An application of weighting approaches to assess the sensitivity of business survey estimates to "the unit problem"

Paul Smith
S3RI, University of Southampton, SO17 1BJ, UK p.a.smith@soton.ac.uk

Abstract

"The unit problem" in business statistics is that there is no consistent, natural unit for a business structure. Administrative processes quite often mandate the creation of particular units, but there is often no easy way to relate them to each other. Since the statistical view of units for sampling purposes is often derived from administrative systems, it can be quite difficult to define what population is being estimated, or to assess how this differs from the target concept.

In this paper we consider a variety of weighting approaches as possible estimation strategies for complex units in a business survey. Standard estimation approaches use units as defined on the business register, including their auxiliary information. In some structural surveys, information is gathered on more than one type of unit, for example on enterprises and local units. Basic classification information is available at both levels, so it is possible to construct an estimator at either level separately, or, using the integrated weighting approach of Lemaître and Dufour (1989), for both levels simultaneously. I will examine what sort of effect this has on estimates and the associated standard errors. Where surveys repeat on the same populations and there is some trading of units back and forth, it is also possible to conceive of the whole as a longitudinal study. There are some real challenges in treating a succession of cross-sectional samples as a longitudinal approach, but I investigate the possibility that changes in the population resulting from acquisition/disposal of whole 'units' (for some suitable definition) can be investigated using generalised weight share methods (Lavallée 2007) to estimate the impact of structural differences.

Introduction

"The unit problem" in business statistics is that there is no consistent, natural unit for a business structure. Businesses are composed of different types of units, at different levels, and there are often webs of relationships and types of control between different units, and a range of autonomy in the management of separate units. This leads to complex structures of ownership and relationships. Administrative processes quite often mandate the creation of particular units – for tax, employment law, and health and safety, for example – but there is often no easy way to relate units from one system to units in another system. In the UK there are examples where a single PAYE unit (for employee-related tax) covers all the employees across several production units (which pay value added tax, VAT).

In order to make official surveys practical, it is necessary to impose a statistical view of a business structure, generally derived from administrative systems. This uses some rules to derive a reasonably consistent data structure from the administrative units, and to classify them in an appropriate system. Classification can be different at different levels of a hierarchical business structure, and there are rules for how higher level classifications are derived from lower level ones, using a 'wholly or mainly' [with activity X] classification system (see Smith 2013 p172 for a worked example). Even this type of structure may not be unique, because there are several possible variables which can be used when measuring "mainly" – in order of preference generally value added, turnover, employment – though there may be other situations where other variables are also interesting. All of these processes provide a statistical population which can be used with reasonable consistency for a range of official statistics.

However, they do result from an application of an assumed structure, and therefore result from a particular model of the world of businesses.

If the target concept is defined as the statistical population concept, then there is no problem, and this is generally the approach that statistical offices (NSIs) take. However, it can be quite difficult to define what population of units is being estimated in more specific terms, or to assess how this differs from a target concept concept defined in another way (eg economically).

In this paper we examine ways of dealing with differences between different target concepts measured on the same population, using these differences as initial estimates of the size of errors induced by the unit problem. We start with the issue of estimation level, and then consider what happens as business structures evolve with the birth, death, acquisition and disposal of units through time. In order to structure the arguments, we will concentrate on differences among two types of statistical units, the reporting unit (RU), which is approximately a business, possibly operating in multiple locations, under common control, and the local unit (LU), which is a single location. The methods could be applied to units defined in many different ways.

Estimating at different levels

The RU and LU structure gives us options for estimation. Many business surveys use the RUs as the main units for estimation, and then the whole activity is included with the classification and other properties of the RU. There are also examples where the LUs are the basis for estimates. Historically, turnover and employment statistics in the UK were produced by different departments using different methods, and as a result turnover statistics (which used sampling much earlier) have been estimated from the RUs and employment statistics (which were based on a periodic employment census until 1992) have been estimated from LUs.

It becomes a natural question to ask what the difference between these two approaches is; this difference clearly captures an element of error in the units problem from classifying and estimating at different levels.

If we have information on the whole population (eg from a business register), then (implicitly) estimating for that population will give every unit in the population a weight of 1. Then the difference is $diff_{pop} = \sum_{RU \in A} 1.x_j - \sum_{LU \in A} 1.x_l$ for a particular set of interest A. Where A = the whole population, this difference is clearly zero, but for particular industries, the difference represents the difference in total activity using the RU classification from total activity using the LU classification.

This is all very well for register variables, but we typically do not have information on the whole population of businesses by both methods, and may want to estimate this difference from sample data. There is a challenge over whether the data are available at both levels, but let us assume that they are (the Business Register and Employment Survey in the UK is an example of a survey that has both kinds of information). Then we can calibrate to register information at either RU level or at LU level, and produce estimates from which we can estimate the difference $diff_{samp} = \sum_{RU \in A} w_j x_j - \sum_{LU \in A} w_l x_l$.

However, these estimators do not use all of the available information; we can calculate weights that calibrate to register information on LUs and RUs simultaneously, using the method of Lemaître & Dufour (1987). This removes part of the difference between the estimates arising from differences in the auxiliary information, and gives $diff_{LD} = \sum_{RU \in A} w_j^{LD} x_j - \sum_{LU \in A} w_l^{LD} x_l$.

To illustrate this process we take a synthetic simple example population with 100 local units split among 53 reporting units. The RUs are divided into a take-all stratum (with 100 or more employees) and a sampled stratum. 25 units are sampled over both strata, with the sampled stratum varying in size according to how much of the sample is CE.

Method	Register employment (known)	Turnover (estimated)
population total (known)	5958	*65431*
RU estimation	5941 *	65766
LU estimation	5958	65724
Lemaitre-Dufour RU & LU estimation	5958	66249

Table 1: Calibrated estimates for total turnover using initial weights derived from various weighting models. Numbers in italics are known only because the population is known in this example, and would not be seen in practice. *one poststratum has no sample units.

NACE	1	2	3	4	5
RU pop total	18378	9005	37650	253	146
RU	18260	9431	37910	164	0
L-D RU NACE	17956	9454	38528	311	0
LU pop total	16018	8446	38429	372	2165
LU	15814	8487	38927	276	2220
L-D LU NACE	15684	8618	39425	230	2293

Table 4: NACE category breakdown for estimates in Table 3. Note that the breakdown by RU classification and the breakdown by LU classification are different in general, so RU and LU rows are given separately for ease of comparison.

In the example given, the small size of the population and sample means that two of the RU-level constraints need to be dropped – one automatically because it has no RUs in the sample and therefore attempts to calibrate a zero to a positive value (which doesn't work), and one which has such small samples that the weights cannot cope and there is no solution. Once this is done, the calibration works, but produces one negative weight. The weights are substantially more variable.

The difference in the turnover population totals (only known as I have a simulated population) varies by NACE from about 120 to 2300, and these are reflected in the estimates. Comparing the estimates gives an estimator of the difference. We can also consider that the different estimation models, particularly the Lemaître-Dufour approach which uses all the information on auxiliary variables together, reflect an element of the differences involved in imposing a particular model of the business structure.

Longitudinal weighting of changing business structures

The approach to weighting for different levels of business gives us some information about the error in a static population from imposing a particular statistical organisation on the information available. However, business structures are also dynamic, with whole businesses being sold to become part of an existing business, or some changes in parts of businesses, being started up, closed down or transferred between different RUs (a disposal and an acquisition). This dynamic structure is hard to deal with consistently in longitudinal estimation (particularly in the sample case where it is very difficult to keep track of the original sample selection probabilities), but there are methods which are

designed to make weighting in these sorts of populations practical. We use the generalised weight share method (GWSM) of Lavallée (2007) as a way to deal with these challenges.

Now we look to see the difference between estimation on an initial population structure of units with estimation on a later population structure of units. Again, if we have information on both RUs and LUs we can examine the effect. In real situations the change in structure from changes in ownership will be confounded with changes in variables, which are measured at two (or more) time points; for simplicity let's take the time points as t_0 and t_1. In order to remove the effect of the changing variable we can calculate the difference using data from only t_0 applied with structures at both times (say s_0 and s_1). This gives $diff_{GWSM} = \sum_{s_0} w_{l0}^{GWSM} x_l - \sum_{s_1} w_{l1}^{GWSM} x_l$, now calculated with the LU data.

For example, let's take the same simple population example as before. Again for simplicity, let's take a population without births and deaths (though the method can be extended to these naturally, with some difficulties over time periods, as births only have data in the second period, and deaths only have data in the first period. We ignore this issue for the moment). We take five of the LUs, and reallocate them to RUs (with a small chance that they will end up returning to the same RU); these LUs change hands between RUs, some of them representing a change of ownership of a single LU, some moving from one RU to another leaving both parent RUs intact, but with different LU compositions. This is quite typical of what would be measured by a longitudinal or rotating survey of RUs.

Using one such example, with calibration of the LU register employment to the known totals within LU NACE classes, we obtain the results in Table 3.

Method	Register employment (known)	t_0 turnover (estimated)
population total (known)	5958	*65431*
t_0 LU estimation	5958	65619
t_1 LU estimation	5958	65641

Table 3: Calibrated estimates for total turnover using initial weights derived from the generalised weight share method. Numbers in italics are known only because the population is known in this example, and would not be seen in practice.

NACE	1	2	3	4	5
population total	*16018*	*8446*	*38429*	*372*	*2165*
t_0 LU estimation	15991	8454	38695	276	2203
t_1 LU estimation	15990	8454	38721	273	2203

Table 4: NACE (for LUs) category breakdown for estimates in Table 3.

There are lots of variants on the way in which calibration can be done, first within the GWSM (see Lavallée 2007 section 7.4), and second using the methods of integrative weighting as in the previous section. We leave these differences and options for further research, but note that these variants will give us a range of estimates of the measurement error arising from the unit problem, and that it may not be clear which provides the best approach in given circumstances.

The differences in this example are relatively small, as only a few LUs move between RUs, and only some of these appear in the sample. But again the estimates show an element of the variability (difference) imposed by taking a particular business unit structure. In this case the RU sample is unchanged, but trading of local units causes differences in the estimates.

Discussion

"The unit problem" arises because of differences in the way a model of business structures is applied. For register variables, these differences can be estimated directly using frame information (provided the frame stays up to date and all the structures of interest can be derived from the register). But for survey variables, typically more directly of interest, the weighting approaches provide a way to estimate the differences in outputs caused by the different structures. The two methods here – one based on the unit level, and one based on a longitudinal set of structures – provide examples of how this approach can be used.

The next stage is to apply these methods to a real dataset, to examine the impacts of realistic differences in business structures, and realistic rates of changes in ownership of local units, to obtain some estimates for differences caused by the differences in structural definitions.

References

Lavallée, P. (2007) *Indirect Sampling*. Springer-Verlag: New York.

Lemaître, G. & Dufour, J. (1987) An integrated method for weighting persons and families. *Survey Methodology* **13** 199-207.

Smith, P. (2013) Sampling and estimation for business surveys. Pp165-218 in G. Snijkers, G. Haraldsen, J. Jones & D.K. Willimack *Designing and conducting business surveys*. Wiley: Hoboken, New Jersey.