Timo Laukkanen
Business Trends/ Business Register
email: timo.laukkanen@stat.fi
tel: +358 9 1734 3388

03 September 2009

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

Abstract

Administrative data include even 96% of the data items received to Statistics Finland, which emphasises the importance of its use. 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. 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 tells the wage bill. An OLS model is applied to reach the number of salaried employees measured as full-time equivalent 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 of entrepreneurs, which is done separately with the help of various administrative data.

Introduction

There are two major concepts of number of persons employed in business register world. Full-time equivalent (FTE) variable presents labour input converted to full-time input, where full-time is considered as average working hours within the industry. Headcount (HC) correspond actual number of persons employed in certain period, without making distinction between full and part-time workers. Three persons working half time throughout the year would count 3 as HC but only 1.5 in FTE. For one enterprise one person can be more than 1 FTE but only 1 HC. For one person the sum of HC figures can also be more than 1 but in this case there must be more than one employer.

Finnish BR employment figures (paid employees and entrepreneurs in FTE and HC) production is based on tax-data on annual wages and salaries. Allocation of LeU level figures to establishments is done using direct data collection to enterprises having more than one LKAU. Tax administrations annual wages declaration data (files provided by wage paying legal unit) contain link between business ID code (BID) and employees personal ID code (PIN). This data cover also actual wages and salaries figures as well as some additional information (flags for principal and secondary jobs, substitute payer, benefits, etc.). In addition to these, BR and employment statistics databases provide information covering businesses and individual employee information such as age, gender, workplace location, education, occupation, etc. 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.

Timo Laukkanen
Business Trends/ Business Register
email: timo.laukkanen@stat.fi
tel: +358 9 1734 3388

03 September 2009

FTE estimation

The method is designed for forecasting the labour input of persons by legal unit. It is based on specifying for every person an expected value of wages, 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 wages.

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)

For the estimating, a regression model is formed in which a person's (i) annual wages (w_i) are 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

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

i = (1...N) and ϵ_i is random error

Thus parameters $\beta_0 \ldots \beta_p$ 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 ($x_1 \ldots x_p$).

$E(w_i) = f(k_1 \ldots k_7)$, 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 the 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)

Before fitting the estimation model to data certain procedures are carried out for ensuring the correctness of the wages 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

In addition to this median annual wages by occupational category were calculated and observations which fell outside of certain range were removed. Bias in total estimate was stud-

Timo Laukkanen
Business Trends/ Business Register
email: timo.laukkanen@stat.fi
tel: +358 9 1734 3388

03 September 2009

ied by comparing the estimates produced using the model to corresponding figures obtained from survey data.

The regression equation obtained by placing the coefficient estimates in the model was (example 2000)

$E(w_i)$ = Wages = 53101 + 380.81*AGE + 744.43*GEN + 1418.86*GEO2 + 5184,41*GEO3 – 415.02*SEC2 + 4990.97*SEC3 + 8834.34*SEC4 + 2210.34*EDU3 + 3540.46*EDU4 + 7391.76*EDU5 + 4824.42*EDU6 + 13760*EDU7 + 20588*EDU8 + 7491,62*OCC2 + 13742*OCC3 + 26882*OCC4 + 34587*OCC5 + 46985*OCC6 + 65237*OCC7 + 85792*OCC8 + 107367*OCC9 + 128568*OCC10 + 170885*OCC11 + 4958,18*NAC2 + 7311,81*NAC3 + 10646*NAC4 + 14304*NAC5 + 18012*NAC6 + 26375*NAC7 + 26937*NAC8

F-test 62216.456 p = [0.0001]
Rate of determination 74.67

Headcount estimation

In same way as 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.

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 defined as “12 months” or “not known”, and “<12 months”. Below table illustrating the decision making with different combinations

<table>
<tr><th colspan="2">If....</th><th>... then</th></tr>
<tr><th>Type of person =</th><th>LeU operating period</th><th>HC impact =</th></tr>
<tr><td>“Full-time job”</td><td>“12 months” or “not known”</td><td rowspan="2">1
(Full HC)</td></tr>
<tr><td>“Part time job”</td><td>“12 months” or “not known”</td></tr>
<tr><td>“Change of full-time job”</td><td>“12 months” or “not known”</td><td rowspan="2">$\frac{PIN\_LeUeur}{PIN\_TOTeur} \times PIN\_FTEinput$
(total labour input divided to different LeU’s according to share of persons total wages)</td></tr>
<tr><td>“Part time job</td><td>“12 months” or “not known”</td></tr>
<tr><td>“Change of full-time job”</td><td>“<12 months”</td><td rowspan="2">$MIN\left(\frac{\frac{PIN\_LeUeur}{PIN\_TOTeur} \times PIN\_FTEinput}{\frac{LeU\_opertime}{12}}, 1\right)$
PIN- LeU labour input scaled with LeU operating time</td></tr>
<tr><td>“Temporary job”</td><td>“<12 months”</td></tr>
</table>

Timo Laukkanen
Business Trends/ Business Register
email: timo.laukkanen@stat.fi
tel: +358 9 1734 3388

03 September 2009

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

Estimation of the entrepreneur employment figures

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.

In addition to pension insurance scheme declarations, other used data are: 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.

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

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.

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.

Statistical upper limit on entrepreneur labour input

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

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

Timo Laukkanen
Business Trends/ Business Register
email: timo.laukkanen@stat.fi
tel: +358 9 1734 3388

03 September 2009

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 inpit

$TM_j = PM_j - WI_j$,

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

$LeU_j\ Y_j = \min(TM_j - SM_j)$

Further, the pool of non-linked PIN's in insurance files is allocated to small businesses (number of employees<=3) which were not linked on this estimation and where average turnover per person figure by industry is indicating too small labour input. In year 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 (10% of all entrepreneur input).

Employment figures allocation to local-kind-of-activity-units

Allocation of the employment input to local kind of activity units (LKAU) is based on direct data collection carried out by BR and by employment statistics. BR is surveying every multi-LKAU unit annually (with minor exceptions on some multi-site non-profit organisations) for maintenance of the LKAU structure in LeU's as well as total number of personnel in these. Statistics Finland's Regional Employment Statistics unit is collecting employee level information about persons employed in different organisations (legal units, local and central government units). Annual wages and salaries data (described earlier) provide link between organisation and person (BID-PIN) but since Finnish legislation does not recognise LKAU's (or any other 'statistical' units except LeU), these data are not sufficient for providing link between individual employee and his/her worksite. This 'missing link' is created by employment statistics direct data collection simultaneously carried out with BR inquiry for multi-site LeUs. Central government organisations LKAU and employee information is received centrally from State Treasury and local government information trough The Local Government Pensions Institution.

Employee-LKAU link enable compilation of versatile social and business statistics by combining this information to register and other administrative data covering persons (population register, occupational register, education register), organisations (business register with LeU/LKAU) and their point of activities (BR LKAU, register of houses and dwellings). Employee - LKAU link enables also compilation of the register based population census.