# Analysis on nonresponse bias for the Swedish Labour Force Surveys (LFS)

Fredrik Olsson, Statistics Sweden, Fredrik.olsson@scb.se

**Abstract**

All statistics have some degree of uncertainty that have an effect on the accuracy. The focus in this paper will be on object nonresponse and its effect on the nonresponse bias in the Swedish Labour Force Surveys (LFS).

An analysis has been conducted by approximating LFS variables with variables from different registers. The variables being analyzed are; employed, unemployed, not in the labour force, employees, students, group after income and not in employment, education or training (NEET). Two different estimates are produced for each register variable by using the General Regression Estimator currently in use in the Swedish LFS, one by using the respondents and the other by using the sample. These estimates are used in order to estimate bias, relative bias and confidence intervals for the biases. All estimates are produced for the total population aged 16-74 and for relevant subgroups.

Results for age group 16-74 in December 2015, age group 16-24 in year 2014 for NEET, show that the relative bias is significantly different from zero for almost all of the analyzed register variables, the exception is unemployed. Looking at the relative bias for employed, unemployed and not in the labour force over time, year 2011-2015, one can see that it has been relatively stable over the time period. Therefore, there is no clear indication of an increasing nonresponse bias as the nonresponse increases.

In order to analyze the nonresponse in estimates of change, estimates of change comparing corresponding months from one year to the next have been computed for employed and unemployed for year 2013-2015. For unemployed, the estimates based on the respondents are similar to the estimates based on the sample. The same pattern is in general also seen for employed with exception for subgroups after education.

**Keywords:** Swedish Labour Force Surveys, Nonresponse bias, General Regression Estimator, Register variables

## Introduction

All statistics are affected by uncertainty; in SCB-FS 2016:17 “Statistics Sweden’s regulations regarding quality for official statistics”, the term quality is described and an important dimension of the quality of statistics is their accuracy. Information on accuracy is a prerequisite for users to be able to use the statistics in a correct way. The accuracy in the statistics depends largely on the chosen estimation procedure and how well it takes into account uncertainty that can be traced to the sources of uncertainty related to sample, frame coverage, measurement, nonresponse, data processing and model assumptions. This report addresses nonresponse error and its effect on the quality of statistics in the Swedish Labour Force Surveys (LFS) with the aim of validating the accuracy of the LFS’s estimates.

The chosen approach is a register-based analysis. Two main reasons determined the choice of a register based analysis. One of the reasons is that Statistics Sweden has good access to register information that makes it possible to use relevant register variables. Another reason is that it is not considered possible to conduct a nonresponse follow-up according to the Hansen-Hurwitz method[[1]](#footnote-1) with an acceptable quality at a justifiable cost. This paper is a short version of the report *Analysis on nonresponse bias for the Swedish Labour Force Surveys (LFS) (Statistics Sweden, 2018).*

## Method

In this section, we describe a general method for quantifying bias due to nonresponse. First, let $U\_{T}$ denote a target population and let $U\_{F}$ be the frame population from which a sample $s⊆U\_{F}$ is drawn according to a sampling design $p.$ Let $y$ denote a survey variable of interest and let $y\_{k}$ be the value of this variable for object $k$ in the population. During the data collection, objects respond according to an existing but unknown response distribution $RD$ which generates the realized response set $r⊆s.$ For more information on this view on nonresponse, see Särndal, Swensson and Wretman (1991), chapter 15, sections 5 and 6.

Further, let $\hat{θ}\_{ys}$ be the estimator of the total of $y$ that would have been used if the all objects in the sample set were observed. However, in practice values of $y$ are only observed for the respondents and since the response distribution is unknown a response model $RM$ must be specified in order to be able to estimate the total of $y$. Let $\hat{θ}\_{yr} $be the estimator of the total of $y$ based on the response set using the chosen response model.

The expected deviation of the estimator $\hat{θ}\_{yr}$ from the total of $y$ in the target population

$$B\left(\hat{θ}\_{yr}\right)=E\left(\hat{θ}\_{yr}\right)-\sum\_{k\in U\_{T}}^{}y\_{k}=E\_{p}[E\_{RD}\left(\hat{θ}\_{yr}\right|s)]-E\_{p}(\hat{θ}\_{ys})+E\_{p}(\hat{θ}\_{ys})-\sum\_{k\in U\_{F}}^{}y\_{k}+\sum\_{k\in U\_{F}}^{}y\_{k}-\sum\_{k\in U\_{T}}^{}y\_{k}$$

can be divided into three parts where

$$B\_{FIMP}=\sum\_{k\in U\_{F}}^{}y\_{k}-\sum\_{k\in U\_{T}}^{}y\_{k}$$

is the bias due to frame imperfection,

$$B\_{EST}=E\_{p}(\hat{θ}\_{ys})-\sum\_{k\in U\_{F}}^{}y\_{k}$$

is the bias due to the estimation and

$$B\_{NR}=E\_{p}[E\_{RD}\left(\hat{θ}\_{yr}\right|s)]-E\_{p}(\hat{θ}\_{ys})$$

is the bias due to nonresponse on which will be the focus in this paper.

Let $A$ denote the assumption that the response model agree with the unknown response distribution. Then, under this assumption, it follows that

$$E\_{RM}\left(s\right)≈\hat{θ}\_{ys}$$

which in turn implies that the nonresponse bias $B\_{NR}≈0$.

Let $\hat{V}\left(\hat{θ}\_{ys}\right)$ be an approximately unbiased estimator for the variance of $\hat{θ}\_{ys}$, i.e.

$$\begin{array}{c}E\_{p}\left[\hat{V}\left(\hat{θ}\_{ys}\right)\right]≈ V\_{p}\left(\hat{θ}\_{ys}\right)\#\left(1\right)\end{array}$$

and let $\hat{V}\left(\hat{θ}\_{yr}\right)$ be an approximately unbiased estimator of the variance $\hat{θ}\_{yr}$ under assumption $A$, i.e.

$$\begin{array}{c}E\_{A}\left[\hat{V}\left(\hat{θ}\_{yr}\right)\right]≈ V\_{A}\left(\hat{θ}\_{yr}\right).\#\left(2\right)\end{array}$$

Further, assume that the sample and response sets are large enough so that

$$\begin{array}{c}\hat{θ}\_{ys}\dot{\~}N\left(E\_{p}\left(\hat{θ}\_{ys}\right),V\_{p}\left(\hat{θ}\_{ys}\right)\right) \#\left(3\right)\end{array}$$

and

$$\begin{array}{c}\hat{θ}\_{yr}\dot{\~}N\left(E\_{p}\left(\hat{θ}\_{ys}\right),V\_{A}\left(\hat{θ}\_{yr}\right)\right) .\#\left(4\right)\end{array}$$

Then, under assumption $A$, the covariance between the estimators based on the response and sample sets can be expressed as

$$\begin{array}{c}C\_{A}\left(\hat{θ}\_{yr},\hat{θ}\_{ys}\right)=C\_{p}\left[E\_{RM}\left(s\right),\hat{θ}\_{ys}\right]+E\_{p}\left[C\_{RM}\left(s\right)\right]\\≈C\_{p}\left(\hat{θ}\_{ys},\hat{θ}\_{ys}\right)+0 \\=V\_{p}\left(\hat{θ}\_{ys}\right). \#\left(5\right)\end{array}$$

Hence, from (3)-(5) it follows that under assumption $A$

$$\begin{array}{c}\hat{θ}\_{yr}-\hat{θ}\_{ys}\dot{\~}N\left(0,V\_{A}\left(\hat{θ}\_{yr}\right)-V\_{p}\left(\hat{θ}\_{ys}\right)\right),\#\left(6\right)\end{array}$$

i.e. the difference between the estimators is normally distributed. By combining (6) with the properties in (1) and (2) of the variance estimators, it thus follows that

$\begin{array}{c}T=\frac{\hat{θ}\_{yr}-\hat{θ}\_{ys}}{\sqrt{ \hat{V}\left(\hat{θ}\_{yr}\right)-\hat{V}\left(\hat{θ}\_{ys}\right)}}\dot{\~}N\left(0,1\right).\#(7)\end{array}$under assumption $A$.

If the values of $y$ were known for all objects in the sample set, (7) could thus be used to test the hypothesis $H\_{0}:The response model RM agrees with the unknow response distribution RD$

versus the alternative hypothesis

$$H\_{1}: RM does not agree with RD$$

by rejecting $H\_{0}$ if the observed absolute value $|T\_{OBS}|>z\_{1-α/2}$ with significance level $100α\%$, where $z\_{α}$ denote the $α$-quantile of a standard normal distribution. Moreover, (7) can be used to construct a confidence interval

$$\begin{array}{c}\left(\hat{θ}\_{yr}-\hat{θ}\_{ys}\pm z\_{1-\frac{α}{2}}\sqrt{ \hat{V}\left(\hat{θ}\_{yr}\right)-\hat{V}\left(\hat{θ}\_{ys}\right)}\right)\#\end{array}$$

for the estimated nonresponse bias with confidence level $1-α.$ Unfortunately, if assumption $A$ does not hold, the distributional properties of (7) are not known and hence, it is hard calculate the power for the test based on the test statistic (7).

The value of the survey variable $y$ is of course unknown for every object in the non-response set $s-r$, but if there exists a register variable $z$ which is correlated with the variable $y$ the method above can be used to gain important information. Then, using the method above, one can construct a confidence interval for the bias estimate of the total of $z,$ and test if the response model agrees with the unknown response distribution with respect to the register variable $z$. This information, together with information about the relation between $\hat{θ}\_{yr}$ and $\hat{θ}\_{zr}$, can be used to express a belief if and to what degree the estimator $\hat{θ}\_{yr}$ is associated with nonresponse bias.

## Application to the Swedish LFS

The method described in the previous section have been applied to a various number of register variables. These variables have been chosen in order to approximate different survey variables. The estimated nonresponse bias and the estimated relative bias can provide an indication of how the estimates in the LFS are affected by nonresponse.

An overall analysis has been conducted for a longer period while an in-depth analysis has been conducted for 2015. In the in-depth analysis, nonresponse bias has been studied based on selected study domains on the background variables gender, age, born in Sweden or foreign born and level of education. The analysis has been conducted for nonresponse bias for estimates of level and for estimates of change.

For estimates of level, register variables have been used to classify employed persons according to RAMS[[2]](#footnote-2), unemployed persons according to Af[[3]](#footnote-3), not in the labour force (formed by those who are neither employed according to RAMS or unemployed according to Af), employees according to RAKS[[4]](#footnote-4), students according to RPU[[5]](#footnote-5), three different income groups according to IoT[[6]](#footnote-6) and young people who neither work nor study according to UVAS[[7]](#footnote-7).

To study the possible effect of nonresponse on estimates of change, a variable based on register variables was created to identify employed and unemployed persons on a monthly basis. Based on this variable, estimates of change have been computed by comparing corresponding months in consecutive years. Estimates of change have been calculated based on the response set and the sample set respectively. These analyses have been conducted with the same study domains as those for estimates of level.

## Results

Several of the bias estimates are found to be significantly different from zero and it is thereby not possible to conclude that the statistics are not affected by nonresponse bias. For the total population aged 16-74, the relative bias for employed persons is 1.1 (±0.4) percent, unemployed persons 2.9 (±4.9) percent, those not in the labour force -2.7 (±0.9) percent, employees 1.9 (±0.6) percent, students 10.2 (±2.2) percent, income group 1: -2.1 (±0.8) percent, income group 2: -5.9 (±1.6) percent and income group 3: 5.7 (±0.9) percent. The income group with the lowest relative nonresponse bias, income group 1, is comprised of individuals with an income that is lower than SEK 60,000 for women, lower than SEK 80,000 for men and those who lack information on income.

The size of the nonresponse bias varies over the study domains included in the analysis. The study domains that show the highest level of nonresponse bias is level of education. This is true for all studied register variables. The bias estimates show that estimates for the group with primary and lower secondary education are generally underestimated and that the estimates for the group with post secondary education are overestimated.

For the register variable UVAS, which is used as a proxy for the LFS variable NEET[[8]](#footnote-8), the estimates of level that are based on the response set are consistently lower than the corresponding estimates that are based on the sample set. For this study domain, a relative bias on the scale of 30 percent is observed. This means that compared with the sample set, an underestimation of around 30 percent is obtained for the different background variables.

For unemployed persons, the response and sample set show the same pattern for the estimates of change, and the difference between the response and sample set is small. For employed persons, similar results are obtained except for the background variable education. For employed persons with upper secondary education, the estimate of change based on the response set is systematically lower than the corresponding estimate based on the sample set. For post secondary education, the relationship is the opposite. For this group, the estimate based on the response set is systematically higher than the corresponding estimate based on the sample set.

In summary, for estimates of level, the estimated relative bias, in absolute terms, is generally around 1-3 percent on an aggregated level. The relative bias has been relatively constant in recent years despite the increase in the nonresponse rate. For variables where a larger relative bias is observed, the variables affected are ones that to a higher extent comprise young people. These variables are students and UVAS. Level of education shows a larger relative bias than other study domains, where estimates for those with primary and lower secondary education are generally underestimated and estimates for those with post secondary education are generally overestimated.

For estimates of change, the pattern is largely the same for the response and sample set when unemployed and employed persons are studied. The exception is employed persons by level of education. The estimate of change that is based on the response set is, for the employed persons with upper secondary education, systematically lower than the corresponding estimate based on the sample set. For the employed persons with post secondary education, the estimate of change based on the response set is systematically higher than the corresponding estimate based on the sample set.

## References

Särndal, C-E., Swensson, B., and Wretman, J.H. (1992). Model Assisted Survey Sampling. New York: Springer- Verlag.

Statistics Sweden (2018). Analysis on nonresponse bias for the Swedish Labour Force Surveys (LFS). Background Facts. (https://www.scb.se/publikation/35754)

1. Builds on a sub-sample selected from those classed as nonresponse. From all those belonging to the sub-sample, variable information is collected that is missing upon which analyses regarding nonresponse bias are done. [↑](#footnote-ref-1)
2. RAMS - Register-based labour market statistics [↑](#footnote-ref-2)
3. Af – Swedish Public Employment Service [↑](#footnote-ref-3)
4. RAKS - Activity Statistics based on administrative sources [↑](#footnote-ref-4)
5. RPU - Register on participation in education [↑](#footnote-ref-5)
6. IoT - Register on income and taxation [↑](#footnote-ref-6)
7. UVAS - Register over young people who neither work nor study [↑](#footnote-ref-7)
8. NEET - Not in employment, education or training [↑](#footnote-ref-8)