

Report on: “Consistent estimation of pseudo panels in presence of selection bias”

I suppose a more suitable title would have been “Consistent estimation in pseudo panels in the presence of selection bias.”

The authors claim that they have a procedure which is designed to address a version of the selectivity problem (the need to estimate the outcome equation of interest on a subsample rather than the full sample) that emerges when using a pseudo-panel instead of a true panel. Since the individuals in each cross-section are different, methods designed for true panel data (e.g. within estimator) cannot be used.

Given this motivation, one expects to see an application where the “panel” aspect matters, i.e. a problem where dynamics is of interest. For example, we might want to see how outcomes differ across cohorts, so we turn to a pseudo-panel. Alternatively we might want to engage in an investigation where failure to control for cohort effects would mislead us.

What we find instead is a static problem: the parameter of interest is the rate of return to education, which is assumed to be constant over time. Since years of schooling may be correlated with unobserved heterogeneity, we cannot use LS; we need to adjust for this endogeneity. In addition since participation (to be exact remunerated employment) is non-random, we have to adjust for potential selectivity.

So far so good, but the rest of the paper strays from the path I envisioned. Leaving the selectivity issue aside, just above equation (2) the authors seem to argue that replacing the individual fixed effect by a linear function of cohort dummies (i.e. a mean cohort effect) plus a “nice” error (deviation from the mean) a la Deaton (1985) amounts to using IV. This is not true. Deaton proposes aggregation, which amounts to using cohort specific means of all the variables instead of the individual values. Moffit (1993) shows that this is equivalent to using cohort indicators as instruments. But equation (3) in the paper does not achieve this. It surely is not equivalent to IV, and further, the authors talk about the need to find “proper instruments” for the original covariates. Thus it is not clear what has been achieved by including cohort dummies in the model. Furthermore the assumption that the cohort dummies can legitimately be excluded from the main equation can be challenged.

Turning to selectivity, additional questions arise. A Heckman-Lee type two equation model is envisioned, but presence of unobserved heterogeneity in the selection equation (12) introduces additional challenges. This problem is not resolved. On p. 5 we are told that the individual fixed effect of the selection equation will be subjected to the Deaton decomposition; but somehow it shows up linearly in the estimating equation (17)! Furthermore the authors propose using cohort variables as instruments for handling selectivity (that is, as explanatory variables in the selection equation). Since these are already in the outcome equation (3), they do not provide the sought after exclusion restrictions.

All said, it is not clear (i) how the methodology summarised on p. 6 surmounts the problems in hand, and (ii) how it relates to the earlier “derivations” and arguments.

I thought the empirical section could help, but this was not the case. It is difficult to reconcile Table 1 with the methodology on p.6. Evidently there are two excluded variables which achieve identification (wealth dummy and household size) in the selection equation. Presumably these are endogenous (recall the fixed-effect in the selection equation) so IV probit is used, with 5 cohort dummies as instruments. Then Heckman-Lee lambda (inverse Mills ratio) is constructed, and included in the main equation. The third step of the methodology indicates that IV rather than LS will be used; but where will the instruments come from? We are not told. Furthermore, a marital status dummy which did not appear in the original cross-section equations is included in the pooled equation. Consequently we never know whether the change in the magnitude of the rate of return coefficient is due to selectivity (a type of omitted variable bias), or “another” omitted variable bias.

It is possible that the authors have something to offer here; but I for one was unable to discover it. I recommend starting with the 2010 J of Econometrics article by Semykina and Wooldridge, and thinking about its pseudo-panel version. With the help of the synthesis in Moffit and Ridder (2010), and thinking about the problem in a GMM framework, it may be possible to write a credible empirical piece.