Identification of Causal Effects in the Presence of Nonignorable Missing Outcome Values

Mattei, Alessandra
Department of Statistics, University of Florence
Viale Morgagni, 59
50134 Florence, Italy
E-mail: mattei@ds.unifi.it

Mealli, Fabrizia
Department of Statistics, University of Florence
Viale Morgagni, 59
50134 Florence, Italy
E-mail: mealli@ds.unifi.it

Pacini, Barbara
Department of Statistics and Mathematics, University of Pisa
Via Ridolfi 10,
56124 Pisa, Italy
E-mail: barbara.pacini@sp.unipi.it

Introduction

In the potential outcome approach to causal inference (Rubin, 1974, 1978), a causal inference problem is viewed as a problem of missing data, where the assignment mechanism is explicitly posed as a process for revealing the observed data. The assumptions on the assignment mechanism are crucial for identifying and deriving methods to estimate causal effects. A commonly invoked identifying assumption is unconfoundedness (Rosenbaum and Rubin, 1983), which usually holds by design in randomized experiments. However, even under such assumption, inference on causal effects may be invalidated due to the presence of post-treatment complications, such as noncompliance (Angrist et al., 1996), *truncation by death* (Zhang and Rubin, 2003) and missing outcome values (Frangakis and Rubin, 1999). Here, we focus on identifying causal effects in the presence of missing outcome values, primarily due to nonresponse. Because nonresponse occurs after treatment assignment, respondents are not comparable by treatment status: the observed and unobserved characteristics of respondents in each treatment group are likely to differ and may be associated with the values of the missing outcome, making the missing mechanism nonignorable (e.g., Rubin, 1976; Little and Rubin, 2002).

A relatively recent approach to deal with post-treatment complications within the potential outcome approach is principal stratification, introduced by (Frangakis and Rubin, 2002). In this paper, we apply principal stratification in order to develop a novel approach to deal with nonignorable missing outcome values without imposing any restriction on treatment effect heterogeneity. We rely on the presence of a binary instrument for nonresponse and provide new sufficient conditions for partial identification of causal effects for subsets of units (unions of principal strata) defined by their nonresponse behavior in all possible combinations of treatment and instrument values. The framework allows us to clarify and discuss substantive behavioral assumptions, which may differ from those required by other approaches.

Principal Stratification and its Role for Causal Inference

Principal stratification was first introduced by Frangakis and Rubin (2002), in order to address post-

treatment complications, i.e., events which cannot be ignored when inferring on causal effects, and require adjusting for them, although conditioning on their observed values (e.g., including them in a regression model) may lead to estimating parameters which are not, in general, causal effects. We first introduce *potential outcomes* for one post-treatment variable, Y, and a binary treatment, T. If unit i in the study ($i = 1, \ldots, N$) is assigned to treatment $T_i = t$ ($t = 1$ for treatment and $t = 0$ for no treatment), we denote with $Y_i(T_i = 1) = Y_i(1)$ and $Y_i(T_i = 0) = Y_i(0)$ the two potential outcomes, either of which can be observed depending on the value taken by T. A causal effect of T on Y is defined, on a single unit, as a comparison between $Y_i(1)$ and $Y_i(0)$. The fact that only two potential outcomes for each unit are defined reflects the acceptance of the stable unit treatment value assumption (SUTVA; Rubin, 1980) that there is no interference between units and that both levels of the treatment define a single outcome for each unit. We also denote with $S_i(t)$ the post-treatment potential variable, which represents a response indicator for $Y_i(t)$: the observation of $Y_i(t)$ is missing if $S_i(t) = 0$. To simplify the notation, we will drop the i subscript in the sequel.

Consider the potential response indicators $S(0)$ and $S(1)$. Within each cell defined by values of the covariates, units under study can be stratified into four latent groups, named Principal Strata, according to the joint values $(S(0), S(1))$: stratum 11 : $S(1) = S(0) = 1$ comprises those who would respond under treatment and under control; stratum 10 : $S(1) = 1, S(0) = 0$ comprises those who would respond under treatment but not under control; stratum 01 : $S(1) = 0, S(0) = 1$ comprises those who would not respond under treatment but would respond under control; and stratum 00 : $S(1) = S(0) = 0$ comprises those who would never respond regardless of treatment assignment.

The principal stratum membership, $G = \{11, 10, 01, 00\}$, is not affected by treatment assignment by definition, so it only reflects characteristics of subjects, and can be regarded as a covariate, which is only partially observed in the sample (Angrist et al., 1996).

Note that, although causal effects of the treatment are well defined for the whole population, and thus for all latent groups, only in stratum 11 we can observe $Y(1)$ for some respondent units under treatment and $Y(0)$ for some other respondent units under control. On the contrary, in the other three strata we can observe the outcome only for respondents in at most one of the two treatment arms. What makes stratum 11 interesting is the fact that only in this stratum can we hope to learn something about the causal effect, even if it may not be an interesting stratum *per se*. Although conceptually a different problem, the identification issues in estimating the effect for the stratum of the always respondents are analogous to those related to the identification of the effect of the treatment effect on the always survivors in studies suffering from *truncation by death* (e.g., Zhang and Rubin, 2003).

Some of the assumptions that may be invoked to deal with nonresponse essentially assume that nonresponse is ignorable. For instance, both the missing completely at random (MCAR) assumption and the weaker missing at random (MAR) assumption describe ignorable missing data mechanisms (Rubin, 1976; Little and Rubin, 2002)[1], which are convenient because they allow us to avoid an explicit probability model for nonresponse. If the response probability depends on both observable and unobserved characteristics, then nonresponse is nonignorable.

In the econometric literature alternative ways to deal with nonresponse include instrumental variable assumptions (e.g., Manski, 2003). Plausible instrumental variables for nonresponse can be found relatively easily (unlike finding instruments for other intermediate variables): data collection characteristics, for example, are likely to affect the response probability but not the outcome values. Characteristics of the interviewer (e.g., gender), interview mode, length and design of the questionnaire can be convincing instruments for nonresponse (see, for example, Nicoletti, 2010).

We use a binary instrument for nonresponse in a causal inference framework; in this context complications arise because we have to deal simultaneously with the nonresponse behavior under

[1]Ignorability requires that, in addition to MAR, the parameters of a MAR missing data process be distinct from those of the data distribution.

Table 1. Principal strata with a binary treatment and a binary instrument for nonresponse

G	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
$S(0,0)$	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
$S(0,1)$	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1
$S(1,0)$	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1
$S(1,1)$	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1

treatment and under control.

Identifying Causal Effect with Nonignorable Nonresponse on the Outcome and an Instrumental Variable

Principal stratification with a binary treatment and a binary instrument for nonresponse

We assume that the distributions that are asymptotically revealed by the sampling process are known, or can be consistently estimated, thereby not taking account of specific statistical inference problems related to estimation in finite samples.

In addition to treatment T, whose causal effect on Y is still our primary interest, suppose that units are exposed to an additional treatment Z which is related to nonresponse S but unrelated to the outcome Y. For instance, consider the following simplified example from Janssens et al. (2008), as a potential empirical scenario. A randomized trial to assess the effects of a campaign for AIDS prevention is conducted. Let T be a binary treatment which represents the offer of free condoms. T is randomly assigned to a group of individuals at high risk of HIV infection. The post-assignment HIV infection status Y may be missing due to refusal of some patients to participate in the HIV-test; presumably non-participants are more likely to be HIV-positive than individuals who take the test. The identity of nurses, Z, can be reasonably used as an instrument for nonresponse if (a) the propensity to take the HIV test varies with the nurses; (b) nurses, whose identity cannot affect the result of the test (HIV infection status), are randomly assigned to patients.

In this example the variable Z can be regarded as a treatment, because an intervention on it can be contemplated. The assignment of two binary treatments, T and Z, implies that four potential outcomes can be defined for each post-treatment variable, the primary outcome, Y, and the response indicator, S, in our case: $S(t,z)$, $Y(t,z)$ for $t = 0,1$ and $z = 0,1$. Principal strata are defined according to the joint values of $S(0,0)$, $S(0,1)$, $S(1,0)$, and $S(1,1)$. Because the response indicator is binary, the stratum membership, G, takes on 16 values (see Table 1). Unlike the case discussed in the previous section with no instrument, there is more than one stratum from which we can hope to learn something about the causal effect of T on Y, i.e., all the strata where some units respond under treatment and some units respond under control ($G = 6,7,8,10,11,12,14,15,16$). In this setting, estimands of interest are causal effects for some (union) of these strata. Note that these strata include subjects who are more responsive to the instrument, i.e., are more inclined to respond if properly "encouraged".

Basic Assumptions

Due to the presence of two treatments, assumptions are required on the compound assignment mechanism. Both treatments are assumed randomized conditional on a set of pre-treatment covariates, so that:

Assumption 1

$$T, Z \perp\!\!\!\perp S(0,0), S(0,1), S(1,0), S(1,1), Y(0,0), Y(0,1), Y(1,0), Y(1,1) \mid X$$

and $$0 < Pr(T = t, Z = z \,|\, X) < 1 \quad (t = 0, 1; z = 0, 1).$$

Assumption 1 amounts to assuming that, within cells defined by the values of pre-treatment variables X, the treatment, T, and the instrument, Z, are randomly assigned or, at least, are assigned independently of the relevant post-treatment variables. The second condition is an overlap assumption which guarantees that in large samples we can find treated and control units, as well as units with the different values of the instrument, for all values of X. Define $S^{obs} = \sum_{t=0,1} \sum_{z=0,1} \mathbb{1}\{T = t\}\mathbb{1}\{Z = z\}S(t, z)$ and $Y^{obs} = \sum_{t=0,1} \sum_{z=0,1} \mathbb{1}\{T = t\}\mathbb{1}\{Z = z\}Y(t, z)$ if $S^{obs} = 1$ and $Y^{obs} =$ missing otherwise, where $\mathbb{1}(.)$ is the indicator function.

Assumption 1 implies the following: (a) $S(0,0), S(0,1), S(1,0), S(1,1) \perp\!\!\!\perp T, Z \mid X$, so that G is guaranteed to have the same distribution in each treatment$-$instrument arm, within cells defined by pre-treatment variables; (b) $Y(0,0), Y(0,1), Y(1,0), Y(1,1) \perp\!\!\!\perp T, Z \mid S(0,0), S(0,1), S(1,0), S(1,1), X$, so that potential outcomes are independent of the treatment and the instrument given the principal strata. While it is in general improper to condition on S^{obs}, units exposed to different treatment and/or instrument levels can instead be compared conditional on a principal stratum, G; (c) $Y(0,0), Y(0,1), Y(1,0), Y(1,1) \perp\!\!\!\perp T, Z, S^{obs} \mid S(0,0), S(0,1), S(1,0), S(1,1), X$, so that, conditional on a principal stratum, comparison of respondents exposed to different treatment and/or instrument levels leads to valid inference on causal effects. For the sake of notational simplicity we will omit an explicit indication of conditioning on X in the sequel.

In order to characterize Z as an instrument, we propose the following exclusion-restriction assumption:

Assumption 2 $Y(0,0) = Y(0,1)$ and $Y(1,0) = Y(1,1)$,

which says that the value of the instrument is unrelated to the outcome. We further require that the instrument Z has some effect on S, both under treatment and under control:

Assumption 3 $E(S(0,1) - S(0,0)) \neq 0$, and $E(S(1,1) - S(1,0)) \neq 0$.

Main Identification Results

We now analyze how the presence of an instrument can be exploited to achieve identification of some causal estimands. Some identification assumptions can be stated as forms of monotonicity of S:

Assumption 4 $S(t,0) \leq S(t,1) \quad \forall\, t$,

and

Assumption 5 $S(0,z) \leq S(1,z) \quad \forall\, z$.

Assumption 4 relates to the response behavior with respect to the instrument: for a fixed treatment level, units responding when $Z = 0$ would respond also when $Z = 1$. Assumption 5 relates to the response behavior with respect to the treatment: for a fixed value of the instrument, units responding under control would respond also when treated. These assumptions may often be plausible.

Assumptions 4 and 5 reduce the number of strata to 6 (strata $1, 2, 4, 6, 8$, and 16), implying that the strata containing information on causal effects are strata 6, 8 and 16, so that the goal is to isolate these three strata from the remaining ones.

Under Assumptions 1 through 5, we can derive large sample bounds on the causal effect of T for the union of strata 6, 8, and 16, which include units *reacting to the instrument* under control and/or under treatment (strata 6 and 8) and always respondents (stratum 16). For sake of simplicity, henceforth we focus on average treatment effects. Note that the same identification strategies could be used to identify the entire outcome distribution under both values of the treatment for particular strata.

Let $P_{s|t,z} = \Pr(S^{obs} = s \mid T = t, Z = z)$, $s = 0, 1$, $t = 0, 1$ and $z = 0, 1$, be the conditional distribution of the observed response indicator given the treatment and instrument values, and define $\pi_j = \Pr(G = j)$, $j = 1, 2, 4, 6, 8, 16$. In addition, define $E_{tz1}(Y^{obs}) = E(Y^{obs} \mid T = t, Z = z, S^{obs} = 1)$ and let $E^{\leq\alpha}_{tz1}(Y^{obs})$ and $E^{\geq\alpha}_{tz1}(Y^{obs})$ be the conditional expectations of Y^{obs} in the α ($0 < \alpha < 1$) fraction of the observed respondents ($S^{obs} = 1$) assigned to $T = t$ and $Z = z$ with the smallest and largest values of the outcome variable, Y, respectively. The following proposition is proved in the extended web-version of this paper.

Proposition 1 *If Assumptions 1–5 hold, then the following bounds on the average treatment effect for the union of strata 6, 8 and 16 can be derived:*

$$E^{\leq\pi_{6,8,16|111}}_{111}(Y^{obs}) - E_{011}(Y^{obs}) \leq$$
$$E(Y(T=1) - Y(T=0) \mid G \in \{6,8,16\}) \leq E^{\geq\pi_{6,8,16|111}}_{111}(Y^{obs}) - E_{011}(Y^{obs}), \quad (1)$$

where $\pi_{6,8,16|111} = \Pr(G \in \{6,8,16\} \mid T = 1, Z = 1, S^{obs} = 1) = \frac{P_{1|0,1}}{P_{1|1,1}}$.

The sampling process allows us to identify the conditional distributions, $P_{s|t,z}$, the conditional expectations $E_{tz1}(Y_i^{\text{obs}})$, and the conditional lower and upper trimmed means $E^{\leq\alpha}_{tz1}(Y_i^{\text{obs}})$ and $E^{\geq\alpha}_{tz1}(Y_i^{\text{obs}})$, $0 < \alpha < 1$. Therefore finding estimators for the bounds defined in Proposition 1 is relatively straightforward. For instance, a moment-based estimator can be derived by replacing the means of Y and the strata proportions by their sample counterparts:

$$\widehat{P}_{s|t,z} = \frac{\sum_{i=1}^n \mathbb{1}(T_i{=}t)\mathbb{1}(Z_i{=}z)\mathbb{1}(S_i^{obs}{=}s)}{\sum_i \mathbb{1}(T_i{=}t)\mathbb{1}(Z_i{=}z)} \qquad \widehat{E}_{tz1}(Y_i^{obs}) = \frac{\sum_{i=1}^n \mathbb{1}(T_i{=}t)\mathbb{1}(Z_i{=}z)S_i^{obs}Y_i^{obs}}{\sum_{i=1}^n \mathbb{1}(T_i{=}t)\mathbb{1}(Z_i{=}z)S_i^{obs}}$$

$$\widehat{E}^{\leq\alpha}_{tz1}(Y^{obs}) = \frac{\sum_{i=1}^{[n\alpha]} \mathbb{1}(T_i{=}t)\mathbb{1}(Z_i{=}z)S_i^{obs}Y^{obs}_{(i)}}{\sum_{i=1}^{[n\alpha]} \mathbb{1}(T_i{=}t)\mathbb{1}(Z_i{=}z)S_i^{obs}} \qquad \widehat{E}^{\geq\alpha}_{tz1}(Y^{obs}) = \frac{\sum_{i=n-[n\alpha]+1}^{n} \mathbb{1}(T_i{=}t)\mathbb{1}(Z_i{=}z)S_i^{obs}Y^{obs}_{(i)}}{\sum_{i=n-[n\alpha]+1}^{n} \mathbb{1}(T_i{=}t)\mathbb{1}(Z_i{=}z)S_i^{obs}},$$

$s = 0, 1$, $t, z = 0, 1$, where $[n\alpha]$ is the largest integer not greater than $n\alpha$, and $Y^{obs}_{(i)}$, $i = 1, \ldots, n$, are the ordered statistics. In small samples, bounds can be wrapped in confidence bands to account for sampling variability in various ways (e.g., Imbens and Manski, 2004).

The benefit of using an instrument for nonresponse is due to the fact that more information can be extracted from the data about the causal effects of the treatment. Specifically, in the presence of an instrument for nonresponse, strata containing information on the causal effects are strata 6, 8 and 16, which in general include a larger proportion of units than the group of the always respondents without instrument (stratum 11). The bound on the average treatment effect for the always respondents, $E(Y(1) - Y(0) \mid S(1) = S(0) = 1)$, depends on the proportion of the always respondents (see, for instance, Manski, 2003 and Zhang and Rubin, 2003), as well as the bound on $E(Y(1) - Y(0) \mid G \in \{6, 8, 18\})$ depends on the proportion of strata 6, 8, and 16; therefore, when the instrument is not available or is ignored, we have a loss of information. In other words, the presence of an instrument for nonresponse provides information on the causal effect also for subjects who, without the instrument, would not respond under either the standard treatment or the active treatment (i.e., principal strata 10 and 01), but would respond regardless treatment assignment when assigned to $Z = 1$. If causal effects are homogeneous, this implies using more information to estimate the same causal estimands (leading also to a better precision if the instrument is used in a parametric estimation approach). If causal effects are heterogeneous, this implies estimating an average effect for a larger proportion of units, which has higher chances to mimic the behavior of the target overall population. Therefore, when an instrument for nonresponse is available, using it might help identification and estimation of causal effects. Our discussion suggests that an instrument for nonresponse should be included as a design variable in the planning phase of the study design.

Bounds in Proposition 1 can be tightened if additional assumptions are introduced, in order to either reduce the number of strata or state the equivalence of the distribution of Y across some

strata. Point identification can be also achieved, by combining Assumptions 1-5 with these additional assumptions. These results are available in the extended web-version of this paper.

Concluding Remarks

In this paper, we tackled the problem of identifying treatment effects when some outcome values are missing. Identification results were obtained relying on a binary instrument for nonresponse, within the principal stratification framework. We proposed a set of sufficient assumptions allowing identification of causal estimands for some subpopulations of units (union of principal strata) defined by the nonresponse behavior under all possible combinations of treatment and instrument values. Our results suggest that an instrument for nonresponse should be included as a design variable in the planning phase of the study design, and it should be considered in drawing causal inference in the presence of missing outcome data, whenever it is available.

Using principal stratification, the result of inference is usually a *local* causal effect. An issue that often arises regarding the principal stratification approach is that we cannot univocally identify the group the causal effect refers to. Note, however, that the fact that proper causal effects can only be identified for latent subgroups of units is a limitation created by the missing mechanism, rather than a drawback of the framework of principal stratification. In this paper, the focus on these subgroups was primarily driven by our goal of providing valid causal effect estimates in the presence of nonignorable missing data under a set of credible assumptions. These subgroups may not be *ex ante* the most interesting ones, but the data is in general not informative about effects for other subgroups without extrapolation.

Supplementary Material: The extended web-version of this paper is available under the Working Papers link at the website http://www.ds.unifi.it.

REFERENCES (RÉFERENCES)

Angrist, J.D., Imbens , G.W., Rubin, D.B. (1996). Identification of causal effects using instrumental variables (with discussion). *J. Am. Statist. Assoc.* **91**, 444-472.

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

Frangakis, C.E., Rubin, D.B. (2002). Principal stratification in causal inference. *Biometrics* **58**, 191-199.

Imbens, G.W., Manski, C.F. (2004). Confidence intervals for partially identified parameters. *Econometrica* **72**, 1845-1857.

Little, R.J.A., Rubin, D.B. (2002). *Statistical Analysis with Missing Data.* 2nd Ed, NY: John Wiley.

Manski, C.F. (2003). *Partial Identification of Probability Distributions.* Springer-Verlag.

Janssens, W., van der Gaag, J., Rinke de Wit, T. (2008). Pitfalls in the estimation of HIV prevalence. *Amsterdam Institute for International Development Research Series*, RS 08-03.

Nicoletti, C. (2010). Poverty analysis with missing data: Alternative estimators compared. *Empirical Econ.* **38**, 1-22.

Rosenbaum , P., Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41-55.

Rubin, D.B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psycol.* **66**, 688-701.

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

Rubin, D.B. (1978). Bayesian inference for causal effects. Ann. Statist. 6, 34-58.

Rubin, D.B. (1980). Discussion of "Randomization analysis of experimental data: the Fisher randomization test" by D. Basu, *J. Am. Statist. Assoc.* **75**, 591593.

Zhang, J., Rubin, D.B. (2003). Estimation of causal effects via principal stratification when some outcomes are truncated by "death". *J. Educ. Behav. Statist.* **28**, 353–368.