SIS 2017
Statistics and Data Science: new challenges, new generations

28–30 June 2017
Florence (Italy)

Proceedings of the Conference of the Italian Statistical Society

edited by
Alessandra Petrucci
Rosanna Verde

FIRENZE UNIVERSITY PRESS
2017

SIS 2017. Statistics and Data Science: new challenges, new generations : 28-30 June 2017 Florence (Italy) : proceedings of the Conference of the Italian Statistical Society / edited by Alessandra Petrucci, Rosanna Verde. – Firenze : Firenze University Press, 2017.
(Proceedings e report ; 114)

http://digital.casalini.it/9788864535210

ISBN 978-88-6453-521-0 (online)

Peer Review Process
All publications are submitted to an external refereeing process under the responsibility of the FUP Editorial Board and the Scientific Committees of the individual series. The works published in the FUP catalogue are evaluated and approved by the Editorial Board of the publishing house. For a more detailed description of the refereeing process we refer to the official documents published on the website and in the online catalogue of the FUP (www.fupress.com).

Università degli Studi di Firenze
Firenze University Press
via Cittadella, 7, 50144 Firenze, Italy
www.fupress.com

Index

Alessandra Petrucci, Rosanna Verde (edited by), *SIS 2017. Statistics and Data Science: new challenges, new generations. 28-30 June 2017 Florence (Italy). Proceedings of the Conference of the Italian Statistical Society*
ISBN (online) 978-88-6453-521-0 (online), CC BY 4.0, 2017 Firenze University Press

Using administrative data for statistical modeling: an application to tax evasion

L'uso di dati amministrativi per la modellizzazione statistica: un'applicazione all'evasione contributiva

Maria Felice Arezzo and Giuseppina Guagnano

Abstract Administrative data, gathered by public authorities with a general aim of control, are very precious sources of information because they allow to study phenomena that would remain otherwise unknown. On the other side, administrative data strictly contain the information they were collected for, and to be used for statistical purposes they need to be integrated. This work shows the potentials of the integration of three data sets for statistical modeling: the audits carried out in Italy in 2005 by the National Institute of Social Security on building and costruction companies, the ASIA archive of Istat and the "'Studi di Settore"' of the Italian Revenue Agency.
Abstract *I dati amministrativi, raccolti dalle istituzioni pubbliche per scopi generalmente di controllo, sono fonti informative estremamente preziose in quanto permettono spesso di studiare fenomeni che in altro modo non potrebbero essere conosciuti. D'altro canto, proprio perchè rispondono a finalità specifiche, le indagini amministrative non contengono informazioni aggiuntive rispetto a quelle per le quali sono state pensate. Il lavoro illustra le potenzialità dell'integrazione di tre basi dati da fonte differente: le ispezioni INPS, l'archivio ASIA dell'Istat e gli Studi di settore dell'Agenzia delle entrate. La sperimentazione è stata condotta sulle imprese che operano nel settore delle costruzioni.*

Key words: Administrative data, Sample selection, Response-based sampling

Maria Felice Arezzo
Sapienza University of Rome, Address, e-mail: mariafelice.arezzo@uniroma1.it

Giuseppina Guagnano
Sapienza University of Rome, Address e-mail: giuseppina.guagnano@uniroma1.it

Alessandra Petrucci, Rosanna Verde (edited by), *SIS 2017. Statistics and Data Science: new challenges, new generations. 28-30 June 2017 Florence (Italy). Proceedings of the Conference of the Italian Statistical Society*
ISBN (online) 978-88-6453-521-0 (online),

1 Introduction

Administrative data are archives of great interest as they often contain information available only to public authorities responsible for the control of some phenomena. Almost always, though, these files do not contain information other than those for which they were collected (a typical example are the socio-economic characteristics of the individuals), as the purpose underlying their gathering is not statistical modeling. For this very same reason, administrative data require, on the one side, a throughout pretreatment and validation process and, on the other, the development of statistical methodologies that allow for the drawing of valid inferences.

The purpose of our work is to draw the entire "production chain": a) the creation of a dataset with all relevant variables, b) the evaluation of the dataset quality, c) the development of a statistical method suitable for the data at stake.

The case study is on the detection of the firms which evade worker contributions because they employ off-the-book workers (i.e. employee who are completely unknown to fiscal authorities)

2 Creation of the data set

Our starting point is an administrative dataset on the audits carried out in Italy in 2005 by the National Institute of Social Security (INPS henceforth) on building and construction companies (NACE section: F). It amounts to a total of 31,658 inspections on 28,731 firms. The global amount of firms operating in the building industry in Italy in the same year was $N = 595,226$. Audits data allow to observe the compliant/non-compliant behavior.

Following the idea that the risk of a non-compliant behavior can be predicted by the economic characteristics of the firm, we integrated the information of audits with two other sources of data. The first is the ASIA archive owned by the National Institute of Statistics (ISTAT). It contains data on the legal structure, turnover and number of employee and is a high quality source of data as the information are validated through a very careful process. The second, owned by the Italian Revenue Agency, is the so called 'Studi di Settore' (SS in the following) archive. It contains an exhaustive list of information on corporate organization, firm structure, management and governance.

The three data sets were merged using VAT numbers and/or tax codes. Surprisingly the match rate was only 51% meaning that the number of firms in the merged archive is 14,651.

The original variables were used to build economic indicators which can be grouped in the following different firm's facets: a) 9 indicators for economic dimension, b) 13 for organization, c) 6 for structure, d) 6 for management, e) 11 for performance f) 38 for labor productivity and profitability g) 3 for contracts award mode h) 7 variables for location and type. The final dataset had 93 independent variables observed on 14,651 building companies with a match rate of 51%. The

Table 1 Datasets characteristics

Data Owner	Content	Individual	Dimension
INPS	Inspections outputs (2005)	Inspection	31,658 inspections on 28,731 firms
Revenue Agency	Studi di settore (2005).Models: TG69U, TG75U (SG75U),TG50U (SG50U and SG71U), TG70U	Firm	Universe of firms with at most 5 million euros of income
ISTAT	Asia Archives (2005)	Firm	Universe of firms

variable to be predicted is named Y and it takes value 1 if in a firm there is at least one off-the-book worker and 0 otherwise. In the following we will refer to the final dataset as the integrated db because it gathers and integrate information from different sources.

3 The assessment of the integrated dataset

As we said, the matching rate was 51% which means that we had information on the features of interest for (roughly) half of the firms in original INPS database. We studied inspection coverage and the risk of non complying for different turnover class and corporate designation typologies and over the territory. The idea was to verify if a whole group of firms (for example all the companies in a geographical region) was lost because of the merging process.

We checked for: Regions (20 levels), Number of employee (9 classes), Legal structure (5 levels), Turnover (11 classes); we then made sure that during the matching procedure, no whole groups of individuals were lost.

4 The Model

Under a statistical point of view, there are two main methodological issues arising from the type of data we use. The first is the non-randomness of the inspections and the second is the fraction of inspected firms in the population is low.

SELECTION BIAS IN THE SAMPLE OF INSPECTED FIRM. To detect undeclared work, an inspector audits firms. Inspected firms are not randomly chosen; they are chosen because the inspector thinks that there are some off-the-book workers and s/he has strong incentives to target the "right" firms (i.e. the irregular ones). We can think of the decision to inspect a firm as a rational process in which the inspection is made if the utility to inspect, U^A, (i.e. find undeclared workers and get a benefit) is higher than the utility of non-inspect, $U^{\bar{A}}$. Moreover we can observe the status of the $i-th$ firm (regular or not) only if it has been inspected, otherwise a censoring process intervenes. It is obvious that there is a strong selection bias in the sample of inspected firms.

As it is well known, [3] proposed a useful framework for handling estimation when the sample is subject to a selection mechanism. In the original framework, the outcome variable is continuous and can be explained by a linear regression model (called *output equation*), with a normal random component; in addition to the output equation, a *selection equation* describes the selection rule by means of a binary choice model (probit).

In our framework the output equation defines the compliance decision, so the dependent variable is binary, and the selection equation refers to the decision of inspecting a firm. Just as the inspection decision, the evasion is based on a rational process and it happens if the utility of evading, $U^{\overline{C}}$, is greater than the utility of complying U^C. The corresponding econometric model, in its general form, is:

$$Y_i^* = U_i^{\overline{C}} - U_i^C = \boldsymbol{X}_{1i}\beta + \varepsilon_{1i} \quad (1a)$$

$$A_i^* = U_i^A - U_i^{\overline{A}} = \boldsymbol{X}_{2i}\theta + \varepsilon_{2i} \quad (1b)$$

where $\boldsymbol{X}_i = (\boldsymbol{X}_{1i}, \boldsymbol{X}_{2i})$ is a vector of exogenous variables (namely, $\boldsymbol{X}_{1i}$ for Y_i and $\boldsymbol{X}_{2i}$ for A_i), containing all the relevant covariates.

Since we cannot observe directly the utilities (neither those determining compliance, nor those governing the decision to inspect), we assume that if in equation (1a) $Y_i^* > 0$, the firm does not comply, otherwise it does. Let's define a dummy variable Y_i which we can observe and that denotes the alternative selected:

$$Y_i = \begin{cases} 1 & \text{if } Y_i^* > 0 \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

Similarly, we can define an observable dummy variable A_i for the inspections, such that:

$$A_i = \begin{cases} 1 & \text{if } A_i^* > 0 \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

The p.d.f. of Y_i and A_i is Bernoulli with probability of success respectively equal to ${}_Y\pi$ and ${}_A\pi$ and depending on $\boldsymbol{X}_{1i}\beta$ and on $\boldsymbol{X}_{2i}\theta$. A selection bias exists if $corr(\varepsilon_1, \varepsilon_2) = \rho$ is not null.

As it is known (see for example [2]), the likelihood function for the Heckman's selection model is:

$$L(\eta) = \prod_{i=1}^{n} \Big[1 - {}_A\pi(\boldsymbol{X}_i)\Big]^{1-A_i} \cdot \Big[f(Y_i|A_i = 1) \cdot {}_A\pi(\boldsymbol{X}_i)\Big]^{A_i} \quad (4)$$

where $\eta = (\beta, \theta, \rho)$ is the vector of parameters to be estimated.

THE CASE-CONTROL SETTING. In this sampling design [4], also known as response-based, samples of fixed size are randomly chosen from the two strata identified by the dependent variable A. In particular n_A units are drawn at random from the N_A cases and $n_{\overline{A}}$ from the $N_{\overline{A}}$ controls.

The likelihood function is the product of the two stratum-specific likelihoods and depends on the probability that the individual is in the sample, and on the joint density of the covariates:

$$\prod_{i=1}^{n_A} Pr(\boldsymbol{X}_i|A_i = 1, S_i = 1) \cdot \prod_{i=1}^{n_{\overline{A}}} Pr(\boldsymbol{X}_i|A_i = 0, S_i = 1). \qquad (5)$$

The c-c design is particularly suited in our study because the probability that a firm is inspected is very low and therefore it is much more convenient to directly sample from the two strata (inspected/non-inspected).

A BINARY CHOICE MODEL WITH SAMPLE SELECTION AND CASE-CONTROL SAMPLING SCHEME. In the following we provide the likelihood function under the framework of interest, i.e. a sample selection mechanism with a severe censoring process. The interested reader can find the full proof and the simulation results in [1].

We make the following very general and non restrictive assumptions:

1. we have a set of fully informative and exogenous covariates $\boldsymbol{X}_i = (\boldsymbol{X}_{1i}, \boldsymbol{X}_{2i})$;
2. conditional on the covariates, the probability that an observation is uncensored doesn't depend on its value, i.e. $P(A_i = 1|S_i = 1, \boldsymbol{X}_i, Y_i) = P(A_i = 1|S_i = 1, \boldsymbol{X}_i)$;
3. the set of covariates $\boldsymbol{X}_{1i}$, specific for Y_i, and the set $\boldsymbol{X}_{2i}$, specific for A_i, may have common elements but they cannot fully overlap;
4. the probability of being in the sample does not depends neither on the covariates $\boldsymbol{X}_i$ nor on Y_i. More precisely, letting S_i be a binary variable which takes value 1 if the $i-th$ individual is in the sample and 0 otherwise, it is true that $P(S_i = 1|\boldsymbol{X}_i, Y_i, A_i = a_i) = P(S_i = 1|A_i = a_i)$.

Assumption (1) means that it does not exist correlation between the covariates and the residual terms in equations (1a) and (1b). Assumption 2 is justified because, as the covariates are informative, all the information brought by Y_i is contained in $\boldsymbol{X}_i$. Assumption (3) is necessary for parameters identification (exclusion conditions). Assumption (4) is typical in the response-based sampling framework and no further explanation is required.

Under the conditions stated, the likelihood function for a binary choice model with sample selection under a response-based sampling is:

$$L(\eta) = \prod_{i=1}^{n} f(\boldsymbol{X}_i|S_i = 1) \left\{ (1 - {}_A\pi(\boldsymbol{X}_{2i})) \cdot \frac{N}{N_{\bar{A}}} \right\}^{1-A_i} \tag{6}$$
$$\cdot \left\{ \left[{}_Y\pi(\boldsymbol{X}_{1i}) \cdot \frac{N}{N_A} \frac{n_A}{n_{1A}} \right]^{y_i} \cdot \left[(1 - {}_Y\pi(\boldsymbol{X}_{1i})) \cdot \frac{N}{N_A} \frac{n_A}{n_{0A}} \right]^{1-y_i} \cdot {}_A\pi(\boldsymbol{X}_{2i}) \right\}^{A_i}$$

where ${}_A\pi(\boldsymbol{X}_{2i})$ is the probability that an observation is uncensored and ${}_Y\pi(\boldsymbol{X}_{1i})$ is the probability of observing $Y = 1$ given that the observation is uncensored; as already said, n_A is the number of units sampled from the N_A uncensored observations and $n_{\bar{A}}$ is the number of units sampled from the $N_{\bar{A}}$ censored observations; n_{yA} is the amount of units in the sample having $Y = y$, with $y = 0, 1$.

It's easy to understand that the likelihood (6) is a weighted version of (4), and the weights simply take into account the sampling design. Note also that in the maximization process the term $f(\boldsymbol{X}_i|S_i = 1)$ is non influential, as it does not contain any information on the vector of parameters η, and that in our estimator the only quantities to be known at the population level are N_A and N.

References

1. Arezzo, M.F., Guagnano, G. Response-based sampling for binary choice models with sample selection. Working Paper 149, Department Memotef - Sapienza University of Rome (2017).
2. Cameron, A.C., Trivedi, P.K.: Microeconometrics: Methods and Applications. Cambridge University Press, New York (2005)
3. Heckman, J.J.: Sample selection bias as a specification error. Econometrica, 47(1): 153-162 (1979).
4. Hosmer, D.W., Lemeshow, S.: Applied Logistic Regression. John Wiley & Sons (2013)