A comment on the use of "cluster" corrections in the context of panel data

Martin Wittenberg

School of Economics and DataFirst

Martin.wittenberg@uct.ac.za

August 2013

Many researchers have asked what "cluster" correction would be appropriate for the second (or subsequent) waves of the NIDS panel. This is not a straightforward issue and requires some clarity as to what the correction is designed to achieve.

1. What is the "cluster" correction good for?

"Cluster" corrections, such as those achieved by Stata's `svyset` command, are designed to produce correct standard errors for most estimators if the original sample design has been two-stage sampling in which a primary sampling unit (colloquially referred to as a "cluster") is sampled first and then units (households and individuals in this case) are sub-sampled within them. It is important to understand under what conditions these corrections are necessary.

Let's examine the case of a multiple regression. We need to assume that the process which governs the outcome ***y*** of individual *i* who happens to be in cluster *c* can be written as:

$$y_i = \beta_1 + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \eta_c + \varepsilon_i$$

where $x_2,\ldots,x_k$ are variables that are measured, while η_c represents unmeasured factors that are common to individuals in the cluster. The idiosyncratic individual factors are given by ε.

If we were to analyse our sample with the default assumptions of every statistical package (that our sample was extracted by means of simple random sampling), we would ignore the fact that individuals from the same cluster are more alike than two individuals drawn at random – due to their common "cluster effect" η_c. When we calculate the standard errors we would think that we have more knowledge about the random variation within the population than we really do. With the wrong standard errors we would find "significant" relationships more readily than we should.

Basically if we do not attempt to correct our standard errors for clustering we are making the very strong assumption that there are no "cluster effects" in our data, i.e. individuals from the same sampling unit are as alike on the outcome (controlling for the observables $x_2,\ldots,x_k$) as two individuals drawn at random from the population. Empirically this is just not true for many of the outcomes that we are interested in.

2. "Clusters" in a panel context

The simple story outlined above becomes more complicated when we have panel data. Thus far we could be completely agnostic about what the common cluster effect η_c was about. There are several types of processes that could give rise to common cluster effects, *inter alia*

a) **Place-bound** effects – individuals living in the same neighbourhood share a common infrastructure, common amenities, distance to services
b) **Shared social background** – people tend to sort themselves (or get sorted by discriminatory pressures) into neighbourhoods that are marked by common culture, values, language and attitudes – certainly more alike than individuals drawn at random from the population as a whole

c) **Peer effects** – to the extent to which people interact more within their neighbourhoods than they do with people elsewhere, they can influence each other in ways that make people from an area more alike than individuals from different areas.

Depending on how we view the social process that we are investigating, we would need to deal with the "cluster correction" in different ways. In the case of place-bound processes, we would need to use the current neighbourhood as the "cluster" variable. To the extent to which people move out of their original locations and become spread across South Africa the need for such a correction would diminish over time, although we should at minimum ***svyset*** households as our "cluster" variable.

Social background, however, presumably moves with the individual. Norms and culture do change, but, we might assume, not as quickly as neighbourhood circumstances. In this case the "cluster of origin" would be the appropriate variable to use – two individuals originating in the same cluster would still be more alike than two randomly picked individuals, even if they are no longer residing in the same neighbourhood.

Peer effects, we might presume, depend on actual local interactions, so someone who has migrated out of an area would presumably no longer be exposed to the same influences. This case would end up more similar to the place-bound case considered earlier.

3. What should we do about new individuals in the panel?

The "temporary sample members" that are enumerated as part of the NIDS panel create additional difficulties. If we view the "cluster effects" as being largely place-bound, then there are no major issues, since we would just assign their current location to them. If, on the other hand we take the view that cluster effects are more likely to be based on common background, culture, values, upbringing then we face the problem that we have no historical information on them. The simplest assumption to make is that co-residence is based on assortative processes, so that TSMs will be most similar to the "continuing sample members" that they coreside with. They should therefore "inherit" the cluster from their coresident CSMs.

One additional complication that should be borne in mind, is that even strongly inculcated norms and practices will change with time. It is unlikely that the "cluster effect" η_c will be unvarying over time, particularly if individuals migrate out of their original area and end up coresiding with other individuals.

4. Recommendations for practice

The discussion above was intended to highlight the fact that "correcting for cluster design" is tricky for subsequent waves. Depending on how one views the underlying social processes different strategies can be equally plausible.

Given the short time period between waves 1 and 2 it is probably most plausible to view the "historical clusters" as still capturing much of the unmeasured common influences, i.e. it is difficult to imagine that common social background would have ceased to matter, even for people who have migrated out. I would suspect that this argument would still hold for wave 3.

It is worth noting that Angrist and Pischke (2009) have drawn attention to the fact that "robust standard errors" can sometimes be less conservative than standard errors calculated under the assumption of simple random sampling. Basically the logic underpinning the robust corrections is based on asymptotic arguments which may fail in finite samples. They suggest that one run the analyses with both robust standard errors and the standard ones and then report the larger of the two (i.e. the smaller t-values). That strikes me as sensible advice in this context too – run the analyses with clusters set to the "inherited" clusters, to households only and not set at all and then report the analysis with the most conservative t-values.

References:

Angrist, J.D. and Pischke J-S (2009) ***Mostly Harmless Econometrics: An Empiricist's Companion***, Princeton N.J.: Princeton University Press