United Nations ECE/CES/GE.42/2011/7

Economic and Social Council

Distr.: General
5 August 2011

English only

Economic Commission for Europe

Conference of European Statisticians

Group of Experts on Business Registers
Twelfth session
Paris, 14-15 September 2011
Item 3 of the provisional agenda
Linking statistical business registers across agencies, statistical domains and among countries

Linking administrative and survey data – employment variable for enterprises and establishments in Finnish Business Register

Note by Statistics Finland

Summary

The purpose of this paper is to present the method of linking employees for enterprises and establishments and the method used for estimating the number of employees with the help of administrative and survey data in Finnish Business Register (BR). The administrative data used for this purpose is Tax Administration’s data on annual wages, which is a central source data for employment figures in Finnish Business Register. The data links each individual wage earner to each employer during the reference year and provides the wage bill. A regression model is applied and estimated by OLS. This is how the number of salaried employees measured as full-time equivalent (FTE) is reached for every employing enterprise in Business Register. The headcount measure is compiled from annual wages data using full-time equivalents and another data - Tax Administration’s payment control register for VAT and employer contributions. The paper also covers the method of linking entrepreneurs, which is done separately by using various administrative data.

The views expressed in this document are fully consistent with the vision of the High-Level Group for Strategic Directions in Business Architecture in Statistics, endorsed by the Conference of European Statisticians at its 59th Plenary Session in June 2011 (ECE/CES/2011/1 and ECE/CES/2011/CRP.1).

Please recycle

I. Introduction

1. There are two major concepts of number of persons employed in business register. Full-time equivalent variable presents labour input converted to full-time input, where full-time is considered as average working hours within the industry. Headcount (HC) corresponds to actual number of persons employed in certain period, without making distinction between full and part-time workers. For example, three persons working half time throughout the year would count three as HC but only 1.5 in FTE. For one enterprise one person can be more than one FTE but only one HC. For one person the sum of HC figures can also be more than one but in this case there must be more than one employer.

2. The definitions of employment variables are based on the business registers and structural business statistics (SBS) regulations. The variables for the number of persons employed and number of employees are measured in head counts. These variables are obligatory for business registers with the exception to SBS that calculation of annual averages is not required. The full time equivalent variable for the number of employees is voluntary.

3. From the statistical year 2011 onward the Finnish Business Register is responsible for the production of employment information at enterprise level. Employment statistics conduct surveys on persons, but other units do not undertake employment surveys on enterprises. The benefit of this organization is not only the reduction of response burden on informants, but also the consistency of information on employment by enterprises. Therefore coordination within the statistical office is required to maintain the consistency of relevant concepts. Also, regular contact with the Tax Administration is necessary to follow the changes in tax reporting..

II. Data sources

4. Finnish BR employment figures (paid employees and entrepreneurs in FTE and HC) production is based on tax-data on annual wages and salaries. Tax Administration’s annual wages declaration data consists of files provided by wage paying legal units. This dataset contains the business ID code (BID), employees personal ID code (PIN), the type and value of payment between the two. There are nearly 70 different types of payments in the data, of which not all types are considered salary in the estimation of employment variables. Excluded from salary are payments, such as interest and benefits, which are not paid for work performed for the principal employer. The National Accounts unit determine the concept of salary to maintain consistency with SNA and ESA manuals.

5. Crucial information what annual tax-data does not contain is the starting and ending dates for different working periods. Having these, whole estimation process could be replaced with much simpler process of counting annual averages directly from the data. Salaries paid by enterprises that have started after the 15th of November are also missing for the reference year.

6. A number of registers and surveys are utilized to provide data for the explanatory variables in the estimation equation. The indicator variable on the level of education is retrieved from the registry of degrees maintained by Education Statistics. Similarly, the occupation code indicator is used from administrative register data, which is however supplemented by wage-, employment- and occupation surveys.

7. The survey on enterprises with multiple local kind of activity units (LKAUs) is used in allocating the legal unit (LeU) level figures on employment to establishments. In addition, this survey provides the locations of workplaces in municipalities, which are used as the geographic indicator in the model. Business Register undertakes the bulk of this

survey in co-operation with employment statistics, but both BR and employment statistics have their own questionnaires for enterprises that are not included in the joint survey.

III. FTE estimation

8. The method is designed for forecasting the labour input of persons by legal unit. It is based on specifying an expected value of wages for persons, which depends on indicator variables formed from data on the characteristics of the person in question. The indicator variables are the activity and (institutional-) sector of employing legal unit as well as persons occupational code, educational level, age, gender and location of workplace. Annual labour input for each person is calculated by dividing his/her real annual wages by the expected value of his/her annual wage.

W_{ij} = person's (i) annual wages in legal unit (j), where

i = 1...N, where N = number of wage and salary earners

j = 1...M, where M = number of wage and salary payers (legal units)

9. A regression model is formed in which a person's (i) annual wage (w_i) is explained with data on the characteristics (x_i) of the person (i). The aim of the regression model is to describe the formation of a person's wages (i). Expected value of person's annual wage

$E(w_i) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p + \epsilon_i$, where

i = (1...N) and ϵ_i is random error, n.i.d.

Thus parameters $\beta_0 ... \beta_p$ can be estimated using the least square method (OLS).

The expected values of employees' annual wages are calculated with the help of the parameter estimates (values) and the data on their characteristics ($x_1 ... x_p$).

$E(w_i) = f(k_1 ... k_7)$, where

k_1 = occupation code indicators, formed by dividing 5-digit level occupational codes into 11 groups according to median wages (OCC1-11)

k_2 = activity indicators, formed by dividing 5-digit level activities into eight groups according to median wages (NAC1-8)

k_3 = educational level indicators, formed by dividing educational qualifications into seven groups according to level of education (EDU1-7)

k_4 = classification of sector indicators, where persons are divided to four groups according to sector classification of the (paying) legal unit (business unit, financing, general government, households) (SEC1-4)

k_5 = geographic indicator, where persons are divided to three groups according to the location of their workplace: growth centres (capital region, Tampere, Turku, Oulu), other urban settlements, and scattered settlements (Geo1-3)

k_6 = gender indicator (GEN)

k_7 = age variable (AGE)

10. Before fitting the model to data, certain procedures are carried out for ensuring the correctness of the wage data. Aim is to remove all non-typical working relations such as part-time employees, jobs outside of BR scope (households, non-market value wages jobs), secondary jobs and such. This is done by accepting only relations with:

- Principal job (flagged in Tax data)
- Wages >=40% of the overall median wages

- Continuous principal work relation from previous year

11. In addition to removing all non-typical working relations, median annual wages by occupational category were calculated and outlier observations were removed. Bias in total estimate was studied by comparing the estimates produced using the model to corresponding figures obtained from survey data.

12. The OLS estimation results obtained by placing the coefficient estimates in the model was (example 2007)

$E(w_i)$ = Wages = 12242 + 61*AGE + 1505*GEN + 779*GEO2 + 1440*GEO3 – 66*SEC1 + 1181*SEC2 - 829*SEC4 + 708*EDU3 + 1235*EDU5 + 1864*EDU6 + 3716*EDU7 + 5350*EDU8 + 2200*OCC2 + 5344*OCC3 + 8495*OCC4 + 12943*OCC5 + 17621*OCC6 + 23437*OCC7 + 30050*OCC8 + 37381*OCC9 + 52392*OCC10 + 60173*OCC11 + 1661*NAC2 + 2529*NAC3 + 2961*NAC4 + 3728*NAC5 + 4317*NAC6 + 4925*NAC7 + 6782*NAC8

F-test 160561 p = [0.0001]
Rate of determination 78.82

13. The coefficients have changed only moderately since 2007 and only the coefficient for household sector has been statistically insignificant. The sector coefficients have been the only ones to change their sign regularly.

IV. Headcount estimation

14. Like in FTE estimation, Tax Administrations annual wages and salaries data forms the base for the head count estimation. Previous year data is also utilised for defining if certain working relation has been continuous since previous year. During the FTE estimation the labour input by person and by working relation type (principal and secondary jobs) is compiled and saved for further use in head count estimation. In addition to these also the monthly VAT and PAYE (pay as you earn) data is utilised for defining the period of activity of the legal units in this population.

15. The idea of HC compilation is based on defining work relation types for each employee, number and continuity of jobs person is having as well as evaluating operating period for each legal unit. Employee type in certain LeU can be full-time, part time, change of full time and temporary working relation. Operating time of LeU is classified as “12 months” or “not known”, and “<12 months”. Below table illustrates the decision-making with different combinations.

If....		**... then**
Type of person =	**LeU operating period**	**HC impact =**
“Full-time job”	“12 months” or “not known”	1 (Full HC)
“Part time job”	“12 months” or “not known”	
“Change of full-time job”	“12 months” or “not known”	$\frac{PIN\_LeUeur}{PIN\_TOTeur}\times PIN\_FTEinput$ (total labour input divided to different LeU’s according to share of persons total wages)
“Part time job	“12 months” or “not known”	
“Change of full-time job”	“<12 months”	$MIN\left(\frac{\frac{PIN\_LeUeur}{PIN\_TOTeur}\times PIN\_FTEinput}{\frac{LeU\_opertime}{12}}, 1\right)$ PIN- LeU labour input scaled with LeU operating time
“Temporary job”	“<12 months”	

16. After each employee/LeU HC's are defined, total HC count is compiled simply by aggregating employee level HC's by LeU.

V. Estimation of the entrepreneur employment figures

17. A large part of small enterprises is run mainly on entrepreneurial work input where the income is taken out from the business in different form than in wages and salaries. Estimating number of employees is based on wages and salaries which contain also income paid for entrepreneur him-/herself and this part of entrepreneurs' labour input is counted among number of paid employees (according to BR recommendations). The aim of entrepreneur labour input estimation is to define the amount of work which the entrepreneur or entrepreneur's family has done without actual paid wages. Main data used in this process is the pension insurance scheme files where insured persons are presented by PIN. However, these files do not contain information about actual paying enterprise. Actual problem is to define first in which enterprise each insured person is working and second, what is person's actual labour input.

18. In addition to pension insurance scheme declarations, other used data are as follows: Annual tax declarations (as in FTE and HC), tax files on business partners, tax files on business owners, family relations files from social statistics and FTE estimation files containing labour input per person as paid employee.

19. In the first place number of potential entrepreneurs is defined. PINs from pension insurance files are matched step by step (if found from first file then excluded from other, etc.) with PINs from

- annual tax file, flagged entrepreneur work relation type (principal and secondary)[1]
- secondary ID for BR legal units with positive turnover/ agricultural income
- tax partner files where PIN flagged as active partner
- ownership files, only unique PIN-BID relations accepted

20. For each found pension scheme PIN - BID relation, the number of potential entrepreneur is set to 1. Non-found (pension file) PINs are further linked to other PIN's and BIDs found in previous stages with the help of family relations files. Entrepreneur input as head-count is simply the number of entrepreneurial PIN's found in these stages.

21. On second stage each potential entrepreneurs labour input (=1) is subtracted with the calculated labour input of this person in FTE estimation (0-1). Aggregating these by LeU, the figure represent theoretical upper limit of entrepreneur labour input in this LeU expressed in FTE. For adjusting the labour input against the size of the enterprise, turnover of the LeU is divided by average turnover per person employed (5-digit activity class) multiplied by 1.5 (factor for limiting the maximum entrepreneur input in certain enterprise, based on testing the data). This figure reduced by number of paid employees is used as statistical upper limit for FTE entrepreneurs. Smaller of the two is selected as actual number of entrepreneurs.

[1] If tax-file contains more than one BID as payer of wages, ID having largest turnover will be selected.

Statistical upper limit on entrepreneur labour input

$$SM_j = \mathbf{a} * (LV_j/LVP_d) - \sum_{i=1}^{N} \hat{S}_{ij} ,$$

SM_j = LeU (j) statistical upper limit of entrepreneur input

a=1,5 (factor for limiting the maximum entrepreneur input)

LV_j = LeU (j) turnover

LVP_d = average Turnover/person employed in activity class d,

$\sum_{i=1}^{N} \hat{S}_{ij}$ = LeU (j) FTE of paid employees

Theoretical upper limit on entrepreneur labour input

$TM_j = PM_j – WI_j$,

PM_j = Number of linked pension insured persons/entrepreneurs in LeU (j)

WI_j = linked pension insured persons estimated FTE in LeU (j)

LeU(j) estimated entrepreneur labour input

$LeU_j\ Y_j = \min(TM_j – SM_j)$

22. Further, the pool of non-linked PINs in insurance files is allocated to small businesses (with three or less employees) which were not linked on this estimation and where average turnover per person figure by industry is indicating too small labour input. In 2007 the total number of non-linked unique PINs was around 19,000. In the end, allocated total amount of non-specified entrepreneur working years was 23,000, divided to 66,000 LeUs . This amount was approximately 10% of all entrepreneur input.

VI. Conclusions

23. Estimation based on registry information can be considered a reliable way of estimating the employment variables in BR. However, considering the pivotal role of business register, care has to be taken to maintain the validity of estimation methods and source information. Business register surveys on enterprises are still necessary to regularly validate the estimated figures, but all surveys considered, estimation provides a way to reduce the total response burden on businesses.

24. Co-operation with the users of business register information, such as the National Accounts and information services is similarly important. It has been noted, that estimated statistical information may not always suit the needs of external users of information services and misuse of this information should be avoided. From the National Accounts perspective, it is essential to keep the wage sums in line with the employment figures.