How to deal with educational bias in the LFS – A comparison between the use of educational registers in Denmark and Sweden

Jonas Gielfeldt (Statistics Denmark)
Frida Videll (Statistics Sweden)
2013 Workshop on LFS Methodology, Gdansk

Abstract

The 4th Europe 2020 target regarding education has brought an increasing focus from policy-makers on the educational results of the LFS. This poses questions about both estimation and comparability. A well-known problem in surveys is the educational bias, where more persons with higher-level education than persons with lower-level or non-education tend to respond. Another problem arises when registers are being used because there is often a time lag between the register and the LFS. At the same time there are also issues about the comparability of the estimates on educational level since there are very different approaches throughout the member states and potential different degrees and types of bias.

The purpose of this paper is to show how the use of educational registers can help minimizing the bias regarding educational level. There are however different ways of applying the register information in the LFS. Denmark and Sweden has used the educational register for the above mentioned purpose, but in different ways. These different approaches are described in the paper. Furthermore the paper will show the effects of the changes in the use of educational registers on the estimates concerning the 2020 targets.

The use of educational registers in the Danish and Swedish LFS – similar but yet different

Denmark and Sweden are very comparable countries when it comes to the societal characteristics, eg. demography, socio-economic structure and educational level. More specifically Denmark and Sweden are very similar in the way they measure educational level. Both Denmark and Sweden rely heavily on educational registers. The registers are used to impute information about the highest achieved level of education.

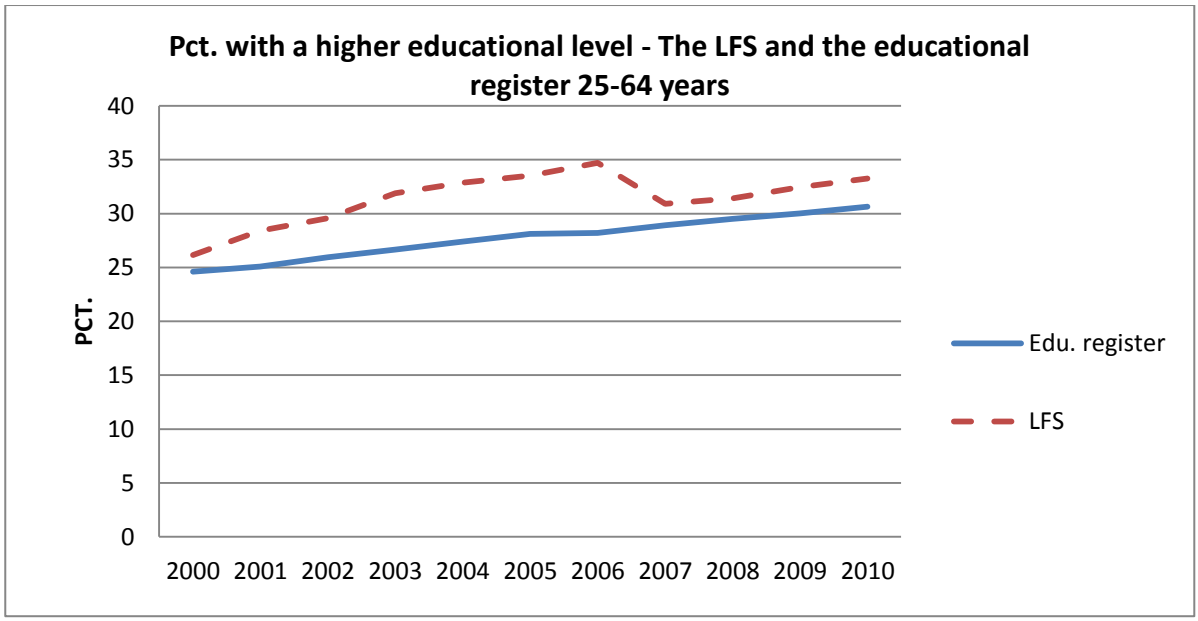
The Danish case

Prior to 2007 the Danish LFS used register information on educational level. However persons who had finished their educations in the last 3 years were asked about their educational level because of a 3-year time lag in the educational register. Auxiliary information on educational level was used only for the group of persons who were registered as unemployed. In 2007 two changes occurred. The time lag was limited to 2 years and only people who had finished their education in the last two years were asked about their educational level. Auxiliary information on educational level was now

introduced for the whole population in the LFS. The implementation of educational registers as auxiliary information for the whole population was a consequence of an increasing difference between the educational level measured in the LFS and the level in the educational register, which indicated an overestimation on the part of the LFS.

Figure 1 shows the development in the Danish LFS. From 2000 and forward there is a growing discrepancy between the educational level measured in the LFS and the level in the educational register. This indicates a bias of higher educated persons in the Danish LFS. In 2007 the educational register starts being used as auxiliary information in the weighting model as well. The level of persons 25-64 years of age in the Danish LFS with a higher education drops from 35 pct. in 2006 to 31 pct. in 2007. This means that the LFS from 2007 and onwards is much closer to the level of the educational register.

Figure 1.



The Swedish case

The primary source for educational level in the Swedish LFS is the educational register. Information on educational level is imputed from the register for each individual that belong to the sample in the LFS. Information about the highest achieved level of education is provided in April each year for individuals aged 16-74. The reference period is the year before with information about educational level from June that year. This gives a time lag of 6 to 18 months between the educational register and the LFS. The difference in age groups, 15-74 in the LFS and 16-74 in the educational register, causes problems with missing values for young individuals.

Individuals that belong to the sample but are missing in the educational register are asked about their highest educational level when they participate in the LFS. This approach helps reduce the number of individuals with unknown educational level but it does not deal with the time lag between the LFS and the educational level. The time lag especially causes problems with incorrect level of education for the young population because of the group's tendency to change their educational level more often than the rest of the population. In order to try to handle this problem the LFS, during the third quarter, asks individuals aged 15-24 questions about their educational level. New individuals in this age group are asked about their educational level in the fourth quarter. This has been done since the third quarter 2012.

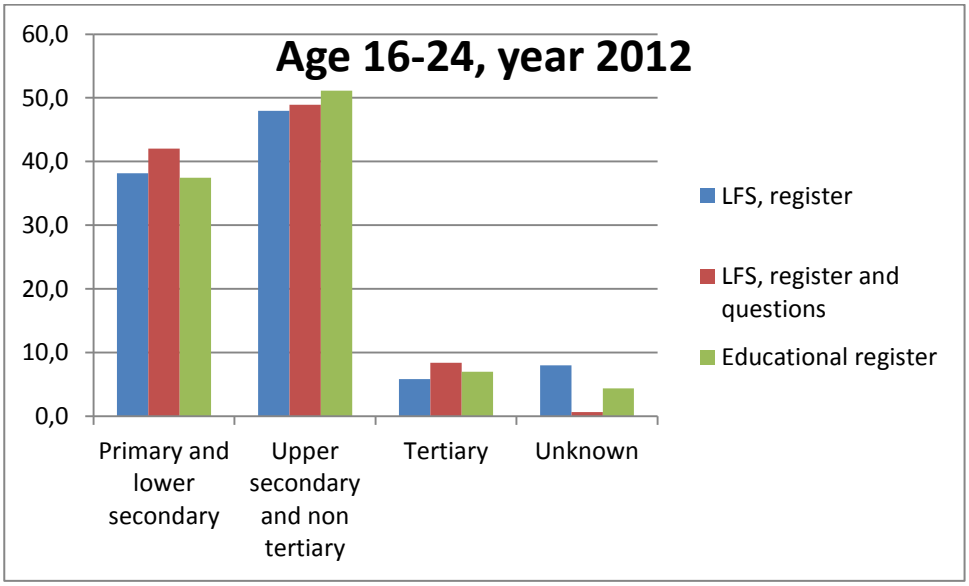
Information on education was introduced in the auxiliary information on an experimental basis to try to improve the LFS-estimates on educational level. This test generated an improvement for the precision in the estimates for education but unfortunately some of the more important estimates on labour force status got a lower precision. Because of the loss in precision for more important estimates it was decided not to include education in the auxiliary information.

Implications on educational estimates - Sweden

Figure 2 shows a comparison between the LFS estimates of educational level and the educational register for the age group 16-24. The reference period for the LFS and the educational register is year 2012. Although the reference period for the two sources is the same the population is defined in different ways. The LFS yearly estimates are constructed as means based on the monthly estimates and the educational register refers to the population on December 31. The first bar shows the estimate in LFS when the only source is the educational register and the second bar shows the LFS-estimate when register and the LFS-questions about educational level are used. The third bar is based on the educational register. The biggest effect when both the register and LFS-questions are being used compared to the other two cases is seen in *Primary and lower secondary* and *Unknown*. Figure 2 shows that by complementing the information from the educational register with information that is received during the LFS interview the proportion of *Unknown* is reduced and the proportion of *Primary and lower secondary* is increased. The proportions for the other two educational groups are also increased but not as much as seen for *Primary and lower secondary*.

The goal with the change has been to improve the information about highest achieved educational level. As a result the proportion of *Primary and lower secondary* is further from the proportion in the educational register compared to the proportion that comes from the LFS estimates without complementary questions. The opposite is true for *Upper Secondary and non tertiary*.

Figure 2.



The implication of the new estimates on the 2020-target - Denmark

The 2007 changes in the weighting model solved the problem concerning the overall overestimation of the educational level, but new challenges occur. A severe educational-bias for the younger age-groups was discovered where there is a tendency towards a higher educational level bias. The problem of skewed bias for different age groups in relation to educational level becomes highly important because the young age groups are the focal point of the Euro-2020 target on education. Partly due to this, the weighting model in the Danish LFS was adjusted in 2011. One of the new elements in this model was that educational level was crossed with age-groups as auxiliary information. Data was consequently revised with the new weighting model back to 2007.

This had quite an impact on the estimates for the younger age groups; the ones that are central to the Europe-2020 target on education. The number of persons in the age group 30-34 years of age, who have tertiary educational attainment, dropped from 43 pct. in 2006 to 38 pct. in 2007. At the same time the numbers of persons who fulfill the definition of being early school leavers corresponding to the definition in the Europe-2020 target increased from 9 pct. in 2006 to 13 pct. in 2007. The improvement of the estimates meant that in 2007 and 2008 the 2020 targets were not fulfilled the 2020 targets in Denmark. This led to controversy with some users and a debate about the tradeoff between quality improvement and comparability started.

Table 1. Changes in the 2020 target estimates of the Danish LFS.

	2001	2002	2003b	2004	2005	2006	2007b	2008	2009	2010	2011
Early school leavers 18-24 years	9	9	10	9	9	9	13	13	11	11	10
Tertiary educational attainment 30-34 years	33	34	38	41	43	43	38	39	41	41	41

b= Break in series

The implication of the new estimates on the 2020-target - Sweden

When both register and questions are being used to compute estimates on early school leavers a decrease of 3 percent is achieved compared to the case where only the register is used. This enhances the results in figure 2 that the time lag between the LFS and the educational register tends to give a distorted description of the educational level for the young population.

The changes concerning estimates on educational level have the biggest impact on the age group 15-24 which gives little or none effect on the 2020 target for tertiary educational attainment 30-34 years as seen in Table 2.

Table 2. Changes in the 2020 target estimates of the Swedish LFS.

	2012	
	LFS, register	**LFS, register and questions**
Early school leavers 18-24 years	10	7
Tertiary educational attainment 30-34 years	48	48

Conclusions

For the Danish LFS it has been a great improvement to the accuracy of the educational level to include educational registers in the weighting model. However the full potential of this is first accomplished when educational groups are crossed with age-groups, thereby handling the educational bias in age-groups better.

The Danish case shows, that before the implementation of educational registers in the weighting model, there is a systematic overestimation of the educational level in the LFS. This raises some questions about the general condition of the estimates of education in the LFS as such. Maybe the LFS is too optimistic in its estimation of the educational level in Europe, and may cause policy-makers to think that the Europe-2020 target on education is reached, when there is actually a large degree of uncertainty to whether this is in fact the

case. The quite drastic changes in the Danish LFS would probably occur in most of the European LFS's, since the educational bias is a very typical bias problem in surveys as such.

The focus for the Swedish LFS has been to reduce the effect of the lag between the educational register and the LFS, this lag has the biggest impact on the younger age groups. The Swedish approach has been to make changes in the questionnaire, no changes have been done in the weighting model. The biggest effects of these changes have been a decrease in the proportion *Unknown education* and an increase in the proportion *Primary and lower secondary education*. There has also been a slight increase in higher educational level and a decrease of early school leavers.

One thing is the challenge with the accuracy for educational level. Another challenge is the very different approaches that member states have when examining the educational level. The Danish and Swedish LFS's are among the most similar when it comes to the use of the educational registers. But even the small differences between the two countries have an impact on the results. Bearing in mind the even larger differences between all the member states, this questions the comparability of educational estimates between the different LFS's. Many LFS does not use registers at all, which possibly leads to overestimation as shown in the Danish case. There is no easy answer to the possible trade-off between estimating more accurate and comparability.