Discussion Papers No. 273, June 2000
Statistics Norway, Division for Statistical Methods and Standards

Ib Thomsen and Li-Chun Zhang

The Effect of Using Administrative Registers in Economic Short Term Statistics: The Norwegian Labour Force Survey as a Case Study

Abstract:
In case of a single survey at one time point, it is well known that combining administrative registers with survey data often substantially improves the quality of statistic production. However, in short term statistics it is as important to measure changes over time as it is to measure the overall level. Using data from the Norwegian Labour Force Surveys and administrative registers, we demonstrate in this paper that the use of registers have little or no effect on the accuracy of estimates of change, both in terms of the sampling variance and the bias introduced by nonresponse. The survey data alone measures the changes over time. One consequence of interest for the rotation design is that a very high proportion of the sample can be retained from one survey to the next without seriously reducing the accuracy of the level estimates. We believe that these findings are relevant to short term statistics in general.

Keywords: Post-stratification, estimation of level and change, design of survey

Address: Li-Chun Zang, Statistics Norway, Division for Statistical Methods and Standards.
E-mail: li.chun zang @ssb.no

Discussion Papers comprise research papers intended for international journals or books. As a preprint a Discussion Paper can be longer and more elaborate than a standard journal article by including intermediate calculation and background material etc.

Abstracts with downloadable PDF files of
Discussion Papers are available on the Internet: http://www.ssb.no

For printed Discussion Papers contact:

Statistics Norway
Sales- and subscription service
N-2225 Kongsvinger

Telephone: +47 62 88 55 00
Telefax: +47 62 88 55 95
E-mail: Salg-abonnement@ssb.no

1 Introduction

Both administrative registers and survey data are common sources of statistic production. It is also well known that the use of administrative registers through techniques like ratio-estimation, post-stratification, raking and/or calibration may lead to substantial reduction in the sampling variance as well as the bias introduced by nonresponse (Bethlehem, 1988; Djerf, 1997; Thomsen and Holmøy, 1998; Zhang, 1999). Most studies in this respect concentrate on a single survey at one time point. However, in short term statistics it is as important to measure changes over time as it is to measure the overall level. In this paper we shall examine in some detail the effects of the combined use of rotating samples and administrative data.

Using data from the Norwegian Labour Force Surveys (LFS) and administrative registers, we demonstrate that the use of registers have little or no effect on the accuracy of estimates of change, both in terms of the sampling variance and the bias introduced by nonresponse. The role of the register is to produce high quality measures of the overall level, while the survey data alone measures the changes over time. One consequence of interest for the rotation design is that a very high proportion of the sample can be retained from one survey to the next without seriously reducing the accuracy of the level estimates. We believe that these findings based on the LFS are relevant to short term statistics in general.

In several countries including Norway, a Register-Employment Status is available for the entire population. These administrative registers are prepared independently of the LFS, and can be linked to the LFS at the individual level. In this case study we focus on the LFS-Employment Status as the survey variable, and use the Register-Employment Status as the auxiliary variable. Both are illstrated in Figure 1 below, where the solid lines connect the quarterly population Register-Employment Rates, and the dashed ones the quarterly sample LFS-Employment Rates.

Figure 1: Register-Employment and LFS-Employment in Norway from 1995 to 1997

There are many reasons why the LFS is necessary in spite of the existence of the Employment Registers, several of which can be seen in Figure 1. First of all there is a clear discrepancy in the overall levels according the two sources. This is largely due to the defintion of the Register-Employment, which is hardly comparable to the LFS-Employment Statistics from other countries. At the end of each calendar year, the Register undergoes a major control which produces unpredicatable outcomes. Throughout the year, the Employment Register is updated based on selv-reports from the employers. Delay in the routine is probably a reason why the Register-Employment Rate is higher in the 4th than the 3rd quarter, which counters the traditional wisdom of economy. At the present stage, we are not able to determine the general pattern of the variations, including such delays, in this selv-governed reporting process.

2 Effects of post-stratification on the variance of the estimators

In studying the combined use of rotating samples and the Register, we shall first concentrate on the *netto LFS-panel* between two successive quarters, i.e. the part of the LFS-sample which has responded in both quarters. Denote by s_0 the netto LFS-panel of size n_0. For anyone in s_0, let y_t (for $t = 1, 2$) be the LFS-Employment status in two successive quarters, where $y_t = 1$ for employment and $y_t = 2$ otherwise. Classified according to (y_1, y_2), the netto LFS-panel forms a 2×2 contingency table, with cell counts n_{ij} for $i, j = 1, 2$, which corresponds to the number of people with LFS-Employment status $(y_1, y_2) = (i, j)$, i.e. $\sum_{i,j=1}^{2} n_{ij} = n_0$. Let p_{ij} be the corresponding cell probability, i.e. $\sum_{i,j=1}^{2} p_{ij} = 1$. Denote by $\hat{p}_1 = (n_{11} + n_{12})/n_0$ the simple sample mean estimator of the LFS-Employment Rate at $t = 1$, and $\hat{p}_2 = (n_{11} + n_{21})/n_0$ that at $t = 2$. The change in LFS-Employment Rate from $t = 1$ to $t = 2$ is estimated by $\hat{p}_2 - \hat{p}_1$, and the average LFS-Employment Rate for $t = 1$ and $t = 2$ by $(\hat{p}_1 + \hat{p}_2)/2$. In particular, $Var(\hat{p}_t) = p_t(1 - p_t)/n_0$ for $t = 1, 2$, and $Cov(\hat{p}_1, \hat{p}_2) = (p_{11} - p_1 p_2)/n_0$. This gives us

$$Var_{ssm}(\hat{\bar{p}}) = \{\bar{p}(1 - \bar{p}) - \alpha/4\}/n_0 \qquad \text{where} \quad \bar{p} = (p_1 + p_2)/2 \quad \text{and} \quad \alpha = p_{21} + p_{12}. \tag{1}$$

where we have used subscript *ssm* to specify the case of simple sample mean; and

$$Var_{ssm}(\hat{p}_2 - \hat{p}_1) = (\alpha - \delta^2)/n_0 \qquad \text{where} \quad \alpha = p_{21} + p_{12} \quad \text{and} \quad \delta = p_{21} - p_{12}, \tag{2}$$

Let x_t (for $t = 1, 2$) be the Register-Employment status in two successive quarters, defined similarly to y_t. According to the values of (x_1, x_2), the netto LFS-panel can be divided into non-overlapping subsamples, denoted by $s_{0,h}$ for $h = 1, ..., H$, i.e. the post-strata. Within each post-stratum, (x_1, x_2) is a constant, and can be used to identify the post-stratum. In particular, *dynamic post-stratification* according to the Register from both quarters gives us post-strata $(x_1, x_2) = (1, 1)$, $(1, 2)$, $(2, 1)$ and $(2, 2)$. Whereas *simple post-stratification* uses the Register from only one of the two quarters, giving us post-strata $(x_1, x_2) = (1, -)$ and $(2, -)$, or $(x_1, x_2) = (-, 1)$ and $(-, 2)$. The marginal proportion of each post-stratum is known for the population, and is denoted by q_h for $h = 1, ..., H$. Let $(\theta_h, \hat{\theta}_h)$ be any parameter and its estimator within post-stratum h. The post-stratified estimator of $\theta = \sum_h q_h \theta_h$ is given by $\hat{\theta} = \sum_h q_h \hat{\theta}_h$. Conditional on the actual sample sizes of the post-strata, denoted by $(n_{0,1}, ..., n_{0,H})$ and $n_{0,h} > 0$, its variance is

$$Var_{pst}(\hat{\theta}|n_{0,1}, ..., n_{0,H}) = \sum_h q_h^2 Var_{ssm}(\hat{\theta}_h|n_{0,h}), \tag{3}$$

where we have used subscript *pst* for the case of post-stratification, and $Var_{ssm}(\hat{\theta}_h|n_{0,h})$ is the corresponding within-stratum variance such as those in (2) and (1). The unconditional variance is obtained by averaging (3) over the distribution of $(n_{0,1}, ..., n_{0,H})$ (Holt and Smith, 1979). Expanding $1/n_{0,h}$ around $E[n_{0,h}]$ gives us $1/E[n_{0,h}]$ as the leading term of $E[1/n_{0,h}]$. Due to the relatively large $E[n_{0,h}]$, the unconditional variance is almost identical with the conditional one in the present case. It is thus instructive to observe that, given $n_{0,h} \doteq n_0 q_h$, we have that

$$Var_{ssm}\{(\hat{p}_1+\hat{p}_2)/2|n_0\} - Var_{pst}\{(\hat{p}_1+\hat{p}_2)/2|n_0\} \doteq (\sum_h q_h \bar{p}_h^2 - \bar{p}^2)/n_0,$$

where $\bar{p}_h$ is obtained from (1) within post-stratum h, and $\bar{p} \doteq \sum_h q_h \bar{p}_h$. Therefore, roughly speaking, the more $\bar{p}_h$ differs from one post-stratum to another, the greater reduction in the variance of the level estimator can be achieved through post-stratification. Meanwhile,

$$Var_{ssm}(\hat{p}_2-\hat{p}_1|n_0) - Var_{pst}(\hat{p}_2-\hat{p}_1|n_0) \doteq (\sum_h q_h \delta_h^2 - \delta^2)/n_0,$$

where δ_h is obtained from (2) within post-stratum h, and $\delta \doteq \sum_h q_h \delta_h$. That is, the reduction in variance of the estimator of change through post-stratification is largely determined by its ability to differentiate δ_h from one post-stratum to another. In particular, notice that, given the size of the netto panel, $\bar{p}$ is a function of $p_{11} - p_{22}$, i.e. the difference between the two diagonal cells; whereas δ is the difference between the two off-diagonal cells. The same interpretation applies to $\bar{p}_h$ and δ_h in each post-stratum.

The next table shows the netto LFS-panel between the third and fourth quarter in 1997:

Year 1997	(4th Quarter)	Register-Employment			
(3rd Quarter)		Yes		No	
Register-Employment	LFS-Employment	Yes	No	Yes	No
Yes	Yes	10913	203	200	89
	No	155	353	15	73
No	Yes	258	27	1209	311
	No	115	42	279	4122

Using the observed $n_{0,h}/n_0$ as q_h, we obtain the following estimates (all values $\times 10^{-6}$):

Post-stratification	$\hat{Var}(\hat{p}_1)$	$\hat{Var}(\hat{p}_2)$	$\hat{Cov}(\hat{p}_1, \hat{p}_2)$	$\hat{Var}(\hat{p}_2 - \hat{p}_1)$	$\hat{Var}(\hat{\bar{p}})$
(-, -)	10.99	11.08	9.27	3.54	10.15
(1,-), (2,-)	5.51	5.91	3.94	3.54	4.83
(-,1), (-,2)	5.69	5.44	3.80	3.53	4.68
(1,1), (1,2), (2,1), (2,2)	5.29	5.32	3.58	3.44	4.44

Post-stratification according to the Register results into an approximate 50% reduction in the variance of the level estimators. Similar effects have been reported in the literature (Djerf, 1997; Zhang, 1999). However, it appears that post-stratification has practically no effect on the variance of the estimator of change. In particular, dynamic post-stratification leads to no noteworthy improvement over simple post-stratification, neither for the level- nor the change-estimators. Notice that $\delta_h \approx -0.004$ in post-stratum (1,1) and -0.005 in post-stratum (2,2), which together contain about 95% of the sample. Another intuitive way of understanding the result is to observe that the correlation coefficient between Register-Change, i.e. $X_2 - X_1$, and LFS-Change, i.e. $Y_2 - Y_1$, was estimated to be 0.164 based on the netto LFS-panel. In contrast, it is about 0.7 between X_t and Y_t, i.e. Register- and LFS-Employment at the same t.

3 Effects of post-stratification on the bias caused by nonresponse

We refer to the part of the LFS-sample which overlaps in two successive quarters as the *brutto LFS-panel*, denoted by s of size n. Given nonresponse, $s_0 \subset s$ and $n_0 < n$. The difference between s_0 and s being those who did not respond in either one or both of these two quarters. Let θ be the population mean of LFS-Employment which is unknown, and $\hat{\theta}(s_0)$ the corresponding sample mean estimator based on the netto LFS-panel, and $\hat{\theta}(s)$ that derived from the brutto LFS-panel which can not be observed. This gives us the identity $\hat{\theta}(s_0) - \theta = \{\hat{\theta}(s_0) - \hat{\theta}(s)\} + \{\hat{\theta}(s) - \theta\}$. The difference between $\hat{\theta}(s)$ and θ arises from sampling, whereas that between $\hat{\theta}(s_0)$ and $\hat{\theta}(s)$ is due to nonresponse. The effect of post-stratification on $\hat{\theta}(s) - \theta$ is well known. To study the effect of post-stratification on reducing the bias caused by nonresponse, therefore, we shall concentrate on $\hat{\theta}(s_0) - \hat{\theta}(s)$.

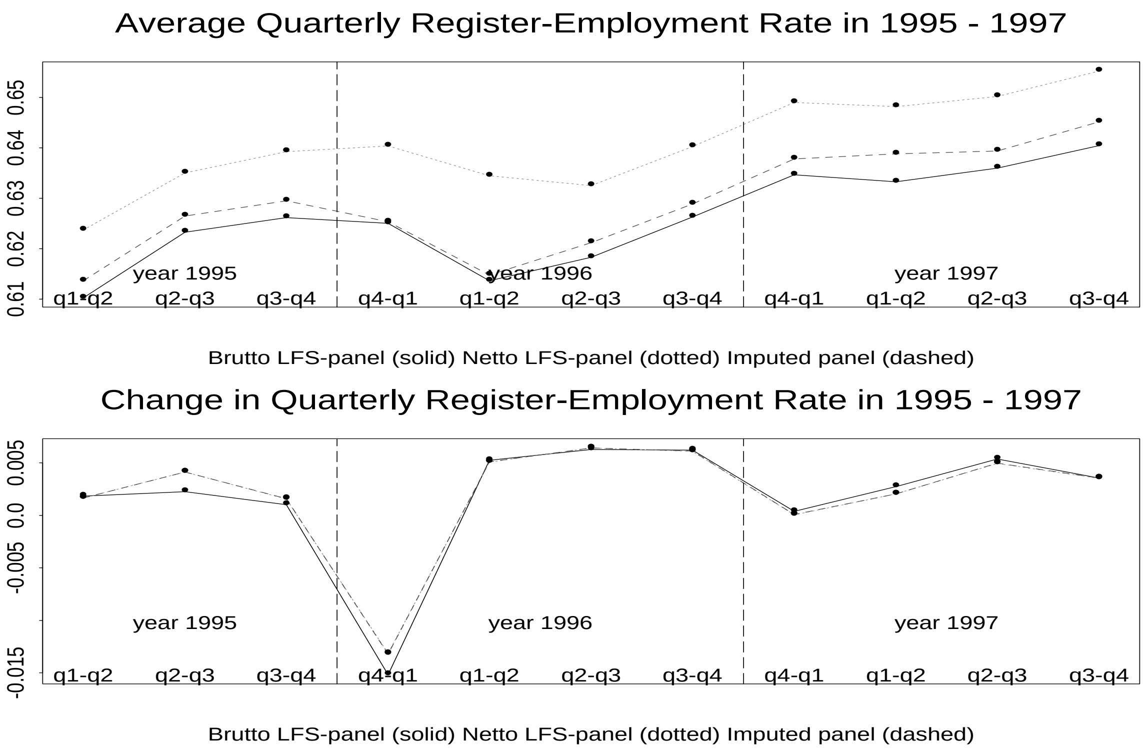

Figure 2: Register-Employment Rate in the Norwegian LFS from 1995 to 1997

Since the Register-Employment status is available for the brutto LFS-panel as well, it seems natural first to examine the difference between the netto and brutto LFS-panel regarding variable Register-Employment. Based on each LFS-panel, we calculated the (sample) Average Quarterly Register-Employment Rate, i.e. the mean Register-Employment Rate of the two quarters involved, and (sample) Change in Quarterly Register-Employment Rate. The difference between the corresponding $\hat{\theta}(s_0)$ and $\hat{\theta}(s)$ then provides an estimate of the bias caused by nonresponse conditional to s. These are given in Figure 2, i.e. solid $\hat{\theta}(s)$ and dotted $\hat{\theta}(s_0)$. Nonresponse her is clearly nonignorable (Rubin, 1976) in the sense that its distribution depends on the object variable Register-Employment. As a consequence the Register-Employment Rate differs from the respondents to the nonrespondents — it is lower among the nonrespondents. The bias of the netto estimator of Change, on the other hand, was much smaller. Let $X_2 - X_1$ be Register-Change. The approximate agreement between the netto Register-Change and the brutto one implies that the latter can be re-constructed out of the former, by proportionally allocating the nonrespondents according to observed frequncy of Register-Change in the netto panel. In other words, nonresponse is approximately independent of Register-Change. Thus, nonresponse seems to depend on Register-Employment, i.e. (X_1, X_2), almost entirely through the mean Register-Employment, i.e. $(X_2 + X_1)/2$, since (i) $(X_2 - X_1, X_2 + X_1)$ is a one-to-one transformation of (X_1, X_2), and (ii) $Cov(X_2 - X_1, X_2 + X_1) = Var(X_2) - Var(X_1) \doteq 0$.

Fay (1986) and Little and Rubin (1987) discussed general approaches to estimation in the presence of nonignorable nonresponse. We have applied the following chained logistic regression model, which was motivated by the particular dependence structure (of nonresponse on Register-Employment) observed above. Examples of similar chained logistic regression models based on the factorizations of the joint probability of (X_1, X_2, R_1, R_2) can be found in Bjørnstad and Sommervoll (1993). Let $R_t = 1$ denote response at t and $R_t = 0$ nonresponse. Let $\text{logit}(\eta)$ denote the logistic transformation of η, i.e. $\text{logit}(\eta) = \log(\eta) - \log(1 - \eta)$, and

$$\begin{aligned}
\text{logit } P[X_1 = 1] &= \beta_1, \\
\text{logit } P[X_2 = 1|x_1] &= \beta_2 + \beta_3 x_1, \\
\text{logit } P[R_1 = 1|(x_1, x_2)] &= \beta_4 + \beta_5(x_1 + x_2), \\
\text{logit } P[R_2 = 1|(x_1, x_2, r_1)] &= \beta_6 + \beta_7(x_1 + x_2) + \beta_8 r_1.
\end{aligned}$$

We assume, through the factorization of $P[R_1, R_2|(x_1, x_2)]$ into $P[R_1|x_1 + x_2]P[R_2|(x_1 + x_2, r_1)]$, that (R_1, R_2) is independent of (X_1, X_2) given $(x_1 + x_2)$. Having fitted the model to the netto LFS-panel, using the EM algorithm, we constructed the *imputed (brutto) panel*, denoted by s^*, conditional to the observed netto panel, by evaluating the expectations at the estimated parameter values. Based on s^*, we obtain $\hat{\theta}(s^*)$ as if s^* had been observed. This gives us the third (dashed) series of estimates in Figure 2. We notice that the estimated Changes based on the imputed panels coincide with those on the netto ones, now that the model assumes nonresponse to be independent of $X_2 - X_1$. Meanwhile, the model has resulted into much reduction in the bias of the level estimator. The discrepency between the imputed panels and brutto ones nevertheless shows that there were things which remained unexplained by the model. This could be the case if the nonrespondents form subgroups with different nonresponse patterns. For instance, people might refuse to participate out of reasons which have nothing to do with their employment status.

We now turn to LFS-Employment which is only observed in the netto LFS-panel. Based on each netto panel, we calculated the sample mean estimator. To apply the dynamic post-stratification, we simply used n_h/n as the marginal proportion of the post-strata. These have been given in Figure 3, i.e. solid for dynamic post-stratification and dotted for netto sample mean, which display a similar pattern as that between $\hat{\theta}(s)$ and $\hat{\theta}(s_0)$ in the case of Register-Employment. In particular, the close agreement between LFS-Change $(Y_2 - Y_1)$ based on the dynamic post-stratification and the netto panel implies that, the latter can be re-constructed from the former, by proportionally allocating the nonrespondents within each post-stratum according to the observed frequncey of $Y_2 - Y_1$ within the same post-stratum. In other words,

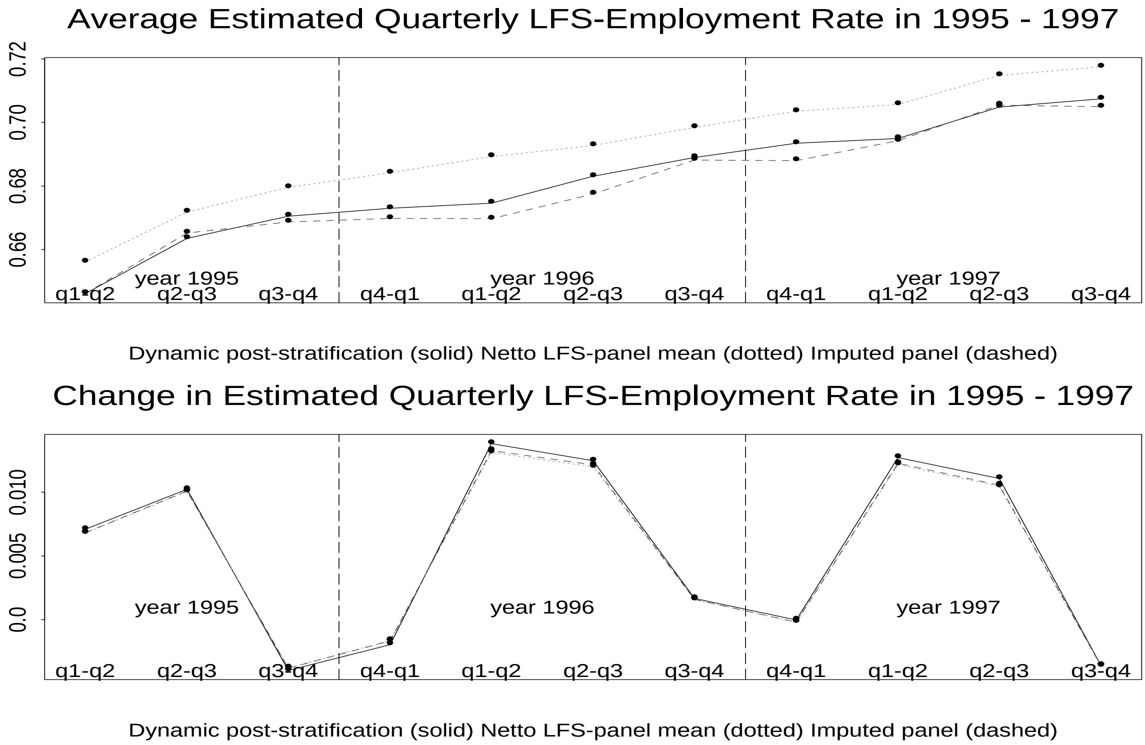

Figure 3: LFS-Employment Rate in the Norwegian LFS from 1995 to 1997

nonresponse is independent of LFS-Change conditional to Register-Employment. To see whether this independence also holds marginally, we applied the nonignorable nonresponse model above to the data, after having replaced (X_1, X_2) with (Y_1, Y_2). That is, we assume that (R_1, R_2) does not depend on $Y_2 - Y_1$, irrespective of (X_1, X_2). This gives us the third (dashed) series of estimates in Figure 3. We notice that the estimated LFS-Change based on the imputed panels largely coincide with those on the netto panel directly, which seems to suggest that nonresponse is independent of LFS-Change also marginally. On the other hand, the dynamic post-stratification had similar effects on the level estimator as the nonignorable nonresponse model, despite that post-stratification rests on the assumption that nonresponse is ignorable within each post-stratum. Due to reasons suggested earlier, we do not expect the nonresponse model to be able to fully adjust the bias in the level estimator. Neither, therefore, is the post-stratified estimator unbiased.

4 Further work

This study has been part of a more comprehensive evaluation of the total survey design of the Norwegian LFS. Three questions concerning the sampling strategy are of particular importance in this connection: (i) Is the sample size adequate? (ii) How should the sample be selected? (iii) How should the excisting administrative registers be used in order to support the sample? These questions are interrelated, but we shall discuss them separately below.

Concerning the size of the sample it is worth noticing the results shown in Figure 4. Here it is seen that the estimate of the Employment Rate is lower by using post-stratification. This

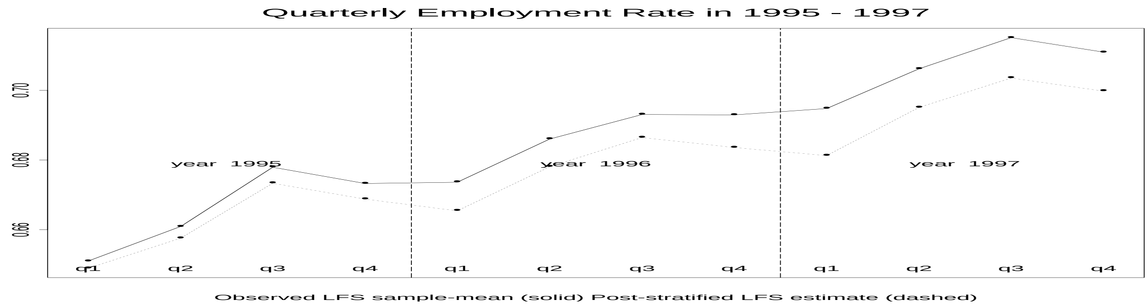

Figure 4: Register-Employment Rate in the Norwegian LFS from 1995 to 1997

decrease is approximately three times the standard error of the estimate. This relatively dramatic difference immidiately raises the question whether the sample size is too large. However. the Labour Force Surveys are multipurpose, and therefore an evaluation of the adequate sample size should include a discussion about which economic indicators are the most important ones produced from the surveys. Furthermore, it should be stated what accuracy, included accuracy of changes, one is aiming at. As seen from the study the accuracy of changes are not affected by the use of post-stratification.

At present a one-stage, equal probability sample of families is used each quarter. The sample is selected from the Central Population Register which include information concerning sex, age and addresses of each person. A question of interest is whether this information can be used to form homogeneous strata. It is well known that young and old persons persons change status on the labour market more often than the rest of the population. It therefore seems of interest to study the feasibility of stratifying the families before selection and overrepresent families with young and old individuals.

Finally, concerning the use of other registers for post-stratification, there are a number of possibilities open. In our opinion it is of particular interest to include the register of unemployed persons, which must be merged with the register presently used for post-stratification. After eventual inconsistencies between the two registers have been identified and decided upon, the new register would form a better basis for post-stratification.

References

Bethlehem, J.G. (1988). Reduction of nonresponse bias through regression estimation. *J. Off. Statist.*, **4**, 251–60.

Bjørnstad, J.F. and Sommervoll, D.E. (1993). Nonresponse models for panel surveys. Technical report, Statistics Norway (Notater 93/18).

Djerf, K. (1997). Effects of post-stratification on the estimates of the Finish Labour Force Surveys. *J. Off. Statist.*, **13**, 29–39.

Fay, R.E. (1986). Causal models for patterns of nonresponse. *J. Amer. Statist. Assoc.*, **81**, 354–65.

Holt, D. and Smith, T.M.F. (1979). Post stratification. *J. Roy. Statist. Soc. A*, **142**, 33–46.

Little, R.J.A. and Rubin, D.B. (1987). *Statistical Analysis with Missing Data.* New York: Wiley.

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, **63**(3), 581–92.

Thomsen, I. and Holmøy, A.M.K. (1998). Combining data from surveys and administrative record systems. The Norwegian experience. *Int. Statist. Rev.*, **66**, 201–21.

Zhang, L.-C. (1999). A note on post-stratification when analyzing binary survey data subject to nonresponse. *J. Off. Statist.*, **15**, 329–34.