Journal of Educational and Behavioral Statistics
Winter 2002, Vol. 27, No. 4, pp. 411–415

Discussion of "Estimation of Intervention Effects with Noncompliance: Alternative Model Specifications" by Booil Jo

Fabrizia Mealli
Università di Firenze

Donald B. Rubin
Harvard University

We thank the editors for the opportunity to offer our comments on this very readable article on recent work on noncompliance. Since Angrist, Imbens, and Rubin (1996), there has been an explosion of interest and activities in noncompliance related methods and applications, which include the bridging of work in different fields such as statistics, economics, epidemiology, sociology, and education. This article contributes nicely to this expanding literature by explicating various assumptions involving covariates that can be used to uniquely identify maximum likelihood estimates in place of exclusion restrictions.

Although we like the article very much, as discussants we focus on points designed to stimulate further discussion. Thus the casual reader may get the mistaken impression that we are more critical than we really are. The topics we address are: (a) summarizing results by "significant" versus "not significant"; (b) handling missing data by listwise deletion; (c) describing subpopulation differences as "impacts" or "effects"; (d) discussing the scientific plausibility of competing models.

Sensitivity of Significant/Not Significant Results

The author's main interest is sensitivity analysis, which we support, but the focus seems to be on sensitivity of "significant" and "not significant" results. For instance:

> The results from the model assuming the exclusion restriction show that the intervention assignment did not have a significant effect on the level of depression of compliers ($\widehat{ITT}_c = -0.361$, effect size = 0.498). The results from the model without assuming the exclusion restriction show that the intervention assignment had a significant effect on the level of depression of compliers ($\widehat{ITT}_c = -0.451$, effect size = 0.622). (p. 399)

Even if the −0.451 estimate were known without any sampling variability, the difference between it and the $\widehat{ITT}_c$ estimate under the exclusion restriction (0.080) would be only 0.4 standard error away from zero (see Table 2 in Jo, 2002). This insensitivity of $\widehat{ITT}_c$ holds despite the apparent superiority of the "without

exclusion" model ($\widehat{ITT}_n = 0.227$, $SE = 0.094$), suggesting the need for this extra parameter. There are more sophisticated ways to assess the different implications of such models; for example, our preference would be to employ the Bayesian paradigm, possibly using posterior predictive checks of the ability of the more parsimonious models to predict important aspects of the observed data (Rubin, 1984; Gelman, Meng, & Stern, 1996).

Dealing with Missing Data by Listwise Deletion

We were somewhat disappointed by Jo's general handling of missing data in the datasets, because the author was so careful when dealing with noncompliance, a special kind of missing data. The method known as "listwise deletion" is quite generally inferior to other methods (e.g., see Little and Rubin, 1987, chapter 3). There has been a tremendous amount of development in missing data methodology in the past quarter century (e.g., since Rubin, 1976; Dempster, Laird, & Rubin, 1977), and recent work has addressed the simultaneous complications of noncompliance and missing data (Frangakis & Rubin, 1999; Baker, 2000). A highly relevant publication in the context of a randomized educational intervention is Barnard, Frangakis, Hill, and Rubin (2002). In these publications, the assumption of latent ignorability of missing data plays a key role: for groups of subjects with the same true compliance status, the missing data are ignorable. Because the software needed to implement fully principled analyses of data with both missing values (in covariates and outcome measures) and noncompliance is not readily available (e.g., Barnard et al., 2002), we are sympathetic to restricting the analysis to people with covariates fully observed, because this limits the inference to subpopulations of people with complete covariates. However, also restricting the analysis to those with complete outcomes is in principle incorrect for any subpopulation, because this implies the comparison of outcomes under treatment for those units who would produce outcome data if treated, with outcomes under control for those units who would produce outcome data if control; this is not generally a comparison leading to causal effects because these two subpopulations are not the same (see Frangakis & Rubin, 2002).

Language that Tends to Confuse Subpopulation Differences and Treatment Effects

Our third point concerns language that can often be easily read to imply causal effects for variables that cannot be manipulated. We feel that it is especially important to avoid such language in the context of social sciences, where it has too often been misused. As Jo explains clearly, the basic idea underlying the proper analysis of data with noncompliance is to distinguish between (a) subpopulations defined by the true compliance status and (b) the effect of treatments on outcomes within those subpopulations. An example of a potentially deceptive statement is:

> The intervention assignment had a positive impact on compliers if they were highly motivated and married. However, high motivation and being married had a negative impact on individuals who would have complied with the intervention if offered, but were assigned to the control condition. (p. 401)

What we believe is actually meant by these two sentences can be explained by the example in Table 1 where, in order to simplify the presentation, depression is assumed to have only two levels, high and low.

TABLE 1
Depression Level of True Compliers by Treatment Assignment and Subpopulation

	Intervention	
True compliers subpopulation	Control	Treatment
NM^2	Low depression	Low depression
M^2	High depression	Low depression

Note. NM^2 = not married and/or not highly motivated, M^2 = married and highly motivated

The meaning of the first sentence is that for the subpopulation of people who are true compliers and highly motivated and married, the intervention reduces depression: this is indeed an impact (causal effect) that can be observed by comparing the two cells in the M^2 row of Table 1. What is meant by the second sentence instead is that, if all members of the subpopulation of true compliers are assigned control, those who are highly motivated and married will have a depression level that is higher than the subpopulation of true compliers who are not highly motivated and married. This is not an "impact" of an intervention but a description of a difference between two subpopulations under the same treatment condition, which can be observed by comparing the two cells in the Control column of Table 1. Similar confusing language appears later and in other places, where a difference between subpopulations under the same treatment condition is described as an "impact" or "effect".

Scientific Rationale for Various Specifications

Although we applaud the specification of different identifying models, we wished that the resulting comparisons of estimates across the models had included more discussion of the scientific plausibility of the models, and thereby of the resulting estimates. Also, Jo's analysis of alternative assumptions involving covariates is restricted to the case where, in addition to monotonicity (i.e., there are no defiers), there are no always-takers. Always-takers are likely to be present in many observational studies and randomized trials—for example in randomized encouragement designs (Hirano et al., 2000), where people in the control arm are not encouraged to take the treatment (flu shot), but it is nevertheless accessible. It is thus interesting to see how Jo's results generalize to that case.

For example, consider the common setting and allow always-takers, i.e.,

$$C_i = \begin{cases} c \text{ (complier)} & \text{if } D_i(1) = 1 \text{ and } D_i(0) = 0 \\ n \text{ (never-taker)} & \text{if } D_i(1) = 0 \text{ and } D_i(0) = 0 \\ a \text{ (always-taker)} & \text{if } D_i(1) = 1 \text{ and } D_i(0) = 1 \end{cases}$$

Also, assume there is only one binary covariate X, and, in addition to the notation introduced in the "Modeling CACE with Covariates" section, let π_{ai} be the probability that the ith individual with covariate x_i is an always-taker, and let $a_i = 1$ if $i \in C(a)$ and $a_i = 0$ if $i \in \{C(c) \cup C(n)\}$. Equation 11 can be modified to represent the outcome variable for an individual with compliance status c_i, n_i, and a_i:

$$\begin{aligned} Y_i = {} & \alpha_n n_i + \alpha_c c_i + \alpha_a a_i + \gamma_n n_i Z_i + \gamma_c c_i Z_i + \gamma_a a_i Z_i \\ & + \gamma_{nx} n_i Z_i x_i + \gamma_{cx} c_i Z_i x_i + \gamma_{ax} a_i Z_i x_i + \lambda_n n_i x_i \\ & + \lambda_c c_i x_i + \lambda_a a_i x_i + \varepsilon_{in} n_i + \varepsilon_{ic} c_i + \varepsilon_{ia} a_i \end{aligned} \tag{11a}$$

where the additional parameters and residual have the same interpretation for always-takers as the already introduced parameters and residuals have for compliers and never-takers. Based on equation (11a), eight directly estimable population means can be expressed in terms of unrestricted model parameters as:

$$\begin{aligned} \mu_{1n,X=0} &= \alpha_n + \gamma_n \\ \mu_{1n,X=1} &= \alpha_n + \gamma_n + \gamma_{nx} + \lambda_n \\ \mu_{0a,X=0} &= \alpha_a \\ \mu_{0a,X=1} &= \alpha_a + \lambda_a \\ \mu_{00,X=0} &= \pi_{n,X=0}\alpha_n + \pi_{c,X=0}\alpha_c \\ \mu_{00,X=1} &= \pi_{n,X=1}(\alpha_n + \lambda_n) + \pi_{c,X=1}(\alpha_c + \lambda_c) \\ \mu_{11,X=0} &= \pi_{a,X=0}(\alpha_a + \gamma_a) + \pi_{c,X=0}(\alpha_c + \gamma_c) \\ \mu_{11,X=1} &= \pi_{a,X=1}(\alpha_a + \gamma_a + \gamma_{ax} + \lambda_a) + \pi_{c,X=1}(\alpha_c + \gamma_c + \gamma_{cx} + \lambda_c), \end{aligned}$$

where $\mu_{jh,X=k}$ is the population mean potential outcome if $Z = j$, $D(j) = h$ and $X = k$.

Because the probabilities of being of a certain compliance status are directly estimable at both values of X, these equations show that the twelve model parameters $\{\alpha_n, \gamma_n, \gamma_{nx}, \lambda_n, \alpha_c, \gamma_c, \gamma_{cx}, \lambda_c, \alpha_a, \gamma_a, \gamma_{ax}, \lambda_a\}$ cannot be identified without some assumptions. The exclusion restriction for always-takers and never-takers (i.e., $\gamma_a = \gamma_{ax} = \gamma_n = \gamma_{nx} = 0$) reduces the number of parameters to eight and allows the identification of CACE. The same is true for the assumption of constant effects of covariates (i.e., $\lambda_a = \lambda_n = \lambda_c$, and $\gamma_{ax} = \gamma_{nx} = \gamma_{cx}$), whereas the assumption of additivity of treatment assignment effect (i.e., $\gamma_{ax} = \gamma_{nx} = \gamma_{cx} = 0$) reduces the number of parameters to nine, which is not sufficient to identify CACE: when there are always-takers, the additivity assumption needs to be combined with other assumptions, for example with an exclusion restriction for either the never-takers or the always-takers. Thus, speculation about which assumptions are more or less plausible when always-takers are also present is even more difficult without substantial experience with both fitting the models and applying them in substantive areas. One small contribution of ours into

this difficult territory is Mealli and Rubin (in press), where we describe how different settings and experimental designs can have very different implications for the plausibility of competing assumptions. An example that makes the exclusion restriction but discusses its plausibility at some length is Barnard et al. (2002).

In conclusion, Jo's article represents a fine contribution to this demanding and growing area of statistical research, and, moreover, it introduces an interesting new class of tools for applied researchers, which illustrates a profitable path for further work.

References

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association, 91,* 444–455.

Baker, S. G. (2000). Analyzing a randomized cancer prevention trial with a missing binary outcome, an auxiliary variable, and all-or-none compliance, *Journal of the American Statistical Association, 95,* 43–50.

Barnard, J., Frangakis, C., Hill, J., & Rubin, D. B. (2002). School choice in New York City: A Bayesian analysis of an imperfect randomized experiment, in C. Gatsonis, B. Carlin and A. Carriquiry (Eds.), *Case studies in Bayesian statistics,* Vol. V, (pp. 3–97). New York: Springer-Verlag.

Dempster, A., Laird, N., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B, 39,* 1–38.

Frangakis, C., & 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*(2), 365–379.

Frangakis, C., & Rubin, D. B. (2002). Principal stratification in causal inference, *Biometrics, 58*(1), 21–29.

Gelman, A., Meng, X., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies, *Statistica Sinica, 6,* 733–807.

Hirano, K., Imbens, G. W., Rubin, D. B., & Zhou, X. H. (2000). Assessing the effect of an influenza vaccine in an encouragement design. *Biostatistics, 1,* 69–88.

Little, R. J. A., & Rubin D. B. (1987). *Statistical analysis with missing data.* New York: John Wiley & Sons.

Mealli, F., & Rubin, D. B. (in press). Assumptions when analyzing randomized experiments with noncompliance and missing outcomes. *Health Services Outcome Research Methodology.*

Rubin, D. B. (1976). Inference and missing data, *Biometrika, 63*(3), 581–592.

Rubin, D. B. (1984). Bayesian justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics, 12,* 1151–1172.

Authors

FABRIZIA MEALLI is Associate Professor of Statistics, University of Florence, Viale Morgagni 59, Florence, Italy; mealli@ds.unifi.it. She specializes in methods for causal inference in observational and experimental studies, hierarchical models, models for transition data and simulation-based estimation methods.

DONALD B. RUBIN is John L. Loeb Professor of Statistics, Harvard University, 1 Oxford Street, Cambridge, MA 02138; rubin@stat.harvard.edu. He specializes in causal inference, missing data, multiple imputation, Bayesian modeling, and applications in diverse settings.