Collecting the household data as a sub-sample – comparing the weighting model in the Danish household survey to the weighting model of the core-LFS

Abstract

The Danish LFS is characterized by having individual persons and not households as the primary unit. As a result of this, the household data are collected as a sub-sample to the core-LFS on individuals. The advantage is a smaller economic burden on Statistics Denmark since the sample-size is much smaller. Some challenges on the other hand are related with the different weighting models between the core-sample and the sub-sample. This is related to two things. Firstly smaller sample-size means a cruder weighting model. As a result of this much more information is added to the weighting model of the core-sample. Secondly the weighting model of the household survey is optimized to number of households in the population and the few variables solely connected to the household unit (de facto 'jobless household'), while the weighting model for the core-sample is optimized for individuals. This can result in differences in the estimates of central variables.

The difference in the Danish approach to collecting the LFS – main target is individuals not households

Denmark is one of the few countries that differ in the way the LFS is collected from the common practice in Europe, since Denmark collects the LFS data on individuals not households. There are two main reasons for this practice. The first reason is related to the mode of collection. When using CAPI mode there is an obvious gain in collecting the household since it is possible to make interviews with the relevant household members. This gain is not present when using CATI or CAWI mode. In Statistics Denmark CATI mode is the primary collecting supplemented with a little CAWI interviews conducted for the household sample.[1]

The first reason is that Statistics Denmark primarily collects the LFS data on a CATI mode. This mode of collection is very suited for collecting the information on individuals, but is in no way suited for collecting household information. The extensive use of CATI collection differs from the common practice in most countries where the collecting mode is CAPI, which on the contrary is very suited for household collection. There has never been a tradition for CAPI collection in the Danish LFS.

The second reason concerns a more fundamental discussion on what is in fact the unit that we measure in the LFS. In the opinion of Statistics Denmark we primarily measure the labor market situation and behavior of individuals, not households. Central indicators like labor market status, actual working time, a wish to work more or less etc. are all related to an individual, not the household per se. In our view there are no substantial reasons for collecting these variables as household variables when it is possible to collect the information on individuals. However there are a few variables, which are only feasible through a household collection. The variable of jobless households is one of these few variables. Since Statistics Denmark, as all NSI's, has to make prioritizations on how to use the resources the best way possible, a solution for us has been to maintain our core-LFS as an individually based survey and instead collecting the household data as

[1] CAPI has only been applied briefly in 2007-2009 in the Danish LFS.

a sub-sample. The household part of the LFS is collected through a mix of collection modes. The core-respondent is collected through CATI, and the other household members are collected through CAWI. This has the advantage of minimizing the costs of collecting the Household data. However it poses some challenges regarding the weighting of the two different samples and the effect this has on the estimates.

The two different weighting models

There are two key differences between the weighting model applied for the core-LFS and for the household part of the LFS. The first difference is the level of complexity. The core-LFS has a quite complex weighting model, which incorporates a lot of auxiliary information.
This is due to a big non-response in the Danish LFS that has its roots in a large number of persons who have research protection in Denmark. At the same time Denmark has a lot of high quality register information available, which limits this problem. However since the household part is collected as a sub-sample with fewer respondents, the weighting model has to be cruder in order for the model not to break down. The core-LFS has a gross sample around 40.000 persons and 22.000 respondents pr. quarter, whereas the household gross sample is around 11.000 persons (not including core-LFS persons) and 6000 respondents.

Secondly there is a difference in the parameters that the two models are optimized for. The core-LFS is optimized for the population, the total number of individuals. The household part is on the contrary optimized with the total number of household as well as the population in Denmark as the target, which means that the parameters are different. The two weighting models are presented in table 1 and 2.

Table 1 – The weighting model for the core-LFS.

	Variables	Groupings
Information is crossed	-age11	11 grp
	-sex	2 grp
	-region	5 grp
Information is crossed	-age6	6 grp
	-education	3 grp
	-socio-economic status	8 grp
	-number of children in the household	4 grp

	-citizenship	4 grp
	-registered as unemployed	12 grp
	-brutto income	4 grp
	-moved	2 grp

Table 2 – The weighting model for the Household-LFS.

	Variables	Groupings
Information is crossed	-age	3 grp
	-sex	2 grp
	-family type	6 grp
	-size of household	4 grp
	A person from Household has moved	2 grp
	-Only danes in household or mixed household	2 grp
	-average age of the household	3 grp
	-brutto household income	4 grp

As is seen there are significant differences between the two weighting models. The core-LFS model has many more groups on age, and age is also crossed with both sex and region as well as with education in a cruder age-grouping. Other variables that are included in the model are for example socio-economic status, register information from the unemployment register. Besides fundamental auxiliary information such as age, sex and income (here on household) the household model primarily uses auxiliary information concerning the composition of the household.

- Family type divides households between single persons with or without children in the household, married couples with or without children in the household and cohabiting couples with or without children in the household.
- Size of household divides the household in four groups (1, 2, 3 or 4 persons and more).

- A person from household has moved is binary, either one person has moved or not
- Only Danes in household or mixed household is also binary, and divides between households with only Danish citizens and households which are mixed between Danish citizens and foreign citizens/only foreign citizens.
- Average age of the household takes an average age of all the persons living in the household and dividing it into three groups (average under 30 years, average 30-44 years and 45 years and older).

The household related auxiliary information is crucial in order to measure both the total number of different household types as well as the total number of individuals on basic sub-groups such as sex and age. However the fact that the household related auxiliary information is introduced combined with the smaller sample size and thereby fewer respondents makes it problematic to introduce more auxiliary information. Education is an example of important auxiliary information missing in the household model. This has an impact on the estimates on education of the household survey that differs from the core-LFS as a result of the difference in the weighting models.

Difference in educational level between the core-LFS and the household-LFS

It is well known that the non-response in surveys as such, and therefore also in the LFS, is unequally distributed on educational level. More persons with lower educations tend to not answer surveys. Since Denmark has a rather large non-response we will miss a large portion of persons with lower educations leading to an overestimation of the overall educational level. In table 3 this is shown on one of the indicators on education.

Table 3 – Output on educational level.

Highest level of education completed (25-64 years) - %	2011 Core-LFS	2011 HH-LFS	2012 Core-LFS	2012 HH-LFS	2013 Core-LFS	2013 HH-LFS
-At most lower secondary level	23,1	19,4	22,1	18,3	21,7	19,1
-Upper secondary level	43,2	41,6	43,1	41	42,8	40,6
-Third level	33,7	39	34,8	40,6	35,4	40,2

As seen in the table there is a systematic difference between the estimates of the core-LFS and the household-LFS on educational level. The estimate on the percentage of persons with lower educations differ 3,4 pct. point over the three years between the core-LFS who consistently measures a higher number of lower educated compared with the household-LFS. The same goes for upper secondary educational level. Here the average difference over the three years is 2 pct. point. The household-LFS also underestimate this number compared to the core-LFS but not as dramatically. Regarding the third level of education the opposite shows. Now the household-LFS measures a much higher share of persons with longer educations. In average the household-LFS has a share that is 5,3 pct. point higher than in the core-LFS.

The same tendency is shown if we look at the educational level of young persons aged 20-24 having at the minimum an upper secondary educational level as well as the early school leavers, persons

aged 18-24 not in education or training. As table 4 shows the tendency seen above is also present here.

Table 4 – Output on youth educational level and early leavers

	2011 Core-LFS	2011 HH-LFS	2012 Core-LFS	2012 HH-LFS	2013 Core-LFS	2013 HH-LFS
- min. ISCED3c long / upper secondary level (20-24 years) - %	70,0	74,9	72,0	74,9	71,8	76,1
-Early leavers from education and training (18-24 years) - %	9,7	7,9	9,1	8,0	8,1	6,9

If we look at the persons aged 20-24 with at least an upper secondary education the core-LFS in average measures the share 4,0 pct. point lower than the household-LFS over the three years. On the contrary the share of early leavers is consequently lower in the household-LFS compared to the core-LFS in average 1,4 pct. point. This also indicates that the household-LFS underestimates lower educated and overestimate the higher educated.

The examples show that the auxiliary information on education in the weighting model of the core-LFS does its job. It handles the non-response we have on low educated and the overrepresentation of the higher educated. However since it is not possible to introduce more auxiliary information in the household weighting model, it simply will collapse, the household model is not a reliable source when it comes to educational level. But educational level is very rarely used as background information in the relational households' indicator. It is however an important background variable for individual data.

The gain – A reduction of the economic burden for Statistics Denmark

Even though there are some difficulties in collecting the household part of the LFS as a sub-sample, there are also significant gains when it comes to the economical aspect of the. Table 5 shows this.

Table 5 – costs saved by using a sub-sample

Costs saved by sub-sampling		
6000 respondents quadrupled	Euro	DKR (7,45)
Number of respondents	24.000	
Current price in average for ca. 6000 respondents on HH	29.600	220.520
Price quadrupled	118.400	882.080
Difference (saved costs)	**88.800**	**661.560**

Note: The calculation is made on the ground that the data is collected through the present mode which is a CATI-CAWI mix. If all interviews were to be collected through CATI alone, the difference would be even higher.

As seen there is a significant cut in costs by applying a sub-sample on the household data. In the Danish case we save 88.800 euros or over half a million Danish kroner each year on this practice. As noted this is grounded on the present collection mode, which is a mix of CATI and CAWI mode. The use of CAWI-mode to collect the other household members reduces the costs significantly. If CATI-mode was applied the price would rise even more, and the difference as well.

Conclusions

Since the Danish LFS is not suited for household collection regarding the mode of collection, and that we find that most variables are more meaningful to collect on individuals the sub-sample solution has some advantages regarding the costs of collecting the household survey. The sub-sample solution however also has consequences for the quality of some indicators and estimates since the smaller sample and smaller amount of respondents sets certain limits for how fine-tuned the weighting model can be constructed.