

# A Method For Short-Term Estimation of Labour Input Using Data from Administrative Sources having Coverage Errors

*Un Metodo per la Stima Congiunturale di Occupazione e Retribuzioni  
mediante l'Utilizzo Statistico di Archivi di Fonte Amministrativa*

Ciro Baldi

Piero Demetrio Falorsi

Alessandro Pallara

Raffaella Succi

ISTAT, Via Depretis 74/b, Rome (Italy)

e-mail: pallara@istat.it

Aldo Russo

Department of Political Institutions and

Social Sciences

University Roma Tre; Rome

**Riassunto:** Il presente lavoro propone un metodo statistico per la stima congiunturale di occupazione, retribuzioni e costo del lavoro mediante l'utilizzo degli archivi amministrativi di fonte INPS. La necessità di sviluppare metodologie per l'utilizzo statistico di tali archivi nasce dall'esigenza degli Istituti Nazionali di Statistica di rispondere alle richieste informative (e ai relativi standard di qualità) del *Regolamento No. 1165/98* della Comunità Europea relativo alle statistiche congiunturali sulle imprese. Tale regolamento richiede, infatti, una diffusione delle informazioni talmente disaggregata che risulta impossibile produrre i dati richiesti mediante l'utilizzo di indagini dirette.

**Keywords:** Administrative data; Short-term indicators; Coverage error; Predictive Approach; Selection bias

## 1. Introduction

In Italy social security (Istituto Nazionale della Previdenza Sociale, INPS) database (SSD) represents the most important administrative source for satisfying the requirements of the European *COUNCIL REGULATION No. 1165/98 concerning short-term business statistics* (STS), regarding provision in each Member state of quarterly estimates of: (a) number of persons employed, (b) wages and salaries and (c) hours worked. The relevance of this project may be thoroughly evaluated in view of the following issues: (i) information requested in STS Regulation involves such a detailed sectoral disaggregation (2-digit level of Nace Rev.1 classification) that it would be impossible to meet all the requirements through direct data collection; (ii) short term statistics on employment and wages have severe coverage problems in Italy, in that, at the moment, only one direct sample survey is currently executed, covering very large businesses (>500 employees) and specific groups of economic activities.

## 2. Parameters of interest and estimation method

Let  $P_t$  be the (finite) population of active enterprises for current time interval  $t$  (e.g., month or quarter). Furthermore, let  $A_t$  be the set of enterprises registered as *active* in the SSD. Because of the presence of overcoverage errors in the administrative source (due mainly to the circumstance that it may take some months before a dead unit cancels out from the social security register), it is  $P_t \subseteq A_t$ . Each enterprise registered in  $A_t$  when sending each month its remittance has to fill a form containing information on: *i*) number of employees, *ii*) wages and salaries, *iii*) social security costs; because of differences among units in the transmission procedures, in each reference period, information for current time interval  $t$  are available only for a *nonrandom* subset of units,  $s_t$  (currently, some 300,000 out of about 1,100,000 units in the population of enterprises with less than 500 employees in the industry and service sector in Italy, hereafter referred to as the *sample*); the subset  $\bar{s}_t$  of units on which we do not observe the variable of interest at time  $t$ , (the *nonsample*) will be available one year after the current reference period.

The parameter of interest are the totals of the variables *employment* and *wages and salaries* for the target population  $P_t$ , namely:

$$Y_t = \sum_{i \in P_t} y_{ti} = \sum_{i \in A_t} y_{ti} \lambda_{ti} \quad (1)$$

where:  $y_{ti}$  denotes the value of the variable of interest  $y$  for the enterprise  $i$  at time  $t$  and  $\lambda_{ti}$  is an indicator variable, which equals 1 if enterprise  $i \in P_t$  (i.e. it is an *active* unit and therefore *included correctly* in administrative register) and equals 0 otherwise.

The proposed estimation method is based on a predictive model (Royall, 1988), assuming that a relation exists between the variable of interest and some auxiliary variables. When modelling the relationship between the *sample* and the *nonsample* proper attention has to be given to two different potential sources of bias.

- non *randomness* of  $s_t$ . The solution adopted in this paper has been to fit the model within subgroups of the population (*regression groups*), assuming that the model coefficients are the same for the units in each regression group (cf. Royall, 1988);
- *coverage errors* in  $\bar{s}_t$ . It may be assumed that a partition of  $A_t$  exists such that in each of the subsets of  $A_t$  the probability of correct inclusion in the administrative source is nearly equal for all the units of the subset; in what follows these subsets will be denoted as *register error groups*.

Hence, let  $A_{tw} = s_{tw} \cup \bar{s}_{tw}$  denote the  $w$ -th *regression group* ( $w = 1, \dots, W$ ). Denote further with  $A_{tr}$  ( $r = 1, \dots, R$ ) the  $r$ -th *register error group*.

It is assumed that for  $w = 1, \dots, W$  the following superpopulation model holds:

$$\begin{aligned} E(y_{ti} | \lambda_{ti} = 1) &= \beta'_{tw} \mathbf{x}_{ti} & \forall i \in A_{tw} \\ \text{Cov}(y_{ti}, y_{t'i'} | \lambda_{ti}, \lambda_{t'i'}) &= \begin{cases} c_{ti} \sigma_{tw}^2 & \text{for } (t = t') \cap (i = i') \cap (\lambda_{ti} = 1) \\ 0 & \text{otherwise} \end{cases} & \forall i \in A_{tw} \end{aligned} \quad (2)$$

where  $\mathbf{x}_{ti}$  is a vector of  $M_{tw}$  auxiliary variables for the enterprise  $i$  at time  $t$ ,  $\boldsymbol{\beta}_{tw}$  is a column vector of  $M_{tw}$  regression coefficients and  $c_{ti}$  is a known constant related to the dimension of the enterprises reported in the administrative source.

Furthermore, we assume that *correct inclusion* in SSD is defined by a sequence of independent Bernoulli trials, independently of the value of  $y$ . These trials are assumed to be identically distributed within each register error group  $A_{tr}$  ( $r = 1, \dots, R$ ). Therefore, it is postulated that for  $r = 1, \dots, R$  the following model holds:

$$E(\lambda_{ti}) = p_{ti} = \theta_{tr} \quad \forall i \in A_{tr} \quad (3)$$

$$\text{Cov}(\lambda_{ti}, \lambda_{t'i'}) = \begin{cases} \theta_{tr}(1 - \theta_{tr}) & \text{for } (t = t') \cap (i = i') \\ 0 & \text{otherwise} \end{cases} \quad \forall i \in A_{tr}$$

With this set-up, the following model is derived for  $w = 1, \dots, W$  and  $r = 1, \dots, R$

$$E(y_{ti} \lambda_{ti}) = \boldsymbol{\beta}'_{tw} \mathbf{x}_{ti} \theta_{tr} \quad \forall i \in A_{t(w,r)} \quad (4)$$

$$\text{Cov}(y_{ti} \lambda_{ti}, y_{t'i'} \lambda_{t'i'}) = \begin{cases} c_{ti} \sigma_{tw}^2 \theta_{tr} + (\boldsymbol{\beta}'_{tw} \mathbf{x}_{ti})^2 \theta_{tr}(1 - \theta_{tr}) & \text{for } (t=t') \cap (i=i'), \forall i \in A_{t(w,r)} \\ 0 & \text{otherwise} \end{cases}$$

where  $A_{t(w,r)} = A_{tw} \cap A_{tr}$ .

The estimates of  $\boldsymbol{\beta}_{tw}$  and  $\theta_{tr}$  have been obtained separately for each set of parameters as:

$$\tilde{\boldsymbol{\beta}}_{tw} = \left[ \sum_{i \in s_{tw}} \mathbf{x}_{ti} \mathbf{x}'_{ti} / c_{ti} \right]^{-1} \sum_{i \in s_{tw}} \mathbf{x}_{ti} y_{ti} / c_{ti} ; \quad \tilde{\theta}_{tr} = \hat{\theta}_{t-12r} = \sum_{i=1}^{N_{t-12r}} \lambda_{t-12i} / N_{t-12r}$$

where  $N_{t-12r}$  denotes the number of units belonging to the  $r$ -th register error group one year earlier the current reference period, being obviously  $E(\hat{\theta}_{t-12r}) = \theta_{t-12r}$ .

Using a predictive approach, having obtained the estimates  $\tilde{\boldsymbol{\beta}}_{tw}$  ( $w = 1, \dots, W$ ) and  $\tilde{\theta}_{tr}$  ( $r = 1, \dots, R$ ) of the parameters involved in models (2)-(4), the estimate of the total  $Y_t$  is given by

$$\tilde{Y}_t = \sum_{w=1}^W \sum_{r=1}^R \left[ \sum_{i \in s_{t(w,r)}} y_{ti} + \sum_{i \in \bar{s}_{t(w,r)}} \tilde{\boldsymbol{\beta}}'_{tw} \mathbf{x}_{ti} \tilde{\theta}_{tr} \right] \quad (5)$$

where  $s_{t(w,r)} = s_t \cap A_{t(w,r)}$  and  $\bar{s}_{t(w,r)} = \bar{s}_t \cap A_{t(w,r)}$ .

### 3. Estimation of the mean square error

Using actual data on the entire population which are available, as mentioned, with a reference period preceding one year the current reference period, having observed the

realized value of the parameter  $Y_t$ , it is possible to obtain, one year after the current time, estimates of the predictive MSE and of the predictive bias as:

$$\text{MSE}(\tilde{Y}_t|Y_t) = (\tilde{Y}_t - Y_t)^2 \quad (6)$$

$$[\tilde{\text{Bias}}(\tilde{Y}_t|Y_t)]^2 = (\tilde{Y}_t - Y_t)^2 - \tilde{V}[(\tilde{Y}_t - Y_t)|Y_t]. \quad (7)$$

In (7) the predictive variance may be calculated from current sample data as

$$\begin{aligned} \tilde{V}[(\tilde{Y}_t - Y_t)|Y_t] = & \sum_{w=1}^W \sum_{r=1}^R \left\{ \sum_{i \in \bar{s}_t(w,r)} c_{ti} \tilde{\sigma}_{tw}^2 \tilde{\theta}_{tr} + \theta_{tr} (1 - \tilde{\theta}_{tr}) \tilde{\mathbf{p}}'_{tw} \mathbf{x}_{ti} \mathbf{x}'_{ti} \tilde{\mathbf{p}}_{tw} + \right. \\ & \left. + \sum_{i \in \bar{s}_t(w,r)} \left[ (2\hat{\theta}_{t-12r}^2 - \hat{\theta}_{t-12r}) \tilde{\sigma}_{tw}^2 \mathbf{x}'_{ti} \left( \sum_{i \in s_{tw}} \mathbf{x}_{ti} \mathbf{x}_{ti}' / c_{ti} \right)^{-1} \mathbf{x}_{ti} + (\hat{\theta}_{t-12r} (1 - \hat{\theta}_{t-12r})) \mathbf{x}'_{ti} \tilde{\mathbf{p}}_{tw} \tilde{\mathbf{p}}'_{tw} \mathbf{x}_{ti} \right] \right\} \end{aligned}$$

with  $\tilde{\sigma}_{tw}^2 = \frac{1}{n_{tw} - M_{tw}} \sum_{i=1}^{n_{tw}} (y_{ti} - \tilde{\mathbf{p}}'_{tw} \mathbf{x}_{ti})^2 / c_{ti}$ , where  $n_{tw}$  is the number of units in

the subset  $s_{tw}$ .

In order to obtain an estimate of the bias for the current time period, it may be useful to use an approach similar to the one typically adopted with *generalised variance functions* (Wolter, 1985), assuming that: (i) the relative squared bias is a decreasing function of the *sampling fraction* ( $n/N$ ); (ii) the model described in section 2 are nearly constant over short time. The relative square bias may then be modelled as

$$\ln \left[ (\tilde{\text{Bias}}(\tilde{Y}_t|Y_t))^2 / \tilde{Y}_t^2 \right] = a_t + b_t \ln(n_t / N_t) + \varepsilon_t$$

where  $a_t$  and  $b_t$  represent fixed coefficients and  $\varepsilon_t$  is a random residual. Assuming that the coefficients  $a_t$  and  $b_t$  are nearly constant over short time, an estimate of the *current* squared bias may be obtained as:

$$(\tilde{\text{Bias}}(\tilde{Y}_t|Y_t))^2 = \tilde{Y}_t^2 \exp(\tilde{a}_{t-1} + \tilde{b}_{t-1} \ln(n_t / N_t))$$

where  $\tilde{a}_{t-1}$  and  $\tilde{b}_{t-1}$  are the estimates of the parameters  $a$  and  $b$  obtained one year before the current time. Therefore at current time an estimation of MSE is given by:

$$\text{MSE}(\tilde{Y}_t|Y_t) = \tilde{V}[(\tilde{Y}_t - Y_t)|Y_t] + (\tilde{\text{Bias}}(\tilde{Y}_t|Y_t))^2.$$

## References

- Royall, R.M. (1988), "The Prediction Approach to Sampling Theory", in: Krishnaiah, P.R., Rao, C.R. (eds.), *Handbook of Statistics*, vol. 6, Elsevier Science Publishers, pp. 399-413.
- Wolter, K.M. (1985), *Introduction to variance estimation*, New-York; Springer-Verlag.