i1} = 1$ if $e_i > 0$, $s_{i1} = 0$ otherwise and let $s_{i2} = 1$ if $e_i < 0$, $s_{i2} = 0$ otherwise. Wilcoxon's signed rank statistic is the sum of the ranks of the absolute differences in the residuals for pairs in which the treated subject had the larger residual, that is, $t(\mathbf{Z}, \mathbf{e}_0) = \sum_{i=1}^{I}\sum_{j=1}^{2} Z_{ij}s_{ij}q_i$, where s_{ij} and q_i are fixed constants under the null hypothesis.

Because $\Pr(\mathbf{Z} = \mathbf{z}) = 2^{-I}$ for each $\mathbf{z} \in \Omega$, it follows under the null hypothesis $H_0: \tau = \tau_0$, that the signed rank statistic $t(\mathbf{Z}, \mathbf{e}_0)$ has as its exact null distribution the usual distribution of Wilcoxon's signed rank statistic. In particular, if there are neither ties nor zero residuals, then its null expectation and variance are $E\{t(\mathbf{Z}, \mathbf{e}_0)\} = \frac{I(I+1)}{4}$ and $\text{var}\{t(\mathbf{Z}, \mathbf{e}_0)\} = \frac{I(I+1)(2I+1)}{24}$.

A confidence interval for τ is obtained by inverting the test, that is, by testing each hypothesis $H_0: \tau = \tau_0$ and retaining those τ_0's that are not rejected (Lehmann, 1986, Section 3.5, Theorem 3.4, page 90). A point estimate $\widehat{\tau}$ is obtained by the general principle suggested by Hodges and Lehmann (1963): equate the statistic to its null expectation $t(\mathbf{Z}, \mathbf{e}_0) = t\{\mathbf{Z}, \varepsilon(\mathbf{Y} - \tau_0\mathbf{V})\} = \frac{I(I+1)}{4}$ and solve for the point estimate $\widehat{\tau}$, with slight allowance for the discreteness of a rank statistic.

A minor technical point deserves mention. The argument just given is correct as it stands. However, the order of the two subjects, $j = 1$ and $j = 2$, in pair i is arbitrary and therefore one might prefer that this order did not affect inferences about τ. For many fitting procedures $\varepsilon(\mathbf{Y} - \tau_0\mathbf{V})$, the issue does not arise, because $t\{\mathbf{Z}, \varepsilon(\mathbf{Y} - \tau_0\mathbf{V})\}$ and its null distribution are unchanged by changing the order within a pair. For instance, this is true if the regression is fitted using common m-estimates and the test statistic is the signed rank statistic. [To show this, suppose $\varepsilon(\mathbf{Y} - \tau_0\mathbf{V})$ is the vector of residuals obtained by m-estimation in which the adjusted responses $\mathbf{Y} - \tau_0\mathbf{V}$ are regressed on $\mathbf{D}$ with an odd $\psi(\cdot)$ function, $\psi(a) = -\psi(-a)$ for all a, so one solves for β in the system of equations $\mathbf{0} = \sum \mathbf{D}_i\psi(Y_i - \tau_0V_i - \mathbf{D}_i^T\widehat{\beta})$; see Huber (1981, Section 7.3). Now reversing the order in pair i changes $\mathbf{D}_i$ to $-\mathbf{D}_i$, changes $Y_i - \tau_0V_i - \mathbf{D}_i^T\beta$ to $-(Y_i - \tau_0V_i - \mathbf{D}_i^T\beta)$, and, because, $\psi(\cdot)$ is odd, $\psi(Y_i - \tau_0V_i - \mathbf{D}_i^T\beta)$ changes to $-\psi(Y_i - \tau_0V_i - \mathbf{D}_i^T\beta)$, so $\mathbf{D}_i\,\psi(Y_i - \tau_0V_i - \mathbf{D}_i^T\beta)$ is unchanged, the solution $\widehat{\beta}$ is unchanged and the residual $\mathbf{Y} - \tau_0\mathbf{V} - \mathbf{D}_i^T\widehat{\beta}$ becomes $-(\mathbf{Y} - \tau_0\mathbf{V} - \mathbf{D}_i^T\widehat{\beta})$, q_i is unchanged, s_{i1} and s_{i2} are interchanged, Z_{i1} and Z_{i2} are interchanged and the signed rank statistic $t(\mathbf{Z}, \mathbf{e}_0) = \sum_{i=1}^{I}\sum_{j=1}^{2} Z_{ij}s_{ij}\,q_i$ is unchanged, as desired.] The situation with local smoothers is different. Suppose $\mathbf{D}_i$ is a scalar, say D_i, and $Y_i - \tau_0V_i$ is regressed on D_i using a local smoother such as Cleveland's Lowess. Now, changing the order within pair i changes the signs of both $Y_i - \tau_0V_i$ and D_i.

4.6 Covariance Adjustment in the Example

In the minimum wage example, there are $I = 66$ pairs of restaurants, in which Y_i is the NJ-minus-PA difference in the post-minus-pre change in the number of full time equivalent employees. There are two unmatched covariates, namely whether the store was company owned and the number of hours the store was open on a weekday; see Table 2.

With $I = 66$ pairs, under the null hypothesis, the signed rank statistic T has null expectation $I(I+1)/4 = 1105.5$ and null variance $I(I+1)(2I+1)/24 = 24505.25$, and standardized deviate $\{T - I(I+1)/4\}/\sqrt{I(I+1)(2I+1)/24}$. One computes $t\{\mathbf{Z}, \varepsilon(\mathbf{Y} - \tau_0\mathbf{V})\}$ for various τ_0 and compares it with this null distribution. If $\tau_0 = 2.065$, the signed rank statistic is 1105, slightly below the null expectation of 1105.5, but if $\tau_0 = 2.0649$, the signed rank statistic is 1106, slightly above the null expectation, so $\widehat{\tau} = 2.065$, that is, contrary to economic theory, a gain of about two employees per restaurant. If $\tau_0 = -0.58$, the signed rank statistic is 1413, yielding a deviate of 1.96. If $\tau_0 = 4.8075$, the signed rank statistic is 798, yielding a deviate of -1.96. Hence, the approximate 95% confidence interval for τ_0 is $[-0.58, 4.81]$, so a loss of about half an employee is plausible, but so is a gain of about five employees. This confidence interval does not suggest dramatic declines in employment in the fast food industry brought on by the increase in New Jersey's minimum wage. Again, these inferences assume there is no hidden bias. How might the conclusions change if hidden biases are present?

5. SENSITIVITY TO HIDDEN BIAS: MATCHED STUDIES WITH COVARIANCE ADJUSTMENT

5.1 Treatment Assignment in an Observational Study

A sensitivity analysis asks how hidden biases of various magnitudes might alter conclusions. Although all studies are sensitive to sufficiently large biases, studies vary markedly in their degree of sensitivity to hidden bias; see the examples in Rosenbaum (1995, Section 4). Here, a simple method of sensitivity analysis

for the signed rank statistic (Rosenbaum 1987, 1991b, 1995, Section 4) is generalized for use with covariance adjustment. Essentially, it is shown that it is appropriate to carry out this standard sensitivity analysis on the residuals from a regression fit. Other methods of sensitivity analysis have been discussed by Cornfield et al. (1959), Greenhouse (1982), Rosenbaum and Rubin (1983), Rosenbaum (1986, 1996), Manski (1990, 1995), Gastwirth (1992), Angrist, Imbens and Rubin (1996), Copas and Li (1997), Gastwirth, Krieger and Rosenbaum (1998), Lin, Psaty and Kronmal (1998) and Berk and De Leeuw (1999). In particular, Rosenbaum (1986) and Lin, Psaty and Kronmal (1998) discussed sensitivity analyses for particular types of covariance adjustment, although from a very different point of view than is taken here.

The chance that subject $j = 1$ in pair i receives the treatment $Z_{i1} = 1$ given that one subject in pair i receives the treatment $Z_{i1} + Z_{i2} = 1$ is $\pi_{i1}/(\pi_{i1} + \pi_{i2})$. The sensitivity analysis model says matched subjects may differ in their chances of receiving the treatment by at most a factor of $\Gamma \geq 1$,

$$\begin{aligned}\Gamma &\geq \frac{\Pr(Z_{i1}=1)}{\Pr(Z_{i2}=1)} = \frac{\pi_{i1}}{\pi_{i2}} \\ &= \frac{\pi_{i1}/(\pi_{i1}+\pi_{i2})}{\pi_{i2}/(\pi_{i1}+\pi_{i2})} \geq \frac{1}{\Gamma} \quad \text{for } i = 1, \ldots, I.\end{aligned} \tag{1}$$

When $\Gamma = 1$, it follows that $\pi_{ij}/(\pi_{i1} + \pi_{i2}) = \frac{1}{2}$ for each i, j, resulting in the randomization distribution. When $\Gamma > 1$, it follows that $\Pr(Z_{ij} = 1)$ is unknown, so there may not be a single inference about the treatment effect τ, but rather a range of inferences (e.g., a range of possible significance levels), reflecting uncertainty about how treatments were assigned, with the range widening as Γ increases. A sensitivity analysis computes the range of possible inferences for several values of Γ, thereby displaying the degree to which hidden biases of various magnitudes might alter the conclusions of the study.

The model (1) may be derived by assuming that the matching has failed to control for an unobserved binary covariate u_{ij} associated with Γ-fold increase in the odds of exposure to treatment. See Rosenbaum (1987, 1995, Section 4) for detailed discussion.

5.2 Sensitivity Analysis for Covariance Adjustment

The procedure is similar to that in Section 4.5 except that the signed rank statistic is no longer governed by its randomization distribution, but rather has a range of distributions implied by (1). Specifically, to test the hypothesis $H_0 : \tau = \tau_0$, compute the signed rank statistic $T = t(\mathbf{Z}, \mathbf{e}_0) = \sum_{i=1}^{I} \sum_{j=1}^{2} Z_{ij} s_{ij} q_i$ exactly as in Section 4.5, where s_{ij} and q_i are fixed under the null hypothesis.

When $\Gamma > 1$, the null distribution of the signed rank statistic is unknown, but is bounded by two known distributions. Let $\overline{\overline{T}}$ be the sum of I independent random variables, $i = 1, \ldots, I$, taking values 0 with probability $\frac{1}{1+\Gamma}$ and the value $(s_{i1} + s_{i2})q_i$ with value $\frac{\Gamma}{1+\Gamma}$. Similarly, let $\overline{T}$ be the sum of I independent random variables, $i = 1, \ldots, I$, taking values 0 with probability $\frac{\Gamma}{1+\Gamma}$ and the value $(s_{i1} + s_{i2})q_i$ with probability $\frac{1}{1+\Gamma}$. Inequality (1) implies that the null distribution of the signed rank statistic T is bounded in the sense of stochastic order by the distributions of $\overline{T}$ and $\overline{\overline{T}}$, that is, $\Pr(\overline{\overline{T}} \geq k) \geq \Pr(T \geq k) \geq \Pr(\overline{T} \geq k)$ for every k; see Rosenbaum (1987, 1995, Section 4) for proof. As a consequence, for each fixed $\Gamma \geq 1$, although the significance level $\Pr(T \geq k)$ is unknown, bounds on its value are easily computed. The sensitivity analysis computes these bounds for several values of Γ to display the sensitivity of the inference to hidden bias.

The expectation and variance of $\overline{T}$ and $\overline{\overline{T}}$ are

$$E(\overline{\overline{T}}) = \frac{\Gamma}{1+\Gamma} \sum (s_{i1} + s_{i2})q_i,$$

$$E(\overline{T}) = \frac{1}{1+\Gamma} \sum (s_{i1} + s_{i2})q_i,$$

$$\operatorname{var}(\overline{\overline{T}}) = \operatorname{var}(\overline{T}) = \frac{\Gamma}{(1+\Gamma)^2} \sum \{(s_{i1} + s_{i2})q_i\}^2.$$

If there are neither ties among the $|e_i|$ nor zero differences $|e_i| = 0$, then $s_{i1} + s_{i2} = 1$ for every i, and then there are standard simplifications $\sum (s_{i1} + s_{i2})q_i = \sum q_i = I(I+1)/2$ and $\sum \{(s_{i1} + s_{i2})q_i\}^2 = \sum q_i^2 = I(I+1)(2I+1)/6$; see Lehmann (1999, problem 84, page 51). Bounds on point estimates and confidence intervals follow immediately in the usual way, that is, by inverting tests to get confidence intervals and using the device of Hodges and Lehmann (1963) to get point estimates; see Rosenbaum (1987, 1993, 1995, Section 4).

5.3 Example: Sensitivity Analysis for Minimum Wage Effects

Table 3 displays a concise sensitivity analysis for Card and Krueger's minimum wage study in Section 4.2. Within pairs exactly matched for restaurant chain and closely matched for starting wages at baseline, covariance analysis makes further corrections for

TABLE 3
Minimum point estimates and maximum P-values

		$H_0:\tau$	
Γ	$\widehat{\tau}_{\text{min}}$	-2	-4
1	2.06	0.0021	0.000029
1.5	0.14	0.069	0.0039
2	−1.22	0.29	0.038

the two unmatched covariates. The anticipation of a negative employment effect by economic theory, together with the neutral to slightly positive differences found by Card and Krueger, suggest that a sensitivity analysis might reasonably ask: Would small hidden biases reconcile economic theory with the Card and Krueger data?

For this reason, Table 3 asks the following three questions repeatedly for several values of Γ in (1). What is the minimum point estimate $\widehat{\tau}_{\text{min}}$ possible subject to the bound (1)? What is the maximum significance level, subject to (1), for testing the hypothesis of a 2 employee decline $H_0:\tau=-2$? What is the maximum one-sided significance level, subject to (1), for testing the hypothesis of a 4 employee decline $H_0:\tau=-4$? For the typical restaurants, the number of full time equivalent employees was about 20, so a decline of 2 employees is about 10% and 4 employees is about 20%.

In Table 3, the row labeled $\Gamma=1$ repeats the analysis in Section 4.6 assuming no hidden biases, because when $\Gamma=1$ in (1), the randomization distribution within pairs is produced. In this case, there is only one point estimate, namely a $\widehat{\tau}=2.06$ employee gain, and there is only one significance level for each hypothesis, namely 0.0021 for $H_0:\tau=-2$ and 0.000029 for $H_0:\tau=-4$. If there were no hidden biases, the point estimate suggests a gain in employment and substantial declines are not plausible.

These results are not materially altered by a small hidden bias of $\Gamma=1.5$. A bias of this magnitude refers to an unobserved variable u, say a binary attribute, strongly related to employment change and about 50% more common in New Jersey than in eastern Pennsylvania—that is, an odds ratio of 1.5 linking u and the state. For $\Gamma=1.5$ in Table 3, the minimum point estimate remains slightly positive and neither hypothesis looks especially plausible.

For a moderate bias of $\Gamma=2$, the minimum possible point estimate compatible with (1) is now somewhat negative, a decline of about $\widehat{\tau}_{\text{min}}=-1.22$ employees. The hypothesis of a two employee decline is now plausible for some π_{ij} satisfying (1), whereas a four employee decline is still implausible for all π_{ij} satisfying (1).

The parameter Γ measures the degree of departure from a randomized experiment. One way, perhaps the best way, to develop a feeling for various magnitudes of Γ is to proceed empirically, looking at past observational studies, noting the value of Γ at which the conclusions become sensitive to hidden bias, that is, the value at which several competing and conflicting interpretations are simultaneously plausible. A number of such examples are given in Rosenbaum (1995, Section 4). By comparison with other studies, the current study is neither sensitive to extremely small hidden biases, as was true of a study of coffee as a cause of myocardial infarction, nor extremely insensitive to large biases, as was true of several studies of smoking as a cause of lung cancer or diethylstilbestrol as a cause of vaginal cancer.

A confidence interval for a parameter quantifies and expresses the uncertainty that is due to sampling variation; however, it does not dispel that uncertainty. In parallel, a sensitivity analysis quantifies and expresses the uncertainty that is due to hidden bias, but does not dispel that uncertainty. In both cases, it is useful to have an objective measure of the degree of uncertainty that is actually present in the study at hand, because the degree of uncertainty varies from study to study and is difficult to appraise without quantitative techniques.

6. INSTRUMENTAL VARIABLES

6.1 Role of Instrumental Variables

In Sections 4 and 5, the effects on employment of changes in minimum wage laws were modelled as a constant, additive effect of the laws themselves. One might reasonably believe that minimum wage laws affect employment primarily, if not exclusively, by changing wages. For example, if one raised the minimum wage to \$5.05 per hour in a region where market forces had already pushed starting wages well above \$5.05, then one might expect to see negligible consequences for employment, whereas a similar change in law in different market conditions might produce different effects. One might possibly believe that what is stable from one circumstance to another is not the effect of increasing the minimum wage as recorded in law, but the effect of increasing the wages paid to employees. If one believed this, one might wish to model

the change in employment in terms of actual changes in wages, using the change in law as an instrumental variable.

Estimation using instrumental variables (IV) has a long history and is a standard topic in econometrics; see Davidson and MacKinnon (1993, Section 7) for one good textbook discussion and see Manski (1995) for a less traditional but insightful discussion. More recently, Angrist, Imbens and Rubin (1996) recast the IV argument, stripping away most modelling and distributional assumptions and expressing the IV argument in terms of potentially observable outcomes. In their approach, they argued that it is sometimes natural to model the effect of an assigned treatment on one outcome, not in terms of the assignment to treatment or control, but rather in terms of the effect of the treatment on a second outcome. Typically, this second outcome is really a measure of the degree to which the assigned treatment is actually delivered. What is unique about the IV argument is that the assigned treatment—the instrument—and the delivered treatment—the secondary outcome—both enter the estimation, but with very different roles.

This issue arises perhaps most clearly in clinical trials with noncompliance, for instance, in which patients consume only part of the drug assigned to them by the experimental protocol. In a randomized clinical trial, the assigned treatment is randomized, but it does not describe the treatment the patient actually received. If noncompliance is ignored, and the assigned treatment groups are compared, then the so-called intent-to-treat estimate is obtained. It accurately estimates the effect of encouraging patients to take a drug, but it does not estimate the effect of the drug. If the assigned treatment is ignored and groups are defined by the dose they actually received, the benefits of randomization are lost and a severely biased estimate of effect may result. The IV estimate uses the randomly assigned treatment as an instrument for the dose of treatment actually received. See Rosenbaum (1996) for a very simple numerical example of how this works. The use of IV in randomized experiments with noncompliance is discussed with varied terminology by Sommer and Zeger (1991), Sheiner and Rubin (1995), Frangakis and Rubin (1999) and Barnard, Frangakis, Hill and Rubin (1999). In Angrist, Imbens and Rubin (1996), the assigned treatment was the Vietnam era draft lottery, while the delivered treatment was actual military service. In randomized clinical trials with noncompliance and in the draft lottery, there is reason to hope that the assigned treatment is randomized or nearly so, and there is good reason to doubt that the delivered treatment is randomized. In these examples, the instrument is valuable because it is randomized, but the secondary outcome is valuable because it reflects the potent part of the treatment, the part that is likely to produce effects. Central to the IV argument is that the instrument affects the outcome only indirectly through the secondary outcome, and the instrument is associated with the secondary outcome; see Angrist, Imbens and Rubin (1996) for some specifics, and see Rosenbaum (1996, 1999b) for methods of exact permutation inference using instrumental variables.

The minimum wage example is both more typical and more problematic than the clinical trial and Vietnam draft lottery examples. Here, the change in New Jersey's minimum wage is the instrument and the change in starting wages is the secondary outcome. The analyses in Sections 4 and 5 are analogous to the intent-to-treat analysis; they focus on the change in law, whether or not the change in law resulted in changes in wages. Here, it is plausible but not certain that changes in minimum wage laws affect employment only indirectly through changes in starting wages. As discussed in Section 4, it is far from clear that the instrument is randomized—that is, that the New Jersey restaurants can safely be viewed as a simple random sample from the finite population of New Jersey and eastern Pennsylvania restaurants. For this reason, one needs to examine the sensitivity of IV estimates to possible departures from random assignment of the instrument. Without covariance adjustment, a simple method of exact permutation inference and sensitivity analysis for IV was discussed in Rosenbaum (1996, 1999b). Here, that method is extended for use with covariance adjustment. All that is involved is the merging of two methods, namely the methods of Sections 4 and 5 above and the methods of permutation inference for IV, so the discussion that follows can be brief.

6.2 Methods for an Instrumental Variable

In addition to the outcome of primary interest, (r_{Cij}, r_{Tij}), whose observed value is R_{ij}, there is a secondary outcome, (w_{Cij}, w_{Tij}), whose observed value is $W_{ij} = Z_{ij}w_{Cij} + (1 - Z_{ij})w_{Tij}$. In the minimum wage example, (r_{Cij}, r_{Tij}) describes the change in full time equivalent employment and (w_{Cij}, w_{Tij}) describes the change in starting wages, where C indicates the control condition in which the minimum wage is not increased and T indicates the treatment condition in which it is increased from \$4.25

to \$5.05. The increase in New Jersey's minimum wage forced many, but not all, New Jersey restaurants to sharply raise the starting wage; indeed, many raised their starting wage from the old minimum to the new minimum, so $w_{Tij} = \$0.80$ for these restaurants.

The effect of the treatment is modelled in terms of the secondary outcome, $r_{Tij} - r_{Cij} = \beta(w_{Tij} - w_{Cij})$, which implies Angrist, Imbens and Rubin's exclusion restriction, and this model says the treatment effect on employment is proportional to the effect on starting wages. For instance, if a restaurant paid \$5.10 an hour before the increase and would raise the wage by \$0.25 whether or not the minimum wage was increased, then $w_{Tij} = w_{Cij} = 0.25$, and this model says this one restaurant should experience no effect on employment, $r_{Tij} - r_{Cij} = \beta(w_{Tij} - w_{Cij}) = 0$. The model of an additive treatment effect, used in Sections 4 and 5, is the special case in which $w_{Tij} = 1$ and $w_{Cij} = 0$ for all i, j, so treatment assignment Z_{ij} is the same as treatment received W_{ij}. Write $L_i = W_{i1} - W_{i2}$ for the observed difference in the secondary outcomes, so L_i is the matched pair difference in the change in starting wages. The adjusted difference $Y_i - \beta L_i$ takes one of two possible values, namely the value $(r_{Ti1} - \beta w_{Ti1}) - (r_{Ci2} - \beta w_{Ci2})$ if the first unit in pair i received the treatment, $Z_{i1} = 1$, or the value $(r_{Ci1} - \beta w_{Ci1}) - (r_{Ti2} - \beta w_{Ti2})$ if the second unit in pair i received the treatment, $Z_{i2} = 1 - Z_{i1} = 1$. Therefore, under the model $r_{Tij} - r_{Cij} = \beta(w_{Tij} - w_{Cij})$, the adjusted difference $Y_i - \beta L_i$ equals

$$\begin{aligned} &Z_{i1}\{(r_{Ti1} - \beta w_{Ti1}) - (r_{Ci2} - \beta w_{Ci2})\} \\ &\quad + (1 - Z_{i1})\{(r_{Ci1} - \beta w_{Ci1}) - (r_{Ti2} - \beta w_{Ti2})\} \\ &\quad = (r_{Ci1} - \beta w_{Ci1}) - (r_{Ti2} - \beta w_{Ti2}) \\ &\quad = \widetilde{y}_{Ci}, \qquad \text{say,} \end{aligned}$$

which is constant, not varying with the treatment assignment Z_{ij}. Write $\mathbf{L} = (L_1 \ldots L_I)^T$ and $\widetilde{\mathbf{y}}_C = (\widetilde{y}_{C1} \ldots \widetilde{y}_{CI})^T$.

To test $H_0 : \beta = \beta_0$, fit the adjusted responses, $\mathbf{Y} - \beta_0\mathbf{L}$, using $\mathbf{D}$ in a model without a constant term, obtaining residuals $\widetilde{\mathbf{e}}_0 = \varepsilon(\mathbf{Y} - \beta_0\mathbf{L}) = (\widetilde{e}_1 \ldots \widetilde{e}_I)^T$, which equal the fixed $\varepsilon(\widetilde{\mathbf{y}}_C)$ if the null hypothesis is true. Compute the signed rank statistic, $T = t(\mathbf{Z}, \widetilde{\mathbf{e}}_0) = \sum_{i=1}^{I}\sum_{j=1}^{2} Z_{ij}\, s_{ij}\, q_i$, as in Section 4.5, where s_{ij} and q_i are fixed under the null hypothesis because they are functions of fixed quantities. Under the null hypothesis, in the absence of hidden bias, this signed rank statistic has its usual randomization distribution, whereas under the sensitivity analysis model (1), bounds on its distribution are the same as in Section 4.2.

One obtains a two-sided 95% confidence set for β by testing each hypothesis $H_0 : \beta = \beta_0$, retaining in the set those values of β_0 not rejected in a two-sided 0.05-level test. The 95% confidence interval is the shortest interval containing this set.

6.3 Identifying Conditions, Long Confidence Intervals, Rejecting the IV Specification

Discussions of IV methods typically involve an identifying assumption which asserts, in one form or another, that the instrument Z_{ij} is positively related to the secondary outcome W_{ij}. For instance, this assumption would be true if the treatment Z_{ij} had a positive effect on the secondary outcome $w_{Tij} > w_{Cij}$ for all i, j, because $W_{ij} = Z_{ij}w_{Tij} + (1 - Z_{ij})w_{Cij}$, and the assumption would be false if the treatment did not affect the secondary outcome $w_{Tij} = w_{Cij}$ for all i, j. Without some such assumption, typical IV methods are inconsistent, and there is often concern about the performance of IV methods when the correlation between Z_{ij} and W_{ij} is low. No such assumption was made or needed in Section 6.2 and this merits some discussion.

As noted in Rosenbaum (1999b) in a simpler but essentially parallel context, in exact permutation inference for IV, the assumption that Z_{ij} and W_{ij} are correlated is not needed. However, when Z_{ij} and W_{ij} are unrelated or weakly related, the confidence interval for β may be quite long, perhaps infinite in length, possibly a half line or the entire line. This is familiar in nonparametric inference: when relevant information is limited, a nonparametric interval maintains 95% coverage by becoming longer, perhaps infinite in length. For example, the standard nonparametric 95% confidence interval for the 99% quantile from a sample of size 20 will be a half line, properly reflecting the fact that, without distributional assumptions, 20 observations help to provide a lower bound but not an upper bound on the 99% quantile. In the same way, when Z_{ij} and W_{ij} are weakly related, the confidence interval for β may be long or infinite; that is, the hypothesis test may fail to reject $H_0 : \beta = \beta_0$ for every β_0.

This is a desirable property of permutation methods when compared to conventional methods. The data at hand speak to the question of whether Z_{ij} and W_{ij} are sufficiently strongly related to use IV methods. Better than assuming, perhaps incorrectly, that Z_{ij} and W_{ij} are sufficiently strongly related to use IV methods, the permutation inference correctly reflects the observed

relationship between Z_{ij} and W_{ij} and requires no assumptions. This is desirable for two reasons. First, an assumption is replaced by an observation. Second, cases at the margin, where Z_{ij} and W_{ij} are related but perhaps not as strongly as one might hope, are not forced into an artificial dichotomy of "identified" or "unidentified"; rather, they result in an interval that is appropriately somewhat longer.

The IV model may be incorrect or misspecified, and evidence of this may arise in the following way. The IV model says the effect of the treatment on the primary outcome is proportional to the effect of the treatment on the secondary outcome, that is, $r_{Tij} - r_{Cij} = \beta(w_{Tij} - w_{Cij})$. Just as the test of $H_0 : \beta = \beta_0$ may fail to reject every β_0, returning the entire line as a confidence interval, it may alternatively reject every hypothesis $H_0 : \beta = \beta_0$, returning the empty set as the confidence interval. For instance, we would expect an empty confidence set in a sufficiently large study if the model $r_{Tij} - r_{Cij} = \beta(w_{Tij} - w_{Cij})$ were wrong in the sense that the treatment had a large effect on the primary outcome, say $r_{Tij} - r_{Cij} = \tau \gg 0$, but it had no effect on the secondary outcome, $w_{Tij} = w_{Cij}$ for all i, j, so that no value of β can represent this effect. An empty confidence set for β should be viewed as a rejection of the IV model $r_{Tij} - r_{Cij} = \beta(w_{Tij} - w_{Cij})$.

In the econometric terminology commonly used to describe IV methods, one might say that permutation methods automatically appraise the identifying assumption, perhaps returning a confidence interval of infinite length if the assumption is inappropriate, and automatically conduct a specification test, perhaps returning an empty confidence interval if the specification is incorrect.

6.4 Instrumental Variable Estimates in the Minimum Wage Example

In Table 2, in the first pair, $i = 1$, of Burger King restaurants, the starting wage in New Jersey rose from the old minimum of \$4.25 to the new minimum of \$5.05, so $W_{11} = \$0.80$, whereas in the Pennsylvania restaurant the starting wage rose from \$4.25 to \$4.40, so $W_{12} = \$0.15$, and the difference is $L_1 = W_{11} - W_{12} = \$0.65$; that is, the New Jersey restaurant raised its wage by \$0.65 more than did the Pennsylvania restaurant. The coefficient β attempts to model the effect of the minimum wage increase in terms of the changes in wages, here \$0.65, rather than in terms of the dichotomous change in the law, because the change in law forced larger changes in wages on some restaurants than on others.

TABLE 4
Minimum point estimates and maximum P-values using IV

		$H_0 : \tau$	
Γ	$\widehat{\beta}_{\min}$	-2.5	-5
1	4.28	0.009	0.001
1.5	0.25	0.17	0.054
2	-2.52	0.50	0.25

Table 4 displays the covariance adjusted estimates and sensitivity analysis from the IV model, with adjustment for the two covariates. Conventional economic theory predicts that a rise in wages will result in a decline in employment, that is, a negative value for β. Using the randomization distribution within pairs, $\Gamma = 1$, the single point estimate is positive, $\widehat{\beta} = 4.28$, or a 4.28 employee increase for a \$1.00 increase in wages. Two hypotheses consistent with conventional theory are tested, namely $H_0 : \beta = -2.5$ and $H_0 : \beta = -5$, which parallel the hypotheses in Table 3. The New Jersey law raised the minimum wage by \$0.80, so these hypotheses correspond to declines of two or four employees in a pair of restaurants with stable wages in Pennsylvania and the full \$0.80 increase in New Jersey, namely $-2.5 \times 0.80 = -2$ and $-5 \times 0.80 = -4$. In the absence of hidden bias, $\Gamma = 1$, neither hypothesis is plausible.

For a small bias of $\Gamma = 1.5$, the minimum point estimate, $\widehat{\beta}_{\min} = 0.25$, is still positive, although the hypothesis $H_0 : \beta = -2.5$ is no longer implausible for some assignment probabilities satisfying (1). For a moderate bias of $\Gamma = 2$, the minimum point estimate, $\widehat{\beta}_{\min} = -2.52$, corresponds to a decline of two employees for an \$0.80 increase in starting wages, and both hypotheses are plausible for some assignment probabilities satisfying (1). In this one example, the IV estimate is about as sensitive to bias as the additive effect estimate in Section 5.

7. MATCHING WITH MULTIPLE CONTROLS AND FULL MATCHING

Various matching structures are often used in place of matched pairs. For instance, if controls are plentiful and inexpensive, but treated subjects are limited or expensive, then matching with multiple controls may increase precision with little increase in cost (Ury, 1975; Smith, 1997). Matching with a variable number of con-

trols per treated subject may remove substantially more bias than matching the same number of controls in a fixed matching ratio (Ming and Rosenbaum, 2000). Examples of matching with multiple controls are in Jick et al. (1973), Cohn et al. (1981) and Smith (1997). Even greater reductions in bias are possible with full matching, in which a treated subject may have several controls or a control may have several treated subjects (Rosenbaum, 1991b; Gu and Rosenbaum, 1993). In particular, full matching may be applied when a fixed data set is available essentially without cost and one wishes to use every available treated and control subject.

In studies with more than two subjects in a matched set, one might replace the signed rank statistic by the extension proposed by Hodges and Lehmann (1962), namely the aligned rank statistic. The procedure is as follows. Matched set i now contains $n_i \geq 2$ subjects, numbered $j = 1, \ldots, n_i$, of which m_i received the treatment, $0 < m_i < n_i$, where $m_i = \sum_{j=1}^{n_i} Z_{ij}$, with $N = \sum n_i$ subjects in total. In a randomized experiment or an observational study without hidden biases and with the propensity score controlled by the matching, the $\binom{n_i}{m_i}$ possible treatment assignments in matched set i are equally likely, each having probability $\binom{n_i}{m_i}^{-1}$. Under the model of an additive treatment effect τ, the hypothesis $H_0 : \tau = \tau_0$ is tested by computing the adjusted responses, $R_{ij} - \tau_0 Z_{ij}$, and aligning them by subtracting their mean in each matched set, yielding aligned, adjusted responses

$$(R_{ij} - \tau_0 Z_{ij}) - \frac{1}{n_i} \sum_{k=1}^{n_i} (R_{ik} - \tau_0 Z_{ik}).$$

The covariates also are aligned, subtracting their means within each matched set. The aligned, adjusted responses are regressed on the aligned covariates, perhaps using robust regression, and the residuals are ranked from 1 to N with average ranks for ties. The sum of the ranks of the residuals for treated subjects is the test statistic.

When there are no covariates, the method just described reduces to the aligned rank statistic of Hodges and Lehmann (1962). On the other hand, if least squares is used to fit the regression, if the residuals themselves are used instead of the ranks and if τ is estimated by the general device of Hodges and Lehmann (1963), so the sum the residuals for treated subjects is equated to zero and solved as a function of τ_0, then the resulting solution $\hat{\tau}$ is the usual least estimate of τ in a regression with treatment indicator, matched set indicators and covariates.

When treatment assignments within matched sets have the randomization distribution, with each assignment having probability $\binom{n_i}{m_i}^{-1}$, the aligned rank statistic has its usual permutation distribution as studied by Hodges and Lehmann (1962). A sensitivity analysis for the aligned rank statistic was proposed and illustrated by Gastwirth, Krieger and Rosenbaum (2000), and it too may be applied to the ranked residuals from the regression. The logic justifying this is the same as in Sections 4 and 5.

8. CONCLUSION

Starting with Fisher's (1935) approach to inference in randomized experiments, an exact, distribution-free theory of covariance adjustment in experiments was developed in Section 2. The method removes a hypothesized treatment effect from the responses, fits these adjusted responses using covariates and applies conventional permutation methods to the resulting residuals. The method is general: the adjustment for covariates may use robust regression fitting or functional smoothers, yet exact inferences result. This method is applicable in randomized experiments.

The method was then extended to observational studies in two cases: with and without hidden biases. If there are no hidden biases, then treatment assignment probabilities are functions—typically unknown functions—of observed covariates. With overt biases but no hidden biases, inferences about treatment effects are possible after appropriate adjustments that reflect the relationship between treatment assignment and observed covariates. In the absence of hidden biases, estimates, tests and confidence intervals were developed in Sections 3 and 4. There is hidden bias if treatment assignment probabilities depend on both observed covariates and a relevant unobserved covariate. A sensitivity analysis displays how inferences might be altered by hidden biases of various magnitudes, as discussed and illustrated in Section 5. The same method works with an instrumental variable, in which the assigned treatment and the delivered treatment are not always the same (Section 6).

ACKNOWLEDGMENTS

This work was supported by grant SES-00-04205 from the Methodology, Measurement and Statistics Program and the Statistics and Probability Program of the U.S. National Science Foundation. The hospitality

and support of the Center for Advanced Study in the Behavioral Sciences are gratefully acknowledged.

REFERENCES

ADICHIE, J. N. (1978). Rank tests of sub-hypotheses in the general linear regression. *Ann. Statist.* **6** 1012–1026.

ADICHIE, J. N. (1984). Rank tests in linear models. In *Handbook of Statistics* **4**: *Nonparametric Methods* (P. R. Krishnaiah and P. K. Sen, eds.) 229–257. North-Holland, New York.

ANGRIST, J., IMBENS, G. and RUBIN, D. (1996). Identification of causal effects using instrumental variables (with discussion). *J. Amer. Statist. Assoc.* **91** 444–472.

BARNARD, J., FRANGAKIS, C., HILL, J. and RUBIN, D. B. (1999). School choice in NY City: A Bayesian analysis of an imperfect randomized experiment. Unpublished manuscript.

BERGSTRALH, E. J., KOSANKE, J. L. and JACOBSEN, S. J. (1996). Software for optimal matching in observational studies. *Epidemiology* **7** 331–332. Available at http://www.mayo.edu/hsr/sasmac/match.sas.

BERK, R. A. and DE LEEUW, J. (1999). An evaluation of California's inmate classification system using a generalized regression discontinuity design. *J. Amer. Statist. Assoc.* **94** 1045–1052.

BOX, G. E. P. and GUTTMAN, I. (1966). Some aspects of randomization. *J. Roy. Statist. Soc. Ser. B* **28** 543–558.

CARD, D. and KRUEGER, A. (1994). Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania. *American Economic Review* **84** 772–793.

CARD, D. and KRUEGER, A. (1995). *Myth and Measurement*: *The New Economics of the Minimum Wage*. Princeton Univ. Press.

CLEVELAND, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *J. Amer. Statist. Assoc.* **74** 829–836.

COCHRAN, W. G. (1965). The planning of observational studies of human populations (with discussion). *J. Roy. Statist. Soc. Ser. A* **128** 234–266.

COHN, P., HARRIS, P., BARRY, W., ROSATI, R., ROSENBAUM, P. R. and WATERNAUX, C. (1981). Prognostic importance of anginal symptoms in angiographically defined coronary artery disease. *American J. Cardiology* **47** 233–237.

COPAS, J. B. and LI, H. G. (1997). Inference for non-random samples (with discussion). *J. Roy. Statist. Soc. Ser. B* **59** 55–95.

CORNFIELD, J., HAENSZEL, W., HAMMOND, E., LILIENFELD, A., SHIMKIN, M. and WYNDER, E. (1959). Smoking and lung cancer: Recent evidence and a discussion of some questions. *J. National Cancer Institute* **22** 173–203.

COX, D. R. (1956). A note on weighted randomization. *Ann. Math. Statist.* **27** 1144–1151.

COX, D. R. (1958). The interpretation of the effects of non-additivity in the Latin square. *Biometrika* **45** 69–73.

COX, D. R. (1970). *The Analysis of Binary Data*. Methuen, London.

DAVIDSON, R. and MACKINNON, J. G. (1993). *Estimation and Inference in Econometrics.* Oxford Univ. Press.

DEHEJIA, R. H. and WAHBA, S. (1999). Causal effects in non-experimental studies: Reevaluating the evaluation of training programs. *J. Amer. Statist. Assoc.* **94** 1053–1062.

FISHER, R. A. (1935). *The Design of Experiments*. Oliver and Boyd, Edinburgh.

FRANGAKIS, C. E. and RUBIN, D. B. (1999). Addressing complications of intention-to-treat analysis in the combined presence of all-or-none treatment-noncompliance and subsequent missing outcomes. *Biometrika* **86** 365–379.

GABRIEL, K. R. and HALL, W. J. (1983). Rerandomization inference on regression and shift effects: Computationally feasible methods. *J. Amer. Statist. Assoc.* **78** 827–836.

GAIL, M. H., TAN, W. Y. and PIANTADOSI, S. (1988). Tests for no treatment effect in randomized clinical trials. *Biometrika* **75** 57–64.

GASTWIRTH, J. L. (1992). Methods for assessing the sensitivity of statistical comparisons used in Title VII cases to omitted variables. *Jurimetrics* **33** 19–34.

GASTWIRTH, J. L., KRIEGER, A. M. and ROSENBAUM, P. R. (1998). Dual and simultaneous sensitivity analysis for matched pairs. *Biometrika* **85** 907–920.

GASTWIRTH, J. L., KRIEGER, A. M. and ROSENBAUM, P. R. (2000). Asymptotic separability in sensitivity analysis. *J. Roy. Statist. Soc. Ser. B* **62** 545–555.

GREENHOUSE, S. (1982). Jerome Cornfield's contributions to epidemiology. *Biometrics Suppl.* 33–45.

GU, X. S. and ROSENBAUM, P. R. (1993). Comparison of multivariate matching methods: Structures, distances and algorithms. *J. Comput. Graph. Statist.* **2** 405–420.

HAJEK, J., SIDAK, Z. and SEN, P. K. (1999). *Theory of Rank Tests*, 2nd ed. Academic Press, New York.

HARVILLE, D. A. (1975). Experimental randomization: Who needs it? *Amer. Statist.* **29** 27–31.

HODGES, J. L. and LEHMANN, E. L. (1962). Rank methods for combination of independent experiments in the analysis of variance. *Ann. Math. Statist.* **33** 482–497.

HODGES, J. L. and LEHMANN, E. L. (1963). Estimates of location based on rank tests. *Ann. Math. Statist.* **34** 598–611.

HOOPER, P. M. (1989). Experimental randomization and the validity of normal-theory inference. *J. Amer. Statist. Assoc.* **84** 576–586.

HUBER, P. J. (1981). *Robust Statistics*. Wiley, New York.

JAECKEL, L. A. (1972). Estimating regression coefficients by minimizing the dispersion of the residuals. *Ann. Math. Statist.* **43** 1449–1458.

JICK, H., MIETTINEN, O. S., NEFF, R. K., SHAPIRO, S., HEINONEN, O. P., and SLONE, D. (1973). Coffee and myocardial infarction. *New England J. Medicine* **289** 63–67.

JURECKOVA, J. (1971). Nonparametric estimate of regression coefficients. *Ann. Math. Statist.* **42** 1328–1338.

KEMPTHORNE, O. (1952). *The Design and Analysis of Experiments*. Wiley, New York. [Reprinted (1973). Krieger, Malabar, FL.]

KEMPTHORNE, O. (1955). The randomization theory of experimental inference. *J. Amer. Statist. Assoc.* **50** 946–967.

KOUL, H. L. (1970). A class of ADF tests of subhypotheses in the multiple linear regression. *Ann. Math. Statist.* **41** 1273–1281.

KRAFT, C. H. and VAN EEDEN, C. (1972). Linearized rank estimates and signed-rank estimates for the general linear hypothesis. *Ann. Math. Statist.* **43** 42–57.

LEHMANN, E. L. (1963). Nonparametric confidence intervals for a shift parameter. *Ann. Math. Statist.* **34** 1507–1512.

LEHMANN, E. L. (1986). *Testing Statistical Hypotheses*, 2nd ed. Wiley, New York.

LEHMANN, E. L. (1998). *Nonparametrics*: *Statistical Methods Based on Ranks*, revised first ed. Prentice-Hall, Upper Saddle River, NJ.

LI, Y. P., PROPERT, K. J. and ROSENBAUM, P. R. (2001). Balanced risk set matching. *J. Amer. Statist. Assoc.* **96** 870–882.

LIN, D. Y., PSATY, B. M. and KRONMAL, R. A. (1998). Assessing the sensitivity of regression results to unmeasured confounders in observational studies. *Biometrics* **54** 948–963.

MANSKI, C. (1990). Nonparametric bounds on treatment effects. *American Economic Review* 319–323.

MANSKI, C. (1995). *Identification Problems in the Social Sciences*. Harvard Univ. Press, Cambridge, MA.

MCKEAN, J. W. and HETTMANSPERGER, T. P. (1978). A robust analysis of the general linear model based on one-step *R*-estimates. *Biometrika* **65** 571–579.

MING, K. and ROSENBAUM, P. R. (2000). Substantial gains in bias reduction from matching with a variable number of controls. *Biometrics* **56** 118–124.

MING, K. and ROSENBAUM, P. R. (2001). A note on optimal matching with variable controls using the assignment algorithm. *J. Comput. Graph. Statist.* **10** 455–463.

MOSES, L. E. (1965). Confidence limits from rank tests. *Technometrics* **7** 257–260.

NEYMAN, J. (1923). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Roczniki Nauk Rolniczych* **X** 1–51 (in Polish). [English transl. *Statistical Science* **5** (1990) 463–480, with discussion].

PURI, M. L. and SEN, P. K. (1969). Analysis of covariance based on general rank scores. *Ann. Math. Statist.* **40** 610–618.

QUADE, D. (1967). Rank analysis of covariance. *J. Amer. Statist. Assoc.* **62** 1187–1200.

RAZ, J. (1990). Testing for no effect when estimating a smooth function by nonparametric regression: A randomization approach. *J. Amer. Statist. Assoc.* **85** 132–138, 1182.

ROBINS, J. M., MARK, S. D. and NEWEY, W. K. (1992). Estimating exposure effects by modeling the expectation of exposure conditional on confounders. *Biometrics* **48** 479–495.

ROBINS, J. M. and RITOV, Y. (1997). Toward a curse of dimensionality appropriate asymptotic theory for semi-parametric models. *Statistics in Medicine* **16** 285–319.

ROBINSON, J. (1973a). The analysis of covariance under a randomization model. *J. Roy. Statist. Soc. Ser. B* **35** 368–376.

ROBINSON, J. (1973b). The large sample power of permutation tests for randomization models. *Ann. Statist.* **1** 291–296.

ROSENBAUM, P. R. (1984). Conditional permutation tests and the propensity score in observational studies. *J. Amer. Statist. Assoc.* **79** 565–574.

ROSENBAUM, P. R. (1986). Dropping out of high school in the United States: An observational study. *J. Educational Statistics* **11** 207–224.

ROSENBAUM, P. R. (1987). Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika* **74** 13–26, **75** 396.

ROSENBAUM, P. R. (1989). Optimal matching for observational studies. *J. Amer. Statist. Assoc.* **84** 1024–1032.

ROSENBAUM, P. R. (1991a). A characterization of optimal designs for observational studies. *J. Roy. Statist. Soc. Ser. B* **53** 597–610.

ROSENBAUM, P. R. (1991b). Discussing hidden bias in observational studies. *Annals of Internal Medicine* **115** 901–905.

ROSENBAUM, P. R. (1993). Hodges–Lehmann point estimates of treatment effect in observational studies. *J. Amer. Statist. Assoc.* **88** 1250–1253.

ROSENBAUM, P. R. (1995). *Observational Studies*. Springer, New York.

ROSENBAUM, P. R. (1996). Comment on "Identification of causal effects using instrumental variables," by J. Angrist, G. Imbens and D. B. Rubin. *J. Amer. Statist. Assoc.* **91** 465–468.

ROSENBAUM, P. R. (1999a). Reduced sensitivity to hidden bias at upper quantiles in observational studies with dilated treatment effects. *Biometrics* **55** 560–564.

ROSENBAUM, P. R. (1999b). Using combined quantile averages in matched observational studies. *Appl. Statist.* **48** 63–78.

ROSENBAUM, P. R. (1999c). Choice as an alternative to control in observational studies (with discussion). *Statist. Sci.* **14** 259–304.

ROSENBAUM, P. R. and RUBIN, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70** 41–55.

ROSENBAUM, P. R. and RUBIN, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *Amer. Statist.* **39** 33–38.

ROSENBAUM, P. R. and SILBER, J. H. (2001). Matching and thick description in an observational study of mortality after surgery. *Biostatistics* **2** 217–232.

RUBIN, D. B. (1973). The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics* **29** 185–203.

RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educational Psychology* **66** 688–701.

RUBIN, D. B. (1977). Assignment to treatment group on the basis of a covariate. *J. Educational Statistics* **2** 1–26.

RUBIN, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Ann. Statist.* **6** 34–58.

RUBIN, D. B. (1979). Using multivariate matched sampling and regression adjustment to control bias in observational studies. *J. Amer. Statist. Assoc.* **74** 318–328.

SEBER, G. A. F. (1977). *Linear Regression Analysis*. Wiley, New York.

SHEINER, L. B. and RUBIN, D. B. (1995). Intention-to-treat analysis and the goals of clinical trials. *Clinical Pharmacology and Therapeutics* **57** 6–15.

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.

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).

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

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).

from the idealized template for causal inference, the controlled randomized experiment. Rosenbaum draws on the strengths of existing approaches (matching and instrumental variables), but allows for fewer distributional assumptions. In addition, he demonstrates straightforward approaches to sensitivity analysis for each method.

My comments will primarily, although not exclusively, pertain to Rosenbaum's estimation strategies in the context of matched pairs, as that is the area with which I am more familiar.

2. STRENGTHS

There are considerable strengths to the methods discussed. I will highlight a few that I view as among the most important.

Within the context of propensity score matching, the distributional assumptions for treatment effect estimates are difficult to derive analytically. For this reason, reliable estimators for the variance of these treatment effect estimates in practice have yet to be defined. Rosenbaum's techniques may provide a robust approach both for estimating average treatment effects as well as for forming confidence intervals in certain observational settings.

With regard to instrumental variables estimation, Rosenbaum's techniques do not require the assumption that the instrument and "treatment" of interest are strongly correlated. Relative strength of the instrument is reflected in the length of the corresponding confidence interval. The issue of the impact on inference of so-called "weak instruments" (see, for example, Nelson and Startz, 1990b, a; Bound, Jaeger and Baker, 1995; Staiger and Stock, 1997) and the inherent complications in judging instrument strength and adjusting confidence intervals accordingly can thus be avoided altogether. His approach also provides greater evidence about model misspecification than might be typical.

The emphasis on sensitivity analysis is much needed. So many of the current debates about the efficacy of competing methodologies in causal inference are at heart debates about the adequacy of the ignorability assumption (Rubin, 1978; Little and Rubin, 1987). Propensity score matching, for instance, generally achieves what it purports to; given sufficient overlap in the distributions of the observed covariates, it balances these covariates across treatment groups. So examples where matching is used and yet bias remains should point to a failure to control for all confounding covariates—that is, violation of the ignorability assumption—rather than a "failure" of the approach. Given the seeming difficulty in satisfying ignorability in many observational settings then, analyses that explore the sensitivity of conclusions to unobserved covariates should ideally be included as an integral part of all causal analyses. Achieving this ideal can be greatly facilitated by the existence of straightforward approaches such as those described by Rosenbaum.

3. ADDITIVE TREATMENT EFFECTS AND MATCHING

The focus on additive treatment effects within the context of matching applications (for example in situations where there is no obvious instrument) is potentially problematic. If the response surfaces across treatments are not parallel this assumption will not hold. Yet the existence of nonparallel, and in particular nonparallel and nonlinear, response surfaces is one of the strongest motivations for the use of matching techniques. If the response surfaces are linear why wouldn't standard regression work just as well for covariance adjustment, even perhaps more efficiently than the techniques proposed? If this is the case, it is only the scenario where the response surfaces are nonlinear and parallel (a seemingly unlikely combination) when there seems to be a more obvious advantage for the approach discussed in Section 4.5 over simpler techniques such as linear regression.

In the absence of the arguably unrealistic assumption of additive treatment effects, however, the choice of the best summary of the individual treatment effects (e.g., average treatment effect and median treatment effect) becomes much more messy (for an interesting discussion of the inadequacy of average treatment effects in a policy context, see Angrist and Dehejia, 2001). However, given that an appropriate statistic is chosen, it appears that permutation tests could be performed for any such choice. Although some statistics might require calculation of the exact permutation distribution, with current computing technology Monte Carlo estimates of these distributions should be fairly trivial to perform.

These concerns point to the more general need for a closer examination of the types of empirical situations within which we might expect to see gains from using techniques such as the ones described in this paper. Optimally, such an exploration would rely either on empirical data in a context where the "true" answer is known (e.g., see LaLonde, 1986; Friedlander and

Robins, 1995; Dehejia and Wahba, 1999; Heckman, Ichimura and Todd, 1997) or on simulations that are well calibrated to real data. The disadvantages of the former strategy are the complications created by missing data typically present in real studies and the difficulty in using an effect estimate, about which we also have uncertainty, as a benchmark. The simulation option can be tricky to perform in ways that do not bias the simulations results toward a particular technique. Either, however, should help us to gain better insight about when competing techniques perform well or poorly in practice.

4. ADDITIONAL COMMENTS

One small weakness in presentation (that admittedly may have been beyond the scope of the paper given its breadth) was that both examples had very small numbers of covariates. In my opinion, the most convincing examples of observational studies (in the absence of a truly convincing instrument) rely on far greater numbers of covariates to support the necessary ignorability assumptions. It was unclear to me how the current conclusions would change if many more covariates were included.

In addition, a remaining concern about the confidence intervals estimated in the matched pair setting via the Rosenbaum approach is that they appear to ignore the variability inherent in creating the matched pairs. The matched pairs are taken as a given. In reality which control units are picked can be strongly influenced by the properties of the covariate distributions of the comparison units that are not chosen as matches. It is unclear, however, how this uncertainty would be reflected in the interval estimates described.

5. CONCLUSION

Rosenbaum describes a rich class of techniques applicable to an important area, causal inference in observational settings, still sorely in need of robust approaches to inference. These methods should at the very least be considered a strong starting point for exploration of new approaches that are less dependent on distributional assumptions (as these are) but in addition can accommodate more realistic complications such as non-additive treatment effects in matched pairs studies.

Comment

James M. Robins

1. INTRODUCTION

I am grateful to the Associate Editor, Alicia Carriquiry, for the opportunity to discuss Paul Rosenbaum's paper. This technically flawless and beautifully motivated paper touches either directly or indirectly upon many major issues in causal inference: superpopulation model-based versus observed study population randomization-based inference; the risks versus the benefits of assuming a constant additive treatment effect; how to find optimal covariance-adjustment procedures for randomization-based inference; the assumptions under which bias due to unmeasured confounding can be corrected by an instrumental variable; the relationship between different sensitivity analysis methodologies; and, finally, the scientific role of sensitivity analysis in the interpretation of observational studies. Rosenbaum has thought deeply about many of these issues, both in the paper under discussion and in previous papers. In my discussion I shall consider each of these issues, because I find them important and often interrelated. Here is a summary of my major points. I argue in Section 4 that the model-based locally efficient semiparametric estimators can provide near optimal covariance adjustment even for randomization-based inference. Second, in Section 4.1, I derive a formal equivalence between Rosenbaum's sensitivity analysis methodology and a special case of a methodology described in Section 3 below that was introduced in Robins (1997, 1998) and Robins, Scharfstein and Rotnitzky (2000).

Third and perhaps most importantly, in Section 5, I argue that Rosenbaum's approach to sensitivity analysis, although logically flawless and mathematically elegant, may be scientifically useless. The problem is as

James M. Robins is Professor, Department of Epidemiology, Harvard University, Boston, Massachusetts (e-mail: robins@hsph.harvard.edu).

follows. Rosenbaum's sensitivity analysis model will only be useful if experts can provide a plausible and logically coherent range for the value of the sensitivity parameter Γ^*, that measures the potential magnitude of hidden bias. Define a measure of hidden bias (i.e., of confounding by unmeasured factors) to be "paradoxical" if its magnitude can increase as we decrease the amount of hidden bias by measuring some of the unmeasured confounders. In Section 5, I prove that Rosenbaum's Γ^* is a "paradoxical" measure. A sensitivity analysis methodology based on a "paradoxical" measure of hidden bias may be scientifically useless because, without prolonged and careful training, users of the methodology may reach grossly misleading, logically incoherent conclusions. I need to emphasize that the same arguments I use here against Rosenbaum's methods apply equally to certain methods based on paradoxical measures developed by myself and colleagues Andrea Rotnitzky and Daniel Scharfstein. Indeed I am just continuing here the criticism in Scharfstein, Rotnitzky and Robins (1999, Section 7.2.3) of our own methods, partly in the hope that Rosenbaum, in his rejoinder, can succeed in convincing me that the aforementioned training is not so difficult as to render both his methods and ours scientifically useless. In Sections 3.2 and 5, I suggest an alternative sensitivity methodology described in Robins, Scharfstein and Rotnitzky (2000) and Scharfstein, Rotnitzky and Robins (1999) that avoids paradoxical measures of hidden bias.

Finally, even in the absence of hidden bias due to unmeasured factors, it is my opinion that whatever Rosenbaum believes may be gained by eliminating the need for a hypothetical superpopulation is offset by Rosenbaum's assumption that individual outcomes are deterministic and that an additive treatment effect model holds. Furthermore, even when there exist interval estimators for the effect of treatment in the observed study population that are valid without the assumption of additivity, arguments given in Section 2 suggest that interval estimators for the treatment effect in the superpopulation are more relevant for medical decision-making. Finally, my alternative sensitivity analysis methodology requires a superpopulation model to be well defined.

2. SUPERPOPULATION VERSUS OBSERVED STUDY POPULATION INFERENCE

Consider an observational or randomized study where there are n subjects, $j = 1, \ldots, n$, with observed data $\{(r_j, Z_j, x_j),\ j = 1, \ldots, n\}$, where r denotes response, Z denotes treatment and x denotes a vector of pretreatment covariates. We shall also assume there exist counterfactual (potential outcome) data $r_{\cdot j} = (r_{zj}; z \in \mathcal{Z})$, where $\mathcal{Z}$ is the set of possible treatments and r_{zj} is subject j's response were treatment z taken. The observed and counterfactual data are linked by the consistency assumption that $r_j = r_{Z_j j}$. We have generalized Rosenbaum by allowing for nondichotomous Z. Following Rosenbaum, we take $(r_{\cdot j}, x_j)$ to be nonrandom and $\mathbf{Z} = (Z_1, \ldots, Z_n)'$ to be random with support $\mathcal{Z}$. Let $\pi_j(z_j)$ denote the marginal density of Z_j. Rosenbaum considers both the independence model in which the density of $\mathbf{Z}$ is $\boldsymbol{\pi}(\mathbf{Z}) = \prod_j \pi_j(Z_j)$ and the *always m-treated* model in which we condition the independence model on the event that exactly m subjects were treated. Let $z = 0$ denote a baseline treatment level. The average effect of treatment z in the finite study population is $\tau_n^*(z) = n^{-1} \sum_j [r_{zj} - r_{0j}]$. Rosenbaum's parameter of interest is $\tau_n^* \equiv \tau_n^*(1)$. When we wish to emphasize the dependence of $\tau_n^*(z)$ and τ_n^* on $\mathbf{r}_\cdot = \{r_{\cdot j},\ j = 1, \ldots, n\}$, we will write $\tau_n^*(z, \mathbf{r}_\cdot)$ and $\tau_n^*(\mathbf{r}_\cdot)$.

An investigator usually wishes to generalize his or her findings from the observed study population to a larger similar population. That is, he or she is interested in the external validity of the inferences to be drawn from the data. For example, an investigator who considers recommending a medical intervention must hope the observed study population is representative of the population of the potential recipients of that intervention. The simplest model is to consider the study population as a random sample of a larger population (Lehmann, 1975). That is, we regard the n study subjects as randomly sampled without replacement from a large superpopulation of N subjects of potential recipients of treatment. Then, in the limit $N \to \infty$ and $n/N \to 0$, and we can model the data on the n study subjects as independent and identically distributed. That is, we have $(R_{\cdot j} = \{R_{zj}, z \in \mathcal{Z}\}, Z_j, X_j),\ j = 1, \ldots, n$, i.i.d. copies of the random vector $(R_\cdot, Z, X)$ drawn from the superpopulation law $F_{R_\cdot, Z, X}$. Let $E(\ )$ denote expectation under $F_{R_\cdot, Z, X}$. Then $\tau^*(z) = E[R_z - R_0]$ is the average causal effect of z in the superpopulation. Under squared error loss and assuming higher values of r are desirable and no treatment–covariate interaction, the as yet untreated $N - n$ members should receive the treatment z_{opt} that maximizes $\tau^*(z)$. Thus one goal is to estimate the superpopulation dose–response function $\tau^*(z)$. Of course, an investigator should recognize the limitations

of the above superpopulation model when (1) as often occurs in randomized clinical trials, he or she believes the dose–response function $\tau(z)$ of potential recipients of the treatment differs from that of the observed study population to an extent unaccountable by sampling variability or (2) as happens rarely, the size of the pool of potential recipients is not much larger than the size of the study population. In situation (1), the superpopulation confidence interval still serves as a useful informal lower bound on uncertainty concerning $\tau^*(z)$ in the population of potential recipients.

To study the relationship between our i.i.d. superpopulation model and Rosenbaum's observed study population model, we will obtain the latter from the former by conditioning as follows. Suppose we assume that $\pi_j(z_j)$ depends on subject j only through $(x_j, r_{\cdot j})$, so that $\pi_j(z_j) = \pi(z_j, x_j, r_{\cdot j})$. This entails no restriction if there are no ties. Rosenbaum considers the distribution of estimators $\widehat{\tau}$ of $\tau_n^* \equiv \tau_n^*(z)$ under $\boldsymbol{\pi}(\mathbf{Z}) = \prod_j \pi_j(Z_j)$. We can precisely reproduce this inference by taking $\pi(z, x, r_{\cdot})$ as our superpopulation conditional density $f(z|x, r_{\cdot})$ for Z given $(X, R_{\cdot})$ and considering the conditional distribution of $\widehat{\tau}$ given $\mathbf{R}_{\cdot} \equiv \{R_{\cdot j},\ j = 1, \ldots, n\} = \mathbf{r}_{\cdot}$ and $\mathbf{X} = \mathbf{x} \equiv \{x_j, j = 1, \ldots, n\}$. In the always m-treated sampling model, we additionally condition on the event $\sum_j Z_j = m$. Note the observed study population causal effect $\tau_n^*(z) = n^{-1}\sum_j [r_{zj} - r_{0j}]$ is fixed given the conditioning event $\mathbf{R}_{\cdot} = \mathbf{r}_{\cdot}$. Statements about consistency and asymptotic normality will refer to an asymptotics in which $n \to \infty$, $N \to \infty$ and $n/N \to 0$.

We now have the machinery required to compare superpopulation and finite population inference. As in Rosenbaum's Section 2.2, assume the always m-treated model with Z dichotomous and treatment completely randomized so $\pi(1, x, r_{\cdot j})$ is a constant π and data on X are not recorded. Unlike Rosenbaum, suppose we estimate the average treatment effect by the difference $\widehat{\tau} = \widehat{P}_1 - \widehat{P}_0$ between the sample mean $\widehat{P}_1 = \sum_j Z_j R_j / m$ in the treated and the sample mean $\widehat{P}_0 = \sum_j R_j(1 - Z_j)/(n - m)$ in the untreated. Consider the variance decomposition

$$\operatorname{var}\left(\widehat{\tau}\,\middle|\,\sum_j Z_j = m\right) = E\left\{\operatorname{var}\left[\widehat{\tau}\,\middle|\,\mathbf{R}_{\cdot}, \sum_j Z_j = m\right]\right\} + \operatorname{var}\left\{E\left[\widehat{\tau}\,\middle|\,\mathbf{R}_{\cdot}, \sum_j Z_j = m\right]\right\}.$$

Now $\widehat{\tau}$ is unbiased for finite study population treatment effect $\tau_n^* = \tau_n^*(\mathbf{r}_{\cdot})$ under the *randomization* distribution $f(\mathbf{Z}|\mathbf{R}_{\cdot} = \mathbf{r}_{\cdot}, \sum_j Z_j = m)$, that is, $E[\widehat{\tau}|\mathbf{R}_{\cdot} = \mathbf{r}_{\cdot}, \sum_j Z_j = m] = \tau_n^*(\mathbf{r}_{\cdot})$. Thus we see that the superpopulation variance $\operatorname{var}(\widehat{\tau}|\sum_j Z_j = m)$ of $\widehat{\tau}$ will be greater than the average finite sample variance $E\{\operatorname{var}[\widehat{\tau}|\mathbf{R}_{\cdot}, \sum_j Z_j = m]\}$ unless $\tau_n^*(\mathbf{r}_{\cdot})$ is the same for all $\mathbf{r}_{\cdot}$ in a set of probability 1. This can occur only if there is an additive treatment effect, that is, in the superpopulation $R_{1j} - R_{0j}$ is a constant τ^* with probability 1. However, if R is a Bernoulli outcome and $0 < \tau_n^*(\mathbf{r}_{\cdot}) < 1$, then the hypothesis of additive treatment effects cannot hold (as the only values of τ_n^* consistent with additivity are 0, -1 and 1). In this setting, the superpopulation variance $\operatorname{var}(\widehat{\tau}|\sum_j Z_j = m)$ of $\widehat{\tau}$ is the usual binomial variance $(p_0(1 - p_0))/(n - m) + (p_1(1 - p_1))/m$, where $p_0 = E[R_0]$ and $p_1 = E[R_1]$. Robins (1988) and Copas (1973) showed that the randomization variance $\operatorname{var}[\widehat{\tau}|\mathbf{R}_{\cdot} = \mathbf{r}_{\cdot}, \sum_j Z_j = m]$ is

$$\frac{p_{n0}(1 - p_{n0})}{n - m} + \frac{p_{n1}(1 - p_{n1})}{m} - s(\kappa),$$

where $p_{nk} = n^{-1}\sum_j r_{kj}$ and $s(\kappa) \equiv -(p_{0n} + p_{1n} - 2p_{0n}p_{1n} - p_{1n}(1 - p_{1n}) - p_{0n}(1 - p_{0n}) - \kappa)$, where κ is not identified and can lie anywhere in the range $|p_{1n} - p_{0n}| \le \kappa \le \min(p_{0n} + p_{1n}, 2 - p_{0n} - p_{1n})$. Furthermore, $s(\kappa) \ge 0$ with equality only when (i) $\kappa = 0$ and $p_{1n} - p_{0n} = 0$ or (ii) $|p_{1n} - p_{0n}| = 1$. Let $\widehat{s}(\kappa)$ be $s(\kappa)$ with $\widehat{P}_1$ and $\widehat{P}_0$ substituted for p_{1n} and p_{0n}. Robins (1988) showed that the interval estimator

$$C(0.95) = (\widehat{P}_1 - \widehat{P}_0) \pm 1.96\left[\frac{\widehat{P}_1(1 - \widehat{P}_1)}{m} + \frac{\widehat{P}_0(1 - \widehat{P}_0)}{n - m} - \widehat{s}(|\widehat{P}_1 - \widehat{P}_0|)\right]$$

is a large sample conservative 95% confidence interval for $\tau_n^*(\mathbf{r}_{\cdot})$ under the randomization distribution as $n \to \infty$, with length less than that of the usual binomial interval with probability approaching 1 whenever $\tau_n^*(\mathbf{r}_{\cdot}) \notin \{0, -1, 1\}$. Note that $|\widehat{P}_1 - \widehat{P}_0|$ is consistent for the minimum possible value of κ. For example, when $m = 100$, $n - m = 100$, $\widehat{P}_1 = 40/100$ and $\widehat{P}_0 = 15/100$, the usual "binomial 95% interval" is 0.250 ± 0.118, while the conservative large sample 95% confidence interval for τ_n^* under the randomization distribution is 0.250 ± 0.102. Thus, for large n, the usual 95% binomial interval is approximately 10% wider than it needs to be to cover τ_n^* at its nominal rate under the randomization distribution. However, the usual 95% binomial interval is the smallest possible interval that will cover the superpopulation parameter $\tau^* = \tau^*(1)$ at its nominal rate in large samples.

Robins (1988) also showed that the interval $(\widehat{P}_1 - \widehat{P}_0) \pm 1.96[(\widehat{P}_0(1-\widehat{P}_0))/m + (\widehat{P}_0(1-\widehat{P}_0))/(n-m)]$, which differs from the usual binomial interval by having $\widehat{P}_0(1-\widehat{P}_0)$ in the numerator of both terms, is a large sample 95% confidence interval for the average causal effect $\sum_j Z_j(r_{1j} - r_{0j})/m$ of treatment on the treated under the randomization distribution. Recently Rosenbaum (2001b) has considered exact inference for this random quantity under the additional assumption that treatment does not both help some subjects and harm others. Finally, a caveat is in order. Although it appears quite advantageous to use $C(0.95)$ in lieu of the standard 95% binomial interval, in fact, the increased precision that comes from using $C(0.95)$ depends wholly on the nonidentifiable assumption that outcomes are deterministic. To see why, note there is no data evidence to contradict the assumption of a stochastic counterfactual model in which r_{1j} and r_{0j} are the outcomes of Bernoulli experiments with counterfactual success probabilities p_{1j} and p_{0j}, respectively. In that case, τ_n^* is redefined as $n^{-1}\sum_j p_{1j} - p_{0j}$ and, even in large samples, $C(0.95)$ is not guaranteed to cover τ_n^* at a rate greater than or equal to 0.95 if any of the p_{1j} and p_{0j} are not exactly equal to either 0 or 1.

2.1 Implications of Possible Nonadditivity for Continuous Responses

In this section, we consider a randomized drug study without data on covariates X and with a dichotomous treatment given independently to each subject with probability π. Hereafter we assume continuous responses in the sense that there are no ties either in the r_{1j} or in the r_{0j} among the n study subjects. As in the binomial case, the additivity assumption $r_{1j} - r_{0j} = \tau^*$ for all subjects j has testable consequences. Specifically, it implies a shift model under which the empirical marginal distribution of the r_{1j} differs from that of the r_{0j} by a shift τ^*. Without going into details, a test of the shift model could, for example, be based on the difference from the zero vector of test statistics $D(\widehat{\tau})$, where $D(\tau) = n^{-1}\sum_j d(r_j - \tau Z_j)(Z_j - \pi)$ and $d(\cdot)$ is a possibly nonlinear vector function chosen by the analyst and $\widehat{\tau}$ is, say, the Hodges–Lehmann estimate of τ^*. Suppose a test of the shift model does not reject; indeed, we shall suppose the shift model is actually true with shift parameter τ^*. Even then, in many settings, the hypothesis of a constant treatment effect is rather biologically implausible due to between-individual differences in bioabsorption, metabolism and so forth. We now consider the consequence of this fact for randomization-based inference. Let $E_{\mathbf{r}.}$ denote conditional expectations given $\mathbf{r}.$. To begin, consider a small example with $n = 2$, $(r_{01}, r_{11}) = (4, 4)$ and $(r_{02}, r_{12}) = (2, 6)$ so the shift parameter τ^* is $6-4 = 4-2 = 2$, but additivity fails. Then the Wilcoxon test will give the wrong level, because, unlike the additive case, (Z_1, Z_2) is not independent of the residuals $(r_1 - \tau^* Z_1, r_2 - \tau^* Z_2)$. [For example, if $Z_1 = Z_2 = 1$, the vector of observed residuals $(r_1 - \tau^* Z_1, r_2 - \tau^* Z_2)$ is $(4-2, 6-2) = (2, 4)$, so the unconditional randomization probability that $Z_1 = Z_2 = 1$ is π^2, but the conditional probability given $(r_1 - \tau^* Z_1, r_2 - \tau^* Z_2) = (2, 4)$ is 1, since $Z_1 = 1$, $Z_2 = 0$ implies $(r_1 - \tau^* Z_1, r_2 - \tau^* Z_2) = (2, 6)$, $Z_1 = 0$, $Z_2 = 1$ implies $(r_1 - \tau^* Z_1, r_2 - \tau^* Z_2) = (4, 4)$ and $Z_1 = Z_2 = 0$ implies $(r_1 - \tau^* Z_1, r_2 - \tau^* Z_2) = (4, 2)$.]

Nonetheless, even though exact inference fails, following Neyman (1935), under the shift model with d taking values in R^1, $D(\tau^*)$ has mean zero, variance $n^{-1}\sum_j E_{\mathbf{r}.}[D_j(\tau^*)^2] - n^{-1}\sum_j \{E_{\mathbf{r}.}[D_j(\tau^*)]\}^2$ and is asymptotically normal. These results follow from the fact that $D(\tau^*)$ is the sum of independent random variables and, under the shift model, the values of $\sum_j E_{\mathbf{r}.}[D_j(\tau^*)^2]$ and of $\sum_j E_{\mathbf{r}.}[D_j(\tau^*)] = 0$ do not depend on whether the additive effect submodel $r_{1j} - r_{0j} = \tau^*$ holds, because these statistics are invariant to permutations of the $(r_j - \tau^* Z_j)$ within each level of z; however $\sum_j \{E_{\mathbf{r}.}[D_j(\tau^*)]\}^2$ is 0 if and only if the additive effect submodel holds. It follows that large sample 95% confidence intervals for the true shift parameter τ^* calculated under the additivity assumption will be conservative (i.e., cover at a rate greater than 95%) if additivity fails and will cover at the nominal rate if additivity holds. Thus, under the shift model, it is appropriate to assume additivity as Rosenbaum did since large sample inference for τ^* assuming additivity is appropriately conservative and additivity cannot be rejected.

Suppose now the superpopulation follows a shift model with parameter τ^*, that is, $\text{pr}(R_1 - \tau^* < t) = \text{pr}(R_0 < t)$ for all t. Then, under our superpopulation model, $Z \coprod (R_j - Z_j\tau^*)$ so $f[\mathbf{Z}|\{(R_j - Z_j\tau^*); j = 1, \ldots, n\}] = \prod_j \pi^{Z_j}(1-\pi)^{1-Z_j}$, justifying Rosenbaum's exact inference methods even when additivity fails, but now for the superpopulation parameter τ^* under hypothetical resampling from the superpopulation (Lehmann, 1975). However, this approach fails under resampling from the randomization distribution, since $Z \not\coprod (R_j - Z_j\tau^*)|\mathbf{R}.$ when the addi-

tivity submodel does not hold. In Sections 3.5 and 4, we extend the results of this section to allow for both measured and unmeasured confounders.

3. SUPERPOPULATION COVARIANCE ANALYSIS

In this section, we describe a superpopulation sensitivity analysis model introduced in Robins (1997, 1998) and Robins, Scharfstein and Rotnitzky (2000, Sections 8.1–8.5) based on a quantile–quantile function treatment effect model that includes the shift model as a special case. In Section 4 we examine its use in Rosenbaum's observed study population sampling model.

Parametrization and identification. Let $F(t|\cdot) = \text{pr}[R < t|\cdot]$ and $F_z(t|\cdot) = \text{pr}[R_z < t|\cdot]$ be the conditional cumulative distribution functions of the observed outcome R and of potential outcome R_z given $\cdot$, and assume that the R_z have continuous distribution functions. Note $F_z[t|Z = z, X = x] = F[t|Z = z, X = x]$. Let

$$h(t, x, z) = F_0^{-1}\{F_z[t|Z = z, X = x]|Z = z, X = x\}$$

be the conditional quantile–quantile function mapping quantiles of $F_z[t|Z = z, X = x]$ into those of $F_0[t|Z = z, X = x]$. Let

$$h^c(t, x, z) = F_0^{-1}\{F_z(t|X = x, Z \neq z)|X = x, Z \neq z\}$$

be defined like $h(t, x, z)$ except conditional on $Z \neq z$ rather than $Z = z$. Note $h(t, x, 0) = h^c(t, x, 0) = t$ and both functions are increasing in their first argument. Further $h(t, x, z) \equiv t$ if and only if the null hypothesis

$$F(t|Z = z, X = x) = F_0(t|Z = z, X = x) \quad \text{for all } z \in \mathcal{Z} \text{ and } x \text{ in the support of } X$$

of no effect of treatment level z compared to no treatment is true among those treated with level z. Here, the triple equal sign indicates the equality of two functions. Also $h(t, x, z) \equiv h^c(t, x, z) \equiv t$ implies the null hypothesis

$$F_z(t|X = x) = F_0(t|X = x) \quad \text{holds for all } z \in \mathcal{Z} \text{ and } x \text{ in the support of } X \tag{1}$$

of no effect of treatment at any level of X. The latter is the null hypothesis of interest as it says that treatment Z will result in exactly the same conditional response distributions. This motivates our desire to model $h(t, x, z)$. Note that were $h(t, x, z)$ identified, then $F_0(t|Z = z, X = x)$ would become identified since, by $h(t, x, z)$ a quantile–quantile function, R_0 and

$$H \equiv h(R, X, Z)$$

have the same conditional distribution given $Z = z$, $X = x$. If the nonidentifiable assumption that $R_0 = H$ with probability 1 holds, we say we have partial rank preservation. If both $h(t, x, z)$ and $h^c(t, x, z)$ were identified, $F_z(t|X = x)$ would be identified *for all z and x*, since it is easy to show that R_z and

$$H_z = RI(Z = z) + I(Z \neq z)h^{-1c}(H, X, z)$$

have the same conditional distribution given $X = x$, where $h^{-1c}(t, x, z)$ is the inverse of $h^c(t, x, z)$ with respect to its first argument.

If the nonidentifiable assumption that $H_z = R_z$ with probability 1 holds for all z, we say we have full rank preservation. If

$$h(t, x, z) = h^c(t, x, z) \tag{2}$$

are the same function, we say we have no treatment interaction (in the sense that, conditional on X, the quantile–quantile transformation required to map R_z into a random variable with the same law as R_0 is the same among those with $Z = z$ as among those with $Z \neq z$).

We can now better understand the additive and shift models. With $\{0, 1\}$ the support of Z, the shift model holds if there is no treatment interaction and $h(t, x, z) = t - \tau^* z$ for some τ^*. The additive model holds if the shift model holds and we have full rank preservation. Under the partial ignorability assumption

$$R_0 \coprod Z|X, \tag{3}$$

$h(t, x, z)$ is identified. Under the stronger ignorability assumption $R. \coprod Z|X$ (as would hold in a randomized trial), there is also no treatment interaction (but there remains no data evidence as to whether partial or full rank preservation holds). Thus when $R. \coprod Z|X$, it would be more robust to try to nonparametrically estimate $h(t, x, z) = h^c(t, x, z)$ from the data than to assume a priori a shift model, much less an additive model. In practice, due to the often high dimension of X, nonparametric estimation of $h(t, x, z)$ is not feasible and instead we assume a parametric model.

DEFINITION. A structural distribution model (SDM) specifies that $h(t, x, z) = h(t, x, z, \tau^*)$, where $h(t, x, z, \tau)$ is a known function increasing in its first argument and satisfying both $h(t, x, 0, \tau) = 0$ and $h(t, x, z, \tau) \equiv t \Leftrightarrow \tau = 0$, so $\tau = 0$ represents the null hypothesis (1).

With $Z \in \{0, 1\}$, a superpopulation shift model is equivalent to the SDM $h(t, x, z, \tau) = t - \tau z$. To be robust, it would often be wise to choose a more flexible model with τ of higher dimension that includes interactions of Z with components of X. Below we discuss estimation of τ. Before doing so, we consider identifying assumptions for $h(t, x, z)$ and $h^c(t, x, z)$ when $f[Z|X, R.] \neq f[Z|X]$. We will see that to identify $h(t, x, z)$, we need only impose assumptions (which we will then vary in a sensitivity analysis) on the nonidentifiable density $f[Z|X, R_0]$ and not on the more complicated density $f[Z|X, R.]$. Consider the model

$$f(z|x, r_0; \lambda, \gamma^*) = \frac{v(z|x; \lambda) \exp(\gamma^* q(r_0, x.z))}{\int v(z|x; \lambda) \exp(\gamma^* q(r_0, x.z))\, d\mu(z)}, \tag{4}$$

where $v(z|x; \lambda)$ is a conditional density with carrying measure $\mu(z)$ that is known except for the parameter λ, $q(r_0, x.z)$ is a known function that as described below encodes the functional form of dependence on hidden biases and γ^* is a known parameter that encodes the magnitude of hidden biases (unmeasured confounding). We will vary γ^* in a sensitivity analysis along with $q(r_0, x.z)$. Note the dependence on r_0 enters through an exponential tilt. In the absence of hidden biases (i.e., $\gamma^* = 0$), $f(z|x, r_0; \lambda, \gamma^*)$ is just $v(z|x; \lambda)$. For dichotomous Z, the model can be written

$$\text{logit pr}(Z = 1|x, r_0; \lambda, \gamma^*) = v_{\text{dic}}(x; \lambda) + \gamma^* q_{\text{dic}}(r_0, x),$$

where

$$v_{\text{dic}}(x; \lambda) = \log\{v(1|x; \lambda)/v(0|x; \lambda)\}$$

and

$$q_{\text{dic}}(r_0, x) = q(r_0, x, 1) - q(r_0, x, 0).$$

Here is some motivation for model (4). Suppose there are unmeasured variables U such that $R_0 \coprod Z|X, U$, but $R_0 \coprod Z|X$ is false because $U \not\coprod Z|X$ and $U \not\coprod R_0|X$. We say that U is an unmeasured confounder (i.e., cause of hidden bias). In that case, even under the causal null hypothesis (1), R and Z will be dependent given X. The magnitude of potential bias in estimation of $h(t, x, z)$ due to not observing U depends on U's conditional dependence with both Z and R_0. The degree of dependence of $f[Z|X, R_0] = \int f[Z|U, X] f(U|X, R_0)\, dU$ on R_0 properly weights the effect of these dependencies and is captured by γ^* and $q(r_0, x, z)$ in model (4).

EXAMPLE. A simple example of $q_{\text{dic}}(r_0, x)$ is r_0 itself. In Section 4.1, we show that Rosenbaum's sensitivity analysis methodology corresponds to the special case in which (i) $q_{\text{dic}}(r_0, x) = I(r_0 > c(x))$ or $I(r_0 < c(x))$ and (ii) we vary $c(x)$ over all possible functions at each choice of γ^*.

Robins, Scharfstein and Rotnitzky (2000, Theorem 8.2) proved that when λ is an infinite-dimensional parameter indexing all possible conditional densities $v(z|x; \lambda)$, model (4) (i) nonparametrically identifies the quantile–quantile function $h(t, x, z)$ and the density $v(z|x; \lambda)$, and yet (ii) places no restrictions on the joint distribution of the observed data. Result (i) proves that knowledge of the dependence of $f[Z|X, R_0]$ on R_0 [through γ^* and $q(r_0, x, z)$] would be sufficient to correct for bias in estimation of $h(t, x, z)$ due to unmeasured confounders. Unfortunately, result (ii) implies that, without additional assumptions, the data will not help us learn about the magnitude γ^* and functional form $q(r_0, x, z)$ of hidden biases, so our only choice is to vary γ^* and $q(r_0, x, z)$ in a sensitivity analysis. Further, result (ii) also implies that under model (4) the data cannot help us learn about $h^c(t, x, z)$ so, unless we follow Rosenbaum and impose no treatment interaction as a default choice, $h^c(t, x, z)$ too must be varied in a sensitivity analysis if we wish to estimate $F_z(t|X = x)$ for $z \neq 0$.

REMARK 1. If the quantile–quantile function $h(t, x, z)$ is known, then the conditional density $f(Z|X, R_0)$ is identified; hence if we had strong prior knowledge of the magnitude of the treatment effect $h(t, x, z)$, we could learn from the data about the degree of dependence of treatment on hidden biases. See Rosenbaum (1995, Chapter 5) for related discussion.

3.1 Estimation, Semiparametric Efficiency and Double Robustness

We now turn to estimation of a SDM under two models that restrict the dimension of λ in model (3). The first model assumes that the true value λ^* of λ is known and that $\gamma^* = 0$, so $f[Z|X, R_0] = f[Z|X] = f(Z|X; \lambda^*)$ is known by design, as would be true in a randomized trial with randomization probabilities possibly depending on X. The second model assumes $v(x|x; \lambda)$ is a given parametric model and $\lambda \in R^p$ is an unknown parameter. In analyzing observational data, this model would often be used when X is high dimensional. In the model with λ finite dimensional, γ^* and q may be identified, but only weakly. Therefore, we recommend they be treated as known rather than estimated and then varied in a sensitivity analysis.

The key to our estimation procedure is the identity $f[Z|X, R_0] = f[Z|X, H]$ mentioned above. This identity implies that under a SDM $h(t, x, z, \tau)$, $f[z|X, R_0] = f[z|X, H(\tau^*)]$, where $H(\tau) = h(R, X, Z, \tau)$. Suppose first Z is dichotomous so $\text{logit}\,\text{pr}(Z = 1|x, r_0; \lambda, \gamma^*) = v_{\text{dic}}(x; \lambda)\gamma^* q_{\text{dic}}(r_0, x)$ and λ is finite-dimensional. Let $\widehat{\tau}$ be the value of τ that makes the maximum likelihood estimator of the "artificial parameter vector" θ equal to zero when maximizing the logistic "likelihood" $\prod_j \mathcal{L}_j(\lambda, \theta, \tau)$ over (λ, θ) with τ and γ^* held fixed, where $\mathcal{L}(\lambda, \theta, \tau) = \Pi^Z(1-\Pi)^{1-Z}$ with $\Pi = \text{expit}(v_{\text{dic}}(X; \lambda) + \gamma^* q_{\text{dic}}(H(\tau), X) + \theta' d_{\text{dic}}(H(\tau), X))$, $\text{expit}(x) = 1/(1 + e^{-x})$ and d_{dic} is a vector function of the dimension of τ chosen by the investigator. In large samples, a $1 - \alpha$ joint confidence interval for τ^* is the set of τ for which the score test of the hypothesis $\theta = 0$ does not reject at level α. [The score test "numerator" is $D(\tau)$ as defined in Section 2.1 except with $\text{expit}(v_{\text{dic}}(X; \widetilde{\lambda}) + \gamma^* q_{\text{dic}}(H(\tau), X))$ replacing π, where $\widetilde{\lambda} = \widetilde{\lambda}(\tau)$ maximizes $\prod_j \mathcal{L}_j(\lambda, 0, \tau)$.] The choice of the function d_{dic} influences the efficiency of the estimate of $\widehat{\tau}$ (and confidence interval width), but not its consistency or asymptotic normality. We discuss the optimal choice of d below. Note we have put the term "likelihood" in parentheses because the "likelihood" $\mathcal{L}(\lambda, \theta, \tau)$ is not related to the true likelihood function for τ. Rather, it is an artificial likelihood which we use as a computational "trick" to obtain $\widehat{\tau}$. In fact, $\widehat{\tau}$ is a semiparametric nonlikelihood-based estimator. Having obtained an estimate $\widehat{\tau}$ of τ^*, we immediately obtain a consistent asymptotically normal (CAN) estimate of the distribution function $F_0(t)$ of R_0 from the empirical distribution function of $H_j(\widehat{\tau})$, $j = 1, \ldots, n$. Under the assumption of no treatment interaction, we obtain a CAN estimate of $F_z(t)$ from the empirical distribution of $H_{zj}(\widehat{\tau}) = R_j I(Z_j = z) + I(Z_j \neq z)h^{-1}(H_j(\widehat{\tau}), X_j, z, \widehat{\tau})$.

The true likelihood function for the data is the product over the n study subjects of $\mathcal{L}_{\text{true}}(\tau, \lambda, \eta) = \{\partial H(\tau)/\partial R\} f\{H(\tau)|X; \eta_1\} f(X; \eta_2) f(Z|X, H(\tau); \lambda)$, where we have suppressed the dependence on the known parameter γ^*, $\partial H(\tau)/\partial R$ is the Jacobian, and η_1 and η_2 are infinite-dimensional parameters indexing all conditional laws of $H(\tau^*)$ given X and marginal laws of X. The semiparametric variance bound (SVB) for estimators of τ is the supremum of the Cramér–Rao variance bounds for τ over all correctly specified parametric submodels for the infinite-dimensional parameter η. Let $S(\tau, \lambda, \eta_1) = s(H(\tau), X, Z, \tau, \lambda, \eta_1) = \partial \log \mathcal{L}_{\text{true}}(\tau, \lambda, \eta)/\partial\tau$ be the score for τ. In Appendix A, we prove that the asymptotic variance of the estimator $\widehat{\tau} = \widehat{\tau}(D_{\text{dic, opt}})$ with $D_{\text{dic, opt}}(\tau, \lambda, \eta_1) \equiv d_{\text{dic, opt}}(H(\tau), X, \tau, \lambda, \eta_1) = s(H(\tau), X, 1, \tau, \lambda, \eta_1) - s(H(\tau), X, 0, \tau, \lambda, \eta_1)$ attains the SVB. Since (λ, η_1) is unknown, we cannot use $d_{\text{dic, opt}}$ in the above algorithm. Therefore, we estimate (λ, η_1) from the data as follows. First we obtain inefficient estimates $\widehat{\lambda}(\tau)$ by maximizing $\prod_j \mathcal{L}_j(\lambda, 0, \tau)$ over λ. We then specify a lower dimensional model submodel $f(H(\tau)|X; \eta_{1\text{sub}})$ that depends on a finite- or infinite-dimensional parameter $\eta_{1\text{sub}}$. Let $\widetilde{\eta}_{1\text{sub}}(\tau)$ be an estimator of $\eta_{1\text{sub}}$ under the submodel such as the maximizer of $\prod_j f(H_j(\tau)|X_j; \eta_{1\text{sub}})$ if $\eta_{1\text{sub}}$ is finite-dimensional. Then, under model (4), the estimator $\widehat{\tau}_{\text{opt}}$ that uses $\widehat{D}_{\text{dic,opt}}(\tau) = D_{\text{dic,opt}}(\tau, \widehat{\lambda}(\tau), \widetilde{\eta}_{1\text{sub}}(\tau))$ in the above algorithm is locally semiparametric efficient at the submodel $f(H(\tau^*)|X; \eta_{1\text{sub}})$. That is, under model (4), $\widehat{\tau}_{\text{opt}}$ is CAN whether or not the submodel $f(H(\tau^*)|X; \eta_{1\text{sub}})$ is correctly specified; if the submodel is correct, $\widehat{\tau}_{\text{opt}}$ attains the SVB for the model.

Further, it follows from Theorem 2 of Robins and Rotnitzky (2001) that the estimator is a locally efficient doubly robust estimator at partial ignorability (i.e., $\gamma^* = 0$). See also van der Laan and Yu (2001). That is, when $\gamma^* = 0$, $\widehat{\tau}_{\text{opt}}$ is doubly robust in the sense that it is CAN if either (but not necessarily both) the model $\text{logit}\,\text{pr}(Z = 1|x, r_0; \lambda) = v_{\text{dic}}(x; \lambda)$ or the model $f(H(\tau^*)|X; \eta_{1\text{sub}})$ is correct. It is locally efficient in the sense that when both models are correct, it has the smallest asymptotic variance of any doubly robust estimator. However, confidence intervals for $\widehat{\tau}_{\text{opt}}$ should be based on a bootstrap estimate of its variance, since the aforementioned interval estimator will not cover at its nominal rate when only model $f(H(\tau^*)|X; \eta_{1\text{sub}})$ is correct. It can be shown that no doubly robust estimator exists when $\gamma^* \neq 0$.

With nondichotomous Z, we proceed as above except now $d_{\text{dic}}(H(\tau), X)$ is replaced by a function $D = d(H(\tau), X, Z)$,

$$\mathcal{L}(\lambda, \theta, \tau) = \frac{v(Z|X;\lambda)\exp(\gamma^* q(H(\tau),X,Z)+\theta' d(H(\tau),X,Z))}{\int v(Z|X;\lambda)\exp(\gamma^* q(H(\tau),X,Z)+\theta' d(H(\tau),X,Z))\,d\mu(Z)}$$

and $D_{\text{opt}}(\tau, \lambda, \eta_1) = S(\tau, \lambda, \eta_1)$. Estimation in the model with λ^* known is as above except, of course, λ^* need not be estimated.

EXAMPLE. Suppose $H(\tau) = R - \tau Z$, where Z may not be dichotomous, and we choose the model $f[H(\tau^*)|X; \eta_{1\text{sub}}] = f[H(\tau^*) - \beta' X; \mu, \sigma^2]$, where $f[\varepsilon; \mu, \sigma^2] = \sigma f_0[(\varepsilon - \mu)/\sigma]$, f_0 is a known density such as $N(0, 1)$ or Cauchy and $\eta_{1\text{sub}} = (\beta', \mu, \sigma^2)'$. Let $s_0(t) = \partial \log f_0(t)/\partial t$, so $s_0(t) = -t$ for the

$N(0,1)$ distribution and $s_0(t) = -2t/(1+t^2)$ for the Cauchy distribution. Let $\dot{q}$ be the derivative of q with respect to its first argument. Then $S(\tau,\lambda,\eta_1) = -Z\sigma^{-1}s_0(\sigma^{-1}\{R-\tau Z-\beta' X-\mu\}) - \gamma^*\{\dot{q}(H(\tau), X, Z) - E[\dot{q}(H(\tau), X, Z)|H(\tau), X; \lambda, \gamma^*]\}$. It follows that, in the normal case with $\gamma^* = 0$ and Z dichotomous, $\widehat{d}_{\text{dic, opt}}(H(\tau), X)$ is minus the residual from the ordinary least squares (OLS) regression of $H(\tau)$ on $(1, X')'$ divided by $\widehat{\sigma}(\tau)$. In contrast, in the Cauchy case, $d_{\text{dic, opt}}(H(\tau), X)$ is a highly nonlinear function of $R - \tau Z - \widetilde{\beta}(\tau)'X - \widetilde{\mu}(\tau)$, and $\widetilde{\beta}(\tau)$ and $\widetilde{\mu}(\tau)$ are not computed by OLS. To not have to decide between choosing f_0 to be $N(0,1)$ or Cauchy, we can follow Bickel (1982) and adapt to f_0 without asymptotic efficiency loss by letting $f[\varepsilon;\mu,\sigma^2] = \sigma f_0[(\varepsilon-\mu)/\sigma]$ be a smooth, completely unknown density (so $\eta_{1\text{sub}}$ is infinite-dimensional) and then computing kernel density and density derivative estimators of the ratio $\{\partial f(\varepsilon;\mu,\sigma^2)/\partial\varepsilon\}/f(\varepsilon;\mu,\sigma^2) = \sigma^{-1}s_0(\sigma^{-1}\{R-\tau Z-\beta' X-\mu\})$ from the estimated residuals $\widetilde{\varepsilon}(\tau) = R - \tau Z - \widetilde{\beta}(\tau)'X$, where, $\widetilde{\beta}(\tau)$ is a robust regression estimator of the coefficient of X from a robust regression of the variable $R-\tau Z$ on $(1, X')'$.

3.2 An Alternative Sensitivity Analysis Model

The biggest challenge in conducting a sensitivity analysis is to choose a parameterization that has an interpretation that can be communicated to relevant subject matter experts with sufficient clarity so they can provide informed opinions. To use model (4), a subject matter expert must be able to offer opinions about the magnitude γ^* and the shape $q(t,x,z)$ of the dependence of $f(z|x,r_0)$ on the potential outcome r_0. When Z is dichotomous, the task can often be made easier by rewriting model (4) using Bayes' theorem as

$$\begin{aligned} &f(r_0|z=1, X=x;\lambda,\gamma^*)\\ &\quad= c(\lambda,\gamma^*)f(r_0|Z=0,X=x)\\ &\qquad\cdot\exp\bigl(v_{\text{dic}}(x;\lambda)+\gamma^* q_{\text{dic}}(r_0,x)\bigr), \end{aligned}$$

where $c(\lambda,\gamma^*)$ is a normalizing constant and $f(r_0|Z=0,X=x)$ is completely unknown. Even then, we suspect that many experts will have less difficulty giving opinions about the quantile–quantile function linking these two distributions than about the densities. Hence, let

$$\ell(t,x,z) = F_0^{-1}\bigl[F_0\{t|X=x,Z=0\}\big|X=x,Z=z\bigr]$$

be the quantile–quantile function that quantifies how the distribution of R_0 among subjects with $X=x$ and $Z=z$ differs from that among subjects with $X=x$ and $Z=0$. Under partial ignorability, $\ell(t,x,z)\equiv t$.

An alternative sensitivity analysis model regards $\ell(t,x,z)$ as known. A simple choice for $\ell(t,x,z)$ would be $t-\gamma^*z$, with γ^* the parameter to be varied in a sensitivity analysis. The model characterized by

$$\ell(t,x,z) \text{ is a known function} \tag{5}$$

is a nonparametric just identified model in the sense that it places no restriction on the joint distribution of the observables (R,X,Z), but identifies the function $h(r,x,z)$ and the law of R_0 given $X=x$ and $Z=x$. Models (5) and (6) are identical at partial ignorability, that is, when $\gamma^*=0$ and $\ell(t,x,z)=t$.

Given a SDM, let $L=\ell(H,X,Z)$ and $L(\tau)=\ell(H(\tau),X,Z)$ with H and $H(\tau)$ as above. Note the distribution of $L|Z,X$ is $F_0(t|X,Z=0)$ so $L\coprod Z|X$ and $L(\tau^*)\coprod Z|X$. It follows that if we impose a parametric model $v(z|x;\lambda)$ for $f(z|x)$, then, for a given user-supplied vector function d, the estimator $\widehat{\tau}$ and the $1-\alpha$ confidence interval for τ^* described above are CAN and cover at the nominal rate when we modify $\mathcal{L}(\lambda,\theta,\tau)$ by eliminating the term containing γ^* and by replacing $H(\tau)$ by $L(\tau)$.

The true likelihood function for model (5) is the product over the study subjects of $\mathcal{L}_{\text{true}}(\tau,\lambda,\eta) = \{\partial L(\tau)/\partial R\}f\{L(\tau)|X;\eta_1\}f(X;\eta_2)f(Z|X;\lambda)$, where η_1 and η_2 are infinite-dimensional. We obtain a locally semiparametric efficient doubly robust estimator $\widehat{\tau}_{\text{opt}}$ using the above algorithm with $L(\tau)$ substituted for $H(\tau)$. In model (5), (X,Z) are always ancillary for τ. In contrast, in model (4), (X,Z) are ancillary only when $\gamma^*=0$ and thus $R_0\coprod Z|X$. It follows from the ancillarity of (X,Z) and Theorem 2 of Robins and Rotnitzky (2001) that, in contrast to model (4), $\widehat{\tau}_{\text{opt}}$ is always doubly robust in model (5), even when $R_0\not\coprod Z|X$. From a statistical standpoint, this is an advantage of model (5) over (4).

3.3 Instrumental Variable Methods

Suppose, following Rosenbaum, we observe the post-Z variable W with support $\mathcal{W}$ and potential outcomes W_z before observing the outcome of interest R with potential outcomes R_{wz} that depend on the levels to which both w and z are set. Robins (1989) considered inference under the randomization assumption $Z\coprod\{R_{zw};w\in\mathcal{W},z\in\mathcal{Z}\}|X$ and the no direct effect of Z assumption $R_{zw}=R_w$ (which Angrist, Imbens and Rubin, 1996, referred to as the exclusion restriction). We now follow the approach of Robins, Greenland and Hu (1999) and extend our sensitivity analysis methodology to this data structure. The randomization and exclusion restrictions are included as

special cases. Specifically, define $F_{zw}(t|X, Z, W) = \text{pr}(R_{zw} < t|X, Z, W)$. Define the quantile–quantile functions $h_1(t, X, Z, W) = F_{Z0}^{-1}\{F_{zw}(t|X, Z, W)|X, Z, W\}$ and $h_0(t, X, Z) = F_{00}^{-1}\{F_{z0}(t|X, Z)|X, Z\}$ and let $H_1 = h_1(R, X, Z, W)$ and $H = h_0(H_1, X, Z)$. It directly follows from these definitions that H and R_{00} have the same distribution given (X, Z). A structural nested distribution model (SNDM) $(h_0(t, X, Z; \tau_0), h_1(t, X, Z, W; \tau_1))$ is a pair of SDM for (h_0, h_1) (Robins, 1997). Each SDM function is identically equal to t if and only if its parameter is zero. Let $\tau^* = (\tau_0^*, \tau_1^*)$ denote the truth. Let $H(\tau) = h_0(H_1(\tau_1), X, Z, \tau_0)$, where $H_1(\tau_1) = h_1(R, X, Z, W, \tau_1)$. The no direct effect of Z(i.e., exclusion restriction) implies that $\tau_0^* = 0$, equivalently $h_0(t, X, Z) \equiv t$, or equivalently $F_{00}(t|X, Z) = F_{Z0}(t|X, Z)$. The hypothesis $R_{zw} = R_{z0}$ for all w and z of no direct effect of W on R implies $\tau_1^* = 0$ or, equivalently, $F_{ZW}(t|X, Z, W) = F_{Z0}(t|X, Z, W)$.

We shall estimate τ^* using generalizations of our sensitivity analysis models (4) and (5). The generalization of model (5) redefines $\ell(t, X, Z)$ as $F_{00}^{-1}\{F_{00}(t|X, Z)|X, Z = 0\}$. The generalization of model (4) simply replaces R_0 by R_{00} in (4). Thus both models impose assumptions about the conditional dependence between Z and R_{00} given X. Neither makes any assumptions about the dependence between W and R_{Z0} given (Z, X). In this sense, both models are firmly in the tradition of standard instrumental variable methods based on the randomization assumption: assumptions about the association of Z with the potential outcomes are used to draw inferences about the effect of the treatment W on the response R.

The randomization assumption implies the partial randomization assumption $R_{00} \coprod Z|X$ which is equivalent to $\ell(t, X, Z) \equiv t$ in the extended model (5) and to $\gamma^* = 0$ in extended model (4). If we impose a parametric model $v(z|x; \lambda)$, then, given user-supplied vector functions d_{dic} or d, the estimator $\widehat{\tau}$ and the $1-\alpha$ confidence interval for τ^* described above under models (4) and (5), respectively, remain CAN and still cover at the nominal rate under the extended models. Robins, Scharfstein and Rotnitzky (2000) describe additional nonidentifiable assumptions, analogous to the assumptions concerning $h^c(t, x, z)$ mentioned above, that are sufficient to identify $F_{zw}(t|X)$.

The likelihood function $\mathcal{L}_{\text{true}}(\tau, \lambda, \eta)$ of extended models (4) and (5) has an additional term, $f(W|H(\tau), X, Z; \eta_3)$ and $f(W|L(\tau), X, Z; \eta_3)$, respectively, with η_3 an unrestricted infinite-dimensional parameter. The optimal function $D_{\text{opt}}(\tau, \lambda, \eta_1, \eta_3)$ is $E[S(\tau, \lambda, \eta_1, \eta_3)|H(\tau), X, Z]$ in extended model (4) and $E[S(\tau, \lambda, \eta_1, \eta_3)|L(\tau), X, Z]$ in extended model (5), where $S(\tau, \lambda, \eta_1, \eta_3) = \partial \log \mathcal{L}_{\text{true}}(\tau, \lambda, \eta)/\partial\tau$. In extended model (4), $\widehat{D}_{\text{opt}}(\tau) = D_{\text{opt}}(\tau, \widehat{\lambda}(\tau), \widetilde{\eta}_{1\text{sub}}(\tau), \widetilde{\eta}_{3\text{sub}}(\tau))$, where $\widehat{\lambda}(\tau)$ is calculated as above and $\widetilde{\eta}_{1\text{sub}}(\tau)$ and $\widetilde{\eta}_{3\text{sub}}(\tau)$ maximize $\prod_j f[H_j(\tau)|X_j; \eta_{1\text{sub}}] f[W_j|H_j(\tau), X, Z; \eta_{3\text{sub}}]$ when $\eta_{1\text{sub}}$ and $\eta_{3\text{sub}}$ index finite-dimensional parametric submodels. For extended model (5), the procedure is identical but with $L_j(\tau)$ substituted for $H_j(\tau)$. In both extended models (4) and (5), $\widehat{\tau}_{\text{opt}}$ is locally semiparametric efficient at the submodels indexed by $\eta_{1\text{sub}}$ and $\eta_{3\text{sub}}$. In extended model (5), $\widehat{\tau}_{\text{opt}}$ is also doubly robust in the sense that it is CAN if either the model $v(Z|X; \lambda)$ is correct or both of the parametric submodels indexed by $\eta_{1\text{sub}}$ and $\eta_{3\text{sub}}$ are correct. In extended model (4), $\widehat{\tau}_{\text{opt}}$ is doubly robust only when $\gamma^* = 0$. Newey (1990) was the first to derive the SVB for model (5) and its extension and model (4) and its extension when $\gamma^* = 0$. Robins (1997, 1998) gave the SVB in model (4) with $\gamma^* \neq 0$.

3.4 Matched Studies

Following Rosenbaum, consider a pair matched study with dichotomous Z, and $n/2$ subject pairs. We let the first component X_{1j} of X_j have the value k if subject j is in the kth pair and we take $v_{\text{dic}}(X; \lambda) = \sum_{k=1}^{n/2} \lambda_{1k} I(X_1 = k) + v(X; \lambda_2)$, where λ is an unknown parameter and $v(X; \lambda_2)$ is a known function that adjusts for variables that were not matched on. Then inference proceeds exactly as above except we now use the conditional logistic likelihood that conditions on the $n/2$-dimensional vector recording the number of treated subjects N_k in each pair k as in Robins, Blevins, Ritter and Wulfsohn (1992) and Rosenbaum (1988). For example, for model (4), this likelihood is $\prod_{k=1}^{n/2} \mathcal{L}_k(\lambda, \theta, \tau)$, where

$$\mathcal{L}_k(\lambda, \theta, \tau) = \left(\frac{\exp(M_{1k}(\lambda,\theta,\tau))^{Z_{1k}} \exp(M_{2k}(\lambda,\theta,\tau))^{1-Z_{1k}}}{\exp(M_{1k}(\lambda,\theta,\tau)) + \exp(M_{2k}(\lambda,\theta,\tau))}\right)^{I(N_k=1)},$$

and, for example, $M_{1k}(\lambda, \theta, \tau) = v(X_{1k}; \lambda_2) + \gamma^* q_{\text{dic}}(H_{1k}(\tau), X_{1k}) + \theta d_{\text{dic}}(H_{1k}(\tau), X_{1k})$ and we have used $1k$ and $2k$ to denote the subjects in pair k. Thus, $\prod_{k=1}^{n/2} \mathcal{L}_k(\lambda, \theta, \tau)$ is the likelihood for $\sum_{k=1}^{n/2} I(N_k = 1)$ independent Bernoulli random variables Z_{1k} with success probabilities

$$\frac{\exp(M_{1k}(\lambda, \theta, \tau))}{\exp(M_{1k}(\lambda, \theta, \tau)) + \exp(M_{2k}(\lambda, \theta, \tau))}.$$

3.5 Exact Superpopulation Inference

In models (4) and (5) and their extensions, small sample exact tests of the hypothesis that τ is the true value of the superpopulation parameter τ^* when (1) Z is dichotomous and (2) $v_{\text{dic}}(X;\lambda) = \lambda' v_{\text{dic}}(X)$ is linear in λ and includes an intercept can be obtained as above except now we use exact tests of the hypothesis $\theta = 0$ based on exact logistic regression as in Rosenbaum's Section 3.1. That is, as in Rosenbaum's Section 3.1, we condition on the sufficient statistic $\mathbf{Z}'\mathbf{v}_{\text{dic}}(\mathbf{X})$, where $\mathbf{v}_{\text{dic}}(\mathbf{X})$ is the $n \times p$ matrix with rows $v_{\text{dic}}(X_j)$.

4. OBSERVED STUDY POPULATION INFERENCE REVISITED

In this section, we show that the superpopulation methodology reviewed in Section 3 above can, with minor modification, be used for observed study population inference as considered by Rosenbaum. Suppose Z is dichotomous, data on W are not obtained, there are no hidden biases and we again wish to estimate effects in the observed study population so the counterfactuals $\mathbf{r}.$ and covariates $\mathbf{x}$ are fixed constants. We assume $f(\mathbf{Z}|\mathbf{r}., \mathbf{x})$ is $\prod_j \pi_j(Z_j)$, where $\pi_j(1) = \text{expit}\{v_{\text{dic}}(x_j;\lambda)\}$ with $v_{\text{dic}}(x_j;\lambda)$ known and λ an unknown finite-dimensional parameter so hidden bias is absent. Let $\widehat{F}_z(t|x)$ be the empirical distribution of the r_{zj} among subjects with x_j equal to x. Redefine $h(t,x,z) = \widehat{F}_0^{-1}\{\widehat{F}_z[t|x]\}$ and let $h(t,x,z,\tau)$ be a parametric SDM. The true parameter value $\tau^* = \tau^*(\mathbf{r}., \mathbf{x})$ is a function of $(\mathbf{r}., \mathbf{x})$ of the observed study. Then, arguing exactly as in Section 2.1, given a correctly specified SDM model, the 95% interval estimators for τ^* described above for model (4) will have coverage strictly greater than 0.95 under $f(\mathbf{Z}|\mathbf{r}., \mathbf{x})$ in large samples unless rank preservation holds [i.e., by definition, $h(r_{1j}, x_j, 1) = r_{0j}$ for all j], in which case the limiting coverage rate is exactly 0.95. Note if, as often will be the case, no two subjects j have the same value of x, then rank preservation trivially holds, because $\widehat{F}_z(t|x_j)$ is a point mass at r_{zj}. If we have rank preservation we can examine effects of hidden bias by assuming $\pi_j(1) = \text{expit}(v_{\text{dic}}(x_j;\lambda) + \gamma^* q_{\text{dic}}(r_{0j}, x_j))$ and varying $(\gamma^*, q_{\text{dic}}(\cdot,\cdot))$ in a sensitivity analysis. Under this model, the above 95% interval estimators for τ^* under model (4) will again cover under $f(\mathbf{Z}|\mathbf{r}., \mathbf{x})$ at their nominal rate in large samples. Rosenbaum assumed rank preservation with $h(t,x,z,\tau) = t - \tau z$.

The issue of what function d_{dic} to choose to construct interval and point estimators $\widehat{\tau} = \widehat{\tau}(d_{\text{dic}})$ for $\tau^*(\mathbf{r}., \mathbf{x})$ remains. A reasonable approach is to entertain the working hypothesis that $(r._j, x_j)$, $j = 1, \ldots, n$, are n i.i.d. realizations from a density F satisfying rank preservation $h(r_{1j}, x_j, 1) = r_{0j}$ on a set with F-probability 1. Then $\text{var}[n^{1/2}(\widehat{\tau}(d_{\text{dic}}) - \tau^*)|\mathbf{R}. = \mathbf{r}., \mathbf{X} = \mathbf{x}]$ will, on a set with F^n-probability 1, converge to the unconditional variance of $n^{1/2}(\widehat{\tau}(d_{\text{dic}}) - \tau^*)$ since $E[\widehat{\tau}(d_{\text{dic}}) - \tau^*|\mathbf{R}. = \mathbf{r}., \mathbf{X} = \mathbf{x}] = 0$ on a set of F^n-probability 1. Here F^n is the n-fold product law derived from F. It follows that if we make the further working hypothesis that $R_0|X$ under F is in a given parametric or semiparametric model $f(r_0|x;\eta_{1\text{sub}})$, then one should use the locally efficient estimator $\widehat{\tau}_{\text{opt}} = \widehat{\tau}(\widehat{d}_{\text{dic, opt}})$ to try to minimize the variance and confidence interval length for large n under $f(\mathbf{Z}|\mathbf{r}., \mathbf{x})$. This methodology for covariance adjustment based on semiparametric efficiency theory, in contrast to the more ad hoc approach of Rosenbaum, provides for active adaptation to, rather than simply robust protection against, skew or heavy-tailed empirical distributions for R_0.

When data on W are available, we suppose $h_1(r_{zwj}, x_j, z, w_{zj}, z) = r_{z0j}$ and $h_0(r_{z0j}, x_j, z) = r_{00j}$ for all j and we have a correctly specified SNDM $(h_0(t, x, z; \tau_0), h_1(t, x, z, w; \tau_1))$ for (h_0, h_1). Note that Rosenbaum chose $\tau_0^* = 0$ and $h_1(t,x,z,w;\tau_1^*) = t - \tau_1^* w$. Then our above confidence intervals for the observed study parameter $\tau^*(\mathbf{r}.., \mathbf{x}, \mathbf{w}.) = \tau^* = (\tau_0^*, \tau_1^*)$ under extended model (4) will cover at their nominal rate in large samples under $f(\mathbf{Z}|\mathbf{r}.., \mathbf{x}, \mathbf{w}.) = \prod_j \pi_j(Z_j)$ with $\pi_j(1) = \text{expit}(v_{\text{dic}}(x_j;\lambda) + \gamma^* \cdot q_{\text{dic}}(r_{00j}, x_j))$, where $\mathbf{r}.. = \{r_{zwj}\}$ and $\mathbf{w}. = \{w_{zj}\}$. Furthermore, $\widehat{d}_{\text{dic, opt}}$ computed as in Section 3.3 should be used to attempt to minimize confidence interval length for large n.

4.1 Relationship to Rosenbaum's Methodology

As in Rosenbaum's Section 3, suppose Z is dichotomous, data on W are unavailable and we have an additive model $h(r_{zj}, x_j, z) = r_{zj} - \tau^* z$, but we allow for hidden bias. In this setting, Rosenbaum (1988) considered the model

$$\text{logit}\{\text{pr}[Z = 1|x_j, u_j]\} = v_{\text{dic}}(x_j,\lambda) + \gamma^* u_j, \quad u_j \in (0,1), \tag{6}$$

where λ is an unknown vector parameter, the u_j are unmeasured covariates and γ^* is the parameter to be varied in a sensitivity analysis. Rosenbaum uses this model to set confidence intervals $(\tau_{\text{lower}}, \tau_{\text{upper}})$ for τ^* using the following pseudo-algorithm. Given γ^*, let

τ_{lower} be the largest value of τ such that a given exact or large sample $\alpha/2$ test of the null hypothesis $\tau^* < \tau$ rejects for all choices of $\{u_j\}$. Let τ_{upper} be the smallest value of τ such that an $\alpha/2$ test of the hypothesis $\tau^* > \tau$ rejects for all choices of $\{u_j\}$. Rosenbaum's interval $(\tau_{\text{lower}}, \tau_{\text{upper}})$ is the same interval that would be computed under my model (4) if we replace the phrases "rejects for all choices of $\{u_j\}$" with "rejects for all functions $q_{\text{dic}}(r_0, x)$ both of the form $I(r_0 > c(x))$ and of the form $I(r_0 < c(x))$, where $c(x)$ is an arbitrary function." When X is low-dimensional and discrete and $v_{\text{dic}}(x_j, \lambda)$ is a saturated model, Rosenbaum has developed tractable algorithms that compute τ_{lower} and τ_{upper} in many important cases. It would be of interest to know if Rosenbaum has ideas as to how to compute τ_{lower} and τ_{upper} when X is high-dimensional.

5. SENSITIVITY ANALYSIS—IS IT SCIENTIFICALLY USEFUL?

Rosenbaum's model (6) and my analogous model (4) will be scientifically useful only if experts can provide a plausible and logically coherent range for the value of the sensitivity parameter $e^{\gamma^*} = \Gamma^*$. I now argue this may be difficult without intensive training. Suppose a covariate X is known to be strongly associated with the outcome R among the untreated $(Z = 0)$. If X is also correlated with treatment Z, then studies in which data on X are available (and can thus be adjusted for) will generally suffer from less hidden bias than studies in which data on X have not been collected. Naively, one would therefore expect that the availability of data on X should narrow the range of γ^* considered plausible. To the surprise of most statisticians and to nearly all subject matter experts, this may not be so, because the meaning of the conditional odds ratio parameter $e^{\gamma^*} = \Gamma^*$ in (4) and (6) depends on the covariates recorded in X. It follows that $e^{\gamma^*} = \Gamma^*$ is a "paradoxical" measure of hidden bias.

To prove this claim, we shall consider an extreme example of a very large study (so sampling variability can be ignored) in which the additive model $r_{1j} - r_{0j} = \tau^*$ holds and X is the only measured covariate. Suppose the empirical distribution of the data (r_j, x_j, z_j), $j = 1, \ldots, n$, is as follows: (i) X is uniformly distributed on $\{1, 2, \ldots, 100\}$, (ii) X and Z are independent, and (iii) given $X = x$ and $Z = z$, R is uniformly distributed on $(\frac{x-1}{100}, \frac{x}{100})$ for both $z = 0$ and $z = 1$ so τ^* will be estimated to be 0 were one to assume no hidden bias $(\Gamma^* = 1)$. Suppose an expert gives $(1/100, 100)$ as his or her plausible range for the parameter $e^{\gamma^*} = \Gamma^*$ in model (6) with $v_{\text{dic}}(x_j, \lambda) = \sum_{k=1}^{100} \lambda_k I(x_j = k)$ saturated. Then in Appendix B, we show that $(\tau_{\text{lower}}, \tau_{\text{upper}}) = (-0.00409, 0.00409)$, the limits corresponding to Γ^* equaling 100 and $\Gamma^* = 1/100$. Suppose, however, that for confidentiality reasons, the expert and the data analyst (i) were not allowed access to the individual data on X, so Rosenbaum's model

$$\text{(7)} \qquad \operatorname{logit} \operatorname{pr}\{Z = 1 | u_j\} = \lambda_0 + \gamma^* u_j, \quad u_j \in (0, 1),$$

with λ_0 an unknown constant must be used in the analysis, but (ii) were told that X and Z were independent and were given both the empirical marginal law of X and empirical conditional law of R given $X = x$ and $Z = 0$. Because, when X and Z are independent, X is intuitively not a confounder and, therefore, the degree of hidden bias should generally be the same regardless of whether data on X have been obtained, an untutored expert might suppose he or she should give the same range of $(1/100, 100)$ for $e^{\gamma^*} = \Gamma^*$ regardless of whether data on X are available. However, in Appendix B, we show that the choice $(1/100, 100)$ in model (6) when data on X are available plus the information available to the expert implies the range $(1/1.03, 1.03)$ for Γ^* in model (7) without data on X. The use of this range for $e^{\gamma^*} = \Gamma^*$ in model (7) again results in the confidence interval $(-0.00409, 0.00409)$ for τ^*. In contrast, use of the incoherent choice $(1/100, 100)$ in model (7) results in the logically incoherent, misleadingly wide interval of $(-0.5, 0.5)$.

Since for logistical reasons the covariates recorded in X vary widely among the various studies of a given treatment–response association, the above example makes it clear that to effectively summarize overall uncertainty or to apply results from one study to help choose a plausible range for another study, we must either abandon Rosenbaum's model (6) and my analogous (4), or we must provide careful guidance and education as to the X-dependent meaning of γ^*. One might hope that model (4) could be saved by using a different model for $\operatorname{pr}(Z = 1 | X, R_0)$ other than the logistic. Based on a related discussion in Section 7.2 of Scharfstein, Rotnitzky and Robins (1999), I doubt that this hope can be fulfilled.

We are still left with the question of whether there exist sensitivity analysis models that are consistent with the naive intuition that data on additional covariates X should not lead to a larger plausible range for the sensitivity parameter. In fact, the alternative sensitivity analysis model of Section 3.2 is such a model

(although it requires that we adopt the superpopulation point of view, as the model will often be undefined or vacuous if we treat counterfactuals $r_\cdot$ and covariates $\mathbf{x}$ as fixed). For example, if we take $\ell(t,x,z) = t - \gamma^* z$ (so $\gamma^* = E[R_0|Z=1, X=x] - E[R_0|Z=0, X=x]$), then in our example with X independent of Z, a plausible range (a,b) for γ^* when data on X are available logically entails, as intuition would suggest, the same range when data on X are unavailable. However, we have a problem peculiar to the academy: the model $\ell(t,x,z) = t - \gamma^* z$ is trivial in the sense that its implications are so self-evident that a paper on the topic would not satisfy the difficulty level required of a freshman college course much less that expected of a *Journal of the American Statistical Association* or *Biometrika* paper. (Although it is possible to hide this triviality behind some fancy math as in Section 3.2 above.) For example, if one obtains a point estimate $\widehat{\tau}$ and interval estimate $(\tau_{\text{lower}}, \tau_{\text{upper}})$ for an additive effect parameter τ^* in the absence of hidden bias ($\gamma^* = 0$), then, for any other value γ^* examined in the sensitivity analysis, we obtain $\widehat{\tau} - \gamma^*$ and $(\tau_{\text{lower}} - \gamma^*, \tau_{\text{upper}} - \gamma^*)$. However, precisely because of its transparent interpretation, I believe that it is often easier for subject matter experts to give their opinions about the plausible magnitude γ^* of the difference in the conditional means of R_0 than to give opinions about the difficult issues of whether the unmeasured confounders u_j are continuous or discrete, single or multidimensional, and the conditional associations of such confounders with treatment and/or outcome. Furthermore, in longitudinal studies with time-varying treatments, there remains a role for sophisticated mathematical analysis as it is a nontrivial exercise to propagate over time an expert's ranges for the treatment–covariate-time specific differences in counterfactual means to quantify overall uncertainty. See Robins (1999) and Robins, Greenland and Hu (1999).

APPENDIX A: EFFICIENT ESTIMATION

In the extended model (4) the nuisance tangent space (NTS) for the finite-dimensional parameters (τ, λ) is the set of all scores for η and equals $\{a_1(H,X) + a_2(X) + a_3(W,H,X,Z); E[a_1(H,X)|X] = E[a_2(X)] = E[a_3(W,H,X,Z)|H,X,Z] = 0\}$. Thus projection of any $M = m(W,H,X,Z)$ on the NTS is $\{M - E[M|H,X,Z]\} + \{E[M|H,X] - E[M|X]\} + \{E[M|X] - E[M]\}$. By definition the joint efficient score for (τ,λ) is the residual from the projection of the scores $M = (S_\tau, S_\lambda)$ for (τ,λ) on NTS, which is $(E[S_\tau|H,X,Z] - E[S_\tau|H,X], S_\lambda)$ since S_λ is orthogonal to NTS. Thus the efficient score for τ alone is $S_{\tau,\text{eff}} = E[S_\tau|H,X,Z] - E[S_\tau|H,X] - E\{(E[S_\tau|H,X,Z] - E[S_\tau|H,X])S_\lambda'\}\{E[S_\lambda S_\lambda']\}^{-1} S_\lambda$, which we write as $S_{\tau,\text{eff}}(\tau^*)$ to indicate it depends on the true value of τ^*. Now a Taylor expansion of the estimating function for $\widehat{\tau}(d_{\text{opt}})$ around λ shows $\widehat{\tau}(d_{\text{opt}})$ is asymptotically equivalent to an estimator solving $0 = \sum_j S_{\tau,\text{eff},j}(\tau)$. But an estimator solving the efficient score equation has the efficient variance. The results for the other models are a special case.

APPENDIX B

Consider first the case where data on X are available and $e^{\gamma^*} = \Gamma^* \in (1/100, 100)$. We shall need the fact that $g(c) = \int_0^1 r_0 \Gamma^{I(r_0>c)}\, dr_0 / \int_0^1 \Gamma^{I(r_0>c)}\, dr_0$ is maximized at $c_{\max} = c_{\max}(\Gamma) = (\Gamma - \Gamma^{1/2})/(\Gamma - 1)$ and that $g(c_{\max}) = c_{\max}$. The empirical means $E[R_0|X=x, Z=0]$ and $E[R_1|X=x, Z=1]$ equal $(x - 1/2)/100$ by the conditional uniformity of R given $X=x$ and $Z=z$. Now from the sensitivity analysis model (6) with data on X and uniformity of the conditional law of R_0 given $X=x$ and $Z=0$,

$$E[R_0|X=x, Z=1] = \frac{\int_{(x-1)/100}^{x/100} r_0 \exp[\gamma^* I(r_0 > c(x))]\, dr_0}{\int_{(x-1)/100}^{x/100} \exp[\gamma^* I(r_0 > c(x))]\, dr_0},$$

which attains the maximum of $(x - 1 + 0.909)/100 = (x - 1 + c_{\max}(100))/100$ at $\Gamma^* = e^{\gamma^*} = 100$ and $q_{\text{dic}}(r_0, x) = I(r_0 > (x - 1 + c_{\max}(100))/100)$. Similarly, $E[R_0|X=x, Z=1]$ attains its minimum of $(x - 0.909)/100 = (x - c_{\max}(100))/100$ at $\Gamma^* = 1/100, q_{\text{dic}}(r_0, x) = I(r_0 < (x - c_{\max}(100))/100)$. Since, under additivity, $\tau^* = E[R_1|X=x, Z=1] - E[R_0|X=x, Z=1]$, we have that $(\tau_{\text{lower}}, \tau_{\text{upper}}) = (-0.00409, 0.00409)$.

Consider now the case with data on X missing. From the information available to the expert, he or she knows that R_0 is uniformly distributed given $Z=0$ on $(0,1)$ and that $E[R_0|Z_1=1] - E[R_0|Z=0]$ attains its maximum of $(0.909 - 0.5)/100 = 0.00409$ at $\Gamma^* = e^{\gamma^*} = 100$ and $q_{\text{dic}}(r_0, x) = I(r_0 > (x - 1 + c_{\max}(100))/100)$ under model (6). Therefore, he or she can conclude that the maximum value of $E[R_0|Z=1]$ is $0.5 + 0.00409$. A similar calculation gives the minimum $0.5 - 0.00409$. Since, under additivity, $\tau^* = E[R_1|Z=1] - E[R_0|Z=1]$ and $E[R_1|Z=1] = 1/2$ from the data, he or she concludes

that $(\tau_{\text{lower}}, \tau_{\text{upper}}) = (-0.00409, 0.00409)$ without using model (7). Suppose, on the other hand, he or she used model (7) to evaluate $E[R_0|Z = 1]$. To be consistent with the maximum value of $0.5 + 0.00409$ obtained based on model (6), he or she must use the value of Γ^* in model (7) that solves $0.5 + 0.00409 = \sup_c(\int_0^1 r_0\Gamma^{*I(r_0>c)}\,dr_0/\int_0^1 \Gamma^{*I(r_0>c)}\,dr_0) = (\Gamma^* - \Gamma^{*1/2})/(\Gamma^* - 1)$, which is $\Gamma^* = 1.032$. A similar calculation for the minimum gives the interval $(1/1.032, 1.032)$ for the parameter Γ^* of model (7). Technically model (7) and model (6) are incompatible; therefore, we simply computed the value of Γ^* in model (7) that would reproduce the limits on $E[R_0|Z = 1]$ calculated under model (6).

Rejoinder

Paul R. Rosenbaum

I would like to thank Joshua Angrist, Guido Imbens, Jennifer Hill and Jamie Robins for their insightful and gracious comments. Over the past decade, Angrist and Imbens have illuminated the concept of instrumental variables through improved, less cluttered, more general theory (Imbens and Angrist, 1994; Angrist and Imbens, 1995; Angrist, Imbens and Rubin, 1996). Hill has clarified broken randomized experiments, which mix elements of experiments and observational studies (Barnard, Du, Hill and Rubin, 1998). Robins has developed an attractive approach to studies with time-dependent treatments (Robins, Blevins, Ritter and Wulfson, 1992; Robins, 1999).

1. DISCONTINUITY

Angrist and Imbens raise the interesting issue of group randomized trials and their parallels in observational studies, suggesting that the Card and Krueger (1994) study is an example. In a group randomized trial, experimental units come in clusters and whole clusters are randomly allocated to treatment or control. I agree with Angrist and Imbens that observational studies resembling such experiments form an interesting, common and relatively unexplored topic; however, I would view Card and Krueger's study as having a stronger relationship with discontinuity designs.

Campbell and Stanley (1963) discussed the "regression-discontinuity design," in which there is a cutpoint for an observed covariate, say L, which is used to assign treatments in a deterministic manner: units with low scores, $L \leq c$, receive treatment; those with high scores, $L > c$, receive the control. If one believed that bias due to L was continuous as a function of L, then in this design it is possible to estimate the effect of the treatment at the cutpoint $L = c$, because there is only a small bias when comparing treated units just below the cutpoint to controls just above it. In other words, the premise is that bias due to L is continuous in L, whereas treatment assignment is discontinuous in L, so a discontinuity in the response surface at $L = c$ provides evidence about the treatment effect.

More recently, discontinuity designs have taken varied forms, sometimes linked to the use of instrumental variables, for instance, the Wald estimator. See Angrist and Krueger (1991), Angrist and Lavy (1999), Black (1999) and Sullivan and Flannagan (2002) for four clever applications, Angrist and Krueger (1999, 2001) for general discussion, and Hahn, Todd and Van der Klaauw (2001) for associated theory. In particular, Black (1999) considered discontinuities defined by geographic boundaries.

Card and Krueger compared employment in New Jersey and Pennsylvania at fast food restaurants such as Burger King before and after New Jersey increased its minimum wage. Their fine study is most compelling for nearby restaurants in similar neighborhoods on opposite sides of the Delaware River, which defines the border between New Jersey and Pennsylvania. The economies along the Delaware are entwined. Cherry Hill, New Jersey, is a suburb of Philadelphia similar to several Pennsylvania suburbs of Philadelphia. Morrisville, Pennsylvania, is a suburb of Trenton, New Jersey, similar to several New Jersey suburbs of Trenton. Lambertville, New Jersey, is similar to nearby New Hope, Pennsylvania. In contrast, parts of southern New Jersey might be compared to appropriate parts of Delaware, while northern New Jersey might be compared to appropriate adjacent parts of New York state. Parts of New Jersey may have no useful controls in adjacent states and might be excluded, for instance, Atlantic City or Hoboken. As in Campbell and Stanley's discontinuity design, the most compelling comparisons are at the policy discontinuity along the state perimeter

when comparable units exist on opposite sides of that perimeter.

Although response surface discontinuities along state perimeters may provide evidence about effects of state policies, more than one policy may change at the border. For example, New Jersey and Pennsylvania may have different minimum wages, but they also have different income taxes. How can one isolate the effects of a single policy? Because Card and Krueger examined employment before and after the wage increase in New Jersey, stable differences in other economic policies do not provide immediately compelling explanations of changes in employment (cf. Rosenbaum, 2001a). Examination of employment changes among certain businesses employing few or no employees directly affected by the minimum wage provides a second opportunity to isolate the minimum wage—those businesses should not be greatly affected—and Card and Krueger performed some analyses of this kind. Comparisons along boundaries with several adjacent states resemble the use of multiple control groups (Rosenbaum, 2002a, Chapter 8); for example, income taxes differ in Pennsylvania and New York, thereby helping to separate income taxes from effects of changing the minimum wage.

2. INFINITE POPULATIONS

The relationship between randomization inference and infinite population models can be formalized in various ways, some yielding mild divergences, as in Robins' discussion, others yielding harmony. A harmonious formalization was given by Lehmann (1986, Section 5.10) building upon Lehmann and Stein (1949) and Fraser (1954). Lehmann considered a stratified sample from a stratified, infinite, continuous population, in which the treated distribution in each stratum is shifted by τ when compared to the control distribution. Lehmann (1986, Theorem 5.10.6) then showed that the *only* distribution-free tests of a hypothesis about τ are permutation or randomization tests. The beautiful proof uses the complete sufficiency of the order statistics: conditioning on the observed responses but not their treatment assignments (i.e., conditioning on the order statistics) eliminates the unknown distribution functions as nuisance parameters. This quickly yields most powerful permutation tests by way of the Neyman–Pearson lemma. In the case of binary responses, Cox and Snell (1989, page 149) presented Fisher's exact randomization test in harmony with a logit model in which a nuisance parameter is eliminated by conditioning; see also Lehmann (1986, Section 4.5).

In his discussion of superpopulations, Robins presumes that the correct standard error of an estimate for an infinite population is an unconditional standard error, but as just noted, in Lehmann (1986) and Cox and Snell (1989), the distributions used for inference with infinite populations are the conditional distributions, the conditioning being needed to eliminate nuisance parameters. In the nonparametric and logit models just described, the infinite population model and the randomization inference agree exactly with each other, yielding exactly the same inference, and this is slightly at odds with the discussion Robins presents.

3. ADDITIVE EFFECTS FOR INDIVIDUALS AND LOCATION SHIFTS FOR DISTRIBUTIONS: STANDARD THEORY

In the discussion, unease about additive treatment effects arises here and there. Unease is unwarranted, I believe, because certain concerns disappear upon close inspection and others submit to commonplace solutions, such as fitting interaction effects. In the current section, I discuss a simple case in conventional terms, and in Section 4 of this reply I discuss the general case in the terms of Section 3 of the paper.

As reviewed in Section 2 of my paper, causal effects are comparisons, such as $r_{Tj} - r_{Cj}$, of two potentially observable responses which are not jointly observable (Neyman, 1923; Rubin, 1974): we observe R_j, Z_j and $\mathbf{x}_j$, $j = 1, \ldots, n$, but not (r_{Tj}, r_{Cj}). This structure is the basis for some of the most celebrated claims in statistics—for example, that randomized experiments produce unbiased estimates of average causal effects—and yet, it is a peculiar structure, because one ends up talking about certain joint distributions, but one observes only certain marginal distributions derived from them. Is it reasonable to use a model of additive effects $r_{Tj} - r_{Cj} = \tau$ given that (r_{Tj}, r_{Cj}) is not jointly observed?

There seem to me to be three issues that need to be distinguished. First, the observable distributions may be perfectly compatible with additivity, but behind the scenes the effect is not additive. Second, there may be visible evidence in observable distributions that the treatment effect is not additive. Third, new data may become available, so that we may be moved from the first situation to the second. Each of these issues can be dealt with in a straightforward manner, providing they are distinguished. In this section, assume the

simplest situation, namely a large randomized trial—that is, large n—with a coarse, discrete covariate $\mathbf{x}_j$ and with treatments assigned completely at random, $\Pr(Z_j = 1) = \Pr(Z_j = 1 | r_{Tj}, r_{Cj}, \mathbf{x}_j) = \frac{1}{2}$, independently for different subjects j, as one could easily do using a table of random numbers; a more general case is considered in Section 4. Because n is large, statements about distributions are correct for theoretical distributions and nearly so for empirical distributions, the two tending toward agreement as $n \to \infty$.

First, if one observes R_j, Z_j and $\mathbf{x}_j$ in such a simple randomized trial, then additivity implies that at any value of $\mathbf{x}_j$, the distribution of observed responses among treated subjects is shifted by τ from the distribution of responses among control subjects, that is, the two distributions have the same shape and dispersions but different locations, although the shapes and dispersions may vary with $\mathbf{x}_j$. These are, by the way, the same conditions that apply to the distributions in Lehmann (1986, Theorem 5.10.6) mentioned in Section 2 of this rejoinder. In this case, the randomization inference for an additive treatment effect $r_{Tj} - r_{Cj} = \tau$, which refers to the unobservable joint distribution, is exactly the same as the randomization inference for Lehmann's constant shift model, which refers only to observable distributions; moreover, the latter is the only nonparametric inference for this problem. One could, then, interpret randomization inferences about the additive effect τ as inferences about the constant shift model, so that the unobservable joint distribution plays no role. That is, if behind the scenes, the unobservable joint distribution is nonadditive in just such a way as to produce a constant shift for the observable marginal distributions, as in one of Robins's examples, then the randomization inference remains correct as a description of the visible constant shift. So far, no problem.

Second, one may observe in the data that the additive shift model is incorrect. In this case, one would, of course, fit a different model. For example, there might be interaction: the magnitude of the shift might be seen to vary with $\mathbf{x}_j$. Suppose, for example, the magnitude of the shift was a function of the first coordinate of $\mathbf{x}_j$, say x_{j1}, which is a binary variable, so $r_{Tj} - r_{Cj} = \tau_1$ if $x_{j1} = 1$ and $r_{Tj} - r_{Cj} = \tau_0$ if $x_{j1} = 0$. Then the situation described in the previous paragraph simply occurs twice, for $x_{j1} = 1$ and for $x_{j1} = 0$, and no fundamentally new problems arise: the option of interpreting the parameters (τ_1, τ_0) of the unobservable joint distribution in terms of observable distributions is still available. Still, no problem. [To fill in a few technical details, the adjusted responses $R_j - Z_j\{\tau_1 x_{j1} + (1 - x_{j1})\tau_0\}$ equal the potential responses under control r_{Cj}. A hypothesis about (τ_1, τ_0) could be tested by computing adjusted responses under the null hypothesis, regressing these on the remaining coordinates of $\mathbf{x}_j$ to obtain residuals, calculating the two independent Wilcoxon rank sum statistics separately for $x_{j1} = 1$ and for $x_{j1} = 0$, and combining these two statistics into a single test with 2 degrees of freedom. The rank sums are independent in this simple case, because under the null hypothesis they are functions of the fixed r_{Cj}'s and the independent Z_j's.]

There is one place where the model of an additive effect for the unobserved joint distribution says more than the constant shift model for the observable distributions. The additive effect model $r_{Tj} - r_{Cj} = \tau$ makes a prediction about what we would see if we measured additional covariates: it says the constant shift model would continue to describe observable distributions once the additional covariates were incorporated. If we obtain additional covariates and that prediction turns out to be true, then we are back in the first situation above. If the prediction turns out to be false, then we are in the second situation above. In either case, the needed tools are available, and there is no problem.

This process of elucidating a theory with intrinsically unobservable features by tracing their observable consequences is called *ontological elimination* by Sklar (2000), who discussed it in general terms.

The next section considers the matter in general terms.

4. ADDITIVE EFFECTS FOR INDIVIDUALS AND LOCATION SHIFTS FOR DISTRIBUTIONS: EXTENSION TO COVARIANCE ADJUSTMENT IN STUDIES FREE OF HIDDEN BIAS

The conventional ideas in the previous section extend easily to the situation discussed in the main paper. Here, I briefly sketch what is involved, then indicate why I prefer the framework in the paper.

Consider the model

$$r_{Tj} = \mathbf{x}_j^T \boldsymbol{\zeta} + \tau + \varepsilon_{Tj}, \qquad r_{Cj} = \mathbf{x}_j^T \boldsymbol{\zeta} + \varepsilon_{Cj}, \tag{1}$$

$$\log \frac{\Pr(Z_j = 1)}{\Pr(Z_j = 0)} = \log \frac{\pi_j}{1 - \pi_j} = \mathbf{x}_j^T \boldsymbol{\lambda}, \tag{2}$$

where the $\mathbf{x}_j$'s are fixed, distinct people j are mutually independent, and the bivariate error vectors $(\varepsilon_{Tj}, \varepsilon_{Cj})$

all have the same bivariate exchangeable distribution and are independent of treatment assignment Z_j. In this case, the treatment effect, $r_{Tj} - r_{Cj} = \tau + \varepsilon_{Tj} - \varepsilon_{Cj}$ is not constant, but is symmetrically distributed about τ. In an experiment, randomization ensures treatment assignment Z_j is independent of $(\varepsilon_{Tj}, \varepsilon_{Cj})$, whereas in an observational study this is an alternative expression of the (often implausible) assumption that the only biases are due to the observed covariates $\mathbf{x}_j$. From (1), the observed responses R_j equal

$$R_j = \mathbf{x}_j^T \boldsymbol{\zeta} + \tau Z_j + E_j,$$

where $E_j = Z_j\varepsilon_{Tj} + (1 - Z_j)\varepsilon_{Cj}$ is independent of Z_j because $(\varepsilon_{Tj}, \varepsilon_{Cj})$ is exchangeable and independent of Z_j. Therefore, the adjusted responses $R_j - \tau Z_j = \mathbf{x}_j^T \boldsymbol{\zeta} + E_j = a_j$ are independent of the Z_j, so $\Pr(\mathbf{Z}|\mathbf{a}) = \Pr(\mathbf{Z})$ is determined from (2).

To test $H_0 : \tau = \tau_0$, assume the null hypothesis for the purpose of testing it, compute the adjusted responses $R_j - \tau_0 Z_j$ which, therefore, equal a_j, and consider the conditional distribution of treatment assignments $\Pr(\mathbf{Z}|\mathbf{a})$. The situation now is identical to the situation in Section 3 of the main paper, with a_j in place of r_{Cj}, and the methods discussed there may be used without change. The important point here is that replacing an additive effect by an effect $r_{Tj} - r_{Cj} = \tau + \varepsilon_{Tj} - \varepsilon_{Cj}$ that is symmetrically distributed about τ did not require changes in the procedures.

To me, the model given by (1) and (2) is less satisfactory than the model in Section 3 of the main paper, even though they both justify the same statistical procedures. The reason is that random assignment of treatments does not, by itself, justify the model (1), but randomization does justify use of the procedures in Section 3 of the paper. In the nice terminology of Angrist and Imbens, Section 3 of my paper is *agnostic* about the covariance adjustment of the responses—it is a fit, not a model—whereas the derivation of the same procedures from (1) and (2) depends upon the model being true. To put this another way, the model (1) blurs the distinction between randomized experiments and observational studies—they both seem to require believing a model. In contrast, the formulation in Section 3 of the paper sharpens the distinction, as randomization creates what is needed for inference in experiments and an important assumption is needed for inference in observational studies. The distinction between experiments and observational studies is important to practice and should not be blurred in theory.

5. NONADDITIVITY AND REDUCED SENSITIVITY TO HIDDEN BIAS

Nonadditivity is related to sensitivity to hidden bias. Other things being equal, larger treatment effects tend to be less sensitive to hidden bias than smaller effects. When the treatment effect is not additive, there may be greater sensitivity to hidden bias where the effect is small and reduced sensitivity where the effect is large. Sensitivity analyses that permit investigation of this issue are given in Rosenbaum (1999a, 2001b, 2002a, Chapter 5, 2002b).

6. THE POSITIVE ASPECT OF THE NEGATIVE LOGIC OF CONFIDENCE INTERVALS

Because confidence intervals are built from hypothesis tests (Lehmann, 1986, Section 3.5), they share the same negative logic: they tell us what is not plausible. A confidence interval summarizes the rejection of certain hypotheses—the information is in these rejections. The points inside the confidence interval live on without endorsement simply as hypotheses not yet rejected. The positive aspect of the negative logic is this: just as one does not have to believe a null hypothesis to learn something by testing it, so too, one does not have to believe a parametric model to learn something from the parameters excluded from a confidence interval. I want to illustrate what I mean in a simple case, and then claim that a confidence interval for an additive treatment effect is informative when an additive effect is interesting, whether or not one is certain the effect is additive. This positive aspect is also related to the issue raised by Angrist and Imbens concerning interpretation of short or empty confidence intervals with instrumental variables when the exclusion restriction is violated.

Imagine a random quantity X with distribution F contained in some set $\mathcal{F}$ of distributions. The set $\mathcal{F}$ contains a subset $\mathcal{C}$ of distributions indexed by a real, scalar parameter, F_θ, $\theta \in \mathbb{R}$—that is, $\mathcal{C} = \{F_\theta : \theta \in \mathbb{R}\} \subset \mathcal{F}$—so one might metaphorically imagine F_θ as tracing a parameterized curve through $\mathcal{F}$, although this metaphor plays no formal role here. Perhaps $F = F_\theta$ for some θ, perhaps not; that is, the true F may lie on the curve or it may fall elsewhere in $\mathcal{F}$. There is a test of size α which uses X to test any null hypothesis $H_0 : F = F_\theta$, yielding a significance level $p(\theta)$; so if it should happen to be true that $F = F_\theta$, then $\Pr\{p(\theta) \leq \alpha\} = \alpha$. The real line divides into two subsets, the "outside" $\mathcal{O} = \{\theta \in \mathbb{R} : p(\theta) \leq \alpha\}$ and the "inside" $\mathcal{I} = \mathbb{R} - \mathcal{O}$. Whether or not $F \in \mathcal{C}$—whether or not the parameterized model is correct—the distributions

outside the confidence set $\{F_\theta : \theta \in \mathcal{O}\}$ are not plausible and that is informative. The inside $\mathcal{I}$ is not particularly interesting: quite often, for each $\theta \in \mathcal{I}$ inside the confidence set, there are many other distributions $\widetilde{F} \in \mathcal{F} - \mathcal{C}$ outside the parameterized family that are very similar to F_θ, so failure to reject F_θ does not convince us that F is in $\{F_\theta : \theta \in \mathcal{I}\}$.

(If our test worked throughout $\mathcal{F}$ with size α, then $\mathcal{F}$ itself could be divided into outside $\widetilde{\mathcal{O}}$ and an inside $\widetilde{\mathcal{I}}$ $1 - \alpha$ confidence set, and our parametric model provides a first step in understanding these two sets of distributions, because $\{F_\theta : \theta \in \mathcal{O}\} \subseteq \widetilde{\mathcal{O}}$ and $\{F_\theta : \theta \in \mathcal{I}\} \subseteq \widetilde{\mathcal{I}}$. Alternatively, sometimes, attributable effects based on pivots permit one-dimensional descriptions of high dimensional confidence sets; see Rosenbaum, 2001b, 2002b.)

In the context of covariance adjustment, suppose one is not certain that the treatment effect is additive, but the model of an additive treatment effect τ is sufficiently plausible to be interesting. Then the outside $\mathcal{O}$ of the confidence set for τ is informative: it tells us that certain specific additive effects are not plausible and that is news. In parallel, with an instrumental variable, if a confidence interval for β is understood as a statement about the values of β that are not plausible, then an empty confidence interval entails no change in perspective: it says that all values of β are not plausible, so the entire parametric family of distributions defined by β is implausible.

7. PARADOXICAL?

Robins writes: "Rosenbaum's approach to sensitivity analysis, although logically flawless and mathematically elegant, may be scientifically useless." As the reader might anticipate, I enthusiastically agree with part of this. Robins also says he believes that methods of sensitivity analysis he himself has proposed are scientifically useless for the same reasons.

Robins says that a measure of hidden bias is "paradoxical ... if its magnitude can increase as we decrease the amount of hidden bias by measuring some of the unmeasured confounders." I disagree. If this were accepted as the definition of paradoxical behavior of a statistical quantity, then regression coefficients of all kinds—linear, logit, proportional hazards—would be paradoxical, which they are not. The magnitude and interpretation of a regression coefficient depends upon which other variables are in the model; the magnitude can increase or decrease as variables are added to the model. This is correct, not paradoxical, behavior for a regression coefficient. For example, if one were fitting a logit model to predict lung cancer, the coefficient of "cigarettes smoked" would presumably increase when the variable "age" is added to the model: although smoking is responsible for most lung cancer, the cancer tends to occur later in life, and is uncommon among young heavy smokers. Smoking is more important as a predictor of lung cancer at a fixed age, say age 60, than it is ignoring age. That is common sense, not paradox. Regression coefficients are a part of a model and they cannot be understood without reference to the other parts of the model. The sensitivity parameter in my discussion, $\gamma = \log(\Gamma)$, can be viewed as the coefficient of an unobserved covariate in a logit regression of treatment assignment on observed covariates and an unobserved covariate; see Rosenbaum (2002a, Section 4.2). The parameter $\gamma = \log(\Gamma)$ is no more or less paradoxical than any other regression coefficient.

Robins also writes: "Rosenbaum's model (6) and my analogous model (4) will be scientifically useful only if experts can provide a plausible and logically coherent range for the value of the sensitivity parameter $e^\gamma = \Gamma$." Here, Robins and I disagree about what a sensitivity analysis says and how it is used. To my mind, a sensitivity analysis simply indicates the magnitude of hidden bias, measured by Γ, that would need to be present to alter the qualitative conclusions of the study. The sensitivity analysis is a fact of the matter, something one calculates from the data at hand, and it does not rest on opinions, expert or otherwise. It is simply a fact that Hammond's (1964) study of the effects of heavy smoking on lung cancer is much less sensitive to hidden bias than Jick et al.'s (1973) study of the effects of coffee on myocardial infarction. The smoking study becomes sensitive at $\Gamma = 6$, while the coffee study becomes sensitive at $\Gamma = 1.3$; see Rosenbaum (2002a, Chapter 4) for details. This does not mean that the coffee study is biased nor does it mean that coffee does not cause myocardial infarction; it simply means that an unobserved covariate weakly related to coffee consumption could explain the observed association in that study. It is useful to know that the smoking study is vastly less sensitive to bias than the coffee study, even though we do not know how much bias is actually present in either study. Bad luck could explain a result significant at level 0.1 or a result significant at 0.0001, but much more bad luck would be required to explain the latter result. In parallel, hidden bias could explain a result sensitive at $\Gamma = 1.3$ or a result sensitive at $\Gamma = 6$, but much more hidden bias would be required to explain the latter result.

ADDITIONAL REFERENCES

ALTONJI, J. G., ELDER, T. E. and TABER, C. R. (2000). Selection on observed and unobserved variables: Assessing the effectiveness of Catholic schools. Working paper 7831, National Bureau of Economic Research.

ANGRIST, J. D. (1998). Estimating the labor market impact of voluntary military service using social security data on military applicants. *Econometrica* **66** 249–288.

ANGRIST, J. D. and DEHEJIA, R. (2001). When is ATE enough? Risk aversion and inequality aversion in evaluating training programs. Technical report, Columbia Univ.

ANGRIST, J. D. and IMBENS, G. W. (1995). Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *J. Amer. Statist. Assoc.* **9** 431–442.

ANGRIST, J. D. and KRUEGER, A. B. (1991). Does compulsory school attendance affect schooling and earnings? *Quarterly J. Economics* **106** 979–1014.

ANGRIST, J. D. and KRUEGER, A. B. (1999). Empirical strategies in labor economics. In *Handbook of Labor Economics* (O. Ashenfelter and D. Card, eds.) **3A** 1277–1366. North-Holland, Amsterdam.

ANGRIST, J. D. and KRUEGER, A. B. (2001). Instrumental variables and the search for identification: From supply and demand to natural experiments. *J. Economic Perspectives* **15** 69–85.

ANGRIST, J. D. and LAVY, V. (1999). Using Maimonides' rule to estimate the effect of class size on scholastic achievement. *Quarterly J. Economics* **114** 533–575.

BARNARD, J., DU, J., HILL, J. and RUBIN, D. B. (1998). A broader template for analyzing broken randomized experiments. *Sociological Methods and Research* **27** 285–317.

BEKKER, P. (1994). Alternative approximations to the distributions of instrumental variable estimators. *Econometrica* **62** 657–681.

BERTRAND, M., DUFLO, E. and MULLAINATHAN, S. (2001). How much should we trust differences-in-differences estimates? Working paper, MIT Dept. Economics.

BICKEL, P. J. (1982). On adaptive estimation. *Ann. Statist.* **10** 647–671.

BLACK, S. (1999). Do better schools matter? Parental valuation of elementary education. *Quarterly J. Economics* **114** 577–599.

BOUND, J., JAEGER, D. and BAKER, R. (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *J. Amer. Statist. Assoc.* **90** 443–450.

BRAUN, T. M. and FENG, Z. (2001). Optimal permutation tests for the analysis of group randomized trials. *J. Amer. Statist. Assoc.* **96** 1424–1432.

CAMPBELL, D. and STANLEY, J. (1963). *Experimental and Quasi-experimental Designs for Research*. Rand McNally, Chicago.

CHAMBERLAIN, G. (1984). Panel data. In *Handbook of Econometrics* (Z. Griliches and M. D. Intriligator, eds.) **2** 1248–1318. North-Holland, Amsterdam.

CHAMBERLAIN, G. and IMBENS, G. (1996). Hierarchical Bayes models with many instrumental variables. Working paper 204, National Bureau of Economic Research.

COPAS, J. B. (1973). Randomization models for matched and unmatched 2×2 tables. *Biometrika* **60** 467–476.

COX, D. R. and SNELL, E. J. (1989). *Analysis of Binary Data*, 2nd ed. Chapman and Hall, New York.

DEHEJIA, R. H. and WAHBA, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *J. Amer. Statist. Assoc.* **94** 1053–1062.

FRASER, D. A. S. (1954). Completeness of order statistics. *Canadian J. Math.* **6** 42–45.

FRIEDLANDER, D. and ROBINS, P. K. (1995). Evaluating program evaluations—new evidence on commonly used nonexperimental methods. *Amer. Econom. Rev.* **85** 923–937.

GOLDBERGER, A. S. (1991). *A Course in Econometrics.* Harvard Univ. Press, Cambridge, MA.

HAAVELMO, T. (1944). The probability approach in econometrics. *Econometrica* **12** (Suppl.) 1–115.

HAHN, J., TODD, P. and VAN DER KLAAUW, W. (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica* **69** 201–209.

HAMMOND, E. C. (1964). Smoking in relation to mortality and morbidity: Findings in first thirty-four months of follow-up in a prospective study started in 1959. *J. National Cancer Institute* **32** 1161–1188.

HECKMAN, J. J., ICHIMURA, H. and TODD, P. (1997). Matching as an econometric evaluation estimator: Evidence from a job training programme. *Rev. Econom. Stud.* **64** 605–654.

IMBENS, G. W. and ANGRIST, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica* **62** 467–475.

IMBENS, G. and ROSENBAUM, P. (2001). Randomization inference with an instrumental variable. Manuscript.

LALONDE, R. (1986). Evaluating the econometric evaluations of training programs with experimental data. *Amer. Econom. Rev.* **76** 604–620.

LEHMANN, E. (1975). *Nonparametrics: Statistical Methods Based on Ranks*. Holden-Day, San Francisco.

LEHMANN, E. L. and STEIN, C. (1949). On the theory of some nonparametric hypotheses. *Ann. Math. Statist.* **20** 28–45.

LIANG, K.-Y. and ZEGER, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73** 13–22.

LITTLE, R. J. A. and RUBIN, D. B. (1987). *Statistical Analysis with Missing Data*. Wiley, New York.

MOULTON, B. R. (1986). Random group effects and the precision of regression estimates. *J. Econometrics* **32** 385–397.

NELSON, C. and STARTZ, R. (1990a). The distribution of the instrumental variables estimator and its t-ratio when the instrument is a poor one. *J. Business* **63** S125–S140.

NELSON, C. and STARTZ, R. (1990b). Some further results on the exact small sample properties of the instrumental variable estimator. *Econometrica* **58** 967–976.

NEYMAN, J. (1935). Statistical problems in agricultural experimentation. *J. Roy. Statist. Soc. Suppl.* **2** 107–154.

NEWEY, W. K. (1990). Semiparametric efficiency bounds. *J. Appl. Econometrics* **5** 99–135.

ROBINS, J. M. (1988). Confidence intervals for causal parameters. *Statistics in Medicine* **7** 773–785.

ROBINS, J. M. (1989). The analysis of randomized and nonrandomized AIDS treatment trials using a new approach to causal inference in longitudinal studies. In *Health Service Research Methodology*: *A Focus on AIDS* (L. Sechrest, H. Freeman and A. Mulley, eds.) 113–159. National Center for Health

Services Research, U.S. Public Health Service, Washington, DC.

ROBINS, J. M. (1997). Causal inference from complex longitudinal data. *Latent Variable Modeling and Applications to Causality. Lecture Notes in Statistics* **120** 69–117. Springer, New York.

ROBINS, J. M. (1998). Correction for non-compliance in equivalence trials. *Statistics in Medicine* **17** 269–302.

ROBINS, J. M. (1999). Association, causation, and marginal structural models. *Synthese* **121** 151–179.

ROBINS, J., BLEVINS, D., RITTER, G. and WULFSOHN, M. (1992). *G*-estimation of the effect of prophylaxis therapy for pneumocystis carinii pneumonia on the survival of AIDS patients. *Epidemiology* **3** 319–336.

ROBINS, J. M., GREENLAND, S. and HU, F.-C. (1999). Rejoinder to comments on "Estimation of the causal effect of a time-varying exposure on the marginal mean of a repeated binary outcome." *J. Amer. Statist. Assoc.* **94** 708–712.

ROBINS, J. M. and ROTNITZKY, A. (2001). Comment on "Inference for semiparametric models: Some questions and an answer," by P. Bickel and J. Kwon. *Statist. Sinica* **11** 920–936.

ROBINS, J. M., SCHARFSTEIN, D. and ROTNITZKY, A. (2000). Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In *Statistical Models in Epidemiology, the Environment and Clinical Trials* (M. E. Halloran and D. Berry, eds.) 1–94. Springer, New York.

ROSENBAUM, P. R. (1988). Permutation tests for matched pairs with adjustments for covariates. *Appl. Statist.* **37** 401–411.

ROSENBAUM, P. R. (2001a). Stability in the absence of treatment. *J. Amer. Statist. Assoc.* **96** 210–219.

ROSENBAUM, P. R. (2001b). Effects attributable to treatment: Inference in experiments and observational studies with a discrete pivot. *Biometrika* **88** 219–231.

ROSENBAUM, P. R. (2002a). *Observational Studies*, 2nd ed. Springer, New York.

ROSENBAUM, P. R. (2002b). Attributing effects to treatment in matched observational studies. *J. Amer. Statist. Assoc.* **97** 183–192.

SCHARFSTEIN, D. O., ROTNITZKY, A. and ROBINS, J. M. (1999). Adjusting for non-ignorable drop-out using semiparametric non-response models. *J. Amer. Statist. Assoc.* **94** 1096–1120.

SKLAR, L. (2000). *Theory and Truth*: *Philosophical Critique within Foundational Science*. Oxford University Press, New York.

STAIGER, D. and STOCK, J. (1997). Instrumental variables regression with weak instruments. *Econometrica* **65** 557–586.

SULLIVAN, J. M. and FLANNAGAN, M. J. (2002). The role of ambient light level in fatal crashes: Inferences from daylight saving time transitions. *Accident Analysis and Prevention* **34** 487–498.

VAN DER LAAN, M. and YU, Z. (2001). Comment on "Inference for semiparametric models: Some questions and an answer," by P. Bickel and J. Kwon. *Statist. Sinica* **11** 910–917.

WRIGHT, P. G. (1928). *The Tariff on Animal and Vegetable Oils*. Macmillan, New York.

YITZHAKI, S. (1996). On using linear regressions in welfare economics. *J. Bus. Econom. Statist.* **14** 478–486.