Principal Stratification, Partial Contingency Table, and Statistical Leverage

Yu Xie and Susan Murphy, University of Michigan

In "Principal Stratification Designs to Estimate Input Data Missing due to Death," Frangakis, Rubin, An, and MaKenzie (hereafter FRAM) propose an analysis to do what may seem impossible: to recover input data that are missing due to death and then use the (observed and missing) input data to predict death. FRAM show that, under certain assumptions, this can be done with the introduction of an additional variable, "treatment," that possesses certain desirable properties.

We organize our comments as follows. First, we present the logic behind FRAM's analysis from the perspective of contingency table analysis. Second, with insights from this perspective, we will consider the implications of FRAM's analysis. Third, we discuss some considerations that should be taken into account in practice.

From Principal Stratification to Statistical Leverage

It appears that FRAM's analysis hinges on the notion of principal stratification (Angrist, Imbens, and Rubin 1996; Frangakis and Rubin 2002), i.e., the idea that discrete subpopulations, or strata, have distinct patterns of response to a treatment (called Z in the paper). For simplicity, we focus on the main case discussed by FRAM: there are only two strata: a stratum of "always survivors" regardless of the treatment, and another stratum of "protectable" patients whose lives can be saved, but who cannot be harmed, by the treatment. Here the principal stratification assumption can be replaced by a less restrictive assumption:

Assumption 2'. If treatment is Z=1 then the person must be alive at 3 months (S=1) or, equivalently, P[S=1|Z=1]=1.

Assumption 2' is true if FRAM's assumption 2 is true, but assumption 2' invokes neither potential outcomes nor principal stratification. The crucial ignorability assumption 1 of FRAM is that the assignment of Z is independent of both stratum membership and input data (A), conditional on covariate X (see below for more on this assumption).

We note that covariate X plays no special role in FRAM's paper except to make the ignorability assumption plausible. Thus, the discussion that follows is conditional on X. In terms of time ordering, the

input data, A, exist prior to the critical event (here an injury), the treatment Z occurs shortly after the injury but prior to death, and S denotes death (here coded as 1 if the subject is alive 3 months after the injury and 0 otherwise). Note that S is always observed so S is S^{obs} in the FRAM paper, and furthermore $S=ZS(1) + (1-Z)S(0)$ where S(z) denotes the potential outcome when the treatment Z=z. Z is always observed, but A is observed only if S=1.

Since Z, S, and A are all binary, we can capture their joint distribution with a 3-way cross-classified contingency table, shown in Table 1. We use F_{ijk} to denote the frequency count in the cross-classified table for the cell Z=i, S=j, and A=k, with i =0, 1, j =0, 1, and k =0, 1. We use the plus sign, "+", in the subscript to denote the subtotal for summation over a particular subscript. Two features stand out in Table 1. First, since all patients who received the treatment (Z=1) survived, the third row (representing Z=1, S=0) contains structural zeros. Second, while we know the subtotal of the first row, representing the situation of Z=0, S=0, we do not know the distribution of A in that row. Indeed, recovering this distribution from patients who had died before the interview is a main research objective here. Due to these two unique features, Table 1 differs from the usual 2x2x2 contingency table. We call such as table as Table 1 a "partial contingency table."

How can we recover the distribution of A in the row in the partial contingency table? We make use of the ignorability assumption in FRAM's approach and our assumption 2' . Assumption 2' sets the third row (Z=1, S=0) to structural zeros so that $F_{1+0}= F_{110}$, and $F_{1+1}= F_{111}$. The independence assumption for the relationship between Z and A means that the odds of A=1 versus A=0 is the same across the two different values of Z. We thus have the following constraint:

$$F_{1+1} /F_{1+0} = F_{111}/ F_{110} = (F_{001} + F_{011})/(F_{000} + F_{010}). \quad (1)$$

We then add to equation (1) the known information that

$$F_{000} + F_{001} = F_{00+} . \quad (2)$$

We can easily solve equations (1) and (2) for two unknowns, F_{000}, F_{001}. In Table 2, we present our numerical results based on the information provided by FRAM for their data from the National Study on the Costs and Outcome of Trauma Centers (NSCOT). There may be small discrepancies between our results and the actual

results, since we recovered counts from FRAM's original results in percentages. Following FRAM, we also treat the illustrative example as if we have population data and thus do not consider statistical inference issues.

From the approach of a contingency table analysis, we see why FRAM's analysis works. We think that our contingency table approach is more intuitive and more straightforward. One advantage of our approach is that equation (1) clearly reveals how the missing information pertaining to the distribution of A for the dead group (Z=0, S=0) is recovered: it compensates the distribution of A among untreated survivors (Z=0, S=1) so that the combined distribution equals that of the treated group (Z=1). Everything else being equal, the distribution of A among the dead (Z=0, S=0) moves in the same direction as the distribution of A in the treated group (Z=1) and in the opposite direction from that of the distribution of A among untreated survivors (Z=0, S=1). We are clearly borrowing information from other related groups. It is as though we are able to move an enormous object by a mechanical lever. Thus, FRAM's approach is an exemplary case of using "statistical leverage."

Implications for Research Objectives

In FRAM's analysis using statistical leverage, an additional treatment variable can recover the missing information about input data. We showed earlier that we were able to fill in the cells of missing data in Table 2 for their numerical example. How well does the recovered information serve the original objectives of the substantive research? To answer this question, let us visit the research objectives that FRAM's analysis is intended to help achieve. The abstract clearly states the two research objective: (1) "to measure 'input' variables, which describe the period before the critical event, and to characterize the distribution of input variables in the cohort"; and (2) "to measure 'output' variables, primarily mortality, after the critical event, and to characterize the predictive (conditional) distribution of mortality given the input variables in the cohort."

If we are to take the first objective literally, it is not necessary to fill in the missing data, as we did in Table 2. By assumption, the distribution of the input variable (A) is independent of Z. Thus, the distribution

of the input variable (A) conditional on Z also describes the unconditional distribution of the input variable (A), as the following is true by the ignorability assumption (assumption 1):

$$P(A = 1|Z=1) = P(A = 1|Z=0) = P(A = 1). \quad (3)$$

Of course, this does not tell us P(A|Z=0, S=0), which can only be recovered after missing values are estimated.

Achieving the second research objective requires an additional assumption; here we use assumption 2'. If we take the stated objective literally, the researcher is interested in the following quantities for the entire population:

$$P(S=0|A=k),\ k=0,1 \quad (4)$$

We can further decompose these quantities by treatment status (Z):

$$\begin{aligned} P(S=0|A=k) &= P(S=0|A=k, Z=0)\ P(Z=0|A=k) + P(S=0|A=k, Z=1)\ P(Z=1|A=k), \quad (5)\\ &= P(S=0|A=k, Z=0)\ P(Z=0) + P(S=0|A=k, Z=1)\ P(Z=1)\\ &= P(S=0|A=k, Z=0)\ P(Z=0). \end{aligned}$$

Note that we obtained the second line of equation (5) by using the independence assumption and the last line of equation (5) by using the information that all subjects survive if treated (Z=1). Because P(Z=0) is unrelated to A, this term is cancelled in the formula for the relative risk, the ratio of conditional probabilities:

$$[P(S=0|A=1)] / [P(S=0|A=0)] = [P(S=0|A=1, Z=0)] / [P(S=0|A=0, Z=0)]. \quad (6)$$

Equation (6) can be estimated using our partial contingency table approach by

$$[F_{001}/(F_{001} + F_{011})] / [F_{000}/(F_{000} + F_{010})]. \quad (7)$$

We present our numerical results using equation (7) for the illustrative example.

Two comments concerning the second research objective are in order. First, if we wish to know the mortality rates by the values of the input variable, it is necessary to know the proportion not receiving treatment in the population, P(Z=0). When the researcher is interested only in the relative risk, or odds-ratio, by the input variable, P(Z=0) can be ignored. Second, the appearance that the group of treated persons (Z=1) do not seem to affect the relative risk in equation (6) is misleading, as these persons affect the estimation of the missing information as part of the “statistical leverage” discussed earlier.

Practical Considerations

Although FRAM's analysis allows the researcher to uncover missing data that are not missing at random through the power of statistical leverage, implementation is not trivial. Below, we discuss some considerations that researchers should take into account when adapting the analysis in practice.

First of all, the researcher needs to carefully consider the treatment variable Z. A number of questions arise:

(a) Is Z an existing treatment in practice or a new intervention as part of the study design?

(b) If the researcher does not manipulate Z, are we comfortable with the assumption that Z and A are independent conditional on covariates?

(c) If the administrator knows the effectiveness of Z, what prevents her/him from "over-prescribing" the treatment to reduce deaths?

(d) Does the effectiveness of Z vary with time, location, population, or the proportion being treated?

While the first two questions are straightforward, as they are concerned with the ignorability assumption, the last two questions need some discussion.

Let us generalize the idea of principal stratification. Suppose the population is not divided into two strata--those who always survive and those who are helped by treatment--but numerous subclasses characterized by the degree to which treatment Z helps survival. That is, the counter-factual response function for person i is a continuous score, depending on the person's latent response function R, $R_i = S_i(1) - S_i(0)$. Under the common assumption of monotonicity (Angrist, Imbens, and Rubin 1996; Frangakis and Rubin 2002), we specify that $0 \leq R_i \leq 1$. Further imagine that because the administrator of Z knows additional information (unknown to the researcher) about patients' and hospitals' conditions, he or she would assign Z to those patients who would benefit most from the treatment. That is, we entertain the possibility that the likelihood of receiving Z is correlated with the amount of treatment effect R. When this is the case, increasing the proportion of Z necessarily results in lowering the average treatment effectiveness of treatment Z, as the composition of the stratum receiving treatment (Z=1) has changed from having a higher average R score towards having a lower average R score (i.e., from benefiting more on average to

benefiting less on average). This discussion illustrates a practical difficulty with the principal stratification approach in general: we do not know individuals' memberships in the various strata, as the existence of the strata can only be inferred from the group level. Thus, we may view principal strata either as distinct subpopulations with distinct response patterns or as aggregations of heterogeneous individuals with somewhat similar response patterns. The latter, nominal perspective is consistent with the view of heterogeneous treatment effects at the individual level. Our concern is that if we accept the nominal perspective, policy or technological changes can change the proportion and at the same time the composition of the group of subjects receiving treatment. Properties of principal strata, nominally defined, are thus not fixed and are subject to change.

We next consider the role of the covariates X. From the perspective of assumptions needed to make FRAM's analysis work, X precedes both A and Z and indeed makes them independent of each other conditional on X. From the perspective of data collection, X was not provided in the interview, as it would, like A, then be truncated by death. Conceptually at least, one would like to condition on a rich set of covariates before accepting the conditional independence assumption. For example, we would like to know a person's medical history, demographics, and family socioeconomic status. Needless to say, it is not possible to condition on them if they are considered part of A instead of X. In other words, an input variable A and a covariate X differ in two respects: (1) X is observed, whereas A is only partially observed; (2) X is to be conditioned on, whereas A and Z are assumed to be conditionally independent. Strict association of partial observability with the conditional assumption is more a practical convenience than a necessary condition justified by science. Conceptually at least, it is possible that we may wish to condition on covariates that may only be partially observed. However, not observing them in practice would force us to convert them into input data (A) that would then need to satisfy the independent assumption.

There is no easy and magic solution to this problem. We recommend that the researcher collect more and better data as a possible remedy. One possibility is to use administrative records (such as the death certificates and medical records). Another possibility is to interview surviving family members for proxy reports. In general, better data can yield far more statistical information than can be achieved through

statistical leverage. In the approach of pushing for better observed data, the boundary between input data and covariates is blurred.

Conclusion

The FRAM analysis is intuitively appealing, and relatively easy to implement. One of the most interesting features of the analysis is that it allows the researchers to impute data that do not satisfy the ignorability assumption alone, but under a **model** that satisfies ignorability.

If the input data were to satisfy the ignorability assumption, the distribution of the input data would be the same between survivors and non-survivors. This is clearly implausible and is rejected by FRAM. Even after introducing a new treatment, FRAM do **not** assume ignorability in the distribution of the input data between survivors and the non-survivors within treatment status. Rather, the ignorability assumption is imposed on the two-way marginal association between the treatment variable and the input variable. This restriction allows FRAM to recover missing input data among non-survivors.

How well FRAM's analysis will work in practice is a substantive question that will depend on concrete applications. At the minimum, the new analysis provides alternative estimates so as to characterize the distribution of input data and the association between the input data and the risk of deaths. This exercise is informative even if one does not necessarily believe that the underlying model is correct, for the alternative estimates provide some sensible and plausible bases for the researcher to critique and improve upon. For this and many other reasons previously discussed, we recommend this article to all who are interested in the topics it covers: missing data, causal inference, principal stratification, and partial contingency table.

References

Angrist, J. D., G.W. Imbens, and D.B. Rubin. 1996. "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association* 91(434): 444-455.

Frangakis, Constantine E and Donald B. Rubin. 2002. "Principal Stratification in Causal Inference." *Biometrics* 58 (1):21–29.

Table 1: Partial Three-Way Cross-Classified Frequency Table by Z, S, and A.

Z	S	A=0	A=1	Sub-Total
0	0	F_{000}	F_{001}	F_{00+}
	1	F_{010}	F_{011}	F_{01+}
1	0			
	1	F_{110}	F_{111}	F_{11+}

Note: Entry F_{ijk} refers to the frequency cross-classified by Z=i, S=j, and A=k. Cells shaded gray are not observed but estimated. Cells shaded by slanted grids are not allowed by assumption.

Table 2: Numerical Example using the NSCOT Data for the Partial Three-Way Cross-Classified Frequency by Z, S, and A, by Covariate X.

X = low injury severity

Z	S	A=0	A=1	Sub-Total
0	0	3	14	17
	1	257	72	329
1	0			
	1	6	2	8

Note: Cells shaded gray are not observed but estimated. Cells shaded by slanted grids are not allowed by assumption. Estimated relative risk of death is: 14.1.

X = high injury severity

Z	S	A=0	A=1	Sub-Total
0	0	18	6	24
	1	95	5	100
1	0			
	1	10	1	11

Note: Cells shaded gray are not observed but estimated. Cells shaded by slanted grids are not allowed by assumption. Estimated relative risk of death is: 3.4.