

Invited Paper Session "Design and Analysis of Evaluation Studies", Discussion

Markus Frölich

Universität St. Gallen, Department of Economics

Dufourstr. 48, SIAW

9000 St. Gallen, Switzerland

markus.froelich@unisg.ch

The three papers by Riccio/Bloom/Hill, Pradhan/Rawlings and Purdon/Skinner are applied microeconomic evaluation studies using a quasi-experimental framework (also called *selection-on-observables* or ignorable treatment approach). Verma/Betti suggest a new sampling design for sampling two types of units that differ in their geographical density.

The paper by Purdon and Skinner is a typical example of a microeconomic impact evaluation by propensity score matching, estimating the effect of participating in *active labour market policies* on subsequent *employment probability* in the *UK*. An interesting aspect of their study is the way how the data is collected. Since the available administrative database is not sufficiently informative, they supplement it with survey data. Importantly, they start surveying the eligible population *before* anyone had entered in treatment, with a 36% non-response rate. About 2000 of the respondents entered treatment during the following 8 months and were surveyed afterwards on various outcome variables. From the non-treated, 2000 individuals were selected by propensity score matching and surveyed. These face-to-face surveys on the outcome variables suffered from a non-response rate of 30%, giving a total non-response rate of 55%. This high non-response rate demonstrates the shortcomings of using surveys compared to a situation where informative administrative data were available (as is the case in many European countries today). On the other hand, starting to collect survey data *ex-ante* (before individuals can be identified as treated or untreated) is certainly superior to surveying retrospectively treated and controls only *ex-post*. Collecting data *ex-ante* does not only avoid recall bias, it may also mitigate differential non-response rates among treated and controls. In addition, covariate information is available for *all* respondents of the first survey (baseline population), and this covariate information can help to analyze non-random non-response in later stages. For example, the 30% non-response rate in the later face-to-face survey is somewhat unsettling, and non-response could be analyzed like an outcome variable given the prior covariate information on respondents and non-respondents. *Ex-ante* surveying also eases the collection of pre-treatment outcome data, which would enable difference-in-difference matching or to 'test' for selection-on-unobservables bias. With *ex-ante* surveying the baseline population is well defined, and dynamic aspects of e.g. changes in eligibility status can better be incorporated.

Pradhan and Rawlings also use propensity score matching to estimate the effects of infrastructure projects financed by the *Social Investment Fund* (FISE) on various *health* and *educational* outcomes in *Nicaragua*. In contrast to the previous paper, the treatment does not apply directly to individuals but rather to whole communities (areas of influence). A household living in a community where a FISE-project was carried out is treated, whereas a household living in a community without such project is non-treated. Since the project communities are non-randomly selected, selection bias has to be taken into account. Because no pre-intervention data is available, identification relies on adjustment for observed covariates (individual and community characteristics). For data collection they make use of the Living Standards Measurement Survey (LSMS) that was conducted in 1998. By modifying the sampling scheme of the LSMS, households living in project communities and selected neighbouring non-project communities (with similar community characteristics) are oversampled. Through this approach, data for the evaluation can be collected at a low cost, and it is ensured that the same questionnaire was

used and at the same time for treated and control observations. When collecting the treatment sample in the areas where projects had been carried out, a distinction between direct and potential (but not direct) beneficiaries was made, and the former group (i.e. those who actually had used the new infrastructure) was oversampled.

Two comparison groups were used for the impact evaluation: (A) all households of the supplemental survey living in communities close to but outside the area of influence of a project, (B) all LSMS households not living in a project community. When using the comparison group A, the treatment impacts were computed as differences in mean outcomes. When using comparison group B, propensity score matching (without replacement) and with *exact* matching on region (and characteristics such as school-age child) was used to adjust for the different characteristics among the treated and the LSMS sample.

As pointed out by Pradhan and Rawlings, the estimated treatment effects are sometimes contradictory for the two comparison groups. This seems to warrant further analysis, since the availability of 2 control groups can be used to assess the plausibility of the selection-on-observables assumption. Therefore the same (propensity score) matching approach should be used for both comparison samples. There seems to be no reason, why a different treatment effect estimator is used for comparison group A than for B. Although comparison sample A was matched on region and some school and health centre characteristics, it may still be rather different in other characteristics. In addition, a more efficient matching estimator could be used and the stratification weights in the treated group be accounted for (as discussed below). With two comparison samples available, a very useful analysis it to estimate mutually the adjusted means for both comparison samples. In other words, to test whether the treatment effect of being in sample A versus B is zero. If it is significantly different from zero, one (or both) group is not a valid comparison group, and the selection-on-observables assumption becomes less credible.

Bloom, Hill and Riccio estimate the impact of *welfare-to-work* programmes for female single parents in the USA on subsequent *earnings*. They are particularly interested in how the impact of these programmes depends on the way the programme is implemented. Randomized experiments were conducted at 59 different sites in the USA, giving 59 different impact estimates, which range from - 1400\$ to 4200\$. This raises the question why some offices perform much better or worse than others and whether these differences are related to the way how these programmes are run at the different sites. Such kind of analysis is very important for policy improvements.

The way how the programme is implemented on site is characterized through 6 indicators: case-load per staff member, degree of personal attention to the client, degree of monitoring, emphasis of quick job entry, and inconsistencies between staff and supervisors. These 6 characteristics do vary substantially between sites. When comparing the impacts across sites, the analysis is necessarily non-experimental, and confounding factors have to be taken account of. Bloom, Hill and Riccio use a two-level hierarchical linear model (to account for clustering) to combine the individual level data on treatment status, earnings outcomes and personal characteristics with the office level data on implementation characteristics, services composition and economic environment. In essence, for the treated they estimate the relationship between earnings and individual characteristics, office level characteristics and economic environment. For the non-treated the coefficients on office level characteristics are restricted to zero! The basic motivation for these zero restrictions seems to be, that the characteristics of programme implementation should not affect the randomized-out. This should be further discussed, because many of the randomized-out (about 20%) also receive certain job search assistance and education services where they interact with the treated. In addition, the providers of these services might also react on how the welfare-to-work programme is implemented. (It could be made clearer that the impact of the randomized treatment actually consists of four different effects: the counselling by the welfare-to-work officer, the increased likelihood of receiving services, differences of the

quality in the services the treated and non-treated have access to, and the differential impact that these services may have on treated and non-treated.) If indeed we expect a zero effect on the non-treated, including the office level characteristics in the equation for the non-treated provides an opportunity to test the plausibility of the selection-on-observables assumption, i.e. whether all confounding factors have been taken account of.

In the data used by Bloom, Hill and Riccio, information on the randomization status is available for all individuals, but information on actual services receipt is available only for a (random) subsample. The analysis is therefore restricted to estimating the impact of being randomized-in versus out. A supplemental analysis using also the information on services received could reveal further interesting insights. One could think of (at least) 4 different groups: 1) randomized-in who received services, 2) randomized-in who did not receive services, 3) randomized-out who received services, 4) randomized-out who did not. Comparing, for example, the outcome difference (adjusted for confounders) between group 2 and 4 to the outcome difference between group 1 and 3, could be indicative of the effect of implementation characteristics on the effectiveness of the actual services provided. Of course, receipt of services is not randomly assigned and confounding factors need to be accounted for. The results might thus be less reliable, but could nevertheless be indicative of which subprocesses (counselling, monitoring, service referral, service provision) are affected most by the way the welfare-to-work programme is implemented.

This last paper demonstrated that even with randomized experiments, non-experimental evaluation methods are often necessary for supplemental analysis. Randomized experiments are important and too infrequently used in Europe, but are often very difficult (and expensive) to conduct when more than 2 or 3 different treatment varieties (e.g. differences in implementation philosophies) shall be compared, which is very important for policy improvements. Various nonparametric evaluation methods have been developed in recent years and the use of matching estimators to adjust for different covariate distributions has become frequent. These matching estimators are very often used for *two purposes* at the same time (for example in Purdon/Skinner and Pradhan/Rawlings): For data collection and for impact estimation. Often a large pool of potential control observations is available and surveying all potential controls would be very costly. Therefore matched pairs of treated and controls are formed and only the matched control observations are surveyed. The average treatment effect is then estimated by the mean outcome difference within pairs. However, since both purposes are rather distinct, different types of matching should be used. While pair-wise matching is a useful strategy for selecting a sample of non-treated for further data collection, it is inefficient for the subsequent estimation. For estimating the treatment effect, *after* data has been collected, a more efficient estimator should be used, e.g. kernel matching, since there is no reason to restrict the matching estimator to integer weights (see e.g. Imbens (2003) or Frölich (2004)).

Another theme relevant to these papers is the relation between sampling design and estimation. The first issue is how to account for non-random sampling in the treatment effect estimator. Incorporating sampling weights in a weighting estimator is straightforward. Incorporating weights in propensity score matching is also possible, but requires various modifications (not only in the estimation of the propensity score, but also in the matching estimator, see e.g. Frölich 2003). The second issue concerns the reverse direction: How should data be sampled to obtain a precise estimate at minimum surveying costs? This question has received rather little research attention since most research on *non-experimental* treatment evaluation supposes that data has already been collected. Research analyzing the precision loss due to different forms of stratified sampling would be very valuable for guiding data collection.

Sampling design is the focus of the paper by Verma and Betti, who propose a particular design for collecting data on two types of units with different geographical density: a general population (households) and a population of institutional units, such as health care facilities. Such type of data are often of relevance in evaluation studies to estimate, for example, the

impact of health services. Household level data on the use of health services and health outcomes is often combined with data on the quality of health care facilities. Employing the same stratified sampling scheme for the population and for the facilities survey, however, may be inappropriate, since for the population survey often rather small areas for the first-stage stratification are used (for lower surveying costs), which may not even contain a single health care facility. Hence, larger areas should be used for the facilities survey than for the population survey, to obtain an acceptable sample size for the former. However, the sample design should be such that the sample of surveyed facilities is 'close' to where the surveyed population lives. Verma and Betti suggest to base the facilities survey on the sampling areas selected for the population survey, but to survey also the facilities in the neighbouring areas. They discuss various ways how these neighbouring areas could be defined in practice. They also derive the sampling weights that are associated with the surveyed facilities and show that a simpler expression can be derived if the definition of the neighbourhood satisfies certain properties.

REFERENCES

Frölich, M. (2004) "Finite Sample Properties of Propensity-Score Matching and Weighting Estimators", forthcoming in *Review of Economics and Statistics*.

Frölich, M. (2003) "Propensity score matching without conditional independence assumption - with an application to the gender wage gap in the UK", mimeo, Universität St.Gallen.

Imbens, G. (2003) "Semiparametric Estimation of Average Treatment Effects under Exogeneity: A Review", mimeo, UC Berkeley.