THE USE OF ADMINISTRATIVE DATA SOURCES FOR LITHUANIAN ANNUAL DATA OF EARNINGS

Milda Šličkutė-Šeštokienė

Statistics Lithuania, Lithuania

e-mail: milda.slickute@stat.gov.lt

1 Abstract

Statistics Lithuania has the full range of labour statistics that meet the timeliness and demands of Eurostat and national needs. The challenge is to keep this quality and timeliness and to publish even more detailed information and at the same time spare costs.

Users need more and more statistical information and at the same time respondents want to get less and less questionnaires. That enforce Statistics Lithuania to seek for new methods for estimation of statistical information required.

This presentation describes the introduction of administrative sources at estimation stage for data of earnings. Generalized Regression estimator of total and ratio is examined. Introduction of administrative sources at estimation stage significantly improved the quality of the statistical estimates and spared the burden and the costs.

2 History of Annual Survey of Earnings

Until 2003 Annual Survey of Earnings (ASE) used to be performed completely enumerating all enterprizes. According to the one of the goals of Statistics Lithuania, to diminish burden for enterprizes as much as possible using administrative sources, Labour statistics division decided to reject ASE and to calculate annual data of earnings for 2004 on the basis of Quarterly Survey of Earnings (QSE) and data of Social Insurance (SI).

It is supposed that usage of administrative sources will diminish the burden for enterprizes as well as for staff of Statistics Lithuania keeping quite good quality of statistical data.

The year 2003 were chosen for simulation and consideration of methods that could be used for estimation, because it is the only year when all three sources (ASE, QSE and SI) are available. The ASE where rejected since 2004 and data of SI become available

since 2003. All methods were analyzed for the year 2003 and it was compared with the real figures of Annual Survey of Earnings 2003.

3 Simulation and results

3.1 Sources available

As mentioned before annual data on earnings 2004 was estimated on the basis of two sources:

- Quarterly Survey of Earnings;
- Data of Social Insurance (administrative source).

Quarterly Survey of Earnings is conducted applying sampling methods. A simple random stratified sample is used. The Horvitz-Thompson estimator is applied to estimate the parameters of interest in each domain. The definitions of main variables of Quarterly Survey on Earnings and Annual Survey of Earnings is the same. The main reasons why two surveys duplicating variables used to be performed are following:

- Quarterly data are required every quarter for national needs;
- Detailed breakdown of annual data requires complete enumeration.

Data of Social Insurance that available for Statistics Lithuania are for the year 2003 and later. Definitions of statistical variables and variables of Social Insurance does not coincide but statistical variables are well correlated with the variables of SI. So it was decided to exploit variables of SI as auxiliary information at estimation stage. High correlation ensure improvement of quality. So it is expected to achieve the breakdown of annual data using the sample of quarterly survey.

Variables of SI analyzed:

- Number of insured persons: average on number of insured persons at the beginning of each quarter and end of each quarter;
- Taxable income per year;
- Days worked.

Coefficient of correlation for variables of Quarterly Survey of Earnings with variables of Social Insurance was calculated at NACE section level for each quarter of the years 2003 and 2004. The distribution of coefficients of correlation is presented in the table bellow.

Coefficients of correlation between variables of QSE and variables of SI, 2003 and 2004

Variable in SI	Coeff of corr	Variable in Quarterly Survey of Earnings				
		Number of employees	Number of full-time units	Gross remuneration	Hours worked	Hours paid
Number of insured	<0.8	0.0	0.0	3.1	0.0	0.0
	0.8-0.9	0.0	1.6	29.7	1.6	1.6
	0.9-1	100.0	98.4	67.2	98.4	98.4
Taxable income	<0.8	2.3	1.6	0.0	0.8	3.9
	0.8-0.9	26.6	19.5	0.0	21.9	19.5
	0.9-1	71.1	78.9	100.0	77.3	76.6
Days worked	<0.8	0.0	0.0	2.3	0.0	0.0
	0.8-0.9	0.0	0.0	28.1	0.0	0.8
	0.9-1	100.0	100.0	69.5	100.0	99.2

From the table above we can see that in most cases coefficient of correlation is higher than 0.9, some of them fall into interval [0.8; 0.9) and only few cases when it is less than 0.8. So it could be affirmed that there exist well-correlated auxiliary variables.

3.2 Notation

The purpose is to examine the estimation for domains using auxiliary information. It was decided to analyze possibility to introduce General Regression Estimator (GREG) for estimation of annual data.

Let us denote $U = \{1, 2, ..., k, ..., N\}$ - the *sample frame*. A probability sample s is drawn from U according to the specified sampling design. The sample size is denoted by n. The first order *inclusion probabilities* are denoted by $\pi_k = P(k \in s)$, the second order inclusion probabilities are denoted $\pi_{kl} = P(k\&l \in s)$. The corresponding *sampling weights* denoted $d_k = 1/\pi_k$ and $d_{kl} = 1/\pi_{kl}$.

Lets denote y - *the variable of interest*. The value of the *auxiliary variable vector* for the k-th element is denoted by $\mathbf{x_k} = (x_{1k}, ..., x_{Jk})$, J - the number of auxiliary variables.

The objective is to estimate the unknown y total:

$$t_y = \sum_{k \in U} y_k \tag{1}$$

when we have observed $(y_k, \mathbf{x_k})$ for $k \in s$ and when $\mathbf{x_k}$ is also known for $k \in U \backslash s$.

Generally, when J auxiliary variables are present, the General Regression Estimator is given by

$$\hat{t}_y^{greg} = \hat{t}_y + \sum_{j=1}^{J} \hat{B}_j (t_{x_j} - \hat{t}_{x_j}) = \hat{t}_y + \hat{\mathbf{B}}'(\mathbf{t_x} - \hat{\mathbf{t}}_\mathbf{x}) \tag{2}$$

where $\hat{t}_y$ is Horvitz-Thompson estimator of t_y, $\mathbf{t_x} = (t_{x_1}, ..., t_{x_J})'$ is the vector of known population total of the J auxiliary variables, and similar for $\hat{\mathbf{t}}_\mathbf{x}$, the vector of estimated population totals of the auxiliary variables. The $\hat{B}_1, ..., \hat{B}_J$ are components of the vector

$$\hat{\mathbf{B}} = \Big(\sum_{i \in s} d_i \mathbf{x_i} \mathbf{x_i'} \Big)^{-1} \Big(\sum_{i \in s} d_i \mathbf{x_i} y_i \Big) \tag{3}$$

The Generalized Regression Estimator can be alternatively written as

$$\hat{t}_y^{greg} = \sum_{i \in s} d_i g_i y_i, \tag{4}$$

where

$$g_i = 1 + (\mathbf{t_x} - \hat{\mathbf{t}}_\mathbf{x})' \Big(\sum_{i \in s} d_i \mathbf{x_i} \mathbf{x_i'} \Big)^{-1} \mathbf{x_i} \tag{5}$$

3.3 Simulation accomplished

The sampling frame and the sample of QSE is the same whole year. That is why it is possible to use the sample of QSE for annual data. The annual number of employees in the sampled enterprizes was calculated as average of four quarters, the gross earnings and hours - as the sum of the quarterly values.

The breakdowns required for Annual Survey of Earnings:

- NACE (two digits or sometimes even more detailed) & economic sectors (49 economic activities and 2 economic sectors), it is also the breakdown of QSE;

- NACE (section level) & size of enterprize & economic sector (15 economic activities, 6 sizes of enterprize and 2 economic sectors);;
- NACE (section level) & county (15 economic activities and 10 counties);
- Municipality (60 municipalities);

Total by $49 * 2 + 15 * 6 * 2 + 15 * 10 + 60 = 488$ partly overlapping domains are required. The sample size in 2004 is 6111 units. It is evident that it is impossible to get reliable data for such detailed breakdown using only the sample of QSE. As the data of SI became available for the statistical purposes it was decided to use this data as auxiliary information in order to calculate estimates by more detailed breakdown.

There are no problems with the first breakdown because it is also the breakdown of QSE and at the moment of sample was foreseen to get the results by this breakdown. The main problem is breakdown by regions because the quarterly survey does not aim to get data for the estimates for regions. If we want to get reliable results by detailed NACE and by regions using only the data from survey we need almost a complete enumeration of enterprizes. That used to be done till 2003.

The high correlation of variables of Social Insurance and variables of Annual Survey of Earnings let us expect that the usage of variables of Social Insurance will allow to switch from census of enterprizes to the sample survey.

The main task is to identify the most reliable vector of auxiliary information. As mentioned in 3.1 three auxiliary variables were analyzed. Also two levels of auxiliary information were examined:

- NACE at section level;
- NACE at section level & region at county level (10 counties).

Combining different auxiliary variables 14 GREG estimators were calculated: 7 possible combinations of three auxiliary variables multiplied by 2 levels of auxiliary information.

Notation of different GREG estimators

Notation	**Auxiliary information used**
G1, G8	Number of employees
G2, G9	Taxable income
G3, G10	Days worked
G4, G11	Number of employees and taxable income
G5, G12	Number of employees and days worked
G6, G13	Taxable income and days worked
G7, G14	All variables

G1 - G7 refer to auxiliary information at NACE section level and G8 - G14 refer to auxiliary information at NACE section level & county.

The main criteria for choosing the most suitable estimator from the list above was variance and distribution of weights g_i presented in formula (5). The variance should be as small as possible and the weights g_i should not be scattered too mush. But unfortunately, as presented in the tables bellow, the smaller the variance the weights g_i are more scattered.

Distribution of weights g_k for different GREG estimators 2004, in per cent

$\mathbf{g_i}$	**G1**	**G2**	**G3**	**G4**	**G5**	**G6**	**G7**	**G8**	**G9**	**G10**	**G11**	**G12**	**G13**	**G14**
< 0.4	0	0	0	0	0	0	0	2	2	67	3	4	65	5
$[0.4; 0.8)$	1	0	1	1	1	1	1	7	7	11	8	8	11	9
$[0.8; 1.2)$	97	97	97	96	95	96	94	59	59	7	56	54	7	52
$[1.2; 1.6)$	2	2	2	3	3	3	4	26	26	7	26	25	8	25
≥ 1.6	0	0	0	0	0	0	0	6	6	8	7	9	9	10

Distribution of the coefficients of variation for different GREG estimators 2004, in per cent

CV	**HT**	**G1**	**G2**	**G3**	**G4**	**G5**	**G6**	**G7**	**G8**	**G9**	**G10**	**G11**	**G12**	**G13**	**G14**
[0; 5)	44	44	44	44	44	43	44	44	70	63	83	82	75	88	86
[5; 10)	16	15	14	15	14	15	13	14	14	18	8	7	11	6	5
[10; 30)	26	26	26	26	27	27	26	26	11	13	6	7	10	3	5
≥ 30	15	16	16	16	16	15	16	16	5	6	3	4	4	4	4

The GREG estimator which was chosen for estimation of Annual Data of Earnings 2004 is G8, it uses only the number of insured persons as auxiliary information and level of auxiliary information is NACE at section level & county. This estimator is a compromise between small variance and scatter of weights g_i: 70% of coefficients of variation are less than 5% and 59% of weights g_i fall into interval [0.8; 1.2).

3.4 Precision gained

The chosen GREG estimator G8 was compared with the HT estimator. Some variables were improved very significantly but some only a bit. In the table bellow it is presented the distribution of coefficients of variation for the variable "Average number of employees" and for all variables altogether.

Distribution of statistical estimates (G8) by size of the coefficient of correlation (CV) 2004

CV	All variables		Average number of employees	
	G8	HT	G8	HT
Mean	9.8	14.4	7.0	18.5
Median	2.5	6.8	1.9	11.9
[0; 5)	70.1	43.7	76.7	32.4
[5; 10)	13.5	15.5	10.0	14.8
[10; 30)	11.4	25.7	8.1	31.4
[30; 50)	1.2	9.5	1.0	13.3
[50; 100)	3.7	5.6	4.3	8.1

We can see from the table above that median value of CV for chosen GREG estimator declines almost three times compare to HT estimator for all variables altogether and more than 6 times for average number of employees. Also the number of estimates with CV less than 5% is significantly higher for GREG estimator compare to HT.

As mentioned above all estimates for 2003 were compared with the real figures of ASE 2003. In fact the frame for ASE is not the same as for QSE but most enterprizes belongs to both frames. So G8 estimates may not coincide with the respective figures from ASE but should be close.

Distribution of statistical estimates for Average Number of Employees by deviation from ASE 2003, in %

Interval of deviation, in %	**Number of statistical estimates, in %**	
	G8	**HT**
[0; 5]	54.1	35.5
(5; 10]	16.9	19.8
(10; 20]	11.6	15.1
(20; 50]	12.8	21.5
50 and	4.7	8.1

From the table above it can be noticed that the G8 estimates for Average Number of Employees are closer to corresponding figures from Annual Survey of Earnings than the HT estimates. Similar situation are found calculating deviation for all other variables.

3.5 Improvements foreseen

Analyzing the results it was noticed some improvements that should be introduced for estimation of further annual data:

- More levels of auxiliary information should be analyzed and fist of all size of enterprize should be included;
- Maybe different auxiliary information should be used for estimation of different variables;
- Because of too detailed breakdown of data regression imputation should be implemented (variables should be imputed for whole population).

4 Conclusions

Introduction of administrative sources for estimation of data of Labour Statistics is undoubted a useful experience. Burden for enterprizes as well as for staff of Labour Statistics was significantly diminished. Approximately 40000 of enterprizes do not need to fill in annual questionnaire on earnings, staff of Statistics Lithuania do not need to enter and check those questionnaires and users are able to get information sooner than they used to. Useful experience enforced to start analysis of possibility to introduce administrative sources for other surveys on earnings.

5 References

[1] Deville, J; and Särndal, C.-E. Calibrated estimators in survey sampling. Journal of American Statistical Association, (1992, 87 p.376-382).

[2] Lundström, S. and Särndal, C.-E. Calibration as a standard method for treatment of nonresponse. Journal of Official Statistics, (1999, 15(2), p. 305-327).

[3] Särndal, C. E. Swensson, B., Wretman, J. Model Assisted Survey Sampling. Springer-Verlag, New York, (1992).