Data Integration Application with Coarsened Exact Matching

Mariana Kotzeva
Bulgarian NSI, University of National and World Economy, mkotzeva@nsi.bg
Roumen Vesselinov
Bulgarian NSI, Sofia University St. Kliment Ohridski, rvesselinov@nsi.bg

Abstract: This paper focused on the problem of integrating data from two distinct sources or groups for statistical analysis. The two groups could be from representative sample and business registers, or in relation to the non-response bias problem. We investigated the properties of some traditional techniques like propensity scores and simple regression and a more advanced method for coarsened exact matching. The main finding of the paper was the suggestion that most methods were comparable in simple cases of bias but in more complicated cases of bias the exact matching approach was superior.

Keywords: exact matching, data integration, generalized log-linear models

1. Introduction

The problem for integrating data from two (or more) distinct sources could be addressed by some of the already established statistical methods for weighting, propensity scores weighting and stratification (Rosenbaum and Rubin, 1983). On the other hand, some more recent new methods for exact matching had emerged (Iacus et al, 2011a, 2011b). It was our intention to test and compare the estimation properties of the traditional and more recent methods with data for Bulgaria.

2. Data Sources

The data used in this paper were from the Bulgarian register of enterprises for 2008. The variables included were as follows: Type of enterprise: 1= Sole proprietor 0 = Limited liability company or Partnership; Foreign ownership: Yes/No; Regions – 6 economic regions in Bulgaria; Labour: Number of employed; Economic sector: 1=Industry; 2=Services; 3=Agriculture; Revenue: in thousand Bulgarian leva, current prices; In-vestment: spending for capital assets, in thousand Bulgarian leva, current prices; also as binary Yes/No investment; and ratio of investment to revenue (limited to between 0 and 1). Indicator (Dummy) variables were created for the categorical variables whenever necessary.
From the population data were excluded enterprises with no employed, or no revenue, or with ratio of investment to revenue greater than one, or extremely large values of revenue or investment. A 5% random sample was drawn from the rest of the population. The final sample size was N=13851.
The classical interpretation (Rosenbaum and Rubin, 1983 and Rosenbaum, 2002) focuses the sample selection bias on the imbalance in the covariates between "Treatment" and "Control" groups. In this paper we treated the problem more broadly. Under "sample selection bias" we understood the problem of integrating data from two

sources (sample and register), or addressing the non-response bias (Matsuo et al, 2010). For this purpose we introduced a bias indicator variable (0/1) where 0 was interpreted as the sample data and 1 as the data from the register of enterprises. We worked with two types of bias, “random” and “non-random”. For the random bias we generated a random variable that assigned the cases (40% to 60% ratio) to the two groups (e.g. sample and register). For the non-random bias we assigned value of 1 to all enterprises with only 1 employed person and 0 for the rest.

3. Methodology

Three different types of models were considered: Model 1 : Regression model with Revenue as dependent variable and Labour as independent; Model 2: Logistic regression model with Investment (Y/N) as dependent variable and Labour as independent; Model 3: Zero-Inflated Poisson (ZIP) Model with Ratio of Investment/Revenue dependent on Labour (in thousands).
The ZIP model was specifically designed (Long, 1997 and Lambert, 1992) to handle count or rate (like in our case) variables with many zeroes. In our sample 71.3% did not have any investment. This is a type of generalized log-linear model or a mixture model with two classes: zero and non-zero. Voung (1989) proposed test to determine whether the ZIP model is to be preferred to the traditional Poisson model.
Four different methods for addressing sample selection bias were implemented in the paper: A. No weighting and no matching; B: Propensity score weighting; C: Propensity score stratification (5 strata); and D: Coarsened exact matching (CEM).
The propensity score methods involved first estimating a logistic regression model with the bias (0/1) as dependent variable and regions, type, foreign ownerships, and economic sector. The predicted values of the models were saved as propensity scores (PS). They were used in two ways, as weights (similar to Matsuo et al, 2010) and by creating 5 strata based on the PS quintiles as suggested by Rosenbaum and Rubin (1983).
CEM is a type of exact matching method which reduces the potential differences between the data from the two data sources (sample and register) by grouping or coarsening the data into bins and exact matching the data and then running the analysis on the matched data. This is type of monotonic imbalance bounding and it has very attractive statistical properties (Blackwell et al, 2009 and Iacus et al, 2011a, 2011b).

4. Results

The analysis was done separately for the random and non-random bias and for the three models using standard methods and the three methods for adjustment of the sample bias.

4.1. Results for Random Bias

This was the case where, some of the data were considered as collected by a survey and some from a register and there was no known pattern or bias related to the source of the data. The results for the random bias estimation are presented in Tables 1, 2 and 3.
For the regression model and the logistic regression model CEM worked as well as the other methods (see Table 1 and 2 respectively). For the ZIP model (Table 3) the PS stratification did not work well, while the other 3 worked similarly well. The conclusion

was that in the case of random bias the use of CEM did not gain much compared to the PS- based methods. The results were comparable.

Table 1: *Random Bias Estimation Results for Model 1.*

Method		Regression Coefficient	P-value	95% CI
A	No weighting and no matching	89.5	<.001	87.6-91.3
B	Propensity Score Weighting	88.7	<.001	85.8-91.5
C	Mean Propensity Score 5 Strata	97.3	<.001	93.7-100.9
D	Coarsened Exact Matching	91.2	<.001	89.4-93.1

Table 2: *Random Bias Estimation Results for Model 2.*

Method		Odds Ratio	P-value	95% CI
A	No weighting and no matching	1.16	<.001	1.15-1.17
B	Propensity Score Weighting	1.16	<.001	1.15-1.18
C	Mean Propensity Score 5 Strata	1.20	<.001	1.16-1.23
D	Coarsened Exact Matching	1.16	<.001	1.15-1.17

Table 3: *Random Bias Estimation Results for Model 3.*

Method		Incidence-Rate Ratio	P-value	95% CI
A	No weighting and no matching	1.69	0.009	1.14-2.51
B	Propensity Score Weighting	1.70	0.096	0.91-3.18
C	Mean Propensity Score 5 Strata*	3.64*	Range too wide.	Range too wide.
D	Coarsened Exact Matching	1.72	0.007	1.16-2.54

* Two extreme results excluded.

4.2 Results for Non-Random Bias

This was the case where, for example, some of the data were collected by a survey and some from a register and there was a known pattern to where the data came from. As in our experiment, the data for small enterprises (only 1 employed person) came only from register, while the data for larger enterprises (more than 1 employed) came from survey. The results for the non-random bias estimation are presented in Tables 4, 5 and 6. For the regression model, CEM showed very different results than the other 3 methods (see Table 4). The coefficient estimate and its 95% CI were below the range of the other methods. Theoretically the exact matching had some advantages over the PS methods so we were more likely to believe the CEM results. So in this case CEM did make a difference.

Table 4: *Non-Random Bias Estimation Results for Model 1.*

Method		Coefficient	P-value	95% CI
A	No weighting and no matching	89.5	<.001	87.6-91.3
B	Propensity Score Weighting	80.8	<.001	78.3-83.3
C	Mean Propensity Score 5 Strata	87.4	<.001	83.6-91.3
D	Coarsened Exact Matching	73.2	<.001	71.7-74.7

Table 5: *Non-Random Bias Estimation Results for Model 2.*

Method		Odds Ratio	P-value	95% CI
A	No weighting and no matching	1.16	<.001	1.15-1.17
B	Propensity Score Weighting	1.19	<.001	1.17-1.22
C	Mean Propensity Score 5 Strata	1.22	<.001	1.18-1.27
D	Coarsened Exact Matching	1.19	<.001	1.18-1.21

Table 6: *Non-Random Bias Estimation Results for Model 3.*

Method		Incidence-Rate Ratio	P-value	95% CI
A	No weighting and no matching	1.69	0.009	1.14-2.51
B	Propensity Score Weighting	1.72	0.118	0.87-3.41
C	Mean Propensity Score 5 Strata	1.37*	Range too wide.	Range too wide.
D	Coarsened Exact Matching	1.67	0.049	1.00-2.78

* Three extreme results excluded.

For the logistic regression model (Table 5) and the ZIP model (Table 6) all the methods except the PS stratification gave similar results.

5. Discussion

The results of this study showed that the theoretical advantages of the CEM and the class of exact matching methods were confirmed by the empirical results. CEM performed equally well as the PS methods and in some cases it gave very distinct results. More empirical work is needed, but in our opinion the exact matching methods for adjustment of sample bias and data integration deserve the attention of researchers and practitioners.

References

Blackwell, M., S. Iacus, G. King, G. Porro (2009) CEM: Coarsened exact matching in Stata, in *The Stata Journal*, Number 4: pp. 524-546.

Iacus, Stefano M., Gary King, and Giuseppe Porro (2011a) Causal Inference Without Balance Checking: Coarsened Exact Matching, in *Political Analysis*, 2011.

Iacus, Stefano M., Gary King, and Giuseppe Porro (2011b) Multivariate Matching Methods That are Monotonic Imbalance Bounding, in *Journal of the American Statistical Association*, 106 (2011): 345-361.

Lambert, D. (1992) Zero-inflated Poisson regression models with an application to defects in manufacturing, in *Technometrics*, Feb; 34(1): pp. 1-14.

Long, J. (1997) Regression Models for Categorical and Limited Dependent Variables. Thousand Oaks, CA: Sage Publications.

Matsuo, H., G. Loosveldt, J. Billiet, F. Berglund, O. Kleven, Measurement and adjustment of non-response bias based on non-response surveys: the case of Belgium and Norway in the European Social Survey Round 3, Survey Research Methods, 2010, Vol. 4, No.3, pp. 165-178.

Rosenbaum, P., D. Rubin (1983) The central role of the propensity score in observational studies for causal effects, in *Biometrika*, 70(1):41-55.

Rosenbaum, P. 2002 Observational Studies, 2nd ed.. NY: Springer-Verlag.

Vuong, Q. (1989) Likelihood Ratio Tests for model selection and non-nested hypotheses, in *Econometrica*, Mar; 57(2):307-333.