EXACT MATCHING FOR DATA INTEGRATION. EVIDENCE FROM BULGARIA.

Roumen Vesselinov[1], *Mariana Kotzeva*[2]

Keywords: exact matching, data integration, logistic regression, zero-inflated poisson model.

1. Data

The data are from the Bulgarian NSI's register of enterprises for 2008. The variables included are 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; Investment: 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.

Exclusions

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.

Sample Selection Bias

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 will treat the problem more broadly. Under "sample selection bias" we will understand 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 introduce a bias variable (0/1) where 0 may be interpreted as the sample data and 1 as the data from register. We work with two types of bias, "random" and "non-random". For the random bias we generate 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 assign value of 1 to all enterprises with only 1 employed person and 0 for the rest.

[1] National Statistical Institute of Bulgaria, 1038 Sofia, 2 P.Volov Str, rvesselinov@nsi.bg

[2] National Statistical Institute of Bulgaria, 1038 Sofia, 2 P.Volov Str, mkotzeva@nsi.bg

2. Methods and Models

Models

Three 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 is 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.

Methods

Four methods for addressing sample selection bias are 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 involve 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 are saved as propensity scores (PS). They are 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, in press).

3. Results

The analysis is 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.

Random Bias

This is the case where, for example, some of the data were collected by a survey and some from a register and there is no known pattern to where the data came from. 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 works as well as the other methods (see Table 1 and 2 respectively). For the ZIP model (Table 3) the PS stratification does not work well, while the other 3 work similarly well. The conclusion is that in the case of random bias the use of CEM does not gain much compared to the PS- based methods. The results are comparable.

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	Propensity Score 5 Strata (Average)	97.3	<.001	93.7-100.9
D	Coarsened Exact Matching	91.2	<.001	89.4-93.1

Table 1. Random Bias Estimation Results for Model 1.

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	Propensity Score 5 Strata (Average)	1.20	<.001	1.16-1.23
D	Coarsened Exact Matching	1.16	<.001	1.15-1.17

Table 2. Random Bias Estimation Results for Model 2.

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	Propensity Score 5 Strata (Average)*	3.64*	Range too wide.	Range too wide.
D	Coarsened Exact Matching	1.72	0.007	1.16-2.54

* Two extreme results excluded.

Table 3. Random Bias Estimation Results for Model 3.

Non-Random Bias

This is the case where, for example, some of the data were collected by a survey and some from a register and there is 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 shows very different results than the other 3 methods (see Table 4). The coefficient estimate and its 95% CI are below the range of the other methods. Theoretically the exact matching has some advantages over the PS methods so we are more likely to believe the CEM results. So in this case CEM does make a difference.

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

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	Propensity Score 5 Strata (Average)	87.4	<.001	83.6-91.3
D	Coarsened Exact Matching	73.2	<.001	71.7-74.7

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

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	Propensity Score 5 Strata (Average)	1.22	<.001	1.18-1.27
D	Coarsened Exact Matching	1.19	<.001	1.18-1.21

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

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	Propensity Score 5 Strata (Average)	1.37*	Range too wide.	Range too wide.
D	Coarsened Exact Matching	1.67	0.049	1.00-2.78

* Three extreme results excluded.

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

Conclusion

The results of this study show that the theoretical advantages of the CEM and the class of exact matching methods were strengthened 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.

Bibliography

Blackwell, M., S. Iacus, G. King, G. Porro, CEM: Coarsened exact matching in Stata, *The Stata Journal*, 2009, 9, Number 4, pp= 524-546.

Iacus, S., G. King, G. Porro, Causal Inference Without Balance Checking: Coarsened Exact Matching, *Political Analysis*, In Press.

Iacus, S., G. King, G. Porro, Multivariate Matching Methods that are Monotonic Imbalance Bounding, *Journal of the American Statistical Association*, In Press.

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

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

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 Eurpean Social Survey Round 3, *Survey Research Methods*, 2010, Vol. 4, No.3, pp. 165-178.

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

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

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