Q2016
European Conference on
Quality in Official Statistics

A latent class model to estimate labour cost from Multi-Source data

Session number 3:

Quality Challenges in Social Statistics: preserving privacy and other issues.

Ugo Guarnera
Silvia Pacini
Roberta Varriale
(Istat)

INE Instituto Nacional de Estadística
eurostat
Madrid, May 31 - June 3

The estimation problem

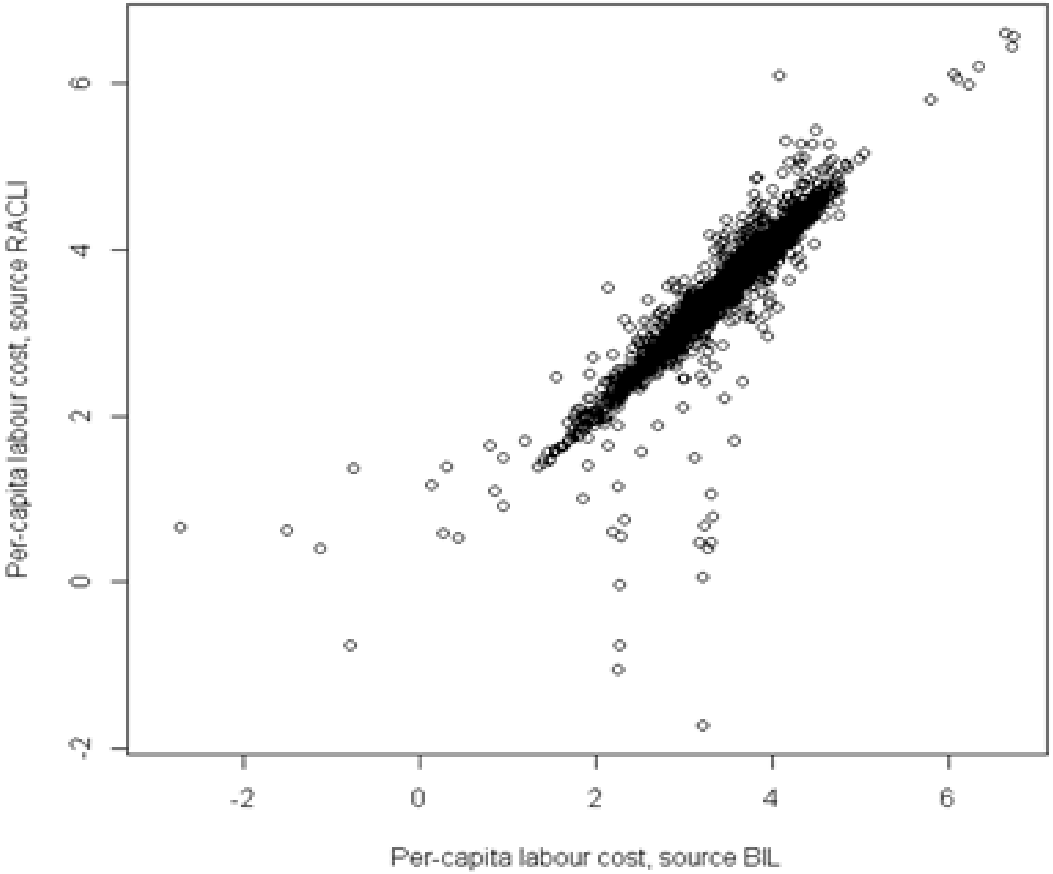
Goal: estimation of Labour Cost (LC) for large enterprises (>100 employees) based on 2 administrative sources:

- S1: Istat register of *wages, worked hours and labour cost* based on social security data (RACLI)
- S2: data from financial statements (BIL)

BIL chosen as reference source for two reasons:

1) more compliant with international regulation on Structural Business Statistics (SBS)
2) financial statements also used for other SBS variables → coherence among items of the enterprise account

BIL vs RACLI


Per-capita labour cost, source RACLI
Per-capita labour cost, source BIL
-2
0
2
4
6
-2
0
2
4
6

Multi-Source inference

Different roles of administrative sources:

- BIL is considered a *primary source of information*, i.e., it provides a direct, although imperfect, measure of the variable of interest (Y variable). However, part of data is affected by measurement errors both of "systematic" and random nature. An important systematic source of error is the inclusion of costs for "*agency workers*" and "*external workers*" in the total labour cost.

- RACLI data is *auxiliary* information (X variable). Although strongly related to the analysed phenomenon, the RACLI variable is considered as conceptually distinct from the target *labour cost* variable. On the other hand, data refer to enterprise employees *stricto sensu*.

The adopted model reflects the different roles of the external sources.

The latent class model

d_i	number of employees of enterprise *i*
Y_i^*	true (unobserbved) per capita (p.c.) value of the labour-cost
Y_i	p.c. labour-cost from BIL
X_i	p.c. labour-cost from RACLI
$C_i^{age(ext)}$	total known cost for agency (external) workers; $V_i^{age(ext)} = C_i^{age(ext)} / d_i$

true data model:

$$Y_i^* = \beta X_i + U_i \qquad U_i \sim N(0, \sigma^2)$$

measurement (error) model:

$$Y_i = Y_i^* + Z_{i,age} V_i^{age} + Z_{i,ext} V_i^{ext} + Z_{i,\varepsilon} \varepsilon_i$$

$$Z_{i,age} \sim Be(\pi_{age})$$

$$Z_{i,ext} \sim Be(\pi_{ext})$$

$$Z_{i,\varepsilon} \sim Be(\pi_{\varepsilon})$$

$$\varepsilon_i \sim N(0, \alpha\sigma^2)$$

model parameters : $\theta = (\beta, \sigma^2; \pi_{age}, \pi_{ext}, \pi_{\varepsilon}, \alpha)$

Observed data distribution

observed (BIL) data model:

$$Y_i = \beta X_i + Z_{i,age} V_i^{age} + Z_{i,ext} V_i^{ext} + Z_{i,\varepsilon} \varepsilon_i + U_i$$

the distribution of the observed BIL data is a mixture of Gaussian distributions, each component being associated with a particular error pattern $\mathbf{Z}_i = (Z_{i,age}, Z_{i,ext}, Z_{i,\varepsilon})$.

We assume *a priori* error independence:

$$p_{z_i} \equiv P(Z_{i,age} = z_{i,age}, Z_{i,ext} = z_{i,ext}, Z_{i,\varepsilon} = z_{i,\varepsilon}) = P(Z_{i,age} = z_{i,age}) P(Z_{i,ext} = z_{i,ext}) P(Z_{i,\varepsilon} = z_{i,\varepsilon})$$

Once the model parameters have been estimated (via EM algorithm), one can obtain estimates of the *posterior probabilities (PP)*

$$\tau_{z_i} \equiv P(Z_{i,age} = z_{i,age}, Z_{i,ext} = z_{i,ext}, Z_{i,\varepsilon} = z_{i,\varepsilon} | X_i, Y_i)$$

and thus the (marginal) prob. corresponding to each type of error.

PPs can be used to adjust BIL data for the erroneous inclusion of labour cost items (systematic errors), or to identify outliers.

Conditional true data distribution

Predictions $\hat{Y}_i$ of the true labour cost can be obtained taking expectations form the conditional distribution of Y_i^* given the available information

$(X_i, Y_i; i = 1,.., n)$

Using Bayes formula, it can be easily shown that :

$$\boxed{E(Y_i^* \mid X_i = x_i, Y_i = y_i) = \sum_{\{z_i\}} \tau_{z_i} \frac{y_i + \alpha z_{i,\varepsilon} \beta x_i - z_{i,age} V_i^{age} - z_{i,ext} V_i^{ext}}{1 + \alpha z_{i,\varepsilon}}}$$

Note that in case $Z_{i,\varepsilon} = 0$ (no random error) , Y_i^* is a deterministic function of Y_i given $Z_{i,age}$ and $Z_{i,ext}$. In fact, the true value could be recovered by simply subtracting $z_{i,age} V_i^{age} + z_{i,ext} V_i^{ext}$ from Y_i.

In case $Z_{i,\varepsilon} = 1$, for large α (large random error), the observed value Y_i provides little additional information with respect to the unconditional expected value βx_i .

Editing BIL labour cost

If the model is used to adjust BIL data for the presence of costs due to agency and external workers, a possible approach (*systematic approach*) is to subtract the expected value of the systematic error component

$$c_i = \hat{p}_{i,age} V_i^{age} + \hat{p}_{i,ext} V_i^{ext}$$

from the observed value Y_i where $\hat{p}_{i,age}$ and $\hat{p}_{i,ext}$ are the estimates of the posterior probabilities $P(Z_{i,age} = 1 | X_i, Y_i)$ and $P(Z_{i,ext} = 1 | X_i, Y_i)$ respectively.

Based on the latter, one can also use a a *classification approach:* subtract the quantity V_i^{age} and /or V_i^{ext} whenever the corresponding posterior probability is above a certain threshold (e.g., 0.5)

Application

n=8,866, year=2012.

# enterprises with agency workers = 4,755 (53.6%)

# enterprises with external workers = 7,165 (80.1%)

Parameter estimates:

$$\boxed{\hat{\beta} = 1.02, \hat{\sigma}^2 = 1.37; \hat{\pi}_{age} = 0.62, \hat{\pi}_{ext} = 0.22, \hat{\pi}_{\varepsilon} = 0.12, \hat{\alpha} = 273}$$

It seems that more than half of the enterprises (62%) having agency workers tend to include their costs in the financial statements. Random error has high impact: 12% of outlying observations , some of them very far from the bulk of the data $(\hat{\alpha} = 273)$.

Estimate of total labour cost based on predictions $(\sum_{i=1}^{n} \hat{y}_i)$ is 4.13% lower than the estimate based on BIL raw data.

However, only 1.13% of the total difference seems caused by systematic error $(\sum_{i=1}^{n} c_i)$.

Conclusions

- According to the current deterministic approach, costs for agency/external workers value are removed if, after correction, difference between BIL and RACLI value is «typical» (evaluated on enterprises having only «proper employees»).

- Modeling true data and errors allows to assess the quality of administrative sources to be possibly used in place of survey data.

- Latent class models provide useful tools like posterior error probabilities and smoothed values (posterior means) to be used for robust estimation, or at least for (selective) editing of influential errors corresponding to large discrepancies between observed and expected values.

- Latent class models are useful in multi-source contexts, i.e., in situations where more than one measure of the target phenomenon is available.

- Gaussian models for true data and errors are easy to manage. Research is needed to extend methodology to more general distributions.