The effect of microaggregation on regression results: An application to Spanish innovation data*

Alberto López†
Universidad Complutense de Madrid

March 2010

Abstract

Microaggregation is a technique for masking confidential data by aggregation. The aim of this paper is to analyze the extent to which microaggregated data can be used for rigorous empirical research. In doing this, I use data from the Technological Innovation Panel (PITEC) and compare regression results using the original and the anonymized data. PITEC is a new firm-level panel data base for innovative activities of Spanish firms. I find that the microaggregation procedure used has a slight effect on the coefficient estimates and their estimated standard errors. The main lesson that can be drawn from this empirical exercise is that the use of anonymized data from PITEC produces reliable results.

JEL Classification: C80; O30

*I acknowledge support from project SEJ2004-02525/ECON. Thanks is due to the INE for access to the data. Errors are mine.

†Departamento de Fundamentos del Análisis Económico I. Facultad de Ciencias Económicas y Empresariales. Universidad Complutense de Madrid. E-mail: alberto.lopez@ccee.ucm.es.

1. Introduction

Observing confidentiality is crucial when collecting data and providing individual level information. Statistical offices have to maintain confidentiality required by data protection laws and, at the same time, its reputation is at stake. On the other hand, researchers require access to individual micro data. The main method used to deal with this problem is the application of masking or anonymization procedures to data (which are also commonly referred to as disclosure control methods). These masking procedures modify the original data in a way that re-identification of individual respondents (i.e. individuals and firms) is almost impossible (or re-identification may still be possible but would involve high cost). At this point, a trade-off between quality of the data for analysis and confidentiality appears. In this sense, the higher the degree of anonymization applied to the data, the less the quality of the data for empirical analysis.

Literature on this topic has focussed on two issues. First, a large body of literature has focussed on the disclosure control[1]. These studies evaluate the validity of the different anonymization procedures to avoid disclosure of confidential information. Second, another strand of literature, which is less developed, analyzes the effect of anonymization procedures on estimation. The aim of these studies is to analyze the extent to which anonymized data can be used instead of the original data and how reliable estimates form anonymized data would be.

This paper is an empirical example to illustrate the effect of one of the most relevant anonymization procedure, microaggregation, on estimation results. In doing this, I use data from the Technological Innovation Panel (PITEC) and compare the results from estimating linear and non-linear models using the original data and the anonymized version. Thus, the final aim of this paper is to determine whether the use of anonymized data from PITEC produces reliable results.

The rest of the paper is organized as follows. Section 2 reviews the main findings on the effect of microaggregation on estimation results, both from a methodological and empirical

[1] See Willenborg and de Waal (2001) for a review.

perspective. Section 3 introduces the data used and Section 4 describes the anonymization process applied. Section 5 analyzes the extent to which anonymized data from PITEC can be used for rigorous empirical research. In doing this, I estimate two linear equations and one non-linear equation, using both the original and the anonymized data from PITEC. Finally, Section 5 concludes.

2. The effect of microaggregation by individual ranking

One of the most commonly used anonymization procedures is microaggregation[2]. Microaggregation is a technique for protecting individual data by aggregation (see Defays and Anwar (1998) for a detail description). In this section, I focus on microaggregation by individual ranking (IR)[3]. Microaggregation by IR is the anonymization procedure chosen by Eurostat, although it is used in combination with other disclosure control techniques designed for masking discrete data (see Eurostat 1996, 1999).

IR is an anonymization procedure for continuous data consisting on three steps: sorting, grouping and replacement with average values. As a first step, for each variable to be anonymized, the data records are ranked in decreasing (or increasing) order. Secondly, data records are grouped (usually the group size is 3 or 5). Finally, each original data value is replaced with its respective group mean. Note that this three-step procedure is applied to each variable to be anonymized.

Regarding the effect of microaggregation by IR on estimation results, it is necessary to distinguish between the effect on linear and no-linear models. In the first case, there exist theoretical (or methodological) analysis on this issue. Recently, Schmid and Schneeweiss (2009)[4] present a theoretical analysis of the effect of the microaggregation by IR on the estimation of arbitrary moments[5], and, hence, on the least squares estimation of a linear model

[2] See, e.g., Adam and Wortmann (1989) and Winkler (2004) for detailed reviews of the various anonymization procedures.

[3] See Schmid and Schneeweiss (2005) for a review of the different microaggregation techniques.

[4] This paper is a generalization of the results obtained in Schmid (2006) and Schmid and Schneeweiss (2008).

[5] Arbitrary moments include first, second, and product moments of the transformed and untransformed

in transformed variables. These authors study both consistency and asymptotic normality. First, they prove the consistency of the empirical moments computed from microaggregated data by IR. And, hence, they prove that any "method of moments" estimator is consistent if computed from the microaggregated data by IR. Moreover, mixed moments between a microaggregated continuous and a non-microaggregated discrete variable can also be consistently estimated. Second, these authors specify conditions and regularity assumptions under which the moments are asymptotically normal. Finally, they provide a simulation study on the theoretical results and an empirical example based on real data.

However, evidence on the effect of microaggregation by IR on no-linear model estimation is, to my knowledge, restricted to empirical examples[6]. For example, Mairesse and Mohnen (2001) compare the estimation results of a generalised tobit model using original and microaggregated data. In doing this, these authors use French CIS 2 data[7], and find that the estimates are "rather similar" whether they use the original or the microaggregated data.

3. The Technological Innovation Panel (PITEC)

The Technological Innovation Panel (PITEC) is a statistical instrument for studying the innovation activities of Spanish firms over time. The data base is being carried out by the INE (The National Statistics Institute), which counts on advice from a group of university researchers and the sponsorship of FECYT and Cotec. The data come from the Spanish Community Innovation Survey (CIS).

PITEC has two main advantages. First, it is designed as a panel survey and contains a huge amount of information related to the innovation activities of Spanish firms. This data base includes information for more than 450 variables and from 2003 to 2007, for the moment. Second, it is a free available data set. The data base is placed at the disposal of

data as special cases.

[6] There exists theoretical evidence on the effect of other anonymization procedure in the presence of nonlinear estimation techniques. For example, Ronning (2005) analyzes the effect of randomized response with respect to some binary dependent variable on the estimation of the probit model. While, Hausman et al (1998) focus on a more general framework under the heading "misclassification".

[7] Eurostat (1999) details the microaggregation process adopted by Eurostat for CIS 2.

researchers on the FECYT[8] web site. Except for the anonymization of a set of variables, the files available on the web site correspond with the "original" files in the hands of the INE. This anonymization is necessary in order to avoid the disclosure problem (i.e., the possibility of identifying firms through the data). This anonymization procedure is described in the next section.

For reasons of opportunity and viability, PITEC started with two samples with data from 2003: a sample of firms with 200 or more employees (sample of big firms, which represented 73% of all firms with 200 or more employees according to data from the DIRCE), and a sample of firms with intramural R&D expenditures. Given the improvements made by the INE in information on firms undertaking R&D activities, there were enlargements of the second sample in 2004 and 2005. Moreover, in 2004, a sample of firms with fewer than 200 employees, external R&D expenditure and no intramural R&D expenditure; and a representative sample of firms with fewer than 200 employees and no innovation expenditure were included.

4. Anonymization procedure applied at the PITEC

The anonymization procedure applied at the PITEC consists, mainly, in a microaggregation by IR. In what follows, I describe this method in detail. The anonymization procedure used implies four modifications:

1. Microaggregation by individual ranking (IR) of six quantitative variables (turnover, exports, investment, number of employees, innovation expenditures and number of R&D employees). IR procedure used slightly departs from that described in Section 2. In this sense, IR is applied using two different procedures for forming groups of observations.

Firstly, the data records are divided into groups according to the firm's industry. For each of the continuous variables mentioned above (and for each industry), the data records are ranked in decreasing order. Then, the arithmetic mean of the five highest observations is calculated. Finally, the value of each "top five" observation is replaced with its cluster

[8] http://sise.fecyt.es/sise-public-web/mostrarCarpetaEstudiosInformes.do.

mean. Note that this procedure is applied for each of the variables in question in each industry. If there are fewer than three firms with positive value of the variable in question in a given industry, this procedure is not applied.

Secondly, for each of the variables in question, the data records are ranked in decreasing order (without considering the records replaced in the previous procedure). Then, the observations are grouped by three and the value of each one is replaced with the cluster arithmetic mean. The last group or the last two groups may have four observations.

In summary, IR applied implies that the available variables are: (i) the mean of the five highest observations, after ranking the data in decreasing order and according to the firm's sector, or (ii) the mean of three or four consecutive observations, after ranking the data in decreasing order.

2. To replace the firm-level observations of the rest of the quantitative variables with the percentage value with respect to the microaggregated value. The variables related to innovation expenditures and R&D personnel are expressed in percentage values. Specifically, intramural R&D expenditures according to the nature of the spending, the source of funding and spending by region, R&D expenditures in biotechnology and the amount of research grants are given as a percentage of the intramural R&D expenditures; the external R&D expenditure by supplier is given as a percentage of external R&D expenditure; the expenditure for each innovation activity and the innovation expenditures by region are given as a percentage of the total innovation expenditure; R&D personnel by activity, by education and by region, and the number of research scholars are given as a percentage of total R&D personnel.

3. The firms' activity (4-digit NACE Code) is replaced with a 56-industry breakdown.

4. In order to avoid the disclosure problem, and considering the sample stratification, the data of a given number of firms has been censored: those firms belonging to an industry in which the number of firms is less than or equal to three, both in the sample and in the population. Once a firm is censored in a given year, it will be censored in previous and subsequent years.

5. The effect of the anonymization procedure applied at the PITEC

The aim of this section is to analyze the extent to which anonymized data from PITEC can be used for rigorous empirical research. In view of the anonymization method applied (consisting mainly in a microaggregation by IR) and the literature reviewed in Section 2, the expected estimation bias is small.

I present the estimation of two linear equations (a sales equation and a labour productivity equation) and one non-linear equation (an innovation cooperation equation), using both the original and the anonymized data from PITEC.

In the first equation, sales are assumed to be a linear function of size, exports, investment in equipment and innovation expenditures. Hence, sales equation can be expressed as follows:

$$\begin{aligned} \log(sales) &= \alpha_1 \log(size) + \alpha_2 \log(exports) + \alpha_3 \log(investment) + \\ &\quad \alpha_4 \log(innovation\ expenditures) + u_1 \end{aligned} \tag{1}$$

The second equation specifies labour productivity as a linear function of export intensity and technological innovation.

$$\log(labour\ productivity) = \beta_1 \log(export\ intensity) + \beta_2\ technological\ innovation + u_2 \tag{2}$$

Finally, I estimate the determinants of innovation cooperation by using the standard probit model. The third equation models the probability of innovation cooperation as a non-linear function depending on size, R&D intensity and a measure of cost factors as hampering factors for innovation.

$$\begin{aligned} P(innovation\ cooperation &= 1) = \Phi(\gamma_1 \log(size) + \gamma_2 \log(R\&D\ intensity) + \\ &\quad \gamma_3\ cost + u3) \end{aligned} \tag{3}$$

where Φ is the standard normal cdf.

In estimating equations (1), (2) and (3), I also include industry dummies[9] and a constant.

[9] Appendix B reports the details on the industry breakdown used to define industry dummies (52 industry dummies).

Moreover, equations (1) and (2) include a dummy for belonging to a group. Appendix A gives details on the variables employed.

In this empirical exercise, I use data from PITEC for the year 2005 and for manufacturing and service sectors. This gives a total sample of 11,160 firms.

Tables 1, 2 and 3 present the results for the estimation of equations (1), (2) and (3), respectively. In each table, estimate *a* presents the results using the original data, while estimate *b* shows the estimations using the anonymized data. All estimates have been rounded to three decimal places.

First, I focus on comparing the results using original and anonymized data. I find that the anonymization procedure used has a slight effect on the coefficient estimates of equations (1), (2) and (3) and their estimated standard errors. Regarding coefficient estimates, maximum aggregation bias arises in estimating a non-linear model (equation (3)). In particular, on the estimation of the effect of R&D intensity on innovation cooperation. Aggregation bias for estimated standard errors is smaller. In this sense, estimated standard errors, rounded to three decimal places, are exactly the same whether the original or the anonymized data is used. The main lesson that can be drawn from this exercise is that the use of anonymized data from PITEC produces reliable results.

Second, I briefly comment on the results obtained for the estimation of equations (1), (2) and (3). I estimate three simple equations explaining sales, labour productivity and innovation cooperation. However, results are consistent with the existing literature. Firstly, innovation expenditure has a positive effect on sales. Moreover, firm's size, exports and investment have the expected positive effect. Secondly, technological innovation and firm's export intensity are associated with higher labour productivity (see, for example, Crepon et al (1998) and Bernard and Jensen (1999), respectively). Thirdly, absorptive capacity of the firm (measured by firm's size and R&D intensity) and the importance of cost as a hampering factor for innovation are significant and positive determinants of innovation cooperation (see, for example, López (2008) for evidence from Spanish manufacturing firms).

6. Conclusions

There exist different techniques for masking confidential data One of the most commonly used anonymization procedures is microaggregation. Microaggregation is a technique for protecting individual data by aggregation. These masking procedures modify the original data in a way that re-identification of individual respondents is almost impossible. At this point, a question arises as to whether the use of anonymized data produces reliable results.

The aim of this paper is to analyze the extent to which microaggregated data can be used for rigorous empirical research. In doing this, I use data from the Technological Innovation Panel (PITEC) and compare regression results using the original and the anonymized data. In particular, I present the estimation of two linear equations (a sales equation and a labour productivity equation) and one non-linear equation (an innovation cooperation equation).

PITEC is a new firm-level panel data base for innovative activities of Spanish firms. This data base is placed at the disposal of researchers in a microaggregated form. The anonymization procedure applied at the PITEC consists, mainly, in a microaggregation by individual ranking.

Preliminary results show that the microaggregation procedure used has a slight effect on the coefficient estimates and their estimated standard errors. The main lesson that can be drawn from this empirical exercise is that the use of anonymized data from PITEC produces reliable results.

Appendix A: Definitions of Variables

Cost: Sum of the scores of importance of the following obstacles to innovation process (number between 1 (high) and 4 (not relevant)): Lack of funds within the firm or group; Lack of finance from sources outside the firm; Innovation costs too high. Rescaled between 0 (not relevant) and 1 (high).

Exports: Firm's total exports.

Export intensity: Ratio between exports and number of employees.

Group: Dummy variable that takes the value 1 if the firm belongs to a group.

Innovation cooperation: Variable which takes the value 1 if the firm cooperates on innovation activities with suppliers, customers, competitors, commercial laboratories/R&D enterprises, universities, or government or private non-profit research institutes.

Innovation expenditures: Total amount of expenditure in innovation activities.

Investment: Physical investment.

Labour productivity: Ratio between sales and number of employees.

R&D intensity: Ratio between intramural R&D expenditure and number of employees.

Sales: Firm's total turnover.

Size: Total number of employees.

Technological innovation: Dummy variable that takes the value 1 if the firm reports having introduced product or process innovations.

Appendix B: Industry definitions (NACE Code)

Food products and beverages (15)

Tobacco products (16)

Textiles (17)

Wearing apparel; dressing and dyeing of fur (18)

Leather and leather products (19)

Wood and wood products (20)

Pulp, paper and paper products (21)

Publishing, printing and reproduction of recorded media (22)

Coke, refined petroleum products and nuclear fuel (23)

Chemicals and chemical products (24, except 244)

Pharmaceuticals, medicinal chemicals and botanical products (244)

Rubber and plastic products (25)

Ceramic tiles and flags (263)

Other non-metallic mineral products (26, except 263)

Ferrous metals (271, 272, 273, 2751, 2752)

Non-ferrous metals (274, 2753, 2754)

Fabricated metal products, except machinery and equipment (28)

Machinery and equipment (29)

Office machinery and computers (30)

Electrical machinery and apparatus (31)

Manufacture of electronic valves and tubes and other electronic components (321)

Radio, television and communication equipment and apparatus (32, except 321)

Medical, precision and optical instruments, watches and clocks (33)

Motor vehicles, trailers and semi-trailers (34)

Building and repairing of ships and boats (351)

Aircraft and spacecraft (353)

Other transport equipment (35, except 351, 353)

Furniture (361)

Games and toys (365)

Manufacturing n.e.c. (36, except 361, 365)

Recycling (37)

Sale, maintenance and repair of motor vehicles and motorcycles; retail sale of automotive fuel (50)

Wholesale trade and commission trade, except of motor vehicles and motorcycles (51)

Retail trade; repair of personal and household goods (52)

Hotels and restaurants (55)

Transport (60, 61, 69)

Supporting and auxiliary transport activities; activities of travel agencies (63)

Post and courier activities (641)

Telecommunications (642)

Financial intermediation (65, 66, 67)

Real estate activities (70)

Renting of machinery and equipment without operator and of personal and household goods (71)

Software consultancy and supply (722)

Computer and related activities (72, except 722)

Research and development (73)

Architectural and engineering activities and related technical consultancy (742)

Technical testing and analysis (743)

Other business activities (74, except 742, 743)

Education (80, except 8030)

Motion picture and video activities (921)

Radio and television activities (922)

Other community, social and personal service activities (80, 85, 90, 91, 92, 93)

References

Adam, N. R. and Wortmann, J. C., (1989), "Security-Control Methods for Statistical Databases: A Comparative Study", *ACM Computing Surveys,* 21(4), 515-556.

Bernard, A.B.and Jensen, J.B., (1999), "Exceptional exporter performance: cause, effect, or both?", *Journal of International Economics,* 47 (1), 1–26.

Crepon, B., Duguet, E. and Mairesse, J., (1998), "Research and Development, Innovation and Productivity: An Econometric Analysis at the Firm Level", *Economics of Innovation and New Technology*, 7(2), 115-156.

Defays, D.and Anwar, M.N., (1998), "Masking microdata using microaggregation", *Journal of Official Statistics*, 14(4), 449-461.

Eurostat (1996), *Manual on Disclosure Control Methods*, 9E, Statistical Office of the European Communities, Luxembourg.

Eurostat (1999), "Annex II.9. Micro-Aggregation Process", in *The Second Community Innovation Survey*, Statistical Office of the European Communities, Luxembourg.

Hausman, J.A., Abrevaya, J.and Scott-Morton, F.M., (1998), "Misclassification of the dependent variable in a discrete-response setting", *Journal of Econometrics,* 87, 239–269.

López, A. (2008), "Determinants of R&D cooperation: Evidence from Spanish manufacturing firms", *International Journal of Industrial Organization*, 26, 113–136.

Mairesse, J.and Mohnen, P., (2001), "To be or not to be innovative: An exercise in measurement", *STI Review*, OECD 27, 103–129.

Ronning, G., (2005), "Randomized response and the binary probit model", *Economics Letters*, 86, 221-228.

Schmid, M., (2006), "Estimation of a linear model under microaggregation by individual ranking", *Journal of the German Statistical Society*, 90 (3), 419-438.

Schmid, M. and Schneeweiss, H., (2005). "The effect of microaggregation procedures on the estimation of linear models: A simulation study", In *Econometrics of Anonymized Micro Data* (W. Pohlmeier, G. Ronning, J. Wagner, eds.), Jahrbficher ffir National-Skonomie und Statistik, 225, No. 5, Lucius & Lucius, Stuttgart.

Schmid, M. and Schneeweiss, H., (2008), "Estimation of a linear model in transformed variables under microaggregation by individual ranking", *AStA Advances in Statistical Analysis*, 92 (4), 359-374.

Schmid, M. and Schneeweiss, H., (2009), "The effect of microaggregation by individual ranking on the estimation of moments", *Journal of Econometrics,* 153, 174 -182.

Willenborg, L. and de Waal, T., (2001), "Elements of Statistical Disclosure Control", *Springer Lecture Notes in Statistics*, vol. 155. Springer, Berlin.

Winkler, W. E., (2004), "Masking and Re-identification Methods for Public-Use Microdata: Overview and Research Problems", In *Proc. Privacy in Statistical Databases,* J. Domingo-Ferrer and V. Torra (Eds.): LNCS 3050, 231–246.

Table 1. Sales equation[1,2]

Dependent variable: Sales (in logs)

	(a) Original Data	(b) Anonymized data
Size (in logs)	0.963	0,958
	(0,009)	(0,009)
Exports (in logs)	0,033	0,033
	(0,001)	(0,001)
Investment (in logs)	0,009	0,010
	(0,002)	(0,002)
Innovation expenditures (in logs)	0,011	0,011
	(0,002)	(0,002)
Group	0,434	0,439
	(0,022)	(0,022)
R^2	0,809	0,809

[1]Robust standard errors between brackets.

[2]Industry dummies included.

Table 2. Labour productivity equation[1,2]
Dependent variable: Sales/Employees (in logs)

	(a) Original Data	(b) Anonymized data
Export intensity (in logs)	0,052	0,052
	(0,002)	(0,002)
Technological innovation	0,040	0,040
	(0,023)	(0,023)
Group	0,427	0,426
	(0,019)	(0,019)
R^2	0,326	0,325

[1]Robust standard errors between brackets.

[2]Industry dummies included.

Table 3. Innovation cooperation equation[1,2]

Dependent variable: Innovation cooperation (dummy variable)

	(a) Original Data	(b) Anonymized data
Size (in logs)	0,049	0,046
	(0,004)	(0,004)
R&D intensity (in logs)	0,030	0,021
	(0,002)	(0,002)
Cost	0,099	0,098
	(0,017)	(0,017)
pseudo-R^2	0,050	0,050

[1]Robust standard errors between brackets. The coefficients are the marginal effect of the independent variable on the probability of cooperation.

[2]Industry dummies included.