Statistics Norway

Statistisk sentralbyrå

Ole Villund

Documents

Labour Force Survey non-response in relation to immigrant origin. Some results from the period 2000-2005

Statistics Norway/Department of Economic Statistics/Division for Labour market

Contents

2   Introduction and background

There is a reasonable consensus that a successful integration of migrants contributes to social cohesion and economic welfare. The social and economic conditions for immigrants are closely connected with the integration into the labour market, and statistics on this subject have great interest. Statistics Norway produces detailed register-based labour market statistics for immigrants, but the Norwegian Labour Force Survey (LFS) contains important socio-economic data for individuals, families and households, which could contribute to results that are not achievable from administrative registers alone.

Non-response among immigrants in the LFS combined with the relatively small proportion of immigrants reduces the precision of labour market estimates for immigrant groups. Even more worrying is that the non-response is biased for several important labour market attributes, and more so for immigrants. This reduces the accuracy further, and especially in subdomain analyses for immigrant groups. There are also indications that the quality is poorer when measuring employment transitions in immigrant groups. The results presented here focus on non-response, while measurement and other quality issues will be discussed in forthcoming documents.

This document is one result of a project intended at improving response rates and quality on analyses about immigrants in the LFS. That project was part of the preparations for the inclusion of the EUROSTAT supplementary module "Labour market situation of migrants and their immediate descendants" in the Norwegian LFS in 2008, partly financed by EUROSTAT grant No.: 32100.2005.004-2006-249

3   Data description and some quality issues

This chapter presents an overview of the data and mentions some quality issues in the Norwegian LFS, with emphasis on immigrants.

The subsequent project parts will explore the practical implementations, and some issues are mentioned as ideas for further investigations. First the various data sources are described, with some information on variables and time scope, in relation to the quality analyses presented in this document.

3.1   Data

The data used for analyzing general non-response are LFS quarterly files for the years 1996-2007. Some variables are collected from register data and the files are only 4th quarter of the years 2000-2005. This limits the possibilities with the merged data sets, which means for instance that seasonal variations, trends and predictions are not evaluated here. In some tables results for several years are presented, this is done to evaluate possible instability more than to search for trends.

The register data referred to in this document are files that combines individual data from several administrative registers on employment and other labour market attributes, as well as wages, pensions, social welfare and rehabilitation programmes. The various administrative registers often contain multiple records per person, and each person can be registered in more than one register at a time. When several registers are merged it can lead to seemingly

inconsistent states for some persons. The combinations of information are harmonized in order to create unique statistical units, and a coherent main status for each person. The resulting files are produced routinely as a common data source for a number of statistics, analyses and micro-data files for research purposes.

Some further analyses requires that more data, such as family structure, immigrant related variables, educational and economical information be merged with these LFS files.

3.2   Coverage

The LFS covers the resident population aged 15-74, 16-74 before 2006. The sampling frame is the Central Population Register (CPR) in which persons are identified with a unique ID-number and grouped into families by the ID of a reference person.

Workers on short-term stay are not defined to be in the sampling frame. Illegal aliens who might be regarded as de facto residents, are not covered. In the analyses of LFS response-data, a few respondents not found in register data are discarded from the calculations.

Residents of any type without a registered telephone number are not contacted in other ways. It is not altogether obvious whether this should be classified as frame error or non-response error.

Persons from countries outside the Nordic countries are regarded as residents of Norway when they have lived here or intend to live here at least 6 months, even though the stay is temporary. The same six-month rule applies to migration from Norway to a country outside the Nordic countries.

Norwegian foreign and consular service staff and Norwegian military personnel posted for duty abroad and their families are counted as residents of Norway. Foreign staff at embassies and consular services, foreign personnel attached to NATO, or their families are not counted as residents of Norway. People living in Svalbard, on Jan Mayen, in Norwegian dependencies or on the Norwegian continental shelf who on departure were registered in the population register of a Norwegian municipality shall still be counted as residents of that municipality.

From March 1987 to January 1994 asylum seekers were usually counted as immigrants and hence also as residents even though the processing of their application for residence had not been completed. Before and after this period, only asylum seekers with residence permits have been registered.

3.3   Sampling

The sampling design is a systematic cluster sampling where all aged 15-74 in the sampled families is selected. Such a design could give a cluster effect due to the intrafamilial covariance on labour market variables. The error because of this is essentially adjusted by the estimation procedures, where the statistical unit is person and not family.

A design that uses family units could possibly introduce sampling error with respect to immigrant groups with different family sizes and household structures than the population in general, but the adjustment mentioned above probably compensate this in the same way for

the individual base statistics. Possible consequences for household statistics are not discussed here.

The LFS sample size in any quarter is about 12.000 families, with about 24.000 persons, which gives a sampling rate of about 1/150. The sampling method is stratified systematic sampling that ensures representation from all counties. An effect of the stratification is that Oslo (a separate county) is underrepresented in the sample. Since immigrants are over represented in Oslo, this effect leads to fewer immigrants with this design.

The survey type is a continuous rotating panel, in which each person participates eight consecutive quarters. The LFS aims to measure the person's situation during a specified reference week, and data are collected in such a way that all weeks are covered, by allocating reference periods to persons evenly throughout the quarter.

The panel data ensures more precise estimates of changes, for variables that are stable in individuals, such as employment. For more transient variables the precision improvement is probably less.

3.4 Measurement

The data collection is computer-assisted interviews with some procedures for automatic data control that are included in the questionnaire interface, such as warnings and preset maximum values. The interviewer is guided through the questionnaire based on the answers given, with logical checks. Several data are copied from registers and previous interviews with the same person. The latter is to shorten the interview time, but can possibly impose an artificial stability for some variables, such as changes in occupational tasks within the "same" job.

The main mode of the LFS is interview by telephone, and primarily the respondent is the same person as the observation unit. If it is not possible to get in contact with the observation unit within reasonable time, one of the other family members is interviewed on behalf of the observation unit. This type of proxy mode causes measurement discrepancies when the respondent does not have sufficient information on the labour market relations of the other family members. Proxy interviewing also causes a bias if hard-to-reach single persons are underrepresented because there are no proxies.

Both the mode bias and family bias can have consequences on the response level and measurement with regards to immigrant to the extent that they have different family sizes or differences in the information of other family members.

The questionnaire is in Norwegian and interviewing by telephone in Norwegian has of course great impact on measurement where the respondent has a different first language. Language related issues or other cultural aspects can have an impact, and cannot be adjusted for with the current methods.

3.5 Response

Participation in the survey is compulsory, but there are no penalties for non-response as this would probably not increase the response <u>level</u> much, or possibly worsen the response <u>quality</u>. The overall participation is relatively good with a response rate of over 85%.

An information letter aimed at improving response precedes the first interview. There have been mentioned cultural aspects of such letters that could dampen the response improvement in certain groups. The use of corporate style logo, and not national emblems, may cause the letter to be regarded as junk mail. This could very well be the case among natives as well. The absence of official stamps or signature by a recognised authority figure could also have greater impact among certain immigrant groups than in the general population. It has also been suggested that other measures intended to improve response, such as money or lottery tickets, actually could decrease response in certain cultures.

The fact that unemployed and persons outside the workforce have poorer response in the LFS is especially worrying, but at the same time can be more understandable. Motivational issues and methods for improving response will be discussed further in the next project phases and documented separately.

3.6   Estimation

The published statistics from the LFS are produced with an estimation procedure that utilizes register data on demographics and employment status. The method uses first a post-stratification with employment in three main economic areas and detailed demographic groups, and finally a calibration of county figures. The method gives precedence to the quality of estimates of employment level.

The focus here is <u>reducing</u> non-response, so evaluation of the current estimation procedures aimed at <u>compensating for</u> non-response is not the main issue here. Some experiments with alternative estimation methods are to be presented in upcoming documents.

The combined register file type is considered a population count, although administrative registers also contain a certain level of missing and erroneous values, and bias due to legal and practical matters. These errors are not adjusted with estimation methods in the register-based statistics, but to some degree reduced by manual revision as well as through combining and "harmonizing" data, which could be seen as automatic revision or deterministic imputation.

3.7   Processing

All the calculations performed here are executed in SAS 9.1.3 on UNIX OSF1. The analysis in this documents required repeated processing of more than 20.000.000 records in several gigabyte datasets. The available machine resources were not limiting to the complexities of modelling or calculations. This is mentioned because such limitations have been an issue only a few years ago. It should now be possible to perform even more complex modelling on large data sets.

4   Theory

This document deals only with unit non-response at a point in time, as one consider partial non-response and intermittent non-response to be different in terms of causal model and possible actions to improve response.

Before turning to the Norwegian LFS, some general concerns are mentioned: about the effect of non-response and possible outcomes of actions to improve response. Non-response has at least these negative consequences:

- Underestimation of population totals.
- Reduced precision of any estimate.
- Introducing estimate bias.

Consider any variable of interest and its true mean (or proportion of a binary variable):

$$\bar{Y} = \sum_{i=1}^{N} \frac{Y_i}{N} \qquad \text{Population mean}$$

In a complete and representative sample this would be an unbiased estimator:

$$\bar{y} = \sum_{i=1}^{n} \frac{y_i}{n} \qquad \text{Sample mean}$$

Or as often is the case, with different selection probabilities:

$$\bar{y} = \frac{\sum_{i=1}^{n} w_i \cdot y_i}{\sum_{i=1}^{n} w_i} \qquad \text{Weighted sample mean}$$

In real life, with the ever-present non-response, one must consider:

$$\bar{y}_n = \left(\frac{r}{n}\right)\bar{y}_r + \left(\frac{n-r}{n}\right)\bar{y}_{n-r} \qquad \text{Composition of sample mean}$$

Then the estimated bias, i.e. difference between response mean and sample mean:

$$\bar{y}_r - \bar{y}_n = \left(\frac{n-r}{n}\right)(\bar{y}_r - \bar{y}_{n-r})$$

Notice that this composition is itself an estimate, since both the true population non-response and sample the non-response mean are unknown. But it shows that the bias is dependent on two factors: the non-response level and the non-response bias. This means, for instance, that a heavy bias is not necessarily a problem if the non-response level is very low. And, perhaps more important in this context: a small bias can be a big problem when combined with a high non-response level.

Another caution should be made in connection with improving response level in a realistic situation with multiple non-response causes. For clarity, consider a non-response bias that has only two components, and write the bias formula shorter:

$$\bar{y}_r - \bar{y}_n = \left(\frac{n-r}{n} \right) (\bar{y}_r - \bar{y}_{n-r}) \Leftrightarrow B = qb$$

Then the two components for non-response causes 1 and 2 can be written:

$B = q_1 b_1 + q_2 b_2$ Bias caused by two components

A general scheme to improve response will likely reduce the non-response differently in subgroups. That could very well mean that the reasons for non-response that are most easily mended will be affected the most. If this subgroup also has the least inherent bias, the overall non-response error will not be reduced, even though the response level increases. Formally:

$\Delta B = \Delta q_1 b_1 + \Delta q_2 b_2$ Bias reduction decomposed

$(\Delta q_1 \gg \Delta q_2) \wedge (b_1 \ll b_2) \Rightarrow \Delta B \approx 0$ Greatest reduction of the least biased group

General schemes to improve response can improve precision when there is a high non-response level. But the component model explained above shows that an unfocused response enhancement programme could still leave a considerable bias. Furthermore some actions that improve response in general, can actually worsen bias in certain groups.

Also to be considered is that the introduction of different languages or mixed modes could have adverse effects on response bias and measurement, although increasing the gross response rate.

5 Results

5.1 Non-response level

This chapter introduces some properties of general non-response in some detail, as a reference, before turning to immigrants where the small proportion limits the useful level of detail.

5.2 Non-response in the general population

This preparatory analysis is focused on improving response, and unit non-response is here defined broadly to include all reasons why any selected individual does not participate at all in the LFS. This is done specifically to include those immigrants who are not reachable because of a shorter or longer stay outside Norway, or more or less moving abroad. Some of those and certain other reasons are, with regards to estimation procedures, usually classified as "out of frame" and <u>not</u> non-response.

The classification of non-response reasons does specify the category "moving abroad", but it is possible that a considerable proportion of overseas stays are classified as "not available, unspecified".

The LFS rotation panel design consists of 8 waves and there is a considerable lower response in the first wave. The first interview is the most comprehensive, in comparison the other seven interviews are much simplified for individuals in a stable state.

There is possible also a slight attrition effect, in that the non-response increases somewhat with the waves 2 – 8.

The apparent level shift from 1996/1997 to the rest of the time series is a break due to changes in the classification in the data collection. The Norwegian LFS went through major changes in 1996, and some effects of this were introduced gradually. The same could be part of the explanation of the level shift in 2006. Bearing this in mind, the general trend may be that total non-response is increasing.

Table 5-1: Yearly average non-response rate by wave. LFS 1996-2007. Percent.

	Total	w 1	w 2	w 3	w 4	w 5	w 6	w 7	w 8
1996	9.5	10.3	9.0	8.4	7.8	12.4	10.4	9.0	8.7
1997	9.9	10.0	9.0	9.0	9.6	10.3	11.2	10.8	9.6
1998	12.5	13.2	11.8	12.1	12.0	12.4	12.4	12.8	13.1
1999	12.1	12.1	11.5	12.2	11.3	11.6	12.4	12.9	12.6
2000	11.9	13.8	11.3	11.5	11.4	11.5	11.6	12.0	12.5
2001	13.6	15.3	13.5	13.1	13.6	13.6	13.4	13.0	13.3
2002	11.7	11.7	11.0	11.5	11.9	11.6	11.8	12.4	11.9
2003	12.3	13.7	11.9	11.9	12.2	12.0	12.3	12.3	12.3
2004	11.5	12.2	10.2	10.5	10.7	11.5	12.0	12.3	12.4
2005	12.7	14.0	12.7	12.1	12.7	12.7	12.4	12.6	12.3
2006	14.0	16.3	14.7	13.7	13.3	13.1	13.6	13.5	14.1
2007 (Q1)	14.7	18.3	14.5	14.1	15.0	14.1	13.5	13.9	14.0

The response is somewhat lower in the second quarter, in which ad-hoc modules have been allocated. This leads to longer interviews and lower response. The second quarter is also affected by summer holidays.

Table 5-2: Quarterly non-response rate. LFS 1996-2007. Percent.

	Total	Q 1	Q 2	Q 3	Q 4
1996	9.5	10.6	10.6	9.0	7.9
1997	9.9	8.1	11.2	9.9	10.3
1998	12.5	10.9	15.4	12.0	11.5
1999	12.1	11.2	14.4	11.3	11.4
2000	11.9	10.6	13.6	11.1	12.5
2001	13.6	11.6	15.7	13.9	13.2
2002	11.7	11.0	13.9	11.3	10.8
2003	12.3	11.3	14.5	11.9	11.6
2004	11.5	10.8	11.9	11.5	11.7
2005	12.7	12.6	14.4	11.7	12.1
2006	14.0	14.4	14.9	13.5	13.2
2007	14.7	14.7			

In general the non-response rates are highest for young adult men and among those with lower educational level, and this pattern is relatively stable over many years.

Table 5-3: Yearly average non-response by age, gender and education level. LFS 1996-2007. Percent.

		1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007 (Q1)
Total		9.5	9.9	12.5	12.1	11.9	13.6	11.7	12.3	11.5	12.7	14.0	14.7
Men		10.5	10.8	13.4	13.0	12.8	14.6	12.7	13.3	12.2	13.5	14.8	15.3
	16-24	8.3	10.1	12.1	10.8	11.1	12.8	10.7	12.6	11.0	12.0	12.3	12.4
	25-39	12.1	11.4	14.8	15.4	15.5	17.7	15.2	15.9	15.5	17.4	18.6	18.9
	40-54	10.0	10.2	12.8	11.8	11.8	13.9	12.8	12.8	11.5	13.3	15.1	15.3
	55-74	10.8	11.0	13.1	12.6	11.8	12.3	10.9	11.3	10.1	10.3	12.0	13.5
Women		8.5	9.0	11.5	11.1	11.0	12.6	10.7	11.2	10.7	11.8	13.3	14.0
	16-24	9.0	10.7	12.6	11.3	11.9	14.4	12.4	13.4	12.3	13.2	14.6	16.5
	25-39	8.6	8.6	11.8	11.7	12.2	14.6	12.2	12.7	13.3	15.9	17.2	17.3
	40-54	7.2	7.6	10.0	9.6	9.0	10.6	8.7	9.6	8.3	9.4	10.4	10.8
	55-74	9.4	9.9	12.1	12.0	11.3	11.2	10.0	9.9	9.3	9.2	11.5	12.7
Lower		13.1	14.6	17.3	17.3	17.1	17.9	15.5	17.0	16.2	17.3	19.2	20.1
Medium		8.2	8.5	11.2	10.8	10.8	12.8	11.4	12.0	11.3	12.7	13.9	14.4
Higher		6.9	6.9	9.4	8.9	8.6	11.1	9.0	8.7	8.1	8.9	9.7	10.6

In the LFS data collection process the reasons for non-participation are classified in detailed categories. The following table presents the distribution of reasons in somewhat aggregated categories, in order to avoid very small groups, with the exception of "language problems". The largest groups are "other" and "Not available for unspecified reasons", which are probably particularly heterogeneous. E.g. individuals who refuse, but who won't even accept an incoming telephone call, will not be classified as "refusal".

One observes that the proportion of language problems has been increasing for a period over ten years when the immigrant population has increased. It is quite possible that a substantial proportion of unspecified non-response reasons among immigrants in reality are related to language or cultural issues.

Table 5-4: Yearly average distribution of non-response by reason. LFS 1996-2007. Percent.

	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007 (Q1)
Total non-response	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Refusal	9.5	24.5	20.2	15.2	10.6	11.8	17.7	16.7	16.8	10.4	13.3	14.2
Sickness, own or family	8.8	9.2	10.1	11.2	10.8	8.7	10.6	11.2	11.8	10.1	9.4	9.0
Absence: work, studies, etc.	8.4	5.5	5.4	5.5	8.6	8.9	8.7	8.8	6.9	8.2	7.4	6.4
Not available, unspecified	20.4	18.9	17.5	21.3	23.6	33.4	25.4	28.1	25.4	31.9	35.2	31.3
Address unknown	10.5	7.9	6.3	9.3	4.0	3.2	3.2	2.0	1.6	0.9	0.5	0.6
Telephone unknown	12.3	10.3	9.1	10.3	14.2	12.2	11.8	11.6	15.2	15.7	14.6	13.8
Language difficulties	1.1	1.4	1.2	1.4	1.3	1.6	1.5	1.7	1.9	2.2	2.6	2.6
Moved abroad	7.9	10.6	10.7	10.6	10.2	9.1	11.1	10.1	11.3	9.0	8.7	8.4
Other	21.0	11.9	19.6	15.4	16.7	11.2	10.0	9.9	9.1	11.6	8.4	13.8

5.3   General non-response in relations to immigrant status

Individuals with immigrant background in general has lower response rate in the LFS. So-called western immigrants have nearly twice the non-response rate of non-immigrants, and non-westerns triple this rate.

It is not obvious that language and cultural differences can explain all the differences, as Nordic immigrants also have twice the non-response rate as native Norwegians. The Nordic countries have close ties and much in common with regards to culture and language. Non-

contact due to travelling, especially to country of origin, could very well be a dominant factor for non-response in this group.

Table 5-5: Non-response by immigrant background. LFS 2000-2005. Percent.

	2000	2001	2002	2003	2004	2005
Total	12.4	13.2	10.8	11.6	11.7	12.1
Immigrant	26.9	28.3	22.5	24.4	25.9	27.2
Non-immigrant	11.6	12.1	9.9	10.6	10.6	10.8
Immigrants by country:						
Western	20.0	23.3	17.2	15.8	20.4	18.8
Non-Western	31.3	31.3	25.3	28.8	28.3	30.4
1 Nordic	18.5	23.7	16.1	15.1	19.9	19.1
2 Western Europe	19.8	21.1	18.3	17.0	20.0	15.8
3 Eastern Europe	28.8	24.8	19.7	14.5	22.3	23.5
4 North America	33.3	29.4	20.0	15.2	27.6	31.6
5 Asia	30.8	32.4	25.5	32.2	27.3	32.1
6 Africa	36.2	39.1	33.3	37.4	39.5	37.7
7 South America	33.9	27.7	27.3	36.5	34.3	18.2

When one considers the distribution of non-response reasons, it seems that immigrants are more difficult to contact because there is a larger proportion of unknown address and/or telephone unknown number.

The figures suggests that Western immigrants is more likely to move abroad, which could very well be connected with the reasons for immigration in the first place.

Language difficulties is, not surprisingly, most prominent in the non-Western immigrant groups. For these groups there should be a potential for increasing response given that interviews could be carried out in a language more close to the respondent's mother tongue.

It is noticeable that explicit refusal is markedly higher in groups without immigrant background. One should of course not disregard latent refusal in other categories, but it is possible that straightforward practical solutions can improve response in immigrant groups, as well as addressing motivation from a cultural or psychological perspective. How one should deal with the refusal among natives is another matter.

Table 5-6: Distribution of non-response by reason and immigrant origin. LFS 2000-2005. Percent.

0 Native and other	2000	2001	2002	2003	2004	2005
Total	100.0	100.0	100.0	100.0	100.0	100.0
A REFUSAL	13.3	15.7	21.3	19.3	18.7	14.9
B SICKNESS	10.7	10.2	13.9	13.2	13.2	11.9
C ABSENCE	6.3	6.0	7.0	6.5	5.9	6.3
D UNAVAILABLE	24.5	32.5	20.9	28.4	27.2	31.8
E ADDRESS UNKNOWN	2.0	3.4	2.8	1.4	1.0	1.0
F PHONE UNKNOWN	13.2	11.9	12.1	11.5	15.9	15.7
G LANGUAGE	0.1	0.2	.	0.0	0.1	0.1
H MOVED ABROAD	10.0	10.1	12.2	10.5	10.8	9.5
I OTHER REASONS	20.0	10.2	9.8	9.1	7.2	8.9

1 Western immigrant	2000	2001	2002	2003	2004	2005
Total	100.0	100.0	100.0	100.0	100.0	100.0
A REFUSAL	5.7	10.3	6.1	2.3	5.5	8.0
B SICKNESS	7.6	3.7	3.1	1.1	5.5	4.0
C ABSENCE	8.6	8.8	10.2	5.6	2.7	6.0
D UNAVAILABLE	18.1	30.2	17.4	31.5	24.6	30.0
E ADDRESS UNKNOWN	7.6	5.9	9.2	7.9	1.8	.
F PHONE UNKNOWN	14.3	22.8	24.5	19.1	35.5	29.0
G LANGUAGE	1.0	2.9	.	1.1	2.7	4.0
H MOVED ABROAD	17.1	12.5	24.5	25.8	18.2	18.0
I OTHER REASONS	20.0	2.9	5.1	5.6	3.6	1.0

2 Non-Western immigrant	2000	2001	2002	2003	2004	2005
Total	100.0	100.0	100.0	100.0	100.0	100.0
A REFUSAL	7.5	7.0	10.7	10.0	9.9	3.8
B SICKNESS	3.4	3.7	4.1	3.4	2.6	3.3
C ABSENCE	7.5	4.7	4.1	9.1	7.1	5.4
D UNAVAILABLE	20.3	30.6	33.0	30.0	29.6	34.1
E ADDRESS UNKNOWN	8.3	5.3	6.3	2.5	1.7	1.2
F PHONE UNKNOWN	19.2	18.3	18.5	21.3	27.3	22.8
G LANGUAGE	16.5	17.9	11.9	12.5	11.9	14.8
H MOVED ABROAD	10.5	8.6	8.5	7.8	6.5	9.2
I OTHER REASONS	6.8	4.0	3.0	3.4	3.4	5.4

The category "Absence" is any specified travel due to work, studies, etc., as opposed to "Unavailable" where the cause of non-contact is not given.

6   Non-response bias

Some non-response is inevitable in every survey. Particularly harmful is non-response that is biased with respect to the variable of interest. Auxiliary data that have strong association with the variable of interest are helpful for reducing the non-response bias at the estimation stage. For the present analysis of immigrant groups, our main concern is whether this bias is worse here than the rest of the population, since some immigrant groups both have higher non-response level and lower employment rate.

The following table shows the distribution of register-based labour market status and immigrant status in the total sample and non-response group. Generally employed persons have a higher response probability in the LFS. This also applies to non-western immigrants, but to a lesser degree. In other words non-western immigrants has lower response probability even among those employed. The absolute size of the bias for employment is thus lower, but the immigrant groups also have a lower employment rate.

Table 6-1: Distribution of employment status and immigrant groups, by non-response. LFS 2000-2005.

	2000		2001		2002		2003		2004		2005	
	Total	Non-r.	Total	Non-r.	Total	Non-r.	Total	Non-r.	Total	Non-r.	Total	Non-r.
Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Other, outside workforce	28.5	44.9	27.8	42.7	27.8	47.6	28.5	47.4	28.3	46.0	28.4	44.0
Unemployed	1.4	2.2	1.7	2.7	2.1	3.4	1.9	2.9	2.1	3.5	1.9	3.5
Employed	70.2	52.9	70.5	54.6	70.1	49.0	69.5	49.7	69.7	50.5	69.7	52.5
Native and other	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Other	27.9	44.0	27.2	41.6	27.4	47.6	28.1	47.0	27.6	45.1	27.7	43.1
Unemployed	1.2	1.9	1.6	2.3	1.8	3.3	1.7	2.5	1.8	3.0	1.6	2.9
Employed	70.9	54.1	71.3	56.1	70.8	49.1	70.3	50.5	70.6	51.8	70.7	54.0
Western imm.	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Other	28.9	43.8	29.3	42.6	27.4	42.9	27.8	48.3	28.3	41.8	28.6	39.0
Unemployed	1.9	2.9	2.2	5.1	2.6	3.1	3.0	7.9	3.1	6.4	2.3	3.0
Employed	69.2	53.3	68.4	52.2	70.0	54.1	69.2	43.8	68.5	51.8	69.2	58.0
Non-Western imm.	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Other	43.1	54.5	40.7	52.8	36.2	48.9	38.8	50.6	40.3	53.4	40.4	50.4
Unemployed	5.2	4.1	5.5	4.7	8.4	4.8	6.3	4.1	6.1	5.4	6.4	7.1
Employed	51.7	41.4	53.7	42.5	55.4	46.3	54.9	45.3	53.6	41.2	53.2	42.6

In evaluation of this, one should bear in mind that "non-western immigrants" is a very diverse group both in terms of culture and labour market integration. The LFS data lends itself only to an aggregate analysis. It is therefore recommendable that register data should be utilized in further studies of this topic.

7   Model of non-response

The immigrant groups in Norway have different distributions with respect to several characteristics than the general population, such as demography, education and employment. Many of these characteristics are also associated with response probability in the LFS. It is therefore prudent to ask to what extent the non-response level among immigrants has to do with immigrant status as opposed to other characteristics. A common approach to control for several variables at the same time is to use some kind of regression model.

This chapter presents some results from a limited model using only the variables available in the merged LFS and combined register data. The first attempts with the simple model shown here is to explore the data, and give some ideas for further investigations, not to give a complete model.

7.1   Categorical variables

The following presents a model with at least some factors known to be important for non-response. There are of course other manifest and latent factors, as well as interactions that call for further investigations.

In the presented simple model the dependent variable is non-response in the LFS, which is to be modelled by a number of categorical variables. In the parameterization process, which is to make "dummy variables" for each category, one can have arbitrarily many categories. There are many possible aggregation levels, but the categories should reflect some real world differences and not cause the size of each category to be too small. In the model presented

here is chosen a conservative low number of categories, as well as relatively few independent variables.

Definitions:

$$s = \begin{cases} 1 \mid \text{non - response} \\ 0 \mid \text{response} \end{cases}$$ Binary indicator of non-response

$$p = P(s \mid \vec{x})$$ Conditional probability of non-response

$$\log\left(\frac{p}{1-p}\right) = \alpha + \vec{\beta}\vec{x}$$ Logistic model of non-response

The parameters α and β are to be estimated, and the x is a vector of the categorical independent variables shown in the next table. The design is a general parameterisation to show all the category values, and the model is a binary logit with Fisher's scoring optimization technique.

Table 7-1: Parameterisation of the independent variables.

Class	Value	Design Variables			
Gender	Men	1	0		
	Women	0	1		
Age	16-24	1	0	0	0
	25-39	0	1	0	0
	40-54	0	0	1	0
	55-74	0	0	0	1
Residence	OSLO	1	0		
	OTHER	0	1		
Education	Low	1	0	0	
	Medium	0	1	0	
	High	0	0	1	
Labour status	Outside workforce	1	0	0	
	Unemployed	0	1	0	
	Employed	0	0	1	
Origin	Immigrant	1	0		
	Non-immigrant	0	1		

The factors included have all significant effect on non-response, and the results indicates that immigrant background indeed has a great influence on the LFS response. This is not surprising in view of the results from the proportional analysis. But the model approach is tried in order to control for the different composition of the population groups. As mentioned several more factors should be explored, but will require merging of even more data sources.

Table 7-2: Analysis of Maximum Likelihood Estimates. Model of non-response. LFS 2005.

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-3.5032	0.0737	2257.9184	<.0001
Men	1	0.1870	0.0410	20.7752	<.0001
Women	0	0	.	.	.
16-24	1	0.2895	0.0680	18.1019	<.0001
25-39	1	0.9645	0.0602	256.4376	<.0001
40-54	1	0.5589	0.0625	79.8997	<.0001
55-74	0	0	.	.	.
OSLO	1	0.5420	0.0587	85.1640	<.0001
OTHER	0	0	.	.	.
Low	1	0.6937	0.0677	104.8777	<.0001
Medium	1	0.4411	0.0558	62.4752	<.0001
High	0	0	.	.	.
Outside workf.	1	0.9223	0.0466	391.8759	<.0001
Unemployed	1	0.7834	0.1205	42.2457	<.0001
Employed	0	0	.	.	.
Immigrant	1	0.7657	0.0608	158.5922	<.0001
Non-immigrant	0	0	.	.	.

If the estimates are converted to odds ratios, which have a more intuitive interpretation, one can say that immigrants have twice (most likely between 1.9 and 2.4) the odds of non-response, when controlled for the differences in the other population characteristics.

Table 7-3: Odds Ratio Estimates. Model of non-response. LFS 2005.

Effect	Point Estimate	95% Wald Confidence Limits	
Men vs Women	1.206	1.112	1.307
16-24 vs 55-74	1.336	1.169	1.526
25-39 vs 55-74	2.624	2.331	2.952
40-54 vs 55-74	1.749	1.547	1.977
OSLO vs OUTSIDE WORKFORCE	1.719	1.532	1.929
1 Low vs 3 High	2.001	1.752	2.285
2 Medium vs 3 High	1.554	1.393	1.734
Outside workforce vs Employed	2.515	2.296	2.756
Unemployed vs Employed	2.189	1.728	2.772
Immigrant vs Non-immigrant	2.150	1.909	2.423

Notice that general methods for improving response could improve response among immigrants to some degree because of the factors that have an effect not related to immigrant status. This means that for instance techniques that are aimed at groups with lower education and/or not employed in general also could increase immigrant response.

This simple model does not give a complete picture of the causes and influences on non-response, and more variables and interactions should be investigated. Notice that several goodness-of-fit tests (Deviance-, Pearson- and Hosmer & Lemeshow) indicate lack of fit. That implies that predictions based on the particular model will have a fairly small accuracy.

There are known interactions between the covariates that make it clear that a model with only main effects is oversimplified. Such interactions include a negative effect on young-age factor, and not employed, and also a positive interaction between low/missing education and residence of Oslo. That means for instance that an immigrant living in Oslo is particularly susceptible to not responding in the LFS, more than each factor taken together. There is also a significant negative interaction between educational level and labour market status. Not surprisingly there is a positive correlation between higher education and employment, both of which increases the response probability.

7.2 Binary factors

One could argue that an entirely binary model is even cruder than the simple categorical model. But since the purpose after all is exploring and not predicting, such a simple model is presented here, as it has some useful features.

In a strictly binary model all the independent variables takes only two values 0 or 1. This enables a comparison of the interactions and also the relative importance of the factors, regardless of number of categories. For an easier overview all the factors are positive, so that they <u>increase</u> the non-response probability. Given that the model is appropriate, the estimated coefficients can directly tell which factors that are more important, and how much more than others. Since these are on a logarithmic scale, converted odds ratio are also included for easier interpretation, for example "1.2" meaning "20% higher odds".

Table 7-4: Model of non-response with binary factors. LFS 2000-2005.

	Estimate					
	2000	**2001**	**2002**	**2003**	**2004**	**2005**
Intercept	-3.039	-2.932	-3.258	-3.227	-3.160	-3.151
Male	0.284	0.269	0.311	0.248	0.118	0.183
Age_25_39	0.495	0.506	0.410	0.512	0.531	0.610
Oslo_residence	0.600	0.697	0.694	0.401	0.524	0.548
Lower_education	0.336	0.339	0.350	0.496	0.512	0.490
Not_employed	0.934	0.859	1.058	0.974	0.920	0.852
Immigrant	0.760	0.790	0.678	0.773	0.814	0.853

	Std.error					
	2000	**2001**	**2002**	**2003**	**2004**	**2005**
Intercept	0.070	0.057	0.064	0.063	0.061	0.060
Male	0.040	0.039	0.043	0.042	0.041	0.041
Age_25_39	0.043	0.042	0.046	0.045	0.044	0.044
Oslo_residence	0.058	0.056	0.060	0.062	0.060	0.058
Lower_education	0.064	0.050	0.057	0.056	0.055	0.054
Not_employed	0.042	0.041	0.044	0.043	0.043	0.042
Immigrant	0.068	0.064	0.068	0.065	0.062	0.059

	Odds ratio					
	2000	**2001**	**2002**	**2003**	**2004**	**2005**
Male	1.328	1.309	1.364	1.282	1.126	1.200
Age_25_39	1.640	1.659	1.506	1.668	1.701	1.840
Oslo_residence	1.822	2.008	2.001	1.493	1.689	1.730
Lower_education	1.400	1.403	1.420	1.642	1.669	1.632
Not_employed	2.544	2.362	2.880	2.648	2.508	2.344
Immigrant	2.139	2.204	1.970	2.165	2.258	2.347

The following correlation matrix of binary factors shows all the interactions between the explaining factors, and the results indicate clear negative interactions between immigrant background and local residence, as well as between education and employment. That means that albeit low education and not-employed status both increases the probability of non-response, the combined risks don't add up. The result for the combination of being not employed and aged 25-29 years has a positive value, which means that this segment has even greater probability of non-response than the added effect of each category.

Table 7-5: Correlation matrix of binary factors. LFS 2000-2005.

		2000	2001	2002	2003	2004	2005
Age_25_39	Immigrant	-0.07	-0.05	-0.07	-0.08	-0.09	-0.09
	Lower_education	**0.11**	**0.13**	**0.13**	**0.14**	**0.13**	**0.12**
	Male	0.03	0.02	0.01	0.00	0.01	0.01
	Not_employed	**0.22**	**0.21**	**0.20**	**0.18**	**0.18**	**0.19**
	Oslo_residence	-0.05	-0.04	-0.04	-0.03	-0.03	-0.02
Immigrant	Age_25_39	-0.07	-0.05	-0.07	-0.08	-0.09	-0.09
	Lower_education	-0.00	-0.01	-0.02	-0.01	-0.02	-0.03
	Male	0.01	0.01	-0.00	0.01	0.01	0.02
	Not_employed	-0.07	-0.06	-0.06	-0.06	-0.07	-0.08
	Oslo_residence	**-0.20**	**-0.18**	**-0.22**	**-0.23**	**-0.21**	**-0.23**
Lower_education	Age_25_39	**0.11**	**0.13**	**0.13**	**0.14**	**0.13**	**0.12**
	Immigrant	-0.00	-0.01	-0.02	-0.01	-0.02	-0.03
	Male	-0.02	-0.03	-0.03	-0.03	-0.04	-0.05
	Not_employed	**-0.14**	**-0.15**	**-0.14**	**-0.14**	**-0.15**	**-0.15**
	Oslo_residence	**0.09**	**0.12**	**0.11**	**0.09**	**0.11**	**0.12**
Male	Age_25_39	0.03	0.02	0.01	0.00	0.01	0.01
	Immigrant	0.01	0.01	-0.00	0.01	0.01	0.02
	Lower_education	-0.02	-0.03	-0.03	-0.03	-0.04	-0.05
	Not_employed	**0.12**	**0.10**	**0.09**	**0.08**	**0.08**	**0.08**
	Oslo_residence	0.00	0.01	0.01	-0.00	-0.00	0.00
Not_employed	Age_25_39	**0.22**	**0.21**	**0.20**	**0.18**	**0.18**	**0.19**
	Immigrant	-0.07	-0.06	-0.06	-0.06	-0.07	-0.08
	Lower_education	**-0.14**	**-0.15**	**-0.14**	**-0.14**	**-0.15**	**-0.15**
	Male	**0.12**	**0.10**	0.09	0.08	0.08	0.08
	Oslo_residence	0.02	0.02	0.03	0.01	0.01	0.01
Oslo_residence	Age_25_39	-0.05	-0.04	-0.04	-0.03	-0.03	-0.02
	Immigrant	**-0.20**	**-0.18**	**-0.22**	**-0.23**	**-0.21**	**-0.23**
	Lower_education	**0.09**	**0.12**	**0.11**	**0.09**	**0.11**	**0.12**
	Male	0.00	0.01	0.01	-0.00	-0.00	0.00
	Not_employed	0.02	0.02	0.03	0.01	0.01	0.01

For reference all the binary correlations themselves are included in the table below. Even though there are significant interactions, a model containing interaction coefficients yields some non-significant parameter estimates. This is perhaps more due to low sample size than model invalidation.

Table 7-6: Binary correlations of selected variables. LFS 2000-2005. Pearson R.

		2000	2001	2002	2003	2004	2005
Nonresponse	Age_25_39	0.05	0.06	0.03	0.05	0.06	0.07
	Immigrant	0.11	0.12	0.10	0.11	0.12	0.14
	Lower_education	0.04	0.05	0.05	0.07	0.07	0.07
	Male	0.03	0.03	0.04	0.03	0.01	0.02
	Not_employed	0.14	0.14	0.16	0.16	0.15	0.14
	Oslo_residence	0.09	0.10	0.09	0.06	0.07	0.08
Age_25_39	Nonresponse	0.05	0.06	0.03	0.05	0.06	0.07
	Immigrant	0.06	0.05	0.05	0.07	0.07	0.07
	Lower_education	-0.15	-0.15	-0.15	-0.17	-0.15	-0.15
	Male	0.00	0.00	0.01	0.01	0.00	-0.00
	Not_employed	-0.18	-0.17	-0.16	-0.15	-0.14	-0.15
	Oslo_residence	0.08	0.07	0.07	0.07	0.07	0.06
Immigrant	Nonresponse	0.11	0.12	0.10	0.11	0.12	0.14
	Age_25_39	0.06	0.05	0.05	0.07	0.07	0.07
	Lower_education	-0.01	-0.00	-0.00	-0.01	0.00	0.00
	Male	-0.01	-0.00	0.00	-0.00	-0.01	-0.02
	Not_employed	0.06	0.06	0.06	0.06	0.07	0.08
	Oslo_residence	0.16	0.16	0.17	0.18	0.18	0.19
Lower_education	Nonresponse	0.04	0.05	0.05	0.07	0.07	0.07
	Age_25_39	-0.15	-0.15	-0.15	-0.17	-0.15	-0.15
	Immigrant	-0.01	-0.00	-0.00	-0.01	0.00	0.00
	Male	0.01	0.02	0.03	0.03	0.04	0.05
	Not_employed	0.17	0.18	0.18	0.18	0.19	0.18
	Oslo_residence	-0.09	-0.11	-0.10	-0.10	-0.10	-0.11
Male	Nonresponse	0.03	0.03	0.04	0.03	0.01	0.02
	Age_25_39	0.00	0.00	0.01	0.01	0.00	-0.00
	Immigrant	-0.01	-0.00	0.00	-0.00	-0.01	-0.02
	Lower_education	0.01	0.02	0.03	0.03	0.04	0.05
	Not_employed	-0.09	-0.07	-0.06	-0.05	-0.06	-0.05
	Oslo_residence	-0.00	-0.01	-0.00	0.00	-0.00	-0.01
Not_employed	Nonresponse	0.14	0.14	0.16	0.16	0.15	0.14
	Age_25_39	-0.18	-0.17	-0.16	-0.15	-0.14	-0.15
	Immigrant	0.06	0.06	0.06	0.06	0.07	0.08
	Lower_education	0.17	0.18	0.18	0.18	0.19	0.18
	Male	-0.09	-0.07	-0.06	-0.05	-0.06	-0.05
	Oslo_residence	-0.01	-0.01	-0.01	-0.01	-0.00	-0.00
Oslo_residence	Nonresponse	0.09	0.10	0.09	0.06	0.07	0.08
	Age_25_39	0.08	0.07	0.07	0.07	0.07	0.06
	Immigrant	0.16	0.16	0.17	0.18	0.18	0.19
	Lower_education	-0.09	-0.11	-0.10	-0.10	-0.10	-0.11
	Male	-0.00	-0.01	-0.00	0.00	-0.00	-0.01
	Not_employed	-0.01	-0.01	-0.01	-0.01	-0.00	-0.00

The results for a interaction model concludes this chapter:

Table 7-7: Model of non-response with binary factors and interactions. LFS 2000-2005.

	Estimate					
	2000	**2001**	**2002**	**2003**	**2004**	**2005**
Intercept	-2.97	-2.88	-3.22	-3.18	-3.10	-3.08
Male	0.29	0.27	0.31	0.25	0.12	0.19
Age_25_39	0.32	0.39	0.27	0.40	0.40	0.45
Oslo_residence	0.59	0.69	0.67	0.34	0.46	0.44
Lower_education	0.33	0.34	0.35	0.49	0.50	0.48
Not_employed	0.81	0.76	1.01	0.91	0.84	0.75
Immigrant	0.80	0.80	0.88	0.82	0.82	0.85
Age_25_39 * Not_employed	0.47	0.35	0.36	0.29	0.34	0.43
Not_employed * Immigrant	-0.21	-0.08	-0.52	-0.28	-0.18	-0.25
Oslo * Immigrant	0.13	0.03	0.13	0.26	0.26	0.40

	Std.error					
	2000	**2001**	**2002**	**2003**	**2004**	**2005**
Intercept	0.07	0.06	0.07	0.06	0.06	0.06
Male	0.04	0.04	0.04	0.04	0.04	0.04
Age_25_39	0.05	0.05	0.06	0.06	0.06	0.06
Oslo_residence	0.06	0.06	0.07	0.07	0.07	0.07
Lower_education	0.06	0.05	0.06	0.06	0.06	0.05
Not_employed	0.05	0.05	0.06	0.05	0.05	0.05
Immigrant	0.11	0.10	0.10	0.10	0.10	0.09
Age_25_39 * Not_employed	0.09	0.09	0.09	0.09	0.09	0.09
Not_employed * Immigrant	0.13	0.13	0.13	0.13	0.12	0.12
Oslo * Immigrant	0.15	0.14	0.14	0.14	0.14	0.13

For immigrants there are still marked effect, but also significant interactions. For the LFS quality the effects of "Oslo residence" and "immigrant" and its interaction are especially worrying.

7.3   Modelling in separate subdomains

In order to improve response in immigrant groups one is interested to discern which factors that have specific influence on the non-response in these groups. A method to approach this is to look at the effect of the factors in separate subdomains. The results presented in the tables below are first the parameter estimates for a model in which immigrant status is a factor. Secondly the parameter estimates for the model without immigrant status, but estimated separately in immigrant and non-immigrant. As the subdomain sample size is small, and there are no clear trends, these estimates are six-year averages.

Table 7-8: Reference model of non-response. LFS 2000-2005.

	2000	**2001**	**2002**	**2003**	**2004**	**2005**
Intercept	-3.243	-3.144	-3.506	-3.440	-3.322	-3.418
Men	0.264	0.254	0.315	0.240	0.091	0.177
16-24 years	0.267	0.376	0.339	0.256	0.307	0.309
25-39 years	0.770	0.822	0.725	0.788	0.786	0.943
40-54 years	0.388	0.416	0.452	0.397	0.377	0.524
Lower education	0.493	0.509	0.537	0.667	0.583	0.641
Medium education	0.304	0.264	0.304	0.424	0.444	0.432
Outside workforce	0.761	0.713	0.930	0.868	0.815	0.757
Unemployed	0.652	0.598	0.743	0.588	0.718	0.783
Immigrant	0.950	0.962	0.878	0.893	0.957	0.980

Table 7-9: Subdomain estimates in model of non-response. LFS Average of 2000-2005.

6-year average	Native	Immigrant
Intercept	-3.357	-2.183
Men	0.207	0.329
16-24 years	0.315	0.138
25-39 years	0.843	0.399
40-54 years	0.445	0.156
Lower education	0.508	0.884
Medium education	0.377	0.229
Outside workforce	0.827	0.691
Unemployed	0.828	0.208

Education has a greater effect for immigrants than natives. This model indicates that gender and education has a somewhat greater impact, and that the labour market connection has less effect in total. But it is a noticeable difference in the labour market status effects in immigrant groups, while it is higher and more similar for natives.

It will be up to the data collecting experts to utilize such background information in planning and the interview process. It has been suggested that person-to-person, house calls etc. be used in some groups or areas, to reach persons that are not reachable on telephone.

7.4 Effect of length of stay

How long a person have been staying in Norway is of course interesting in relation to the overall integration and the labour market relations. It is also of importance for the language skills and cultural adaptation. Length of stay is possibly related to the degree of responsiveness to an official survey, although this relationship may not be linear (or even positive!). It could very well be that the non-response level in the LFS is most dependent upon if one can actually be able to contact someone, such as known telephone number and some basic language skills. A model including Length-of-stay factor shows significant effect, but this factor could very well be deduced from the two attributes mentioned.

Table 7-10: Simple model of non-response, including Length-of-stay. LFS Q4 2005

Parameter	Estimate	Error	F-test	Prob.
Intercept	1.2902	0.1133	129.6312	<.0001
Male	-0.2066	0.0926	4.9724	0.0258
Age 20-35 years	0.2073	0.0959	4.6731	0.0306
OSLO	-0.8995	0.0970	85.9454	<.0001
Not-employed	-0.6343	0.0938	45.7063	<.0001
Length-of-stay	0.0259	0.00366	49.9065	<.0001

Table 7-11: Non-response by Length-of-stay groups. LFS 2000 – 2005

	2001	2002	2003	2004	2005
Total	11.8	9.8	10.3	10.5	10.6
Non-immigrants	10.7	8.8	9.2	9.2	9.1
0- 5 years	35.4	28.4	31.0	31.4	29.3
6-10 years	22.5	20.6	19.7	27.8	26.5
11-20 years	31.3	26.0	27.3	26.9	29.1
20+ years	19.1	19.2	17.0	20.5	21.8

The aggregated analysis shows that even after a very long stay, LFS non-response is markedly higher. These results tell us that for instance language skills are possible not a major factor for non-response. We assume that most immigrants attain a workable Norwegian language after a number of years of residence.

8 Conclusion

We have noted that non-response among immigrants in the LFS combined with the relatively small proportion of immigrants reduces the precision of labour market estimates for immigrant groups.

The results presented here indicates that non-response is biased for important labour market attributes, and more so for immigrants. This reduction of accuracy must be considered in subdomain analyses for immigrant groups. The analysis of non-response presented here contributes to the understanding of non-response in the LFS and could be useful for improving response and data quality.

Other, preliminary, analyses indicate that the quality is poorer when measuring employment transitions in immigrant groups. We hope to complete an analysis of measurement and other quality issues in the same LFS data that is examined here, based on the preliminary analyses, some of which results are included in a work report to EUROSTAT.