Journal of Official Statistics, Vol. 17, No. 2, 2001, pp. 285–294

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

I. Thomsen[1] *and L.-C. Zhang*[1]

In the case of a single survey at one point in time, it is well known that combining administrative registers with survey data often substantially improves the quality of estimation. 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 (LFS) and administrative registers, we demonstrate in this article that the use of registers has little or no additional effect on the accuracy of estimates of change based on the panel part of the survey data, neither in terms of the sampling variance nor in the bias introduced by nonresponse. The main reason is that the administrative register available is not sufficiently up-to-date at the time of production. Indirectly, however, the use of registers can improve the estimator of change through the rotation design of the surveys, since it allows us to deploy a higher overlap proportion in the sample without seriously reducing the accuracy of the level estimates. We believe that these findings are relevant to short term statistics in general, especially when the registers suffer from delays.

Key words: Poststratification; estimation of level and change; survey design.

1. Introduction

Both administrative registers and survey data are common sources of official statistics. It is well known that the use of administrative registers through techniques like ratio-estimation, poststratification, raking and calibration may lead to substantial reduction in the sampling variance of survey estimates 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 point in time. However, in short term statistics it is as important to measure changes over time as it is to measure the overall level. In this article we shall examine in some detail the effects of the combined use of rotating samples and administrative data.

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 through the personal ID-number 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 illustrated in Figure 1, where the solid lines connect the quarterly population Register-Employment

[1] Statistics Norway, P.b. 8131 Dep., N-0033 Oslo, Norway. E-mail: li.chun.zhang@ssb.no

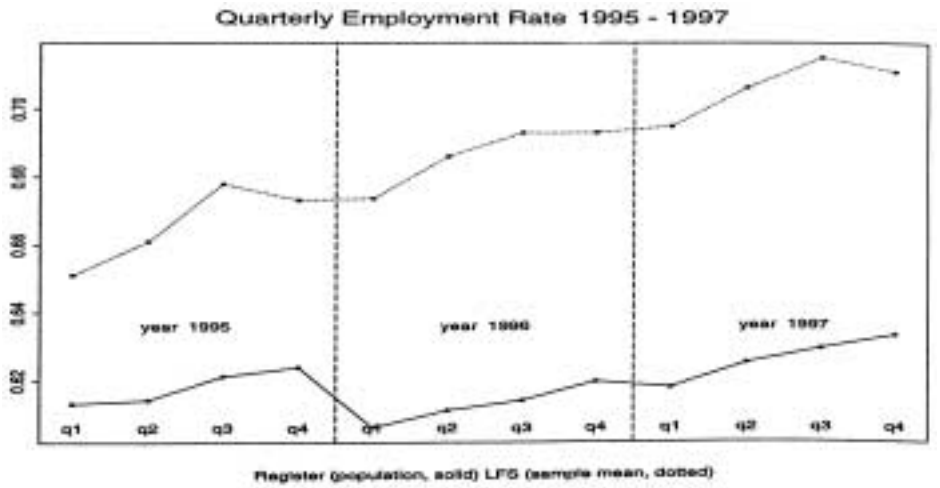

Fig. 1. Register-Employment and LFS-Employment in Norway from 1995 to 1997

Rates, and the dotted ones the quarterly sample LFS-Employment Rates. 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 to the two sources. This is largely due to the definition of the Register-Employment, which is different from the ILO-definition commonly used in the LFS Statistics. At the end of each calendar year, the Register undergoes a major control which produces unpredictable outcomes. Throughout the year, the Employment Register is updated based on reports from employers. Delay in the process 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 present, we are not able to determine the general pattern of the variations, including such delays, in this self-governed reporting process.

Using data from the Norwegian Labour Force Surveys (LFS) and administrative registers, we demonstrate that poststratification of the sample according to information from the registers substantially reduces the effect of the sampling variance of the totals at each point in time. The bias due to nonresponse is substantially reduced. Concerning the measurement of change over time, one must distinguish between the panel part of the data and the rest. We find that there is little or no additional effect from using information from administrative registers, when it comes to the accuracy of estimates of change based on the panel data. The main reason is that the change measured by the registers available at the time of production is of poor quality due to delays. Indirectly, however, the use of registers has an effect on the measurement of change through the design of the surveys. As the accuracy of the estimates at each point in time is increased by the use of the registers, it allows the statistician to deploy a larger overlap proportion in the sample, thereby reducing the sampling variance of the estimator of change over time.

In many countries no personal ID-number is available. In such cases the use of administrative information may have less effect than reported in the present study. Steel (1997) presented some results from the UK. The survey information was linked to the administrative data by asking each person in the sample about his or her status in the register.

Poststratification was then applied using this register status as poststratification variable. The method was found to result in ''considerable overestimate of the number of ILO-unemployed people,'' and almost no effect on the variance of the estimator. The main reason stated was the respondents' confusion about ''the different social security benefits'' that determine the register status.

2. Effects of Poststratification on the Variance of the Estimators

At present the Norwegian LFS uses a stratified sampling design. The strata are made up of the 19 counties in Norway. Within each stratum a fixed number of families are selected with equal probability. (The LFS-population consists of persons between 16 and 74 years old, and the average family size in the Norwegian LFS is below 2.) The sampling fraction varies somewhat from one stratum to another, giving smaller counties higher representation. For simplicity, we shall assume simple random sampling below when calculating the variances of both the standard and the poststratified estimators. The absolute values of the variance estimates are therefore not entirely accurate due to the varying within-stratum sampling fraction, as well as the cluster effect of family. However, we believe that this has very little effect on the conclusions we draw when we compare the methods to each other.

In studying the combined use of rotating samples and the Register, we shall first concentrate on the *net* 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 net 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 net 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, with $\sum_{i,j=1}^{2} p_{ij} = 1$. 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 $\hat{\bar{p}} = (\hat{p}_1 + \hat{p}_2)/2$. Under binomial assumptions, $Var(\hat{p}_t) = p_t(1 - p_t)/n_0$ for $t = 1, 2$, and $Cov(\hat{p}_1, \hat{p}_2) = (p_{11} - p_1 p_2)/n_0$. We have

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

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

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

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 *net* LFS-panel can be divided into nonoverlapping subsamples, denoted by $s_{0,h}$ for $h = 1, \ldots, H$, i.e., the poststrata. Within each poststratum, (x_1, x_2) is a constant, and can be used to identify the poststratum. In particular, dynamic poststratification according to the Register from both quarters gives us poststrata $(x_1, x_2) = (1, 1)$, (1,2), (2,1) and (2,2) whereas *simple poststratification* uses the Register from only one of the two quarters, giving us poststrata

$(x_1, x_2) = (1, -)$ and $(2, -)$, or $(x_1, x_2) = (-, 1)$ and $(-, 2)$. The marginal proportion of each poststratum is known for the population, and is denoted by q_h for $h = 1, \ldots, H$. Let $(\theta_h, \hat{\theta}_h)$ be any parameter and its estimator within poststratum h. The poststratified 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 poststrata, denoted by $(n_{0,1}, \ldots, n_{0,H})$ and $n_{0,h} > 0$, its variance is

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

where we have used subscript *pst* for the case of poststratification, 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 $(n_{0,1}, \ldots, n_{0,H})$ (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

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

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

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

where δ_h is obtained from (2) within poststratum h, and $\delta \doteq \sum_h q_h \delta_h$. That is, the reduction in variance of the estimator of change through poststratification is largely determined by its ability to differentiate δ_h from one poststratum to another. In particular, notice that, given the size of the net 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 poststratum.

Table 1 shows the net LFS-panel between the third and fourth quarter in 1997. The combined effects on the sampling variances of using panel data and poststratification are estimated in Table 2, where we simply set q_h at the observed $n_{0,h}/n_0$. It is seen that poststratification according to the Register results in an approximately 50 percent reduc-

Table 1. The respondents in both the third and fourth quarters in 1997

Year 1997		Register-employment			
		Yes		No	
(3rd Quarter) Register-employment	(4th Quarter) LFS-employment	Yes	No	Yes	No
Yes	Yes	10,913	203	200	89
	No	155	353	15	73
No	Yes	258	27	1,209	311
	No	115	42	279	4,122

Table 2. Combined effects on the sampling variances of survey design and poststratification.

(All values $\times 10^{-6}$)	Independent samples			Panel data			
Method of poststratification	(−,−)	(1,−), (2,−) (−,1), (−,2)	(1,1), (1,2) (2,1), (2,2)	(−,−)	(1,−), (2,−)	(−,1), (−,2)	(1,1), (1,2), (2,1), (2,2)
$\widehat{Var}(\hat{p}_1)$	10.99	5.51	5.29	10.99	5.51	5.69	5.29
$\widehat{Var}(\hat{p}_2)$	11.08	5.44	5.32	11.08	5.91	5.44	5.32
$\widehat{Cov}(\hat{p}_1, \hat{p}_2)$	0	0	0	9.27	3.94	3.80	3.58
$\widehat{Var}(\hat{p}_2 - \hat{p}_1)$	22.07	10.95	10.61	3.54	3.54	3.53	3.44
$\widehat{Var}(\hat{\bar{p}})$	5.52	2.74	2.65	10.15	4.83	4.68	4.44

tion in the variance of the level estimators. Similar effects have been reported in the literature (Djerf 1997; Zhang 1999). For the independent part of the sample it is seen that poststratification has a substantial effect on all the sampling variances. However, it appears that poststratification has practically no effect in addition to the use of panel on the variance of the estimator of change. In particular, dynamic poststratification leads only to relatively small improvement over simple poststratification, both for the level- and the change-estimators. Notice that $\delta_h \approx -0.004$ in poststratum (1,1) and -0.005 in poststratum (2,2), which together contain about 95 percent 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 net 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. We believe that this lack of correlation is largely due to delays in the register available for poststratification at the time of production.

3. Effects of Poststratification on the Bias Caused by Nonresponse

We refer to the part of the LFS-sample which overlaps in two successive quarters as the *gross 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 are persons 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 based on the net LFS-panel, and $\hat{\theta}(s)$ that derived from the gross LFS-panel which is not observed. We have 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 poststratification on $\hat{\theta}(s) - \theta$ is well known. To study the effect of poststratification on reducing the bias caused by nonresponse, we shall concentrate on $\hat{\theta}(s_0) - \hat{\theta}(s)$.

Since the Register-Employment status is available for the gross LFS-panel as well, it seems natural first to examine the difference between the net and gross LFS-panel regarding the 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 on s. The two estimates are given in Figure 2, i.e., solid $\hat{\theta}(s)$ and dotted $\hat{\theta}(s_0)$. Nonresponse here is clearly nonignorable (Rubin

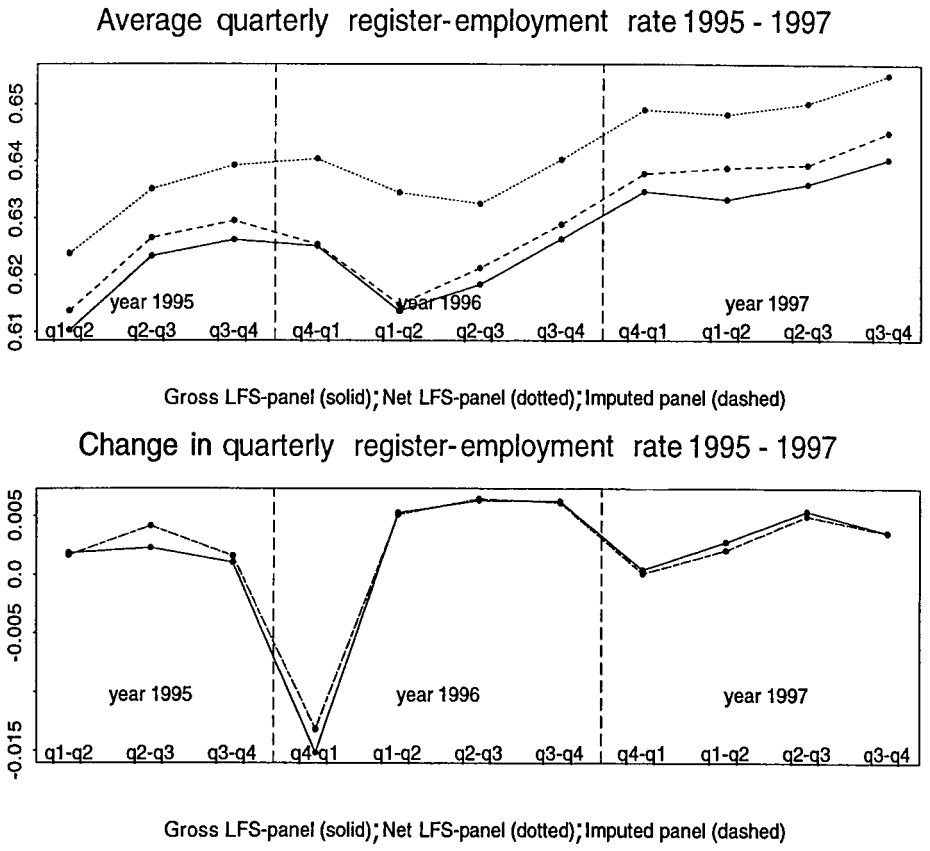

Fig. 2. Register-employment rate in the Norwegian LFS from 1995 to 1997

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 net estimator of Change, on the other hand, was much smaller. Let $X_2 - X_1$ be Register-Change. The approximate agreement between the net Register-Change and the gross one implies that the latter can be reconstructed out of the former, by proportionally allocating the nonrespondents according to observed frequency of Register-Change in the net 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), where $R_t = 1$ denotes response at t and $R_t = 0$ nonresponse, can be found in Bjørnstad and Sommervoll (1993). Let logit(η) denote the logistic transformation of

η, i.e., $\text{logit}(\eta) = \log(\eta) - \log(1-\eta)$, and

$$\text{logit } P[X_1 = 1] = \beta_1$$

$$\text{logit } P[X_2 = 1|x_1] = \beta_2 + \beta_3 x_1$$

$$\text{logit } P[R_1 = 1|(x_1, x_2)] = \beta_4 + \beta_5(x_1 + x_2)$$

$$\text{logit } P[R_2 = 1|(x_1, x_2, r_1)] = \beta_6 + \beta_7(x_1 + x_2) + \beta_8 r_1$$

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 net LFS-panel, using the EM algorithm, we constructed the imputed (gross) panel, denoted by s^*, conditional on the observed net 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 Change estimates based on the imputed panels coincide with those based on the net 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 discrepancy between the imputed panels and gross 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 for reasons which have nothing to do with their employment status.

We now turn to LFS-Employment which is only observed in the net LFS-panel. Based on each net panel, we calculated the sample mean estimator. To apply the dynamic poststratification, we simply used n_h/n as the marginal proportion of the poststrata. These have been given in Figure 3, i.e., solid for dynamic poststratification and dotted for net 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 poststratification and the net panel implies that the latter can be reconstructed from the former, by proportionally allocating the nonrespondents within each poststratum according to the observed frequency of $Y_2 - Y_1$ within the same poststratum. In other words, nonresponse is independent of LFS-Change conditional on 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 LFS-Change estimates based on the imputed panels largely coincide with those based on the net panel directly, which seems to suggest that nonresponse is independent of LFS-Change also marginally. On the other hand, the dynamic poststratification had about the same effects on the level estimator as the nonignorable nonresponse model, despite the fact that poststratification rests on the assumption that nonresponse is ignorable within each poststratum. For 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 poststratified estimator unbiased.

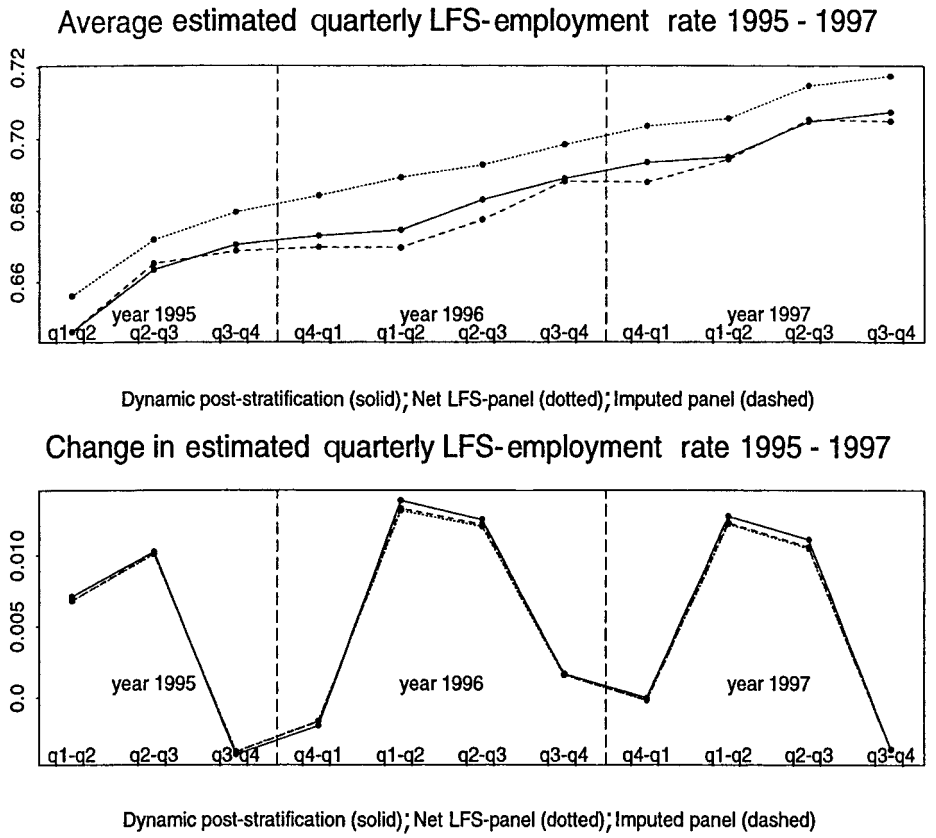

Fig. 3. LFS-employment rate in the Norwegian LFS from 1995 to 1997

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 existing administrative registers be used in order to support the sample? These questions are interrelated, but we shall discuss them separately here.

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 using poststratification. This decrease is approximately three times the standard error of the estimate. This relatively dramatic difference immediately raises the question whether the sample size is too large. However, the Labour Force Surveys are multipurpose. 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, including accuracy of changes, one is aiming at. As can be seen from the study, the accuracy of changes is not affected by the use of poststratification based on the panel part of the survey date.

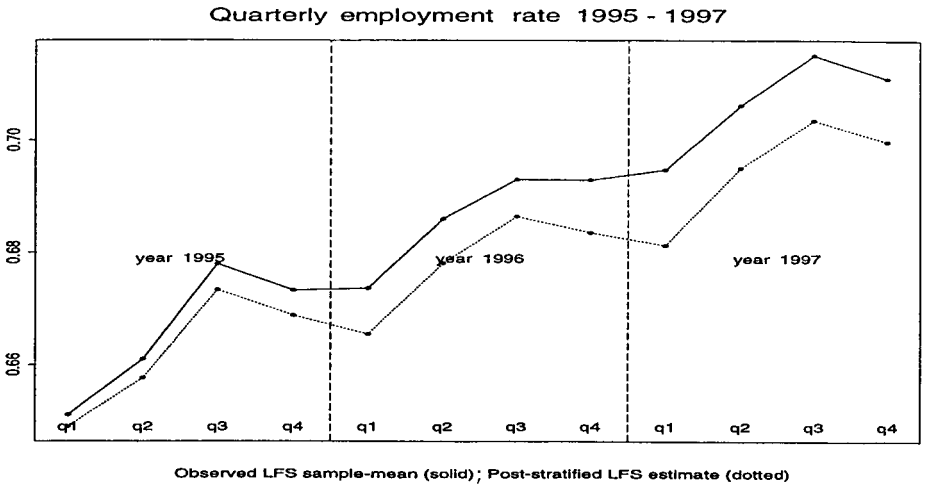

Fig. 4. LFS-employment rate in the Norwegian LFS from 1995 to 1997

At present the sample of families is selected from the Central Address Register (CAR), which is essentially a register of families. The main reason for this is the costs associated with interview. However, it is possible to select individuals from the Central Population Register (CPR) and link them to the CAR to obtain more accurate addresses. Moreover, the CPR also contains information about sex and age of each individual, and therefore the ''structure'' of the family. A question of interest is whether this information can be used to form homogeneous strata. It is well known that young and old people change status on the labour market more often than the rest of the population. It is therefore natural to study the feasibility of stratifying the families before selection and overrepresenting families with young and old individuals.

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

5. References

Bethlehem, J.G. (1988). Reduction of Nonresponse Bias Through Regression Estimation. Journal of Official Statistics, 4, 251–260.

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 Finnish Labour Force Surveys. Journal of Official Statistics, 13, 29–39.

Fay, R.E. (1986). Causal Models for Patterns of Nonresponse. Journal of the American Statistical Association, 81, 354–365.

Holt, D. and Smith, T.M.F. (1979). Post Stratification. Journal of the Royal Statistical Society, 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), 581–592.

Steel, D. (1997). Producing Monthly Estimates of Unemployment and Employment According to the International Labour Office Definition (with discussion). Journal of the Royal Statistical Society, A, 160, 5–46.

Thomsen, I. and Holmøy, A.M.K. (1998). Combining Data from Surveys and Administrative Record Systems. The Norwegian Experience. International Statistical Revue, 66, 201–221.

Zhang, L.-C. (1999). A Note on Post-stratification When Analyzing Binary Survey Data Subject to Nonresponse. Journal of Official Statistics, 15, 329–334.

Received January 2000
Revised September 2000